Improvements on SCORE, Especially for Weak Signals

Jiashun Jin and Shengming Luo
Carnegie Mellon University Pittsburg USA
Zheng Tracy Ke
Harvard University Cambridge USA

Abstract

A network may have weak signals and severe degree heterogeneity, and may be very sparse in one occurrence but very dense in another. SCORE (Ann. Statist. **43**, 57–89, 2015) is a recent approach to network community detection. It accommodates severe degree heterogeneity and is adaptive to different levels of sparsity, but its performance for networks with weak signals is unclear. In this paper, we show that in a broad class of network settings where we allow for weak signals, severe degree heterogeneity, and a wide range of network sparsity, SCORE achieves prefect clustering and has the so-called "exponential rate" in Hamming clustering errors. The proof uses the most recent advancement on entry-wise bounds for the leading eigenvectors of the network adjacency matrix. The theoretical analysis assures us that SCORE continues to work well in the weak signal settings, but it does not rule out the possibility that SCORE may be further improved to have better performance in real applications, especially for networks with weak signals. As a second contribution of the paper, we propose SCORE+ as an improved version of SCORE. We investigate SCORE+ with 8 network data sets and found that it outperforms several representative approaches. In particular, for the 6 data sets with relatively strong signals, SCORE+ has similar performance as that of SCORE, but for the 2 data sets (Simmons, Caltech) with possibly weak signals, SCORE+ has much lower error rates. SCORE+ proposes several changes to SCORE. We carefully explain the rationale underlying each of these changes, using a mixture of theoretical and numerical study.

AMS (2000) *subject classification.* Primary: 62H30, 91C20; Secondary: 62P25.

1 Introduction

Community detection is a problem that has received considerable attention (Karrer and Newman, 2011; Zhang et al., 2020; Bickel and Chen, 2009).

Consider an undirected network $\mathcal{N}$ and let A be its adjacency matrix:

$$A(i,j) = \begin{cases} 1, & \text{if there is an edge connecting node } i \text{ and } j, \\ 0, & \text{otherwise.} \end{cases}$$

Since the network is undirected, A is symmetrical. Also, as a convention, we do not consider self edges, so all diagonal entries of A are 0. We assume the network is connected, consisting of K perceivable non-overlapping communities

$$\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K.$$

Following the convention in many recent works on community detection (e.g., Bickel and Chen 2009; Zhang et al. 2020), we assume K as known and the nodes do not have mixed-memberships, so each node belongs to *exactly* one of the K communities. The community labels are unknown, and the goal is to use (A, K) to predict them. In statistics, this is known as the clustering problem.

See Jin et al. (2020) and reference therein for discussions on how to estimate K, and Jin et al. (2017) for the generalization of SCORE for network analysis in the presence of mixed memberships.

Similar to "cluster", "community" is a concept that is scientifically meaningful but mathematically hard to define. Intuitively, communities are clusters of nodes that have more edges "within" than "across" (Jin, 2015; Zhao et al., 2012). Note that "communities" and "components" are different concepts: two communities may be connected, while two components are always disconnected.

Table 1 presents 8 network data sets which we analyze in this paper. Data sets 2-3 are from Traud et al. (2011, 2012) (see also Chen et al. 2018; Ma et al. 2020), and the other 6 datasets are downloaded from http://www-personal.umich.edu/~mejn/netdata/. For all these data sets, the true labels are suggested by the original authors or data curators, and we use the labels as the "ground truth."

Conceivably, for some of the data sets, some nodes may have mixed memberships (Airoldi et al., 2008; Jin et al., 2017; Zhang et al., 2020). To alleviate the effect, we did some data pre-processing as follows. For the Polbooks data set, we removed all the books that are labeled as "neutral." For the football data set, we removed the 5 "independent" teams. For the UKfaculty data set, we removed the smallest group which only contains 2 nodes. After the pre-processing, our assumption of "no mixed-memberships" is reasonable.

Natural networks have some noteworthy features.

Table 1: The 8 network data sets we analyze in this paper. Note that d_{max}/d_{min} can be as large as a few hundreds, suggesting a severe degree heterogeneity (d_{min}, d_{max} and $\bar{d}$ stand for the minimum degree, maximum degree, and average degree, respectively)

Dataset	Source	#Nodes	K	#Edges	d_{min}	d_{max}	$\bar{d}$
Weblogs	Adamic and Glance (2005) (2005)	1222	2	16714	1	351	27.35
Simmons	Traud et al. (2011)	1137	4	24257	1	293	42.67
Caltech	Traud et al. (2011)	590	8	12822	1	179	43.36
Football	Girvan and Newman (2002)	110	11	570	7	13	10.36
Karate	Zachary (1977)	34	2	78	1	17	4.59
Dolphins	Lusseau et al. (2003)	62	2	159	1	12	5.12
Polbooks	Krebs (unpublished)	92	2	374	1	24	8.13
UKfaculty	Nepusz et al. (2008)	79	3	552	2	39	13.97

- *Node sparsity and severe degree heterogeneity.* Take Table 1 for example, even for networks with only 1222 nodes, the degrees for some nodes can be as large as 351 times higher than those of the others. If we measure the sparsity of a node by its degree, then the sparsity level may range significantly from one node to another.

- *Overall network sparsity.* Some networks are much sparser than others, and the overall network sparsity may range significantly from one network to another.

- *Weak signal.* In many cases, the community structures are subtle and masked by strong noise, where the signal-to-noise ratio (SNR) is relatively low.

It is desirable to have a model that is flexible enough to capture all these features. This is where the DCBM comes in.

1.1. The Degree-Corrected Block Model (DCBM) DCBM is one of the most popular models in network analysis (see for example Karrer and Newman 2011). For each node i, we encode the community label by a K-dimensional vector π_i, such that for all $1 \le i \le n$ and $1 \le k \le K$,

$$i \in \mathcal{C}_k \text{ if and only if all entries of } \pi_i \text{ are 0 except that the } k\text{th entry is 1.}$$

$$(1.1)$$

In DCBM, for a matrix $P \in \mathbb{R}^{K,K}$ and parameters $\theta_1, \theta_2, \ldots, \theta_n$, we assume the upper triangle of the adjacency matrix A contains independent Bernoulli random variables satisfying

$$\mathbb{P}(A(i,j) = 1) = \underbrace{\theta_i \cdot \theta_j}_{\text{node specific}} \times \underbrace{\pi_i' P \pi_j}_{\text{community specific}} . \tag{1.2}$$

Here, P is a symmetrical and (entry-wise) non-negative matrix that models the community structure and $\theta_1, \theta_2, \ldots, \theta_n$ are positive parameters that model the degree heterogeneity. For identifiability, we assume

$$\text{all diagonal entries of } P \text{ equal } to \text{ 1.} \tag{1.3}$$

Writing $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$, $\Theta = \text{diag}(\theta_1, \theta_2, \ldots, \theta_n)$, and $\Pi = [\pi_1, \pi_2, \ldots, \pi_n]'$, we define

$$\Omega = \Theta \Pi P \Pi' \Theta. \tag{1.4}$$

Note that when $i \neq j$, $\Omega(i,j)$ denotes the probability $\mathbb{P}(A(i,j) = 1)$. Let $W \in \mathbb{R}^{n,n}$ be the *centered Bernoulli noise* matrix such that $W(i,j) = A(i,j) - \Omega(i,j)$ when $i \neq j$ and $W(i,j) = 0$ if $i = j$. We have[1]

$$A = \Omega - \text{diag}(\Omega) + W = \text{``main signal''} + \text{``secondary signal''} + \text{``noise''}, \tag{1.5}$$

where $\text{diag}(\Omega)$ stands for the diagonal matrix $\text{diag}(\Omega(1,1), \Omega(2,2), \ldots, \Omega(n,n))$. Note that the rank of Ω is K, so Eq. 1.5 is a low-rank matrix model.

In the special case of[2]

$$\theta_1 = \theta_2 = \ldots = \theta_n = \alpha, \tag{1.6}$$

DCBM reduces to the stochastic block model (SBM). Note that SBM does not model severe degree heterogeneity. The DCBM is also similar to that in Chaudhuri et al. (2012) in some sense.

In DCBM, we allow (θ, Π, P) to depend on n so the model is flexible enough to capture all the three features aforementioned of natural networks.

- A reasonable metric for degree heterogeneity is $\theta_{max}/\theta_{min}$, which is allowed to be large in DCBM. See Table 1.

---

[1] We model $\mathbb{E}[A]$ by $\Omega - \text{diag}(\Omega)$ instead of Ω because the diagonals of $\mathbb{E}[A]$ are all 0. Here, "main signal", "secondary signal", and "noise" refers to Ω, $-\text{diag}(\Omega)$ and W respectively.
[2] For SBM, the diagonal entries of P can be unequal. DCBM has more free parameters, so we have to assume that P has unit diagonal entries to maintain identifiability.

- A reasonable metric for overall network sparsity is $\|\theta\|$, and in DCBM, $\|\theta\|$ depends on n and is allowed to range freely between 1 and $\sqrt{n}$ (up to some multi-log(n) terms),[3] corresponding to the most sparse networks and the most dense networks, respectively.

- A reasonable metric for SNR is $\lambda_K/\sqrt{\lambda_1}$ (see Jin et al. 2019 for an explanation), where λ_k is the kth largest eigenvalue (in magnitude) of Ω. If we allow θ, P, and Π to depend on n, then DCBM is adequate for modeling the weak signal cases where $|\lambda_k|$ may be much smaller than $|\lambda_1|$, $1 < k \leq K$.

In many recent works on community detection, it was assumed that the first K eigenvalues are at the same magnitude. For example, some of these works considered a DCBM model where in Eq. 1.4, we take $P = \alpha_n P_0$. Here, α_n is a scaling parameter that may vary with n and P_0 is a fixed matrix that does not vary with n. In this special case, by similar calculations as in Jin (2015), it is seen that all eigenvalues of Ω are at the same order under mild regularity conditions on (Θ, Π) (e.g., the K communities are balanced; see Jin (2015) for details). Such models do not allow for weak signals, and so are relatively restrictive.

Motivated by these observations, it is desirable to have community detection algorithms that

- accommodate severe degree heterogeneity,

- are adaptive to different levels of overall network sparsity,

- are effective not only for strong signals but also for weak signals.

1.2.   The Orthodox SCORE SCORE, or **S**pectral **C**lustering **O**n **R**atios-of-**E**igenvectors, is a recent approach to community detection proposed by Jin (2015). SCORE consists of three steps.

Orthodox SCORE   Input: adjacency matrix A and the number of communities K. Output: community labels of all nodes.

- *(PCA).* Obtain the first K leading eigenvectors $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K$ of A (we call $\hat{\xi}_k$ the kth leading eigenvector if the corresponding eigenvalue is the kth largest in absolute value).

---

[3]A multi-log(n) term is a term $L_n > 0$ that satisfies "$L_n n^{-\delta} \to 0$ and $L_n n^{\delta} \to \infty$ for any fixed constant $\delta > 0$

- *(Post-PCA normalization).* Obtain the $n \times (K-1)$ *matrix of entry-wise eigen-ratios* by

$$\left[\frac{\hat{\xi}_2}{\hat{\xi}_1}, \frac{\hat{\xi}_3}{\hat{\xi}_1}, \ldots, \frac{\hat{\xi}_K}{\hat{\xi}_1} \right], \tag{1.7}$$

  where the ratio of two vectors should be understood as *the vector of entry-wise ratios.*[4]

- *(Clustering).* Cluster by applying k-means to rows of $\hat{R}$, assuming there are $\leq K$ clusters.

Compared to classical spectral clustering, the main innovation of SCORE is the post-PCA normalization. The goal of this step is to mitigate the effect of degree heterogeneity. The degrees contain very little information of the community structure and pose merely as a nuisance, but severe degree heterogeneity makes different entries of the leading eigenvectors badly scaled. As a result, without this step, SCORE tends to cluster nodes according to their degrees instead of the community structure, and thus have unsatisfactory clustering results. Take the Weblog data for example: with and without this step, the error rates of SCORE are 58/1222 and 437/1222 respectively. See Jin (2015) for more discussions.

SCORE is conceptually simple, easy to use, and does not need tuning. In Jin (2015), Ji and Jin (2016), SCORE was shown to be competitive in clustering accuracy. For computational time, note that in the k-means clustering step of SCORE, one usually uses the Llyod's algorithm (Hastie et al., 2009); and as a result, SCORE is computationally fast and is able to work efficiently for large networks. See Jin (2015) and also Table 4 of the current paper for more discussions.

SCORE is a flexible idea, and can be conveniently extended to many different settings such as network mixed membership estimation (Jin et al., 2017), topic estimation in text mining (Ke and Wang, 2017), state aggregation in control theory and reinforcement learning (Duan et al., 2018), analysis of hyper graphs (Ke et al., 2020), and matrix factorization in image processing.

---

[4]For example, $\frac{\hat{\xi}_2}{\hat{\xi}_1}$ is the n-dimensional vector $(\frac{\hat{\xi}_2(1)}{\hat{\xi}_1(1)}, \frac{\hat{\xi}_2(2)}{\hat{\xi}_1(2)}, \ldots, \frac{\hat{\xi}_2(n)}{\hat{\xi}_1(n)})'$. Note that we may choose to threshold all entries of the $n \times (K-1)$ matrix by $\pm \log(n)$ from top and bottom (Jin, 2015), but this is not always necessary. For all data sets in this paper, thresholding or not only has a negligible difference.

