Statistics and Its Interface Volume 12 (2019) 123–133

Spectral clustering-based network community detection with node attributes*

Fengqin Tang†, Yuanyuan Wang, Jinxia Su, and Chunning Wang

Identifying communities is an important problem in network analysis. Various approaches have been proposed in the literature, but most of them either rely on the topological structure of the network or the node attributes, with few integrating both aspects. Here we propose a community detection approach based on spectral clustering combining information on both the network structure and node attributes (SpcSA). Some of the attributes may not describe the communities we are trying to detect correctly. These irrelevant attributes can add noise and lower the overall accuracy of community detection. To determine how much each attribute contributes to community detection, our method introduces a mechanism by which attribute weights can adjust themselves. We demonstrate the effectiveness of the proposed method through numerical simulation and with real-world data.

AMS 2000 subject classifications: Primary 62-07; secondary 68U20, 91D30.
Keywords and phrases: Spectral clustering, Community detection, Stochastic block model, Node attributes, Normalized mutual information.

1. INTRODUCTION

Network analysis has become increasingly popular in many fields, including social sciences, physics, computer science, statistics and so on. A vital problem in network analysis is community detection which aims to cluster nodes into distinct groups based on observed edges or node attributes. Community detection can uncover interactional relationships or similar properties between nodes in the network. Many methods have been proposed with different emphases on statistical models and criteria for measuring the strength of community detection. Relevant statistical models include latent position cluster model ([1]), the stochastic block model (SBM, [2]) and its extensions, such as valued edges ([3]), overlapping clusters ([4, 5, 6]) and degree heterogeneity ([7]). Examples of criteria include normalized cuts ([8]), modularity ([9, 10]), among many others. Most of the methods only focus on the structural information and the partitions achieve a cohesive internal structure. To learn more, readers can refer to [11, 12, 13], and the references therein.

Real-world data sets are usually obtained from multiple sources. Here we refer to two forms of data in real applications. One is an adjacency matrix A describing the structure of the network. If two nodes are structurally related, the corresponding entry in the matrix is nonzero. The other is a multidimensional vector X describing attributes. A structural-based network is often described by a graph consisting of objects as nodes and connections between nodes as edges. Throughout this paper, we consider a simple graph that does not contain self-loops or multiple edges, thus the adjacency matrix is symmetric and unweighted. Structural information describes the pairwise similarity between objects while their latent characteristics are called attributes. These two sources of information can be combined to uncover latent communities in a network. Therefore, an ideal approach would be to generate clusters that have a cohesive intra-cluster structure with homogeneous attributes.

Several works combine the two sources of information for community detection. [14] introduces an integrated K-means-Laplacian clustering method combining K-means clustering on data attributes and spectral clustering on pairwise relations. The contribution of the two sources of information in the clustering process is balanced by a trade-off parameter. [15] proposes covariate-assisted spectral clustering (CASC) where a covariate function is added to the regularized graph Laplacian weighted by a tuning parameter. [16] presents communities from the edge structure and node attributes, based on a generative model for networks with node attributes. They further assume that the network structure and the attributes are dependent and the attributes have binary values. [17] proposes a method called SA-Cluster which measures both structural and attribute similarities using a unified distance measure. SA-Cluster can automatically learn the degree to which structural and individual attributes contribute to network detection with the assumption that the attributes have binary values. [18] introduces a joint criterion for community detection (JCDC) with node attributes under the assumption that the structure is assortative, which could be seen as a covariate reweighted Newman-Girvan with a convex objective

*The project was sponsored by National Natural Science Foundation of China (11301236), Natural Science Foundation of the AnHui Higher Education Institutions of China (KJ2017A377, KJ2017A376) and Anhui Provincial Natural Science Foundation (1608085QG169).
†Corresponding author.

function. A model is defined to be assortative if the within cluster probability of linkage is larger than the probability of linking between clusters. Other related works include [19, 20, 21, 22, 23, 24, 25, 26], among many others.

The above literature has made various assumptions about the nature of the communities or attributes. For example, they assume that communities are always assortative or that the attributes have binary values ([18, 16]). A crucial problem in detecting communities based on both structural and attribute similarities is how to balance these two sources of information, since they may be independent, or even conflicting. A possible solution is to balance the two sources of information with tuning parameters ([14, 15]), but setting those parameters is difficult. Other recent efforts have been devoted to learning the contributions of node attributes to identifying communities from data, rather than using all available attributes in the same way, [16, 17, 18] are examples of efforts along this line.

Inspired by the existing works, we seek to integrate structural and attribute similarities by altering the edge weights based on the similarity of the corresponding node attributes and simultaneously learn the impact of different attributes. [18] offers an approach that is perhaps the closest to ours. They propose a joint community detection criterion by redefining the similarity between nodes i, j as $A_{ij}(\omega_n - e^{-\sum_l D_{ijl}\beta_l})$, where $A_{ij} = 1$ if there is a structural edge between nodes i, j and 0; otherwise, D_{ijl} is the distance between nodes i, j along the l-th attribute and β_l is the weight of the l-th attribute. The parameter ω_n controls the influence of attributes. Thus, the method requires users to infer the relative importance of the two sources of information in advance and to determine a proper value of ω_n for improving clustering quality. The method might not work well if an improper value is chosen. To solve this problem, we define similarity in (2) below and use the normalized cut as the objective function.

Due to its computational ease and competitive performance, spectral clustering has recently gained popularity in numerous fields ([27]). We propose a community detection approach based on spectral clustering with both structural information and node attributes (SpcSA). The proposed algorithm contains an additional parameter in the Gaussian kernel function which can be estimated by a simple method. The proposed method can also handle categorical attribute data and numerical attribute data. Because some attributes may be irrelevant and hence mask the underlying clustering structure, to improve clustering quality we embed in the spectral clustering a mechanism by which attribute weights can adjust themselves. The self-adjustment mechanism can automatically learn the degree of contribution of each node attribute. Our algorithm assigns higher values to the attributes having smaller variance within a cluster and to the attributes having larger variance between clusters. Experimental results show that our method is still effective in cases where one of the information sources is weak.

