Submitted to the Annals of Statistics
arXiv: math.ST/1212.FILL IN HERE

FAST COMMUNITY DETECTION BY SCORE

BY JIASHUN JIN*

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 **S**pectral **C**lustering **O**n **R**atios-of-**E**igenvectors (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 X be the adjacency matrix of the network. We first obtain the K leading eigenvectors, 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 much more satisfactory than those by the classical spectral methods. Also, compared to modularity methods, SCORE is computationally much faster and has smaller error rates.

We develop a theoretic framework where we show that under mild conditions, the SCORE stably yields successful 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, partic-

*Supported in part by NSF grant DMS-1208315.

AMS 2000 subject classifications: Primary 62H30, 91C20; secondary 62P25.

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

ularly 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).

Large complex network data sets pose an array of new problems where statisticians can make substantial contributions and scientific discoveries. In particular, solidly founded principled approaches are needed to make predictions and to detect structures.

A problem that is of major interest is "network community detection" [7, 8, 9, 14, 22, 23, 24, 25, 32, 33]. 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 blog 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" [5], 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 [3]. 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 more or less the same degrees. However, this conflicts with the empirical observation that in many natural networks, the degrees follow approximately a power-law distribution [12, 19].

In a different line of development, there are the p^* model and the exponential random graph model (ERGM) [12]. 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 [18], which has become increasingly popular in network analysis [7, 8, 18, 30, 33]. 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 [12, 19] 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 X 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

- X is symmetric, with 0 on the diagonals (so there is no self connections);
- The coordinates on the upper triangular $\{X(i,j) : 1 \le i < j \le n\}$ are independent Bernoulli random variables satisfying

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

As n ranges, 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 notations are 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 these notations, we can rewrite

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

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

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

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 "Wigner"-type matrix [29]. Note also that, while it seems $\mathbf{1}_k$ are known, they are not as they depend on the unknown community partition that is of primary interest to us in this paper.

1.2. *Where is the information: spectral analysis heuristics.* In [28], John Tukey mentioned an idea that can be served 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

$$X = \Omega - \mathrm{diag}(\Omega) + W.$$

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 $\mathrm{diag}(\Omega)$ 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 $\theta^{(k)}$ be the $n \times 1$ vector such that

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

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 k-th community.

Definition 1.1. *We call an eigenvalue λ of a matrix A simple if the algebraic multiplicity of λ is 1 [15].*

In most part of the paper, the eigenvalues of interest are simple. The following lemma is a special case of Lemma 2.1, which is proved in Section 7 (note that Θ is the diagonal matrix as in (1.1)).

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

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

and the associated eigenvectors η_1 and η_2 (with possible non-unit norms) are

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

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

$$r_0(i) = \begin{cases} 1, & i \in V^{(1)}, \\ -\Big(\frac{ad_1^2 - cd_2^2 + \sqrt{(ad_1^2 - cd_2^2)^2 + 4b^2d_1d_2}}{2bd_1d_2}\Big)^2, & i \in V^{(2)}. \end{cases} \tag{1.2}$$

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 many approaches (e.g., the modularity approach to be introduced below) 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 **S**pectral **C**lustering **O**n **R**atios-of-**E**igenvectors (SCORE).

- *(a).* Let $\hat{\eta}_1$ and $\hat{\eta}_2$ be the two unit-norm eigenvectors of X 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 [13]. 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 some times.

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.

1.4. *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 16714 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 [31]. 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 [31, 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 $\hat{r}$.

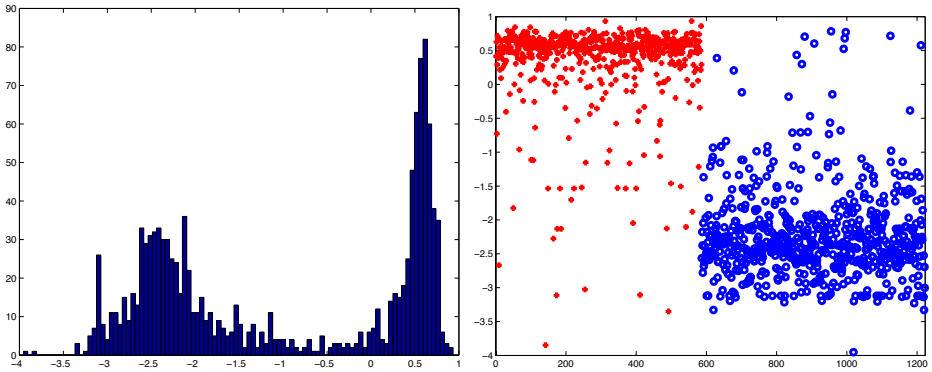

FIG 1. *The vector $\hat{r}$ (web blogs data). Left: histogram of $\hat{r}$. Right: plot of the entries of $\hat{r}$ versus the node indices (red cross: liberal; blue circle: conservative).*

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" [13]).

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 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 [31], 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.

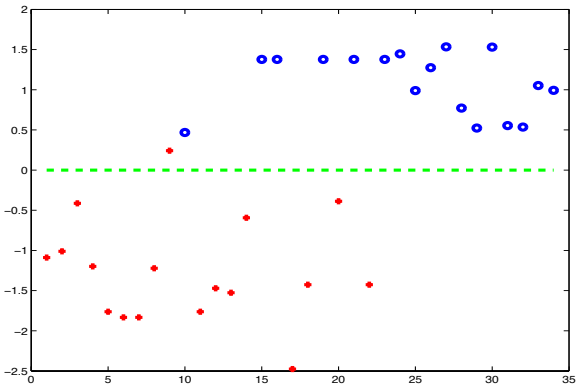

FIG 2. *Plot of the entries of $\hat{r}$ versus the node indices (results are based on karate club network; red cross: Mr. Hi's group; blue circle: John's group).*

1.5. *Comparison with classical spectral clustering methods.* As a spectral clustering approach, the success of the SCORE prompts the following question: would classical spectral methods work well too? By classical spectral method, we mean the following procedure.

- *(a').* Obtain the two leading (unit-norm) eigenvectors $\hat{\eta}_1$ and $\hat{\eta}_2$ of X.
- *(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 [25].

- *(a").* Obtain an $n \times n$ diagonal matrix S defined by $S(i,i) = \sum_{j=1}^{n} X(i,j)$, $1 \le i \le n$.
- *(b").* Apply (a')-(b') to $S^{-1/2} X S^{-1/2}$.

We call the two procedures *ordinary PCA (oPCA)* and *normalized PCA (nPCA)*, respectively; PCA stands for Principle Component Analysis.

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 X, based on the web blogs data. The coordinates of two vectors are highly skewed to the right, reflecting serious degree heterogeneity.

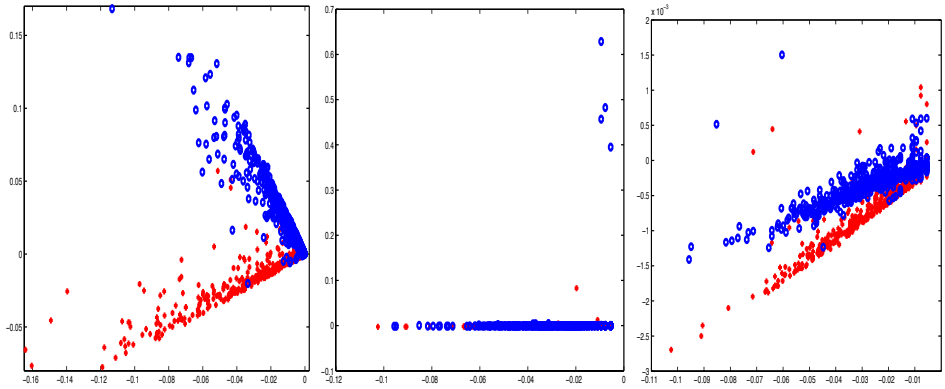

FIG 3. *Left: plot of the first leading eigenvector of X (x-axis) versus the second leading eigenvector of X (y-axis). Middle: plot of the first leading eigenvector of $S^{-1/2}XS^{-1/2}$ (x-axis) versus the second leading eigenvector of $S^{-1/2}XS^{-1/2}$ (y-axis). Right: zoom in of the middle panel.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

Somewhat surprisingly, though nPCA intends to correct degree heterogeneity, the correction is not particularly successful. 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 $S^{-1/2}XS^{-1/2}$. It is seen that some of the coordinates of $\hat{\eta}_2$ are unduly large, compared to the remaining coordinates.

The underlying reason for the unsatisfactory behavior of nPCA is twofold. First, many nodes in the web blogs data have very small degrees, so S only contain relatively poor estimates for the heterogeneity parameters. Second, even when the heterogeneity parameters are given, it is not always helpful to correct them as in nPCA, because when we try to correct the degree heterogeneity, we tend to increase the noise level at the same time. See Section 2.8 for detailed discussion.

1.6. *Comparison with modularity methods.* Modularity methods are well-known approaches to community detection. The methods have many variants, including the well-known approach by [18]. For this paper, we use the recent approach by Zhao *et al* [33], which can be viewed as a variant of the modularity method in [18]; the approach is reported to have similar behavior to that in [18].

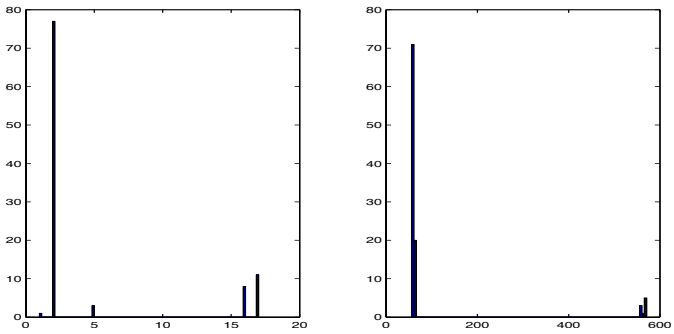

FIG 4. *Histogram of errors by modularity methods for the karate data (left) and the web blogs data. The results are based on* 100 *independent repetitions.*

In principle, modularity methods are computationally NP-hard [3], as it searches exhaustively over all possible community partitions, and pick the one that optimizes the so-called modularity functional. To mitigate this difficulty, many heuristic algorithms are proposed to approximate the theoretic optimizer. In particular, Zhao *et al.* [33] proposes a heuristic algorithm of this type which they call the *tabu algorithm*.

Compared to the SCORE and the classical spectral approaches, the tabu algorithm is computationally much more expensive, and is increasingly so when the size or complexity of the network increases. The tabu algorithm is also relatively unstable: like most modularity methods, the tabu algorithm depends on the initial guess of the community partition, and the algorithm may not converge to the true partition with a "bad" initial guess. In theory, the instability can be alleviated by increasing the number of searches, but this is at the expense of substantially longer computational time.

The performance of the tabu algorithm for the karate network and web blogs network is illustrated in Figure 4 (left: karate; right: web blogs), where for each network, the histogram is based on 100 independent repetitions (the error rates are random for each depends on the initial guess of the community partition, generated randomly).

The most prominent problem of the tabu algorithm (and modularity methods in general) 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

	SCORE			PCA		Modularity
	$t=0$	$t=-0.7$	$t=-0.6$	ordinary	normalized	
web blogs	82	58	55	437	600	104.5 (SD: 145.4)
karate	1	1	0	1	1	4.9 (SD: 5.7)

TABLE 1

Comparison of number of errors. For SCORE, thresholding at $t=-0.7$ is the same as k-means; $t=-0.6$ is the ideal threshold. The result of modularity method depends on the starting point and is random, where mean and standard deviation (SD) are computed based on 100 *independent repetitions. The web blogs data has* 1222 *nodes and the karate data has* 34 *nodes.*

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.

In summary, it is fair to say that compared to SCORE, the tabu algorithm underperforms, especially as it is computationally much more expensive.

That the tabu algorithm 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 tabu 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.

1.7. *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 four algorithms: SCORE, oPCA, nPCA, and tabu. SCORE behaves much more satisfactory than the two PCA methods. It also outperforms the tabu algorithm: it is much faster, more stable, and has a smaller error rate.

The paper is closely related to [25] (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 [25] is on BM where K is large. Our analysis is also very different from that in [33], for we use a much broader model for the degree heterogeneity parameters $\theta(i)$.

1.8. *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 Sections 2.7, we address the stability of the SCORE, and in Section 2.8, we re-investigate the nPCA by comparing the so-called Signal Noise Ratio of the nPCA and that of the oPCA. In Section 3, we suggest some extensions of the SCORE. The main results are proved in Section 4, where we outline main technical devises required for the proofs. Numeric investigation is continued in Section 5, where we compare the SCORE, two PCA approaches, and the tabu algorithm on simulated data. Section 6 discusses connection between SCORE and existing literatures. Secondary lemmas are proved in Section 7.

1.9. *Notations.* 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 $1/c_0 \leq a_n/b_n \leq c_0$ for sufficiently large n.