1.3.   Contribution of this Paper For the three features aforementioned, SCORE accommodates severe degree heterogeneity and is adaptive to different levels of overall network sparsity. However, when it comes to weak signals, there are at least two problems that are not answered.

- What is the theoretical behavior of SCORE in the presence of weak signals?

- In challenging application problems where the SNR is small, is it possible to improve SCORE to have better real data performance, without sacrificing its good properties above?

Note that in the literature, the theoretical analysis on SCORE has been largely focused on the case where the signals are relatively strong; see for example (Jin, 2015).

In this paper, we analyze SCORE theoretically, especially for the weak signal settings. We show that for a broad class of settings where we allow for weak signals, severe degree heterogeneity, and a wide range of overall network sparsity, SCORE attains an exponential rate of convergence for the Hamming error. We also show that, when the SNR is appropriately large, SCORE fully recovers the community labels except for a small probability. The proof uses the most recent advancement on entry-wise bounds (a kind of large-deviation bounds) for the leading eigenvectors of the adjacency matrix (Abbe et al., 2019; Jin et al., 2017).

The theoretical analysis here assures that SCORE continues to work well for weak signal settings. This of course does not rule out the possibility that a further improved version may perform better in real data analysis.

As a second contribution of the paper, we propose SCORE+ as an improved version of SCORE, especially for networks with weak signals. We compare SCORE+ with SCORE and several other recent algorithms using the 8 data sets in Table 1. For the 6 data sets where the signals are relatively strong (the clustering errors of all methods considered are relatively low), SCORE+ and SCORE have comparable performance. For the 2 data sets (Simmons and Caltech) where the signals are relatively weak (the clustering errors of all methods considered are relatively high), SCORE+ improves SCORE significantly, and has the lowest error rates among all methods considered in the paper.

SCORE+ proposes several changes to SCORE. We carefully explain the rationale underlying each of these changes, using a mixture of theoretical and numerical study. A much deeper understanding requires advanced tools in random matrix theory that have not yet been developed, so we leave the study along this line to the future.

1.4. Content and Notations In Section 2, we analyze the orthodox SCORE with some new theoretical results. We show that SCORE attains exponential rates in Hamming clustering errors and achieves perfect clustering provided that the SNR is reasonably large. In Section 3, we propose SCORE+ as an improved version of SCORE. We compare the performance of SCORE+ with SCORE and several recent approaches on community detection using the 8 data sets in Table 1, and show that SCORE+ has the best overall performance. SCORE+ proposes several changes to SCORE. We explain the rationale underlying each of these changes, and especially why SCORE+ is expected to have better performance than SCORE for networks with weak signals. Section 4 proves the main results in Section 2.

In this paper, for any numbers $\theta_1, \theta_2, \ldots, \theta_n$, $\theta_{max} = \max\{\theta_1, \theta_2, \ldots, \theta_n\}$, and $\theta_{min} = \min\{\theta_1, \theta_2, \ldots, \theta_n\}$. Also, $\mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_n)$ denotes the $n \times n$ diagonal matrix with θ_i being the i-th diagonal entry, $1 \leq i \leq n$, For any vector $a \in \mathbb{R}^n$, $\|a\|_q$ denotes the Euclidean ℓ^q-norm, and we write $\|a\|$ for simplicity when $q = 2$. For any matrix $P \in \mathbb{R}^{n,n}$, $\|P\|$ denotes the matrix spectral norm, and $\|P\|_{\max}$ denotes the maximum ℓ^2-norm of all the rows of P. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we say $a_n \asymp b_n$ if there are constants $c_2 > c_1 > 0$ such that $c_1 a_n \leq b_n \leq c_2 a_n$ for sufficiently large n.

2   SCORE: Exponential Rate and Perfect Clustering

We provide new theoretical results for the orthodox SCORE, which significantly improves those in Jin (2015). For the "weak signal" case, the theory in Jin (2015) is not applicable but our theory applies. For the "strong signal" case, compared with Jin (2015), our theory provides a faster rate of convergence for the clustering error and weaker conditions for perfect clustering.

Consider a sequence of DCBM indexed by n, where (K, θ, Π, P) are all allowed to depend on n. Suppose, for a constant $c_1 > 0$,

$$\|P\|_{\max} \leq c_1, \qquad \|\theta\| \to \infty, \qquad \text{and} \qquad \theta_{\max} \leq c_1. \qquad (2.8)$$

Recall that $\mathcal{C}_1, \ldots, \mathcal{C}_K$ denote the K true communities. For $1 \leq k \leq K$, let $n_k = |\mathcal{C}_k|$ be the size of community k, and let $\theta^{(k)} \in \mathbb{R}^n$ be the vector such that $\theta_i^{(k)} = \theta_i$ if $i \in \mathcal{C}_k$ and $\theta_i^{(k)} = 0$ otherwise. We assume, for a constant $c_2 > 0$,

$$\max_{1 \leq k \leq K}\{n_k\} \leq c_2 \min_{1 \leq k \leq K}\{n_k\}, \qquad \text{and} \qquad \max_{1 \leq k \leq K}\{\|\theta^{(k)}\|\} \leq c_2 \min_{1 \leq k \leq K}\{\|\theta^{(k)}\|\}. \qquad (2.9)$$

Introduce a diagonal matrix $G \in \mathbb{R}^{K,K}$ by

$$G = K\|\theta\|^{-2} \cdot \text{diag}(\|\theta^{(1)}\|^2, \|\theta^{(2)}\|^2, \ldots, \|\theta^{(K)}\|^2).$$

Let μ_k denote the kth largest eigenvalue (in magnitude) of the $K \times K$ matrix $G^{1/2}PG^{1/2}$, and let η_k denote the corresponding eigenvector. We assume, for a constant $c_3 \in (0,1)$ and $c_4 > 0$,

$$\min_{2 \leq k \leq K} |\mu_1 - \mu_k| \geq c_3|\mu_1|, \quad \text{and } \eta \text{ is a positive vector to } \min_{1 \leq k \leq K}\{\eta_1(k)\} > 0 \tag{2.10}$$

These conditions are mild. For Eq. 2.8, recall that $\|\theta\|$ measures the overall network sparsity, and the interesting range of $\|\theta\|$ is between 1 and $\sqrt{n}$, up to some multi-log(n) terms (see footnote 3). Therefore, it is mild to assume $\|\theta\| \to \infty$. Condition (2.9) requires that the communities are balanced in size and in total squared degrees, which is mild.

Condition (2.10) is also mild. The most challenging case for network analysis is when the matrix P gets very close to the matrix of all ones, where it is hard to distinguish one community from another. The condition only rules out the less relevant cases such as the network is disconnected or approximately so. These are strong signal cases where it is relatively easy to distinguish one community from another. Note that the condition is satisfied if all entries of P are lower bounded by a constant or if K is fixed and P converges to a fixed irreducible matrix (see Section A.2 of Jin et al. 2017 for a discussion of this condition).

In a hypothesis testing framework, Jin et al. (2019) has pointed out that a reasonable metric for SNR in a DCBM is $|\lambda_K|/\sqrt{\lambda_1}$, where we recall that λ_k denotes the kth largest eigenvalue (in magnitude) of $\Omega = \Theta\Pi P\Pi'\Theta$ and that λ_1 is always positive (Jin, 2015; Jin et al., 2017). We also introduce a quantity to measure the severity of degree heterogeneity:

$$\alpha(\theta) = (\theta_{\min}/\theta_{\max}) \cdot (\|\theta\|/\sqrt{\theta_{\max}\|\theta\|_1}) \quad \in \quad (0,1].$$

The smaller $\alpha(\theta)$, the more severe the degree heterogeneity. When $\theta_{\max} \asymp \theta_{\min}$, $\alpha(\theta)$ is bounded below from 0 by a constant. In the presence of severe degree heterogeneity, $\alpha(\theta)$ gets close to 0. We shall see that the clustering power of SCORE depends on

$$s_n = \alpha(\theta) \cdot (|\lambda_K|/\sqrt{\lambda_1}), \tag{2.11}$$

which is a combination of the SNR and the severity of degree heterogeneity.

Let $\widehat{\Pi} = [\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n]'$ denote the matrix of estimated community labels by the orthodox SCORE. Define the Hamming error rate (per node) for clustering by

$$\text{Hamm}(\widehat{\Pi}, \Pi) = \frac{1}{n} \sum_{i=1}^n 1\{\hat{\pi}_i \neq \pi_i\}, \qquad \text{up to a permutation on columns of } \widehat{\Pi}.$$

The next theorem is proved in Section 4.

THEOREM 2.1. *Consider a sequence of DCBM indexed by n, where* Eqs. 2.8 *–*2.10 *hold. Let s_n be as in* Eq. 2.11. *There exist appropriately small constants $a_1, a_2 > 0$, which depend on the constants c_1-c_4 in the regularity conditions, such that, as long as $s_n \geq a_1^{-1} K^4 \sqrt{\log(n)}$, for sufficiently large n:*

$$\mathbb{E}\left[\text{Hamm}(\widehat{\Pi}, \Pi)\right] \leq \frac{2K}{n} \sum_{i=1}^n \exp\left(-a_2\theta_i \cdot \min\left\{\frac{(|\lambda_K|/\sqrt{\lambda_1})^2 \|\theta\|^2}{K^2 \|\theta\|_3^3}, \frac{(|\lambda_K|/\sqrt{\lambda_1})\|\theta\|}{K\theta_{\max}}\right\}\right)$$
$$+ o(n^{-3}).$$

Theorem 2.1 implies that the Hamming clustering error of SCORE depends on $(\theta_1, \ldots, \theta_n)$ in an exponential form. This significantly improves the bound in Jin (2015), which depends on $(\theta_1, \ldots, \theta_n)$ in a polynomial form. Additionally, Theorem 2.1 suggests that the nodes with smaller θ_i have large contributions to the Hamming clustering error, i.e., it is more likely for the algorithm to make errors on low-degree nodes.

The clustering error has an easy-to-digest form in special examples.

Corollary 2.1. *Consider a special DCBM, where*

$$P = (1 - b)I_K + b\mathbf{1}_K \mathbf{1}_K', \qquad \pi_i \overset{iid}{\sim} \text{Uniform}(\{e_1, e_2, \ldots, e_K\}).$$

Suppose K is fixed and θ satisfies that $\theta_{\max} \leq C\theta_{\min}$. There exist appropriately small constants $\tilde{a}_1, \tilde{a}_2 > 0$, such that, as long as $(1 - b)\|\theta\| \geq a_1^{-1} K^4 \sqrt{\log(n)}$, for sufficiently large n,

$$\mathbb{E}\left[\text{Hamm}(\widehat{\Pi}, \Pi)\right] \leq \frac{2K}{n} \sum_{i=1}^n \exp\left(-a_2 \frac{\theta_i}{\bar{\theta}} \cdot \frac{(1 - b)^2 \|\theta\|^2}{K^3}\right) + o(n^{-3}).$$

In this special example, $\|\theta\|^2$ characterizes the average node degrees, and $(1-b)$ captures the "similarity" across communities. The clustering power of SCORE is governed by $s_n \asymp (1-b)\|\theta\|^2$. The bound in Corollary 2.1 matches

with the minimax bound in Gao et al. (2018), except for the constant $2K$ in front and the constant a_2 in the exponent.[5] It was shown in Gao et al. (2018) that the exponential error rate can be attained by first applying spectral clustering and then conducting a refinement, where the refinement step was motivated by technical convenience. In fact, numerical studies suggest that spectral clustering alone can attain exponential error rates. Theorem 2.1 and Corollary 2.1 provide a rigorous theoretical justification.

The next theorem states that SCORE can exactly recover the community labels with high probability, provided that the SNR is appropriately large.

THEOREM 2.2. *Consider a sequence of DCBM indexed by n, where* Eqs. 2.8 –2.10 *hold. Let s_n be as in* Eq. 2.11. *If $s_n \geq CK^4\sqrt{\log(n)}$ for a sufficiently large constant $C > 0$, then we have that, up to a permutation on columns of $\widehat{\Pi}$,*

$$\mathbb{P}(\widehat{\Pi} \neq \Pi) = o(n^{-3}).$$

Furthermore, if K is finite and $\theta_{\max} \leq C\theta_{\min}$, then the above is true as long as $|\lambda_K|/\sqrt{\lambda_1}$ exceeds a sufficiently large constant.

The condition on s_n cannot be significantly improved. Take the case of fixed K and with moderate degree heterogeneity ($\theta_{\max} \leq C\theta_{\min}$) for example. In this case, the condition becomes $|\lambda_K|/\sqrt{\lambda_1} \geq C$ for a large enough constant $C > 0$. It was shown in Jin et al. (2019) that, if we allow $|\lambda_K|/\sqrt{\lambda_1} \to 0$, then we end up with a class of models that is too broad so we can find two sequences of DCBM models with different (fixed) K but are indistinguishable from each other. In such settings, successful clustering is impossible.

In the literature, the exponential error rate and the perfect clustering property were mainly obtained for non-spectral methods (e.g., Gao et al. 2018; Chen et al. 2018). While spectral methods are practically popular, its theoretical analysis is challenging, since it requires sharp entry-wise bounds for eigenvectors. A few existing works either focus on SBM which does not allow for degree heterogeneity (e.g., Abbe et al. 2019; Su et al. 2019) or restrict to the "strong signal" case and dense networks (e.g., Liu et al. 2019). Our results are new as we provide the first exponential rate result and perfect clustering result for spectral methods that accommodate severe degree heterogeneity, sparse networks, and weak signals.

