Optimal Network Membership Estimation under Severe Degree Heterogeneity

Zheng Tracy Ke
Department of Statistics, Harvard University

and

Jingming Wang
Department of Statistics, Harvard University

August 3, 2024

Abstract

Real networks often have severe degree heterogeneity, with the maximum, average, and minimum node degrees differing significantly. This paper examines the impact of degree heterogeneity on statistical limits of network data analysis. Introducing the heterogeneity distribution (HD) under a degree-corrected mixed-membership network model, we show that the optimal rate of mixed membership estimation is an explicit functional of the HD. This result confirms that severe degree heterogeneity may decelerate the error rate, even when the overall sparsity remains unchanged.

To obtain a rate-optimal method, we modify an existing spectral algorithm, Mixed-SCORE, by adding a pre-PCA normalization step. This step normalizes the adjacency matrix by a diagonal matrix consisting of the bth power of node degrees, for some $b \in \mathbb{R}$. We discover that $b = 1/2$ is universally favorable. The resulting spectral algorithm is rate-optimal for networks with arbitrary degree heterogeneity. A technical component in our proofs is entry-wise eigenvector analysis of the normalized graph Laplacian.

Keywords: DCMM, entry-wise eigenvector analysis, graph Laplacian, least favorable configuration, leave-one-out, Mixed-SCORE, random matrix theory, SCORE.

1 Introduction

Networks are widely observed in social sciences, machine learning, economics, physics, and bioinformatics. One of the core problems in network data analysis is community detection, which aims to cluster nodes into socially meaningful groups. Mixed membership estimation [4] is a "soft clustering" problem that allows a node to have memberships in multiple communities. Given an undirected network with n nodes, let $A \in \mathbb{R}^{n\times n}$ be the adjacency matrix, where $A(i,j) \in \{0,1\}$ indicates if there is an edge between two nodes i and j. Suppose there are K communities. Each node has a Probability Mass Function (PMF) $\pi_i \in \mathbb{R}^K$ such that

$$\pi_i(k) \text{ is the fractional weight of node } i \text{ on community } k, \qquad 1 \leq k \leq K.$$

The goal is estimating $\pi_1, \pi_2, \ldots, \pi_n$ from A. One example of networks with mixed memberships is the political blog network [2]. Each node is a political blog active during the 2004 U.S. Presidential Election, and each edge is a link between two blogs in a snapshot on one day before the election. There are two communities, conservative and liberal. However, for most bloggers, they are neither extremely conservative nor extremely liberal; their positions in-between can be captured by mixed membership [24]. Mixed membership estimation has also been used to learn research interests of statisticians from co-citation networks [20] and understand developmental brain disorders from gene co-expression networks [27].

Several methods have been proposed for estimation and inference of network mixed memberships. To name a few, the Bayesian approach in *Airoldi et al.* [4] puts a Dirichlet prior on π_i's and uses variational inference to get the posteriors, the spectral approaches by *Zhang et al.* [34] and *Jin et al.* [24] estimate π_i's from the leading eigenvectors of A, and *Fan et al.* [13] and *Bhattacharya et al.* [7] studied the hypothesis testing and uncertainty quantification for individual membership vectors.

Despite these encouraging progresses, there is an open problem: how to deal with severe *degree heterogeneity* in real-world networks. An observed static network is often a snapshot of a dynamic process involving node birth and edge formation. The preferential attachment (“rich get richer”) phenomenon [6] is quite common in the underlying dynamic process; as a consequence, the degrees of a small set of nodes can be a magnitude larger than the degrees of other nodes. Meanwhile, there are a large number of nodes whose degrees are quite small, most of which are the “young” nodes in the underlying dynamic process. [23] reported the maximum, minimum, and average degrees of some frequently-studied network datasets and found that the three quantities, $d_{\max}$, $d_{\min}$, and d_{mean}, could be significantly different.

The severe degree heterogeneity in real-world networks inspires a question: How does degree heterogeneity affect the accuracy of mixed membership estimation? Suppose we have two networks whose average node degrees are comparable with each other. In one network, the node degrees are at the same order. In the other network, node degrees are significantly different from each other. Do we achieve the same accuracy of mixed membership estimation on these two networks? We provide the first result that connects degree heterogeneity and the optimal rate of mixed membership estimation. Under a degree-corrected mixed membership (DCMM) model [24], we propose to summarize degree heterogeneity by a cumulative distribution function $F_n(\cdot)$, defined using normalized degree parameters in DCMM, and we show that the optimal rate of mixed membership estimation is determined by a baseline rate (which is related to the average degree) and $F_n(\cdot)$ (which captures degree heterogeneity).

Knowing the optimal rate, the next question is if there exists an algorithm that achieves the optimal rate under arbitrary degree heterogeneity. Existing methods include the variational inference approach [4], the spectral approach [28], and nonnegative matrix factorization [33]. There are also works studying mixed membership estimation for dynamic networks [14], networks with different node types [17], and tensor models [3]. However, most of these works assume no degree heterogeneity, hence not applicable in our setting. OCCAM [34] is a method that allows for degree heterogeneity, but it restricts the fraction of mixed nodes to be relatively small. Mixed-SCORE [24] is the only existing method that allows for severe degree heterogeneity and an arbitrary fraction of mixed nodes. We then focus on Mixed-SCORE and compare its rate with the aforementioned lower bound. We discover that Mixed-SCORE may be non-optimal when degree heterogeneity is severe. Inspired by some non-trivial theoretical insights, we improve Mixed-SCORE by adding a (pre-PCA) Laplacian normalization and a (post-PCA) node trimming. The resulting method is called Mixed-SCORE-Laplacian (MSL). We show that MSL is rate-optimal under arbitrary degree heterogeneity.

Another theoretical result we provide is the node-wise error of MSL. For each node i, we derive a large-deviation bound for $\|\hat{\pi}_i - \pi_i\|_1$. The bound is a monotone decreasing function of θ_i. It shows that the estimation error is not evenly distributed among nodes: higher-degree nodes receive more accurate estimation of π_i. Such node-wise error bounds will be useful for downstream tasks such as using $\hat{\pi}_i$ to rank node preferences or build prediction models.

The backbone of our analysis is entry-wise eigenvector analysis for normalized adjacency matrices. We consider a degree-normalization of A to $L = H^{-b}AH^{-b}$ for any $b \geq 0$, where H is a diagonal matrix containing regularized node degrees. We study each entry of the leading eigenvectors of L and show how its large-deviation is related to b. Such analysis reveals the effects of different degree-normalizations and inspires us to use the Laplacian normalization. Fixing the Laplacian normalization, we derive entry-wise large-deviation bounds for leading eigenvectors, and use this result to derive the node-wise error and rate-optimality of MSL.

In summary, we offer the first rate-optimality study of mixed membership estimation for arbitrary degree heterogeneity. Our contributions include (i) a degree-heterogeneity-aware lower bound, (ii) a rate-optimal spectral algorithm, (iii) node-wise estimation error bounds, and (iv) entry-wise eigenvector analysis for the normalized graph Laplacian. To demonstrate the practical implications of our work, we apply our method to a political blog network [2] and a coauthorship network [19].

Related literature

Mixed membership estimation may be viewed as "soft" community detection, but it is fundamentally different from community detection. Community detection is a clustering problem, and its optimality can be achieved by a post-spectral-clustering 'majority vote' [15]. However, optimality of mixed membership estimation has much higher requirement on the entrywise signal-to-noise ratio in eigenvectors, hence a more challenging problem.

A lower bound for mixed membership estimation under moderate degree heterogeneity was given in a short note [22]. However, this lower bound is not sharp for severe degree heterogeneity. Using more sophisticated proofs, we provide sharp lower bounds under arbitrary degree heterogeneity. We also have many other contributions - a new algorithm, node-wise error bounds, and entry-wise eigenvector analysis, but none of which were seen in [22].

One of our ideas in methodology development is using the normalized graph Laplacian. This matrix has been used in community detection [29, 23]. There are interesting theoretical studies explaining why it improves community detection perfromance, such as large-deviation bounds for matrix spectral norms [8] and analysis of global clustering quality metrics [30]. However, these results are insufficient to conclude that Laplacian is the correct normalization for optimal mixed membership estimation under arbitrary degree heterogeneity. In Section 3, we show that an optimal method has to attain the correct node-wise errors simultaneously for all nodes. Hence, any conclusion of optimality requires both *node-wise error analysis* and *degree-heterogeneity-aware lower bounds*. Neither results existed in the literature.

A technical component of our work is the entry-wise eigenvector analysis for normalized graph Laplacian. In the literature, entry-wise eigenvector analysis was conducted for eigenvectors of the adjacency matrix under a stochastic block model (SBM), such as in *Abbe et al.* [1]. However, a direct extension of the analysis in [1] does not work for our purpose, because: (i) the Laplacian matrix no longer has independent entries, (ii) our model allows for severe degree heterogeneity, and (iii) we need tighter bounds than those in [1]. We overcome these hurdles by new proof ideas, which will be explained in Section 5.

Our theory is for the asymptotic regime where the average node degree grows faster than $\log(n)$, and the rate-optimality of our proposed method is subject to a logarithmic factor. It is an interesting question to remove such a logarithmic gap, but the study is very challenging (see Section 4.2 for more discussions). We leave it to future work.

The remainder of this paper is organized as follows. Section 2 states the model, asymptotic settings, and lower bounds. Section 3 introduces the new algorithm. Section 4 contains the main results, including entry-wise eigenvector analysis, node-wise errors, rate-optimality, and extensions to general loss functions. Section 5 explains the techniques used in entry-wise eigenvector analysis. Section 6 and Section 7 present the simulations and real-data results, respectively. Discussions are in Section 8, and all proofs are in the supplementary material.

2 The model, asymptotic settings, and lower bounds

We model the network with a degree-corrected mixed membership (DCMM) model [24]. DCMM generalizes the stochastic block model (SBM) by incorporating both degree heterogeneity and mixed membership, and its special cases include the well-known mixed membership stochastic block model (MMSBM) [4] and degree-corrected block model (DCBM) [26]. We present an information-theoretic lower bound under DCMM. This lower bound depends on the distribution of node degree parameters, so it reveals the fundamental impact of degree heterogeneity on mixed membership estimation.

2.1 The DCMM model and the asymptotic settings

Let K be the number of communities and let $\Pi = [\pi_1, \pi_2, \ldots, \pi_n]' \in \mathbb{R}^{n\times K}$ be the membership matrix. DCMM uses a symmetric non-negative matrix $P \in \mathbb{R}^{K\times K}$ to model the community structure and a positive vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$ to model the degree heterogeneity among nodes. The upper triangular entries of A are independent Bernoulli variables satisfying

$$\mathbb{P}\big(A(i,j) = 1\big) = \theta_i\theta_j \cdot \pi_i' P \pi_j, \qquad 1 \le i < j \le n. \tag{1}$$

To guarantee parameter identifiability, we follow [24] to require that P is non-singular and has unit diagonals. Write $\Theta = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_n)$.[1] It is seen that

$$A = \Omega - \mathrm{diag}(\Omega) + W, \qquad \text{where} \quad \Omega := \Theta\Pi P\Pi'\Theta \quad \text{and} \quad W := A - \mathbb{E}[A]. \tag{2}$$

[1] Given a vector $v \in \mathbb{R}^n$, we use $\mathrm{diag}(v)$ or $\mathrm{diag}(v_1, v_2, \ldots, v_n)$ to denote the $n \times n$ diagonal matrix whose ith diagonal entry is equal to v_i, for $1 \le i \le n$.

We call node i a *pure node* if π_i is degenerate (i.e., it has only one nonzero coordinate that is equal to 1, with the other coordinates being zero) and a *mixed node* otherwise. The DCMM model is identifiable provided that each community has at least one pure node [24].

Given A and K, we are interested in estimating Π. We use n as the driving asymptotic parameter and study the optimal error rate as $n \to \infty$. A challenge is that DCMM has many nuisance parameters, all of which affect the error rate. We hope to define a few summarizing quantities so that the optimal rate only depends on these quantities.

An illustrating example *(Random-parameter DCMM).* Let $g(\cdot)$ be a distribution on the standard simplex of $\mathbb{R}^K$ such that $\mathbb{E}_g[\pi] = \frac{1}{K}\mathbf{1}_K$ and $\lambda_{\min}(\mathbb{E}_g[\pi\pi']) \geq \frac{c}{K}$, for a constant $c > 0$. Let $F(\cdot)$ be a distribution with support in $(0, \infty)$ and mean equal to 1. For $a_n, p_n \in (0, 1)$,

$$P = (1 - p_n)I_K + p_n\mathbf{1}_K\mathbf{1}_K', \qquad \pi_i \overset{iid}{\sim} g(\cdot), \qquad \theta_i \overset{iid}{\sim} a_n \cdot F(\cdot).$$

Here, $(1-p_n)$ captures the 'dissimilarity' between communities. The average node degree is at the order na_n^2 (if K is finite), so a_n captures the sparsity of the network. The distribution $F(\cdot)$ controls degree heterogeneity (e.g., when F is a point mass, there is no degree heterogeneity). We expect that the optimal rate depends on $(1 - p_n)$, a_n and $F(\cdot)$.

Using insights from this example, we now define the parameter class for a general DCMM, where we introduce a few quantities that serve as the counterpart of $(1 - p_n, a_n, F(\cdot))$. Write $\bar{\theta} = n^{-1}\sum_{i=1}^n \theta_i$ and $\eta_i = \theta_i/\bar{\theta}$, $1 \leq i \leq n$.

