The Annals of Statistics
2015, Vol. 43, No. 1, 57–89
DOI: 10.1214/14-AOS1265

FAST COMMUNITY DETECTION BY SCORE

BY JIASHUN JIN[1]

Carnegie Mellon University

Consider a network where the nodes split into K different communities. The community labels for the nodes are unknown and it is of major interest to estimate them (i.e., community detection). *Degree Corrected Block Model* (DCBM) is a popular network model. How to detect communities with the DCBM is an interesting problem, where the main challenge lies in the degree heterogeneity.

We propose a new approach to community detection which we call the Spectral Clustering On Ratios-of-Eigenvectors (SCORE). Compared to classical spectral methods, the main innovation is to use the entry-wise ratios between the first leading eigenvector and each of the other leading eigenvectors for clustering. Let A be the adjacency matrix of the network. We first obtain the K leading eigenvectors of A, say, $\hat{\eta}_1, \ldots, \hat{\eta}_K$, and let $\hat{R}$ be the $n \times (K-1)$ matrix such that $\hat{R}(i,k) = \hat{\eta}_{k+1}(i)/\hat{\eta}_1(i)$, $1 \le i \le n$, $1 \le k \le K-1$. We then use $\hat{R}$ for clustering by applying the k-means method.

The central surprise is, the effect of degree heterogeneity is largely ancillary, and can be effectively removed by taking entry-wise ratios between $\hat{\eta}_{k+1}$ and $\hat{\eta}_1$, $1 \le k \le K-1$.

The method is successfully applied to the web blogs data and the karate club data, with error rates of $58/1222$ and $1/34$, respectively. These results are more satisfactory than those by the classical spectral methods. Additionally, compared to modularity methods, SCORE is easier to implement, computationally faster, and also has smaller error rates.

We develop a theoretic framework where we show that under mild conditions, the SCORE stably yields consistent community detection. In the core of the analysis is the recent development on Random Matrix Theory (RMT), where the matrix-form Bernstein inequality is especially helpful.

1. Introduction. Driven by the emergence of online "networking communities" (e.g., Facebook, LinkedIn, MySpace, Google+) and by the growing recognition of scientifically central networked phenomena (e.g., gene regulatory networks, citation networks, road networks), we see today a great demand for methods to infer the presence of network phenomena, particularly in the presence of large datasets. Tools and discoveries in this area could potentially reshape scientific data analysis and even have impacts on daily life (friendship, marketing, security).

Received August 2013; revised August 2014.

[1]Supported in part by NSF Grant DMS-12-08315.

MSC2010 subject classifications. Primary 62H30, 91C20; secondary 62P25.

Key words and phrases. Community detection, Degree Corrected Block Model (DCBM), Hamming distance, k-means method, moderate deviation, modularity, PCA, social network, sparsity, spectral analysis.

A problem that is of major interest is "network community detection" [4, 8, 9, 17, 27–29, 31, 36, 37]. Given an n-node (undirected) graph $\mathcal{N} = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of nodes and E is the set of edges. We believe that V partitions into a small number of (disjoint) subsets or "communities". The nodes within the same community share some common characteristics. The community labels are unknown to us and the main interest is to estimate them.

An iconic example is the web blogs data [1], which was collected right after the 2004 presidential election. Each node of the network is a web blogs about US politics, and each edge indicates a hyperlink between them (we neglect the direction of the hyperlink so that the graph is undirected). In this network, there are two perceivable communities: political liberal and political conservative. It is believed that the web blogs share some common political characteristics (liberal or conservative, one supposes) that are significantly different between two communities, but are not significantly different among the nodes in the same community.

1.1. *Degree corrected block model* (*DCBM*). In the spirit of "all models are wrong, but some are useful" [6], we wish to find a network model that is both realistic and mathematically tractable.

The stochastic block model (BM) is a classic network model. The BM is mathematically simple and relatively easy to analyze [4]. However, it is too restrictive to reflect some prominent empirical characteristics of real networks. For example, the BM implies that the nodes within each community have the same expected degrees, and that the distribution of degrees within the community is Poisson. However, this conflicts with the empirical observation that in many natural networks, the degrees follow approximately a power-law distribution [14, 23].

In a different line of development, there are the p^* model and the exponential random graph model (ERGM) [14]. Compared to the BM, these models are more flexible, but unfortunately are also more complicated and so comparably much harder to analyze.

DCBM is a recent model proposed by [22], which has become increasingly popular in network analysis [4, 8, 22, 34, 37]. Compared to the BM, DCBM allows for degree heterogeneity and is much more realistic: for each node, it uses a free parameter to model the degree.

The comparison of DCBM with the p^* model and the ERGM [14, 23] is not obvious, given that all of them use a large number of parameters. However, in sections below, we propose a new spectral method where we show that in the DCBM, the degree heterogeneity parameters are largely ancillary: as far as community detection concerns, it is almost unnecessary to estimate these heterogeneity parameters. For this reason, the DCBM is much easier to analyze than the p^* or the ERGM model.

Perhaps the easiest way to describe the DCBM is to start with the case of two communities (discussion on the case of K communities is in Section 2). Recall

that $\mathcal{N} = (V, E)$ denotes an undirected network. We suppose the nodes split into two (disjoint) communities as follows:

$$V = V^{(1)} \cup V^{(2)}.$$

Let A be the $n \times n$ adjacency matrix of $\mathcal{N}$. In the DCBM, we fix $(n+3)$ positive parameters (a, b, c) and $\{\theta^{(n)}(i)\}_{i=1}^n$ and assume that:

- A is symmetric, with zeros on the diagonal (so there is no self-connections).
- The elements of the upper triangular $\{A(i, j) : 1 \le i < j \le n\}$ are independent Bernoulli random variables satisfying

$$P\big(A(i, j) = 1\big) = \theta^{(n)}(i)\theta^{(n)}(j) \begin{cases} a, & i, j \in V^{(1)}, \\ c, & i, j \in V^{(2)}, \\ b, & \text{otherwise.} \end{cases}$$

As n changes, we assume (a, b, c) are fixed but $\theta^{(n)}(i)$ may vary with n. The superscript "n" becomes tedious, so for simplicity, we drop it from now on. We call $\{\theta(i) : 1 \le i \le n\}$ the *degree heterogeneity parameters* or *heterogeneity parameters* for short.

For identifiability, we assume

$$\max\{a, b, c\} = 1, \qquad \theta_{\max} \le g_0,$$

where $\theta_{\max} = \max_{1 \le i \le n}\{\theta(i)\}$ and $g_0 \in (0, 1)$ is a constant.

It is probably more convenient if we rewrite the model in the matrix form. The following notation is associated with the heterogeneity parameters $\{\theta(i)\}_{i=1}^n$ and are frequently used in this paper. Let θ and Θ be the $n \times 1$ vector and the $n \times n$ diagonal matrix defined as follows:

$$\theta = \big(\theta(1), \theta(2), \ldots, \theta(n)\big)', \qquad \Theta(i, i) = \theta(i), \qquad 1 \le i \le n. \tag{1.1}$$

Moreover, for $k = 1, 2$, let $\mathbf{1}_k$ be the $n \times 1$ indicator vector such that $\mathbf{1}_k(i) = 1$ if $i \in V^{(k)}$ and 0 otherwise. With this notation, we can rewrite

$$A = E[A] + W, \qquad W \equiv A - E[A],$$

where $E[A]$ denotes the expectation of A (also an $n \times n$ matrix), and

$$E[A] = \Omega - \operatorname{diag}(\Omega), \qquad \Omega \equiv \Theta\big[a\mathbf{1}_1\mathbf{1}' + c\mathbf{1}_2\mathbf{1}_2' + b\big(\mathbf{1}_1\mathbf{1}_2' + \mathbf{1}_2\mathbf{1}_1'\big)\big]\Theta.$$

Note that the entries in the upper triangular of W are independently (but not identically) distributed as centered-Bernoulli; such W is known as a generalized Wigner matrix [11].

REMARK. While it seems $\mathbf{1}_k$ are known, they are not for they depend on $V^{(k)}$—the unknown community partitions that are of primary interest.

1.2. *Where is the information*: *Spectral analysis heuristics*. In [33], John Tukey mentioned an idea that can serve as a general guideline for statistical inference. Tukey's idea is that before we tackle any statistical problem, we should think about "which part of the data contains the information": the "best" procedure should capture the most direct information containing the quantity of interest.

In our setting, the quantities of the interest are the community labels. Recall that

$$A = \Omega - \operatorname{diag}(\Omega) + W.$$

Seemingly, Ω contains the most direct information of the community labels: the matrix W only contains noisy and indirect information of the labels, and the matrix $\operatorname{diag}(\Omega)$ only has a negligible effect, compared to that of Ω.

In light of this, we take a close look on Ω. For $k = 1, 2$, let $\theta^{(k)} \equiv \theta^{(n,k)}$ be the $n \times 1$ vector such that (recall that θ is the shorthand for $\theta^{(n)}$)

$$\theta^{(k)}(i) = \theta(i) \qquad \text{if } i \in V^{(k)}$$

and

$$\theta^{(k)}(i) = 0 \qquad \text{otherwise, } 1 \le i \le n.$$

For any vector x, let $\|x\|$ denote the ℓ^2-norm. Write for short

$$d_k = \|\theta^{(k)}\|/\|\theta\|, \qquad k = 1, 2.$$

Note that $\|\theta^{(k)}\|$ can be interpreted as the *overall degree intensities* of the kth community.

In most part of the paper, the eigenvalues of interest are simple (i.e., algebraic multiplicity 1 [18]). The following lemma is a special case of Lemma 2.1, which is proved in the supplementary material [19], Appendix C [Θ is the diagonal matrix in (1.1)].

LEMMA 1.1. *If $ac \neq b^2$, then Ω has two simple nonzero eigenvalues*

$$\tfrac{1}{2}\|\theta\|^2\Big(ad_1^2 + cd_2^2 \pm \sqrt{(ad_1^2 - cd_2^2)^2 + 4b^2d_1^2d_2^2}\Big),$$

and the associated eigenvectors η_1 and η_2 (with possible nonunit norms) are

$$\Theta\Big(bd_2^2 \cdot \mathbf{1}_1 + \tfrac{1}{2}\Big[cd_2^2 - ad_1^2 \pm \sqrt{(ad_1^2 - cd_2^2)^2 + 4b^2d_1^2d_2^2}\Big] \cdot \mathbf{1}_2\Big).$$

The key observation is as follows. Let r be the $n \times 1$ vector of the coordinate-wise ratios between η_1 and η_2 (up to normalizations)

$$r(i) = \frac{\eta_2(i)/\|\eta_2\|}{\eta_1(i)/\|\eta_1\|}, \qquad 1 \le i \le n.$$

Define the $n \times 1$ vector r_0 by

$$r_0(i) = \begin{cases} 1, & i \in V^{(1)}, \\ -\Big(\dfrac{ad_1^2 - cd_2^2 + \sqrt{(ad_1^2 - cd_2^2)^2 + 4b^2 d_1 d_2}}{2bd_1 d_2}\Big)^2, & i \in V^{(2)}. \end{cases} \tag{1.2}$$

Then by Lemma 1.1 and basic algebra,

$$r \propto r_0.$$

We are now ready to answer Tukey's query on "where is the information": the sign vector of r is the place that contains the most direct information of the community labels.

The central surprise is that, as far as community detection concerns, the *heterogeneity parameters* $\{\theta(i)\}_{i=1}^n$ are largely ancillary, and their influence can be largely removed by taking the coordinate-wise ratio of η_1 and η_2 as above [though r still depends on (θ, n), but the dependence is only through the overall degree intensities d_1 and d_2]. This allows us to successfully extract the information containing the community labels without any attempt to estimate the heterogeneity parameters.