2. METHODOLOGY

2.1 Preliminaries

Let network (or graph) $G = (V, E)$ have vertices $V = \{1, 2, \cdots, n\}$ and edges $E \subseteq V \times V$. Let A be the adjacency matrix with $A_{ij} = 0$ if there is no structural relation between nodes i and j and $A_{ij} = 1$ otherwise. Assume that the number of clusters (or groups, or communities) K is known, and let n_k be the cluster size of the k-th group. We aim to partition n nodes into K groups with $\sum_{k=1}^{K} n_k = n$ excluding the overlap case. Let $E_1, E_2, \ldots, E_K$ be the estimated communities corresponding to labels $\vec{e} = \{e_1, e_2, \ldots, e_n\}$, i.e. $E_k := \{i : e_i = k\}$. Given the adjacency matrix A of the graph G, the goal of the normalized cut (Ncut) is to find a partition of the overall graph that minimizes the following function

$$\text{Ncut}(A) = \sum_{k=1}^{K} \frac{\sum_{i \in E_k} \sum_{j \notin E_k} A_{ij}}{\sum_{i \in E_k} \sum_{j=1}^{n} A_{ij}}. \tag{1}$$

Our goal is to use the criterion (1) by adding node attributes to improve the efficiency of community detection. The edges representing structural relations between nodes are weighted according to the attribute similarities between their end nodes, thus reducing the original unweighted graph to a weighted graph. One cannot expect all attributes to be useful for detecting the communities we are interested in. Some attributes play significant roles in specifying groups while others do not. Therefore, we can improve the quality of community detection by giving more weight to relevant attributes and giving less weight to irrelevant attributes in the process of identifying communities. The larger the weight of an attribute is, the greater its contribution to the clustering process.

2.2 Methods

We reconstruct the adjacency matrix by considering attributes information as $W = (w_{ij})$, $w_{ij} = f(A_{ij}, X_i, X_j)$, where $f(\cdot)$ is a nonnegative function and $X_i \in \mathbb{R}^m$ is the attribute vector of node i. Specifically, we set

$$w_{ij} = A_{ij} \exp\Big\{ - \frac{\sum_{l=1}^{m} \beta_l D_{ijl}}{2\sigma^2} \Big\}, \tag{2}$$

where $\vec{\beta} = \{\beta_1, \ldots, \beta_m\}$ is a vector that defines the weights of different attributes, D_{ijl} is the distance between nodes i and j along the l-th attribute, $D_{ij} = (D_{ij1}, D_{ij2}, \ldots, D_{ijm})$, and σ is a parameter in the Gaussian similarity function controlling the width of the neighborhoods. The degree matrix T is defined as the diagonal matrix with the (i, i)-th element

Algorithm 1 SpcSA Algorithm

Input: Adjacency matrix A, node attributes X, a parameter σ of Gaussian function, cluster number K.

Output: K clusters $E_1, \ldots, E_K$ and the attribute weights $\vec{\beta} = (\beta_1, \ldots, \beta_m)$.

Algorithm:

1: Initialize $\vec{\beta}^{(0)}$ to compute $W^{(0)}$; for $t = 1, 2, \ldots$;
2: Compute the normalized Laplacian matrix $L^{(t)} = (T^{-1/2})^{(t-1)} W^{(t-1)} (T^{-1/2})^{(t-1)}$;
3: Compute K eigenvectors $u_1^{(t)}, \cdots, u_K^{(t)}$ of $L^{(t)}$ corresponding to the K eigenvalues of $L^{(t)}$ that are largest in absolute value. Put them in a matrix $U = [u_1^{(t)}, \ldots, u_K^{(t)}] \in \mathbb{R}^{n \times K}$;
4: Normalize each row of U to 1 and put the normalized version in an $n \times K$ matrix $\widehat{U}$;
5: Determine $E^{(t)} = [E_1^{(t)}, \ldots, E_K^{(t)}]$ by performing K-means on the rows of $\widehat{U}$;
6: Optimize $\vec{\beta}^{(t)}$ using equations (5), (6), (7) based on $E^{(t)}$;
7: Repeat steps 3-6 until (3) converges.

$T_{ii} = \sum_{k=1}^{n} w_{ik}, 1 \le i \le n$. We rewrite (1) as

$$\text{Ncut}(\vec{e}, \vec{\beta}) = \sum_{k=1}^{K} \frac{\sum_{i \in E_k} \sum_{j \notin E_k} w_{ij}}{\sum_{i \in E_k} \sum_{j=1}^{n} w_{ij}}. \tag{3}$$

Our goal is to minimize (3) over both cluster labels $\vec{e}$ and attribute weights $\vec{\beta}$. Here we use the Gaussian similarity function to describe the attribute similarities and w_{ij} has a multiplicative form. Other similarity functions, such as the logit exponential function, perform similarly in the simulations. Meanwhile, some other forms for integrating structural and attribute information, such as $w_{ij} = A_{ij} + \alpha g(X_i, X_j)$, where $g(\cdot)$ is the attribute similarity function and α is a tuning parameter, will be discussed in a future study.

2.3 Algorithm

The following subproblem optimizes objective function (3) over cluster labels $\vec{e}$ and attribute weights $\vec{\beta}$:

- Problem E: assign fixed weights $\hat{\vec{\beta}}$ to optimize the reduced problem $\text{Ncut}(\vec{e}, \hat{\vec{\beta}})$;
- Problem $\vec{\beta}$: assign fixed cluster labels $\hat{\vec{e}}$ to optimize the reduced problem $\text{Ncut}(\hat{\vec{e}}, \vec{\beta})$.

Once the weights are fixed, the weighted matrix $W = (w_{ij})$ is computed from (2) immediately. We then apply standard spectral clustering to W and the details are illustrated in Algorithm 1. Notice that $\text{Ncut}(\vec{e}, \vec{\beta})$ is a non-convex function of $\{\vec{\beta}\}$. In order to obtain the desired result, we need either to search for an alternative surrogate criterion to (3) or assign a sequence of good initial values of $\{\vec{\beta}\}$. Unfortunately, both strategies are difficult to implement.

The goal of cluster analysis is to partition n data objects into subgroups such that those objects in the same group are more similar to each other than they are to objects in other groups. However, it is unrealistic to require all attributes to simultaneously achieve minimum distance within clusters and maximum distance between clusters. In addition, certain attributes are more relevant than others to clustering. Inspired by this, we adopt an adaptive weight-adjusting method to measure the relative importance of attributes. Denote by $w_l = \sum_{k=1}^{K} \sum_{i \in E_k} \sum_{j \in E_k} D_{ijl} A_{ij}$ the total distance between the connected nodes within a cluster with respect to the l-th attribute and denote by $b_l = \sum_{k=1}^{K} \sum_{i \in E_k} \sum_{j \notin E_k} D_{ijl} A_{ij}$ the total distance between the connected nodes in different clusters along the l-th attribute. Thus, the goal of the subproblem $\vec{\beta}$ is to maximize the following function

$$N(\beta) = \frac{\sum_{l=1}^{m} \beta_l b_l}{\sum_{l=1}^{m} \beta_l w_l} \tag{4}$$

subject to

$$\begin{cases} \sum_{l=1}^{m} \beta_l = 1 \\ \beta_l \ge 0 \qquad \text{for } l = 1, \ldots, m. \end{cases}$$