Definition 2.1. *The heterogeneity distribution (HD) is the empirical distribution associated with $\eta_1, \eta_2, \ldots, \eta_n$, whose CDF is $F_n(t) = n^{-1}\sum_{i=1}^n 1\{\theta_i \leq t \cdot \bar{\theta}\}$.*

Since the diagonals of P are fixed at 1 for identifiability, the average node degree is $\asymp n\bar{\theta}^2$, implying that $\bar{\theta}$ captures network sparsity. In addition, $F_n(\cdot)$ captures degree heterogeneity. Here, $\bar{\theta}$ and $F_n(\cdot)$ serve as the counterpart of a_n and $F(\cdot)$ for a general DCMM model.

Let $d_i = \sum_j A(i,j)$ be the degree of node i and $\bar{d} = n^{-1}\sum_i d_i$ be the average node degree.

Write $D_\theta = \mathrm{diag}(\mathbb{E}d_1, \mathbb{E}d_2, \ldots, \mathbb{E}d_n) + (\mathbb{E}\bar{d})I_n$. We define a "community-overlap" matrix

$$G = K \cdot \Pi' \Theta D_\theta^{-1} \Theta \Pi \quad \in \quad \mathbb{R}^{K \times K}. \tag{3}$$

Note that $G(k,\ell) = \frac{K}{n}\sum_{i=1}^n \frac{n\theta_i^2}{\mathbb{E}(d_i+\bar{d})}\pi_i(k)\pi_j(\ell)$, where for most nodes i, $\frac{n\theta_i^2}{\mathbb{E}(d_i+\bar{d})} \asymp 1$. Therefore, when $k = \ell$, a large value of $G(k,k)$ implies that a significant fraction of nodes have nonzero membership on community k, so $\frac{n}{K}G(k,k)$ is the "effective size" of this community; when $k \neq \ell$, a large value of $G(k,\ell)$ indicates a big "overlap" between community k and community ℓ. Our study suggests that the counterpart of $(1-p_n)$ for a general DCMM should be defined through eigenvalues of PG. In detail, let $\lambda_k(PG)$ denote the kth largest (in magnitude) right eigenvalue of PG, for $1 \leq k \leq K$. Define

$$\beta_n := |\lambda_K(PG)|. \tag{4}$$

We use β_n as the counterpart of $(1-p_n)$ for a general DCMM. In the aforementioned random-parameter DCMM example, $G \sim K \cdot \mathbb{E}_g[\pi\pi']$, and $\beta_n \asymp 1 - p_n$ (this example also suggests that G is a global quantity that is insensitive to degree heterogeneity).

Later in this section, we will present an information-theoretical lower bound for mixed membership estimation, which depends on n, K, β_n (capturing community dis-similarity), $\bar{\theta}$ (capturing network sparsity), and $F_n(\cdot)$ (capturing degree heterogeneity).

2.2 Regularity conditions

Our results require some regularity conditions. Let G and $\lambda_k(PG)$ be the same as in Section 2.1. For each $1 \leq k \leq K$, denote by $\eta_k \in \mathbb{R}^K$ the kth right eigenvector of PG (associated with the eigenvalue $\lambda_k(PG)$). Fix positive constants c_1-c_4.

Condition 2.1. *The DCMM parameters (θ, Π, P) satisfy the following requirements:*

(a) $\|G\| \leq c_1$, $\|G^{-1}\| \leq c_1$, *and* $\min_{1\leq k\leq K}\{\sum_{i=1}^n \theta_i\pi_i(k)\} \geq c_1 K^{-1}\|\theta\|_1$.

(b) $\max_{k\neq 1}\{\lambda_k(PG)\} \leq (1-c_2)\cdot\lambda_1(PG)$, *and* $\min_{1\leq k\leq K}\{\sum_{1\leq \ell\leq K} P(k,\ell)\} \geq c_2 K$.

(c) η_1 *is a positive vector, satisfying that* $\min_{1\leq k\leq K}\{\eta_1(k)\} \geq c_3 \max_{1\leq k\leq K}\{\eta_1(k)\}$.

(d) Each community has at least one pure node whose θ_i is lower bounded by $c_4\bar{\theta}$.

Condition (a) requires that the matrix G defined in (3) is well-conditioned and the degree parameters are balanced across communities. These are commonly assumed in the literature [24]. In Condition (b), the first is an eigen-gap condition. Since PG is a nonnegative matrix, by Perron's theorem [16], $\max_{k\neq 1}\{\lambda_k(PG)\}$ is strictly smaller than $\lambda_1(PG)$. Here, we further assume that $\max_{k\neq 1}\{\lambda_k(PG)\}$ is smaller than $(1-c_2)\cdot\lambda_1(PG)$, which is mild. The second requirement in (b) restricts that the sum of each row of P is at the order of K. This is for convenience of presentation and can be easily relaxed (see Section J of the supplement). For Condition (c), note that η_1 is a positive vector when PG is irreducible (by Perron's theorem [16]). Hence, this condition is only slightly stronger than requiring the irreducibility of PG. [2] Condition (d) is related to the identifiability of DCMM, which requires each community to have at least one pure node. Here we put a slightly stronger requirement: there is at least one "moderate-degree" pure node per community.

We also define a class of θ. We hope this class to be broad enough to cover various kinds of degree heterogeneity. We note that the EHD (see Definition 2.1) can have many different situations: First, $F_n(\cdot)$ may be discrete (i.e., θ_i's take only finitely many values), or converge to a continuous CDF as $n\to\infty$. Second, $F_n(\cdot)$ may have an unbounded support, to allow for $\theta_{\max}\gg\bar{\theta}$. Finally, the density of $F_n(\cdot)$ may have different shapes in the neighborhood of zero, to control the fraction of extremely-low-degree nodes. We introduce the following class of θ. Despite that its definition is technical, this class indeed covers all the above cases.

Definition 2.2. *Given constants $\varrho\in(0,1)$ and $a_0\in(0,1)$ and any sequence $x_n\to 0$, let $\mathcal{G}_n(\varrho, a_0, x_n)$ be the collection of $\theta\in\mathbb{R}^n$ such that there exists $c_n>0$ satisfying $\varrho c_n\geq x_n^2$, $F_n(c_n)\leq 1-a_0$, and $\int_{\varrho c_n}^{c_n}\frac{1}{\sqrt{t\wedge 1}}dF_n(t)\geq a_0\int_{x_n^2}^{\infty}\frac{1}{\sqrt{t\wedge 1}}dF_n(t)$.*

In Section A.2 of the supplementary material, we justify the broadness of $\mathcal{G}(\varrho, a_0, x_n)$. We

[2] When PG is reducible, with probability $1-o(1)$, the network splits into multiple disconnected components. In that case, we will conduct mixed membership estimation separately on each component, and each sub-problem still has an irreducible PG.

show that the requirements in Definition 2.2 are satisfied with high-probability if $\theta_i \overset{iid}{\sim} a_n \cdot F$, where $a_n > 0$ is a scalar and $F(\cdot)$ is a fixed, finite-mean distribution which has its support in $(0, \infty)$ and belongs to one of the following cases: (i) $F(\cdot)$ is a discrete distribution; (ii) $F(\cdot)$ is a continuous distribution with support in $[c, \infty)$, for some $c > 0$; (iii) $F(\cdot)$ is a continuous distribution supported in $(0, \infty)$, and its density $f(t)$ satisfies that $\lim_{t\to 0+}\{t^\omega f(t)\} = C$, for some constants $\omega \neq 1/2$ and $C > 0$.

2.3 A θ-dependent lower bound

For any estimator $\hat{\Pi} = [\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n]'$, we measure its performance by the ℓ^1-loss:

$$\mathcal{L}(\hat{\Pi}, \Pi) = \min_T \Big\{ \frac{1}{n} \sum_{i=1}^n \|T\hat{\pi}_i - \pi_i\|_1 \Big\}, \tag{5}$$

where the minimum is taken over column permutation of $\hat{\Pi}$. Given any $\theta \in \mathbb{R}^n_+$, $K \geq 2$, and $\beta_n \in (0,1)$, let $\mathcal{Q}_n(\theta) = \mathcal{Q}_n(\theta; K, \beta_n)$ be the collection of (Π, P) such that $|\lambda_K(PG)| \geq \beta_n$ and that (θ, Π, P) satisfies Condition 2.1. The meaning of β_n here is slightly different from the one in (4) —we use β_n as the respective lower bound of $|\lambda_K(PG)|$, by the convention of minimax analysis. Define the minimax error for a particular θ as

$$\mathcal{L}^*_n(\theta) := \inf_{\hat{\Pi}} \sup_{(\Pi,P)\in\mathcal{Q}_n(\theta)} \mathbb{E}\mathcal{L}(\hat{\Pi}, \Pi). \tag{6}$$

Define the 'baseline' rate as

$$err_n = K^{3/2}\beta_n^{-1}(n\bar{\theta}^2)^{-1/2} \tag{7}$$

and the 'degree-heterogeneity-aware' rate as

$$err_n(\theta) := \int \min\Big\{ \frac{err_n}{\sqrt{t \wedge 1}}, 1 \Big\} dF_n(t) = \frac{1}{n}\sum_{i=1}^n \min\left\{ \frac{K\sqrt{K}}{\beta_n\sqrt{n\bar{\theta}(\bar{\theta}\wedge\theta_i)}},\ \ 1 \right\}. \tag{8}$$

Theorem 2.1 (A θ-dependent lower bound)**.** *Fix constants $(c_1, c_2, c_3, c_4, \varrho, a_0)$ and positive sequences (K, β_n) such that $K^{3/2}\beta_n^{-1}(n\bar{\theta}^2)^{-1/2} \to 0$. There exists a constant $C_1 > 0$ such that simultaneously for all $\theta \in \mathcal{G}_n(\varrho, a_0, err_n)$, $\mathcal{L}^*_n(\theta) \geq C_1 err_n(\theta)$.*

Delving into $err_n(\theta)$, the individual contribution of node i in the lower bound is $\tau_n(\theta_i) \equiv \min\{\frac{K\sqrt{K}}{\beta_n\sqrt{n\bar{\theta}(\bar{\theta}\wedge\theta_i)}}, 1\}$. In Figure 1, we plot the curve of $\tau_n(\theta_i)$ versus θ_i. The 'minimum with

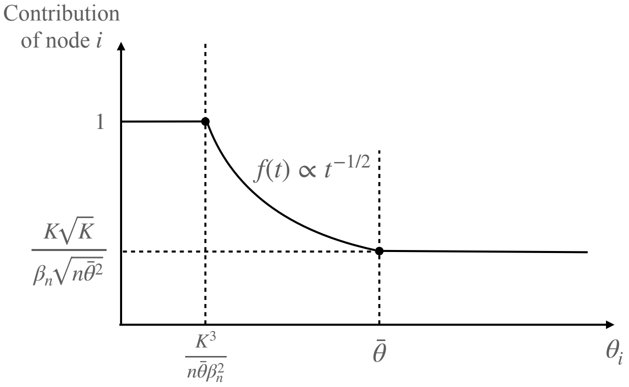

Figure 1: Visualization of the contributions of individual nodes in the minimax lower bound. The x-axis represents θ_i, and the y-axis corresponds to the summand in (8).

1' reflects the naive bound $\|\hat{\pi}_i - \pi_i\|_1 \leq 2$. When $\frac{K\sqrt{K}}{\beta_n\sqrt{n\bar{\theta}(\bar{\theta}\wedge\theta_i)}} \leq 1$, $\tau_n(\theta_i)$ depends on the minimum of θ_i and $\bar{\theta}$. This is because the error at π_i comes from (i) noise in the ith row of A (controlled by θ_i) and (ii) the errors of estimating global parameters such as P (controlled by $\bar{\theta}$). When θ_i is small, (i) dominates (ii); but when θ_i is large, (ii) dominates (i).

We revisit the example of random-parameter DCMM in Section 2.1, where P has equal off-diagonal entries, $\pi_i \overset{iid}{\sim} g(\cdot)$, and $\theta_i \overset{iid}{\sim} a_n \cdot F(\cdot)$. Then, $err_n = \frac{K\sqrt{K}}{(1-p_n)\sqrt{na_n^2}}$, and $err_n(\theta) = \int \min\{\frac{err_n}{\sqrt{t\wedge 1}}, 1\}dF(t)$. The effect of degree heterogeneity is seen from comparing $err_n(\theta)$ with the baseline rate. We consider a few examples of $F(\cdot)$:

Proposition 2.1. *Let $\epsilon \in (0,1)$, $\alpha > 0$ and $c_{\min} > 0$ be constants, and let $L \geq 1$ be a constant integer. Let δ_x denote the point mass at x. Suppose $x_1 < x_2 < \ldots < x_L$, $\sum_{\ell=1}^L \epsilon_\ell x_\ell = 1$, and as $n \to \infty$, $x_1 \gg err_n^2$.*

$$\int \min\Big\{\frac{err_n}{\sqrt{t\wedge 1}}, 1\Big\}dF(t) \asymp \begin{cases} err_n, & \textit{when } F = \mathrm{Uniform}([1-\epsilon, 1+\epsilon]), \\ err_n, & \textit{when } F = \frac{\alpha-1}{\alpha c_{\min}}\mathrm{Pareto}(c_{\min}, \alpha), \textit{ with } \alpha > 1, \\ err_n^{(1\wedge 2\alpha)}, & \textit{when } F = \frac{\beta}{\alpha}\mathrm{Gamma}(\alpha, \beta), \textit{ with } \alpha \neq 1/2, \\ err_n \cdot \max_{1\leq\ell\leq L}\big\{\frac{\epsilon_\ell}{\sqrt{x_\ell\wedge 1}}\big\}, & \textit{when } F = \sum_{\ell=1}^L \epsilon_\ell\delta_{x_\ell}. \end{cases}$$

Among the examples in Proposition 2.1, the uniform distribution corresponds to moderate degree heterogeneity, where $\theta_{\max} \asymp \bar{\theta} \asymp \theta_{\min}$. The Pareto distribution is an example of severe degree heterogeneity, where $\theta_{\max} \gg \bar{\theta} \asymp \theta_{\min}$. The Gamma distribution indicates even more severe heterogeneity: $\theta_{\max} \gg \bar{\theta} \gg \theta_{\min}$; and the optimal rate is slower than the baseline rate when $\alpha < 1/2$. The last example is a discrete distribution, for which the optimal rate is slower than the baseline rate if there are a considerable fraction of extremely small θ_i's.