Compared to approaches where we attempt to estimate the heterogeneity parameters, our approach has advantages. The reason is that many real-world networks (e.g., web blogs network) are sparse in the sense that the degrees for many nodes are small. If we try to estimate the heterogeneity parameters of such nodes, we get relatively large estimation errors which may propagate to subsequent studies.

1.3. *SCORE: A new approach to spectral community detection.* The above observations motivate the following procedure for community detection, which we call Spectral Clustering On Ratios-of-Eigenvectors (SCORE).

(a) Let $\hat{\eta}_1$ and $\hat{\eta}_2$ be the two unit-norm eigenvectors of A associated with the largest and the second largest eigenvalues (in magnitude), respectively.

(b) Let $\hat{r}$ be the $n \times 1$ vector of coordinate-wise ratios: $\hat{r}(i) = \hat{\eta}_2(i)/\hat{\eta}_1(i)$, $1 \le i \le n$.

(c) Clustering the labels by applying the k-means method to the vector $\hat{r}$, assuming there are ≤ 2 communities in total.

The key insight is that, under mild conditions, we expect to see that

$$\hat{\eta}_1 \approx \eta_1/\|\eta\|_1, \qquad \hat{\eta}_2 \approx \eta_2/\|\eta_2\|,$$

where η_1 and η_2 are the two eigenvectors of Ω as in Lemma 1.1. Comparing with (1.2), we expect to have

$$\hat{r} \approx r \propto r_0.$$

In step (c), we use the k-means method. Alternatively, we could use the hierarchical clustering method [16]. For most of the numeric study in this paper, we

use the k-means package in MATLAB. In comparison, the performance of the k-means method and the hierarchical method are mostly similar, and that of the latter is slightly worse sometimes.

Note that since $\hat{r}$ is one-dimensional, both methods are equivalent to *simple thresholding*. That is, for some threshold t, we classify a node i, $1 \le i \le n$, to one community if $\hat{r}(i) > t$, and to the other community otherwise. Seemingly, the simplest choice is $t = 0$. Alternatively, one could use a recursive algorithm to determine the threshold: (a) estimate the community labels by applying the simple thresholding to $\hat{r}$ with $t = 0$, (b) update the threshold with the estimated labels, say, following (1.2) with (a, b, c, d_1, d_2) estimated, (c) repeat (a)–(b) with the threshold updated recursively.

The computational complexity of SCORE mostly comes from obtaining the two leading eigenvectors. For many social network data sets, the adjacency matrix is very sparse, and the computational complexity for obtaining the leading eigenvectors is only slightly larger than $O(n^2)$, using the simple power method. See [28], page 8581, for more discussions.

1.4. *Consistency of SCORE.* In Section 2, we extend SCORE to the case where we have K communities ($K \ge 2$), and investigate the theoretic properties. The main results are presented in Theorems 2.1 and 2.2. In the case of $K = 2$, these theorems simplify to that of tight bounds on $\|\hat{r} - r\|$ and on the Hamming error of community detection (i.e., expected number of nodes where the estimated label does not match with the true label). A direct results of the two theorems is that, under some regularity conditions, SCORE is (weakly) consistent [37] in community detection, in the sense that the Hamming error is much smaller than n.

The focus of the study is to identify a class of θ that is *as broad as possible* over which SCORE is uniformly consistent. To do so, we choose not to impose much *structural* assumptions on θ [such as an i.i.d. model for $\theta(i)$]. In fact, the regularity conditions we need for the consistency are conditions that only depend on the ℓ^q-norms and extreme coordinates of θ, not on the structure of θ. The reasons for doing so is two-fold:

- The structure of θ is largely unknown. For example, the correlation structures among different coordinates of θ is hard to model and is hard to estimate.
- As far as community detection is concerned, the role of θ is largely ancillary, and the effect of θ can be largely removed by using SCORE.

Note that our theoretic study is different from Zhao et al. [37] [where a scaled i.i.d. model is used for $\theta(i)$], and from [13, 31] [where the focus is on BM so all $\theta(i)$ are equal]. On the other hand, compared to those in [13, 31, 37], since we choose not to impose much structural assumptions on θ, our results (and the regularity conditions) generally have more complicate forms. To better compare our results with that in [13, 31, 37], we interpret our results and regularity conditions in Section 2.8 with a (scaled) i.i.d. model.

1.5. *Applications to the web blogs data and the karate club data.* We investigate the performance of the SCORE with two well-known networks: the web blogs network and the karate club network. The web blogs network is introduced earlier in the paper. The network has a giant component which we use for the analysis. The giant component consists of 1222 nodes and 16,714 edges. Each blog is manually labeled either as liberal or conservative in [1] which we use as the ground truth. The karate club network can be found in [35]. The network consists of 34 nodes and 136 edges, where each node represents a member in the club. Due to the fission of the club, the network has two perceivable communities: Mr. Hi's group and John's group. All members are labeled in [35], Table 1, which we use as the ground truth.

Consider the web blogs network first. In the left panel of Figure 1, we plot the histogram of the vector $\hat{r}$, which clearly shows a two mode pattern, suggesting that there are two underlying communities. In the right panel of Figure 1, we plot the entries of $\hat{r}$ versus the indices of the nodes, with red crosses and blue circles representing the nodes from the liberal and conservative communities, respectively; the plot shows that the red crosses and blue circles are almost completely separated from each other, suggesting that two communities can be nicely separated by applying simple thresholding to $\hat{r}$.

The error rate of the SCORE is reasonably satisfactory. In fact, if we use the procedure following steps (a)–(c), the error rate is $58/1222$. The error rate stays the same if we replace the k-means method in (c) by the hierarchical method (for both methods, we use the built-in functions in MATLAB; the linkage for the hierarchical method is chosen as "average" [16]).

Alternatively, we can use simple thresholding in step (c). In fact, the k-means method is equivalent to simple thresholding with $t = -0.7$. Moreover, the error rate is $82/1222$ if we set $t = 0$, and the error rate is $55/1222$ if we set $t = -0.6$ (this is the "ideal threshold," the threshold we would choose if we know the true

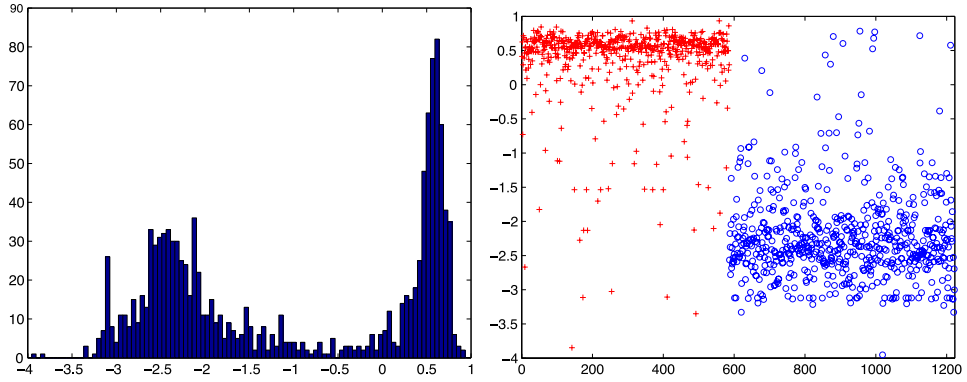

FIG. 1. *The vector $\hat{r}$ (web blogs data). Left: histogram of $\hat{r}$. Right: plot of the entries of $\hat{r}$ versus the node indices (red cross: liberal; blue circle: conservative).*

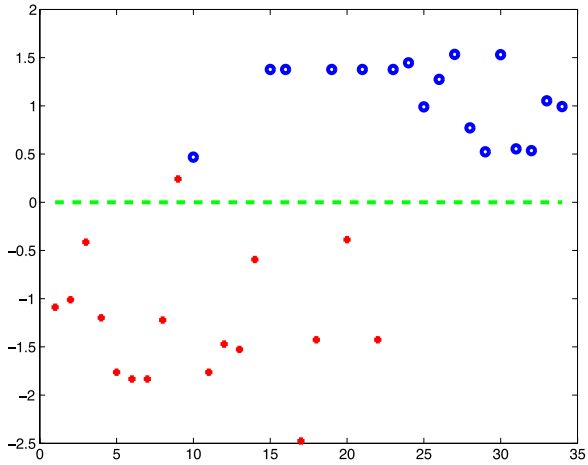

FIG. 2. *Plot of the entries of $\hat{r}$ versus the node indices* (*results are based on karate club network*; *red cross*: *Mr. Hi's group*; *blue circle*: *John's group*).

labels; if only). The results are tabulated in Table 1, along with error rates by some other methods, to be discussed below.

We consider the karate network next. Similarly, in Figure 2, we plot the coordinates of $\hat{r}$ associated with the karate data versus the node indices, with red crosses and blue circles representing the nodes from the group of Mr. Hi and the group of John [35], respectively. Our method has an error rate of $1/34$ if in step (c) we either use the k-means method or the simple thresholding with $t = 0$ (the error rate is $0/34$ if we set t as the "ideal threshold"). See Table 1 for details.

1.6. *Comparison with classical spectral clustering methods.* The success of SCORE (which is a spectral method) prompts the question whether classical spectral methods work well, too. Below are two classical spectral methods:

(a′) Obtain the two leading (unit-norm) eigenvectors $\hat{\eta}_1$ and $\hat{\eta}_2$ of A.
(b′) Viewing $(\hat{\eta}_1, \hat{\eta}_2)$ as a bivariate data set with sample size of n, apply the k-means method assuming there are at most two communities.

Alternatively, one may use the following variation, which is studied in [31].

(a″) Obtain an $n \times n$ diagonal matrix S by $S(i,i) = \sum_{j=1}^{n} A(i,j)$, $1 \le i \le n$.
(b″) Apply (a′)–(b′) to $S^{-1/2} A S^{-1/2}$.

We call the two procedures *ordinary Principle Component Analysis* (*oPCA*) and *normalized PCA* (*nPCA*), respectively.[2]

[2]oPCA and nPCA are also called spectral clustering on the adjacency matrix and on the graph Laplacian, respectively. See [8, 10], for example.

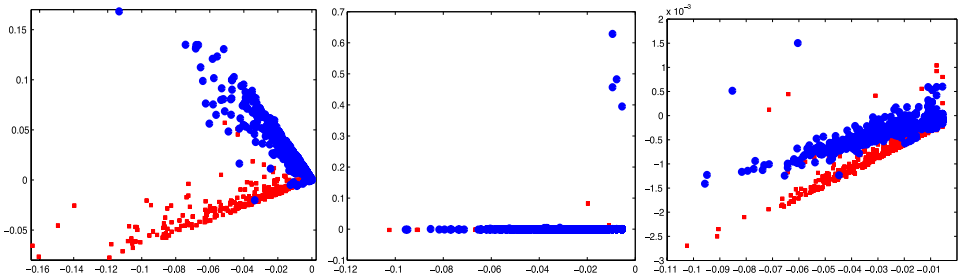

FIG. 3. *Left*: *plot of the first leading eigenvector of* A (x*-axis*) *versus the second leading eigenvector of* A (y*-axis*). *Middle*: *plot of the first leading eigenvector of* $S^{-1/2}AS^{-1/2}$ (x*-axis*) *versus the second leading eigenvector of* $S^{-1/2}AS^{-1/2}$ (y*-axis*). *Right*: *zoom-in of the middle panel. Results are based on the web blogs data, with red representing liberal and blue representing conservative.*

It turns out that both PCA approaches work unsatisfactorily. In fact, for the web blogs data, the error rates of oPCA and nPCA are $437/1222$ and $600/1222$, respectively, and for the karate data, the error rates are $1/34$ for both methods. See Table 1 for details.