As is pointed out by [28], the maximization for (4) is $\beta_l = 1$ if $\frac{b_l}{w_l} > \frac{b_i}{w_i}$ for any $i \ne l, 1 \le i, l \le m$ and 0 otherwise. The solution shows that the clustering quality is improved only by the most important attribute, even though other attributes may be informative. For example, in an online social network, each user is characterized by his/her age, gender, ethnic background and so on. Some of these attributes are correlated with the communities we are interested in. Thus, naturally we would set $\{\vec{\beta}\}$ unequal values to further define the relative influence of various attributes. [28] introduces an adaptive weight-adjusting approach, that is, $\beta_l^{(t+1)}$ depends on $\beta_l^{(t)}$ and an adjustment margin $\overline{\Delta\beta}_l^{(t)}$, where $\beta_l^{(s)}$ is the weight value at the s-th iteration. Recall that $\sum_{l=1}^{m} \beta_l^{(t)} = 1, \beta_l^{(t)} \ge 0, t \ge 1$. Simple calculations yield

$$\beta_l^{(t+1)} = \frac{1}{2}(\beta_l^{(t)} + \overline{\Delta\beta}_l^{(t)}), \tag{5}$$

where

$$\overline{\Delta\beta}_l^{(t)} = \frac{\Delta\beta_l^{(t)}}{\sum_{l=1}^{m} \Delta\beta_l^{(t)}}. \tag{6}$$

With respect to the contribution of the l-th attribute, the adjustment margin $\Delta\beta_l^{(t)}$ is estimated from

$$\Delta\beta_l^{(t)} = \frac{b_l^{(t)}}{w_l^{(t)}}. \tag{7}$$

By (5), if $\overline{\Delta\beta}_l^{(t)} > \beta_l^{(t)}$, then $\beta_l^{(t+1)} > \beta_l^{(t)}$, i.e., $\beta_l^{(t+1)}$ tends to be larger than $\beta_l^{(t)}$ as $\overline{\Delta\beta}_l^{(t)}$ increases. It can be seen from (6) and (7) that a large discrepancy between the between-cluster distance and the within-cluster distance along the l-th attribute indicates a large $\overline{\Delta\beta}_l^{(t)}$. The objective function is iteratively improved by applying this weight-adjustment method and spectral clustering. Since the relevance of attributes is not known in advance, we initialize all attributes with equal weights, i.e., $\beta_l^{(0)} = \frac{1}{m}, 1 \le l \le m$. Experiments show that the results of the proposed algorithm are insensitive to the values of $\beta_l^{(0)}, 1 \le l \le m$. The pseudo-code of the SpcSA algorithm is illustrated in Algorithm 1.

2.4 Selection of the parameter σ

We apply the Gaussian function to compute the attribute similarity between nodes in the network. The Gaussian similarity function models local neighborhoods by adjusting the values of σ. An appropriate σ must be chosen. Most of the data points should have large weights connecting their local neighborhoods while positive but negligible weights connecting the nodes far away. Unfortunately, it is unknown how appropriate values of σ may be found. [29] shows several rules of thumb which are frequently used to choose σ. One of these is to equate σ to the length of the longest edge in a minimal spanning tree of the fully connected graph constructed by node attributes X. We use the minimal spanning tree heuristic to determine σ in the experiments on both simulated networks and real-world networks. Note that there are other rules of thumb for choosing σ but we find the proposed method effective.

Let us illustrate the performance of SpcSA against different values of σ via the following experiment. We generate a degree-corrected stochastic block model with two clusters, where one cluster has 100 nodes and the other has 50 nodes. The probability of an edge lying between nodes i and j within the same cluster is $\theta_i\theta_j p$ and that of an edge lying between node i in one cluster and node j in another cluster is $\theta_i\theta_j vp$, $0 < v < 1$. We set 95% of the nodes in each cluster with the degree-corrected parameter $\theta_i = 1$ and 5% of the nodes with $\theta_i = 10$. We generate four attributes for each node. The first two are generated from a multivariate normal distribution $N(\vec{u}, \mathbf{I})$ for one cluster and from $N(-\vec{u}, \mathbf{I})$ for the other cluster, where $\vec{u} = (u, u+0.5)$ and $\mathbf{I}$ is an identity matrix. The remaining attributes are generated from the uniform distribution. We set $u \in [0.2, 0.8]$, $p = 0.1$ and $v = 0.5$.

We measure the performance of our method by normalized mutual information (NMI) as σ varies. NMI considers

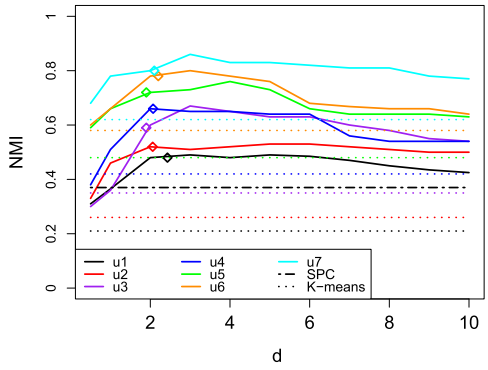


Figure 1. Estimation accuracy measured by NMI as a function of $d, d = 2\sigma^2$. The seven graph lines show the average NMI values of SpcSA for networks with $u1 = 0.2, u2 = 0.3, u3 = 0.4, u4 = 0.5, u5 = 0.6, u6 = 0.7, u7 = 0.8$ (bottom to top). The dashed line corresponds to SPC, which does not depend on d or $ui (i = 1, \ldots, 7)$. The seven dotted lines correspond to K-means. Their colors match those of the graph lines. The values of $\diamond$ correspond to the d values obtained using the minimal spanning tree method.

information-theoretic concepts and is based on the fact that if two clusters are similar to each other, then only a small amount of additional information is needed to infer one clustering assignment from the other. NMI is defined as

$$\mathrm{NMI} = \frac{-2\sum_{i=1}^{|C_1|}\sum_{j=1}^{|C_2|} M_{ij}\log(\frac{nM_{ij}}{M_{i;}M_{;j}})}{\sum_{i=1}^{|C_1|} M_{i;}\log(\frac{M_{i;}}{n}) + \sum_{j=1}^{|C_2|} M_{;j}\log(\frac{M_{;j}}{n})}$$

where n is the number of nodes and $|C_1|, |C_2|$ represent the size of the ground-truth clusters and the number of estimated clusters respectively. M_{ij} indicates the number of nodes belonging to community i that are estimated to be in community j, $M_{i;}$ denotes the number of nodes in the ground-truth community i and $M_{;j}$ indicates the number of nodes in the estimated community j. The larger the NMI value is, the higher the clustering quality of an algorithm. When the NMI measure takes its maximum value of one, the estimated results are identical to the ground truth. When NMI = 0, the results completely disagree with the ground truth. Note that the NMI measure can also be applied to compare two sets of estimated clustering results and not necessarily a comparison with ground truth data. [30] and [31] present related work.