3 The Mixed-SCORE-Laplacian (MSL) algorithm

We shall introduce an algorithm to achieve the lower bound in Theorem 2.1. Most methods for mixed membership estimation [4] assume that θ_i's are equal (i.e., no degree heterogeneity). Mixed-SCORE [24] is one of the few methods that deal with degree heterogeneity, but its error rate only matches with the lower bound in Theorem 2.1 for some special θ. We modify Mixed-SCORE to achieve the lower bound for any θ in the regular class in Definition 2.2.

3.1 A general node embedding, and review of Mixed-SCORE

We define a general node embedding approach, which extends the SCORE embedding [21]. Given constants $\tau > 0$ and $b \geq 0$, let

$$L = H^{-b} A H^{-b}, \qquad \text{where} \quad H = \mathrm{diag}(d_1, d_2, \ldots, d_n) + \tau \bar{d} \cdot I_n. \tag{9}$$

Here L is a normalized version of A, where the entries of A are re-weighted by node degrees. τ is a ridge-regularization parameter to avoid extreme entries in H^{-b}; we usually set $\tau = 1$. The most important parameter is b, which controls the level of re-weighting. When $b = 0$, L is the adjacency matrix; when $b = 1/2$, L is the normalized graph Laplacian. Let $\hat{\lambda}_1, \cdots, \hat{\lambda}_K$ be the K largest eigenvalues (in magnitude) of L, and let $\hat{\xi}_1, \cdots, \hat{\xi}_K \in \mathbb{R}^n$ be the corresponding eigenvectors. Define an $n \times (K-1)$ matrix $\hat{R}$ by

$$\hat{R}(i,k) = \hat{\xi}_{k+1}(i)/\hat{\xi}_1(i), \qquad 1 \leq i \leq n,\ 1 \leq k \leq K-1. \tag{10}$$

When the network is connected, by Perron's theorem, $\hat{\xi}_1$ is always a strictly positive vector, so $\hat{R}$ is well-defined. Let $\hat{r}_1', \hat{r}_2', \ldots, \hat{r}_n'$ denote the rows of $\hat{R}$. We use $\hat{r}_i$ as a $(K-1)$-dimensional

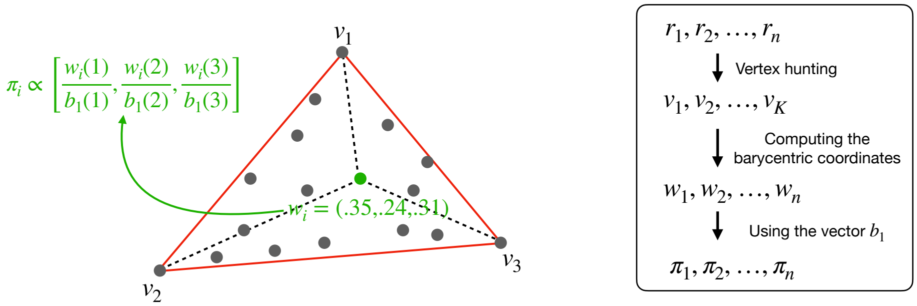
$\pi_i \propto \left[\frac{w_i(1)}{b_1(1)}, \frac{w_i(2)}{b_1(2)}, \frac{w_i(3)}{b_1(3)}\right]$
v_1
v_2
v_3
$w_i = (.35,.24,.31)$
$r_1, r_2, \ldots, r_n$
Vertex hunting
$v_1, v_2, \ldots, v_K$
Computing the barycentric coordinates
$w_1, w_2, \ldots, w_n$
Using the vector b_1
$\pi_1, \pi_2, \ldots, \pi_n$

Figure 2: The main insight of mixed membership estimation. In the left panel, the black dots are $r_1, r_2, \ldots, r_n$. The red triangle is the Ideal Simplex, with K vertices $v_1, v_2, \ldots, v_K$. Each r_i has a barycentric coordinate w_i in the simplex, and w_i is related to the target membership vector π_i through a vector b_1. In the right panel, we present the steps of estimating π_i's from the node embeddings.

embedding of node i. We call them the (general) SCORE embeddings of nodes. The original SCORE embedding [21] is a special case with $b = 0$ in (9), but we allow a general b here.

The main idea of Mixed-SCORE [24] is to estimate π_i from the simplex geometry associated with node embeddings. For illustration, we consider an oracle case where $A = \Omega$. The eigen-pairs $(\hat{\lambda}_k, \hat{\xi}_k)$ and the embedded vectors $\hat{r}_i$ now become non-stochastic, and we re-write them as (λ_k, ξ_k) and r_i. Figure 2 displays the point cloud $\{r_i\}_{i=1}^n$. We observe that this point cloud is contained in a simplex (called the Ideal Simplex) with vertices $v_1, v_2, \ldots, v_K$. For any pure node i, the corresponding r_i falls on one vertex. We can recover the Ideal Simplex as long as there is a pure node for each community. Such a step is called Vertex Hunting [24]. By definition, each point in the simplex can be uniquely written as a convex combination of its K vertices, and the vector containing the combination coefficients is called the *barycentric coordinate* of this point in the simplex. We use $w_i \in \mathbb{R}^K$ to denote the barycentric coordinate of r_i in the Ideal Simplex. After vertex hunting, we can obtain w_i immediately from r_i and the recovered vertices. [24] observed that π_i is linked to w_i via $\pi_i \propto [\text{diag}(b_1)]^{-1} w_i$, where $b_1 \in \mathbb{R}^K$ is defined by $b_1(k) = [\lambda_1 + v_k' \text{diag}(\lambda_2, \ldots, \lambda_K) v_k]^{-1/2}$. We thus estimate π_i's as follows: First, we compute b_1 from the K vertices and the eigenvalues of L. Next, for each $1 \leq i \leq n$, let $\pi_i^* = [\text{diag}(b_1)]^{-1} w_i$ and $\hat{\pi}_i = \pi_i^* / \|\pi_i^*\|_1$. The whole estimation procedure is summarized in Figure 2. In Section A.1 of the supplement, we show that when $A = \Omega$,

$\hat{\Pi} = \Pi$ (up to a column permutation of $\hat{\Pi}$). The proof is similar to that in [24], except that they fixed $b = 0$ in (9) while we allow for an arbitrary b.

In the real case, A is a noisy version of Ω. The above procedure has a natural extension to the real case. When $b = 0$, it gives rise to the conventional Mixed-SCORE [24]. However, the rate-optimality of Mixed-SCORE was unknown (due to the lack of a θ-dependent lower bound in the literature). In Sections 3.2-3.3, we first use the lower bound in Theorem 2.1 to conclude with non-optimality of Mixed-SCORE under severe degree heterogeneity, and then we modify the conventional Mixed-SCORE to obtain a rate-optimal algorithm.

3.2 A theory-guided approach towards rate-optimality

The lower bound in Theorem 2.1 decomposes into node-wise contributions:

$$err_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \tau_n(\theta_i), \qquad \tau_n(\theta_i) := \min\left\{1, \quad \frac{K\sqrt{K}}{\beta_n\sqrt{n\bar{\theta}(\bar{\theta}\wedge\theta_i)}}\right\}. \tag{11}$$

Therefore, if an estimate $\hat{\Pi}$ satisfies $\|\hat{\pi}_i - \pi_i\|_1 \leq C\tau_n(\theta_i)$ simultaneously for all nodes i, then $\mathcal{L}(\hat{\Pi}, \Pi) \leq C err_n(\theta)$. In other words, "rate-optimality for arbitrary θ" reduces to the task of achieving the correct node-wise errors.

We can use (11) to check if an existing method is optimal. Let $\hat{\pi}_i^{\mathrm{MSCORE}}$ be the estimated π_i from the conventional Mixed-SCORE. [24] showed that with high probability ($\theta_{\min}$ and $\theta_{\max}$ denote the respective minimum and maximum of $\theta_1, \theta_2, \ldots, \theta_n$),

$$\|\hat{\pi}_i^{\mathrm{MSCORE}} - \pi_i\|_1 \leq C\min\left\{1, \quad \frac{K\sqrt{K}}{\beta_n\sqrt{n\bar{\theta}\theta_{\min}}}\sqrt{\frac{\theta_{\max}^3\bar{\theta}^2\log(n)}{\theta_{\min}(n^{-1}\|\theta\|^2)^2}}\right\}. \tag{12}$$

When θ_i's are at the same order (i.e., moderate degree heterogeneity), the right hand sides of (11) and (12) are the same up to a logarithmic factor. However, the two rates do not match for a general θ vector, implying non-optimality under severe degree heterogeneity.

How to improve Mixed-SCORE to achieve the node-wise error rate in (11)? Recall that the conventional Mixed-SCORE fixed $b = 0$ in (9). Our proposal is using a theory-guided choice of b. This problem is connected to many attempts in network data analysis that try to find a good normalization of the adjacency matrix. To name a few, [8] proposed to apply spectral clustering algorithms to $H^{-1}A$ (where A and H are the same as in (9)) and they justified their approach by providing a sharp concentration inequality for the spectral norm of $H^{-1}A$ under a DCBM model. [30] studied the leading eigenvectors of a graph Laplacian (a special case of (9) with $b = 1/2$ and $\tau = 0$) under the SBM model, and they showed that the Laplacian normalization can improve an eigenvector-based clustering-quality metric by a constant factor. [29] studied a spectral clustering algorithm under the DCBM model, which used the matrix in (9) with $b = 1/2$ and a constant τ, and they provided a large-deviation bound for $\|L - \mathbb{E}L\|$. However, these results are insufficient to provide insights for choosing b in our problem, because our target rate-optimality is hinged on the *node-wise errors*. The previous works focused on analyzing either the spectral norm of L [8, 29] or some quantities that aggregate all errors in a cluster [30]. Instead, we consider the Mixed-SCORE algorithm with a general value of b and study the impact of b on *node-wise errors*.

Our approach has two main components: (i) Perturbation analysis of the Mixed-SCORE algorithm, to link the node-wise error with the noise level in $\hat{r}_i$. This part is similar to the analysis in [24] and is kept brief. (ii) Sharp eigenvector analysis, to show the influence of b on the node-wise noise level in $\hat{r}_i$. In this part, a key step is deriving entry-wise large-deviation bounds for leading eigenvectors of L. Our analysis starts from the leave-one-out trick [1] and applies new strategies to handle the degree normalization in L. The analysis for $b = 1/2$ is presented in Section 5 in full detail. The analysis for a general b is discussed in Section I of the supplementary material. In this section, due to space constraint, we only describe the high-level ideas and then present the final conclusions.