The main reason why the two PCA methods perform unsatisfactorily is that different coordinates of the two leading eigenvectors are heavily affected by the degree inhomogeneity; see Lemma 1.1. In the left panel of Figure 3, we display the two leading eigenvectors of A, based on the web blogs data. The coordinates of two vectors are highly skewed to the left, reflecting serious degree heterogeneity. Compare [30] where a similar phenomenon is observed.

Somewhat surprisingly, though nPCA intends to correct degree heterogeneity, the correction is not particularly successful, partially because that the adjacency matrix of the web blogs data is very sparse. In the right two panels of Figure 3 (the rightmost panel is the zoom-in version of the panel to its left), we plot the two leading eigenvectors of $S^{-1/2}AS^{-1/2}$. It is seen that some of the entries of $\hat{\eta}_2$ are very large (compared to other entries).

Note that the unsatisfactory performance of oPCA (or nPCA) does not mean that the two leading eigenvectors of A (or $S^{-1/2}AS^{-1/2}$) are not "cluster-able". It only means that we need to pre-process the eigenvectors in a way so that some conventional methods (such as the k-means) can cluster well; SCORE provides a convenient pre-processing approach.

1.7. *Comparison with other spectral methods.* Newman [28] proposes a different spectral method, Spectral Modularity (SM), which we have applied to the weblog data and the karate data. The resultant error rates are $69/1222$ and $1/34$, respectively, compared to $58/1222$ and $1/34$ by SCORE. In Section 3, we further compare this method with SCORE with simulated data; see details therein.

Note that Newman's method is different from SCORE, especially when there are 3 or more communities. Note also that theoretically Newman's method is not

fully analyzed. In comparison, SCORE is fully analyzed in Section 2, where we discuss community detection for the general case of K communities.

1.8. *Comparison with the profile likelihood approach.* The profile likelihood (PL) approach is a well-known method for community detection [4, 22, 37]. The method was first proposed by Karrer and Newman [22] and was later carefully analyzed by Zhao et al. [37]; see [37], equation (2.3), for details.

In principle, PL is computationally NP-hard [4] (and so are many modularity methods; see, e.g., [37]), as it searches exhaustively over all possible community partitions, and pick the one that optimizes the so-called functional of profile likelihood. To mitigate this difficulty, many heuristic algorithms are proposed to approximate the theoretic optimizer, among which is the so-called *tabu algorithm* [37].

We have compared SCORE with the PL (implemented with the tabu search; the code is generously shared by authors of [37]). In comparison, PL is computationally much more expensive, and is increasingly so when the size or complexity of the network increases. The algorithm is also relatively unstable: it depends on the initial guess of the community partition, so it may not converge to the true partition with a "bad" starting point. The instability can be alleviated by increasing the number of searches, but that is at the expense of substantially longer computational time.

The error rates of PL for two data sets are illustrated in Figure 4 (left: karate; right: web blogs), in terms of the histograms based on 100 independent repetitions (the error rates are random for they depend on the initial guess of the community partition, generated randomly).

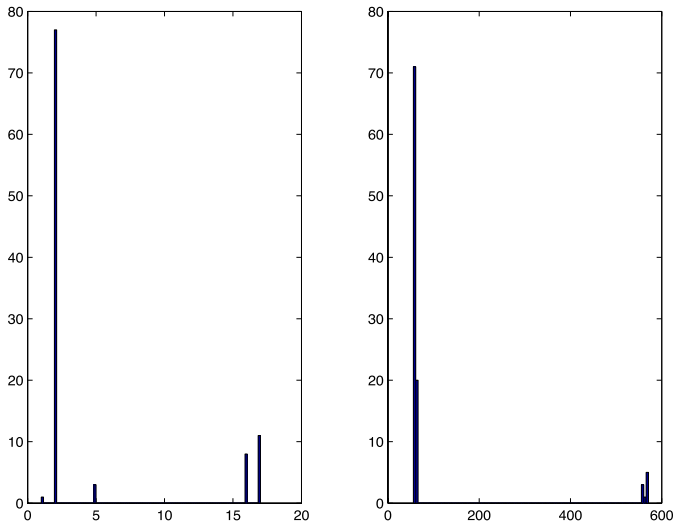

FIG. 4. *Histogram of errors by PL for the karate data* (*left*) *and the web blogs data. The results are based on* 100 *independent repetitions.*

TABLE 1
Comparison of number errors. For SCORE, we consider three different threshold choices. The result of PL depend on the starting point and is random, where means and standard deviations (SD) are computed based on 100 *independent repetitions*

	SCORE			PCA		
	$t = 0$	k-means	Ideal	Ordinary	Normalized	PL
Weblogs	82	58	55	437	600	104.5 (SD:145.4)
Karate	1	1	0	1	1	4.9 (SD:5.7)

The most prominent problem of PL (and many modularity methods [37]) is that, in quite a few repetitions (9 out of 100 for the web blogs data, and 19 out of 100 for the karate data), the algorithm fails to converge to the true community partition and yields poor results. For the karate data, the number of clustering errors have a mean of 4.85 and a standard deviation of 5.7. For the web blogs data, the number of clustering errors have a mean of 104.5 and a standard deviation of 145.5. If we remove the "outliers" (the 9 outlying repetitions for the web blogs data and the 19 outlying repetitions of the karate data), then for the karate data, the errors have a mean of 2.1 and a standard deviation of 0.6, and for the web blogs data, the mean is 59, and the standard deviation of 2.4. See Table 1.

That PL is more stable for the web blogs data than the karate data is unexpected (as the karate data has a relatively small size, we expect that it is relatively easy for the PL to find the true community partition). One possible explanation is that the communities in the former is more strongly structured, so the algorithm converges faster for the web blogs data than for the karate data.

In the above experiments, we use a random start. From time to time, one would like to first pick a fast algorithm to estimate the labels, and then use the estimated labels to start the PL. We have tried this approach where we start with the estimated labels by oPCA, SCORE and nPCA. The error rates (based on 100 independent repetitions) for web blogs data are 62 ± 2.0, 62 ± 0.0 and 569.4 ± 3.7, correspondingly, and those for the karate data set are 2 ± 0.0 for all three choices of start.

On one hand, this confirms that PL performs well given a good start. On the other hand, it is usually hard to pick a good start (or to evaluate how well a start is) in practice, and the performance of PL could be unsatisfactory given a poor start.

To conclude this section, we mention some data analysis results (on one or both data sets) in the literature, where the error rates are reported in different forms. The web blog data was analyzed in [22], where it was reported that the normalized mutual information (NMI) between the true labels and the estimated labels is 0.72. In comparison, SCORE yields an NMI of 0.725. In [37], the error rate is reported in terms of Adjusted Random Index (ARI) between the true labels and the estimated labels. The ARI of SCORE is 0.819 for the web blog data and 0.8823 for the karate

data. The results are similar to those reported in [2], page 16. The web blog data is also analyzed in [4], with an error rate of 61/1228.

1.9. *Summary.* We propose SCORE as a new approach to network community detection when a DCBM is reasonable. The main innovation is to use the coordinate-wise ratios of the leading eigenvectors for clustering. In doing so, we have taken advantage of the fact that the degree heterogeneity parameters $\theta(i)$ are merely nuisance and we can largely remove their effects without actually estimating them.

We have used the karate club data and the web blogs data to investigate the performances for several algorithms including SCORE, oPCA, nPCA, Newman's SM and PL. First, SCORE behaves much more satisfactory than the two PCA approaches. Second, SCORE is different from Newman's SM in both the idea and in the algorithm (especially when $K > 2$), and has a smaller error rate for analyzing the web blogs data. Third, SCORE also has advantages over the PL: the good performance of PL depends on a good start, so it can be unstable; also, computationally, PL is comparably slower than SCORE, but it does not outperform in terms of error rates. Finally, SCORE is conceptually simple and easy to implement, so there is ample space for extensions in the future work.

The paper is closely related to [31] (see also [9]), but is different in important ways. The focus of this paper is on DCBM where the number of communities K is small, while the focus of [31] is on BM where K is large. Our analysis is also different from that in [37], for we use a nonstructured model for the degree heterogeneity parameters $\theta(i)$; see Sections 1.4 and 2 for more discussions.

1.10. *Content.* The remaining part of the paper is organized as follows. In Section 2, we consider a K-community network with a fixed integer $K \geq 2$. By delicate spectral analysis as in Sections 2.1–2.6, we lay out the framework under which the SCORE yields consistent estimates of the community labels. In Section 2.7, we address the stability of the SCORE, where Lemmas 2.7–2.8 contain key ingredient for proving the main theorems. In Section 2.8, we compare our results with that in [13, 31, 37] using a (scaled) i.i.d. model for $\theta(i)$. In the supplementary material [19], Appendix A, we suggest some extensions of the SCORE. The main results are proved in the supplementary material [19], Appendix B. where we outline main technical devices required for the proofs. Numeric investigation is continued in Section 3, where we compare SCORE, oPCA, nPCA, Newman's SM and PL with simulated data. Section 4 discusses connection between SCORE and existing literatures. Secondary lemmas are proved in the supplementary material [19], Appendix C.

1.11. *Notation.* In this paper, for two vector u, v with the same size, (u, v) denotes their inner product. For any fixed $q > 0$ and any vector x, $\|x\|_q$ denotes the ℓ^q-norm. The subscript is dropped for simplicity if $q = 2$. For any matrix M,

$\|M\|$ denotes the spectral norm and $\|M\|_F$ denotes the Frobenius norm. For two positive sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we say $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$, and we say $a_n \asymp b_n$ if there is a constant $c_0 > 1$ such that $1/c_0 \le a_n/b_n \le c_0$ for sufficiently large n.

In this paper, the notation θ and Θ are always linked to each other, where θ denotes the $n \times 1$ vector of degree heterogeneity parameters and Θ denotes the $n \times n$ diagonal matrix satisfying that $\Theta(i,i) = \theta(i)$, $1 \le i \le n$. Also, $\theta_{\min} = \min_{1\le i\le n}\{\theta(i)\}$ and $\theta_{\max} = \max_{1\le i\le n}\{\theta(i)\}$. For a vector ξ, when all coordinates are positive, we use $OSC(\xi)$ to denote the oscillation $\max_{1\le i,j\le n}\{\xi(i)/\xi(j)\}$. Throughout the paper, C denotes a generic positive constant that may vary from occurrence to occurrence.

2. Main results. In this section, we consider the community detection problem where the network $\mathcal{N} = (V, E)$ has K communities. Throughout the paper, $K \ge 2$ is a known integer. See Section 4 for discussion on the case where K is unknown. The section contains the main theoretic results of the paper, and is relatively long. Therefore, it is necessary to give a road map and to illustrate key ideas behind the main results.

First, in Section 2.1, we extend the DCBM and SCORE from the case of $K = 2$ to the case of $K \ge 2$. We then carry out spectral analysis on Ω and A, in Sections 2.3 and 2.4, respectively. In these sections, we derive explicit formulas for the leading eigenvalues and leading eigenvectors for Ω and A. The leading eigenvectors of Ω and A (denoted by $\hat{\eta}_k$ and η_k, $1 \le k \le K$, resp.) have very similar formulas, where the differences are bounded by terms depending on the generalized Wigner matrix W. The study of these terms boils down to that of controlling $\|\hat{\eta}_k - \eta_k\|$ and $\|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|$. We present tight bounds on these two quantities in Lemmas 2.7–2.8. Such bounds form the bases of Theorems 2.1–2.2, which are the main result of the paper. Theorems 2.1 and 2.2 are presented in Sections 2.5 and 2.6, respectively.