In this paper, the notations 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 $\Theta(i, i) = \theta(i)$, $1 \leq i \leq n$. Also, $\theta_{min} = \min_{1\leq i\leq n}\{\theta(i)\}$ and $\theta_{max} = \max_{1\leq i\leq n}\{\theta(i)\}$. For a vector ξ, when all coordinates are positive, we use $OSC(\xi)$ to denote the oscillation $\max_{1\leq i,j\leq n}\{\xi(i)/\xi(j)\}$. 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. We start by describing the SCORE and the DCBM for K-community networks, followed by spectral analysis on Ω and on X, as well as the consistency of SCORE. We then address the stability of SCORE and conclude by re-investigating the normalized PCA.

Throughout the paper, $K \geq 2$ is a known integer. See Section 6 for discussion on the case where K is unknown.

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:

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

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

$$\mathcal{M}_{n,K-1,K} = \{M : n \times (K-1) \text{ matrix that has } \le K \text{ distinct rows}\}. \tag{2.3}$$

For convenience, we use the following terminology, so that whenever we say "leading eigenvectors" or "leading eigenvalues", we are comparing the *magnitudes* of the eigenvalues, neglecting the ± 1 signs.

DEFINITION 2.1. *Fix $1 \le k \le n$ and an $n \times n$ symmetric X. We say λ_k is the k-th leading eigenvalue of X if λ_k has the k-th largest magnitude among all eigenvalues of X, and we say ξ_k is the k-th leading eigenvector if it is an eigenvector of X associated with λ_k.*

The SCORE for K-community networks contains the following steps.

- Obtain the K (unit-norm) leading eigenvectors of X: $\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_K$.
- Fixing a threshold T_n, define an $n \times (K-1)$ matrix $\hat{R}^*$ such that for all $1 \le i \le n$ and $1 \le k \le K-1$,
  (2.4)
  $$\hat{R}^*(i,k) = \begin{cases} \hat{R}(i,k), & \text{if } |\hat{R}(i,k)| \le T_n, \\ T_n, & \text{if } \hat{R}(i,k) > T_n, \\ -T_n, & \text{if } \hat{R}(i,k) < -T_n, \end{cases} \quad \text{where } \hat{R}(i,k) = \frac{\hat{\eta}_{k+1}(i)}{\hat{\eta}_1(i)}.$$
- Let M^* be the matrix satisfying
  $$M^* = \operatorname{argmin}_{M \in \mathcal{M}_{n,K-1,K}} \|\hat{R}^* - M\|_F^2.$$
  Write $M^* = [m_1, m_2, \ldots, m_n]'$ so that m_i' is the i-th 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 $1 \le i_1 < \ldots < i_K \le n$. 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 less than K distinct rows, we let $\hat{V}^{(\ell)} = \emptyset$ for all $(k+1) \le \ell \le K$. Second, the choice of the threshold T_n is flexible, and for convenience, we take

$$T_n = \log(n) \tag{2.5}$$

in this paper. We impose thresholding in (2.4) 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 (2.4) is rarely necessary. Last, provided that the largest K eigenvalues are all simple, the vectors $\hat{\eta}_k$ 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.

2.2. *DCBM when there are K communties.* The DCBM for K-community networks is a direct generalization of the DCBM for two-community networks. As before, we assume that the adjacency matrix X satisfies

$$X = E[X] + W, \qquad W = X - E[X], \quad E[X] = \Omega - \mathrm{diag}(\Omega), \tag{2.6}$$

where Ω is a symmetric matrix, and W is the symmetric "Wigner"-type matrix where all diagonals are 0 and the upper triangular entries are independent centered-Bernoulli random variables with parameters $\Omega(i,j)$.

In the core of DCBM is a $K \times K$ matrix

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

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

$$\Omega(i,j) = \theta(i)\theta(j)A(k,\ell), \qquad \text{if } i \in V^{(k)} \text{ and } j \in V^{(\ell)}. \tag{2.7}$$

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

$$\max_{1\le i,j\le K} A(i,j) = 1, \qquad 0 < \theta_{min} \le \theta_{max} \le g_0, \tag{2.8}$$

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

$$A \text{ is symmetric, non-singular, non-negative and irreducible.} \tag{2.9}$$

A matrix is non-negative if all coordinates are non-negative. See [15, 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, A) as fixed. Consequently, there is a constant $C = C(A) > 0$ such that $\|A^{-1}\| \le C$, where $\|\cdot\|$ denotes the spectral norm.

DEFINITION 2.2. *We call Model (2.6)-(2.9) the K-community DCBM.*

The DCBM we use is similar to that in [33] (see also [18]), but is different in important ways. In their asymptotic analysis, Zhao *et al.* [33], model $\theta(i)$ as random variables that have the same means and take only finite values. In our setting, we treat $\theta(i)$ as non-random and only impose some mild regularity conditions and moderate deviations conditions (see below). Additionally, Zhao *et al.* [33] need certain conditions on A which we don't require. For example, in the special case of $K = 2$, they require A to be positive definite, but we don't. See [33, Page 7] for details.

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

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

Similar as before, for $1 \le k \le K$, we let $\theta^{(k)}$ be the $n \times 1$ vectors such that

$$\theta^{(k)}(i) = \theta(i) \text{ or } 0, \quad \text{according to } i \in V^{(k)} \text{ or not.} \tag{2.10}$$

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 DAD, where D and A are as above. The following lemma characterizes the leading eigenvalues and leading eigenvectors of Ω, and is proved in Section 7.

LEMMA 2.1. *Suppose all eigenvalues of DAD are simple. Let $\lambda_1/\|\theta\|^2$, $\lambda_2/\|\theta\|^2$, ..., $\lambda_K/\|\theta\|^2$ 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 λ_1, λ_2, ..., λ_K, with the associated (unit-norm) eigenvectors being*

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

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* X. In this section, we characterize the leading eigenvalues and leading eigenvectors of X.

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 DAD 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

$$\big(\log(n)\theta_{max}\|\theta\|_1\big)/\|\theta\|^4 \to 0, \qquad \text{as } n \to \infty. \tag{2.11}$$

Combining (2.11) with basic algebra, it follows that

$$\log(n)/\|\theta\|^2 \to 0, \qquad \big(\log(n)\|\theta\|_1\|\theta\|_3^3\big)/\|\theta\|^6 \to 0, \tag{2.12}$$

which are frequently used in the proof section. The following lemma characterizes the spectral norm of $W - \mathrm{diag}(\Omega)$.

LEMMA 2.2. *If (2.11) holds, then with probability at least* $1 + o(n^{-3})$,

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

Lemma 2.2 is proved in Section 7, where the recent result by [27] on matrix-form Bernstein inequality is very helpful.

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

$$\mathrm{eigsp}(A) = \min_{1\leq i\leq K-1} |\lambda_{i+1} - \lambda_i|, \qquad \lambda_1 > \lambda_2 \ldots > \lambda_K. \tag{2.13}$$

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

$$\mathrm{eigsp}(DAD) \geq C. \tag{2.14}$$

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

$$\max_{1\leq i,j\leq K}\{\|\theta^{(i)}\|/\|\theta^{(j)}\|\} \leq C. \tag{2.15}$$

As a result, D has a bounded condition number. Together with (2.9), this implies that all eigenvalues of DAD are bounded away from 0 by a constant $C > 0$. Combining these with Lemma 2.1, the following lemma is a direct result of Lemma 2.2 and basic algebra (e.g., [2, Page 473]), the proof of which is omitted.

LEMMA 2.3. *Consider a DCBM where (2.11), (2.14), and (2.15) hold. Let* $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K$ *be the leading eigenvalues of* X*, and let* $\lambda_1/\|\theta\|^2, \lambda_2/\|\theta\|^2, \ldots, \lambda_K/\|\theta\|^2$ *be the nonzero eigenvalues of* DAD*, both sorted descendingly in magnitudes. With probability at least* $1+o(n^{-3})$*, the* K *leading eigenvalues of* X *are all simple, and*

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

A direct result of Lemma 2.3 is that, with probability at least $1 + o(n^{-3})$,

$$\lambda_k \asymp \|\theta\|^2, \qquad \text{for all } 1 \leq k \leq K. \tag{2.16}$$

This result is frequently used in the proof section.

Next, we study the leading eigenvectors. From now on, we assume Conditions (2.14)-(2.15) hold, and let $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K$ be the K leading eigenvalues as in Lemma 2.3. For $1 \leq k \leq K$, whenever $\hat{\lambda}_k$ is not an eigenvalue of $W - \mathrm{diag}(\Omega)$, let $B^{(k)}$ be the $K \times K$ matrix

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

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 (2.11), (2.14), and (2.15) hold. Let $\{\hat{\lambda}_k\}_{k=1}^K$ be the eigenvalues of X with the largest magnitudes. There is an event with probability at least $1 + o(n^{-3})$ such that over the event, for each $1 \leq k \leq K$, $\hat{\lambda}_k$ is simple, and the associated eigenvector is given by*

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

where $\hat{a}_k$ is an (unit-norm) eigenvector of $DADB^{(k)}$, and $\hat{\lambda}_k/\|\theta\|^2$ is the unique eigenvalue of $DADB^{(k)}$ 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 [15, 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 (2.4). 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 non-stochastic counterpart of $\hat{R}^*$ by

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

note that $\|\eta_k\| = 1$. Unlike $\hat{R}$, $|R(i,k)| \leq C$ for all i and k (see Lemma 2.1), so it is unnecessary to impose thresholding as that in (2.4).

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).

Towards this end, we need to put some conditions on θ (or equivalently on Θ), namely the Moderate Deviation conditions on Vectors (MDV) and the Moderate Deviation conditions on Matrices (MDM), which are used to control the moderate deviations of norms of vectors, say, $W\theta$, and norms of matrices, say, $\Theta^{-1}W$, respectively. The MDV requires that

(2.17)
$$\sum_{i=1}^{n}\max\big\{\theta(i),\frac{\log(n)\theta_{max}^2}{\|\theta\|_3^3}\big\}\le C\|\theta\|_1,\ \sum_{i=1}^{n}\max\big\{\frac{1}{\theta(i)},\frac{\log(n)\theta_{max}^2}{\theta^2(i)\|\theta\|_3^3}\big\}\le\sum_{i=1}^{n}\frac{C}{\theta(i)},$$

for some constant $C>0$, and the MDM requires that

$$\frac{\log(n)}{\theta_{min}^2}\le\max\{\frac{1}{\theta_{min}}\|\theta\|_1,\theta_{max}\sum_{i=1}^{n}\frac{1}{\theta(i)}\}. \tag{2.18}$$

The following short hand notation is used many times below.

$$err_n=\frac{\|\theta\|_3^3}{\|\theta\|^6}\Big[\sum_{i=1}^{n}\frac{1}{\theta(i)}+\frac{\log(n)}{\theta_{min}}\big(\frac{\|\theta\|_1}{\|\theta\|^2}\big)^2\Big]. \tag{2.19}$$

The following theorem is the corner stone for characterizing the behavior of SCORE, and is proved in Section 4.

THEOREM 2.1. *Consider a DCBM where the three regularity conditions (2.11), (2.14), and (2.15), and the two moderate deviation conditions (2.17) and (2.18) hold. If $T_n=\log(n)$ is as in (2.5), then as $n\to\infty$, with probability at least $1+o(n^{-2})$, we have that*

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

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

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

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

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

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

Especially, let $\hat{\ell}^{sc}=\hat{\ell}^{sc}(X,T_n,n)$ 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. Towards this end, let

$$S_K = \{\pi : \pi \text{ is a permutation of the set } \{1, 2, \ldots, K\}\}. \tag{2.20}$$

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

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

With these notations, 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 k-th community:

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

The following theorem is proved in Section 4, which says that SCORE is consistent under mild conditions, and is the main result of the paper.

THEOREM 2.2. *Consider a DCBM where both the three regularity conditions (2.11), (2.14), and (2.15) and the two moderate deviation conditions (2.17) and (2.18) hold. Suppose as $n \to \infty$,*

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

where err_n is as in (2.19). For the estimated label vector $\hat{\ell}^{sc}$ by the SCORE where the threshold $T_n = \log(n)$ is as in (2.5), there is a constant $C > 0$ such that for sufficiently large n,

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

Similarly, for general T_n, the theorem continues to hold if we replace the right hand side by $C \log(n) T_n^2 err_n$.

2.7. *Stability of SCORE.* The performance of SCORE hinges on the matrix $\hat{R}$ defined in (2.4):

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

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 a lemma on $\hat{\eta}_1$, which says a coordinate of $\hat{\eta}_1$ can never be exactly 0, as long as the network is connected.

LEMMA 2.5. *Let X 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.*

Lemma 2.5 is the direct result of Perron's theorem [15, Section 8.2] on non-negative matrices, so we omit the proof.

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 Section 7 (note that the i-th coordinate of $\Theta^{-1}\eta_1$ is $\eta_1(i)/\theta(i)$).

LEMMA 2.6. *Consider a DCBM where (2.14)-(2.15) 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 Section 7.