Denote $d = 2\sigma^2$. For each configuration, the results are averaged over 50 replications presented in Figure 1. Since SPC uses only the network information, the performance of SPC does not rely on u or d, i.e., the corresponding average NMI values are constant. The performance of K-means uses node attributes and depends on u but not d, thus there are

seven dotted lines of different colors representing different u-corresponding NMI values. For simplicity, denote by d^* the corresponding d value obtained by the minimal spanning tree method. From the plot, we can see that for most $ui(i = 1, \ldots, 7)$, the NMI values returned by SpcSA at d^*($\diamond$ in the plot) are the highest. In the cases of $u3 = 0.4$ and $u7 = 0.8$, the highest NMI values of SpcSA corresponding to d values are in neighborhood of d^*, not far from d^*. In general, SpcSA performs well as σ is chosen using the minimal spanning tree heuristic described above. In the future, we will theoretically justify the approaches to choosing the parameter.

2.5 Consistency analysis

We develop an iterative algorithm to solve the subproblems E and $\vec{\beta}$. The attribute weights adjust themselves in such a way that reduces the objective function (3). We briefly illustrate this property as follows: Given nodes i, j, k, assume $A_{ij} = A_{ik} = 1$ and that each node has two attributes. In the $(t-1)$-th iteration, if $w_{ij}^{(t-1)} > w_{ik}^{(t-1)}$, i.e.,

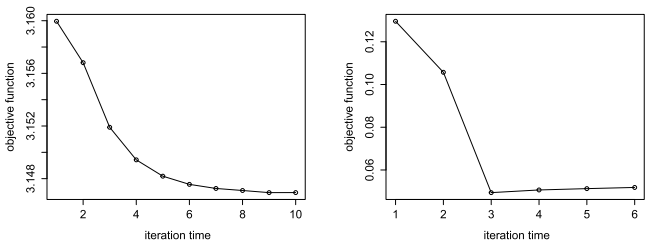
$$\frac{w_{ij}^{(t-1)}}{w_{ik}^{(t-1)}} = \exp\Big\{ -\frac{1}{2\sigma^2}\Big(\beta_1^{(t-1)}(D_{ij1} - D_{ik1}) + \beta_2^{(t-1)}(D_{ij2} - D_{ik2})\Big)\Big\} > 1,$$

then,

$$\beta_1^{(t-1)}(D_{ij1} - D_{ik1}) + \beta_2^{(t-1)}(D_{ij2} - D_{ik2}) < 0. \quad (8)$$

If the first attribute simultaneously possesses large between-cluster distance and low within-cluster distance in the t-th iteration, its weight will be increased, i.e., $\beta_1^{(t)} > \beta_1^{(t-1)}$. Since in an iteration the weights satisfy $\sum_l \beta_l^{(t)} = 1$, some of the attribute weights are increased while others decrease. Therefore, $\beta_2^{(t)} < \beta_2^{(t-1)}$. In general, the distribution of the small-weighted attribute is somewhat similar to the uniform distribution. The difference between D_{ij1} and D_{ik1} is more pronounced compared to the difference between D_{ij2} and D_{ik2}. It is more reasonable to assume $D_{ij1} - D_{ik1} < 0$ rather than $D_{ij1} - D_{ik1} > 0$, because the latter conflicts with (8). It follows from (8) that $D_{ij2} - D_{ik2} < -\frac{\beta_1^{(t-1)}(D_{ij1} - D_{ik1})}{\beta_2^{(t-1)}}$. Then

$$\begin{aligned}\frac{w_{ij}^{(t)}/w_{ik}^{(t)}}{w_{ij}^{(t-1)}/w_{ik}^{(t-1)}} &= \exp\Big\{ -\frac{1}{2\sigma^2}\Big((\beta_1^{(t)} - \beta_1^{(t-1)})(D_{ij1} - D_{ik1}) + (\beta_2^{(t)} - \beta_2^{(t-1)})(D_{ij2} - D_{ik2})\Big)\Big\} \\ &\geq \exp\Big\{ -\frac{1}{2\sigma^2}(D_{ij1} - D_{ik1})\Big(\beta_1^{(t)} - \frac{\beta_2^{(t)}}{\beta_2^{(t-1)}}\beta_1^{(t-1)}\Big)\Big\} \\ &> 1. \quad (9)\end{aligned}$$



(a) World trade network (b) Lawyer friendship network

Figure 2. Cluster Convergence.

Equation (9) implies that, compared with the $(t-1)$-th iteration, nodes i, j are more likely to be grouped together than nodes i, k in the t-th iteration. In summary, we can conclude that the objective function (3) converges when weights self-adjust.

The convergence properties of spectral clustering have been studied by many including [27, 32, 33]. The consistency of our proposed method follows from [27]. We apply the proposed algorithm on two real-world data sets to illustrate the consistency issue. Figures 2(a) and 2(b) show the clustering convergence of the objective function in the world trade network and the lawyer friendship network, respectively. Both plots show that the clustering objective function keeps decreasing when the proposed algorithm alternatively optimizes over the cluster labels and the attribute weights. The objective function converges very quickly, usually in six to ten iterations. The implementation of comparable clustering algorithms on these two data sets will be illustrated in the next section.

3. EMPIRICAL STUDIES

3.1 Numerical simulation

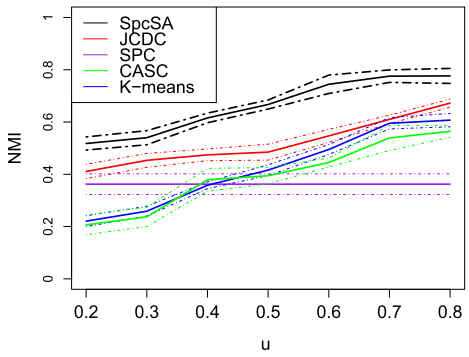
Here we evaluate the performance of the SpcSA algorithm compared with four other algorithms CASC, JCDC, spectral clustering (SPC) and K-means–using both synthetic data sets and real-life data sets. SpcSA, JCDC and CASC all use both structural and node attribute information. SPC and K-means either depends on the structural adjacency matrix or node attributes.