Our analysis uses some results on the spectral analysis of the adjacency matrices from our work (Jin et al., 2017), especially the entry-wise large-deviation bounds for empirical eigenvectors. We refer interesting readers to

---

[5]When translating the bound in Gao et al. (2018), we notice that θ_i there have been normalized, so that their θ_i corresponds to our $(\theta_i/\bar{\theta})$.

a detailed description in Jin et al. (2017). It is understood that the main technical difficulty of analyzing spectral methods lies in entry-wise analysis of eigenvectors. Recent progress includes (but does not limit to) Abbe et al. (2019), Fan et al. (2019), Jin et al. (2017), Liu et al. (2019), Mao et al. (2020) & Su et al. (2019).

3   SCORE+, a Refinement Especially for Weak Signals

We propose SCORE+ as a refinement of SCORE for community detection. SCORE+ inherits the appealing features of SCORE. It improves the performance of SCORE in real applications, especially for networks with weak signals.

3.1. SCORE+ Recall that under DCBM,

$$A = \Omega - \mathrm{diag}(\Omega) + W = \text{``main signal''} + \text{``secondary signal''} + \text{``noise''},$$

where the "main signal" matrix Ω equals to $\Theta\Pi P\Pi'\Theta$ and has a rank K. SCORE+ is motivated by several observations about SCORE.

- Due to severe degree heterogeneity, different rows of the "signal" matrix and the "noise" matrix are in very different scales. We need two normalizations: a pre-PCA normalization to mitigate the effects of degree heterogeneity on the "noise" matrix, and a post-PCA normalization (as in SCORE) on the "signal" matrix; we find that an appropriate pre-PCA normalization is Laplacian regularization.[6] See Section 3.3.2 for more explanations.

- The idea of PCA is dimension reduction: We project rows of A to the K-dimensional space spanned by the first K eigenvectors of A, $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K$, and reduce A to the $n \times K$ matrix of projection coefficients:

$$[\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_K] \equiv [\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K] \cdot \mathrm{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K).$$

Therefore, in SCORE, it is better to apply the post-PCA normalization to $[\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_K]$ instead of $[\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K]$; the two post-PCA normalization matrices (old and new) satisfy

$$\left[\frac{\hat{\eta}_2}{\hat{\eta}_1}, \frac{\hat{\eta}_2}{\hat{\eta}_1}, \ldots, \frac{\hat{\eta}_K}{\hat{\eta}_1}\right] = \left[\frac{\hat{\xi}_2}{\hat{\xi}_1}, \frac{\hat{\xi}_2}{\hat{\xi}_1}, \ldots, \frac{\hat{\xi}_K}{\hat{\xi}_1}\right] \cdot \mathrm{diag}\left(\frac{\hat{\lambda}_2}{\hat{\lambda}_1}, \frac{\hat{\lambda}_3}{\hat{\lambda}_1}, \ldots, \frac{\hat{\lambda}_K}{\hat{\lambda}_1}\right).$$

---

[6]This is analogous to the Students' t-test, where for n samples from an unknown distribution, the t-test uses a normalization for the mean and a normalization for the variance.

In effect, the new change is using eigenvalues to re-weight the columns of $\left[\frac{\hat{\xi}_2}{\hat{\xi}_1}, \frac{\hat{\xi}_2}{\hat{\xi}_1}, \ldots, \frac{\hat{\xi}_K}{\hat{\xi}_1}\right]$. See Section 3.3.3 for more explanations.

- In SCORE, we only use the first K eigenvectors for clustering, which is reasonable in the "strong signal" case, where all the nonzero eigenvalues of the "signal" matrix are much larger than the spectral norm of the "noise" matrix (in absolute value). In the "weak signal" case, some nonzero eigenvalues of the "signal" can be smaller than the spectral norm of the "noise", and we may need one or more additional eigenvectors of A for clustering. In Section 3.3.4, we have an in-depth study on the weak signal case; see details therein.

SCORE+ Input: A, K, a ridge regularization parameter $\delta > 0$ and a threshold $t > 0$. Output: class labels for all n nodes.

- *(Pre-PCA normalization with Laplacian).* Let $D = \text{diag}(d_1, d_2, \ldots, d_n)$ where d_i is the degree of node i. Obtain the graph Laplacian with ridge regularization by

$$L_\delta = (D + \delta \cdot d_{max} \cdot I_n)^{-1/2} A (D + \delta \cdot d_{max} \cdot I_n)^{-1/2}, \qquad \text{where } d_{max} = \max_{1 \leq i \leq n}\{d_i\}.$$

Note that the ratio between the largest diagonal entry of $D + \delta d_{max} I_n$ and the smallest one is smaller than $(1 + \delta)/\delta$. Conventional choices of δ are 0.05 and 0.10.

- *(PCA, where we retain possibly an additional eigenvector).* We assess the aforementioned "signal weakness" by $1 - [\hat{\lambda}_{K+1}/\hat{\lambda}_K]$, and include an additional eigenvector for clustering if and only if

$$(1 - [\hat{\lambda}_{K+1}/\hat{\lambda}_K]) \leq t, \qquad \text{(a conventional choice of } t \text{ is 0.10).}$$

- *(Post-PCA normalization).* Let M be the number of eigenvectors we decide in the last step (therefore, either $M = K$ or $M = K + 1$). Obtain the matrix of entry-wise eigen-ratios by

$$\hat{R} = \left[\frac{\hat{\eta}_2}{\hat{\eta}_1}, \frac{\hat{\eta}_3}{\hat{\eta}_1}, \ldots, \frac{\hat{\eta}_M}{\hat{\eta}_1}\right], \qquad \text{where } \hat{\eta}_k = \hat{\lambda}_k \hat{\xi}_k, 1 \leq k \leq M. \qquad (3.12)$$

- *(Clustering).* Apply classical k-means to the rows of $\hat{R}$, assuming $\leq K$ clusters.

The code is available at http://zke.fas.harvard.edu/software.html.

Compared to SCORE, SCORE+ (a) adds a pre-PCA normalization step, (b) may select one more eigenvectors for later use if necessary, and (c) uses eigenvalues to re-weight the columns of $\left[\frac{\hat{\xi}_2}{\hat{\xi}_1}, \frac{\hat{\xi}_3}{\hat{\xi}_1}, \ldots, \frac{\hat{\xi}_K}{\hat{\xi}_1}\right]$. In Section 3.3, we further explain the rationale underlying these refinements.

3.2. Numerical Comparisons We compare SCORE+ with a few recent methods: Orthodox SCORE, the convexified modularity maximization (CMM) method by Chen et al. (2018), the latent space model based (LSCD) method by Ma et al. (2020), the normalized spectral clustering (OCCAM) method for potentially overlapping communities by Zhang et al. (2020), and the regularized spectral clustering (RSC) method by Qin and Rohe (2013). For each method, we measure the clustering error rate by

$$\min_{\{\tau:\text{ permutation over }\{1,2,\ldots,K\}\}} \frac{1}{n} \sum_{i=1}^{n} 1\{\tau(\hat{\ell}_i) \neq \ell_i\},$$

where ℓ_i and $\hat{\ell}_i$ are the true and estimated labels of node i.

The error rates are in Table 2, where for SCORE+, we take $(t, \delta) = (0.1, 0.1)$. For the three relatively large networks (Weblog, Simmons, Caltech), the error rates of SCORE+ are the best among all methods, and for the other networks, the error rates are close to the best. Especially, SCORE+ provides a commendable improvement for the Simmons and Caltech data sets. In Section 3.3.4, we show that the Simmons and Caltech data sets are "weak signal" networks, and all other networks are "strong signal" networks.

RSC (Qin and Rohe, 2013) is an interesting method that applies the idea of SCORE to the graph Laplacian. It can be viewed as adding a pre-PCA

Table 2: Error rates on the 8 datasets listed in Table 1. For SCORE+, we set $(t, \delta) = (0.1, 0.1)$

Dataset	SCORE	SCORE+	CMM	LSCD	OCCAM	RSC
Weblogs	58/1222	**51/1222**	62/1222	58/1222	65/1222	64/1222
Simmons	268/1137	**127/1137**	137/1137	134/1137	266/1137	244/1137
Caltech	183/590	**98/590**	124/590	106/590	189/590	170/590
Football	5/110	6/110	7/110	21/110	**4/110**	5/110
Karate	**0/34**	1/34	**0/34**	1/34	**0/34**	**0/34**
Dolphins	**0/62**	2/62	1/62	2/62	1/62	1/62
Polbooks	**1/92**	2/92	**1/92**	3/92	3/92	3/92
UKfaculty	2/79	2/79	7/79	1/79	5/79	**0/79**

The method that attains the best performance are in bold

normalization step to SCORE (but it does not include other refinements as in SCORE+). For three of the data sets (Simmons, Caltech, UKfaculty), the modification provides a small improvement, and for three of the data sets (Weblogs, Dolphins, Polbooks), the modification hurts a little bit. The performance of OCCAM is more or less similar to that of SCORE and RSC, which is not surprising, because OCCAM is also a normalized spectral method.

The error rates of CMM and LSCD are comparable with that of SCORE+ in most data sets, except that CMM and LSCD have unsatisfactory results for UKfaculty and Football, respectively. For the three small data sets (Karate, Dolphins, Polbooks), the three methods have similar error rates, with CMM being slightly better. For the three large data sets (Weblogs, Simmons, Caltech), SCORE+ is better than LSCD, and LSCD is better than CMM.

LSCD is an iterative algorithm which solves a non-convex optimization with rank constraint. Since the algorithm only provides a local optimum, the difference between this local optimum and the global optimum may be large, especially for large K. This partially explains why LSCD performs unsatisfactorily on Football, for which data set $K = 11$. CMM first solves a convexified modularity maximization problem to get an $n \times n$ matrix $\hat{Y}$ and then applies k-median to rows of $\hat{Y}$. The matrix $\hat{Y}$ targets on approximating a rank-K matrix, but for UKfaculty, the output $\hat{Y}$ has a large $(K + 1)$th eigenvalue. This partially explains why CMM performs unsatisfactorily on this data set.

SCORE+ has two tuning parameters (t, δ), but each of which is easy to set, guided by common sense. Moreover, SCORE+ is relatively insensitive to the ridge regularization parameter δ: in Table 3, we investigate SCORE+ by setting $t = 0.10$ and letting δ range from 0.025 to 0.2 with an increment of 0.025. The results suggest SCORE+ is relatively insensitive to different choices of δ. In Section 3.3.4, we discuss further how to set the tuning parameter t.

Computationally, SCORE and OCCAM are the fastest, SCORE+ and RSC are slightly slower (the extra computing time is mostly due to the pre-PCA step), and CMM and LSCD are significantly slower, especially for large networks. For comparison of computing time, it makes more sense to use networks larger than those in Table 1. We simulate networks from the DCBM model in Section 1.1. In a DCBM with n nodes and K communities, the upper triangle of A contains independent Bernoulli variables, with

$$\mathbb{E}[A] = \Omega - \text{diag}(\Omega), \qquad \text{and} \qquad \Omega = \Theta\Pi P\Pi'\Theta,$$

where P is a $K \times K$ symmetric nonnegative matrix, $\Theta = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_n)$ with $\theta_i > 0$ being the degree parameters, and Π is the $n \times K$ label matrix. For simulations, we let n range in $\{1000, 2000, 4000, 7000, 10000\}$, and for each fixed n,

- for $c_n = 3\log(n)/n$ and $(\alpha, \beta) = (5, 4/5)$, generate θ_i such that (θ_i/c_n) are *iid* from $\mathrm{Pareto}(\alpha, \beta)$;

- fix $K = 4$ and let Π be the matrix where the first, second, third, and last quarter of rows equal to e_1, e_2, e_3, e_4, respectively;

- consider two experiments, where respectively, the P matrix is

$$\begin{bmatrix} 1 & 1/3 & 1/3 & 1/3 \\ 1/3 & 1 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1 & 1/3 \\ 1/3 & 1/3 & 1/3 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 2/3 & .1 & .1 \\ 2/3 & 1 & .5 & .5 \\ .1 & .5 & 1 & .5 \\ .1 & .5 & .5 & 1 \end{bmatrix};$$

  the value of $|\lambda_K(P)|/\lambda_1(P)$ is 0.333 for the left and 0.083 for the right, so that they represent the "strong signal" case and "weak signal" case, respectively.

The error rates and computing time are reported in Table 4 (both error rates and computing time are the average of 10 independent repetitions).

In summary, SCORE+ compares favorably over other methods both in error rates and in computing times, either for networks with "strong signals" or "weak signals".