The main regularity conditions we need are (2.9), (2.12)–(2.13), and (2.15), to be introduced. Think the adjacency matrix A as the sum of the "signal matrix" $E[A]$ and the "noise matrix" $A - E[A]$. Condition (2.9) ensures that the spectral norm of the noise matrix is smaller than that of the signal matrix. Conditions (2.12)–(2.13) ensure that $E[A]$ has well-spaced leading eigenvalues, so the associated leading eigenvectors are robust to noise corruption. Condition (2.15) ensures several sharp large-deviations inequalities for vectors/matrices associated with $\|\hat{\eta}_k - \eta_k\|$ and $\|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|$ (some of these inequalities have only become available very recently).

The goal of our analysis is different from that in [31, 37]; see Section 1.4 for more discussion. The focus here is to explore how broadly (in terms of θ) SCORE is consistent, so we choose not to impose much *structural* assumptions on θ. Regularity conditions we impose only depend on the ℓ^q-norms of θ and the extreme

entries of θ. As a result, some of these conditions have relatively complicate form. In Section 2.8, we revisit such conditions with a (scaled) i.i.d. model, and show that these conditions can be reduced to simple forms, similar to those found in the literature (e.g., [37]).

2.1. *SCORE when there are K communities.* Given an (undirected) network $\mathcal{N} = (V, E)$, we assume the network splits into K different communities. That is, the set of nodes V partitions to K different (disjoint) subsets:

$$V = V^{(1)} \cup V^{(2)} \cup \cdots \cup V^{(K)}.$$

Let A be the adjacency matrix of $\mathcal{N}$, and introduce

$$\mathcal{M}_{n,K-1,K} = \big\{M : n \times (K-1) \text{ matrix that has } \le K \text{ distinct rows}\big\}. \tag{2.1}$$

SCORE for K-community networks contains the following steps (for convenience, when we say "leading eigenvectors" or "leading eigenvalues", we are comparing the *magnitudes* of the eigenvalues, neglecting the ± 1 signs):

- Obtain the K (unit-norm) leading eigenvectors of A: $\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_K$.
- Fixing a threshold T_n, define an $n \times (K-1)$ matrix $\hat{R}^*$ such that for all $1 \le i \le n$ and $1 \le k \le K-1$,

$$\hat{R}^*(i,k) = \begin{cases} \hat{R}(i,k), & \text{if } |\hat{R}(i,k)| \le T_n, \\ T_n, & \text{if } \hat{R}(i,k) > T_n, \\ -T_n, & \text{if } \hat{R}(i,k) < -T_n \end{cases} \tag{2.2}$$

$$\text{where } \hat{R}(i,k) = \frac{\hat{\eta}_{k+1}(i)}{\hat{\eta}_1(i)}.$$

- Let M^* be the matrix satisfying

$$M^* = \mathop{\mathrm{argmin}}_{M \in \mathcal{M}_{n,K-1,K}} \|\hat{R}^* - M\|_F^2.$$

  Write $M^* = [m_1, m_2, \ldots, m_n]'$ so that m_i' is the ith row of M^*. Note that M^* has at most K distinct rows, say, $m_{i_1}', m_{i_2}', \ldots, m_{i_K}'$ for some indices $1 \le i_1 < \cdots < i_K \le n$. We partition all nodes into K communities $\hat{V}^{(1)}, \hat{V}^{(2)}, \ldots, \hat{V}^{(K)}$ such that $\hat{V}^{(k)} = \{1 \le j \le n : m_j = m_{i_k}\}$.

Note that the last step is the classical k-means method. We make the following remarks. First, when M^* has k distinct rows for some $k < K$, we let $\hat{V}^{(\ell)} = \varnothing$ for all $(k+1) \le \ell \le K$. Second, the choice of the threshold T_n is flexible, and for convenience, we take

$$T_n = \log(n) \tag{2.3}$$

in this paper. We impose thresholding in (2.2) mainly for technical convenience in the proof of Theorem 2.2. Numeric study in this paper suggests that no coordinate of $\hat{R}$ would be unduly large and so the thresholding procedure in (2.2) is rarely

necessary. Last, since $\hat{\eta}_k$ are real-valued unit-norm eigenvectors, $1 \le k \le K$, so by basic algebra, provided that the largest K eigenvalues are all simple, $\hat{\eta}_k$ are uniquely determined, up to a factor of ± 1. Correspondingly, all columns of $\hat{R}$ are uniquely determined, up to a factor of ± 1; these factors do not affect the clustering results.

REMARK. In SCORE, we apply the k-means algorithm to the $n \times (K-1)$ matrix $\hat{R}^*$. When $K = 2$, the algorithm reduces to simple thresholding so the computational cost is relatively low. For larger K, the algorithm is NP hard. For numeric study, a conventional approach is to use some heuristic methods. In this paper, we use the build-in k-means package in MATLAB, which is one of such heuristic methods.

2.2. *DCBM when there are K communities.* As before, we assume that the adjacency matrix A satisfies

$$A = E[A] + W, \qquad E[A] = \Omega - \mathrm{diag}(\Omega), \tag{2.4}$$

where Ω is symmetric, and $W \equiv A - E[A]$ is the generalized Wigner matrix. In the core of DCBM is a $K \times K$ matrix

$$P = \big(P(i,j)\big)_{1 \le i,j \le K}.$$

For positive parameters $\{\theta(i)\}_{i=1}^n$ as before, we extend the $n \times n$ matrix Ω to a more general form such that

$$\Omega(i,j) = \theta(i)\theta(j)P(k,\ell) \qquad \text{if } i \in V^{(k)} \text{ and } j \in V^{(\ell)}. \tag{2.5}$$

Similarly, for identifiability, we fix a constant $g_0 \in (0,1)$ and assume that

$$\max_{1 \le i,j \le K} P(i,j) = 1, \qquad 0 < \theta_{\min} \le \theta_{\max} \le g_0, \tag{2.6}$$

where $\theta_{\min} = \min_{1 \le i \le n}\{\theta(i)\}$ and $\theta_{\max} = \max_{1 \le i \le n}\{\theta(i)\}$. Throughout this paper, we assume

$$P \text{ is symmetric, nonsingular, nonnegative and irreducible.} \tag{2.7}$$

A matrix is nonnegative if all coordinates are nonnegative. See [18], page 361, for the definition of irreducible.

In the analysis below, we use n as the driving asymptotic parameter, and allow the vector θ [and so also the matrix Θ; see (1.1)] to depend on n. However, we keep (K, P) as fixed. Consequently, there is a constant $C = C(P) > 0$ such that $\|P^{-1}\| \le C$, where $\|\cdot\|$ denotes the spectral norm.

The DCBM we use is similar to that in [37] (see also [22]), but is different in important ways. In their asymptotic analysis, Zhao et al. [37], model $\theta(i)$ as random variables that have the same means and take only finite values. In our setting, we treat $\theta(i)$ as nonrandom and only impose some mild regularity conditions and

moderate deviations conditions (see below). Additionally, Zhao et al. [37] need certain conditions on P which we do not require. For example, in the special case of $K = 2$, they require P to be positive definite, but we do not. See [37], page 7, for details.

2.3. *Spectral analysis of* Ω. We start by characterizing the leading eigenvalues and eigenvectors of Ω. Recall that

$$\theta = \big(\theta(1), \ldots, \theta(n)\big)', \qquad V = V^{(1)} \cup V^{(2)} \cup \cdots \cup V^{(K)}.$$

Similarly as before, let $\theta^{(k)}$ be the $n \times 1$ vectors such that

$$\theta^{(k)}(i) = \theta(i) \text{ or } 0, \qquad \text{according to } i \in V^{(k)} \text{ or not, } 1 \le k \le K. \tag{2.8}$$

Let D be the $K \times K$ diagonal matrix of the *overall degree intensities*

$$D(k,k) = \|\theta^{(k)}\|/\|\theta\|, \qquad 1 \le k \le K;$$

note that D depends on θ and so it also depends on n.

The spectral analysis on Ω hinges on the $K \times K$ matrix DPD, where D and P are as above. The following lemma characterizes the leading eigenvalues and leading eigenvectors of Ω, and is proved in the supplementary material [19], Appendix C.

LEMMA 2.1. *Suppose all eigenvalues of DPD are simple. Let* $\lambda_1/\|\theta\|^2$, $\lambda_2/\|\theta\|^2, \ldots, \lambda_K/\|\theta\|^2$ *be such eigenvalues, arranged in the descending order of the magnitudes, and let* $a_1, a_2, \ldots, a_K$ *be the associated (unit-norm) eigenvectors. Then the K nonzero eigenvalues of* Ω *are* $\lambda_1, \lambda_2, \ldots, \lambda_K$, *with the associated (unit-norm) eigenvectors being*

$$\eta_k = \sum_{i=1}^{K} \big[a_k(i)/\|\theta^{(i)}\|\big] \cdot \theta^{(i)}, \qquad k = 1, 2, \ldots, K.$$

Note that (a_k, η_k) are uniquely determined up to a factor of ± 1; such factors do not affect clustering results.

2.4. *Spectral analysis of* A. In this section, we characterize the leading eigenvalues and leading eigenvectors of A.

Consider the eigenvalues first. The study contains two key components, one is to characterize the spectral norm of the noise matrix W, and the other is to impose some conditions on the eigen-spacing of the matrix DPD so that the space spanned by the K leading eigenvectors of Ω are stable up to noise corruption.

For the first component, recall that θ may depend on n. We suppose

$$\big(\log(n)\theta_{\max}\|\theta\|_1\big)/\|\theta\|^4 \to 0 \qquad \text{as } n \to \infty. \tag{2.9}$$

Combining (2.9) with basic algebra, it follows that

$$\log(n)/\|\theta\|^2 \to 0, \qquad \big(\log(n)\|\theta\|_1\|\theta\|_3^3\big)/\|\theta\|^6 \to 0, \tag{2.10}$$

which are frequently used in the proofs in the supplementary material [19]. The following lemma characterizes the spectral norm of $W - \mathrm{diag}(\Omega)$, and is proved in the supplementary material [19], Appendix C, where the recent result by [32] on matrix-form Bernstein inequality is very helpful.

LEMMA 2.2. *If* (2.9) *holds, then with probability at least* $1 + o(n^{-3})$,

$$\|W - \mathrm{diag}(\Omega)\| \le 4\sqrt{\log(n)\theta_{\max}\|\theta\|_1}.$$

We wish that the K leading eigenvalues of A are properly spaced and all of them are bounded away from 0. To ensure that, we need some mild conditions on DPD. In detail, for any symmetric $K \times K$ matrix P, we denote the minimum gap between adjacent eigenvalues of P by

$$\mathrm{eigsp}(P) = \min_{1\le i\le K-1} |\lambda_{i+1} - \lambda_i|, \qquad \lambda_1 > \lambda_2 > \cdots > \lambda_K. \tag{2.11}$$

When any of the eigenvalues of P is not simple, $\mathrm{eigsp}(P) = 0$ by convention. We assume that there is a constant $C > 0$ such that for sufficiently large n,

$$\mathrm{eigsp}(DPD) \ge C. \tag{2.12}$$

Additionally, we assume the degrees in each communities have comparable "overall degree intensities", in that there is a constant $h_2 > 0$ such that

$$\max_{1\le i,j\le K}\{\|\theta^{(i)}\|/\|\theta^{(j)}\|\} \le h_2. \tag{2.13}$$

As a result, D has a bounded condition number. Recalling that P is a fixed matrix with $\|P^{-1}\| \le C$, combining this with (2.7) gives that all eigenvalues of DPD are bounded away from either 0 or ∞ by some constants. Combining these with Lemma 2.1, the following lemma is a direct result of Lemma 2.2 and basic algebra (e.g., [3], page 473), the proof of which is omitted.