Introduce $H_0 = \mathrm{diag}(\mathbb{E}d_i + \tau\mathbb{E}\bar{d})$ and $L_0 = H_0^{-b}\Omega H_0^{-b}$ as the respective population counterpart of H and L. Let λ_k be the kth largest eigenvalue (in magnitude) of L_0, and let $\xi_k \in \mathbb{R}^n$ be the corresponding eigenvector. Write $\Xi = [\xi_1, \xi_2, \ldots, \xi_K]$ and $R = [\mathrm{diag}(\xi_1)]^{-1}[\xi_2, \ldots, \xi_K]$. Denote by r_i' and $\hat{r}_i'$ the ith row of R and $\hat{R}$, respectively. If we supply R to Mixed-SCORE (see Figure (2)), we obtain $\hat{\Pi} = \Pi$. Alternatively, if we supply $\hat{R}$, we obtain a noisy estimate of Π. Mimicking the perturbation analysis in [24], we can show that under mild conditions, with overwhelming probability, there exists an $K \times K$ orthogonal matrix O such that

$$\|\hat{\pi}_i - \pi_i\|_1 \leq C\|\hat{r}_i - r_i\| \leq C\frac{\|e_i'(\hat{\Xi}O - \Xi)\|}{\xi_1(i)}, \qquad \text{simultaneously for all } 1 \leq i \leq n. \tag{13}$$

As b changes, both $\xi_1(i)$ and $\|e_i'(\hat{\Xi}O - \Xi)\|$ will change. The dependence of $\xi_1(i)$ on b follows from elementary linear algebra. The dependence of $\|e_i'(\hat{\Xi}O - \Xi)\|$ on b is difficult to study. It requires knowing row-wise large-deviation bounds for $\hat{\Xi}$ (see Section 4.1 and Section 5). A common strategy [1] is approximating each $\hat{\xi}_k$ by its first-order proxy defined as follows: Note that $L\hat{\xi}_k = \hat{\lambda}_k\hat{\xi}_k$ implies $\hat{\xi}_k(i) = \hat{\lambda}_k^{-1}e_i'L\hat{\xi}_k$. When the signal-to-noise ratio is sufficiently large, $\|\hat{\xi}_k - \xi_k\|$ and $\|H - H_0\|$ are appropriately small. It follows that $\hat{\xi}_k(i) \approx \hat{\lambda}_k^{-1}e_i'L\xi_k = \hat{\lambda}_k^{-1}e_i'H^{-b}AH^{-b}\xi_k \approx \lambda_k^{-1}e_i'H_0^{-b}AH_0^{-b}\xi_k$. We define $\hat{\xi}_k^* = \lambda_k^{-1}H_0^{-b}AH_0^{-b}\xi_k$ as a proxy to $\hat{\xi}_k$ and write $\hat{\Xi}^* = [\hat{\xi}_1^*, \hat{\xi}_2^*, \ldots, \hat{\xi}_K^*]$. Now, $\hat{\Xi}^*$ is a linear mapping of A, easier to study. We relegate detailed discussions to Section I of the supplement but only present the final results:

$$\|\hat{\pi}_i - \pi_i\|_1 \lesssim \frac{C\sqrt{\log(n)}}{\sqrt{n\theta_i\bar{\bar{\theta}}}} \cdot \delta(b, F_n), \qquad \text{where } \delta(b, F_n) = \frac{\sqrt{\int t^{3-4b}dF_n(t)}}{\int t^{2-2b}dF_n(t)}. \tag{14}$$

Given (14), we have two important observations: First, the influence of b on all n nodes is the same, through a function $\delta(b, F_n)$ that is independent of i. Therefore, we can select the best b by simply minimizing $\delta(b, F_n)$. Second, there exists a universal choice of b that minimizes $\delta(b, F_n)$ simultaneously for all $F_n(\cdot)$. To see this, we note that $\int tF_n(t) = 1$, by Definition 2.1. The Cauchy-Schwarz inequality implies $\int t^{2-2b}dF_n(t) \leq \sqrt{\int t^{3-4b}dF_n(t)}\sqrt{\int tdF_n(t)} = \sqrt{\int t^{3-4b}dF_n(t)}$, so $\delta(b, F_n)$ can never be smaller than 1. Meanwhile, if we set $b = 1/2$, $\delta(b, F_n)$ becomes exactly 1, regardless of $F_n(\cdot)$. Therefore, $b = 1/2$ is the universally best choice.

From now on, we fix $b = 1/2$. Then, L is the normalized graph Laplacian, and $\delta(b, F_n) = 1$. However, the right hand side of (14) is still slightly different from $\tau_n(\theta_i)$. In Figure 1 we have seen that $\tau_n(\theta_i)$ has three regions; in fact, (14) matches $\tau_n(\theta_i)$ only in the middle region (which is also the most subtle region). To match $\tau_n(\theta_i)$ in all three regions, we need some minor refinements of the estimation procedure, as described in Section 3.3 below.

3.3 The proposed algorithm

Let L be the normalized Laplacian matrix, corresponding to $b = 1/2$ in (9). Let $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n$ be the node embeddings as in (10). By (14), the noise level in $\hat{r}_i$ is closely related to θ_i. When θ_i is too small, we should delete such $\hat{r}_i$, as it brings in more noise than signals; additionally, we can get information of the order of θ_i from observed node degrees. To this end, given any constants $c > 0$ and $\gamma \in (0, 1)$, define:

$$\hat{S}_n(c) = \big\{i : \; d_i \geq c \cdot K\hat{\lambda}_K^{-2} \log(n)\big\}, \qquad \hat{S}_n^*(c, \gamma) = \hat{S}_n(c) \cap \big\{i : \; d_i \geq \gamma \cdot \bar{d}\big\}. \tag{15}$$

Now, all nodes partition into three non-overlapping subsets: $M_1 = \hat{S}_n^*(c, \gamma)$, which consists of high-degree and moderate-degree nodes, $M_2 = \hat{S}_n(c) \setminus \hat{S}_n^*(c, \gamma)$, which consists of low-degree nodes, and $M_3 = \{1, 2, \ldots, n\} \setminus \hat{S}_n(c)$, consisting of extremely low-degree nodes. For nodes in M_3, their $\hat{r}_i$'s are too noisy to contain useful information of π_i. We delete all such nodes from the embedded point cloud. For nodes in M_2, we exclude them from the vertex hunting step. From Figure 2, we can see that not all $\hat{r}_i$'s are needed for estimating the simplex. We apply a vertex hunting algorithm (e.g., [5]) on those $\hat{r}_i$'s of nodes in M_1. After vertex hunting, we estimate π_i from $\hat{r}_i$, for all nodes in $M_1 \cup M_2$. See Algorithm 1. The pseudo-code, together with details of the vertex hunting algorithm, is in Section B of the supplementary material.

Table 1: A summary of node trimming. All nodes are used to obtain $\hat{R}$. All nodes receive an estimate $\hat{\pi}_i$, but for extremely low-degree nodes, we assign a trivial $\hat{\pi}_i$ rather than estimating it. The three cases here correspond to the three regions in Figure 1.

Subset	Degree range	Node embedding	Vertex hunting	Non-trivial $\hat{\pi}_i$
M_1	$\big[\gamma\bar{d},\ n\big]$	yes	yes	yes
M_2	$\big[cK\hat{\lambda}_K^{-2}\log(n),\ \gamma\bar{d}\big]$	yes	no	yes
M_3	$\big[0,\ cK\hat{\lambda}_K^{-2}\log(n)\big]$	yes	no	no

4 Entywise eigenvector analysis, node-wise errors, and rate-optimality

We present the theoretical properties of Mixed-SCORE-Laplacian (MSL). The backbone of our analysis is an entry-wise large-deviation bound for leading eigenvectors of the normalized

```
Algorithm 1 Mixed-SCORE-Laplaccian (a high-level description).
Input: $K$, $A$, and tuning parameters $\tau = 1$, $c = 0.5$, and $\gamma = 0.05$ (default values).
  1. Node Embedding: Fix $b = 1/2$ in (9) and obtain $\hat{R} = [\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n]'$ as in (10).
  2. Trimming: Let $\hat{S}_n(c)$ be as in (15). For any $i \notin \hat{S}_n(c)$, set $\hat{\pi}_i = K^{-1}\mathbf{1}_K$ and delete such nodes from all the remaining steps.
  3. (De-noised) Vertex Hunting: Let $\hat{S}^*_n(c, \gamma)$ be as in (15). Run the successive projection algorithm [5] on $\{\hat{r}_i : i \in \hat{S}^*_n(c, \gamma)\}$ to obtain $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$.
  4. Membership Estimation: Obtain $\hat{b}_1 \in \mathbb{R}^K$ by $\hat{b}_1(k) = [\hat{\lambda}_1 + \hat{v}_k'\mathrm{diag}(\hat{\lambda}_2, \ldots, \hat{\lambda}_K)\hat{v}_k]^{-1/2}$. For each $i \in \hat{S}_n(c)$, solve its barycentric coordinate $\hat{w}_i \in \mathbb{R}^K$. Let $\hat{\pi}^*_i \in \mathbb{R}^K$ be such that $\hat{\pi}^*_i(k) = \max\{\hat{w}_i(k)/\hat{b}_1(k), 0\}$. Obtain $\hat{\pi}_i$ by normalizing $\hat{\pi}^*_i$ to have a unit $\ell^1$-norm.
Output: $\hat{\Pi}$.
```

graph Laplacian, which is presented in Section 4.1. The node-wise errors and rate-optimality are in Section 4.2. We also extend all results to a more general loss function in Section 4.3.

4.1 Entry-wise eigenvector analysis for graph Laplacian

Fix $b = 1/2$ and $\tau = 1$ in the definition of L in (9). For $1 \leq k \leq K$, let $\hat{\lambda}_k$ be the kth largest eigenvalue (in magnitude) of L, and let $\hat{\xi}_k$ be the associated eigenvector. Define

$$L_0 = H_0^{-1/2}\Omega H_0^{-1/2}, \qquad \text{where} \quad H_0 = \mathbb{E}[H]. \tag{16}$$

For $1 \leq k \leq K$, let λ_k be the kth largest eigenvalue (in magnitude) of L_0, and let ξ_k be the corresponding eigenvector. Write $\Xi_1 = [\xi_2, \ldots, \xi_K]$, $\Xi = [\xi_1, \Xi_1]$, $\hat{\Xi}_1 = [\hat{\xi}_2, \ldots, \hat{\xi}_K]$, and $\hat{\Xi} = [\hat{\xi}_1, \hat{\Xi}_1]$. By Condition 2.1(b), there is a gap between λ_1 and the other eigenvalues; hence, by sin-theta theorem [11], we can obtain upper bounds for $\|\hat{\xi}_1 - \xi_1\|$ and $\min_{O_1} \|\hat{\Xi}_1 O_1 - \Xi_1\|_F$, where the minimum is taken over all orthogonal matrices $O_1 \in \mathbb{R}^{(K-1)\times(K-1)}$. However, these bounds are insufficient for controlling node-wise errors of MSL. We now bound each entry of $(\hat{\xi}_1 - \xi_1)$ and each row of $(\hat{\Xi}_1 O_1 - \Xi_1)$. Let $e_i \in \mathbb{R}^n$ denote the ith standard basis of $\mathbb{R}^n$.

Theorem 4.1 (Row-wise large-deviation bounds for $\hat{\Xi}$). *Consider the DCMM model (1)-(2), where Condition 2.1(a)-(c) are satisfied. Suppose $K^3 \log(n)/(n\bar{\theta}^2\beta_n^2) \to 0$ as $n \to \infty$. With* *probability $1-o(n^{-3})$, there exists $\omega \in \{1,-1\}$ and an orthogonal matrix $O_1 \in \mathbb{R}^{(K-1)\times(K-1)}$ such that simultaneously for all $1 \leq i \leq n$,*

$$|\omega\hat{\xi}_1(i) - \xi_1(i)| \leq C\sqrt{\frac{K\theta_i\log(n)}{n^2\bar{\theta}^3}}\left(1+\sqrt{\frac{\log(n)}{n\bar{\theta}\theta_i}}\right), \tag{17}$$

$$\|e_i'(\hat{\Xi}_1 O_1 - \Xi_1)\| \leq C\sqrt{\frac{K^3\theta_i\log(n)}{n^2\bar{\theta}^3\beta_n^2}}\left(1+\sqrt{\frac{\log(n)}{n\bar{\theta}\theta_i}}\right). \tag{18}$$

Theorem 4.1 is one of our major technical contributions. We sketch the proof ideas in Section 5 and relegate the full proof to the supplementary material. In this theorem, we make an assumption: $K^3\log(n)/(n\bar{\theta}^2\beta_n^2) \to 0$. In a simple case where K is finite and $\beta_n^{-1} = O(1)$, this assumption reduces to $n\bar{\theta}^2 \gg \log(n)$. It means the average degree only needs to grow faster than $\log(n)$. Hence, Theorem 4.1 can handle sparse networks up to logarithmic degree regime. Similarly, all other theorems in this section cover sparse networks.

4.2 Node-wise errors and rate-optimality

We apply Theorem 4.1 to establish the theoretical properties of MSL. First, we give a large-deviation bound for each row of $\hat{R}$. Define $R \in \mathbb{R}^{n\times(K-1)}$ by

$$R(i,k) = \xi_{k+1}(i)/\xi_1(i), \qquad 1 \leq i \leq n, 1 \leq k \leq K-1. \tag{19}$$

Let $r_1', r_2', \ldots, r_n'$ be the rows of R. It is seen that $\hat{r}_i' = [\hat{\xi}_1(i)]^{-1}\cdot e_i'\hat{\Xi}_1$ and $r_i' = [\xi_1(i)]^{-1}\cdot e_i'\Xi_1$. Therefore, we can control the difference between $\hat{r}_i$ and r_i by applying Theorem 4.1:

Lemma 4.1. *Consider the DCMM model in (1)-(2), where Condition 2.1(a)-(c) are satisfied. Fix $b = 1/2$ and $\tau = 1$ in (9). Suppose $K^3\log(n)/(n\bar{\theta}^2\beta_n^2) \to 0$ as $n \to \infty$. For any constant $c_0 > 0$, define $S_n(c_0) := \{1 \leq i \leq n :\ n\bar{\theta}\theta_i\beta_n^2 \geq c_0K^3\log(n)\}$. With probability $1-o(n^{-3})$, there exists an orthogonal matrix $O_1 \in \mathbb{R}^{K-1,K-1}$ such that, simultaneously for $i \in S_n(c_0)$,*

$$\|O_1'\hat{r}_i - r_i\| \leq C\sqrt{\frac{K^3\log(n)}{n\bar{\theta}(\bar{\theta}\wedge\theta_i)\beta_n^2}}. \tag{20}$$

The statement holds for any $c_0 > 0$, except that the constant C will depend on c_0.