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* $1 \le k \le K$,

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

Recall that err_n is defined in (2.19). 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 i-th enty of $\hat{\eta}_1$ well-behaved if $|\hat{\eta}_1(i)/\eta_1(i) - 1| \le c_0$ (say). Let

$$\hat{S} = \hat{S}(c_0, \hat{\eta}_1, \eta_1; X, \Omega, n) = \{1 \le i \le n : |\hat{\eta}_1(i)/\eta_1(i) - 1| \le c_0\}. \tag{2.21}$$

The following lemma is a direct result of Lemma 2.6 and Lemma 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 (2.21). Then with probability at least* $1 + o(n^{-3})$, $|V \setminus \hat{S}| \le C \log(n) err_n$.

Therefore, as long as $\log(n) err_n/n \to 0$ 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 $1 + o(n^{-2})$,

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

Using terminology in the literature of variable selection [10], this says that the SCORE has the *oracle property*, means that it achieves *exact recovery* with overwhelming probabilities.

2.8. *Remarks on normalized PCA (nPCA).* We revisit the normalized PCA (nPCA) discussed in Section 1.5. In the DCBM, on may expect that the normalization in the nPCA reduces the effects of degree heterogeneity. Somewhat surprisingly, this is not always the case. The point is that, when

we try to correct the degree heterogeneity, we tend to inflate the noise level at the same time.

Perhaps the best way to appreciate this is to consider the "oracle" situation where $\theta(i)$ are known. In this case, nPCA is almost equivalent to applying the ordinary PCA (oPCA) to the matrix $\Theta^{-1/2}X\Theta^{-1/2}$. Write

$$\Theta^{-1/2}X\Theta^{-1/2} = \Theta^{-1/2}\Omega\Theta^{-1/2} + \Theta^{-1/2}[W - \mathrm{diag}(\Omega)]\Theta^{-1/2}.$$

We call the ratio between the smallest eigenvalue of $\Theta^{-1/2}\Omega\Theta^{-1/2}$ (in magnitude) and the spectral norm of $\Theta^{-1/2}[W-\mathrm{diag}(\Omega)]\Theta^{-1/2}$ the *Signal Noise Ratio* for nPCA, denoted by $nSNR$.

Similarly, we use $oSNR$ to denote the Signal Noise Ratio in the case without normalization (i.e., the ratio between the smallest eigenvalue of Ω (in magnitude) and the spectral norm of $W-\mathrm{diag}(\Omega)$). If the normalization helps, then we should have $nSNR > oSNR$.

Now, first, by Lemmas 2.1-2.2, we can roughly say that

$$oSNR = \|\theta\|^2/\sqrt{\log(n)\theta_{max}\|\theta\|_1}. \tag{2.22}$$

Additionally, we have the following observations.

- Write $\Theta^{-1/2}\Omega\Theta^{-1/2} = \Theta^{1/2}[\sum_{i,j=1}^K A(i,j)\mathbf{1}_i\mathbf{1}_j']\Theta^{1/2}$. A direct extension of Lemma 2.1 is that, if λ_k is a nonzero eigenvalue of $\Theta^{-1/2}\Omega\Theta^{-1/2}$, then $\lambda_k \asymp \|\theta\|_1$, $1 \le k \le K$.
- A direct extension of Lemma 2.2 is that, $\|\Theta^{-1/2}[W-\mathrm{diag}(\Omega)]\Theta^{-1/2}\| \le C\sqrt{\log(n)n}$, with probability at least $1+o(n^{-3})$.

Therefore, roughly say,

$$nSNR = \|\theta\|_1/\sqrt{\log(n)n}. \tag{2.23}$$

Comparing the right hand sides of (2.22) and (2.23), in order for $nSNR \gg oSNR$, we need

$$\frac{\|\theta\|_1}{\sqrt{\log(n)n}} \gg \frac{\|\theta\|^2}{\sqrt{\log(n)\theta_{max}\|\theta\|_1}} \qquad \Longleftrightarrow \qquad \theta_{max}\|\theta\|_1^3 \gg n\|\theta\|^4.$$

Unfortunately, $\theta_{max}\|\theta\|_1^3 \ll n\|\theta\|^4$ in many situations. Therefore, the normalization in nPCA does not always help, even in such an oracle situation.

3. Variants of SCORE. The key idea underlying the SCORE is that, in a broad context, the K leading eigenvectors of the adjacency matrix X approximate those of the non-stochastic matrix Ω, where the latter are

$$\Theta\Big(\sum_{\ell=1}^k [a_k(\ell)/\|\theta^{(\ell)}\|]\mathbf{1}_\ell\Big), \qquad k = 1, 2, \ldots, K.$$

	SCORE	SCOREq	
		$q = 1$	$q = 2$
web blogs ($n = 1222$)	58	61	64
karate ($n = 34$)	1	1	1

TABLE 2
Comparison of number of errors for the SCORE and the SCOREq.

It is seen that

- The information of the community labels is contained in the term within the bracket, which depends on $\{\theta(i)\}_{i=1}^n$ only through the overall degree intensities $\|\theta^{(k)}\|/\|\theta\|$.
- The diagonal matrix Θ does not contain any information of the community labels.
- Therefore, $\{\theta(i)\}_{i=1}^n$ are almost nuisance parameters, the effect of which can be removed by many *scaling invariant* mappings, to be introduced below.

DEFINITION 3.1. *Let $W \subset R^K$ be a subset such that when $x \in W$, then $ax \in W$ for any $a > 0$. We call a mapping $\mathbb{M}$ from W to R^K scaling invariant if $\mathbb{M}(ax) = \mathbb{M}(x)$ for any $a > 0$ and $x \in W$.*

The following are some examples of scaling invariant mappings.

- (a). $W = \{x \in R^K, x(1) \neq 0\}$, and $\mathbb{M}x = x/x(1)$; $x(1)$ is the first coordinate of x.
- (b). $W = R^K \setminus \{0\}$, $\mathbb{M}x = x/\|x\|_q$, where $q > 0$ is a constant.

Given a scaling invariant mapping $\mathbb{M}$, we have the following extension of SCORE.

- Obtain the K leading (unit-norm) eigenvectors of X. Arrange them in an $n \times K$ matrix $\hat{R}$ as follows so that ξ_i' is the i-th row of $\hat{R}$, $1 \leq i \leq n$:
$$\hat{R} = [\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_K] = (\xi_1, \xi_2, \ldots, \xi_n)'.$$
- Obtain an $n \times K$ matrix $\hat{R}^*$ where the i-th row of $\hat{R}^*$ is $(\mathbb{M}\xi_i)'$.
- Apply the k-means method to the matrix $\hat{R}^*$ for clustering with $\leq K$ classes.

For example, if we view each row of $\hat{R}$ as a point in R^k and apply $\mathbb{M}$ in (a), then we have the $\hat{R}$ matrix in (2.4) associated with the original SCORE (except for that in (2.4), the first column is removed for it is the vector of 1 and is thus non-informative for clustering).

For another example, we take $\mathbb{M}$ as the mapping in (b), and call the resultant procedure SCOREq, where q is the parameter in the mapping $x \to x/\|x\|_q$.

We have investigated the performances of SCORE and SCOREq with the karate club data and the web blogs data, where we pick $q = 1$ and $q = 2$. The performances are largely similar, but SCORE is slightly better for the web blogs data. See Table 2 for details. In Section 5, we further investigate these methods on simulated data.

A natural question is what could be the best $\mathbb{M}$. In principle, this can be studied using the techniques developed in this paper, but the study involves higher order asymptotics and would be rather tedious. For this reason, we leave it for the future work.

4. Proof of the main theorems. In this section, we prove Theorems 2.1-2.2. The key for the proof is Lemmas 2.7-2.8, which contain bounds on $\|\hat{\eta}_k - \eta_k\|$ and $\|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|$, respectively. To show Lemmas 2.7-2.8, we need tight moderate deviation bounds on matrices and vectors involving the noise matrix W. Below, we first describe such moderate deviation bounds, and then give the proofs for Theorems 2.1-2.2. The proofs of Lemmas 2.7-2.8 are given in Section 7.

4.1. *Moderate deviation inequalities on vectors and matrices.* The following theorem is proved in [27], which is the extension of the well-known Bernstein's inequality from the case of random variables to the case of random matrices. Recall that $\|\cdot\|$ denotes the spectral norm.

THEOREM 4.1. *Consider a finite sequence $\{Z_k\}$ of independent, random, symmetric (real-valued) $n \times p$ matrices. Assume that each random matrix satisfies $E[Z_k] = 0$ and $\|Z_k\| \le h_0$ almost surely. Then for all $t \ge 0$,*

$$P(\|\sum_k Z_i\| \ge t) \le (n+p)\exp\Big(-\frac{t^2/2}{\sigma^2 + h_0 t/3}\Big),$$

where $\sigma^2 = \max\{\|\sum_k E[Z_k Z_k']\|, \|\sum_k E[Z_k' Z_k]\|\}$.

The following lemma provides moderate deviation bounds on $\|\Theta^{-1}W\|$.

LEMMA 4.1. *If the Moderate Deviation condition on Matrices (2.18) holds, then $\|\Theta^{-1}W\|^2 \le C\log(n)\max\{\theta_{min}^{-1}\|\theta\|_1, \theta_{max}\sum_{i=1}^n \frac{1}{\theta(i)}\}$, with probability at least $1 + o(n^{-3})$.*

The following lemma provides moderate deviation bounds on the norms of various vectors. The lemma is proved in Section 7, where the classical Bennett's inequality is the key [26] (recall that $\theta^{(k)}$ is defined in (2.10)).

LEMMA 4.2. *If the Moderate Deviation condition on Vectors (2.17) holds, then with probability at least* $1 + o(n^{-3})$, *for all* $1 \le k, \ell \le K$,

- $\|W\theta^{(k)}\|^2 \le C\log(n)\|\theta\|_3^3\|\theta\|_1$.
- $\|\Theta^{-1}W\theta\|^2 \le C\log(n)\|\theta\|_3^3\sum_{i=1}^n \frac{1}{\theta(i)}$.
- $|(\theta^{(k)})'W\theta^{(\ell)}|^2 \le C\log(n)[\|\theta\|_3^6 + \log(n)\theta_{max}^4]$.

We are now ready to show the two main theorems.

4.2. *Proof of Theorem 2.1.* Let $\{\eta_i\}_{i=1}^K$ and $\{\hat{\eta}_i\}_{i=1}^K$ be as in Lemma 2.1 and Lemma 2.4, respectively. Let $\hat{S}$ be the set of "well-behaved" nodes as in (2.21), where $c_0 = 1/2$ for simplicity. By Lemmas 2.7-2.9, there is an event E_n such that $P(E_n^c) = o(n^{-3})$ and that over E_n, for all $1 \le k \le K$,

(4.24)
$$\|\hat{\eta}_k - \eta_k\|^2 \le C\log(n)\|\theta\|_1\|\theta\|_3^3/\|\theta\|^6, \quad \|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|^2 \le C\log(n)err_n,$$

and

$$|V \setminus \hat{S}| \le C\log(n)err_n. \tag{4.25}$$

Especially, combining (4.24) and (2.12), $\|\hat{\eta}_k\| \sim \|\eta_k\| = 1$. To show the claim, it is sufficient to show that over the event E_n, $\|\hat{R}^* - R\|_F^2 \le C\log^3(n)err_n$.

To this end, we write

$$\|\hat{R}^* - R\|_F^2 = U_1 + U_2,$$

where U_1 is the sum of squares of the ℓ^2-norms of all "ill-behaved" rows of $\hat{R}^* - R$, and U_2 is that of all "well-behaved" rows.

Consider U_1. For any $i \notin \hat{S}$ and $1 \le k \le K-1$, it is seen that $|\hat{R}^*(i,k)| \le T_n$ and $|R(i,k)| \le |\eta_{k+1}(i)/\eta_1(i)| \le C$, where $T_n = \log(n)$ and we have used Lemma 2.1 and (2.15). Combining these with Lemma 2.9 and (2.14),

$$U_1 \le C\log^2(n)|V \setminus \hat{S}| \le C\log^3(n)err_n. \tag{4.26}$$

Consider U_2. Recall that for any $i \in \hat{S}$,

$$|\hat{\eta}_1(i)/\eta_1(i) - 1| \le 1/2. \tag{4.27}$$

Since $|R(i,k)| \le C$, $|\hat{R}^*(i,k) - R(i,k)| \le |\hat{R}(i,k) - R(i,k)|$. Write

$$\hat{R}(i,k) - R(i,k) = \frac{\|\hat{\eta}_1\|}{\|\hat{\eta}_{k+1}\|}\frac{\hat{\eta}_{k+1}(i)}{\hat{\eta}_1(i)} - \frac{\eta_{k+1}(i)}{\eta_1(i)} = \frac{\|\hat{\eta}_1\|}{\|\hat{\eta}_{k+1}\|}(I + II + III), \tag{4.28}$$

where

$$I = (\hat{\eta}_{k+1}(i) - \eta_{k+1}(i))/\eta_1(i), \qquad II = \hat{\eta}_{k+1}(i)(\eta_1(i) - \hat{\eta}_1(i))/(\hat{\eta}_1(i)\eta_1(i)),$$

and

$$III = (1 - \|\hat{\eta}_{k+1}\|/\|\hat{\eta}_1\|)\eta_{k+1}(i)/\eta_1(i).$$

Recall that $\|\hat{\eta}_k\| \sim 1$ for all $1 \le k \le K$.