LEMMA 2.3. *Consider a DCBM where* (2.9), (2.12) *and* (2.13) *hold. Let* $\hat\lambda_1, \hat\lambda_2, \ldots, \hat\lambda_K$ *be the leading eigenvalues of* A, *and let* $\lambda_1/\|\theta\|^2, \lambda_2/\|\theta\|^2, \ldots, \lambda_K/\|\theta\|^2$ *be the nonzero eigenvalues of* DPD, *both sorted descendingly in magnitudes. With probability at least* $1 + o(n^{-3})$, *the* K *leading eigenvalues of* A *are all simple, and*

$$\max_{1\le k\le K}\{|\hat\lambda_k - \lambda_k|\} \le 4\sqrt{\log(n)\theta_{\max}\|\theta\|_1}.$$

Combining Lemma 2.3 with (2.9), with probability at least $1 + o(n^{-3})$,

$$\hat{\lambda}_k \asymp \|\theta\|^2 \qquad \text{for all } 1 \le k \le K. \tag{2.14}$$

This result is frequently used in the proof section in the supplementary material [19].

Next, we study the leading eigenvectors. From now on, we assume conditions (2.12)–(2.13) hold, and let $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K$ be the K leading eigenvalues as in Lemma 2.3. For $1 \le k \le K$, whenever $\hat{\lambda}_k$ is not an eigenvalue of $W - \mathrm{diag}(\Omega)$, let $B^{(k)}$ be the $K \times K$ matrix

$$B^{(k)}(i,j) = \big(\|\theta^{(i)}\|\|\theta^{(j)}\|\big)^{-1}\big(\theta^{(i)}\big)'\big[I_n - \big(W - \mathrm{diag}(\Omega)\big)/\hat{\lambda}_k\big]^{-1}\theta^{(j)}, \qquad 1 \le i, j \le K.$$

If $\hat{\lambda}_k$ is an eigenvalue of $W - \mathrm{diag}(\Omega)$, let $B^{(k)}$ be the $K \times K$ matrix of 0.

LEMMA 2.4. *Consider a DCBM where (2.9), (2.12) and (2.13) hold. Let $\{\hat{\lambda}_k\}_{k=1}^K$ be the eigenvalues of A with the largest magnitudes. There is an event with probability at least $1 + o(n^{-3})$ such that over the event, for each $1 \le k \le K$, $\hat{\lambda}_k$ is simple, and the associated eigenvector is given by*

$$\hat{\eta}_k = \sum_{\ell=1}^{K}\big(\hat{a}_k(\ell)/\|\theta^{(\ell)}\|\big)\big[I_n - \big(W - \mathrm{diag}(\Omega)\big)/\hat{\lambda}_k\big]^{-1}\theta^{(\ell)},$$

where $\hat{a}_k$ is an (unit-norm) eigenvector of $DPDB^{(k)}$, and $\hat{\lambda}_k/\|\theta\|^2$ is the unique eigenvalue of $DPDB^{(k)}$ that is associated with $\hat{a}_k$.

We remark that $\hat{\eta}_k$ do not necessarily have unit norms, and they are uniquely determined up to a scaling factor. Among them, $\hat{\eta}_1$ is particularly interesting, where provided that the network $\mathcal{N} = (V, E)$ is connected, then all entries of $\hat{\eta}_1$ are strictly positive (or strictly negative). Also, the associated eigenvalue $\hat{\lambda}_1$ is always strictly positive. These results are due to Perron's powerful theorem [18], page 508; see Section 2.7 for more discussion.

2.5. *Characterization of the matrix $\hat{R}^*$.* We now characterize the matrix $\hat{R}^*$, defined as in (2.2). Let $\eta_1, \eta_2, \ldots, \eta_K$ be the K leading (unit-norm) eigenvectors of Ω as in Lemma 2.1. Define an $n \times (K-1)$ matrix R as a nonstochastic counterpart of $\hat{R}^*$ by

$$R(i,k) = \eta_{k+1}(i)/\eta_1(i), \qquad 1 \le k \le K-1, 1 \le i \le n;$$

note that $\|\eta_k\| = 1$. Unlike $\hat{R}$, $|R(i,k)| \le C$ for all i and k (see Lemma 2.1), so it is unnecessary to impose thresholding as that in (2.2).

We wish to characterize $\|\hat{R}^* - R\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm. To do so, we need to characterize $\|\hat{\eta}_k - \eta_k\|$ and $\|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|$ [the latter is

necessary because in the definition of $R(i,k)$, we have $\eta_1(i)$ on the denominator, which will be shown to be at the magnitude of $\theta(i)$; see Section 2.7 for details].

To derive tight bounds on $\|\hat{\eta}_k - \eta_k\|$ and $\|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|$, we need the following regularity condition, which requires that for sufficiently large n,

$$\log(n)\theta_{\max}^2/\theta_{\min} \le \|\theta\|_3^3. \tag{2.15}$$

This condition can be replaced by more relaxed conditions, B.1–B.2, to be introduced in the supplementary material [19], Appendix B; see details therein.

Given the above regularity conditions, we show in Lemmas 2.7–2.8 (to be introduced in Section 2.7) that

$$\|\hat{\eta}_k - \eta_k\|^2 \ll \|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|^2, \qquad \|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|^2 \le C\log(n)\mathrm{err}_n,$$

where

$$\mathrm{err}_n = \frac{\|\theta\|_3^3}{\|\theta\|^6}\Big[\sum_{i=1}^n \frac{1}{\theta(i)} + \frac{\log(n)}{\theta_{\min}}\Big(\frac{\|\theta\|_1}{\|\theta\|^2}\Big)^2\Big], \tag{2.16}$$

and the right-hand side are bounds derived from Taylor expansions of $[\Theta^{-1}(\hat{\eta}_k - \eta_k)]$ and Bernstein inequalities on random matrices; see the supplementary material [19], Appendix C.12, for details. It is now not surprising that the leading term of $\|\hat{R}^* - R\|_F^2$ is determined by $\|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|$ (and so by err_n). The following theorem is the corner stone for characterizing the behavior of SCORE, and is proved in the supplementary material [19], Appendix B.

THEOREM 2.1. *Consider a DCBM where the regularity conditions* (2.9), (2.12), (2.13) *and* (2.15) *hold. If* $T_n = \log(n)$ *is as in* (2.3), *then as* $n \to \infty$, *with probability at least* $1 + o(n^{-2})$, *we have that*

$$\|\hat{R}^* - R\|_F^2 \le C\log^3(n)\mathrm{err}_n.$$

For general choice of T_n, the result continues to hold if we replace the right-hand side by $C\log(n)T_n^2\mathrm{err}_n$.

2.6. *Hamming errors of SCORE.* Recall that $V = V^{(1)} \cup V^{(2)} \cup \cdots \cup V^{(K)}$ is the true community partition. Introduce the $n \times 1$ vector ℓ of true labels such that

$$\ell(i) = k \quad \text{if and only if} \quad i \in V^{(k)}, \qquad 1 \le i \le n.$$

For any community detection procedure, there is a (disjoint) partition $V = \hat{V}^{(1)} \cup \hat{V}^{(2)} \cdots \cup \hat{V}^{(K)}$, so we can similarly define the $n \times 1$ vector of estimated labels by

$$\hat{\ell}(i) = k \quad \text{if and only if} \quad i \in \hat{V}^{(k)}, \qquad 1 \le i \le n.$$

Especially, let $\hat{\ell}^{\mathrm{sc}} = \hat{\ell}^{\mathrm{sc}}(A, T_n, n)$ be the vector of estimated labels by SCORE.

For any $\hat{\ell}$, the expected number of mismatched labels is

$$H_p(\hat{\ell}, \ell) = \sum_{i=1}^{n} P\big(\hat{\ell}(i) \neq \ell(i)\big).$$

With that being said, we must note that the clustering errors should not depend on how we label each of the K communities. Toward this end, let

$$S_K = \big\{\pi : \pi \text{ is a permutation of the set } \{1, 2, \ldots, K\}\big\}. \tag{2.17}$$

Also, for any label vector ℓ where the coordinates take value from $\{1, 2, \ldots, K\}$ and any $\pi \in S_K$, let $\pi(\ell)$ denote the $n \times 1$ label vector such that

$$\pi(\ell)(i) = \pi\big(\ell(i)\big), \qquad 1 \leq i \leq n.$$

With this notation, a proper way to measure the performance of $\hat{\ell}$ is to use the Hamming distance as follows:

$$\mathrm{Hamm}_n(\hat{\ell}, \ell) = \min_{\pi \in S_K} H_p\big(\hat{\ell}, \pi(\ell)\big).$$

For $k = 1, 2, \ldots, K$, let n_k be the size of the kth community:

$$n_k = \big|V^{(k)}\big|.$$

The following theorem is proved in the supplementary material [19], Appendix B, and is the main result of the paper.

THEOREM 2.2. *Consider a DCBM where both the regularity conditions* (2.9), (2.12), (2.13) *and* (2.15) *hold. Suppose as* $n \to \infty$,

$$\log^3(n)\mathrm{err}_n / \min\{n_1, n_2, \ldots, n_K\} \to 0,$$

where err_n *is as in* (2.16). *For the estimated label vector* $\hat{\ell}^{\mathrm{sc}}$ *by the SCORE where the threshold* $T_n = \log(n)$ *is as in* (2.3), *there is a constant* $C > 0$ *such that for sufficiently large* n,

$$\mathrm{Hamm}_n\big(\hat{\ell}^{\mathrm{sc}}, \ell\big) \leq C \log^3(n)\mathrm{err}_n.$$

Similarly, for general T_n, the theorem continues to hold if we replace the right-hand side by $C \log(n) T_n^2 \mathrm{err}_n$. Theorem 2.2 says that SCORE is (weakly) consistent under mild conditions; see [37] for difference between strong consistency and weak consistency.

2.7. *Stability of SCORE.* The performance of SCORE hinges on the matrix $\hat{R}$ defined in (2.2):

$$\hat{R}(i,k) = \hat{\eta}_{k+1}(i)/\hat{\eta}_1(i), \qquad 1 \le i \le n, 1 \le k \le K-1.$$

Seemingly, SCORE could be unstable if the denominator $\hat{\eta}_1(i)$ is small (or even worse, equals to 0) for some i. Fortunately, this is not the case, and under mild conditions, for most i (or for all i with slightly stronger conditions), $\hat{\eta}_1(i) \asymp \eta_1(i) \asymp \theta(i)$. Below, we further characterize the vector $(\hat{\eta}_k - \eta_k)$, with emphasis on the case of $k = 1$.

We start with the following lemma, which is the direct result of Perron's theorem [18], Section 8.2, on nonnegative matrices, and which says that a coordinate of $\hat{\eta}_1$ can never be exactly 0, as long as the network is connected.

LEMMA 2.5. *Let A be the adjacency matrix of a network $\mathcal{N} = (V, E)$, let $\hat{\lambda}_1$ be the eigenvalue with the largest magnitude, and let $\hat{\eta}_1$ be the associated eigenvector where at least one coordinate is positive. If $\mathcal{N}$ is connected, then both $\hat{\lambda}_1$ and all coordinates of $\hat{\eta}_1$ are strictly positive.*

Next, for any $n \times 1$ vector ξ with strictly positive coordinates, define the *coordinate oscillation* by

$$OSC(\xi) = \max_{1 \le i,j \le n} \{\xi(i)/\xi(j)\}.$$

The following lemma is proved in the supplementary material [19], Appendix C [note that the ith coordinate of $\Theta^{-1}\eta_1$ is $\eta_1(i)/\theta(i)$].