Lemma 4.1 only considers nodes in $S_n(c_0)$. For a node $i \notin S_n(c_0)$, its degree is so small that $\hat{r}_i$ is too noisy to contain useful information of π_i. In MSL, the set $S_n(c_0)$ is estimated by $\hat{S}_n(c)$, and those $\hat{r}_i$'s with $i \notin \hat{S}_n(c)$ are discarded.

Next, we present the node-wise error of MLS.

Theorem 4.2 (Node-wise errors and rate-optimality). *Consider the DCMM model in* (1)-(2)*, where Condition 2.1(a)-(c) hold, and additionally, Condition 2.1(d) is satisfied. Fix $b = 1/2$ and $\tau = 1$ in* (9)*. Suppose $K^3 \log(n)/(n\bar{\theta}^2\beta_n^2) \to 0$ as $n \to \infty$. Let $\hat{\Pi}$ be the output of Algorithm 1, in which (c, γ) are positive constants such that $\gamma < c_4$ (c_4 is the same as in Condition 2.1(d)). With probability $1-o(n^{-3})$, there exists a permutation T on $\{1, 2, \dots, K\}$, such that simultaneously for all $1 \leq i \leq n$,*

$$\|T\hat{\pi}_i - \pi_i\|_1 \leq C \min\left\{\sqrt{\frac{K^3 \log(n)}{n\bar{\theta}(\bar{\theta} \wedge \theta_i)\beta_n^2}},\ 1\right\}, \tag{21}$$

In addition, let $\mathcal{L}(\hat{\Pi}, \Pi)$ be the ℓ^1-loss in (5)*, and $err_n(\theta)$ be as in* (8)*. Then,*

$$\mathbb{E}\mathcal{L}(\hat{\Pi}, \Pi) \leq C err_n(\theta)\sqrt{\log(n)}.$$

Up to a logarithmic factor, the node-wise error in (21) matches with the right hand side of (11) for every θ_i, and the optimality of MSL follows immediately.

An interesting question is if we can further remove the $\sqrt{\log(n)}$-factor in the upper bound. This factor arises from the entry-wise eigenvector analysis in Section 4.1. We conjecture that it is not easy to remove, due to the nature of the leave-one-out trick in eigenvector analysis (see Section 5). Another question is if the requirement $K^3 \log(n)/(n\bar{\theta}^2\beta_n^2) \to 0$ can be further relaxed to allow for bounded degrees. We conjecture that it is impossible. In the community detection literature, there are many results for the bounded-degree regime. However, mixed membership estimation is a more challenging task than community detection. To our best knowledge, there has been no result for the bounded-degree regime.

Theorem 4.2 not only shows the optimality of MSL under a global loss but also provides insights on local errors. For example, if we only want to estimate one particular π_i, should we pre-process the network by removing low-degree nodes or grouping nodes of similar degrees? This is answered by (21). Suppose i is a moderate-degree node, with $\theta_i = O(\bar{\theta})$ and $n\bar{\theta}\theta_i\beta_n^2 \gg K^3\log(n)$. Then, (21) becomes $\|T\hat{\pi}_i - \pi_i\|_1 \leq C(n\bar{\theta}\theta_i\beta_n^2)^{-1/2}\sqrt{K^3\log(n)}$. The total effect of other nodes at $\hat{\pi}_i$ is carried by $n\bar{\theta} = \theta_i + \sum_{j:j\neq i}\theta_j$. As a result, to achieve the best accuracy of $\hat{\pi}_i$, we should not remove any node or group nodes of similar degrees. [3]

4.3 Extension to other loss functions

We extend the results in Section 2 and Section 4.2 to a general loss function:

$$\mathcal{L}(\hat{\Pi}, \Pi; p, q) = \min_T \Big\{ \Big(\frac{1}{n}\sum_{i=1}^n (\theta_i/\bar{\theta})^p \|T\hat{\pi}_i - \pi_i\|_q^q \Big)^{1/q} \Big\}, \quad \text{for } p \geq 0 \text{ and } q \geq 1. \tag{22}$$

The ℓ^1-loss in (5) is a special case with $p = 0$ and $q = 1$. When $p > 0$, the estimation errors are re-weighted by degree parameters. In many real applications, the π_i of high-degree nodes are more interesting, so this general loss metric is relevant. We are particularly interested in the case of $p = 1/2$ and $q = 1$. We write $\mathcal{L}^w(\hat{\Pi}, \Pi) := \mathcal{L}(\hat{\Pi}, \Pi; 1/2, 1)$ and call it the *weighted ℓ^1-loss*. Same as before, let err_n denote the baseline rate in (7)

Corollary 4.1. *Suppose conditions of Theorem 4.2 hold. Let $\hat{\Pi}$ be the estimator from Mixed-SCORE-Laplacian and $\mathcal{L}(\hat{\Pi}, \Pi; p, q)$ be the loss metric in (22), for $p \geq 0$ and $q \geq 1$. Then,*

$$\mathcal{L}(\hat{\Pi}, \Pi; p, q) \leq C\sqrt{\log(n)} \Big(\int t^p \min\Big\{ \frac{err_n^q}{(t \wedge 1)^{q/2}}, 1 \Big\} dF_n(t) \Big)^{1/q}. \tag{23}$$

Furthermore, in the special case of $p = 1/2$ and $q = 1$, $\mathbb{E}\mathcal{L}^w(\hat{\Pi}, \Pi) \leq C\sqrt{\log(n)}\, err_n$.

For the loss metric $\mathcal{L}^w(\hat{\Pi}, \Pi)$, we provide a matching lower bound as follows. It suggests that Mixed-SCORE-Laplacian is still rate-optimal (for arbitrary θ) under this loss metric.

Theorem 4.3. *Fix constants c_1-c_4. Given (n, K) and $\theta \in \mathbb{R}^n$ such that $F_n(err_n^2) \leq \check{c}$, for a constant $\check{c} \in (0, 1)$. Let $\mathcal{Q}_n(\theta)$ be the collection of (Π, P) satisfying Condition 2.1. There is a constant $C > 0$ such that, for all sufficiently large n, $\inf_{\hat{\Pi}} \sup_{(\Pi,P)\in\mathcal{Q}_n(\theta)} \mathbb{E}\mathcal{L}^w(\hat{\Pi}, \Pi) \geq Cerr_n$.*

[3] This argument only says that we should use *all* nodes to compute $\hat{R}$, but it does not imply we should use all rows of $\hat{R}$ in the subsequent steps. There is no conflict with the node trimming strategy in Section 3.3.

5 Proof of Theorem 4.1: Challenges, and our strategy

Theorem 4.1 is connected to the literature of "entry-wise eigenvector analysis" for random graphs [1, 9, 12, 13, 24, 28, 31] but has significant differences. Most existing works (a) study eigenvectors of the adjacency matrix but not the normalized Laplacian, (b) consider a network model that has no degree heterogeneity or only mild degree heterogeneity ($\theta_{\min} \asymp \theta_{\max}$), and (c) derive the "2-to-infinity" bounds, i.e., upper bounds for $\max_i |\omega\hat{\xi}_1(i) - \xi_1(i)|$ or $\max_i \|e_i'(\hat{\Xi}_1 O_1 - \Xi_1)\|$, instead of obtaining different bounds for different i. As a result, Theorem 4.1 cannot be deduced from any of the existing results. In what follows, Section 5.1 reviews the leave-one-out strategy for entry-wise eigenvector analysis, Section 5.2 explains the technical challenges, and Section 5.3 contains a proof sketch of Theorem 4.1.

5.1 A review of the leave-one-out approach

Abbe et al. [1] introduced the leave-one-out trick to study the eigenvectors of the adjacency matrix A when the network follows an SBM model (a special case of DCMM when θ_i's are equal to each other and when π_i's are all degenerate). Let $\hat{\lambda}_k^*$ be the kth largest eigenvalue (in magnitude) of A and let $\hat{\xi}_k^* \in \mathbb{R}^n$ be the corresponding eigenvector. Let λ_k^* be the kth largest eigenvalue (in magnitude) of Ω and let $\xi_k^* \in \mathbb{R}^n$ be the associated eigenvector. Write $\hat{\Xi}^* = [\hat{\xi}_1^*, \ldots, \hat{\xi}_K^*]$ and $\Xi^* = [\xi_1^*, \ldots, \xi_K^*]$. Let $\|\cdot\|_{2\to\infty}$ denote the maximum row-wise ℓ^2-norm of a matrix. [1] derived a large-deviation bound for $\min_O \|\hat{\Xi}^* O - \Xi^*\|_{2\to\infty}$, where the minimum is over $K \times K$ orthogonal matrices. To explain their proof strategies, we consider $K = 1$. Their goal reduced to obtaining a bound for $\|\hat{\xi}_1^* - \xi_1^*\|_\infty$, up to a sign flip of $\hat{\xi}_1$. By definition, $A\hat{\xi}_1^* = \hat{\lambda}_1^*\hat{\xi}_1^*$. It is seen that

$$\hat{\xi}_1^*(i) = (1/\hat{\lambda}_1^*)\, e_i' A\hat{\xi}_1^* = (1/\hat{\lambda}_1^*) \sum_{j: j\neq i} \hat{\xi}_1^*(j) A(i,j). \tag{24}$$

If $\hat{\xi}_1^*$ is independent of the ith row of A, then the right hand side is nothing but a weighted sum of independent Bernoulli variables and can be easily analyzed. However, $\hat{\xi}_1^*$ is dependent of the ith row of A. The leave-one-out trick creates a proxy of $\hat{\xi}_1^*$ that is independent of the ith row of A. Fix i and let $\tilde{\xi}_1^*$ be the leading eigenvector of the matrix $\tilde{A}$, obtained by setting the ith row and the ith column of A to zero. If $\tilde{\xi}_1^*$ is sufficiently close to $\hat{\xi}_1^*$, then

$$\hat{\xi}_1^*(i) \approx (1/\hat{\lambda}_1^*) \sum_{j:j\neq i} \tilde{\xi}_1^*(j) A(i,j). \tag{25}$$

Since $\tilde{\xi}_1^*$ is independent of the ith row of A, the right hand side of (25) is now easy to analyze. Although this leave-one-out trick is easy to describe, it does require tedious analysis to control the "approximation" in (25). [1] proposed to control the difference between the left and right hand sides of (25) using $\|\hat{\xi}_1^* - \tilde{\xi}_1^*\|$ and then apply the sin-theta theorem to bound $\|\hat{\xi}_1^* - \tilde{\xi}_1^*\|$.

5.2 Challenges in proving Theorem 4.1 and how to overcome them

Suppose we want to extend the analysis to study eigenvectors of the normalized Laplacian matrix L, assuming that the network follows a DCMM model. Recall that $L = H^{-\frac{1}{2}} A H^{-\frac{1}{2}}$ and $L_0 = H_0^{-\frac{1}{2}} \Omega H_0^{-\frac{1}{2}}$, where H is a diagonal matrix constructed from degrees. The first challenge is that entries in the upper triangle of L are no longer independent of each other. If we mimick [1] to define a proxy of $\hat{\xi}_1$ as the leading eigenvector of the matrix obtained by setting the ith row & column of L to zero, this vector is still dependent of the ith row of A.

Our solution is to create a proxy to L such that the effect of the ith row of A is removed. Recall that $W = A - \mathbb{E}[A]$. Let $W^{(i)}$ be the matrix obtained by setting the i-th row/column of W to zero. Introduce $A^{(i)} = \Omega - \mathrm{diag}(\Omega) + W^{(i)}$ and $\tilde{H}^{(i)} = \mathrm{diag}\big(A^{(i)} \mathbf{1}_n\big) + n^{-1}(\mathbf{1}_n' A^{(i)} \mathbf{1}_n) I_n$. It can be shown that $(A^{(i)}, \tilde{H}^{(i)})$ are independent of the ith row of A, and they are appropriately close to (A, H). Define

$$\overline{L}^{(i)} := (\tilde{H}^{(i)})^{-\frac{1}{2}} A^{(i)} (\tilde{H}^{(i)})^{-\frac{1}{2}}. \tag{26}$$

Let $(\overline{\lambda}_1^{(i)}, \overline{\xi}_1^{(i)})$ be the first eigen-pair of $\overline{L}^{(i)}$. Then, $\overline{\xi}_1^{(i)}$ is the proxy of $\hat{\xi}_1$ we seek for.

Given $\overline{\xi}_1^{(i)}$, is it easy to study $\hat{\xi}_1(i) - \xi_1(i)$ by mimicking the analysis in [1]? Unfortunately, the answer is no. Since $\hat{\xi}_1 = (1/\hat{\lambda}_1)H^{-1/2}AH^{-1/2}\hat{\xi}_1$, we can re-write $\hat{\xi}_1(i)$ as

$$\hat{\xi}_1(i) = \frac{1}{\hat{\lambda}_1\sqrt{H(i,i)}}\left[\sum_{j:j\neq i}\frac{\overline{\xi}_1^{(i)}(j)}{\sqrt{H(j,j)}}A(i,j) + \overline{e}_i\right], \tag{27}$$

where $\overline{e}_i$ is the approximation error to $(1/\hat{\lambda}_1)e_i'H^{-1/2}AH^{-1/2}\hat{\xi}_1$ by replacing $\hat{\xi}_1$ by $\overline{\xi}_1^{(i)}$ in this expression. To analyze the right hand side of (27), we face two more challenges:

- On the right hand side of (27), $H(j,j)$ is still dependent of the ith row of A. We may replace it with $\tilde{H}^{(i)}(j,j)$. However, this replacement affects every term in the sum, and its effect is not easy to control. Even if we can control it, we are still unable to show that the resulting quantity is close enough to $\xi_1(i)$, because the effect of $|\tilde{H}^{(i)}(j,j) - H_0(j,j)|$ is still non-negligble.