Now, first, by Lemma 2.6,

$$|I| \le C|\hat{\eta}_{k+1}(i) - \eta_{k+1}(i)|/\theta(i). \tag{4.29}$$

Second, write $II = IIa + IIb$, where

$$IIa = \frac{\eta_{k+1}(i)(\eta_1(i) - \hat{\eta}_1(i))}{\hat{\eta}_1(i)\eta_1(i)}, \qquad IIb = \frac{[\hat{\eta}_{k+1}(i) - \eta_{k+1}(i)][\eta_1(i) - \hat{\eta}_1(i)]}{\hat{\eta}_1(i)\eta_1(i)}.$$

By Lemma 2.1, Lemma 2.6, and (4.27), $|\eta_{k+1}(i)/[\hat{\eta}_1(i)\eta_1(i)]| \le C/\theta(i)$ and $|[\eta_1(i) - \hat{\eta}_1(i)]/[\hat{\eta}_1(i)\eta_1(i)]| \le C/\theta(i)$. Therefore, $|IIa| \le C|\hat{\eta}_1(i) - \eta_1(i)|/\theta(i)$ and $|IIb| \le C|\hat{\eta}_{k+1}(i) - \eta_{k+1}(i)|/\theta(i)$. Combining these gives

$$|II| \le C\big[|\hat{\eta}_1(i) - \eta_1(i)| + |\hat{\eta}_{k+1}(i) - \eta_{k+1}(i)|\big]/\theta(i). \tag{4.30}$$

Third, recalling $\|\eta_k\| = 1$ and $\|\hat{\eta}_k\| \sim 1$ for all $1 \le k \le K$ and using triangle inequality,

$$\big|1 - \|\hat{\eta}_{k+1}\|/\|\hat{\eta}_1\|\big| \lesssim \big|\|\hat{\eta}_1\| - \|\hat{\eta}_{k+1}\|\big| \le \big|\|\hat{\eta}_1\| - \|\eta_1\|\big| + \big|\|\hat{\eta}_{k+1}\| - \|\eta_{k+1}\|\big|,$$

where the right hand side does not exceed $\|\hat{\eta}_1 - \eta_1\| + \|\hat{\eta}_k - \eta_k\|$. At the same time, recall that $\eta_{k+1}(i)/\eta_1(i) \le C$. Combining these gives

$$|III| \le C[\|\hat{\eta}_1 - \hat{\eta}_{k+1}\|] \le C[\|\hat{\eta}_1 - \eta_1\| + \|\hat{\eta}_{k+1} - \eta_{k+1}\|]. \tag{4.31}$$

Inserting (4.29)-(4.31) into (4.28), $|\hat{R}(i,k) - R(i,k)|$ does not exceed

$$C\Big(\frac{1}{\theta(i)}\big[|\hat{\eta}_1(i) - \eta_1(i)| + |\hat{\eta}_{k+1}(i) - \eta_{k+1}(i)|\big] + \|\hat{\eta}_1 - \eta_1\| + \|\hat{\eta}_{k+1} - \eta_{k+1}\|\Big).$$

Therefore, over the event E_n,

$$U_2 \le C\sum_{k=1}^{K}\big(\|\Theta^{-1}(\hat{\eta}_k - \eta_k)\|^2 + n\|\hat{\eta}_k - \eta_k\|^2\big).$$

Combining this with (4.24) gives

$$U_2 \le C\log(n)[err_n + n\|\theta\|_3^3\|\theta\|_1/\|\theta\|^6]. \tag{4.32}$$

Note that $n\|\theta\|_1/\|\theta\|^2 \le \sum_{i=1}^{n}(1/\theta(i))$. Therefore, $n\|\theta\|_3^3\|\theta\|_1/\|\theta\|^6 \le err_n$ by definitions. Combining this with (4.26) and (4.32) gives the claim. □

4.3. *Proof of Theorem 2.2.* Without loss of generality, assume $K > 2$. The proof for the case $K = 2$ is the same, except for that $\hat{R}^*$, R, and M^* are vectors rather than matrices, so that we have to change the terminology slightly. The following lemma is proved in Section 7.

LEMMA 4.3. *The $n \times (K-1)$ matrix R has exactly K distinct rows, and the ℓ^2-distance between any two distinct rows is no smaller than $\sqrt{2}$.*

We now show Theorem 2.2. For $1 \leq i \leq n$, let $\hat{r}_i$, r_i, and c_i denote the i-th row of $\hat{R}^*$, R, and M^* correspondingly. Fixing $\delta = \sqrt{2}/3$, we introduce a subset of V by $W = \{1 \leq i \leq n : \|\hat{r}_i - m_i\| \leq \delta, \|r_i - m_i\| \leq \delta\}$. Recalling that V partitions to K communities $V^{(1)}, V^{(2)}, \ldots, V^{(K)}$, we note that W has a similar partition $W = W^{(1)} \cup W^{(2)} \cup \ldots W^{(K)}$, where $W^{(k)} = V^{(k)} \cap W$, $1 \leq k \leq K$.

By Theorem 2.1, there is an event B such that $P(B^c) = o(n^{-2})$, and over the event B,

$$\|\hat{R}^* - R\|_F^2 \leq C \log^3(n) err_n. \tag{4.33}$$

Note that the $n \times (K-1)$ matrix R has exactly K unique rows so $R \in \mathcal{M}_{n,K-1,K}$. By how the k-means procedure is constructed, $\|\hat{R}^* - M^*\|_F \leq \|\hat{R}^* - R\|_F$, and so $\|R - M^*\|_F \leq \|\hat{R}^* - R\|_F + \|\hat{R}^* - M^*\|_F \leq 2\|\hat{R}^* - R\|_F$. Combining this with (4.33),

$$\|M^* - R\|_F^2 \leq 4\|\hat{R}^* - R\|_F^2 \leq C \log^3(n) err_n. \tag{4.34}$$

Combining this with (4.33), it follows from the definition of W that $|V \backslash W| \leq C \log^3(n) err_n$. Comparing this with the desired claim, it is sufficient to show that all nodes in W are correctly labeled; equivalently, this is to show that for any $i, j \in W$ such that $i \in W^{(k)}$ and $j \in W^{(\ell)}$,

$$m_i = m_j \text{ if and only if } k = \ell. \tag{4.35}$$

We now show (4.35). First, by definitions and (4.33)-(4.34), the cardinality of $(V^{(k)} \setminus W^{(k)})$ does not exceed $\delta^{-2} \sum_{i \in V^{(k)}} (\|\hat{r}_i - r_i\|^2 + \|m_i - r_i\|^2) \leq C \log^3(n) err_n$. Recall that we assume $\log^3(n) err_n \ll \min\{n_1, n_2, \ldots, n_k\}$. Combining these, $W^{(k)}$ is non-empty. Second, by Lemma 4.3 and definitions, for any $i, j \in W$ such that $i \in W^{(k)}$, $j \in W^{(\ell)}$, where $1 \leq k, \ell \leq K$ and $k \neq \ell$,

$$\|m_i - m_j\| \geq \|r_i - r_j\| - (\|m_i - r_i\| + \|m_j - r_j\|) \geq \delta. \tag{4.36}$$

Therefore, if $m_i = m_j$ for some $i, j \in W$, then there is a $1 \leq k \leq K$ such that $i, j \in W^{(k)}$. Suppose we pick one node j_k from each $W^{(k)}$, $1 \leq k \leq K$. By

(4.36), the K row vectors $\{m_{j_1}, ..., m_{j_K}\}$ are distinct. Note that the matrix M^* has at most K distinct rows, so if $i, j \in W^{(k)}$ for some $1 \leq k \leq K$, then $m_i = m_j$. Combining these gives (4.35). □

5. Simulations. We have conducted a small-scale simulation study. The goal is to select a few representative cases to investigate the performance of SCORE, two PCA approaches, and the modularity methods.

We compare 6 different algorithms, including three variants of SCORE, two PCA approaches, and one modularity method. The three variants of the SCORE are the original SCORE (SCORE), SCOREq with $q = 1$ (SCORE1), and SCOREq with $q = 2$ (SCORE2). The two PCA approaches are the ordinary PCA (oPCA) and the normalized PCA (nPCA). For modularity method, we use the tabu algorithm (tabu). Note that SCORE, oPCA, nPCA and tabu are studied in Section 1.

For each simulation experiment, we fix integers (n, K) and let $m = n/K$; for simplicity, we pick (n, k) such that m is also an integer. Fix a $K \times K$ matrix A and an $n \times 1$ positive vector $\tilde{\theta}$. For $k = 1, 2, \ldots, K$, we let $V^{(k)} = \{1 + m(k-1), 2 + m(k-1), \ldots, mk\}$, and let $\mathbf{1}_k$ be the indicator vector of $V^{(k)}$ as before. Each simulation experiment contains the following steps.

- (a). Randomly permute the coordinates of $\tilde{\theta}$ and denote the resultant vector by θ. Let Θ be the $n \times n$ diagonal matrix such that $\Theta(i,i) = \theta(i)$, $1 \leq i \leq n$. Define the $n \times n$ matrix Ω by $\Omega = \Theta[\sum_{k,\ell=1}^{K} A(k,\ell)\mathbf{1}_k\mathbf{1}_\ell']\Theta$.
- (b). Generate a symmetric $n \times n$ matrix W where all diagonals are 0, and for all $1 \leq i < j \leq n$, $W(i,j)$ are independent centered-Bernoulli with parameters $\Omega(i,j)$. Let $\tilde{X} = \Omega - \mathrm{diag}(\Omega) + W$, which can be viewed as the adjacency matrix of a network, say, $\mathcal{N} = (V, E)$.
- (c). Remove all nodes that are not connected to any other nodes in $\mathcal{N}$, and denote the resultant network by $\mathcal{N}_0 = (V_0, E_0)$. Let X 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 6 aforementioned algorithms to X. Record the Hamming errors of all methods under investigations.
- (e). Repeat (b)-(d) for *rep* times, where *rep* is a preselected integer.

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 (2.4) 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. The simulation includes 4 different experiments, which we now describe separately.

Experiment 1. In this experiment, we investigate how SCORE, oPCA, nPCA, and tabu perform with the classical stochastic Block Model (BM). We choose $(n, K, rep) = (1000, 2, 50)$, A as the 2×2 matrix with 1 on

Methods	oPCA	nPCA	tabu	SCORE
Mean (SD)	.058 (.009)	.055 (.010)	.050 (.065)	.058 (.009)

TABLE 3
Comparison of mean error rates (Experiment 1). In each cell, the number in the bracket is the corresponding standard deviation (SD); same below.

Methods	SCORE	SCORE1	SCORE2
Mean (SD)	.107 (.01)	.107 (.01)	.107 (.01)
	.054 (.008)	.054 (.008)	.054 (.008)
	.112 (.009)	.111 .008	.111 (.008)
	.122 (.119)	.071 (.006)	.071 (.006)

TABLE 4
Comparison of mean error rates (Experiment 2; from top to bottom: 2a-2d).

the diagonals and 0.5 on the off-diagonals, and $\tilde{\theta}$ as the vector where all coordinates are 0.2. This is a relatively easy case and all methods perform satisfactory and have similar error rates. See Table 3 for the results.

It is noteworthy that in one of the repetitions, tabu 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 much larger.

Experiment 2. In this experiment, we compare SCORE, SCORE1, and SCORE2. The experiment contains four sub-experiments, Experiment 2a-2d, with different network sizes, degrees of heterogeneity, and numbers of communities, etc..

In Experiment 2a, we investigate with a BM model, where all coordinates of $\tilde{\theta}$ are 0.1. Additionally, we take $(n, K, rep) = (2000, 2, 100)$, and A as the 2×2 matrix where two diagonals are 1 and two off-diagonals are 0.4. In Experiment 2b, we take $(n, K, rep) = (800, 2, 100)$, and A as the 2×2 matrix where the diagonals are 1 and two off-diagonals are .5. Also, we take $\tilde{\theta}$ to be the vector where $\tilde{\theta}(i) = .025 + .475 \times (i/n)$. In Experiment 2c, we take $(n, K, rep) = (1200, 2, 100)$ and let A be the same as in Experiment 2b, but we take $\tilde{\theta}$ to be the vector where $\tilde{\theta}(i) = .025 + .475 \times (i/n)^2$. In Experiment 2d, we take $(n, K, rep) = (1500, 3, 100)$, and A as the 3×3 matrix where the diagonals are 1, $A(1, 2) = .4$, $A(2, 3) = .4$, and $A(1, 3) = .05$. Also, we take $\tilde{\theta}(i) = .015 + .785 \times (i/n)^2$. The results are reported in Table 4, where it suggests that the behavior of three versions of SCORE are similar in various settings, except that in the last case, SCORE slightly underperforms than SCORE1 and SCORE2.