LEMMA 2.6. *Consider a DCBM where (2.12)–(2.13) holds. We have*

$$OSC(\Theta^{-1}\eta_1) \le C.$$

The following lemmas constitute the key component of the proof of Theorem 2.1, but can also be used to obtain upper bounds for the number of "ill-behaved" coordinates of $\hat{\eta}_1$. These lemmas are proved in the supplementary material [19], Appendix C.

LEMMA 2.7. *Consider a DCBM where the conditions of Theorem 2.1 hold. With probability at least $1 + o(n^{-3})$, for all $1 \le k \le K$,*

$$\|\hat{\eta}_k - \eta_k\|^2 \le C \log(n) \|\theta\|_1 \|\theta\|_3^3 / \|\theta\|^6.$$

LEMMA 2.8. *Consider a DCBM where the conditions of Theorem 2.1 hold. With probability at least $1 + o(n^{-3})$, for all $1 \le k \le K$,*

$$\|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|^2 \le C \log(n) \mathrm{err}_n.$$

Recall that err_n is defined in (2.16). For Lemma 2.8, a weaker bound is possible if we simply combine Lemma 2.7 and the fact that $\|\Theta^{-1}\| \le 1/\theta_{\min}$. The current bound is much sharper, especially when only a few $\theta(i)$ are small.

We now obtain an upper bound on the number of "ill-behaved" entries of $\hat{\eta}_1$. Recall that $OSC(\Theta^{-1}\eta_1) \le C$. Fixing a constant $c_0 \in (0,1)$, we call the ith enty of $\hat{\eta}_1$ well-behaved if $|\hat{\eta}_1(i)/\eta_1(i) - 1| \le c_0$ (say). Let

$$\hat{S} = \hat{S}(c_0, \hat{\eta}_1, \eta_1; A, \Omega, n) = \big\{1 \le i \le n : \big|\hat{\eta}_1(i)/\eta_1(i) - 1\big| \le c_0\big\}. \tag{2.18}$$

The following lemma is a direct result of Lemmas 2.6 and 2.8, so we omit the proof.

LEMMA 2.9. *Consider a DCBM where the conditions of Theorem 2.1 hold. Fix $c_0 \in (0,1)$ and let $\hat{S}$ be as in (2.18). Then with probability at least $1 + o(n^{-3})$, $|V \setminus \hat{S}| \le C\log(n)\mathrm{err}_n$.*

Therefore, as long as $\log(n)\mathrm{err}_n/n \to 0$ when $n \to \infty$, the fraction of "ill-behaved" coordinates of $\hat{\eta}_1$ tends to 0 and is negligible.

In principle, provided that some stronger conditions are imposed, the techniques in this paper (especially those in the proof of Lemmas 2.7–2.8) can be used to show that with probability at least $1 + o(n^{-3})$,

$$\max_{1\le i\le n}\left|\frac{\hat{\eta}_1(i)}{\eta_1(i)} - 1\right| \le c_0,$$

where $c_0 \in (0,1)$ is a constant. In this case, Theorem 2.2 can be strengthened into that of with probability at least $1 + o(n^{-2})$,

$$\mathrm{Hamm}_n(\hat{\ell}^{\mathrm{sc}}, \ell) = 0.$$

Using terminology in the literature on variable selection [12], this says that SCORE has the *oracle property*, means that it achieves *exact recovery* with overwhelming probabilities.

2.8. *Remarks on the regularity conditions.* In the main results, Theorems 2.1–2.2, we have imposed the following regularity conditions: (2.9), (2.15) and (2.12)–(2.13). The last two can be roughly translated to that the leading eigenvalues of *DPD* are well spaced and are well understood. For this reason, we only discuss (2.9) and (2.15).

The seeming complexity of (2.9) and (2.15) is due to that we choose *not* to impose much structural assumptions on θ. The reasons for doing so are carefully explained in Section 1.4. On the other hand, if we choose to impose some structural assumptions on θ, these conditions can be much simplified. In this section, we illustrate this with an (scaled) i.i.d. model for θ.

In detail, let F be a distribution defined over $(0,\infty)$ and that does not vary with n, and let $g_n > 0$ be a (nonrandom) scaling factor. We model

$$\theta = g_n \cdot \mu \qquad \text{where } \mu = \big(\mu(1),\ldots,\mu(n)\big)' \text{ and } \mu(i) \overset{\text{i.i.d.}}{\sim} F;$$

we allow $g_n \to 0$ as $n \to \infty$, but require $\sqrt{n}g_n \to \infty$. For $q \neq 0$, let $m_q(F) = \int_0^\infty x^q\,dF(x)$. Suppose F satisfies some regularity conditions. By basic statistics, except for negligible probabilities,

$$\|\theta\|_q^q = g_n^q\|\mu\|_q^q \sim g_n^q m_q(F)n, \qquad q = -1, 1, 2, 3. \tag{2.19}$$

Similarly, let $\mu_{\max}$ and $\mu_{\min}$ be the largest and smallest entry of μ, respectively. Using (2.19), (2.9) and (2.15) are satisfied if

$$\frac{\log(n)\mu_{\max}^2/\mu_{\min}}{ng_n^2} \to 0. \tag{2.20}$$

Also, note that when (2.20) holds,

$$\mathrm{err}_n \le C\big(ng_n^2\big)^{-2}.$$

Below are some examples where (2.20) holds.

- (*Finite support*). The support of F is contained in $[a,b]$ for constants $b > a > 0$. This includes the Block Model as a special case (e.g., [13, 31]), where F is a point mass at 1. In this case, (2.20) reduces to $\log(n)/[ng_n^2] \to 0$. This is similar to that in [37], page 7, where F is supported on M different points $x_1, x_2, \ldots, x_M$.
- (F *is log-normal*). A frequently used model is when F is the CDF of e^X, where $X \sim N(u,\sigma^2)$. In this case, except for a probability of $o(n^{-k})$, $\mu_{\min} \ge \exp(u - \sigma\sqrt{2k\log(n)})$ and $\mu_{\max} \le \exp(u + \sigma\sqrt{2k\log(n)})$, and $\mu_{\max}^2/\mu_{\min} \ll n^\delta$ for any fixed $\delta > 0$ and large n. Therefore, condition (2.20) holds as long as $g_n \ge n^{-\vartheta}$ for some $0 \le \vartheta < 1/2$.
- (*Polynomial tails*). The most difficult case is when F has polynomial tails. If $F(x) \ge c_1 x^{q_1}$ as $x \to 0$ and $(1 - F(x)) \le c_2 x^{-q_2}$ as $x \to \infty$ for constants c_1, c_2, q_1, q_2, where q_1 and q_2 are sufficiently large so that (2.19) continues to hold. In this case, except for probability $O(n^{-k})$, $\mu_{\min} \ge Cn^{-(k+1)/q_1}$ and $\mu_{\max} \le Cn^{-(k+1)/q_2}$, and (2.20) holds as long as $\log(n)\cdot n^{1-(k+1)(2q_2^{-1}+q_1^{-1})}g_n^2 \to 0$. The discussion applies to the case where the support of F is contained in $[a,\infty)$ and that $(1 - F(x)) \le c_1\exp(-c_2x^{c_3})$ as $x \to \infty$, where a, c_1, c_2, c_3 are positive constants. In such cases, we can view $q_1 = q_2 = \infty$.

In conclusion, under the (scaled) i.i.d. model, Theorem 2.1 holds if we replace (2.9) and (2.15) by (2.20), where err_n is simplified as in $C(ng_n^2)^{-2}$; these results are very similar to those in [37].

3. Simulations. We have conducted a small-scale simulation study. The goal is to select a few representative cases to investigate the performances of the procedures we discussed in the preceding sections.

The simulation includes 6 different experiments, where we compare 5 different algorithms: SCORE, oPCA, nPCA, the profile likelihood approach (PL), Newman's Spectral Modularity method (Newman SM) and pseudo Likelihood (pseudo).

PL is realized with a tabu algorithm, which needs the input of an initial label vector. We consider two approaches to setting initial label vector: we generate the label vector randomly or set it as the estimated label vector by the SCORE. We consider the first approach in Experiments 1–5 and both approaches in Experiment 6. To differentiate two approaches, PL with the second approach by PL, and PL with the second approach by PL-SCORE.

For each simulation experiment, we choose integers n, K, and *rep*, representing the size of the network, the number of communities, and the number of repetitions for simulations, correspondingly. Fix a $K \times K$ matrix P. We generate an $n \times 1$ vector ℓ taking values from $\{1, 2, \ldots, K\}$, representing the vector for community labels, and an $n \times 1$ vector θ representing the degree heterogeneities. For $k = 1, 2, \ldots, K$, we let $V^{(k)} = \{1 \le i \le n : \ell_i = k\}$, and let $\mathbf{1}_k$ be the indicator vector of $V^{(k)}$ as before. Each simulation experiment contains the following steps:

(a) Let Θ be the $n \times n$ diagonal matrix such that $\Theta(i, i) = \theta(i)$, $1 \le i \le n$. Define the $n \times n$ matrix Ω by $\Omega = \Theta[\sum_{k,\ell=1}^{K} P(k, \ell)\mathbf{1}_k\mathbf{1}_\ell']\Theta$.

(b) Generate a symmetric $n \times n$ matrix W where all diagonals are 0, and for all $1 \le i < j \le n$, $W(i, j)$ are independent centered-Bernoulli with parameters $\Omega(i, j)$. Let $\tilde{A} = \Omega - \mathrm{diag}(\Omega) + W$, which can be viewed as the adjacency matrix of a network, say, $\mathcal{N} = (V, E)$.

(c) Let $\mathcal{N}_0 = (V_0, E_0)$ be the giant component of $\mathcal{N} = (V, E)$. Let A be the adjacency matrix of $\mathcal{N}_0$, and let n_0 be the size of $\mathcal{N}_0$.

(d) Apply all or a subset of the 5 aforementioned algorithms to A. Record the Hamming error rates of all methods under investigations.

(e) For integer *rep* mentioned above, repeat (b)–(d) for *rep* times.

Hamming error rate is defined as the ratio between the Hamming errors and n_0. In our study, n_0 is usually very close to n so we do not report the exact values. Also, we set the threshold T_n in (2.2) as ∞ so that we do not truncate any coordinates of $\hat{R}$ as usually none of them is unduly large; setting $T_n = \log(n)$ gives almost the same results. We now describe the each experiment in detail.

EXPERIMENT 1. In this experiment, we investigate how SCORE, oPCA, nPCA, and PL perform with the classical stochastic Block Model (BM). We choose $(n, K, rep) = (1000, 2, 50)$, P as the 2×2 matrix with 1 on the diagonals and 0.5 on the off-diagonals, and θ as the vector where all coordinates are 0.2. Also, we

TABLE 2
Comparison of mean error rates (Experiment 1). In each cell, the number in the bracket is the corresponding standard deviation (SD)

Methods	oPCA	nPCA	PL	SCORE
Mean (SD)	0.058 (0.009)	0.055 (0.010)	0.050 (0.065)	0.058 (0.009)

generate the label vector ℓ randomly by $(\ell_i - 1) \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)$. This is a relatively easy case and all methods perform satisfactory and have similar error rates. See Table 2 for the results.

It is noteworthy that in one of the repetitions, PL fails to converge and has an error rate of 49.8%. Such outlying cases are observed in most experiments below; sometimes the fraction of outlying cases is larger.

EXPERIMENT 2. In this example, we compare the performance of oPCA, nPCA, PL and SCORE (we use a slight variant of SCORE, SCOREq with $q = 2$; see the supplementary material [19]. The performance of this variant is generally similar to that of SCORE, but is slightly better in this experiment) for the case where we have three communities. We take $(n, K, rep) = (1500, 3, 25)$, and P as the 3×3 symmetric matrix where we have 1 on the diagonals, $P(1,2) = 0.4$, $P(2,3) = 0.4$ and $P(1,3) = 0.05$. We take θ as the vector such that $\theta(i) = 0.015 + 0.785 \times (i/n)^2$, $1 \le i \le n$, and generate the label vector ℓ randomly such that $\ell_i = 1, 2, 3$ with equal probabilities. The results are reported in Table 3, which suggest that SCORE outperforms nPCA and oPCA. The error rates of SCORE and PL are similar, but SCORE is comparably more stable than PL.