- The control of $\overline{e}_i$ is much more challenging than similar steps in [1], for three reasons: First, since [1] studies the eigenvectors of A instead of those of L, the proxy eigenvector is defined in a more straightforward way; consequently, the proxy error has a simpler form. Second, in the setting of [1], the network model has no degree heterogeneity, and the target bound for each entry of $\hat{\xi}_1^*$ is at the same order; then the desirable bound for $\overline{e}_i$ is also the same for all i. However, in our setting, the desirable bound for $\overline{e}_i$ may be significantly different for different i, so we must have *better* control of each $\overline{e}_i$. Third, the idea in [1] for controlling $\overline{e}_i$ is based on studying the ℓ^2-error between the empirical eigenvector and its proxy. However, for our problem, it is impossible to control $\overline{e}_i$ from studying $\|\hat{\xi}_1 - \overline{\xi}_1^{(i)}\|$, because the entries of $\hat{\xi}_1 - \overline{\xi}_1^{(i)}$ can be at different orders due to degree heterogeneity, and analysis based on $\|\hat{\xi}_1 - \overline{\xi}_1^{(i)}\|$ is not sharp.

To overcome the first challenge, we introduce another proxy eigenvector. Let $\tilde{H}^{(i)}$ be the same as in (26). Let $(\tilde{\lambda}_1^{(i)}, \tilde{\xi}_1^{(i)})$ be the first eigen-pair of the following matrix:

$$\tilde{L}^{(i)} := (\tilde{H}^{(i)})^{-\frac{1}{2}}\Omega(\tilde{H}^{(i)})^{-\frac{1}{2}}. \tag{28}$$

Our idea is to use $\tilde{\xi}_1^{(i)}$ as a proxy to ξ_1, and study $|\hat{\xi}_1(i) - \tilde{\xi}_1^{(i)}(i)|$ rather than $|\hat{\xi}_1(i) - \xi_1(i)|$. In the first bullet point above, we have mentioned that the effect of $|\tilde{H}^{(i)}(j,j) - H_0(j,j)|$ is nonnegligible when we try to bound $|\hat{\xi}_1(i)-\xi_1(i)|$ directly. Comparing $\tilde{L}^{(i)}$ and L_0, the difference is that H_0 is replaced by $\tilde{H}^{(i)}$. Hence, when we bound $|\hat{\xi}_1(i)-\tilde{\xi}_1^{(i)}(i)|$, there is no longer any term caused by $|\tilde{H}^{(i)}(j,j)-H_0(j,j)|$; and the issue is partially resolved. It still remains to bound $|\tilde{\xi}_1^{(i)}(i)-\xi_1(i)|$. This seems to involve $|\tilde{H}^{(i)}(j,j)-H_0(j,j)|$ again. Fortunately, since L_0 and $\tilde{L}^{(i)}$ are both low-rank matrices, we can study the difference between $\tilde{\xi}_1^{(i)}$ and ξ_1 in a different way to avoid using $|\tilde{H}^{(i)}(j,j)-H_0(j,j)|$ explicitly (see Lemma 5.1 below).

To overcome the second challenge, we notice that under severe degree heterogeneity, the noise level is different at different entries of $\hat{\xi}_1-\overline{\xi}_1^{(i)}$, and the same happens for $\overline{\xi}_1^{(i)}-\tilde{\xi}_1^{(i)}$. The key question is how to "track" such heterogeneous noise levels in every step of our analysis of $\overline{e}_i$ to ensure that the resulting bounds are good for all $1\leq i\leq n$. We map out the analysis by two key technical lemmas, Lemmas 5.2-5.3. Recall that we focus on studying $|\hat{\xi}_1(i)-\tilde{\xi}_1^{(i)}(i)|$ (see the previous paragraph). These two lemmas together establish an inequality:

$$\begin{aligned}|\hat{\xi}_1(i)-\tilde{\xi}_1^{(i)}(i)| &\leq c_{1n}(\theta_i)+c_{2n}(\theta_i)\cdot\|(\tilde{H}^{(i)})^{-\frac{1}{2}}(\hat{\xi}_1-\tilde{\xi}_1^{(i)})\|_\infty && (29)\\ &\leq c_{1n}(\theta_i)+c_{2n}(\theta_i)\cdot\|(\tilde{H}^{(i)})^{-\frac{1}{2}}H_0^{\frac{1}{2}}\|\cdot\|H_0^{-\frac{1}{2}}(\hat{\xi}_1-\tilde{\xi}_1^{(i)})\|_\infty, && (30)\end{aligned}$$

where $c_{1n}(\theta_i)$ and $c_{2n}(\theta_i)$ are some explicit sequences. We note that (30) follows immediately from (29). Given (30), we first multiply $H_0^{-\frac{1}{2}}(i,i)$ on both hand sides to obtain a bound for $\|H_0^{-\frac{1}{2}}(\hat{\xi}_1-\tilde{\xi}_1^{(i)})\|_\infty$ and then plug this bound into (30) again to get the bound for $|\hat{\xi}_1(i)-\tilde{\xi}_1^{(i)}(i)|$. The non-trivial part is deriving (29) — the backbone of our leave-one-out analysis. The proof of (29) "tracks" the heterogeneous noise levels by repeatedly using the idea of "re-weighting": For example, Lemma 5.2 introduces a re-weighted noise matrix $\Delta^{(i)}:=(\tilde{H}^{(i)})^{-\frac{1}{2}}WH^{-\frac{1}{2}}$, and Lemma 5.3 controls every term by using $(\tilde{H}^{(i)})^{-\frac{1}{2}}(\hat{\xi}_1-\tilde{\xi}_1^{(i)})$ rather than the unweighted vector $\hat{\xi}_1-\tilde{\xi}_1^{(i)}$. In fact, in the proof of these lemmas, we have properly "re-weighted" almost every intermediate vector and matrix by either $H^{-\frac{1}{2}}$, or $H_0^{-\frac{1}{2}}$, or $(\tilde{H}^{(i)})^{-\frac{1}{2}}$.

5.3 Proof sketch of Theorem 4.1

We prove the claim about $\hat{\xi}_1$ in Theorem 4.1. The proof of the claim for $\hat{\Xi}_1$ follows the same vein but is slightly more complicated, which is relegated to the supplementary material.

Let $\bar{\xi}_1^{(i)}$ and $\tilde{\xi}_1^{(i)}$ be as defined in (26) and (28), respectively. As mentioned, $\bar{\xi}_1^{(i)}$ is a proxy to $\hat{\xi}_1$, and $\tilde{\xi}_1^{(i)}$ is a proxy to ξ_1. The next lemma controls the difference between ξ_1 and $\tilde{\xi}_1^{(i)}$.

Lemma 5.1. *Suppose the conditions of Theorem 4.1 hold. We pick the sign of ξ_1 such that $\xi_1(1) \geq 0$. For each $1 \leq i \leq n$, we pick the sign of $\tilde{\xi}_1^{(i)}$ such that $\tilde{\xi}_1^{(i)}(1) \geq 0$. With probability $1 - o(n^{-3})$, simultaneously for all $1 \leq i, j \leq n$,*

$$|\tilde{\xi}_1^{(i)}(j) - \xi_1(j)| \leq CK^{\frac{1}{2}}\kappa_j \cdot \Big(\sqrt{\frac{\theta_j}{\bar{\theta}}} \wedge 1\Big), \qquad \textit{where} \quad \kappa_j := \sqrt{\frac{\log(n)}{n\bar{\theta}^2}}\sqrt{\frac{\theta_j}{n\bar{\theta}}}. \tag{31}$$

Next, we study the difference between $\tilde{\xi}_1^{(i)}(i)$ and $\hat{\xi}_1(i)$. We use $\bar{\xi}_1^{(i)}$ as a bridge quantity. The following lemma is proved in the supplementary material:

Lemma 5.2. *Suppose the conditions of Theorem 4.1 hold. For $1 \leq i \leq n$, let κ_i be as in (31), and define $\Delta = \Delta^{(i)} := (\tilde{H}^{(i)})^{-\frac{1}{2}}WH^{-\frac{1}{2}}$. We pick the signs of ξ_1 and $\tilde{\xi}_1^{(i)}$ in the same way as in Lemma 5.1, and we pick the sign of $\bar{\xi}_1^{(i)}$ such that $\mathrm{sgn}((\tilde{\xi}_1^{(i)})'\bar{\xi}_1^{(i)}) = 1$. With probability $1 - o(n^{-3})$, there exists some $w \in \{1, -1\}$ such that, simultaneously for all $1 \leq i \leq n$,*

$$|w\hat{\xi}_1(i) - \tilde{\xi}_1^{(i)}(i)| \leq C\sqrt{K}\kappa_i + C|e_i'\Delta\hat{\xi}_1|, \tag{32}$$

$$|e_i'\Delta\hat{\xi}_1| \leq |e_i'\Delta\tilde{\xi}_1^{(i)}| + |e_i'\Delta(\bar{\xi}_1^{(i)} - \tilde{\xi}_1^{(i)})| + \frac{C}{\sqrt{n\bar{\theta}^2}}\|w\hat{\xi}_1 - \bar{\xi}_1^{(i)}\|. \tag{33}$$

Write $\bar{\xi}_1^{(i)} = \bar{\xi}_1$, $\tilde{\xi}_1^{(i)} = \tilde{\xi}_1$ and $\tilde{H}^{(i)} = \tilde{H}$ for short. The role of Lemma 5.2 is to reduce the analysis of $|\hat{\xi}_1(i) - \tilde{\xi}_1(i)|$ to the analyses of $|e_i'\Delta\tilde{\xi}_1|$, $|e_i'\Delta(\bar{\xi}_1 - \tilde{\xi}_1)|$ and $\|\hat{\xi}_1 - \bar{\xi}_1\|$. Among the three quantities, the first two are relatively easy to bound, because $e_i'\Delta \approx e_i'H_0^{-1/2}WH_0^{-1/2}$, which is independent of $(\bar{\xi}_1, \tilde{\xi}_1)$; To control the third term $\|\hat{\xi}_1 - \bar{\xi}_1\|$, we need the following lemma, whose proof can be found in the supplementary material:

Lemma 5.3. *Under the assumptions in Lemma 5.2, with probability $1 - o(n^{-3})$, for the same $w \in \{1, -1\}$ in Lemma 5.2, simultaneously for all $1 \leq i \leq n$,*

$$|e_i'\Delta\tilde{\xi}_1^{(i)}| \leq C\widetilde{\kappa}_i, \qquad \textit{where} \quad \widetilde{\kappa}_i := \frac{1}{n\bar{\theta}}\sqrt{\frac{\log(n)}{n\bar{\theta}^2}}\sqrt{n\bar{\theta}\theta_i \vee \log(n)}, \tag{34}$$

$$|e_i\Delta(\bar{\xi}_1^{(i)} - \tilde{\xi}_1^{(i)})| \leq C\widetilde{\kappa}_i\Big(1 + n\bar{\theta}\big\|(\tilde{H}^{(i)})^{-\frac{1}{2}}(w\hat{\xi}_1 - \tilde{\xi}_1^{(i)})\big\|_\infty\Big) + \frac{C\log(n)}{n\bar{\theta}^2}\|w\hat{\xi}_1 - \bar{\xi}_1^{(i)}\|, \tag{35}$$

$$\|w\hat{\xi}_1 - \bar{\xi}_1^{(i)}\| \leq C\,\widetilde{\kappa}_i\Big(1 + n\bar{\theta}\|(\tilde{H}^{(i)})^{-\frac{1}{2}}(w\hat{\xi}_1 - \tilde{\xi}_1^{(i)})\|_\infty\Big) + \frac{C}{\sqrt{n\bar{\theta}^2}}|w\hat{\xi}_1(i) - \tilde{\xi}_1^{(i)}(i)|. \tag{36}$$

We now use Lemmas 5.1-5.3 to prove the first claim, (17), in Theorem 4.1. In Lemma 5.2, although the vector $\tilde{\xi}_1^{(i)}$ depends on i, the scalar $w \in \{\pm 1\}$ is shared by $1 \le i \le n$; similarly, the w in Lemma 5.3 is also shared by all $1 \le i \le n$. Therefore, we can assume $w = 1$ in all claims, without loss of generality. When there is no confusion, we write $\bar{\xi}_1^{(i)} = \bar{\xi}_1$, $\tilde{\xi}_1^{(i)} = \tilde{\xi}_1$ and $\tilde{H}^{(i)} = \tilde{H}$ for short. We plug (34)-(36) into (32)-(33) and note that $\kappa_i \le \widetilde{\kappa}_i$. It gives

$$|\hat{\xi}_1(i) - \tilde{\xi}_1(i)| \le C\sqrt{K}\,\widetilde{\kappa}_i + C\widetilde{\kappa}_i n\bar{\theta}\|\tilde{H}^{-\frac{1}{2}}(\hat{\xi}_1 - \tilde{\xi}_1)\|_\infty + \frac{C\sqrt{\log(n)}}{n\bar{\theta}^2}|\hat{\xi}_1(i) - \tilde{\xi}_1(i)|.$$