Experiment 3. In this example, we compare the performance of oPCA,

Methods	oPCA	nPCA	tabu	SCORE2
Mean (SD)	.378 (.041)	.165 (.084)	.0636 (.123)	.0695 (.004)

TABLE 5
Comparison of mean error rates (Experiment 3).

nPCA, tabu, SCORE2 in the case where we have three communities. We take $(n, K) = (1500, 3)$ and so $m = n/K = 500$. We take $rep = 25$ instead of $rep = 50$ as the tabu algorithm is sort of time consuming. We take A as the 3×3 symmetric matrix where we have 1 on the diagonals, $A(1,2) = 0.4$, $A(2,3) = 0.4$, and $A(1,3) = 0.05$. We take $\tilde{\theta}$ as the vector such that $\tilde{\theta}(i) = .015 + .785 \times (i/n)^2$, $1 \leq i \leq n$. The results are reported in Table 5, which suggest that SCORE2 outperforms nPCA and oPCA in terms of error rates. The error rates of SCORE2 and the tabu are similar, but SCORE2 is comparably much more stable than the tabu.

Experiment 4. In this experiment, we investigate how the heterogeneity parameters affect the performance of oPCA, nPCA, tabu and SCORE. We take $(n, K, rep) = (1000, 2, 50)$, and A to be the 2×2 matrix that has 1 on the diagonals and 0.5 on the off-diagonals. Fix $c_0 = 0.5$ and $d_0 = 0.02$. The experiment contains three sub-experiments, Experiment 4a-4c. In Experiment 4a, we take $\tilde{\theta}$ to be the vector such that $\tilde{\theta}(i) = d_0 + (c_0 - d_0)(i/n)$, $1 \leq i \leq n$. In Experiment 4b, we take $\tilde{\theta}$ to be the vector such that $\tilde{\theta}(i) = d_0 + (c_0 - d_0)(i/n)^2$, $1 \leq i \leq n$. In Experiment 4c, we take $\tilde{\theta}$ to be the vector such that $\tilde{\theta}(i) = c_0$ for $1 \leq i \leq n/2$ and $\tilde{\theta}(i) = 0.02$ for $n/2 < i \leq n$. Note that the heterogeneity effects are mild in Experiment 4a, but are much more severe than that in Experiment 4b-4c.

The results are tabulated in Table 6. The error rates of oPCA and nPCA are usually higher than that of the tabu and the SCORE. The average error rates of the tabu and the SCORE are similar, but the tabu usually has a much larger standard deviation (some times ten times as large). The instability of the tabu algorithm is due to that it depends on an initial guess (generated randomly), and when the initial guess is "bad", the tabu may fail to converge to the true labels.

In conclusion, the SCORE methods (include the original SCORE and two variants) have error rates much smaller than those of the two PCA approaches. The error rates of the SCORE and the tabu are usually comparable on average, but the stand deviation of the latter is usually a few times larger, showing that the tabu is comparably less stable. Additionally, the computation time of the tabu is much longer than that of the SCORE in matlab code. Therefore, it is fair to say that the SCORE outperforms both

Methods	oPCA	nPCA	tabu	SCORE
Mean (SD)	.066 (.021)	.066 (.107)	.042 (.064)	.043 (.006)
	.292 (.014)	.431 (.122)	.138 (.080)	.140 (.010)
	.254 (.034)	.476 (.049)	.139 (.074)	.130 (.010)

TABLE 6
Comparison of error rates (Experiment 4; from top to bottom: 4a-4c).

the two PCA approaches and the tabu algorithm.

6. 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.

Compared to many existing methods for DCBM (e.g. [18, 25, 33]), a very different 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 [33]. Zhao *et al* [33, Page 6] models the heterogeneity parameters as random variables that assume only finite values and have the same means, which is relatively restrictive. We model the heterogeneity parameter as non-stochastic 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.* [33] impose certain conditions on the $K \times K$ core matrix A which we don't 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 [17], we have extended the method to bipartite networks and have seen nice results on the 110-th 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 [12]).

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 [21, 20]), 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 non-trivial and we leave it to the future work.

Intellectually, this work is connected to the recent interest on low-rank matrix recovery and matrix completion; see for example [6]. 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 [16], where we show that spectral clustering can be very effective in cancer clustering with microarray 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 [4].

7. Proofs. In this section, we prove all the lemmas in the preceding sections.

7.1. *Proof of Lemmas 2.1.* Fix $1 \le k \le K$. Let λ_k be the nonzero eigenvalue of Ω with the k-th largest magnitude, let η_k be one of the (unit-norm) eigenvector associated with λ_k, and let a_k be the $K \times 1$ vector such that $a_k(i) = (\theta^{(i)}/\|\theta^{(i)}\|, \eta_k)$. In our model, we can rewrite Ω as

$$\Omega = \|\theta\|^2 \sum_{i,j=1}^{K} (DAD)(i,j)\Big(\frac{\theta^{(i)}}{\|\theta^{(i)}\|}\Big)\Big(\frac{\theta^{(j)}}{|\theta^{(j)}\|}\Big)', \tag{7.37}$$

and so by basic algebra and notations,

$$\Omega\eta_k = \|\theta\|^2 \sum_{i,j=1}^{K} (DAD)(i,j)(\frac{\theta^{(j)}}{\|\theta^{(j)}\|}, \eta_k)\frac{\theta^{(i)}}{\|\theta^{(i)}\|} = \|\theta\|^2 \sum_{i,j=1}^{K} (DAD)(i,j)a_k(j)\frac{\theta^{(i)}}{\|\theta^{(i)}\|}. \tag{7.38}$$

At the same time, since $\Omega\eta_k = \lambda_k\eta_k$,

$$a_k(i) = (\theta^{(i)}/\|\theta^{(i)}\|, \eta_k) = \frac{1}{\lambda_k}(\theta^{(i)}/\|\theta^{(i)}\|, \Omega\eta_k). \tag{7.39}$$

Note that $\{\theta^{(i)}/\|\theta^{(i)}\|\}_{i=1}^K$ is an orthonormal base. Inserting (7.38) to the right hand side of (7.39) gives $a_k(i) = (\|\theta\|^2/\lambda_k)\sum_{j=1}^K (DAD)(i,j)a_k(j)$, or in matrix form,

$$DADa_k = (\lambda_k/\|\theta\|^2)a_k. \tag{7.40}$$

This says that $\lambda_k/\|\theta\|^2$ is an eigenvalue of DAD and a_k is one of the associated eigenvector. Moreover, inserting (7.40) into to the right hand side of (7.38) and recalling $\Omega\eta_k = \lambda_k\eta_k$,

$$\eta_k = \frac{1}{\lambda_k}\Omega\eta_k = \sum_{i=1}^K a_k(i)\theta^{(i)}/\|\theta^{(i)}\|,$$

and so $\|a_k\|^2 = \|\eta_k\|^2 = 1$. By our assumptions, all eigenvalues of DAD are simple. It follows that a_k is unique determined up to a factor of ± 1. □

7.2. *Proof of Lemma 2.2.* By the definition of DCBM and (2.8), we have that $\|\mathrm{diag}(\Omega)\| \le \theta_{max}^2$, where $\theta_{max} \le g_0 < 1$. Note that by (2.11)-(2.12), $\theta_{max}\|\theta\|_1 \to \infty$. It follows that

$$\|\mathrm{diag}(\Omega)\| = o(\sqrt{\log(n)\theta_{max}\|\theta\|_1}).$$

Therefore, to show the claim, it is sufficient to show that with probability at least $1 + o(n^{-3})$,

$$\|W\| \le 3\sqrt{\log(n)\theta_{max}\|\theta\|_1}.$$

Let e_i be the $n \times 1$ vector such that $e_i(j) = 1$ if $i = j$ and 0 otherwise. Write $W = \sum_{1\le i<j\le n} Z^{(i,j)}$, where $Z^{(i,j)} = W(i,j)[e_ie_j' + e_je_i']$. Let $\sigma^2 = \|\sum_{1\le i<j\le n} E[(Z^{(i,j)})^2]\|$. By elementary statistics and (2.8), $E[W^2(i,j)] \le \theta(i)\theta(j)$. At the same time,

$$E[(Z^{(i,j)})^2] = E[W(i,j)^2]\cdot[e_ie_j' + e_je_i']^2 = E[W(i,j)^2]\cdot[e_ie_i' + e_je_j'].$$

Combining these gives that $\sigma^2 \le \theta_{max}\|\theta\|_1$. Fix $q > 0$. Applying Theorem 4.1 with $Z^{(i,j)} = W(i,j)[e_ie_j' + e_je_i']$, $h_0 = 1$, $\sigma^2 = \|\sum_{i<j} E[(Z^{(i,j)})^2]\|$, and $t = \sqrt{2q\log(n)\theta_{max}\|\theta\|_1}$,

$$P(\|W\| \ge \sqrt{2q\log(n)\theta_{max}\|\theta\|_1}) \le 2n\exp\left[-\frac{q\log(n)}{1 + (1/3)\sqrt{2q\log(n)/(\theta_{max}\|\theta\|_1)}}\right].$$

Note that $\theta_{max}\|\theta\|_1 \ge \|\theta\|^2$, and that $\|\theta\|^2/\log(n) \to \infty$ as in the assumption (2.11). It follows that $q\log(n)/(\theta_{max}\|\theta\|_1) \to 0$, and the claim follows by taking $q = 9/2$.

7.3. *Lemma 2.4.* Let $\lambda_1 > \lambda_2 > \ldots > \lambda_K$ be the K nonzero eigenvalues of Ω. By Lemma 2.3 and (2.14)-(2.15), for all $1 \le k \le K-1$, $|\lambda_{k+1} - \lambda_k| \ge C\|\theta\|^2$, At the same time, by Lemma 2.2 and (2.11), it follows from basic algebra (e.g., [2, Page 473]) that with probability at least $1 + o(n^{-3})$,

$$|\hat{\lambda}_k - \lambda_k|/\|\theta\|^2 = o(1), \tag{7.41}$$

and so all the K eigenvalues are simple.

Now, fixing $1 \le k \le K$, let $\hat{\eta}_k$ be an eigenvector (the norms of which are not necessarily 1) associated with $\hat{\lambda}_k$. Writing for short $\hat{\theta}^{(i,k)} = [I_n - (W - \mathrm{diag}(\Omega))/\hat{\lambda}_k]^{-1}\theta^{(i)}$, we let $\hat{b}_k$ be the $K \times 1$ vector such that

$$\hat{b}_k(i) = (\theta^{(i)}/\|\theta^{(i)}\|, \hat{\eta}_k), \qquad 1 \le i \le K,$$

and let

$$\hat{a}_k = (B^{(k)})^{-1}\hat{b}_k.$$

Since $X\hat{\eta}_k = \hat{\lambda}_k\eta_k$ and $X = \Omega + (W - \mathrm{diag}(\Omega))$, it follows that

$$\hat{\eta}_k = [\hat{\lambda}_k I_n - (W - \mathrm{diag}(\Omega))]^{-1}\Omega\hat{\eta}_k.$$

Recall that (e.g., (7.37)) $\Omega = \|\theta\|^2 \sum_{i,j=1}^K (DAD)(i,j)(\theta^{(i)}/\|\theta^{(i)}\|)(\theta^{(j)}/\|\theta^{(j)}\|)'$. Combining these and rearranging,

$$\hat{\eta}_k = (\frac{\|\theta\|^2}{\hat{\lambda}_k})\sum_{i=1}^K\Big[\Big(\sum_{j=1}^K (DAD)(i,j)\hat{b}_k(j)\Big)\cdot\frac{\hat{\theta}^{(i,k)}}{\|\theta^{(i)}\|}\Big]. \tag{7.42}$$

Recall that $B^{(k)}(\ell,i) = (\theta^{(\ell)})'[I_n - (W - \mathrm{diag}(\Omega))/\hat{\lambda}_k]^{-1}\theta^{(i)}/[\|\theta^{(\ell)}\|\cdot\|\theta^{(i)}\|] \equiv (\theta^{(\ell)})'\hat{\theta}^{(i,k)}/[\|\theta^{(\ell)}\|\cdot\|\theta^{(i)}\|]$. Taking the inner product of two sides in (7.42) with $\theta^{(\ell)}/\|\theta^{(\ell)}\|$, it follows from the definitions of $\hat{b}_k$ that for any $1 \le \ell \le K$,

$$\hat{b}_k(\ell) = (\frac{\|\theta\|^2}{\hat{\lambda}_k})\sum_{i,j=1}^K B^{(k)}(\ell,i)(DAD)(i,j)\hat{b}_k(j) = (\frac{\|\theta\|^2}{\hat{\lambda}_k})\sum_{j=1}^n (B^{(k)}DAD)(\ell,j)\hat{b}_k(j),$$

or in matrix form,

$$\hat{b}_k = (\frac{\|\theta\|^2}{\hat{\lambda}_k})B^{(k)}DAD\hat{b}_k.$$

This means that $\hat{\lambda}_k/\|\theta\|^2$ is an eigenvalue of $B^{(k)}DAD$ and $\hat{b}_k$ is one of the associated eigenvector. Recall that $\hat{a}_k = (B^{(k)})^{-1}\hat{b}_k$. By basic algebra, $\hat{\lambda}_k/\|\theta\|^2$ is an eigenvalue of $DADB^{(k)}$, and $\hat{a}_k$ is one of the associated eigenvectors. Especially,

$$DAD\hat{b}_k = DADB^{(k)}\hat{a}_k = [\hat{\lambda}_k/\|\theta\|^2]\hat{a}_k. \tag{7.43}$$