3.1.1 Simulation 1

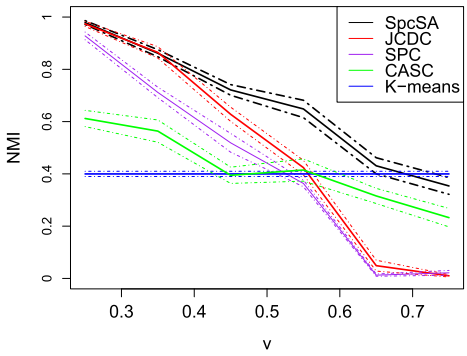
The first simulation investigates the sensitivity of all methods to attribute information, given structural information of the network. We generate the node attributes and a degree-corrected stochastic block model in the same setting as that in Section 2.4 with $v = 0.5, p = 0.1$. Each node has four attributes where the first two attributes are generated from a multivariate normal distribution $N(\vec{u}, \mathbf{I})$ in one cluster and $N(-\vec{u}, \mathbf{I})$ in the other cluster, and the remaining two attributes obey uniform distribution for all nodes. Here, $\vec{u} = (u, u + 0.5)$ and $u \in [0.2, 0.8]$, and the diagonal

Table 1. Notation in simulations 1-3

	Simulation 1	Simulation 2	Simulation 3
cluster number	2	2	2
cluster sizes	$(100, 50)$	$(100, 50)$	$(100, 100)$
attribute number	4	4	2
u	$[0.2, 0.8]$	1	/
v	0.5	$[0.25, 0.75]$	/
p	0.1	0.1	/



(a) NMI vs. attribute information

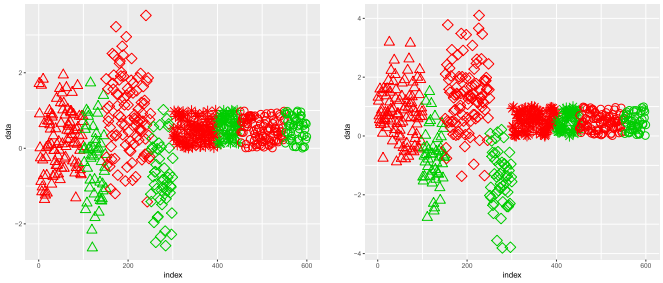


(b) NMI vs. structural information

Figure 3. (a) NMI values of different methods as u grows, where dashed lines represent the corresponding confidence bands. (b) NMI values of different methods as v grows, where dashed lines represent the corresponding confidence bands.

covariance matrix $\mathbf{I}$ is an identity matrix. The smaller the value of u is, the smaller the contribution of an attribute to identifying groups. We summarize the notation used in 3.1.1-3.1.3 in Table 1.

We plot the average NMI values for all methods from 50 random runs against an increasing u in Figure 3(a) as well as the 95% confidence intervals. The average NMI value of SPC is constant because the algorithm only relies on the adjacency matrix. As u increases from 0.2 to 0.8, the contributions of the two relevant attributes increase as well in the clustering process. Since K-means uses only the attributes,



(a) Node attribute distribution with $u = 0.3$ (b) Node attribute distribution with $u = 0.8$

Figure 4. (a) Node attribute distribution with $u = 0.3$, where red indicates the first cluster and green indicates the second cluster; the symbols represent different node attributes: "△" is the first attribute, "◇" is the second attribute, "∗" is the third attribute and "○" is the fourth attribute. (b) Data distribution with $u = 0.8$, where colors and shapes match those in (a).

Table 2. Clustering performance of different methods with $u = 0.3$

Methods	NMI	Attribute weights
SpcSA	0.63	(0.25,0.50,0.13,0.12)
JCDC	0.45	E_1:(0.002,0.004,0,0)
		E_2:(0, 0.09, 0, 0)
CASC	0.2	$\omega = 0.005$
K-means	0.25	(0.25, 0.25, 0.25, 0.25)
SPC	0.38	/

the average NMI values of K-means increase obviously as u grows.

We further observe that the attributes play different roles in clustering for the different normal distributions from which they are generated. To illustrate this, given $u = 0.3$ and $u = 0.8$, the data distribution of the four attributes are shown in Figures 4(a) and 4(b). Furthermore, for each fixed u, we record the average attribute weights of different methods. To conserve space, we only show the results of two instances where $u = 0.3$ and $u = 0.8$ in Tables 2 and 3, respectively.

By Figure 4, we can see that no matter what value u takes, the third and fourth node attributes do not provide judgment information for identifying communities. The attribute weights estimated by SpcSA reflect this fact, as shown in Tables 2 and 3 (i.e., (0.13, 0.12) in Table 2 and (0.10, 0.10) in Table 3). It should be pointed out that our algorithm cannot handle sparse cases, and because the sum of weights is 1, the irrelevant attribute weights are generally greater than zero. As shown in Figures 4(a) and 4(b), both the first attribute and the second attribute contain useful information for identifying communities, but the first relevant attribute does not contribute as much as the second relevant attribute does. This is consistent with the estimated

Table 3. Clustering performance of different methods with $u = 0.8$

Methods	NMI	Attribute weights
SpcSA	0.85	(0.28,0.52,0.10,0.10)
JCDC	0.65	E_1:(0.08, 0.59, 0, 0)
		E_2:(0, 0.07, 0, 0)
CASC	0.6	$\omega = 0.001$
K-means	0.51	(0.25, 0.25, 0.25, 0.25)
SPC	0.38	/

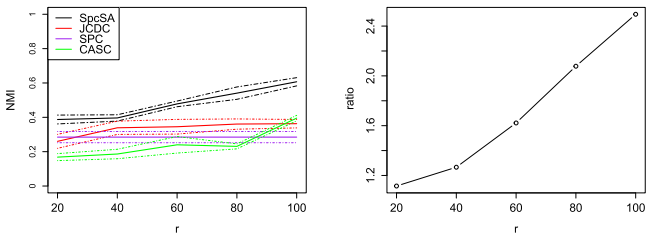
attribute weights returned by SpcSA in Tables 2 and 3 (i.e., (0.25, 0.50) in Table 2 and (0.28, 0.52) in Table 3).

Note that JCDC optimizes over the weights of attributes by the gradient ascent method and assigns relatively small weights to all attributes as the attribute information is not strong enough (Table 2 with $u = 0.3$). CASC balances the structural information and the node attributes by a tuning parameter ω and treats the degree of contributions of the attributes equally. Thus, due to the two irrelevant attributes, the results of CASC are unsatisfactory as the attribute information is weak. Overall, the confidence bands imply that SpcSA is stable as u varies and the average NMI values imply that SpcSA outperforms other methods even in cases where one of the data sources is uninformative or both of them are uninformative.

3.1.2 Simulation 2

The second simulation investigates the performance of five methods as measured by the strength of structural information, given fixed node attributes. A two-cluster network structure is generated similar to the first simulation with $u = 1, p = 0.1$ and $v \in [0.25, 0.75]$. The smaller the value of v is, the stronger the structural information becomes. We use NMI to measure the effectiveness of these algorithms as well. We plot the average NMI values for all methods from 50 random runs against an increasing v in Figure 3(b) as well as the 95% confidence intervals.

Since K-means only relies on attributes, its average NMI values are constant as v varies. Spectral clustering depends on the adjacency matrix, thus its average NMI values vary dramatically. CASC improves spectral clustering by using attribute information, the performance of CASC is not bad when the structural information is weak. On the other hand, since CASC cannot detect the degree of contribution of each attribute, due to the two irrelevant attributes, the performance of CASC is not prominent when the structural information is strong. JCDC assigns different weights to node attributes in different clusters. The algorithm standardizes all similarity measures along each attribute and truncates the weights of all attributes to their positive half. JCDC might not be able to highlight relevant attributes via attribute weights as shown in Tables 2 and 3. Figure 3(b) indicates that SpcSA gives a higher average NMI than other clustering algorithms and performs stably as v varies in terms of the confidence bands.