3.3. *Rationale Underlying the Key Components of SCORE+* SCORE+ contains 4 components: the post-PCA normalization that was originally proposed in SCORE, and 3 proposed refinements (pre-PCA normalization using the Laplacian regularization, reweighing the leading eigenvectors by

Table 3: Community detection errors of SCORE+ for different δ (t is fixed at 0.10)

δ	Polblogs	Karate	Dolphins	Football	Polbooks	UKfaculty	Simmons	Caltech
0.025	57	1	0	6	2	1	127	99
0.05	54	1	1	6	2	2	**117**	100
0.075	**51**	1	1	6	2	2	121	99
0.10	**51**	1	2	6	2	2	127	**98**
0.125	53	1	3	6	2	2	134	101
0.15	54	1	3	6	2	2	137	101
0.175	56	0	3	6	2	3	141	104
0.20	58	0	3	6	2	3	142	105

The method that attains the best performance are in bold

Table 4: Comparison of error rates and computation time on simulated data. Top: Experiment 1 ("strong signal"). Bottom: Experiment 2 ("weak signal")

n	SCORE	SCORE+	CMM	LSCD	OCCAM	RSC
1000	0.08 (0s)	0.04 (3s)	**0.03** (32s)	0.05 (35s)	0.06 (0s)	0.04 (0s)
2000	0.06 (1s)	**0.03** (4s)	**0.03** (240s)	0.04 (180s)	0.07 (1s)	**0.03** (1s)
4000	0.16 (2s)	**0.02** (7s)	0.03 (1930s)	0.03 (967s)	0.14 (2s)	**0.02** (2s)
7000	0.04 (6s)	**0.01** (22s)	0.02 (10500s)	0.02(2900s)	0.05 (6s)	0.02 (13s)
10000	0.03 (9s)	**0.01** (39s)	**0.01** (31000s)	0.02 (6000s)	0.03 (9s)	**0.01** (21s)
1000	0.37 (1s)	**0.07** (3s)	0.10 (47s)	0.17 (40s)	0.37 (0s)	0.32 (0s)
2000	0.31 (1s)	**0.05** (4s)	0.07 (313s)	0.06 (194s)	0.31 (1s)	0.32 (1s)
4000	0.30 (1s)	**0.05** (7s)	0.06 (2130s)	0.06 (960s)	0.30 (1s)	0.30 (2s)
7000	0.26 (4s)	**0.03** (22s)	0.05 (10800s)	0.05 (2900s)	0.27 (4s)	0.28 (13s)
10000	0.27 (9s)	**0.03** (39s)	0.04 (32150s)	0.04 (6100s)	0.28 (9s)	0.29 (21s)

The method that attains the best performance are in bold

eigenvalues, and recruiting one more eigenvector for use when the eigengap is small). We now explain the rationale of each of these components.

Recall that under DCBM,

$$A = \Omega - \mathrm{diag}(\Omega) + W = \text{"main signal"} + \text{"secondary signal"} + \text{"noise"},$$

where the "main signal" matrix Ω equals to $\Theta\Pi P\Pi'\Theta$ and has a rank K. Let $\xi_1, \xi_2, \ldots, \xi_K$ be the eigenvectors of Ω associated with K largest eigenvalues in magnitude. Write $\Xi = [\xi_1, \xi_2, \ldots, \xi_K] = [\Xi_1, \Xi_2, \ldots, \Xi_n]'$.

3.3.1. Rationale Underlying the Post-PCA Normalization. The rationale underlying the post-PCA normalization was carefully explained in Jin (2015), so we keep the discussion brief here. Under DCBM, Jin (2015) observed that

$$\Xi_i = \theta_i \cdot q_i, \qquad \text{where } \{q_1, \ldots, q_n\} \text{ take only } K \text{ distinct values in } \mathbb{R}^K.$$

Without θ_i's, we can directly apply k-means to rows of Ξ_i. Now, with the degree parameters, Jin (2015) considered the family of *scaling invariant mappings (SIM)*, $\mathbb{M} : \mathbb{R}^K \to \mathbb{R}^K$, such that $\mathbb{M}(ax) = \mathbb{M}(x)$ for any $a > 0$ and $x \in \mathbb{R}^K$, and proposed the post-PCA normalization

$$\Xi_i \qquad \mapsto \qquad \mathbb{M}(\Xi_i), \qquad 1 \le i \le n.$$

The scaling-invariance property of $\mathbb{M}$ ensures $\{\mathbb{M}(\Xi_1),\ldots,\mathbb{M}(\Xi_n)\}$ take only K distinct values, so that we can apply k-means. Two examples of SIM include:

- $\mathbb{M}(x) = (x_2/x_1, x_3/x_1,\ldots, x_n/x_1)'$, i.e., normalizing Ξ_i by its first entry;
- $\mathbb{M}(x) = \|x\|_p^{-1} x$, i.e., normalizing Ξ_i by its L_p-norm.

The first one was recommended by Jin (2015) and is commonly referred to as SCORE. The second one is a variant of SCORE and was proposed in the supplement of Jin (2015).

In the more general DCMM model with mixed membership, Jin et al. (2017) discovered that the post-SCORE matrix is associated with a low-dimensional simplex geometry and developed SCORE into a simplex-vertex-hunting method for mixed-membership estimation. Interestingly, although each normalization in the scaling invariant family proposed by Jin (2015) works for DCBM, only the SCORE normalization produces the desired simplex geometry under DCMM.

3.3.2.   Why the Laplacian is the Right Pre-PCA Normalization. The target of SCORE is to remove the effect of degree heterogeneity in the "main signal" matrix Ω. However, the "noise" matrix $W = A - \mathbb{E}[A]$ is also affected by degree heterogeneity and requires a proper normalization. We note that, since PCA only retains a few leading eigenvectors which are driven by "signal," the "noise" is largely removed after conducing PCA. Therefore, one has to use a pre-PCA operation to normalize the "noise" matrix.

Our idea is to re-weight the rows and columns of A by node degrees. Let D be the diagonal matrix where $D(i,i)$ is the degree of node i. There are many ways for pre-PCA normalization, and simple choices include

- $A \mapsto D^{-1/2} A D^{-1/2}$.
- $A \mapsto D^{-1} A D^{-1}$.

Which one is the right choice?

Given an arbitrary positive diagonal matrix H, write

$$H^{-1}AH^{-1} = \underbrace{H^{-1}\Omega H^{-1}}_{\text{``signal''}} + \underbrace{H^{-1}[A - \mathbb{E}[A] - \text{diag}(\Omega)]H^{-1}}_{\text{``noise''}}.$$

The best pre-PCA normalization is such that, despite severe degree heterogeneity, the variances of all entries of the "noise" matrix are at the same order (Jin and Ke, 2018). Under DCBM, by direct calculations,

variance of (i, j)-entry of "noise" $\asymp \dfrac{\theta_i \theta_j}{h_i^2 h_j^2}$ $\implies$ we hope $h_i \propto \sqrt{\theta_i}$.

At the same time, the node degrees satisfy

$$d_i \qquad \propto \qquad \theta_i, \qquad \text{approximately.}$$

Therefore, the right choice is $h_i \propto \sqrt{d_i}$, i.e., we should use the pre-PCA normalization of $A \mapsto D^{-1/2} A D^{-1/2}$. See Mihail and Papadimitriou (2002) for a similar finding. For better practical performance, we add a ridge regularization.

Besides normalizing the "noise" matrix, the pre-PCA normalization also changes the "signal" matrix from Ω to $D^{-1/2} \Omega D^{-1/2}$. Fortunately, the new "signal" matrix has a similar form as $\Omega = \Theta \Pi P \Pi' \Theta$, except that Θ is replaced by $D^{-1/2} \Theta$, so the post-PCA normalization of SCORE is still valid.

3.3.3. Why $\hat{\eta}_k$ is the Appropriate Choice in Post-PCA Normalization. In the post-PCA normalization, SCORE+ constructs the matrix of entry-wise eigen-ratios using $\hat{\eta}_1, \ldots, \hat{\eta}_K$, where each $\hat{\eta}_k$ is $\hat{\xi}_k$ weighted by the corresponding eigenvalue. There are many ways of weighting the eigenvectors, and simple choices include

- $[\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K] \cdot \text{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K)$.

- $[\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K] \cdot \text{diag}\left(\sqrt{\hat{\lambda}_1}, \sqrt{\hat{\lambda}_2}, \ldots, \sqrt{\hat{\lambda}_K} \right)$.

Why do we choose the first one?

We briefly explained it in Section 3.1 using the perspective of projecting rows of data matrix to the span of $\hat{\xi}_1, \ldots, \hat{\xi}_K$. We now take a different perspective. Recall that L_δ is the regularized graph Laplacian, by Abbe et al. (2019), the first-order approximations of eigenvectors are

$$\hat{\xi}_k \approx \frac{1}{\lambda_k} L_\delta \xi_k \approx \xi_k + \frac{1}{\lambda_k} (L_\delta - \mathbb{E}[L_\delta]) \xi_k.$$

Intuitively speaking, since each ξ_k has a unit-norm, the "noise" vector $(L_\delta - \mathbb{E}[L_\delta]) \xi_k$ is at the same scale for different k; it implies that the noise level in different eigenvectors is proportional to $1/\lambda_k$. This means $\hat{\xi}_1$ is less noisy than $\hat{\xi}_2$, and $\hat{\xi}_2$ is less noisy than $\hat{\xi}_3$, and so on. By weighing the eigenvectors by $\hat{\lambda}_k$, the noise level in $\hat{\eta}_1, \ldots, \hat{\eta}_K$ is approximately at the same order.

In most theoretical studies, $\lambda_1, \ldots, \lambda_K$ are assumed at the same order, so whether or not to re-weight the eigenvectors does not affect the rate of convergence. However, in many real data, the magnitudes of the first a few eigenvalues can be considerably different, so such a weighting scheme does improve the numerical performance.

3.3.4. When We Should Choose One More Eigenvector for Inference.
In SCORE+, we retain M eigenvectors in the PCA step for later uses, where

$$M = \begin{cases} K, & 1 - (\hat{\lambda}_{K+1}/\hat{\lambda}_K) > t, \\ K + 1, & \text{otherwise.} \end{cases}$$

For the 8 data sets in Table 1, if we choose $t = 0.1$ as suggested, then
$M = K + 1$ for the Simmons and Caltech data sets, and $M = K$ for all
others. The insight is that, if a data set fits with the "strong signal" profile,
then we use exactly K eigenvectors for clustering, but if it fits with the
"weak signal" profile, we may need to use more than K eigenvectors. Our
analysis below shows that the Simmons and Caltech data sets fit with the
"weak signal" profile, while all other data sets fit with the "strong signal"
profile.

We illustrate our points with the scree plot and the Rayleigh quotient.
Let $\ell \in \mathbb{R}^n$ be the true community label vector, and let

$$S_k = \{1 \le i \le n : \ell_i = k\}, \qquad 1 \le k \le K.$$

For any vector $x \in \mathbb{R}^n$, define *normalized Rayleigh quotient* (Fisher, 1936):

$$Q(x) = 1 - \frac{\text{Within-Class-Variance}}{\text{Total Variance}} = \frac{\text{Between-Class-Variance}}{\text{Total Variance}},$$

where Total Variance, Within-Class Variance, and Between-Class-variance
are $\sum_{i=1}^n (x_i - \bar{x})^2$, $\sum_{k=1}^K \sum_{i \in S_k} (x_i - \bar{x}_k)^2$, and $\sum_{k=1}^K (|S_k| \cdot (\bar{x}_k - \bar{x})^2)$, respec-
tively ($\bar{x}$ is the overall mean of x_i and $\bar{x}_k$ is the mean of x_i over all $i \in S_k$).
Rayleigh quotient is a well-known measure for the clustering utility of x.
Note that $0 \le Q(x) \le 1$ for all x, $Q(x) = 1$ when $x = \ell$, and $Q(x) \approx 0$ when
x is a randomly generated vector.

Fix $\delta = 0.1$. Let $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{K+1}$ be the $(K + 1)$ eigenvalues of L_δ with
largest magnitude and let $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{K+1}$ be the corresponding eigenvectors.
Below are some features that help differentiate a "strong signal" setting from
a "weak signal" setting.

- In the scree plot, we expect to see a relatively large gap between $\hat{\lambda}_K$
  and $\hat{\lambda}_{K+1}$ when the "signal" is strong, and a relatively small gap if the
  "signal" is relatively weak.

- In a "strong signal" setting, we expect to see that the Rayleigh quotient
  $Q(\hat{\xi}_k)$ is relatively large for $k = K$, but is relatively small for $k = K + 1, K + 2$, etc. In a "weak signal" setting, we may observe that a
  relatively large Rayleigh quotient $Q(\hat{\xi}_k)$ for $k = K + 1, K + 2$, etc., and
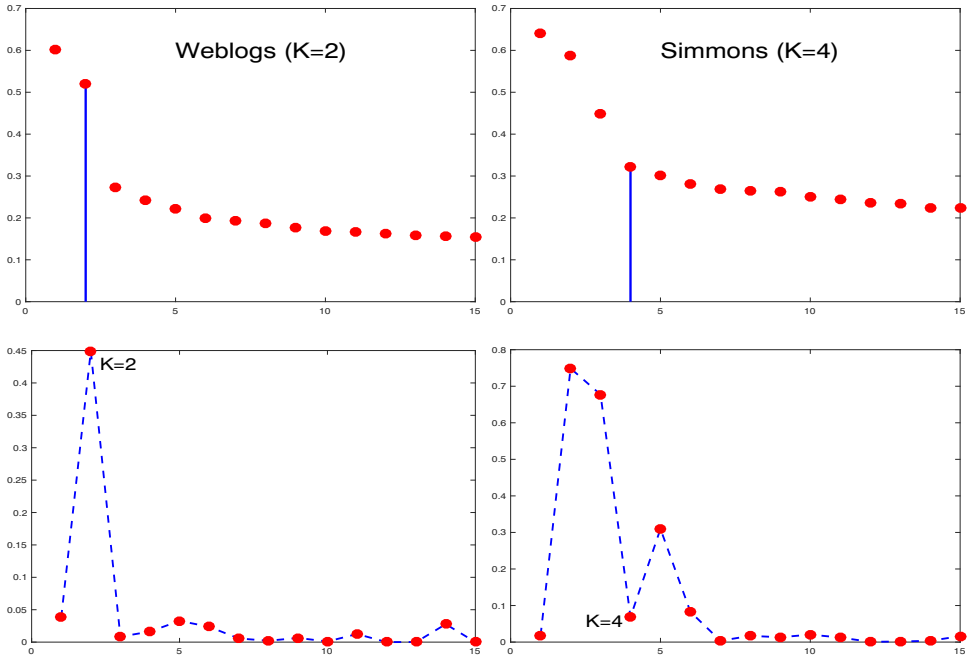 $Q(\hat{\xi}_K)$ can be relatively small.