Inserting (7.43) into (7.42) and rearranging,

$$\hat{\eta}_k = \sum_{i=1}^{K} \hat{a}_k(i) \cdot \hat{\theta}^{(i,k)}/\|\theta^{(i)}\|.$$

We now check the uniqueness of $\hat{\lambda}_k$. By Lemma 7.1 to be introduced below and (2.12), $\|B^{(k)} - I_K\|_F \leq C\log(n)\|\theta\|_1\|\theta\|_3^3/\|\theta\|^6 = o(1)$. By similar argument as in (7.41), $\mathrm{eigsp}(DAD) \geq C$. Combining these with basic algebra (e.g., [2, Page 473]), all eigenvalues of $DADB^{(k)}$ are simple, and $\hat{b}_k$ (and so $\hat{\lambda}_k$, $\hat{a}_k$, and $\hat{\eta}_k$) are uniquely determined up to some scaling factors. If we further require $\|\hat{a}_k\| = 1$, then $\hat{a}_k$, $\hat{b}_k$, and $\hat{\eta}_k$ are all uniquely determined, up to a common factor that takes values from $\{-1, 1\}$. This gives the claim. □

7.4. *Proof of Lemma 2.6.* Recall that D is a diagonal matrix where the k-th diagonal is the k-th coordinate of the $K \times 1$ vector $d^{(n)}$, equalling $\|\theta^{(k)}\|/\|\theta\|$, $1 \leq k \leq K$; the superscript "(n)" emphasizes the dependence on n (same below). By Lemma 2.1, $\Theta^{-1}\eta_1 = \sum_{k=1}^{K}[a_1^{(n)}(k)/\|\theta^{(k)}\|]\mathbf{1}_k$, where $a_1^{(n)}$ is the eigenvector associated with the largest eigenvalue of DAD. By (2.14), to show the lemma, it suffices to show that for sufficiently large n,

$$OSC(a_1^{(n)}) \leq C. \tag{7.44}$$

Note that in the special case where $d^{(n)}$ does not depend on n, the claim follows directly by Perron's theorem [15, Page 508], since DAD is non-negative and irreducible. Consider the general case where $d^{(n)}$ may depend on n. If (7.44) does not hold, then we can find a subsequence of $n \in \{1, 2, \ldots, \}$ such that along this sequence, there are two $K \times 1$ vectors d_0 and a such that (a) $OSC(a_1^{(n)}) \to \infty$, (b) $d^{(n)} \to d_0$, and (c) $a_1^{(n)} \to a$. By the condition (2.15), $OSC(d_0) \leq C$, and a direct use of of Perron's theorem [15, Page 508] implies that $OSC(a) \leq C$. This contradicts with (a). The contradiction proves (7.44) and the claim follows. □

7.5. *Proof of Lemmas 2.7-2.8.* Write $\hat{\theta}^{(i,k)} = [I_n - (W - \mathrm{diag}(\Omega))/\hat{\lambda}_k]^{-1}\theta^{(i)}$ for short. In our notations,

$$\eta_k = \sum_{i=1}^{K}[a_k(i)/\|\theta^{(i)}\|]\theta^{(i)}, \qquad \hat{\eta}_k = \sum_{i=1}^{K}[\hat{a}_k(i)/\|\theta^{(i)}\|]\hat{\theta}^{(i,k)}, \tag{7.45}$$

where a_k are the eigenvectors of DAD and $\hat{a}_k$ are the eigenvectors of $DAD(B^{(k)})$. To show the claim, we first characterize $\|\hat{a}_k - a_k\|$, and then characterize $\|\hat{\theta}^{(i,k)} - \theta^{(i)}\|$.

Consider $\|\hat{a}_k - a_k\|$ first. Let I_K be the $K \times K$ identity matrix. The following lemma is proved below (implicitly, we assume that in Lemma 7.1, the conditions of Lemmas 2.7-2.8 hold; same for Lemmas 7.3-7.4).

LEMMA 7.1. *With probability at least* $1 + o(n^{-3})$,

$$\|B^{(k)} - I_K\|_F \leq C \log(n)(\|\theta\|_1 \cdot \|\theta\|_3^3)/\|\theta\|^6.$$

Note that by (2.11), the right hand side tends to 0 as $n \to \infty$.

We also need a lemma on *eigenvector sensitivity*. Suppose U and Err are both symmetric $K \times K$ matrix where $\|Err\| < (1/2)\text{eigsp}(U)$, so that all the eigenvalues of U and $U + Err$ are simple. Let $\lambda_1^{(1)} > \lambda_2^{(1)} > \ldots > \lambda_K^{(1)}$ and $\lambda_1^{(2)} > \lambda_2^{(2)} > \ldots > \lambda_K^{(2)}$ be the eigenvalues of U and $U + Err$, respectively, and let $\xi_2^{(1)}, \xi_2^{(1)}, \ldots, \xi_K^{(1)}$ and $\xi_1^{(1)}, \xi_2^{(2)}, \ldots, \xi_K^{(2)}$ be the corresponding (unit-norm) eigenvectors, of U and $U + Err$, respectively. The following lemma is proved below.

LEMMA 7.2. *If* $\|Err\| < \text{eigsp}(U)/2$, *then for any* $1 \leq k \leq K$, $\|\xi_k^{(1)} - \xi_k^{(2)}\| \leq 2\sqrt{2}\frac{\|Err\|}{\text{eigsp}(U)}$.

Note that $\|DAD\| \leq C$. Using Lemma 7.1 and basic algebra, with probability at least $1 + o(n^{-3})$,

$$\|DAD(B^{(k)}) - DAD\| \leq C\|(B^{(k)}) - I_K\| \leq C \log(n)(\|\theta\|_1 \cdot \|\theta\|_3^3)/\|\theta\|^6.$$

Applying Lemma 7.2 with $U = DAD$ and $Err = DAD[(B^{(k)}) - I_K]$, it follows from the eigen-space condition (2.14) that with probability at least $1 + o(n^{-3})$, for $1 \leq k \leq K$,

$$\|\hat{a}_k - a_k\| \leq C\|Err\| \leq C \log(n)(\|\theta\|_1 \cdot \|\theta\|_3^3)/\|\theta\|^6.$$

By (2.11), the right hand side tends to 0, so

$$\|\hat{a}_k - a_k\|^2 \leq \|\hat{a}_k - a_k\| \leq C \log(n)(\|\theta\|_1 \cdot \|\theta\|_3^3)/\|\theta\|^6. \tag{7.46}$$

Next, we consider $\|\hat{\theta}^{(i,k)} - \theta^{(k)}\|$. The following lemmas are proved below.

LEMMA 7.3. *With probability at least* $1 + o(n^{-3})$, *for all* $1 \leq k, i \leq K$,

$$\|\hat{\theta}^{(i,k)} - \theta^{(i)}\|^2 \leq C \log(n)\|\theta\|_1\|\theta\|_3^3/\|\theta\|^4.$$

LEMMA 7.4. *With probability at least* $1+o(n^{-3})$, *for any* $1 \le k, i \le K$,

$$\|\Theta^{-1}(\hat{\theta}^{(i,k)} - \theta^{(i)})\|^2 \le C\log(n)\frac{\|\theta\|_3^3}{\|\theta\|^4}\Big[\sum_{i=1}^n \frac{1}{\theta(i)} + \frac{1}{\theta_{min}}\frac{\log(n)\|\theta\|_1^2}{\|\theta\|^4}\Big].$$

We now show Lemmas 2.7-2.8. Consider Lemma 2.7 first. By (7.45) and basic algebra,

$$\|\hat{\eta}_k - \eta_k\|^2 \le C(I+II),$$

where $I = \sum_{i=1}^K \hat{a}_k^2(i)(\|\hat{\theta}^{(i,k)} - \theta^{(i)}\|^2/\|\theta^{(i)}\|^2)$, and $II = \sum_{i=1}^K (\hat{a}_k(i) - a_k(i))^2 = \|\hat{a}_k - a_k\|^2$. Since $\hat{a}_k$ has unit norm and $\|\theta^{(i)}\|^2 \ge C\|\theta\|^2$, combining these with (7.46) and Lemma gives

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

and the claim follows.

Next, consider Lemma 2.8. Similarly, $\|\Theta^{-1}[\hat{\eta}_k - \eta_k]\|^2 \le C(I+II)$, where $I = \sum_{i=1}^K \hat{a}_k^2(i)(\|\Theta^{-1}[\hat{\theta}^{(i,k)} - \theta^{(i)}]\|^2/\|\theta^{(i)}\|^2)$, and $II = \sum_{i=1}^K |\hat{a}_k(i) - a_k(i)|^2\|\mathbf{1}_i\|^2/\|\theta^{(i)}\|^2$. By Lemma 7.4 and similar argument,

$$\|\Theta^{-1}[\hat{\eta}_k - \eta_k]\|^2 \le C\log(n)\frac{\|\theta\|_3^3}{\|\theta\|^6}\Big[\sum_{i=1}^n \frac{1}{\theta(i)} + \frac{\log(n)}{\theta_{min}}\frac{\|\theta\|_1^2}{\|\theta\|^4} + \frac{n\|\theta\|_1}{\|\theta\|^2}\Big].$$

Note that $n\|\theta\|_1 \le \|\theta\|^2 \sum_{i=1}^n \frac{1}{\theta(i)}$, it follows that

$$\|\Theta^{-1}[\hat{\eta}_k - \eta_k]\|^2 \le C\log(n)\frac{\|\theta\|_3^3}{\|\theta\|^6}\Big[\sum_{i=1}^n \frac{1}{\theta(i)} + \frac{\log(n)}{\theta_{min}}\frac{\|\theta\|_1^2}{\|\theta\|^4}\Big],$$

and the claim follows by the definition of err_n; see (2.19). □

7.6. *Proof of Lemma 4.1.* Similar to that in the proof of Lemma 2.2, let e_i be the $n \times 1$ vector such that $e_i(j) = 1$ if $i = j$ and 0 otherwise. Write $\Theta^{-1}W = \sum_{i<j} Z^{(i,j)}$, where $Z^{(i,j)} = W(i,j)\Theta^{-1}[e_ie_j' + e_je_i']$. Let

$$\sigma^2 = \max\{\|\sum_{1\le i<j\le n} E[Z^{(i,j)}(Z^{(i,j)})']\|,\ \|\sum_{1\le i<j\le n} E[(Z^{(i,j)})'Z^{(i,j)}]\|\}.$$

First, by (2.8) and basic statistics, $E[W^2(i,j)] \le \theta(i)\theta(j)$. It is seen

$$E[Z^{(i,j)}(Z^{(i,j)})'] = E[W^2(i,j)]\Theta^{-1}[e_ie_j' + e_je_i']^2\Theta^{-1} = E[W^2(i,j)]\Theta^{-1}[e_ie_i' + e_je_j']\Theta^{-1},$$

which is a diagonal matrix, where the i-th diagonal $\le \theta(j)/\theta(i)$, the j-th diagonal $\le \theta(i)/\theta(j)$, and all other diagonals are 0. Therefore, $\sum_{1\le i<j\le n} E[Z^{(i,j)}(Z^{(i,j)})']$

is a diagonal matrix where the i-th coordinate does not exceed $\|\theta\|_1/\theta(i)$, and the matrix norm of which $\le \theta_{min}^{-1}\|\theta\|_1$.