(a) (b)

Figure 5. (a) Performance of different methods as r varies. The black, red and green lines show the average NMI values of SpcSA, JCDC and CASC respectively as r grows, whereas the horizontal line shows the average NMI value of SPC. The dotted lines represent the corresponding confidence bands. (b) Ratio of the weight of the relevant attribute to the weight of the irrelevant attribute returned by SpcSA against increasing r values.

3.1.3 Simulation 3

The first two simulations, which use synthetic data sets, assume that the network structure and the node attributes share common community labels. In this part, we analyze the performance of three different methods (SpcSA, JCDC and CASC) under the assumption that the community labels for the network structure are not completely the same as the labels generating the node attributes. We create the network using the standard stochastic block model with 200 nodes. Assume that the network has two clusters of equal size, and the within-group and between-group edges occur with probabilities of 0.1 and 0.05 respectively. For each node, we generate two attributes, where one is a relevant attribute and the other is an irrelevant attribute whose distribution is the same for all nodes. For the relevant attribute, the node attributes follow $N(1, 1)$ in the first community, and r node attributes follow $N(-1, 1)$ and $100-r$ node attributes follow the uniform distribution in the second community. We vary r from 20 to 100 and investigate the performance of three different methods (SpcSA, JCDC and CASC). Small values of r imply that the relevant attribute is weakly correlated with the desired division. For each configuration, we repeat the simulation 50 times and report the results in Figure 5(a).

When the connection between the relevant attribute and communities is stronger, i.e., the attributes contain useful information for clustering, the performance of SpcSA is enhanced by using the attributes. Since the community structure of the network is not very informative, even though the attributes and the community structure strongly agree with each other, the clustering quality of JCDC does not improve by much, and CASC performs similarly.

Figure 5(b) shows the ratio of the weight of the relevant attribute to the weight of the irrelevant attribute returned by SpcSA as r grows, i.e., $ratio = \frac{the\ weight\ of\ the\ relevant\ attribute}{the\ weight\ of\ the\ irrelevant\ attribute}$. When r takes a small

Table 4. NMI (cluster number) values returned by comparing the detected partition of SPC to three attributes, respectively

Attributes	Position in 1994	Position in 1980	Continent
NMI	0.07(3)	0.17(4)	0.47(5)

value, the relevant attribute is useless for detecting communities, thus its weight is close to the irrelevant one. As r approaches 100, the relevant attribute plays a more important role in the clustering process than the irrelevant attribute does, thus the ratio increases. In short, our method automatically combines the available information from the network structure and node attributes to detect communities more efficiently than any algorithm based on either the network structure or node attributes alone, as well as than certain detection methods combining both sources of information.

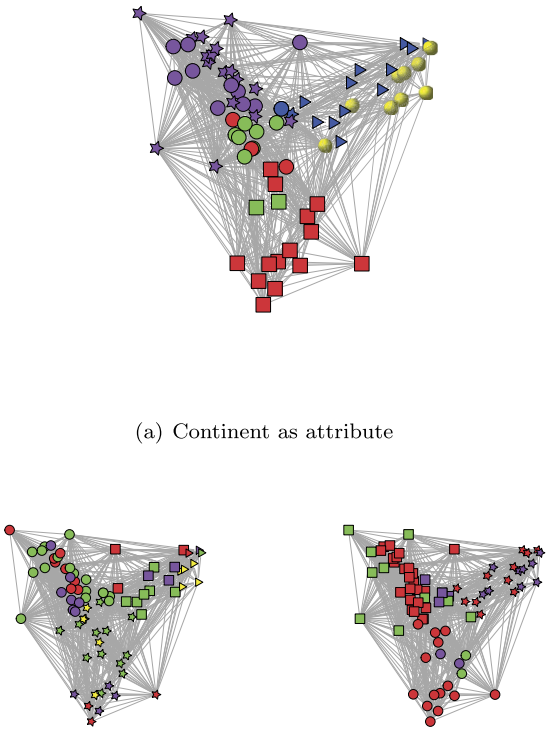
3.2 Real-life data sets

We use two real network data sets for evaluation in the following experiments. For most real networks, the ground-truth communities are typically unknown. This allows the networks to be divided in more ways than one ([23]). In practice, it is common to use some of the observed node attributes as ground-truth communities. In the following, we demonstrate that the SpcSA algorithm can use the representative attributes effectively to steer the analysis toward a desired direction. Simultaneously, the proposed algorithm can reveal which attributes are relevant and which attributes are irrelevant. We also perform a comparison with JCDC, CASC, SPC and K-means.

3.2.1 The world trade network

The data set comes from [34] and describes the world trade network among 80 countries. Each country represents a node in the network, and there is an edge between two countries if and only if the absolute amount of imports satisfies some mild conditions. Each node has several attributes, including the continent (Africa, Asia, Europe, North America, South America, and Oceania), the world-system position of the countries in 1980 (core, strong semi-periphery, weak semi-periphery, and periphery) and 1994 (core, semi-periphery, and periphery). Since the true labels of the clusters are unknown, we apply SPC to the network adjacency matrix and choose the continents, position in 1980 and position in 1994 as reference points, respectively. The corresponding cluster numbers of spectral clustering are $K = 5, 4$ and 3. Note that the three Oceania countries are omitted because of the masked structural information. The estimated results are shown in Table 4.

From Table 4, partitions by continent show an apparent community structure, thus we choose the continents as reference labels for comparisons. To illustrate this more clearly, we apply SpcSA to detect communities of the network that