Figure 1: A typical "strong signal" dataset (Weblogs, left panels) and a typical "weak signal" dataset (Caltech, right panels). The top two figures display the absolute eigenvalues. We observe there is a relatively large gap between $|\hat{\lambda}_K|$ and $|\hat{\lambda}_{K+1}|$ in a "strong signal" profile and a relatively small gap in a "weak signal" profile. The bottom two plots display the Rayleigh quotients $Q(\hat{\xi}_k)$. We observe that $Q(\hat{\xi}_k)$ for $k = K + 1, K + 2, \ldots$ are all small in a "strong signal" profile but some of them are large in a "weak signal profile

The points are illustrated in Fig. 1 with the Weblog data and Simmons data, which are believed to be a typical "strong signal" dataset and a typical "weak signal" dataset, respectively. We note that the first eigenvector consists of global information of L_δ and it alone does not have much utility for clustering. Therefore, the corresponding Rayleigh quotient $Q(\hat{\xi}_1)$ is usually small. In SCORE (e.g., Eq. 3.12), we use $\hat{\xi}_1$ for normalization, but not directly for clustering.

Table 5 shows the Rayleigh quotients of all 8 datasets. We found that the $(K + 1)$th eigenvector contains almost no information of community labels, except for Caltech and Simmons. This agrees with our findings that Caltech and Simmons fit with the "weak signal" profile.

Table 5: Rayleigh quotient $Q(\hat{\xi}_k)$ for 8 networks. The first four rows are for eigenvectors of the adjacency matrix, and the last four rows are for eigenvectors of the regularized graph Laplacian. Except for Simons and Caltech, the $(K+1)$th eigenvector of all other datasets contains almost no information

Dataset	Polblogs	Karate	Dolphins	Football	Polbooks	UKfaculty	Simmons	Caltech
Eigen(K)	**0.36**	**0.76**	**0.60**	**0.45**	**0.63**	**0.80**	0.04	0.25
Eigen(K+1)	0.02	0.07	0.00	0.01	0.01	0.11	**0.20**	**0.47**
Eigen(K+2)	0.06	0.05	0.01	0.00	0.01	0.00	0.13	0.06
Eigen(K+3)	0.01	0.01	0.01	0.01	0.01	0.00	0.15	0.11
Eigen(K)	**0.45**	**0.81**	**0.79**	**0.48**	**0.79**	**0.89**	0.07	0.32
Eigen(K+1)	0.02	0.02	0.00	0.22	0.01	0.06	**0.31**	**0.54**
Eigen(K+2)	0.03	0.01	0.00	0.00	0.00	0.00	0.08	0.03
Eigen(K+3)	0.01	0.01	0.02	0.02	0.00	0.00	0.00	0.09

The eigenvector associated with the largest $RQ(\xi_k)$ are in bold

How to choose between $K = M$ and $K = M + 1$? The scree plot could potentially be a good way to estimate how much information is contained in each eigenvector. If the Kth and $(K+1)$th eigenvalues are close, it is likely that the $(K+1)$th eigenvector also contains information. To measure "closeness", we propose to use the quantity $1 - [\hat{\lambda}_{K+1}/\hat{\lambda}_K]$ with a scale-free tuning parameter $t = 0.1$. This seems to work well on all the 8 datasets. See Table 6.

4. Proofs

We now prove Theorems 2.1, Corollary 2.1, and Theorem 2.2.

Table 6: The quantity $1 - [\hat{\lambda}_{K+1}/\hat{\lambda}_K]$ for 8 network data sets, where $\hat{\lambda}_k$ are from the adjacency matrix (left) and the regularized graph Laplacian (right). With a threshold $t = 0.1$, this criterion successfully selects $M = K + 1$ for Simmons and Caltech and $M = K$ for all others

Dataset	Adjacency matrix	Regularized graph Laplacian
Polblogs	0.5997	0.5223
Karate	0.4140	0.1768
Dolphins	0.1863	0.2027
Football	1.9255	0.1414
Polbooks	0.5034	0.2246
UKfaculty	0.3139	0.3336
Simmons	**0.0804**	**0.0533**
Caltech	**0.0777**	**0.0236**

Values below the threshold 0.1 are in bold

4.1. Analysis of Empirical Eigenvectors Recall that λ_k and $\hat{\lambda}_k$ denote the kth largest eigenvalue (in magnitude) of Ω and A, respectively, and ξ_k and $\hat{\xi}_k$ denote the respective eigenvectors. Define

$$\beta_n = |\lambda_K(G^{1/2}PG^{1/2})|, \quad \text{where} \quad G = K\|\theta\|^{-2} \cdot \text{diag}\left(\|\theta^{(1)}\|^2, \ldots, \|\theta^{(K)}\|^2\right).$$

Let $\|M\|_{2\to\infty}$ denote the maximum row-wise ℓ^2-norm of a matrix M. The key technical tool we need in the proof is the following lemma:

THEOREM 4.1. *Under conditions of Theorem 2.1, write* $\hat{\Xi}_0 = [\hat{\xi}_2, \hat{\xi}_3, \ldots, \hat{\xi}_K]$, $\Xi_0 = [\xi_2, \xi_3, \ldots, \xi_K]$, *and* $\Lambda_0 = \text{diag}(\lambda_2, \lambda_3, \ldots, \lambda_K)$. *With probability* $1 - o(n^{-3})$, *there exists an orthogonal matrix* $O \in \mathbb{R}^{K-1,K-1}$ *(which depends on A and is stochastic) such that*

$$\|\hat{\Xi}_0 O - A\Xi_0\Lambda_0^{-1}\|_{2\to\infty} \leq \frac{C\theta_{\max}K^5\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\beta_n\|\theta\|^3}.$$

Theorem 4.1 is an extension of equation (C.71) in Jin et al. (2017) (this equation appears in the proof of Lemma 2.1 of Jin et al. (2017)). Lemma 2.1 of Jin et al. (2017) assumes that $|\lambda_2|, |\lambda_3|, \ldots, |\lambda_K|$ are at the same order, but here we allow them to be at different orders. The proof also needs some modification.

Remark In the bound in Theorem 4.1, the power of K can be further reduced by adding mild regularity conditions on $|\lambda_2|, |\lambda_3|, \ldots, |\lambda_K|$. For example, if we assume $|\lambda_2|, |\lambda_3|, \ldots, |\lambda_K|$ can be grouped into $s = O(1)$ groups such that $\kappa(I) \leq C$ for each group (see the statement of Lemma 4.1 for the definition of $\kappa(I)$), then the power of K can be reduced from K^5 to $K\sqrt{K}$. In fact, the setting in Jin et al. (2017) corresponds to a special case of $s = 2$.

PROOF OF THEOREM 4.1. Re-arrange the $(K-1)$ eigenvalues in the descending order, i.e., $\lambda_{(2)} \geq \lambda_{(3)} \geq \ldots \geq \lambda_{(K)}$. We first assume that all these eigenvalues are positive and use the following procedure to divide them into groups:

- Initialize: $k = 1$ and $m = 2$.

- Compute the eigen-gaps g_s, where $g_s = \lambda_{(s)} - \lambda_{(s+1)}$, for $m \leq s \leq K-1$, and $g_K = \lambda_{(K)}$. Let $s^* = \min\{m \leq s \leq K : g_s \geq K^{-1}\lambda_{(m)}\}$. Since $\lambda_{(m)} = \sum_{s=m}^{K} g_s$, such s^* must exist.

- Group $\lambda_{(m)}, \lambda_{(m+1)}, \ldots, \lambda_{(s^*-1)}, \lambda_{(s^*)}$ together as the kth group.

- If $s^* = K$, terminate; otherwise, increase k by 1, reset $m = s^* + 1$, and repeat the above steps to obtain the next group.

For each group k, let I_k be the corresponding index set in the original order, i.e., group k consists of eigenvalues λ_j for all $j \in I_k$. Define the eigengap associated with group k as

$$\delta(I_k) = \min \left\{ \min_{\substack{1 \leq i,\, j \leq K \\ i \in I_k,\; j \notin I_k}} |\lambda_i - \lambda_j|, \quad \min_{i \in I_k} |\lambda_i| \right\}. \tag{4.13}$$

The above grouping procedure, as well as the first inequality of condition Eq. 2.10, guarantees that

$$\max_{j \in I_k} |\lambda_j| \leq K \cdot \delta(I_k), \qquad \text{for each group } k. \tag{4.14}$$

When some of the $(K-1)$ eigenvalues are negative, we first partition these eigenvalues into two subsets, one consists of positive eigenvalues, and the other consists of negative ones. We directly apply the grouping procedure in the first subset. In the second subset, we take absolute values, sort in the descending order, and then apply the above grouping procedure. Finally, we combine the two collections of groups. The resulting groups still satisfy Eq. 4.14.

We then prove the following technical lemma, which extends Theorem 2.1 of Abbe et al. (2019) and Lemma C.3 of Jin et al. (2017).

Lemma 4.1. *Let $M \in \mathbb{R}^{n,n}$ be a symmetric random matrix, where $\mathbb{E}M = M^*$ for a rank K_0 matrix M^*. Let d_k^* and d_k be the kth largest nonzero eigenvalue of M^* and M, respectively, and let η_k^* and η_k be the corresponding eigenvector, respectively, $1 \leq k \leq K_0$. Consider a partition*

$$\{1, 2, \ldots, K_0\} = I \cup \left(\cup_{k=1}^N I_k \right), \qquad \text{where} \quad I = \{s+1, s+2, \ldots, s+r\},$$

with s and r being two integers such that $1 \leq r \leq K_0$ and $0 \leq s \leq K_0 - r$, and where each of $I_1, I_2, \ldots, I_N$ contains consecutive indices. For any index subset $B \subset \{1, 2, \ldots, K_0\}$, define

$$\delta(B) = \min \left\{ \min_{i \in B, j \notin B} \{|d_i^* - d_j^*|\}, \; \min_{i \in B} |d_i^*| \right\}, \qquad \text{and} \qquad \kappa(B) = \left(\max_{i \in B} \{|d_i^*|\} \right) / \delta(B).$$

Let $D = \mathrm{diag}(d_{s+1}, \ldots, d_{s+r})$, $D^* = \mathrm{diag}(d_{s+1}^*, \ldots, d_{s+r}^*)$,

$$U = [\eta_{s+1}, \eta_{s+2}, \ldots, \eta_{s+r}], \qquad \text{and} \qquad U^* = [\eta_{s+1}^*, \eta_{s+2}^*, \ldots, \eta_{s+r}^*].$$

Let $M^*_{m,\cdot}$ denote the m-th row of M^*, for $1 \leq m \leq n$. Suppose for a number $\gamma > 0$, the following assumptions are satisfied:

- *A1 (Incoherence): $\max_{1 \leq m \leq n} \|M^*_{m,\cdot}\| \leq \gamma \Delta^*$, where $\Delta^* = \delta(I)$.*

- *A2 (Independence): For any $1 \leq m \leq n$, the entries of the m-th row and column of M are independent with the other entries.*

- *A3 (Spectral norm concentration): For a number $\delta_0 \in (0,1)$, $\mathbb{P}(\|M - M^*\| \leq \gamma \Delta^*) \geq 1 - \delta_0$.*

- *A4 (Row concentration): There is a number $\delta_1 \in (0,1)$ and a continuous non-decreasing function $\varphi(\cdot)$ with $\varphi(0) = 0$ and $\varphi(x)/x$ being non-increasing in $\mathbb{R}^+$ such that, for any $1 \leq m \leq n$ and non-stochastic matrix $Y \in \mathbb{R}^{n,r}$,*

$$\mathbb{P}\left(\|(M - M^*)_{m,\cdot}Y\|_2 \leq \Delta^* \|Y\|_{2\to\infty} \varphi\left(\frac{\|Y\|_F}{\sqrt{n}\|Y\|_{2\to\infty}}\right)\right) \geq 1 - \delta_1/n.$$

With probability $1 - \delta_0 - 2\delta_1$, for an orthogonal matrix $O \in \mathbb{R}^{r,r}$,

$$\|UO - MU^*(D^*)^{-1}\|_{2\to\infty}$$
$$\leq C\left[\kappa(\kappa + \varphi(1))(\gamma + \varphi(\gamma)) + \widetilde{\kappa}\gamma\right] \cdot \|\widetilde{U}^*\|_{2\to\infty}, \tag{4.15}$$

where $\widetilde{U}^* = [\eta_1, \eta_2, \ldots, \eta_{K_0}]$ and $\widetilde{\kappa} = \sum_{1 \leq k \leq N} \kappa(I_k)$.