EXPERIMENT 3. In this experiment, we investigate how the heterogeneity parameters affect the performances of oPCA, nPCA, SCORE, PL, and Newman's SM. The experiment contains two experiments, Experiments 3(a)–3(b):

Experiment 3(a). In this experiment, we take $(n, K, rep) = (1000, 2, 50)$, P as the 2×2 matrix that has 1 on the diagonals and 0.5 on the off-diagonals. Also, we generate ℓ randomly by $(\ell_i - 1) \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)$. Fixing $c_0 = 0.5$ and $d_0 =$

TABLE 3
Comparison of mean error rates when there are three communities (Experiment 2). In each cell, the number in the bracket is the corresponding standard deviation (SD)

Methods	oPCA	nPCA	PL	SCORE
Mean (SD)	0.378 (0.041)	0.165 (0.084)	0.0636 (0.123)	0.0695 (0.004)

TABLE 4

Comparison of error rates [*Experiment* 3(a)]. *In each cell, the number in the bracket is the corresponding standard deviation* (*SD*). *From top to bottom*: $\theta(i) = d_0 + (c_0 - d_0)(i/n)$, $\theta(i) = d_0 + (c_0 - d_0)(i/n)^2$ *and* $\theta(i) = c_0 1\{i \le n/2\} + d_0\{i > n/2\}$; $(c_0, d_0) = (0.5, 0.02)$

Methods	**oPCA**	**nPCA**	**PL**	**SCORE**
Mean (SD)	0.066 (0.021)	0.066 (0.107)	0.042 (0.064)	0.043 (0.006)
	0.292 (0.014)	0.431 (0.122)	0.138 (0.080)	0.140 (0.010)
	0.254 (0.034)	0.476 (0.049)	0.139 (0.074)	0.130 (0.010)

0.02, we investigate three different choices of θ. In the first one, $\theta(i) = d_0 + (c_0 - d_0)(i/n)$, $1 \le i \le n$. In the second one, $\theta(i) = d_0 + (c_0 - d_0)(i/n)^2$, $1 \le i \le n$. In the last one, $\theta(i) = c_0 1\{i \le n/2\} + d_0 1\{i > n/2\}$, $1 \le i \le n$. Note that the heterogeneity effects are mild for the first choice of θ, but are much more severe in the other two choices. The results are tabulated in Table 4. The error rates of oPCA and nPCA are usually higher than that of PL and the SCORE. The average error rates of PL and SCORE are similar, but PL usually has a much larger standard deviation. The instability of the PL algorithm is due to that it depends on an initial guess (generated randomly), and when the initial guess is "bad," PL may fail to converge to the true labels.

Experiment 3(b). In this experiment, we take $(n, K, rep) = (1000, 2, 25)$ and P as the 2×2 matrix with 1 on the diagonals and 0.5 on the off-diagonals. We set the label vector ℓ by $\ell_i = 1\{i \le n/4\} + 2 \cdot 1\{n/4 < i \le n\}$. We first generate θ by $\log(\theta(i)) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, $1 \le i \le n$, where $\sigma = 0.2 \times [1, \sqrt{2}, \sqrt{3}, 2, \sqrt{5}]$, and then normalize θ by $\theta = 0.9 \times \theta/\theta_{\max}$. The results are reported in Figure 5, which suggest that SCORE has better performance than oPCA, Newman's SM and PL.

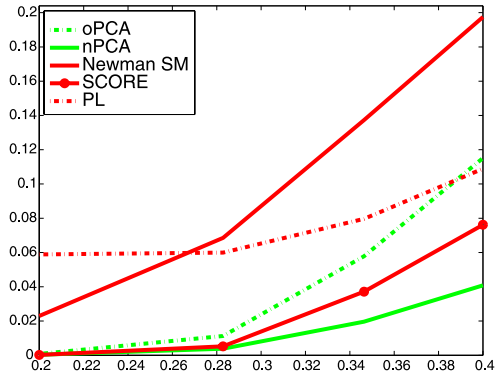

FIG. 5. *Comparison of Hamming error rates* [*Experiment* 3(b)]. *The heterogeneity vector* θ *satisfies* $\log(\theta(i)) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, $1 \le i \le n$. *x-axis*: σ. *y-axis*: *Hamming errors*.

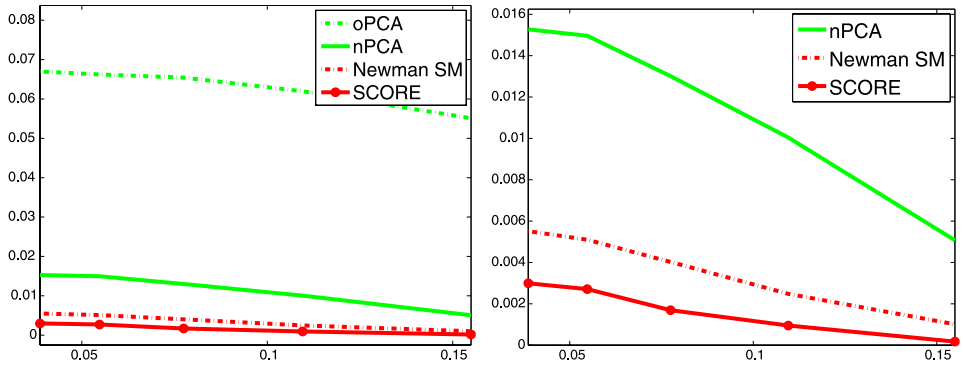

FIG. 6. *Comparison of Hamming error rates* (*Experiment* 4). *Right panel*: *zoom-in of the left panel*. *x-axis*: *parameter* d_0; *see Experiment* 4 *for details*. *y-axis*: *Hamming errors*.

Somewhat surprisingly, in this particular setting, nPCA has the best performance among all these procedures.

EXPERIMENT 4. In this experiment, we study the performances of the procedures for a larger n, and investigate how the heterogeneity affects the performances of the above procedures. We take $(n, K, rep) = (4000, 2, 25)$, P be the symmetric matrix with $P(1,1) = 3$, $P(2,2) = 1$ and $P(1,2) = 0.5$, and set the label vector ℓ by $\ell_i = 1\{i < n/4\} + 2 \cdot 1\{n/4 < i \le n\}$. For $c_0 = 0.5$ and each d_0 in $0.0025 \times [1, 3, 5, 7, 9]$, we take $\theta(1\!:\!n/4) = d_0 + (c_0 - d_0) \times (4i/n)$ and $\theta(1+n/4\!:\!n) = d_0 + (c_0 - d_0) \times [4i/(3n)]^2$. We further normalize two sub-vectors $\theta(1\!:\!n/4)$ and $\theta(1+n/4\!:\!n)$ by dividing their vector ℓ^1-norm, respectively. Finally, the whole vector θ is normalized by $\theta = 0.8 \times \theta/\theta_{\max}$. In this setting, $\theta_{\max} = c_0$ and $\theta_{\min} \approx d_0$, so when d_0 decrease, the level of heterogeneity increase.

The performance of oPCA, nPCA, SCORE and Newman's SM are reported in Figure 6. In this experiment, $n = 4000$ and PL is found to be rather time consuming so we do not include it in the experiment. The results suggest that the problem is increasingly harder as d_0 decrease, and SCORE has the best performance among all 4 methods.

EXPERIMENT 5. In this experiment, we take $(n, k, rep) = (1200, 2, 25)$, and θ to be the vector with $\theta(i) = d_0 + (c_0 - d_0) \times (i/n)^2$, where $d_0 = 0.025$ and $c_0 = 0.5$, and investigate the performances of all the above procedures. The experiment contains two sub-experiments: Experiment 5(a) and 5(b):

Experiment 5(a). In this experiment, we investigate how the ratios between the diagonals and off-diagonals of P (which can be thought of as a measure for how the classes are separated) affect the performances of all 5 procedures. The label vector ℓ is randomly generated by $(\ell_i - 1) \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)$. For each a in $0.1 \times (3, 4, 5, 6, 7)$, we take P as the 2×2 matrix which has 1 on the diagonals and a on

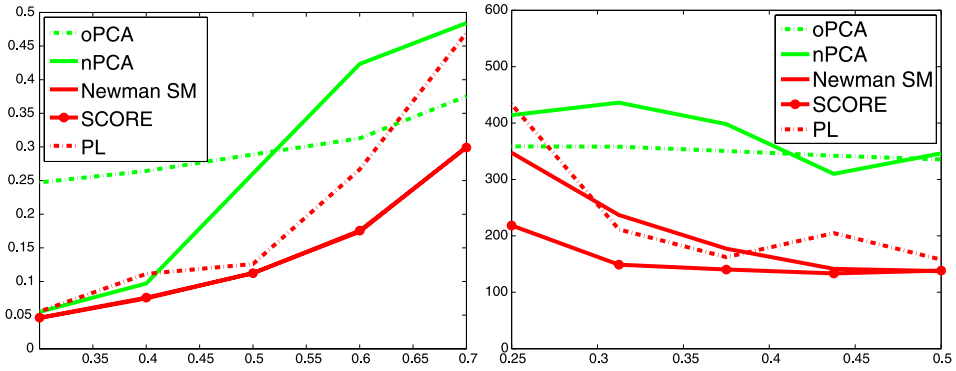

FIG. 7. *Left*: *Experiment* 5(a). *x-axis*: *parameter a* (*the off-diagonals of P*); *y-axis*: *Hamming error rates in community detection*; *the curve for Newman SM is very close to that of SCORE and is hard to distinguish visually. Right*: *Experiment* 5(b). *x-axis*: *parameter* ε (*probability that a node comes from the community with a smaller size*); *y-axis*: *Hamming error rates*.

the off-diagonals. The results are reported in Figure 7 (left panel), which suggest that the problems become increasingly harder as a increase. In this experiment, SCORE and Newman's SM have very similar error rates, which are smaller than those of nPCA, oPCA and PL.

Experiment 5(b). In this experiment, we investigate how the probabilities of classes affect the performances of the procedures. We take P as the 2×2 matrix which has 1 on the diagonals and 0.5 on the off-diagonals. For each ε in $0.25 + 0.0625 \times [0, 1, 2, 3, 4]$, we generate the label vector randomly by $(\ell_i - 1) \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\varepsilon)$. The results are in Figure 7 (right panel), which suggest that the problems become easier as ε increase (so that two communities become increasingly more balanced), and that SCORE has the best performance among all 5 procedures.

EXPERIMENT 6. In this experiment, we compare SCORE with PL and PL-SCORE (note that these are two versions of PL, where in the first one, the initial label vector is generated randomly, and in the second one, the initial label vector is set by the estimate of SCORE). We also include the pseudo Likelihood algorithm (pseudoL) by Amini et al. [4] for comparison.

For parameters (a, c_0, d_0) and $\gamma \in \{0.5, 1, 1.5, 2, 2.5\}$, we take $(n, k, rep) = (1200, 2, 100)$, let P be the 2×2 matrix which has 1 on the diagonal and a elsewhere, generate the label vector ℓ by $(\ell_i - 1) \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)$, and let $\theta(i) = d_0 + (c_0 - d_0) \times (i/n)^\gamma$, $1 \le i \le n$. The experiment contains two parts, Experiment 6(a) and 6(b), where we take (a, c_0, d_0) to be $(0.8, 0.5, 0.2)$ and $(0.6, 0.5, 0.025)$, respectively.