Since $\sqrt{\log(n)} \ll n\bar{\theta}^2$, we can rearrange the above inequality to get

$$|\hat{\xi}_1(i) - \tilde{\xi}_1(i)| \le C\sqrt{K}\,\widetilde{\kappa}_i + C\widetilde{\kappa}_i n\bar{\theta}\|\tilde{H}^{-\frac{1}{2}}(\hat{\xi}_1 - \tilde{\xi}_1)\|_\infty. \tag{37}$$

We further apply Lemma 5.1 and note that $\kappa_j \le \widetilde{\kappa}_j$ for all j and that with high probability, $\|\tilde{H}^{-\frac{1}{2}}H_0^{\frac{1}{2}}\| \le C$ (see the supplementary material). It yields that $|\hat{\xi}_1(i) - \xi_1(i)| \le |\hat{\xi}_1(i) - \tilde{\xi}_1(i)| + |\tilde{\xi}_1(i) - \xi_1(i)| \le C\sqrt{K}\,\widetilde{\kappa}_i + C\widetilde{\kappa}_i n\bar{\theta}\|H_0^{-\frac{1}{2}}(\hat{\xi}_1 - \tilde{\xi}_1)\|_\infty$, which leads to

$$|\hat{\xi}_1(i) - \xi_1(i)| \le C\sqrt{K}\,\widetilde{\kappa}_i + C\widetilde{\kappa}_i n\bar{\theta}\big(\|H_0^{-\frac{1}{2}}(\hat{\xi}_1 - \xi_1)\|_\infty + \|H_0^{-\frac{1}{2}}(\tilde{\xi}_1 - \xi_1)\|_\infty\big).$$

To bound the second term in the bracket, we apply Lemma 5.1 again and note that $H_0(j,j) \asymp n\bar{\theta}(\bar{\theta} \vee \theta_j)$ (see Section D.1 of the supplementary material) and $\kappa_j[H_0(j,j)]^{-\frac{1}{2}} \le \widetilde{\kappa}_j[H_0(j,j)]^{-\frac{1}{2}} \le \sqrt{\frac{\log(n)}{n\bar{\theta}^2}}\frac{1}{n\bar{\theta}}$. It gives $\|H_0^{-\frac{1}{2}}(\tilde{\xi}_1 - \xi_1)\|_\infty \le C\sqrt{\frac{K\log(n)}{n\bar{\theta}^2}}\frac{1}{n\bar{\theta}}$. Plugging it into the above inequality, we obtain that

$$|\hat{\xi}_1(i) - \xi_1(i)| \le C\sqrt{K}\,\widetilde{\kappa}_i + C\widetilde{\kappa}_i n\bar{\theta}\|H_0^{-\frac{1}{2}}(\hat{\xi}_1 - \xi_1)\|_\infty. \tag{38}$$

We multiply $H_0^{-\frac{1}{2}}(i,i)$ on both hand sides and use $\widetilde{\kappa}_j[H_0(j,j)]^{-\frac{1}{2}} \le \sqrt{\frac{\log(n)}{n\bar{\theta}^2}}\frac{1}{n\bar{\theta}}$ again. It leads to that $\big|e_i' H_0^{-\frac{1}{2}}(\hat{\xi}_1 - \xi_1)\big| \le CK^{\frac{1}{2}}\sqrt{\frac{\log(n)}{n\bar{\theta}^2}}\frac{1}{n\bar{\theta}} + C\sqrt{\frac{\log(n)}{n\bar{\theta}^2}}\,\|H_0^{-\frac{1}{2}}(\hat{\xi}_1 - \xi_1)\|_\infty$. This inequality holds for every $1 \le i \le n$. Since $\hat{\xi}_1 - \xi_1$ does not depend on i, if we take a maximum over i, then both the left and right hand sides contain a term related to $\|H_0^{-\frac{1}{2}}(\hat{\xi}_1 - \xi_1)\|_\infty$. Noticing that $\sqrt{\frac{\log(n)}{n\bar{\theta}^2}} = o(1)$, we re-arrange the terms to get $\|H_0^{-\frac{1}{2}}(\hat{\xi}_1 - \xi_1)\|_\infty \le CK^{\frac{1}{2}}\sqrt{\frac{\log(n)}{n\bar{\theta}^2}}\frac{1}{n\bar{\theta}} \le CK^{\frac{1}{2}}\frac{1}{n\bar{\theta}}$. Plugging it into (38) gives $|\hat{\xi}_1(i) - \xi_1(i)| \le CK^{\frac{1}{2}}\widetilde{\kappa}_i$. The claim (17) follows immediately by plugging in the definition of $\widetilde{\kappa}_i$ in Lemma 5.3. □

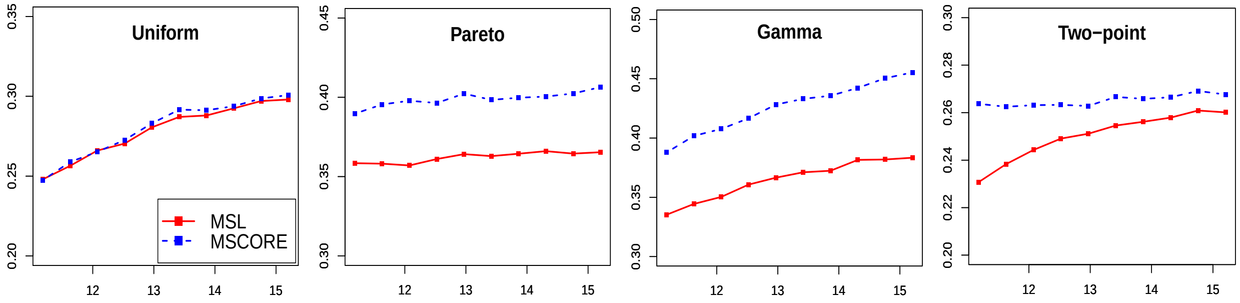

Figure 3: MSL v.s. MSCORE ($n = 2000$, $K = 2$, x-axis is $\sqrt{n\bar{\theta}^2}$, and y-axis is the ℓ^1-loss in (5), averaged over 100 repetitions). We consider four different cases of degree heterogeneity.

6 Simulations

We conduct two experiments: Experiment 1 compares Mixed-SCORE-Laplacian (MSL) with Mixed-SCORE [24]. Recall that MSL is a modification of the conventional Mixed-SCORE. We hope to demonstrate that our proposed modification improves the numerical performance under severe degree heterogeneity. Experiment 2 investigates the node-wise errors of MSL. We hope to show that the node-wise error indeed varies with θ_i as specified in Theorem 4.2. In addition, we also study the effect of K in Section C of the supplementary material. Both MSL and Mixed-SCORE require plugging in a vertex hunting (VH) algorithm as the input. We use successive projection [5] as the default VH algorithm, except in Experiment 2. Since this experiment studies the very delicate node-wise errors, we follow [24] to use an improved VH algorithm; see Section B.1 of the supplementary material for details.

Experiment 1: Comparison with the conventional Mixed-SCORE. Fix $(n, K) = (2000, 2)$. Let $P = \beta_n I_2 + (1 - \beta_n)\mathbf{1}_2\mathbf{1}_2'$. To generate θ, given a distribution $F(\cdot)$ and $b_n > 0$, we draw $\theta_1^0, \theta_2^0, \ldots, \theta_n^0 \overset{iid}{\sim} F(\cdot)$ and then let $\theta_i = b_n \cdot n\theta_i^0/\|\theta^0\|_1$ for $1 \leq i \leq n$. To generate Π, we first set $\pi_i = (1, 0)'$ and $\pi_i = (0, 1)'$ each for 15% of nodes, and then let $\pi_i = (t_i, 1 - t_i)'$ for the remaining 70% of nodes, with $t_i \overset{iid}{\sim} \text{Uniform}([0, 1])$. The distribution $F(\cdot)$ controls degree heterogeneity. In light of Proposition 2.1, we consider four choices of $F(\cdot)$: in Experiment 1.1, $F = \text{Uniform}([0.3, 5])$; in Experiment 1.2, $F = \text{Pareto}(10, 0.3)$, with a truncation at 5000; in Experiment 1.3, $F = \text{Gamma}(1/3, 1)$; in Experiment 1.4, $F = 0.05\delta_9 + 0.95\delta_{0.1}$, where

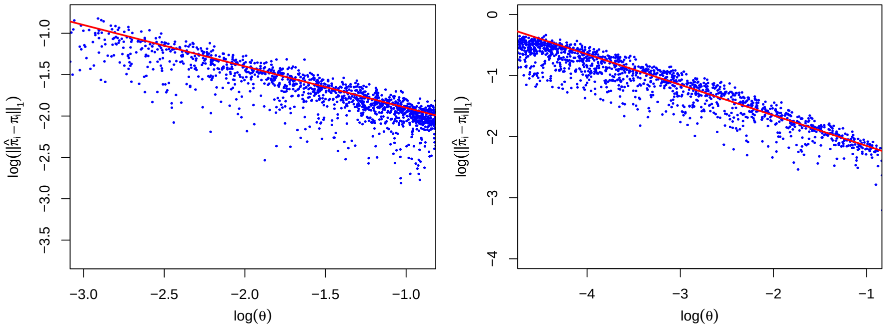

Figure 4: Node-wise errors of MSL. θ_i's are generated from $\mathcal{U}([0.3, 5])$ and Pareto$(10, 0.3)$, respectively for the left and the right. In each plot, the slope of the red line is $-1/2$.

δ_x is a point mass at x. The parameter b_n $(= \bar{\theta})$ controls network sparsity, and β_n controls 'community dis-similarity'. Write SNR $:= \sqrt{n\bar{\theta}^2}(1 - P(1,2)) = b_n\beta_n\sqrt{n}$. We let nb_n range from 500 to 680 with a grid of 20, and change β_n accordingly by fixing SNR $= 10$. Figure 3 reports the ℓ^1-loss (averaged over 100 repetitions) for four cases of degree heterogeneity and various levels of network sparsity. In Section C of the supplementary material, we also report the weighted ℓ^1-loss. Except in Experiment 1.1, degree heterogeneity is severe in the other three cases, and we always observe a significant improvement of MSL over MSCORE.

Experiment 2: Node-wise errors of Mixed-SCORE-Laplacian. Fix $(n, K) = (2000, 2)$. We first generate (θ, Π) as in Experiment 1. Fixing (θ, Π), we then generate 100 networks, and compute the average of $\|\hat{\pi}_i - \pi_i\|_1$ on these 100 repetitions, for each $1 \leq i \leq n$. In Experiment 2.1, F is Uniform$([0.3, 5])$; in Experiment 2.2, F is Pareto$(10, 0.3)$. Fix $\sqrt{n}\, b_n = 15$ and SNR $= b_n\beta_n\sqrt{n} = 10$. Theorem 4.2 says that the error at $\hat{\pi}_i$ is (approximately) proportional to $(\theta_i \wedge \bar{\theta})^{-1/2}$. In Figure 4, we plot $\log(\|\hat{\pi}_i - \pi_i\|_1)$ versus $\log(\theta_i)$, for $\theta_i \leq \bar{\theta}$. Under both moderate degree heterogeneity (left panel) and severe degree heterogeneity (right panel), the plot fits a straight line with slope $-1/2$ well. This verifies the claim of Theorem 4.2.

7 Real data

We use real-world examples to denomstrate the practical implications of our study. The first data set, the political blog network [2], contains "true" community labels. We use this data set to compare the node embeddings with and without pre-PCA normalization. The second data set is a co-authorship network of statisticians [19]. We apply both the conventional Mixed-SCORE and our proposed MSL with $K = 2$ and compare their performances.

The political blog network [2]. Each node is a blog during 2004 U.S. presidential election, and there is an edge between two nodes if there is a link (one-way or bilateral) between two blogs. A manual label of "conservative" or "liberal" was given to each blog by [2]. We use the giant component of the network, which has $n = 1222$ nodes. In addition, we construct a new network by adding two outlier nodes; for each of them, we connect it to 80% of the original nodes (randomly selected). The original network already shows severe degree heterogeneity ($d_{\text{min}} = 1$, $d_{\text{mean}} = 27.35$, and $d_{\text{max}} = 351$). Degree heterogeneity in the new network is even more severe ($d_{\text{max}} = 978$). For each network, we compute the SCORE embeddings in (10) with $K = 2$, where $\hat{\xi}_1$ and $\hat{\xi}_2$ are either eigenvectors of A (no pre-PCA normalization) or L (with pre-PCA normalization). The histograms of $\hat{r}_i$'s in these four cases (i.e., two networks, with/without pre-PCA normalization) are displayed in Figure 5. Since most bloggers are not extremely conservative/liberal, this network has mixed memberships. However, the manual labels by [2] are binary. To compare different embeddings, we calculate (a) Rayleight quotient $Rq := \frac{\pi^*(1-\pi^*)(\bar{r}_{(1)}-\bar{r}_{(2)})^2}{\pi^* v_{(1)}+(1-\pi^*)v_{(2)}}$, where π^* is the fraction of nodes in the "conservative" group, $\bar{r}_{(1)}$ and $v_{(1)}$ are the mean and variance of $\hat{r}_i$'s within this group, and $\bar{r}_{(2)}$ and $v_{(2)}$ are those within the "liberal" group, and (b) the k-means clustering errors on $\hat{r}_i$'s (treating the manual labels as the ground truth) in each of the four cases. When calculating RQ and clustering error, we exclude the outliers and only use $\hat{r}_i$'s of the original nodes. The left two panels of Figure 5 are for the original network (no outlier). The SCORE embeddings with pre-PCA normalization yield better RQs and smaller clustering errors. The advantage is more significant in the new network with outliers. These outliers have extremely high degrees and bring a lot of noise into the eigenvectors of A. Consequently, the RQ drops significantly and the clustering error has a big increase. In comparison, the pre-PCA normalization can properly "down-weight" high-degree nodes in PCA and make the SCORE embeddings more robust to outliers.