We now prove Lemma 4.1. The proof is a light modification of the proof of Lemma C.3 of Jin et al. (2017). Fix $1 \leq m \leq n$. Let $M^{(m)}$ be the matrix by setting the mth row and the mth column of M to be zero. Let $\eta_1^{(m)}, \eta_2^{(m)}, \ldots, \eta_n^{(m)}$ be the eigenvectors of $M^{(m)}$. Write $U^{(m)} = [\eta_{s+1}^{(m)}, \ldots, \eta_{s+r}^{(m)}]$. Let $H = U'U^*$, $H^{(m)} = (U^{(m)})'U^*$ and $V^{(m)} = U^{(m)}H^{(m)} - U^*$. We aim to prove

$$\|M_{m\cdot}V^{(m)}\| \leq 6(\kappa + \widetilde{\kappa})\gamma\Delta^*\|\widetilde{U}^*\|_{2\to\infty}$$
$$+ \Delta^*\varphi(\gamma)\left(4\kappa\|UH\|_{2\to\infty} + 6\|U^*\|_{2\to\infty}\right). \tag{4.16}$$

Once Eq. 4.16 is obtained, the proof is almost identical to the proof of (B.26) in Abbe et al. (2019), except that we plug in Eq. 4.16 instead of (B.32) in Abbe et al. (2019). This is straightforward, so we omit it. What remains is

to prove Eq. 4.16. In the proof of (Abbe et al. 2019, Lemma 5), it is shown that

$$
\begin{aligned}
\|M_{m\cdot}V^{(m)}\| &\leq \|M_m^*V^{(m)}\| + \|(M-M^*)_{m\cdot}V^{(m)}\|, \\
\|(M-M^*)_{m\cdot}V^{(m)}\| &\leq \Delta^*\varphi(\gamma)\left(4\kappa\|UH\|_{2\to\infty} + 6\|U^*\|_{2\to\infty}\right).
\end{aligned}
$$

Combining them gives

$$
\|M_{m\cdot}V^{(m)}\| \leq \|M_{m\cdot}^*V^{(m)}\| + \Delta^*\varphi(\gamma)\left(4\kappa\|UH\|_{2\to\infty} + 6\|U^*\|_{2\to\infty}\right). \quad (4.17)
$$

We further bound the first term in Eq. 4.17. Define

$$
I_0 = \cup_{\{1\leq k\leq N:\, \max_{j\in I_k}|d_j^*|>\max_{j\in I}|d_j^*|\}} I_k.
$$

In other words, I_0 is the union of groups of eigenvalues such that the largest absolute eigenvalue in that group is larger than $\|D^*\|$. Let $\widetilde{M}^* = \sum_{j\in I_0} d_j^*\eta_j^*(\eta_j^*)'$.

$$
\begin{aligned}
\|M_{m\cdot}^*V^{(m)}\| &\leq \|\widetilde{M}_{m\cdot}^*V^{(m)}\| + \|(M_{m\cdot}^* - \widetilde{M}_{m\cdot}^*)V^{(m)}\| \\
&\leq \|\widetilde{M}_{m\cdot}^*V^{(m)}\| + \|M^* - \widetilde{M}^*\|_{2\to\infty}\|V^{(m)}\| \\
&\leq \|\widetilde{M}_{m\cdot}^*V^{(m)}\| + 6\gamma\|M^* - \widetilde{M}^*\|_{2\to\infty},
\end{aligned}
$$

where the last line uses $\|V^{(m)}\| \leq 6\gamma$, by (B.12) of Abbe et al. (2019). Note that $M^* - \widetilde{M}^* = \sum_{j\notin I_0} d_j^*\eta_j^*(\eta_j^*)'$. By definition of I_0, for any $j\notin I_0$, $|d_j^*| \leq \max_{i\in I}|d_i^*| \leq \kappa\Delta^*$. It follows that

$$
\|M^* - \widetilde{M}^*\|_{2\to\infty} \leq \left(\max_{j\notin I_0}|d_j^*|\right)\|\widetilde{U}^*\|_{2\to\infty} \leq \kappa\Delta^*\|\widetilde{U}^*\|_{2\to\infty}.
$$

Combining the above gives

$$
\|M_{m\cdot}^*V^{(m)}\| \leq \|\widetilde{M}_{m\cdot}^*V^{(m)}\| + 6\kappa\gamma\Delta^*\|\widetilde{U}^*\|_{2\to\infty}. \quad (4.18)
$$

Without loss of generality, we assume all groups except for I are contained in I_0, i.e., $I_0 = \cup_{k=1}^N I_k$. Let $D_k^* = \text{diag}(d_j^*)_{j\in I_k}$, $U_k^* = [\eta_j^*]_{j\in I_k}$, $U_k = [\eta_j]_{j\in I_k}$, $U_k^{(m)} = [\eta_j^{(m)}]_{j\in I_k}$, and $H_k^{(m)} = (U_k^{(m)})'U_k^*$. Then,

$$
\widetilde{M}^* = \sum_{k=1}^N U_k^*\Lambda_k^*(U_k^*)'.
$$

Similar to (B.12) of Abbe et al. (2019), we have $\|U_k^{(m)}H_k^{(m)} - U_k^*\| \leq 6\gamma_k$, where γ_k is defined in the same way as γ but is with respect to the eigen-gap

of group k, which is $\Delta_k^* \equiv \delta(I_k)$. It is not hard to see that $\gamma_k = \gamma\Delta^*/\Delta_k^*$. Therefore,

$$\|U_k^{(m)}H_k^{(m)} - U_k^*\| \le 6\gamma\Delta^*/\Delta_k^*, \qquad 1 \le k \le N. \tag{4.19}$$

By mutual orthogonality of eigenvectors, $(U_k^{(m)})'U^{(m)} = 0$, and $(U_k^*)'U^* = 0$. Additionally, we have $\|U_k^{(m)}\| = 1$ and $\|H_k^{(m)}\| \le 1$. It follows that

$$
\begin{aligned}
\|\widetilde{M}_{m\cdot}^* V^{(m)}\| &\le \sum_{k=1}^N \|e_m'[U_k^*\Lambda_k^*(U_k^*)'][U^{(m)}H^{(m)} - U^*]\| \\
&= \sum_{k=1}^N \left\|e_m'[U_k^*\Lambda_k^*(U_k^*)']U^{(m)}H^{(m)}\right\| \quad \text{(by mutual orthogonality)} \\
&\le \sum_{k=1}^N \left\|e_m'[U_k^*\Lambda_k^*(U_k^*)']U^{(m)}\right\| \\
&= \sum_{k=1}^N \left\|e_m'U_k^*\Lambda_k^*(U_k^* - U_k^{(m)}H_k^{(m)})'U^{(m)}\right\| \quad \text{(by mutual orthogonality)} \\
&\le \sum_{k=1}^N \left\|e_m'U_k^*\Lambda_k^*(U_k^* - U_k^{(m)}H_k^{(m)})'\right\| \\
&\le \sum_{k=1}^N \|U_k^*\|_{2\to\infty} \cdot \|\Lambda_k^*\| \cdot \|U_k^* - U_k^{(m)}H_k^{(m)}\| \\
&\le \sum_{k=1}^N 6(\|\Lambda_k\|^*/\Delta_k^*) \cdot \gamma\Delta^*\|U_k^*\|_{2\to\infty} \quad \text{(by Eq. 4.19)} \\
&\le 6\gamma\Delta^* \cdot \left(\sum_{k=1}^N \kappa(I_k)\right) \cdot \|\widetilde{U}^*\|_{2\to\infty} \quad \text{(note that} \|U_k^*\|_{2\to\infty} \le \|\widetilde{U}^*\|_{2\to\infty}\text{).}
\end{aligned}
$$

We plug it into Eq. 4.18 and use the definition of $\widetilde{\kappa}$. It gives

$$\|M_{m\cdot}^* V^{(m)}\| \le 6(\kappa + \widetilde{\kappa})\gamma\Delta^*\|\widetilde{U}^*\|_{2\to\infty}. \tag{4.20}$$

Combining Eq. 4.20 with Eq. 4.17 gives Eq. 4.16. Then, the proof of Lemma 4.1 is complete.

We now apply Lemma 4.1 to prove the claim. For the groups in Eq. 4.14, they satisfy that

$$\kappa(I_k) \le K, \qquad \text{and} \qquad \delta(I_k) \ge K^{-1}|\lambda_K| \ge K^{-2}\beta_n\|\theta\|^2.$$

We fix I to be one of the groups, and let $I_1, \dots, I_N$ be the remaining groups. We apply Lemma 4.1 to $M = A$, $M^* = \Omega = \text{diag}(\Omega) + (A - \mathbb{E}A)$, and

$$\Delta^* \asymp \beta_n K^{-2}\|\theta\|^2, \qquad \gamma \asymp \frac{K^2\sqrt{\theta_{\max}\|\theta\|_1}}{\beta_n\|\theta\|^2}.$$

We construct $\varphi(\gamma)$ in the same way as in Lemma C.3 of Jin et al. (2017). It satisfies that $\varphi(\gamma) \leq C\gamma\sqrt{\log(n)}$. Similarly as in the proof of Lemma C.3, we can show that conditions A1-A4 are satisfied. Write

$$\hat{\Xi}_{01} = [\hat{\xi}_i]_{i\in I}, \qquad \Xi_{01} = [\xi_i]_{i\in I}, \qquad \text{and} \quad \Lambda_1 = \text{diag}(\lambda_i)_{i\in I}.$$

It follows from Eq. 4.15 that there exists an orthogonal matrix $O \in \mathbb{R}^{|I|\times|I|}$ such that

$$\|\hat{\Xi}_{01}O - A\Xi_{01}\Lambda_1^{-1}\|_{2\to\infty} \leq C\frac{K^4\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\beta_n\|\theta\|^2}\|\Xi\|_{2\to\infty}.$$

By Lemma B.2 of Jin et al. (2017), $\|\Xi\|_{2\to\infty} = O(\sqrt{K}\|\theta\|^{-1}\theta_{\max})$. Plugging it into the above inequality, we find that

$$\|\hat{\Xi}_{01}O - A\Xi_{01}\Lambda_1^{-1}\|_{2\to\infty} \leq \frac{C\theta_{\max}K^4\sqrt{K\theta_{\max}\|\theta\|_1\log(n)}}{\beta_n\|\theta\|^3}. \qquad (4.21)$$

The above inequality holds for each group. Note that $\hat{\Xi}_0$ is obtained by putting such $\hat{\Xi}_{01}$ together. When $B = [B_1, B_2, \ldots, B_N]$, it holds that $\|B\|_{2\to\infty} \leq \sqrt{\sum_k \|B_k\|_{2\to\infty}^2} \leq \sqrt{K}\max_k \|B_k\|_{2\to\infty}$. Combining it with Eq. 4.21 gives the claim.

4.2.   Proof of Theorem 2.1 The rationale of SCORE guarantees that the rows of R take only K distinct values $v_1, v_2, \ldots, v_K \in \mathbb{R}^{K-1}$. Below, we first derive a crude high-probability bound for $\text{Hamm}(\hat{\Pi}, \Pi)$ without using Theorem 4.1. This bound implies that each k-means center is close to one of the true v_k. Next, we use Theorem 4.1 to derive a sharper bound for $\mathbb{E}[\text{Hamm}(\hat{\Pi}, \Pi)]$.

We start from deriving a crude bound for $\text{Hamm}(\hat{\Pi}, \Pi)$. Let β_n be the same as in Section 4.1. By Lemma B.1 of Jin et al. (2017), $C^{-1}K^{-1}\|\theta\|^2 \leq \lambda_1 \leq \|\theta\|^2$, and $|\lambda_K| \asymp K^{-1}\beta_n\|\theta\|^2$. It follows that

$$C^{-1}\sqrt{K}(|\lambda_K|/\sqrt{\lambda_1}) \leq \beta_n\|\theta\| \leq CK(|\lambda_K|/\sqrt{\lambda_1}).$$

Therefore, the assumption $s_n \geq a_1^{-1}K^4\sqrt{\log(n)}$ guarantees that

$$\frac{\theta_{\max}K^4\sqrt{K\theta_{\max}\|\theta\|_1\log(n)}}{\beta_n\theta_{\min}\|\theta\|^2} \leq a_1. \qquad (4.22)$$

Let O be the orthogonal matrix in Theorem 4.1. By Lemma 2.1 of Jin et al. (2017), with probability $1 - o(n^{-3})$, there exists $\omega \in \{\pm 1\}$ such that

$$\|\omega\hat{\xi}_1 - \xi_1\|_\infty \leq \frac{C\theta_{\max}K\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\|\theta\|^3}, \qquad \|\hat{\Xi}_0 O - \Xi_0\|_F \leq C\frac{K\sqrt{K\theta_{\max}\|\theta\|_1}}{\beta_n\|\theta\|^2}. \quad (4.23)$$

By Lemma B.2 of Jin et al. (2017), $\xi_1(i) \geq C^{-1}\theta_i/\|\theta\| \geq C^{-1}\theta_{\min}/\|\theta\|$. By choosing a_1 appropriately small, the condition on s_n guarantees that $\|\hat{\xi}_1 - \xi_1\|_\infty \leq \xi_1(i)/3$, for any $1 \leq i \leq n$. Then, we can use a proof similar to that in Lemma C.5 of Jin et al. (2017) to show that, with probability $1 - o(n^{-3})$, there exists an orthogonal matrix H such that

$$\sum_{i=1}^n \|H\hat{r}_i - r_i\|^2 \leq \frac{\|\hat{\Xi}_0 O - \Xi_0\|_F^2}{(\min_{1 \leq i \leq n} \theta_i/\|\theta\|)^2} \leq \frac{CK^3\theta_{\max}\|\theta\|_1}{\theta_{\min}^2\beta_n^2\|\theta\|^2}.$$

Since $\theta_{\max}^2 \geq \|\theta\|^2/n$, we can further write that

$$\sum_{i=1}^n \|H\hat{r}_i - r_i\|^2 \leq \frac{CnK^3\theta_{\max}^3\|\theta\|_1}{\theta_{\min}^2\beta_n^2\|\theta\|^4} \leq Cn \cdot \frac{a_1^2}{K^6\log(n)}, \qquad (4.24)$$

where the last inequality is from Eq. 4.22. Recall that the rows of R take only K distinct values $v_1, \ldots, v_K$. By Lemma B.3 of Jin et al. (2017), there exists a constant $c_0 > 0$ such that, for all $1 \leq k \neq \ell \leq K$,

$$\|v_k - v_\ell\| \geq c_0\sqrt{K} \qquad \text{and} \qquad \|v_k\| \leq C\sqrt{K}. \qquad (4.25)$$

Furthermore, in the proof of Theorem 2.2 of Jin (2015), it was shown that the k-means solution satisfies that

$$\text{Hamm}(\widehat{\Pi}, \Pi) \leq (3/\delta)^2 \sum_{i=1}^n \|H\hat{r}_i - r_i\|^2,$$

where δ is the minimum distance between two distinct rows of R. Combining the above gives

$$\text{Hamm}(\widehat{\Pi}, \Pi) \leq Ca_1^2 \cdot \frac{n}{K^7\log(n)}, \qquad (4.26)$$

where C is a constant that does not depend on a_1.