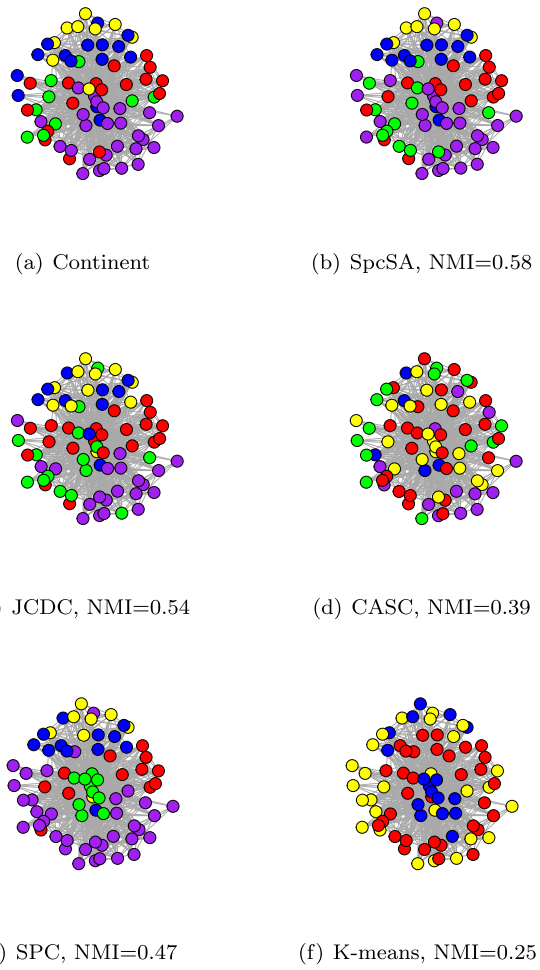


(a) Continent as attribute

(b) Position in 1980 as attribute (c) Position in 1994 as attribute

Figure 6. (a)The partition of the world trade network using continent as attribute, where different vertex shapes indicate detected partition and different colors indicate continent partition, where red is Asia, purple is Europe, yellow is North America, green is Africa and blue is South America. (b)-(c) The partitions of the world trade network using position in 1980 and position in 1994 as attributes, respectively, different colors indicate attribute partitions and different vertex shapes indicate detected partitions.

uses the three node attributes as auxiliary attributes, respectively. Figure 6 shows the result of applying SpcSA to the network three times. Each time, we use a single attribute ($m = 1$) and set $\beta_1 = 1$ in the SpcSA algorithm. In Figure 6(a), we use the continent as attribute and set $K = 5$. The plot shows a nearly perfect partition of the continents except the Europe group. Figure 6(b) shows the result from a four-way partition ($K = 4$) of the network using SpcSA with position in 1980 as attribute. The result indicates that the position in 1980 labels are mixed across the inferred communities. Finally, we use the position in 1994 as attribute and set $K = 3$. As shown in Figure 6(c), the three groups estimated by the proposed algorithm do not correlate well with the attribute. In conclusion, there is a strong correlation between the community structure of the network and the continents. Thus, it makes sense to compare partitions found by different community detection algorithms to the



(a) Continent (b) SpcSA, NMI=0.58

(c) JCDC, NMI=0.54 (d) CASC, NMI=0.39

(e) SPC, NMI=0.47 (f) K-means, NMI=0.25

Figure 7. (a) Network colored by continent, where red is Asia, purple is Europe, yellow is North America, green is Africa and blue is South America. (b)-(f) Results on the world trade network estimated using different methods. Colors match those in (a) in the best way possible.

continents. The position in 1994 and that in 1980 are treated as two auxiliary attributes.

We compare the partitions found by different methods to continent in Figure 7. From Figure 7, we can see that all methods except SpcSA fail to estimate the disassortative nature of African trade ([18]). In the proposed method, we suggest selecting the eigenvectors of the Laplacian matrix corresponding to the largest absolute eigenvalues to reveal the clusters. [27] points out that the eigenvectors related to the negative eigenvalues having large absolute values can discover disassortative clusters. The NMI values indicate that SpcSA outperforms the other algorithms. Further, the weights returned by SpcSA of the two attributes (position in 1994 and position in 1980) are 0.35 and 0.65. This is similar to the estimated results as shown in Table 4.

Table 5. NMI (cluster number) values returned by comparing the detected partition of SPC to seven attributes, respectively

Attributes	Status	Gender	Practice	School
NMI	0.44(2)	0.05(2)	0.087(2)	0.12(3)
Attributes	Office	Age	Year	
NMI	0.12(2)	0.20(2)	0.25(2)	

Table 6. Attribute weights returned by SpcSA in the lawyers friendship network

Attributes	Gender	Practice	School	Office	Age	Year
NMI	0.11	0.03	0.11	0.10	0.15	0.5

3.2.2 The lawyers friendship network

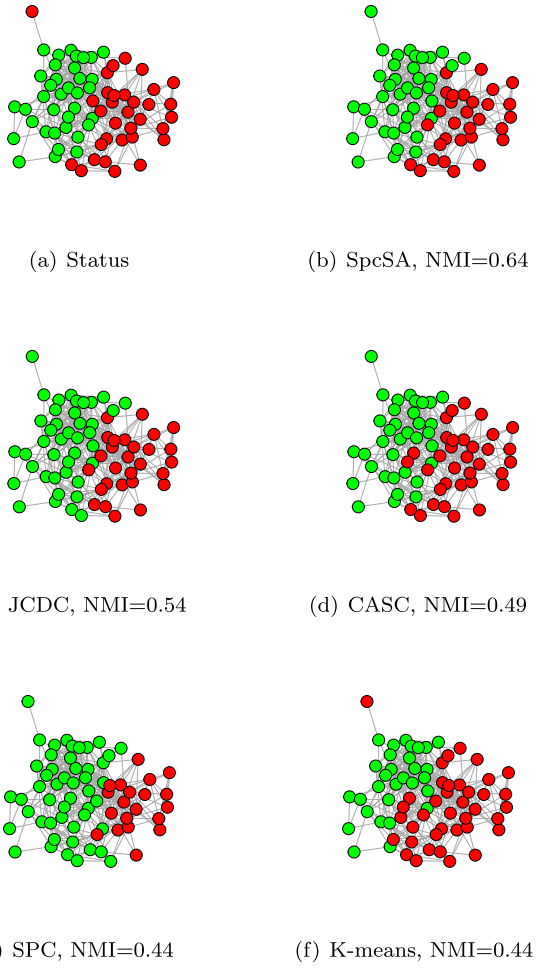
This data set represents the friendship among 71 lawyers in a Northeastern US corporate ([35]). Nodes denote lawyers and edges denote their friendship. Each lawyer has seven attributes: formal status (partner, associate), office (Boston, Hartford, Providence), gender, practice (litigation, corporate), law school (Harvard, Yale, others), age and number of years with the firm. True cluster labels of the lawyers are unknown and we adopt a similar method to the one in analyzing the world trade network. Along with the idea of analyzing Figure 7, we apply SpcSA to detect communities of the lawyer friendship network that uses the seven node attributes as auxiliary attributes, respectively. We find that the status correlates well with the community structure. To conserve space, the details are omitted here.

A comparison between the NMI values of spectral clustering on the network is shown in Table 5, with the seven attributes as reference points. The number of clusters in spectral clustering is equal to the levels of the category attributes, except for the office attribute because its smallest partition only contains two nodes after data preprocessing. We set $K = 2$ for the office case and the two numeric attributes (age and the number of years with the firm), although other values of K may be feasibility. From Table 5, we see that choosing the status partition as a reference point is meaningful. The other attributes are taken into account in the clustering process.

NMI values of the communities estimated using different methods are shown in Figure 8. SpcSA gives the highest NMI value among the methods mentioned above. The estimated weights of different attributes are shown in Table 6. The relative importance of each attribute as shown in Table 6 is similar to that in Table 5.