Second, we similarly have

$$E[(Z^{(i,j)})'Z^{(i,j)}] = E[W^2(i,j)][e_ie_j' + e_je_i']\Theta^{-2}[e_ie_j' + e_je_i'],$$

which is a diagonal matrix where the i-th coordinate does not exceed $\theta(i)/\theta(j)$, the j-th coordinate $\le \theta(j)/\theta(i)$. As a result, $\sum_{1\le i<j\le n} E[(Z^{(i,j)})'Z^{(i,j)}]$ is a diagonal matrix where the i-th coordinate does not exceed $\theta(i)\sum_{j=1}^n(1/\theta(j))$, and the matrix norm $\le \theta_{max}\sum_{i=1}^n(1/\theta(i))$. Combining these gives

$$\sigma^2 \le \max\Big\{\frac{1}{\theta_{min}}\|\theta\|_1,\ \theta_{max}\sum_{i=1}^n\frac{1}{\theta(i)}\Big\} \equiv \sigma_0^2.$$

Fix $q > 0$. Applying Theorem 4.1 with $h_0 = 1/\theta_{min}$ and $t = \sigma_0\sqrt{2q\log(n)}$ gives

$$P(\|\Theta^{-1}W\| \ge \sigma_0\sqrt{2q\log(n)}) \le 2n\exp\Big[-\frac{q\log(n)}{1+(1/3)\sqrt{2q\log(n)}\theta_{min}^{-1}/\sigma_0}\Big].$$

By (2.18), $\log(n)\theta_{min}^{-2} \le \sigma_0^2$, and the claim follows by picking q to be a sufficiently large constant. □

7.7. *Proof of Lemma 4.2.* Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with $|Y_k| \le b$, $E[Y_k] = 0$, and $\mathrm{var}(Y_k) \le \sigma_k^2$ for $1 \le k \le n$. Write for short $\sigma^2 = \sigma_1^2 + \ldots \sigma_n^2$. We claim that with probability at least $1 + o(1/n^3)$,

$$|\sum_{i=1}^n Y_i|^2 \le C\log(n)\max\{\sigma^2, \log(n)b^2\}. \tag{7.47}$$

In detail, using Bennett's Lemma [26, Page 851], for all $\lambda > 0$,

$$P(|\sum_{i=1}^n Y_i| \ge \lambda) \le \begin{cases} 2\exp(-\frac{c_0}{2\sigma^2}\lambda^2), & \lambda b \le \sigma^2, \\ 2\exp(-\frac{c_0}{2}\frac{\lambda}{b}), & \lambda b \ge \sigma^2. \end{cases}$$

where $c_0 = \psi(1)$, with ψ as in [26, Page 851]; note that $c_0 \approx 0.773$. Now, when $\sigma/b \ge 2\sqrt{2\log(n)}$, we take $\lambda = 2\sqrt{2\log(n)}\sigma$. It is seen $\lambda b \le \sigma^2$, and so

$$P(\Big|\sum_{i=1}^n Y_i\Big| \ge \lambda) \le 2\exp(-4c_0\log(n)) = o(n^{-3}).$$

When $\sigma/b < 2\sqrt{2\log(n)}$, we take $\lambda = 8b\log(n)$. It is seen $\lambda b \geq \sigma^2$. It follows that

$$P(\big|\sum_{i=1}^{n} Y_i\big| \geq \lambda) \leq 2\exp(-4c_0\log(n)) = o(n^{-3}).$$

Combining these, except for a probability of $o(n^{-3})$,

$$\big|\sum_{i=1}^{n} Y_i\big| \leq 2\sqrt{2\log(n)}\sigma 1\{\sigma/b \geq 2\sqrt{2\log(n)}\} + 8b\log(n)1\{\sigma/b < 2\sqrt{2\log(n)}\},$$

and (7.47) follows.

We now show Lemma 4.2. The last item follows directly from (7.47), and the proofs for first two items are similar, so we only show the second item. Let e_i be the $n \times 1$ vector such that $e_i(j) = 0$ if $i = j$ and 0 otherwise. Write $\|W\theta^{(k)}\|^2 = \sum_{i=1}^{n}(e_i'W\theta^{(k)})^2$. For each fixed i, applying (7.47) to $Y_j = W(i,j)\theta^{(k)}(j)$, $b = \theta_{max}$, and $\sigma^2 = E[(\sum_{j\in V^{(k)}}\theta(j)W(i,j))^2]$, we have that with probability at least $1 + o(1/n^3)$,

$$|e_i'W\theta^{(k)}|^2 \leq C\log(n)\max\{\sigma^2, \log(n)\theta_{max}^2\}.$$

Now, direct calculation shows that $\sigma^2 \leq \theta(i)\|\theta\|_3^3$. It follows that with probability at least $1 + o(n^{-2})$ that

$$\|W\theta^{(k)}\|^2 \leq C\log(n)\sum_{i=1}^{n}\max\{\theta(i)\|\theta\|_3^3, \log(n)\theta_{max}^2\},$$

and the claim follows by the first MDV assumption in (2.17). □

7.8. *Proof of Lemma 4.3.* We expand R to be an $n \times K$ matrix by adding a column of ones to the left. For notational simplicity, we still call the matrix by R. It is sufficient to show that R has exactly K distinct rows, and the ℓ^2-distance for each pair of such distinct rows is no smaller than $\sqrt{2}$.

With the new notations, since Θ is a diagonal matrix, for any $1 \leq i \leq n$ and $1 \leq k \leq K$,

$$R(i,k) = \frac{\eta_k(i)}{\eta_1(i)} = \frac{(\Theta^{-1}\eta_k)(i)}{(\Theta^{-1}\eta_1)(i)},$$

where $\eta_1, \eta_2, \ldots, \eta_K$ are the K leading eigenvectors of Ω. Combining this with Lemma 2.1 and recalling that $d_j = \|\theta^{(j)}\|/\|\theta\|$,

$$R(i,k) = \frac{\sum_{j=1}^{K} a_k(j)\mathbf{1}_j(i)/d_j}{\sum_{j=1}^{K} a_1(j)\mathbf{1}_j(i)/d_j},$$

which equals to $a_k(\ell)/a_1(\ell)$ if and only if node i belongs to the ℓ-th community $V^{(\ell)}$, $\ell = 1, 2, \ldots, K$. It is now evident that R has K distinct rows, each is one of the following row-vectors:

$$\frac{1}{a_1(\ell)}\big(a_1(\ell), a_2(\ell), \ldots, a_K(\ell)\big), \qquad \ell = 1, 2, \ldots, K.$$

Fix $k \neq \ell$. The square of the ℓ^2-distance between the vector $\frac{1}{a_1(k)}\big(a_1(k), \ldots, a_K(k)\big)$ and the vector $\frac{1}{a_1(\ell)}\big(a_1(\ell), \ldots, a_K(\ell)\big)$ is

$$\frac{1}{a_1^2(k)}\sum_{j=1}^K a_j^2(k) + \frac{1}{a_1^2(\ell)}\sum_{j=1}^K a_j^2(\ell) - \frac{2}{a_1(k)a_1(\ell)}\sum_{j=1}^K a_j(k)a_j(\ell).$$

Since $a_1, a_2, \ldots, a_K$ form an orthonormal base, $\sum_{j=1}^K a_j^2(k) = 1$, $\sum_{j=1}^K a_j^2(\ell) = 1$, and $\sum_{j=1}^K a_j(k)a_j(\ell) = 0$. Therefore, the square of the ℓ^2-distance between these two vectors is $a_1^{-2}(k) + a_1^{-2}(\ell)$ and the claim follows since $|a_1(k)| \leq 1$ and $|a_1(\ell)| \leq 1$. □

7.9. *Proof of Lemma 7.1.* Write for short $U = \mathrm{diag}(\Omega)$ and $H = (W - \mathrm{diag}(\Omega))/\hat{\lambda}_k$. For $1 \leq i, j \leq K$, $B^{(k)}(i,j) = (\theta^{(i)})'[I_n - H]^{-1}\theta^{(j)}/(\|\theta^{(i)}\| \cdot \|\theta^{(j)}\|)$. By (2.15), for all $1 \leq i \leq K$, $\|\theta^{(i)}\| \asymp \|\theta\|$. All we need to show is

$$\big|(\theta^{(i)})'[(I_n - H)^{-1} - I_n]\theta^{(j)}\big| \leq C\log(n)\|\theta\|_1\|\theta\|_3^3/\|\theta\|^4, \qquad 1 \leq i, j \leq K; \tag{7.48}$$

note that $(\theta^{(i)})'\theta^{(j)} = \|\theta^{(i)}\|^2$ if $i = j$ and 0 otherwise.

Write

$$(\theta^{(i)})'[(I_n - H)^{-1} - I_n]\theta^{(j)} = I + II, \tag{7.49}$$

where

$$I = (\theta^{(i)})'H\theta^{(j)}, \qquad II = (\theta^{(i)})'H[I_n - H]^{-1}H\theta^{(j)}.$$

Consider I first. By $H = (W - U)/\hat{\lambda}_k$, we have

$$I = \frac{1}{\hat{\lambda}_k}(Ia - Ib), \tag{7.50}$$

where $Ia = (\theta^{(i)})'W\theta^{(j)}$, $Ib = (\theta^{(i)})'U\theta^{(j)}$. First, by (2.8) and that all $\theta(i) \leq 1$, $|Ib| \leq \|\theta\|_4^4 \leq \|\theta\|_3^3$ with probability at least $1 + o(n^{-3})$. Second, by Lemma 4.2, $Ia \leq C\sqrt{\log(n)}\max\{\|\theta\|_3^3, \sqrt{\log(n)}\theta_{max}^2\}$. Last, by (2.16), with probability at least $1 + o(n^{-3})$, $|\hat{\lambda}_k| \asymp \|\theta\|^2$. Inserting these into (7.50) gives

$$|I| \leq C\sqrt{\log(n)}\max\{\|\theta\|_3^3, \sqrt{\log(n)}\theta_{max}^2\}/\|\theta\|^2. \tag{7.51}$$

Consider II next. First, by Schwartz inequality, $|II| \leq \|(I_n-H)^{-1/2}H\theta^{(i)}\| \cdot \|(I_n-H)^{-1/2}H\theta^{(j)}\|$. Second, by (2.11) and Lemma 2.3, with probability at least $1+o(n^{-3})$, $\|I_n-H\|^{-1/2} \lesssim 1$ and $\hat{\lambda}_k \asymp \|\theta\|^2$. Therefore, for any $1 \leq i \leq K$, $\|(I_n-H)^{-1/2}H\theta^{(i)}\|^2$ does not exceed

$$\frac{1}{\hat{\lambda}_k^2}\|(I_n-H)^{-1/2}(W-U)\theta^{(i)}\|^2 \leq C\|(W-U)\theta^{(i)}\|^2/\|\theta\|^4 \leq (IIa+IIb)/\|\theta\|^4,$$

where $IIa = \|W\theta^{(i)}\|^2$ and $IIb = \|U\theta^{(i)}\|^2$. Now, on one hand, by Lemma 4.2, with probability at least $1+o(n^{-3})$,

$$\|W\theta^{(i)}\|^2 \leq C\log(n)\|\theta\|_1\|\theta\|_3^3.$$

On the other hand, by basic algebra and that $\|\theta^{(i)}\|_\infty < 1$,

$$\|U\theta^{(i)}\|^2 \leq \|\theta\|_6^6 \leq \|\theta\|_3^3.$$

Note that by (2.11)-(2.12), $\log(n)\|\theta\|_1 \geq \|\theta\|^4 \gg 1$. Combining these gives that with probability at least $1+o(1/n^2)$,

$$|II| \leq C\log(n)\|\theta\|_1\|\theta\|_3^3/\|\theta\|^4. \tag{7.52}$$

Inserting (7.51) and (7.52) into (7.49) gives that with probability at least $1+o(1/n^2)$,

$$\big|(\theta^{(i)})'[(I_n-H)^{-1}-I_n]\theta^{(j)}\big| \leq \frac{C\log(n)}{\|\theta\|^4}[\max\{\|\theta\|_3^3, \sqrt{\log(n)}\theta_{max}^2\}\|\theta\|^2+\|\theta\|_1\|\theta\|_3^3]. \tag{7.53}$$

Write $\max\{\|\theta\|_3^3, \sqrt{\log(n)}\theta_{max}^2\}\|\theta\|^2 \leq \|\theta\|_3^3\|\theta\|^2 + \sqrt{\log(n)}\theta_{max}^2\|\theta\|^2$. First, since $\|\theta\|_\infty < 1$, $\|\theta\|^2\|\theta\|_3^3 \leq \|\theta\|_1\|\theta\|_3^3$. Second, by (2.11), $\sqrt{\log(n)}\theta_{max}^2\|\theta\|^2 \leq \|\theta\|^4 \leq \|\theta\|_3^3\|\theta\|_1$. Inserting these into (7.53) gives (7.48) and the claim follows. □

7.10. *Proof of Lemma 7.2.* By the assumptions and elementary algebra, it is seem that $(\lambda_k^{(1)}, \xi_k^{(1)})$ and $(\lambda_k^{(2)}, \xi_k^{(2)})$ take real values. By definitions, $(U+Err)\xi_k^{(2)} = \lambda_k^{(2)}\xi_k^{(2)}$, and so

$$(\xi_k^{(2)})'[\lambda_k^{(2)}I_K - U]^2\xi_k^{(2)} = (\xi_k^{(2)})'(Err)^2\xi_k^{(2)}.$$

Since $\{\xi_1^{(1)}, \xi_2^{(1)}, \ldots, \xi_K^{(1)}\}$ constitute an orthonormal base, we have

$$(\xi_k^{(2)})'[\lambda_k^{(2)}I_K-U]^2\xi_k^{(2)} = \sum_{i=1}^K(\lambda_k^{(2)}-\lambda_i^{(1)})^2(\xi_k^{(2)}, \xi_i^{(1)})^2 \geq \sum_{i\neq k}(\lambda_k^{(2)}-\lambda_i^{(1)})^2(\xi_k^{(2)}, \xi_i^{(1)})^2.$$

Combining these gives

$$\sum_{i\neq k}(\lambda_k^{(2)}-\lambda_i^{(1)})^2(\xi_k^{(2)},\xi_i^{(1)})^2 \leq (\xi_k^{(2)})'(Err)^2\xi_k^{(2)} \leq \|Err\|^2. \tag{7.54}$$