This crude bound Eq. 4.26 is enough for studying the k-means centers. By Eq. 4.26, the total number of misclustered nodes is $O(n/[K^7\log(n)])$. Also, Condition Eq. 2.9 implies that each true cluster has at least $c_2^{-1}K^{-1}n$ nodes. This means that each cluster has only a negligible fraction of misclustered nodes. Particularly, each true cluster $\mathcal{C}_k$ is associated with one and only one k-means cluster, which we denote by $\hat{\mathcal{C}}_k$; furthermore, we have $|\hat{\mathcal{C}}_k \backslash \mathcal{C}_k| = O(n/[K^7\log(n)])$ and $|\hat{\mathcal{C}}_k \backslash \mathcal{C}_k| = O(n/[K^7\log(n)])$. The cluster center $\hat{v}_k$ of the cluster $\hat{\mathcal{C}}_k$ satisfies that

$$\hat{v}_k = \frac{1}{|\hat{\mathcal{C}}_k|} \sum_{i \in \hat{\mathcal{C}}_k} \hat{r}_i.$$

Note that $r_i = v_k$ for $i \in \mathcal{C}_k$. It follows that

$$\left\| H\hat{v}_k - v_k \right\| \leq \frac{1}{|\hat{\mathcal{C}}_k|} \left\| \sum_{i \in \hat{\mathcal{C}}_k} (H\hat{r}_i - r_i) \right\| + \frac{1}{|\hat{\mathcal{C}}_k|} \left\| \sum_{i \in \hat{\mathcal{C}}_k \setminus \mathcal{C}_k} (r_i - v_k) \right\|.$$

By Eq. 4.25, $\|r_i - v_k\| = O(\sqrt{K})$. Furthermore, $|\hat{\mathcal{C}}_k| \gtrsim c_2^{-1} K^{-1} n$ and $|\hat{\mathcal{C}}_k \setminus \mathcal{C}_k| = O(n/[K^7 \log(n)])$. Combing them with the Cauchy-Schwarz inequality, we find that

$$\begin{aligned}
\|H\hat{v}_k - v_k\| &\leq \frac{1}{\sqrt{|\hat{\mathcal{C}}_k|}} \cdot \sqrt{\sum_{i \in \hat{\mathcal{C}}_k} \|H\hat{r}_i - r_i\|^2} + \frac{|\hat{\mathcal{C}}_k \setminus \mathcal{C}_k|}{c_2^{-1} K^{-1} n} \cdot O(\sqrt{K}) \\
&\leq \frac{1}{\sqrt{c_2^{-1} K^{-1} n}} \cdot O\left(\sqrt{\frac{n}{K^6 \log(n)}} \right) + O\left(\frac{1}{K^5 \sqrt{K} \log(n)} \right) \\
&= O\left(\frac{1}{K^2 \sqrt{K \log(n)}} \right).
\end{aligned}$$

The right hand side is $o(\sqrt{K})$. Let c_0 be the same as in Eq. 4.25. Then, for sufficiently large n,

$$\|H\hat{v}_k - v_k\| \leq c_0 \sqrt{K}/8, \qquad \text{for all } 1 \leq k \leq K. \qquad (4.27)$$

Next, we use Theorem 4.1 to get the desired bound for $\mathbb{E}[\text{Hamm}(\widehat{\Pi}, \Pi)]$. Let D be the event that Eq. 4.27 holds. We have shown that $\mathbb{P}(D^c) = o(n^{-3})$. It follows that

$$\mathbb{E}[\text{Hamm}(\widehat{\Pi}, \Pi)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(\hat{\pi}_i \neq \pi_i) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(\hat{\pi}_i \neq \pi_i, D) + o(n^{-3}). \quad (4.28)$$

It remains to bound the probability of making a clustering error on i, when the event D holds. Suppose $i \in \mathcal{C}_k$. On the event D, if $\|H\hat{r}_i - r_i\| \leq c_0 \sqrt{K}/4$, then

$$\|H\hat{r}_i - H\hat{v}_k\| \leq c_0 \sqrt{K}/4 + c_0 \sqrt{K}/8 \leq 3c_0 \sqrt{K}/8,$$

while for any $\ell \neq k$,

$$\|H\hat{r}_i - H\hat{v}_\ell\| \geq \|v_k - v_\ell\| - c_0 \sqrt{K}/4 - c_0 \sqrt{K}/8 \geq 5c_0 \sqrt{K}/8.$$

Then, node i must be clustered into $\hat{\mathcal{C}}_k$, i.e., there is no error on i. This implies that

$$\mathbb{P}(\hat{\pi}_i \neq \pi_i, D) \leq \mathbb{P}\left(\|H\hat{r}_i - r_i\| > c_0 \sqrt{K}/4 \right). \qquad (4.29)$$

We further study the right hand side of Eq. 4.29. Let $(\hat{\xi}_1, \hat{\Xi}_0, \omega, O)$ be the same as in Eq. 4.23. Fix i. Let $\hat{\Xi}'_{0,i} \in \mathbb{R}^{K-1}$ and $\Xi'_{0,i} \in \mathbb{R}^{K-1}$ denote the ith row of $\hat{\Xi}_0$ and Ξ_0, respectively. Then,

$$H\hat{r}_i = \frac{1}{\omega\hat{\xi}_1(i)}O'\hat{\Xi}_{0,i}, \qquad r_i = \frac{1}{\xi_1(i)}\Xi_{0,i}.$$

It is seen that

$$H\hat{r}_i - r_i = \frac{1}{\omega\hat{\xi}_1(i)}(O'\hat{\Xi}_{0,i} - \Xi_{0,i}) - \frac{\omega\hat{\xi}_1(i) - \xi_1(i)}{\omega\hat{\xi}_1(i)}r_i.$$

By Lemma B.2 of Jin et al. (2017), $\xi_1(i) \geq C^{-1}\theta_i/\|\theta\|$. Combining it with Eq. 4.23 gives $\|\omega\hat{\xi}_1 - \xi_1\|_\infty = o(\xi_1(i))$. It follows that $\omega\hat{\xi}_1(i) \geq \xi_1(i)/2 \geq C^{-1}\theta_i/\|\theta\|$. Additionally, $\|r_i\| \leq C\sqrt{K}$, by Eq. 4.25. Therefore, with probability $1 - o(n^{-3})$,

$$
\begin{aligned}
\|H\hat{r}_i - r_i\| &\leq \frac{C\|\theta\|}{\theta_i}\left(\|O'\hat{\Xi}_{0,i} - \Xi_{0,i}\| + \sqrt{K}|\omega\hat{\xi}_1(i) - \xi_1(i)|\right) \\
&\leq \frac{C\|\theta\|}{\theta_i}\left(\|O'\hat{\Xi}_{0,i} - \Lambda_0^{-1}\Xi'_0 A_{\cdot,i}\| + \|\Lambda_0^{-1}\Xi'_0 A_{\cdot,i} - \Xi_{0,i}\| + \sqrt{K}|\omega\hat{\xi}_1(i) - \xi_1(i)|\right) \\
&\leq \frac{C\|\theta\|}{\theta_{\min}}\left(\|\hat{\Xi}_0 O - A\Xi_0\Lambda_0^{-1}\|_{2\to\infty} + \sqrt{K}\|\omega\hat{\xi}_1 - \xi_1\|_\infty\right) + \frac{C\|\theta\|}{\theta_i}\|\Lambda_0^{-1}\Xi'_0 A_{\cdot,i} - \Xi_{0,i}\|.
\end{aligned}
$$

We plug in Theorem 4.1 and the first inequality of Eq. 4.23. It yields

$$
\begin{aligned}
\|H\hat{r}_i - r_i\| &\leq \frac{C\theta_{\max}K^5\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\beta_n\theta_{\min}\|\theta\|^2} + \frac{C\|\theta\|}{\theta_i}\|\Lambda_0^{-1}\Xi'_0 A_{\cdot,i} - \Xi_{0,i}\| \\
&\leq C\sqrt{K}\cdot a_1 + \frac{C_1\|\theta\|}{\theta_i}\|\Lambda_0^{-1}\Xi'_0 A_{\cdot,i} - \Xi_{0,i}\|,
\end{aligned}
$$

where the second inequality is from Eq. 4.22 and the constant C_1 does not depend on a_1. By choosing an appropriately small a_1, we can make the first term $\leq c_0\sqrt{K}/8$. It follows that

$$\mathbb{P}(\hat{\pi}_i \neq \pi_i, D) \leq \mathbb{P}\left(\frac{C_1\|\theta\|}{\theta_i}\|\Lambda_0^{-1}\Xi'_0 A_{\cdot,i} - \Xi_{0,i}\| > c_0\sqrt{K}/8\right) + o(n^{-3}). \quad (4.30)$$

Note that $A = \Omega + W - \text{diag}(\Omega)$, where $W = A - \mathbb{E}A$ and $\Omega = \Theta\Pi P\Pi'\Theta = \Xi\Lambda\Xi'$. In particular,

$$\Lambda_0^{-1}\Xi'_0\Omega = \Xi'_0.$$

It follows that

$$\Lambda_0^{-1}\Xi'_0 A_{\cdot,i} = \Lambda_0^{-1}\Xi'_0[\Omega + W - \text{diag}(\Omega)]_{\cdot,i} = \Xi_{0,i} + \Lambda_0^{-1}\Xi'_0 W_{\cdot,i} - \Omega(i,i)\Lambda_0^{-1}\Xi_{0,i}.$$

Note that $\Omega(i,i) \leq \theta_i^2$. Additionally, $\|\Lambda_0^{-1}\| = |\lambda_K|^{-1} \asymp K\beta_n^{-1}\|\theta\|^{-2}$ and $\|\Xi_{0,i}\| \leq \|\Xi_0\| \leq 1$. It follows that

$$
\begin{aligned}
\frac{C_1\|\theta\|}{\theta_i}\|\Lambda_0^{-1}\Xi_0'A_{\cdot,i} - \Xi_{0,i}\| &\leq \frac{C_1\|\theta\|\|\Lambda_0^{-1}\|}{\theta_i}\left(\|\Xi_0'W_{\cdot,i}\| + \theta_i^2\right) \\
&\leq \frac{C_2K}{\theta_i\beta_n\|\theta\|}\|\Xi_0'W_{\cdot,i}\| + \frac{C_2K\theta_{\max}}{\beta_n\|\theta\|},
\end{aligned}
$$

where $C_2 > 0$ is a constant that does not depend on a_1. The second term is $O(K\beta_n^{-1}\|\theta\|^{-1})$. At the same time, the left hand side of Eq. 4.22 is lower bounded by $K^4\sqrt{K\log(n)}/(\beta_n\|\theta\|)$. Therefore, Eq. 4.22 implies that the second term is $O(1/[K^3\sqrt{K\log(n)}]) = o(\sqrt{K})$. Particularly, for sufficiently large n, the second term is $\leq c_0\sqrt{K}/16$, i.e.,

$$
\frac{C_1\|\theta\|}{\theta_i}\|\Lambda_0^{-1}\Xi_0'A_{\cdot,i} - \Xi_{0,i}\| \leq \frac{C_2K}{\theta_i\beta_n\|\theta\|}\|\Xi_0'W_{\cdot,i}\| + (c_0/16)\sqrt{K}.
$$

We plug it into Eq. 4.30 to get

$$
\begin{aligned}
\mathbb{P}(\hat{\pi}_i \neq \pi_i, D) &\leq \mathbb{P}\left(\frac{C_2K}{\theta_i\beta_n\|\theta\|}\|\Xi_0'W_{\cdot,i}\| > (c_0/16)\sqrt{K}\right) + o(n^{-3}) \\
&= \mathbb{P}\left(\|\Xi_0'W_{\cdot,i}\|^2 > \frac{c_0^2}{16^2C_1^2} \cdot \frac{\theta_i^2\beta_n^2\|\theta\|^2}{K}\right) + o(n^{-3}) \\
&= \mathbb{P}\left(\sum_{k=2}^{K}(e_i'W\xi_k)^2 > \frac{c_0^2}{16^2C_1^2} \cdot \frac{\theta_i^2\beta_n^2\|\theta\|^2}{K}\right) + o(n^{-3}) \\
&\leq \sum_{k=2}^{K}\mathbb{P}\left(|e_i'W\xi_k| > \frac{c_0}{16C_1} \cdot \frac{\theta_i\beta_n\|\theta\|}{K}\right) + o(n^{-3}), \quad (4.31)
\end{aligned}
$$

where the last inequality is because of the probability union bound.