The results are summarized in Figure 8, which suggest the following. First, SCORE significantly outperforms pseudoL in the setting of Experiment 6(a) (left

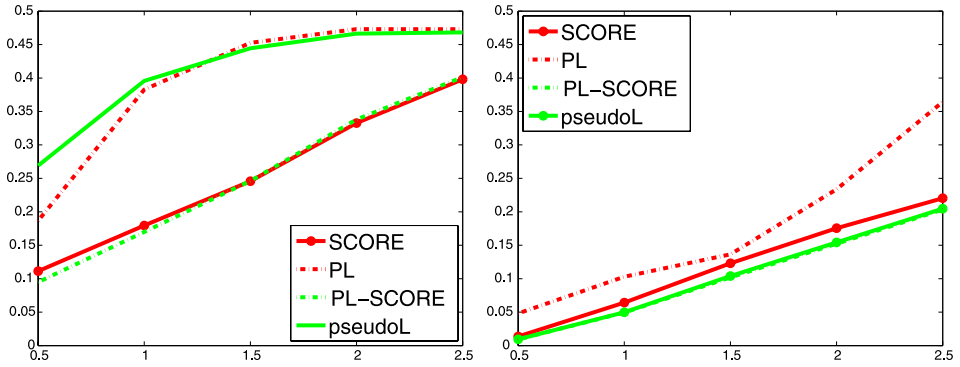

FIG. 8. *Experiment* 6(a) (*left*; $a = 0.8$) *and* 6(b) (*right*; $a = 0.6$). *x-axis*: *parameter* γ; *y-axis*: *Hamming error rates in community detection. In Experiment* 6(b), *the curve associated with PL-SCORE is very close to that of pseudoL and it is hard to distinguish them.*

panel) and mildly underperforms pseudoL in the setting of Experiment 6(b) (right panel). Second, among the four procedures, PL-SCORE has the smallest error rates. In contrast, the error rates of PL are much larger, which suggests that for the initial label vector in the tabu algorithm, letting it be the estimate of the SCORE is substantially better than letting it be a randomly generated label vector. However, the improvements of PL-SCORE over SCORE are negligible in the setting of Experiment 6(a) (in fact, the improvements are negative in some cases) and are only mild in the setting of Experiment 6(b). The possible reason is that SCORE may have already provided a good estimate in such settings. Note that PL-SCORE is computationally more demanding than SCORE, especially when n is large.

In conclusion, the error rates of SCORE are smaller than those of oPCA, nPCA, and Newman's SM in many settings. The error rates of SCORE are similar to those of PL in some settings, but can also be better in other settings. In the numeric experiments considered here, we find that PL is computationally slower than oPCA, nPCA and Newman's SM. It is noteworthy that SCORE is conceptually simple and easy to implement. This leaves ample space for future work where we can improve SCORE in various directions, and combine SCORE with other methods (such as PL) for better algorithms. On the other hand, we must note that that SCORE does not always have the best performance; see, for example, Figure 5. It would be very interesting to study under which settings SCORE has the best performance, and under which settings, some other methods presented in this paper may have a better performance. Seeming, there is no easy answer to this question, and addressing it needs rather delicate analysis. For reasons of space, we leave such studies to the future.

4. Discussion. We propose SCORE as a novel spectral approach to community detection with a DCBM. The method is largely motivated by the observation

that the degree heterogeneity parameters of the DCBM are largely ancillary. If we obtain the first K leading eigenvectors of the adjacency matrix and arrange them in an $n \times K$ matrix $\hat{R}$, then the heterogeneity can be largely removed by applying a scaling-invariant mapping to each row of $\hat{R}$. SCORE is one of such methods.

An interesting feature of SCORE is that it does not attempt to estimate the heterogeneity parameters or to correct the heterogeneity. This is especially important when many nodes of the network are sparse, in which case the estimates of the heterogeneity parameters are inaccurate and the estimation errors can largely affect subsequent studies. Additionally, when we tend to correct the heterogeneity effects, we also tend to inflate the noise level, resulting a smaller Signal Noise Ratio in spectral analysis.

The theoretic conditions required for the success of the SCORE is very different from that in Zhao [37]. Zhao et al. [37], page 6, models the heterogeneity parameters as random variables that assume only finite values and have the same means, which is relatively restrictive. For practical concerns, we choose not to put much structural assumptions on the heterogeneity parameters, and model them as nonstochastic vectors that may vary with the size of the network, and we only need some conditions on regularity and moderate deviations for consistency. Additionally, Zhao et al. [37] impose certain conditions on the $K \times K$ core matrix P which we do not require.

The work can be extended to various directions. First, SCORE can be extended to a large class of methods that utilize a scaling-invariant mapping that operates on $\hat{R}$ row by row. Second, the DCBM can be generalized to more realistic models, where the spectral methods could continue to work well. For example, in work in progress [21], we have extended the method to bipartite networks and have seen nice results on the 110th Senate and House voting network. Third, the ideas developed here can be used to tackle some other problems in network analysis (e.g., linkage prediction [14]).

In this paper, we have assumed the number of communities K as known. In many applications (e.g., the web blogs data and the karate club data), we have a good idea on how many perceivable communities are there, and such an assumption makes sense. In some other applications (e.g., coexpression genetic network [25, 26]), the situation is more complicated and we may not have a good idea on how large K is. Community detection for the case where K is unknown is an unsolved problem, even for low-dimensional clustering problems. A possible approach is to try our methods for different K, and see for which K the results give the best fit to the data. The study along this line is nontrivial and we leave it to the future work.

We have imposed two regularity conditions in Section 2.4: (2.12)–(2.13). These conditions ensure that the gaps between the leading eigenvalues of Ω is lower bounded by $c\|\theta\|^2$ for some constant $c > 0$, which in turn are used to bound the differences between the eigenvalues/eigenvectors of Ω and those of A. On one

hand, such conditions can be relaxed, and the lower bound $c\|\theta\|^2$ can be replaced by a term of smaller order. On the other hand, when the gap between two adjacent (nonzero) eigenvalues of Ω is small (e.g., when the network is sparse [24]), it is known that the associated eigen-vectors are unstable and are vulnerable to noise corruption; see [15], Section 7.2.5. In this case, it is usually hard to ensure the stability of the eigenvectors of Ω, without more assumptions. The study along this line involves delicate theory on the stability of eigenvectors [15], and is nontrivial. For this reason, we leave it for future study.

Intellectually, this work is connected to the recent interest on low-rank matrix recovery and matrix completion; see, for example, [7]. In the area of low-rank matrix recovery, there is a tendency of using the so-called methods of nuclear-norm penalization to replace spectral clustering. Our finding says the contrary: spectral clustering can be effective, and what it takes to make it effective is some careful adjustment. Such findings are resonated in our forthcoming manuscript [20], where we show that spectral clustering can be very effective in cancer clustering with micro-array data provided that we add a careful feature selection step. In spirit, this is connected to several recent papers by Boots and Gordon; see, for example, [5].

Acknowledgements. The author would like to thank Joel Tropp and Roman Vershynin for helpful pointers, and Stephen E. Fienberg and Peter G. Hall for encouragement.

SUPPLEMENTARY MATERIAL

Supplementary material for "Fast communication detetion by SCORE" (DOI: 10.1214/14-AOS1265SUPP; .pdf). Owing to space constraints, the technical proofs are relegated a supplementary document. The supplementary document contains three parts: Appendices A, B and C. Appendix A discusses possible variants of SCORE, Appendix B contains proofs of the main theorems and Appendix C contains proofs of the secondary lemmas.

REFERENCES

[1] ADAMIC, L. and GLANCE, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the* 3*rd International Workshop on Link Discovery* 36–43. ACM, New York.

[2] AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859

[3] BAI, Z. and SILVERSTEIN, J. W. (2009). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, New York. MR2567175

[4] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[5] BOOTS, B. and GORDON, G. (2011). Online spectral identification of dynamical systems. In *NIPS Workshop on Sparse Representation and Low-Rank Approximation*. Sierra Nevada, Spain.

[6] BOX, G. E. P. and DRAPER, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York. MR0861118
[7] CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000
[8] CHAUDHURI, K., FAN, C. and TSIATAS, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition of model. *J. Mach. Learn. Res.* **35** 1–23.
[9] CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253
[10] CHUNG, F. R. K. (1997). *Spectral Graph Theory*, 1st ed. *CBMS Regional Conference Series in Mathematics* **92**. AMS, Providence, RI. MR1421568
[11] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Bulk universality for generalized Wigner matrices. *Probab. Theory Related Fields* **154** 341–407. MR2981427
[12] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581
[13] FISHKIND, D., SUSSMAN, D., TANG, M. and VOGELSTEIN, J. (2012). Consistent adjacency-spectral clustering partitioning for the stochastic model when the model parameters are unknown. Available at arXiv:1205.0309.
[14] GOLDENBERG, A., ZHENG, A., FIENBERG, S. and AIROLDI, E. (2009). A survey of statistical network models. *Faund. Trends Mach. Learn.* **2** 129–233.
[15] GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720
[16] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. Springer, New York. MR1851606
[17] HOFF, P. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*. Cambridge.
[18] HORN, R. A. and JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR0832183
[19] JIN, J. (2014). Supplement to "Fast community detection by SCORE". DOI:10.1214/14-AOS1265SUPP.
[20] JIN, J. and WANG, W. (2012). Optimal spectral clustering by Higher Criticism thresholding. Working manuscript.
[21] JIN, J. and ZHANG, Q. (2012). New spectral methods for community detection with bipartite networks. Working manuscript.
[22] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206
[23] KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data*: *Methods and Models*. Springer, New York. MR2724362
[24] KRZAKALA, F., MOORE, C., MOSSEL, E., NEEMAN, J., SLY, A., ZDEBOROVA, L. and ZHANG, P. (2012). Spectral redemption: Clustering sparse networks. Available at arXiv:1306.5550.
[25] LIU, H., XU, M., GU, H., GUPTA, A., LAFFERTY, J. and WASSERMAN, L. (2011). Forest density estimation. *J. Mach. Learn. Res.* **12** 907–951. MR2786914
[26] NAYAK, R., KEARNS, M., SPIELMAN, R. and CHEUNG, V. (2009). Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Res.* **19** 1953–1962.
[27] NEWMAN, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* (3) **74** 036104, 19. MR2282139
[28] NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582.
[29] PERRY, P. and WOLFE, P. (2012). Null models for network data. Available at arXiv:1201.5871.

[30] PRAKAH, B. A., SRIDHARAN, A., SESHADRI, M., MACHIRAJU, S. and FALOUTSOS, C. (2010). Eigenspokes: Surprising patterns and scalable community chipping in large graphs. In *Advances in Knowledge Discovery and Data Mining* 435–448. Springer, Berlin.

[31] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

[32] TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

[33] TUKEY, J. W. (1965). Which part of the sample contains the information? *Proc. Natl. Acad. Sci. USA* **53** 127–134. MR0172387

[34] YAN, X., JENSEN, J., KRZAKALA, F. et al. (2014). Model selection for degree-corrected block model. *J. Stat. Mech. Theor. Exp.* **2014** P05007.

[35] ZACHARY, W. (1977). An information flow model for conflict and fission in small groups. *J. Anthropo. Res.* **33** 452–473.

[36] ZHANG, S. and ZHAO, H. (2012). Community identification in networks with unbalanced structure. *Phys. Rev. E* **85** 066114.

[37] ZHAO, Y., LEVINA, L. and ZHU, J. (2011). Consistency of community detection in network under degree-corrected stochastic block models. Available at arXiv:1110.3854v3.

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: jiashun@stat.cmu.edu