4. DISCUSSION

In many cases, structural and attribute information are related and combining them can improve the accuracy of community detection in a network. We propose an integrated clustering that combines the topological structure and node attributes based on spectral clustering. This



(a) Status (b) SpcSA, NMI=0.64

(c) JCDC, NMI=0.54 (d) CASC, NMI=0.49

(e) SPC, NMI=0.44 (f) K-means, NMI=0.44

Figure 8. (a) Network colored by status, where green is partner and red is associate. (b)-(f) Results on the lawyers friendship network estimated using different methods. Colors match those in (a) in the best way possible.

method can detect the weights of different attributes, and the weights are estimated according to the importance of each attribute to the clustering quality. To overcome the challenge of the non-convexity of the objective function which can be treated as an extension of the normalized cut in spectral clustering, we propose an adaptive weight-adjustment method to iteratively improve the objective function. Taking into account of the structural information, the adjustment margin of an attribute weight is controlled by the separations between clusters and the separations within the same cluster. At each iteration, the weights are updated by adding an adjustment margin. The experimental results for both numerical and real data sets show that our method outperforms other methods mentioned in the paper. Note that our method does not truly result in sparse clustering, since all variables have non-zero weights. This issue may be further tackled in future studies.

ACKNOWLEDGEMENTS

The project was sponsored by National Natural Science Foundation of China (11301236), Natural Science Foundation of the AnHui Higher Education Institutions of China (KJ2017A377, KJ2017A376) and Anhui Provincial Natural Science Foundation (1608085QG169). The authors would like to thank an associate editor and an anonymous referee for their careful reading and helpful comments. The authors also express sincere thankfulness for Professor Bingyi Jing for his valuable suggestions.

Received 13 November 2017

REFERENCES

[1] Handcock, M. S., Raftery, A. E. and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A*, 170(2): 301–354. MR2364300

[2] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks*, 5(2): 109–137. MR0718088

[3] Mariadassou, M., Robin, S. and Vacher. C. (2010). Uncovering latent structure in valued graphs: a variarion approach. *he Annals of Applied Statistics*, 4(2): 715–742. MR2758646

[4] Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008). Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep): 1981–2014.

[5] Latouche, P., Birmelé, E. and Ambroise, C. (2009). Overlapping stochastic block models. *arXiv preprint arXiv:0910.2098.*

[6] Zhang, Y., Levina, E. and Xu, X. (2014). Detecting overlapping communities in networks using spectral methods. *arXivpreprint arXiv: 1412.3432.*

[7] Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1): 016107. MR2788206

[8] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 888–905.

[9] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69, 026113. MR2282139

[10] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23): 8577–8582.

[11] Fienberg, S. E. (2010). Introduction to papers on the modeling and analysis of network data. *Annals of Applied Statistics*, 1–4. MR2758081

[12] Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21(4): 825–839. MR3005799

[13] Goldenberg, A., Zheng, A. X., Fienberg, S. E. and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2): 129–233.

[14] Wang, F., Ding, C. and Li, T. (2009). Integrated KL (K-means-Laplacian) Clustering: A New Clustering Approach by Combining Attribute Data and Pairwise Relations. *Proceedings of the 2009 SIAM International Conference on Data Mining*, 38–48.

[15] Binkiewicz, N., Vogelstein, J. T. and Rohe, K. (2017). Covariate-assisted spectral clustering. *Biometrika*, 104(2): 361–377. MR3698259

[16] Yang, J., McAuley, J. and Leskovec, J. (2013). Community detection in networks with node attributes. *IEEE International Conference on Data Mining*, 1151–1156.

[17] Zhou, Y., Cheng, H. and Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2(1): 718–729.

[18] Zhang, Y., Levina, E. and Zhu, J. (2016). Community Detection in Networks with Node Features. *Electronic Journal of Statistics*, 10(2): 3153–3178. MR3571965
[19] Cao, X., Chang, X. and Xu, Z. (2014). Community Detection for Clustered Attributed Graphs via a Variational EM Algorithm. *Proceedings of the 2014 International Conference on Big Data Science and Computing, ACM*, 8.
[20] Cheng, H., Zhou, Y. and Yu, J. X. (2011). Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(2): 12.
[21] Tian, Y., Hankins, R. A. and Patel, J. M. (2008). Efficient aggregation for graph summarization. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 567–580.
[22] Gunnemann, S., Farber, I. and Raubach, S. (2013). Spectral Subspace Clustering for Graphs with Feature Vectors. *IEEE International Conference on Data Mining*, 231–240.
[23] Newman, M. E. J and Clauset, A. (2016). Structure and inference in annotated networks. *Nature Communications*, 7: 11863.
[24] Sengupta, S. and Chen, Y. (2017). A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2): 365–386. MR3763696
[25] Xu, Z., Ke, Y. and Wang, Y., et al. (2012). A model-based approach to attributed graph clustering. *Proceedings of the 2012 ACM SIGMOD international conference on management of data. ACM*, 505–516.
[26] Zanghi, H., Volant, S. and Ambroise, C. (2010). Clustering based on random graph model embedding vertex features. *Pattern Recognition Letters*, 31(9): 830–836.
[27] Rohe, K., Chatterjee, S. and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic block model. *Annals of Applied Statistics*, 1878–1915. MR2893856
[28] Tsai, C. Y. and Chiu, C. C. (2008). Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm. *Computational Statistics and Data Analysis*, 52(10): 4658–4672. MR2521612
[29] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416. MR2409803
[30] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5): 75–174. MR2580414
[31] Danon, L., Duch, D, J., Arenas, A. and Díaz-Guilera, A. (2007). *Large scale structure and dynamics of complex networks: from information technology to finance and natural science.* World Scientific, Singapore, 93–113.
[32] Sarkar, P. and Bickel P. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *Annals of Applied Statistics*, 43(3): 962–990. MR3346694
[33] Von, Luxburg. U., Belkin, M. and Bousquet, O. (2008). Consistency of Spectral Clustering. *The Annals of Statistics*, 36(2): 555–586. MR2396807
[34] De, Nooy. W., Mrvar, A. and Batagelj, V. (2011). *Exploratory social network analysis with Pajek.* Cambridge University Press.
[35] Lazega, E. (2001). *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership.* Oxford University Press.

Fengqin Tang
School of Mathematics and Statistics, Lanzhou University, Lanzhou, 730000, P.R. China
School of Mathematical Sciences, Huaibei Normal University, Huaibei, 235000, P.R. China
E-mail address: tfq05@163.com

Yuanyuan Wang
School of Mathematics and Statistics, Lanzhou University, Lanzhou, 730000, P.R. China
E-mail address: wangyuanyuan1021@hotmail.com

Jinxia Su
School of Mathematics and Statistics, Lanzhou University, Lanzhou, 730000, P.R. China
E-mail address: jinxiasu@lzu.edu.cn

Chunning Wang
School of Mathematics and Statistics, Lanzhou University, Lanzhou, 730000, P.R. China
E-mail address: w.wangchunning@sohu.com