It remains to get a large deviation inequality for $|e_i'W\xi_k|$. Note that

$$
e_i'W\xi_k = \sum_{1\leq j\leq n: j\neq i} \xi_k(j)W(i,j).
$$

The summands are independent, and $|\xi_k(j)W(i,j)| \leq |\xi_k(j)| \leq C\sqrt{K}\theta_j/\|\theta\| \leq C\sqrt{K}\theta_{\max}/\|\theta\|$ (the bound of $|\xi_k(j)|$ is from Lemma B.2 of Jin et al. (2017)). We shall apply Bernstein's inequality. Note that $\sum_j \xi_k^2(j)\mathrm{Var}(W(i,j)) \leq \sum_j \xi_k^2(j)\|P\|_{\max}\theta_i\theta_j \leq C\sum_j(K\theta_j^2/\|\theta\|^2)\theta_i\theta_j \leq \theta_i \cdot CK\|\theta\|_3^3/\|\theta\|^2$. It follows from Bernstein's inequality that

$$\mathbb{P}\left(|e_i'W\xi_k| > t\right) \leq 2\exp\left(-\frac{t^2/2}{\theta_i \cdot CK\|\theta\|_3^3/\|\theta\|^2 + (t/3) \cdot C\sqrt{K}\theta_{\max}/\|\theta\|}\right), \qquad \text{for all } t > 0.$$

We plug in $t = (c_0/16C_1) \cdot K^{-1}\theta_i\beta_n\|\theta\|$. It follows that

$$\mathbb{P}\left(|e_i'W\xi_k| > \frac{c_0}{16C_1} \cdot \frac{\theta_i\beta_n\|\theta\|}{K}\right) \leq 2\exp\left(-\frac{K^{-2}\theta_i^2\beta_n^2\|\theta\|^2}{C_3\theta_i \cdot K\|\theta\|_3^3/\|\theta\|^2 + C_4\theta_i \cdot K^{-1/2}\theta_{\max}\beta_n}\right)$$

$$\leq 2\exp\left(-a_2\theta_i \cdot \min\left\{\frac{\beta_n^2\|\theta\|^4}{K^3\|\theta\|_3^3}, \frac{\beta_n\|\theta\|^2}{K\sqrt{K}\theta_{\max}}\right\}\right)$$

$$\leq 2\exp\left(-a_2\theta_i \cdot \min\left\{\frac{(|\lambda_K|/\sqrt{\lambda_1})^2\|\theta\|^2}{K^2\|\theta\|_3^3}, \frac{(|\lambda_K|/\sqrt{\lambda_1})\|\theta\|}{K\theta_{\max}}\right\}\right). \tag{4.32}$$

where C_3, C_4 are constants that depend on (c_0, C_2, C), $a_2 = \min\{C_3, C_4\}$, and the last inequality is due to $\beta_n \geq \sqrt{K}(|\lambda_K|/\sqrt{\lambda_1})$. We plug it into Eq. 4.31 to get

$$\mathbb{P}(\hat{\pi}_i \neq \pi_i, D) \leq 2K\exp\left(-a_2\theta_i \cdot \min\left\{\frac{(|\lambda_K|/\sqrt{\lambda_1})^2\|\theta\|^2}{K^2\|\theta\|_3^3}, \frac{(|\lambda_K|/\sqrt{\lambda_1})\|\theta\|}{K\theta_{\max}}\right\}\right) + o(n^{-3}).$$

Combining it with Eq. 4.28 gives the desired claim.

4.3. Proof of Corollary 2.1 We use an intermediate result in the proof of Theorem 2.1, which is the second last line of Eq. 4.32. We plug it into Eq. 4.31 and Eq. 4.28 to get

$$\mathbb{E}[\mathrm{Hamm}(\widehat{\Pi}, \Pi)] \leq 2K\sum_{k=1}^{K}\exp\left(-a_2\theta_i \cdot \min\left\{\frac{\beta_n^2\|\theta\|^4}{K^3\|\theta\|_3^3}, \frac{\beta_n\|\theta\|^2}{K\sqrt{K}\theta_{\max}}\right\}\right), \tag{4.33}$$

where $\beta_n = |\lambda_K(G^{1/2}PG^{1/2})|$, with $G = K\|\theta\|^{-2}\mathrm{diag}(\|\theta^{(1)}\|^2, \ldots, \|\theta^{(K)}\|^2)$. In this example, from the way π_i is generated, by elementary probability, $\|G - I_K\| = O(\sqrt{\log(K)/n})$; moreover, the first eigenvalue of P is $(1-b)+Kb$, and other eigenvalues are all equal to $(1-b)$. It follows that

$$\beta_n \asymp 1 - b.$$

Additionally, since $\theta_{\max} \leq C\theta_{\min}$, we have $\|\theta\|_3^3 \asymp \|\theta\|^2\bar{\theta}$. It follows that

$$\frac{\beta_n^2\|\theta\|^4}{K^3\|\theta\|_3^3} \asymp \frac{n}{\bar{\theta}}\frac{(1-b)^2\|\theta\|^2}{K^3}, \qquad \frac{\beta_n\|\theta\|^2}{K\sqrt{K}\theta_{\max}} \asymp \frac{1}{\bar{\theta}}\frac{(1-b)\|\theta\|}{K\sqrt{K}}.$$

The first term is dominating. We plug it into Eq. 4.33. The claim follows immediately.

4.4. Proof of Theorem 2.2 We have shown in Eq. 4.29 that there is an event D such that $\mathbb{P}(D^c) = o(n^{-3})$ and that on the event D,

$$\|H\hat{r}_i - r_i\| \le c_0\sqrt{K}/4 \qquad \Longrightarrow \qquad \hat{\pi}_i = \pi_i.$$

Therefore, it suffices to show that, with probability $1 - o(n^{-3})$,

$$\|H\hat{r}_i - r_i\| \le c_0\sqrt{K}/4. \tag{4.34}$$

In the equation above Eq. 4.30 and the equation above Eq. 4.31, we have shown that, as long as a_1 in Theorem 2.1 is properly small,

$$
\begin{aligned}
\|H\hat{r}_i - r_i\| &\le (c_0/8)\sqrt{K} + \frac{C_1\|\theta\|}{\theta_i}\|\Lambda_0^{-1}\Xi_0' A_{.,i} - \Xi_{0,i}\| \\
&\le (c_0/8)\sqrt{K} + \frac{C_2 K}{\theta_i\beta_n\|\theta\|}\|\Xi_0' W_{.,i}\| + (c_0/16)\sqrt{K} \\
&\le (3c_0/16)\sqrt{K} + \frac{C_2 K}{\theta_i\beta_n\|\theta\|}\sqrt{\sum_{k=2}^{K}(e_i' W\xi_k)^2}. \tag{4.35}
\end{aligned}
$$

We then apply Eq. 4.32. In order for the exponent of the right hand side of Eq. 4.32 to be at the order of $\log(n)$, we need

$$\theta_{\min} \cdot \frac{(|\lambda_K|/\sqrt{\lambda_1})^2\|\theta\|^2}{K^2\|\theta\|_3^3} \ge C\log(n), \quad \text{and} \quad \theta_{\min} \cdot \frac{(|\lambda_K|/\sqrt{\lambda_1})\|\theta\|}{K\theta_{\max}} \ge C\log(n), \tag{4.36}$$

for a large enough constant $C > 0$. Note that the condition on s_n implies

$$\frac{\theta_{\min}^2\|\theta\|^2(|\lambda_K|/\sqrt{\lambda_1})^2}{K^8\theta_{\max}^3\|\theta\|_1} \ge C\log(n),$$

for a large constant $C > 0$. It is straightforward that this condition guarantees Eq. 4.36. Then, the right hand side of Eq. 4.32 can be $o(n^{-3})$. In other words, with probability $1 - o(n^{-3})$,

$$|e_i' W\xi_k| \le cK^{-1}\theta_i\beta_n\|\theta\|, \tag{4.37}$$

where the constant $c_1 > 0$ can be arbitrarily small by setting the constant C in the assumption of s_n to be sufficiently large. We plug it into Eq. 4.35 to get, with probability $1 - o(n^{-3})$,

$$\|H\hat{r}_i - r_i\| \le (3c_0/16)\sqrt{K} + C_2 c\sqrt{K}.$$

Since c can be made arbitrarily small by increasing C in the assumption of s_n, we choose a large enough C such that $C_2 c < (c_0/16)\sqrt{K}$. Then, Eq. 4.34 is satisfied. The claim follows directly.

References

ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2019). Entrywise eigenvector analysis of random matrices with low expected rank. Ann. Statist. (to appear).

ADAMIC, L A and GLANCE, N (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43.

AIROLDI, E., BLEI, D., FIENBERG, S. and XING, E. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9**, 1981–2014.

BICKEL, P. J. and CHEN, A (2009).

CHAUDHURI, K., CHUNG, F. and TSIATAS, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Proceedings of the 25th annual conference on learning theory, JMLR workshop and conference proceedings*, vol. 23, pp. 1–35.

CHEN, Y., LI, X. and XU, J. (2018). Convexified modularity maximization for degree-corrected stochastic block models. *Ann. Statist.* **46**, 1573–1602.

DUAN, Y., KE, Z. T. and WANG, M. (2018). State aggregation learning from Markov transition data. In *NIPS workshop on probabilistic reinforcement learning and structured control.*

FAN, J., FAN, Y., HAN, X. and LV, J. (2019). SIMPLE: statistical inference on membership profiles in large networks. arXiv:1910.01734.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188.

GAO, C., MA, Z., ZHANG, A.Y. and ZHOU, H.H. (2018). Community detection in degree-corrected block models. *Ann. Statist.* **46**, 2153–2185.

GIRVAN, M and NEWMAN, M EJ (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 12, 7821–7826. National Acad Sciences.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The elements of statistical learning, 2nd edn.* Springer, Berlin.

JI, P. and JIN, J (2016). Coauthorship and citation networks for statisticians (with discussion). *Ann. Appl. Statist.* **10**, 4, 1779–1812.

JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43**, 57–89.

JIN, J. and KE, Z. T. (2018). Optimal membership estimation, especially for networks with severe degree heterogeneity. Manuscript.

JIN, J., KE, Z. T. and LUO, S. (2017). Estimating network memberships by simplex vertex hunting. arXiv:1708.07852.

JIN, J., TRACY KE, Z. and LUO, S. (2019). Optimal adaptivity of signed-polygon statistics for network testing. arXiv:1904.09532.

JIN, J., KE, Z. T., LUO, S. and WANG, M. (2020). Optimal approach to estimating K in social networks. Manuscript.

KARRER, B. and NEWMAN, M. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107.

KE, Z. T. and WANG, M. (2017). A new SVD approach to optimal topic estimation. arXiv:1704.07016.

KE, Z. T., SHI, F. and XIA, D. (2020). Community detection for hypergraph networks via regularized tensor power iteration. arXiv:1909.06503.

LUSSEAU, D, SCHNEIDER, K, BOISSEAU, O J, HAASE, P, SLOOTEN, E and DAWSON, S M (2003). The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **54**, 4, 396–405. Springer.

LIU, Y., HOU, Z., YAO, Z., BAI, Z., HU, J. and ZHENG, S. (2019). Community detection based on the ℓ_∞ convergence of eigenvectors in dcbm. arXiv:1906.06713.

MA, Z., MA, Z. and YUAN, H. (2020). Universal latent space model fitting for large networks with edge covariates. *J. Mach. Learn. Res.* **21**, 1–67.

MAO, X., SARKAR, P. and CHAKRABARTI, D. (2020). Estimating mixed memberships with sharp eigenvector deviations. J. Amer. Statist. Assoc. (to appear), 147.

MIHAIL, M. and PAPADIMITRIOU, C. H. (2002). On the eigenvalue power law. In *International workshop on randomization and approximation techniques in computer science*, pp. 254–262. Springer, Berlin.

NEPUSZ, T, PETRÓCZI, A, NÉGYESSY, L and BAZSÓ, F (2008). Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E* **77**, 1, 016107. APS.

QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. Adv. Neural Inf. Process. Syst. 3120–3128.

SU, L., WANG, W. and ZHANG, Y. (2019). Strong consistency of spectral clustering for stochastic block models. *IEEE Trans. Inform. Theory* **66**, 324–338.

TRAUD, A. L., KELSIC, E. D., MUCHA, P. J. and PORTER, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* **53**, 526–543.

TRAUD, A. L., MUCHA, P. J. and PORTER, M. A. (2012). Social structure of facebook networks. *Physica A* **391**, 4165–4180.

ZACHARY, W W (1977). An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 4, 452–473. University of New Mexico.

ZHANG, Y., LEVINA, E. and ZHU, J. (2020). Detecting overlapping communities in networks using spectral methods. *SIAM J. Math. Anal.* **2**, 265–283.

ZHAO, Y, LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40**, 2266–2292.

JIASHUN JIN
SHENGMING LUO
CARNEGIE MELLON UNIVERSITY,
PITTSBURG, PA, USA
E-mail: jiashun@stat.cmu.edu

ZHENG TRACY KE
HARVARD UNIVERSITY, CAMBRIDGE, MA,
USA