By the assumption of $\|Err\| \leq (1/2)\text{eigsp}(U)$, for all $1 \leq i \leq K$ and $i \neq k$, $|\lambda_k^{(2)} - \lambda_i^{(1)}| \geq (1/2)\text{eigsp}(U)$. Inserting this into (7.54) gives

$$\sum_{i\neq k}(\xi_k^{(2)},\xi_i^{(1)})^2 \leq 4\|Err\|^2/\text{eigsp}(U)^2.$$

Since $(\xi_k^{(2)},\xi_k^{(1)})^2 = 1-\sum_{i\neq k}(\xi_k^{(2)},\xi_i^{(1)})^2$, it follows that $(\xi_k^{(2)},\xi_k^{(1)})^2 \geq 1 - 4\|Err\|^2/\text{eigsp}(U)^2$, and the claim follows by basic algebra. □

7.11. *Proof of Lemma 7.3.* Write for short $U = \text{diag}(\Omega)$ and $H = (W - \text{diag}(\Omega))/\hat{\lambda}_k$. Similarly, write

$$\hat{\theta}^{(i,k)} - \theta^{(i)} = [(I_n - H)^{-1} - I_n]\theta^{(i)} = (I_n - H)^{-1}H\theta^{(i)}.$$

By Lemma 2.3, with probability at least $1 + o(n^{-3})$, $\|H\| = o(1)$,

$$\|\hat{\theta}^{(i,k)} - \theta^{(i)}\| \leq \|(I_n - H)^{-1}\|\|H\theta^{(i)}\| \lesssim \|H\theta^{(i)}\|. \tag{7.55}$$

Next, by (2.16), with probability at least $1 + o(n^{-3})$, $\hat{\lambda}_k \asymp \|\theta\|^2$. It follows from basic algebra that

$$\|H\theta^{(i)}\|^2 \leq \frac{1}{\hat{\lambda}_k^2}(I + II) \leq C(I + II)/\|\theta\|^4, \tag{7.56}$$

where $I = \|W\theta^{(i)}\|^2$ and $II = \|U\theta^{(i)}\|^2$. Note that, first, since $\|\theta\|_\infty < 1$,

$$II \leq \|\theta\|_6^6 \leq \|\theta\|_3^3. \tag{7.57}$$

Second, by the assumption (2.17) and Lemma 4.2, with probability at least $1 + o(n^{-3})$,

$$I \leq C\log(n)\|\theta\|_1\|\theta\|_3^3. \tag{7.58}$$

Combining (7.57)-(7.58) with (7.55)-(7.56) gives

$$\|\hat{\theta}^{(i,k)} - \theta^{(i)}\|^2 \leq C\log(n)(\|\theta\|_1 + 1)\|\theta\|_3^3/\|\theta\|^4. \tag{7.59}$$

By (2.11) and basic algebra, $\|\theta\|_1 \geq \|\theta\|^2 \geq \log(n)$, and so $\|\theta\|_1 + 1 \leq 2\|\theta\|_1$. Inserting this into (7.59) gives the claim. □

7.12. *Proof of Lemma 7.4.* Write for short $U = \mathrm{diag}(\Omega)$ and $H = (W - \mathrm{diag}(\Omega))/\hat{\lambda}_k$. Since $(I_n - H)^{-1} - I_n = H + H(I_n - H)^{-1}H$, it follows from definitions and basic algebra that

$$\|\Theta^{-1}[\hat{\theta}^{(i,k)} - \theta^{(i)}]\|^2 = \|\Theta^{-1}[(I_n - H)^{-1} - I_n]\theta^{(i)}\|^2 \leq 2(I + II), \tag{7.60}$$

where

$$I = \|\Theta^{-1}H\theta^{(i)}\|^2, \qquad II = \|\Theta^{-1}H(I_n - H)^{-1}H\theta^{(i)}\|^2.$$

Consider I first. Similarly, by Lemma 2.3, with probability at least $1 + o(n^{-3})$, $\hat{\lambda}_k \asymp \|\theta\|^2$, and so

$$I = \frac{1}{\hat{\lambda}_k^2}\|\Theta^{-1}W\theta^{(i)} - \Theta^{-1}U\theta^{(i)}\| \leq \frac{C}{\|\theta\|^4}[Ia + Ib], \tag{7.61}$$

where

$$Ia = \|\Theta^{-1}W\theta^{(i)}\|^2, \qquad Ib = \|\Theta^{-1}U\theta^{(i)}\|^2.$$

Now, first, since $\|\theta\|_\infty < 1$,

$$Ib \leq \|\theta\|_4^4 \leq \|\theta\|_3^3. \tag{7.62}$$

Second, by (2.17) and Lemma 4.2, with probability at least $1 + o(n^{-3})$,

$$Ia \leq C\log(n)\|\theta\|_3^3 \sum_{i=1}^n (1/\theta(i)). \tag{7.63}$$

Inserting (7.62)-(7.63) into (7.61) gives that with probability at least $1 + o(n^{-3})$,

$$I \leq C\log(n)\frac{\|\theta\|_3^3}{\|\theta\|^4} \cdot \Big[1 + \sum_{i=1}^n \frac{1}{\theta(i)}\Big] \leq C\log(n)\frac{\|\theta\|_3^3}{\|\theta\|^4} \cdot \Big[\sum_{i=1}^n \frac{1}{\theta(i)}\Big]. \tag{7.64}$$

Next, we analyze II. By definitions, $II = \frac{1}{\hat{\lambda}_k^4}\|\Theta^{-1}(W-U)(I_n-H)^{-1}(W-H)\theta^{(i)}\|^2$. Recalling that $\hat{\lambda}_k \asymp \|\theta\|^2$ with probability at least $1 + o(n^{-3})$, and so by basic algebra,

$$II \leq \frac{1}{\|\theta\|^8}\|\Theta^{-1}(W - U)(I_n - H)^{-1}(W - U)\theta^{(i)}\|^2 \leq \frac{1}{\|\theta\|^8} IIa \cdot IIb, \tag{7.65}$$

where $IIa = \|\Theta^{-1}(W - U)(I - H)^{-1}\|^2$ and $IIb = \|(W - U)\theta^{(i)}\|^2$.

Consider IIa first. By Lemma 2.2, with probability $1+o(n^{-3})$, $\|H\| = o(1)$. Therefore,

$$IIa \lesssim \|\Theta^{-1}(W-U)\|^2 \leq C\big[\|\Theta^{-1}W\|^2 + \|\Theta^{-1}U\|^2\big].$$

Next, by (2.18) and Lemma 4.1, we have with probability at least $1+o(n^{-3})$,

$$\|\Theta^{-1}W\|^2 \leq C\log(n)\max\{\theta_{max}\sum_{i=1}^n \frac{1}{\theta(i)},\ \frac{1}{\theta_{min}}\|\theta\|_1\}.$$

At the same time, it is seen $\|\Theta^{-1}U\|^2 \leq 1$, which is much smaller than the right hand side of the equation above. Combining these gives

$$IIa \leq C\log(n)\max\{\theta_{max}\sum_{i=1}^n \frac{1}{\theta(i)},\ \frac{1}{\theta_{min}}\|\theta\|_1\}. \tag{7.66}$$

Next, we consider IIb. Write

$$IIb = \|(W-U)\theta^{(i)}\|^2 \leq C\big[\|W\theta^{(i)}\|^2 + \|U\theta^{(i)}\|^2\big].$$

On one hand, by Lemma 4.2, with probability at least $1+o(n^{-3})$,

$$\|W\theta^{(i)}\|^2 \leq C\log(n)\|\theta\|_1\cdot\|\theta\|_3^3.$$

On the other hand, since $\|\theta\|_\infty < 1$, by definitions and direct calculations,

$$\|U\theta^{(i)}\|^2 \leq \|\theta\|_4^4 \leq \|\theta\|_3^3.$$

Recall that that (2.11) implies $\|\theta\|_1 \geq \|\theta\|^2 \geq 1$. Combining these gives

$$IIb \leq C\log(n)[1+\|\theta\|_1]\|\theta\|_3^2 \leq C\log(n)\|\theta\|_1\cdot\|\theta\|_3^3. \tag{7.67}$$

Inserting (7.66)-(7.67) into (7.65) gives

$$II \leq C\log^2(n)\frac{\|\theta_1\|\cdot\|\theta\|_3^3}{\|\theta\|^8}\max\{\theta_{max}\sum_{i=1}^n \frac{1}{\theta(i)},\ \frac{1}{\theta_{min}}\|\theta\|_1\}. \tag{7.68}$$

Inserting (7.64) and (7.68) into (7.60), $\|\Theta^{-1}[\hat{\theta}^{(i,k)}-\theta^{(i)}]\|^2$ does not exceed

$$\frac{C\log(n)\|\theta\|_3^3}{\|\theta\|^4}\Big[\sum_{i=1}^n \frac{1}{\theta(i)} + \frac{\log(n)\|\theta\|_1}{\|\theta\|^4}\max\{\theta_{max}\sum_{i=1}^n \frac{1}{\theta(i)},\ \frac{1}{\theta_{min}}\|\theta\|_1\}\Big].$$

By (2.11), $\log(n)\theta_{max}\|\theta\|_1/\|\theta\|^4 \to 0$, and so

$$\|\Theta^{-1}[\hat{\theta}^{(i,k)}-\theta^{(i)}]\|^2 \leq \frac{C\log(n)\|\theta\|_3^3}{\|\theta\|^4}\Big[\sum_{i=1}^n \frac{1}{\theta(i)} + \frac{1}{\theta_{min}}\frac{\log(n)\|\theta\|_1^2}{\|\theta\|^4}\Big],$$

and the claim follows. □

Acknowledgements. The author would like to thank Joel Tropp and Roman Vershynin for helpful pointers.

REFERENCES

[1] Adamic L, Glance N (2005) The political blogosphere and the 2004 U.S. election. In *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem.*

[2] Bai Z, Silverstein J (2009) *Spectral analysis of large dimensional random matrices.* Springer, NY.

[3] Bickel P, Chen A (2009) A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106**, 21068-21073.

[4] Boots B, Gordon G (2011) Online spectral identification of dynamical systems. *NIPS workshop on Sparse Representation and Low-rank Approximation.*

[5] Box G, Draper R (1987). *Empirical model-building and response surfaces.* Wiley.

[6] Candes E, Li X, Ma Y, Wright J (2011) Robust Principal Component Analysis? *J. ACM* **58**(3).

[7] Chaudhuri K, Fan C, Tsiatas A (2012). Spectral clustering of graphs with general degrees in the extended planted partition of model. *J. Mach. Learn. Res.*, 1-23.

[8] Chen A, Amin A, Bickel P, Levina E (2012) Fitting community models for large networks. *arXiv: 1207.2340.*

[9] Choi D, Wolfe P, Airoldi E (2012) Stochastic blockmodels with growing number of classes. *Biometrika* **99**, 273-284.

[10] Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

[11] Girvan M, Newman ME (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**, 7821-7826.

[12] Goldenberg A, Zheng A, Fienberg S, Airoldi E (2009) A survey of statistical network models. *Foundations and Trends in Machine Learning* **2**, 129-233.

[13] Hastie T, Tibshirani R, Friedman J (2001). *The Elements of Statistical Learning.* Springer.

[14] Hoff P (2007) Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems* **19**, Cambridge.

[15] Horn R, Johnson C (1985) *Matrix analysis.* Cambridge Press.

[16] Jin J, Wang W (2012) Optimal spectral clustering by Higher Criticism thresholding. *Working Manuscript.*

[17] Jin J, Zhang Q (2012) New spectral methods for community detection with bipartite networks. *Working Manuscript.*

[18] Karrer B, Newman K (2011) Stochastic blockmodels and community structures in networks. *Physical Review E* **83**, 016107.

[19] Kolaczyk E (2009) *Statistical analysis of network data: methods and models.* Springer, NY.

[20] Liu H, Xu M, Gu H, Gupta A, Lafferty J, Wasserman L (2011). Forest density estimation. *J. Mach. Learn. Res.* **12**, 907-951.

[21] Nayak R, Kearns M, Spielman R, Cheung V (2009) Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Res.* **19**, 1953-1962.

[22] Finding community structure in network using the eigenvectors of matrices. *Phys. Rev. E* **74**(3), 036104.

[23] Mewman M (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103**(23), 8577-8582.

[24] Perry P, Wolfe P (2012) Null models for network data. *arXiv:1201.5871*.

[25] Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39**, 1878-1915.

[26] Shorack GR, Wellner JA (1986) *Empirical processes with applications to statistics*. John Wiley & Sons.

[27] Tropp J (2012) User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12**, 389-434.

[28] Tukey JW (1965) Which part of the sample contains the information? *Proc. Natl. Acad. Sci.* **153**, 127-134.

[29] Tulino A, Verdu S (2004) *Random matrix theory and wireless communications*. now, NL.

[30] Yan X, Jensen J, Krzakala F, *et al.* (2012) Model selection for degree-corrected block model.

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

[32] Zhang S, Zhao H (2012) Community identification in networks with unbalanced structure. *Phys. Rev. E.* **85**(6), 066114.

[33] Zhao Y, Levina L, Zhu J (2011). Consistency of community detection in network under degree-corrected stochastic block models. *arXiv: 1110.3854v3*.

J. Jin
Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213
USA
E-mail: jiashun@stat.cmu.edu