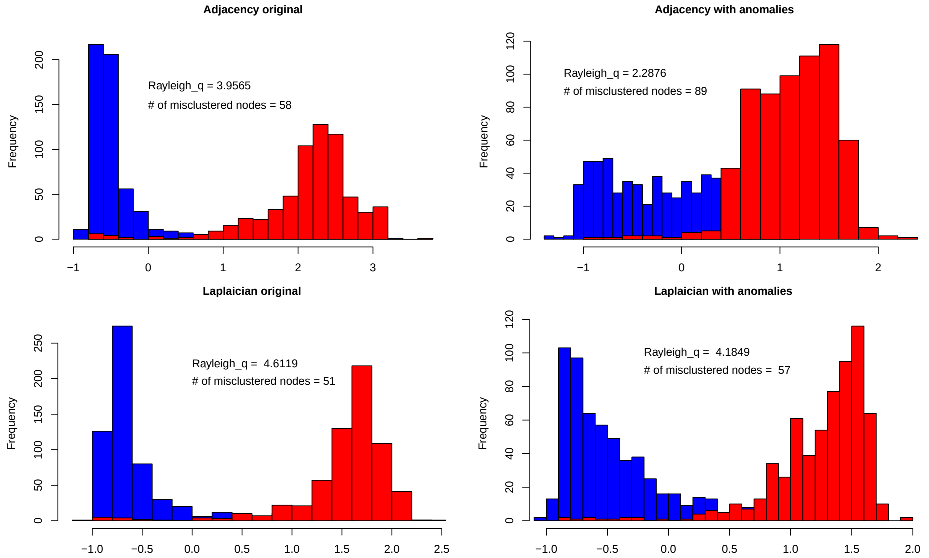

Figure 5: Comparison of the SCORE embeddings without pre-PCA normalization (top) and with pre-PCA normalization (bottom). Since $K = 2$, each embedding is a point in $\mathbb{R}$. The colors correspond to the true community labels. The left two panels are based on the original political blog network, and the right two panels are based on the network with outliers.

The co-authorship network of statisticians [19]. Several co-authorship networks were constructed by [19]. One of them is a 236-author network, in which two authors (nodes) are connected by an edge if they co-authored ≥ 2 papers during 2003-2012 in four core statistical journals. [19] discorvered that there are two communities, "Carroll-Hall" and "North-Carolina", in this network. [24] further studied this network and found strong evidence of mixed memberships: many members of the Fan-group (Jianqing Fan and collaborators) have nearly half-half memberships in two communities. Figure 6 shows the estimated membership in the "Carroll-Hall" for all 236 authors, where the x-axis is the Mixed-SCORE estimate, the y-axis corresponds to the MSL estimate, and the dashed line is $y = x$. We make several observations. First, the membership of Jianqing Fan is still near 50%. Second, most pure nodes of the "North-Carolina" community are still pure. Last, Raymond Carroll and Peter Hall are no longer 100% pure in the "Carroll-Hall" community. Meanwhile, authors such as Jing Qin are now pure nodes of this community. This does not conflict with the discoveries in [19, 24]: In fact, they claimed that this is a group of authors interested in nonparametric and semi-parametric statistics and "Carroll-Hall" was used as a short name. We note that

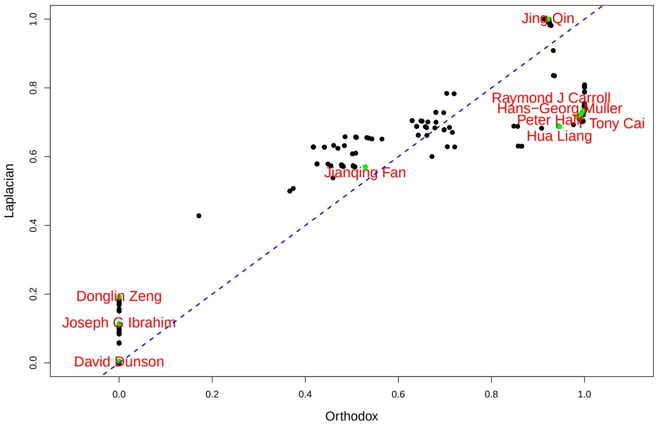

Figure 6: The estimated membership in the "Caroll-Hall" community. The x-axis and y-axis correspond to MSCORE and MSL, respectively. The 10 highest-degree nodes are marked in green with the author names displayed by side.

both Raymond Carroll and Peter Hall have high degrees. MSL properly down-weights the contributions of high-degree nodes before conducting PCA, so we expect it to yield more accurate estimates. Indeed, the results are more consistent with our common insights: highly collaborative authors usually have diversified interests, hence "mixed" memberships.

8 Discussions

Severe degree heterogeneity has been widely observed in real networks [6], but its effect on network model estimation and inference is largely unclear. This paper provides the first result of optimal mixed membership estimation under (nearly) arbitrary degree heterogeneity. Our contributions include (i) a degree-heterogeneity-aware lower bound, (ii) an optimal spectral algorithm, (iii) node-wise error rates, and (iv) entry-wise eigenvector analysis.

In this paper, we assume K is known. If K is unknown, how to estimate K from data is an interesting problem. Existing works of estimating K (e.g., [32, 25, 10, 18]) focus on the community detection setting and cannot be applied directly. Our theoretical results on eigenvalues of L (see Section D of the supplementary material) suggest an estimator of K:

$$\hat{K} = \#\big\{1 \leq j \leq n: \ \hat{\lambda}_j^2 \geq \bar{d}^{-1}\log^2(n)\big\}, \quad \text{where } \bar{d} \text{ is avarege node degree.}$$

We can show that when $K^3 \log^2(n)/(n\bar{\theta}^2\beta_n^2) \to \infty$, $\hat{K} = K$ with probability $1 - o(1)$. Hence, all the results in this paper continue to hold when K is estimated from data.

We focus on undirected networks, but our results are extendable to directed networks. We can extend the DCMM model to directed networks by introducing two degree parameters, θ_i^{in} and θ_i^{out}, and two mixed membership vectors π_i^{in} and π_i^{out}, for each node i. The degree heterogeneity in this model is captured by two CDFs, F_n^{in} and F_n^{out}. We conjecture that the optimal rates in estimating Π^{in} and Π^{out} are different, and they are both functions of n, K, $\bar{\theta}^{in}$, $\bar{\theta}^{out}$, and F_n^{in} and F_n^{out}. We also conjecture that the optimal spectral method requires a normalization on the adjacency matrix as $L = (H^{out})^{-b_1} A (H^{in})^{-b_2}$, where H^{in} and H^{out} are diagonal matrices consisting of incoming and outgoing node degrees of nodes, and the choice of (b_1, b_2) is inspired by node-wise error analysis. We leave this to future work.

Uncertainty quantification of $\hat{\Pi}$ is another interesting problem. [7] derived a novel finite-sample expansion for the $\hat{\pi}_i$ from conventional Mixed-SCORE, which yields valid confidence interval of π_i. They also proposed a ranking inference procedure for individual membership's profiles. Our results in this paper suggest that when there is severe degree heterogeneity, MSL may significantly improve the conventional Mixed-SCORE. It is thus interesting to develop uncertainty quantification for the π_i's from the MSL algorithm. This requires second-order expansions for the leading eigenvectors of graph Laplacian. We also leave it to future work.

The SCORE-family of algorithms [21, 24] can be extended to weighted networks, because the rationale of such methods is hinged on the structure of Ω. For example, let $\Omega = \Theta\Pi P \Pi'\Theta$ as before. If $A(i,j) \sim \text{Poisson}(m\Omega(i,j))$ or $A(i,j) \sim N(\Omega(i,j), \sigma^2)$, we can still apply Mixed-SCORE (with or without normalizations) to get a consistent estimate of Π. The question is how to achieve the optimal error rate. The rate-optimality result in this paper is tied to the Bernoulli edge generation (e.g., the mean and variance of a Bernoulli variable are of the same order). For Poisson edges, we conjecture that similar optimality result can be achieved. For normal edges, we conjecture that no degree normalization is needed, when the variances of all edges are the same.

References

[1] Abbe, E., J. Fan, K. Wang, and Y. Zhong (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist. 48*(3), 1452–1474.

[2] Adamic, L. A. and N. Glance (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.

[3] Agterberg, J. and A. Zhang (2022). Estimating higher-order mixed memberships via the 2-to-infinity tensor perturbation bound. *arXiv:2212.08642*.

[4] Airoldi, E., D. Blei, S. Fienberg, and E. Xing (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res. 9*, 1981–2014.

[5] Araújo, M. C. U., T. C. B. Saldanha, R. K. H. Galvao, T. Yoneyama, H. C. Chame, and V. Visani (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems 57*(2), 65–73.

[6] Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *Science 286*(5439), 509–512.

[7] Bhattacharya, S., J. Fan, and J. Hou (2023). Inferences on mixing probabilities and ranking in mixed-membership models. *arXiv:2308.14988*.

[8] Chaudhuri, K., F. Chung, and A. Tsiatas (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pp. 35–1. JMLR Workshop and Conference Proceedings.

[9] Chen, Y., Y. Chi, J. Fan, and C. Ma (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning 14*(5), 566–806.

[10] Dall'Amico, L., R. Couillet, and N. Tremblay (2021). A unified framework for spectral clustering in sparse graphs. *J. Mach. Learn. Res. 22*(217), 1–56.

[11] Davis, C. and W. M. Kahan (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM J. Numer. Anal. 7*(1), 1–46.

[12] Erdős, L., A. Knowles, H.-T. Yau, and J. Yin (2013). Spectral statistics of erdős–rényi graphs i: Local semicircle law. *Ann. Probab. 41*(3B), 2279–2375.

[13] Fan, J., Y. Fan, X. Han, and J. Lv (2022). SIMPLE: Statistical inference on membership profiles in large networks. *J. R. Stat. Soc. Ser. B. 84*(2), 630–653.

[14] Fu, W., L. Song, and E. P. Xing (2009). Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th annual international conference on machine learning*, pp. 329–336.

[15] Gao, C., Z. Ma, A. Y. Zhang, and H. H. Zhou (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res. 18*(1), 1980–2024.

[16] Horn, R. and C. Johnson (1985). *Matrix Analysis*. Cambridge University Press.

[17] Huang, W., Y. Liu, and Y. Chen (2020). Mixed membership stochastic blockmodels for heterogeneous networks. *Bayesian Analysis 15*(3), 711–736.

[18] Hwang, N., J. Xu, S. Chatterjee, and S. Bhattacharyya (2023). On the estimation of the number of communities for sparse networks. *J. Amer. Statist. Soc.*, 1–16.

[19] Ji, P. and J. Jin (2016). Coauthorship and citation networks for statisticians (with discussions). *Ann. Appl. Statist. 10*(4), 1779–1812.

[20] Ji, P., J. Jin, Z. T. Ke, and W. Li (2022). Co-citation and co-authorship networks of statisticians (with discussions). *J. Bus. Econom. Statist. 40*, 469–485.

[21] Jin, J. (2015). Fast community detection by score. *Ann. Statist. 43*(1), 57–89.

[22] Jin, J. and Z. T. Ke (2017). A sharp lower bound for mixed-membership estimation. *arXiv:1709.05603*.

[23] Jin, J., Z. T. Ke, and S. Luo (2022). Improvements on SCORE, especially for weak signals. *Sankhya A 84*(1), 127–162.

[24] Jin, J., Z. T. Ke, and S. Luo (2024). Mixed membership estimation for social networks. *J. Econometrics 239*(2), 105369.

[25] Jin, J., Z. T. Ke, S. Luo, and M. Wang (2023). Optimal estimation of the number of network communities. *J. Amer. Statist. Assoc. 118*(543), 2101–2116.

[26] Karrer, B. and M. Newman (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E 83*(1), 016107.

[27] Liu, F., D. Choi, L. Xie, and K. Roeder (2018). Global spectral clustering in dynamic networks. *Proc. Natl. Acad. Sci. 115*(5), 927–932.

[28] Mao, X., P. Sarkar, and D. Chakrabarti (2021). Estimating mixed memberships with sharp eigenvector deviations. *J. Amer. Statist. Assoc. 116*(536), 1928–1940.

[29] Qin, T. and K. Rohe (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Adv. Neural Inf Process. Syst.*, pp. 3120–3128.

[30] Sarkar, P. and P. J. Bickel (2015). Role of normalization in spectral clustering for stochastic blockmodels. *Ann. Statist. 43*(3), 962–990.

[31] Tang, M. and C. E. Priebe (2018). Limit theorems for eigenvectors of the normalized laplacian for random graphs. *Ann. Statist. 46*(5), 2360–2415.

[32] Wang, Y. R. and P. J. Bickel (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist. 45*(2), 500.

[33] Yang, J. and J. Leskovec (2013). Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 587–596.

[34] Zhang, Y., E. Levina, and J. Zhu (2020). Detecting overlapping communities in networks using spectral methods. *SIAM J. Math. Data Sci. 2*(2), 265–283.