Estimating Mixed Memberships with Sharp Eigenvector Deviations

Xueyu Mao, Purnamrita Sarkar an Deepayan Chakrabarti
The University of Texas at Austin

March 30, 2020

Abstract

We consider the problem of estimating community memberships of nodes in a network, where every node is associated with a vector determining its degree of membership in each community. Existing provably consistent algorithms often require strong assumptions about the population, are computationally expensive, and only provide an overall error bound for the whole community membership matrix. This paper provides uniform rates of convergence for the inferred community membership vector of *each* node in a network generated from the Mixed Membership Stochastic Blockmodel (MMSB); to our knowledge, this is the first work to establish per-node rates for overlapping community detection in networks. We achieve this by establishing sharp row-wise eigenvector deviation bounds for MMSB. Based on the simplex structure inherent in the eigen-decomposition of the population matrix, we build on established corner-finding algorithms from the optimization community to infer the community membership vectors. Our results hold over a broad parameter regime where the average degree only grows poly-logarithmically with the number of nodes. Using experiments with simulated and real datasets, we show that our method achieves better error with lower variability over competing methods, and processes real world networks of up to 100,000 nodes within tens of seconds.

Keywords: Overlapping community detection, clustering, networks, asymptotic analysis

1 Introduction

In most real-world networks, a node belongs to multiple communities. In an university, professors have joint appointments to multiple departments; a movie like "Dirty Harry" in the Netflix recommendation network belongs to action, thriller, and the drama genre according to Google; in a book recommendation network like goodreads.com, "To Kill a Mockingbird" can be classified as a classic, historical fiction, young-adult fiction, etc. The goal of community detection is to consistently infer each node's community memberships from just the network structure.

A well-studied variant of this problem assumes that each node belongs to a single community. For instance, under the Stochastic Blockmodel (SBM) [1], the probability of a link between two nodes depends only on their respective communities. Thus, provably consistent inference under the Stochastic Blockmodel involves finding the unknown cluster membership of each node (see [2, 3, 4]) and these are not immediately applicable for the general problem where a node may belong to multiple communities to different degrees.

In this paper, we work with the popular Mixed Membership Stochastic Blockmodel (MMSB) [5]. This generalizes the Stochastic Blockmodel by letting each node i have different degrees of membership in all communities. In particular, each node i is associated with a community membership vector $\boldsymbol{\theta}_i \in \mathbb{R}^K$ ($\boldsymbol{\theta}_i \geq 0, \|\boldsymbol{\theta}_i\|_1 = 1$), drawn from a Dirichlet prior. The model for generating the symmetric adjacency matrix is as follows: [1]

$$\begin{aligned} &\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha}) && \boldsymbol{\alpha} \in \mathbb{R}_+^K, \quad i \in [n] \\ &\mathbf{P} := \rho\boldsymbol{\Theta}\mathbf{B}\boldsymbol{\Theta}^T \qquad \mathbf{A}_{ij} = \mathbf{A}_{ji} \sim \text{Bernoulli}(\mathbf{P}_{ij}) && i, j \in [n] \end{aligned} \tag{1}$$

The matrix $\boldsymbol{\Theta}$ has $\boldsymbol{\theta}_i^T$ as its i^{th} row. For identifiability we assume $\max_{ij} \mathbf{B}_{ij} = 1$. When $\mathbf{B}$ has higher values on its diagonal as compared to the off-diagonal, edges are likely between nodes that have a high membership in the same community. These are called assortative

[1] Note that self-loops are allowed here for simplicity of analysis. Without them, the analysis gets cumbersome, leading to a negligible error term added to all our bounds, and we skip it for ease of exposition.

communities. In contrast, in disassortative settings, off-diagonal elements are larger than diagonal elements. Bipartite graphs are an extreme case of this. The smallest singular value of $\mathbf{B}$, denoted by $\lambda^*(\mathbf{B})$, is a measure of the separation between communities. A larger $\lambda^*(\mathbf{B})$ corresponds to more well-separated communities. The parameter ρ controls the the expected average degree of nodes $O(n\rho)$. We allow both ρ and $\lambda^*(\mathbf{B})$ to go to zero with increasing number of nodes n. The quantity $\alpha_0 = \sum_{a=1}^{K} \alpha_a$ controls the level of overlap between members of different communities. As $\alpha_0 \to 0$, the MMSB model degenerates to the Stochastic Blockmodel. The goal of community detection under the MMSB model is to recover $\mathbf{\Theta}$ and $\mathbf{B}$ from the observed adjacency matrix $\mathbf{A}$.

Prior work on this problem include MCMC [5] and computationally efficient variational approximation methods [6] (SVI) which do not have any guarantees of consistency. Other interesting network models for overlapping communities and non-negative matrix factorization style inference methods which do not have theoretical guarantees include [7, 8, 9, 10]. A notable family of algorithms that has been shown to be theoretically consistent uses tensor-based methods [11, 12]. However, these are typically hard to implement, and provide overall error bounds for the columns of the estimated $\mathbf{\Theta}$ matrix.

Recently Mao et al. [13] have proposed a provably consistent geometric algorithm (GeoNMF) for MMSB with diagonal $\mathbf{B}$ and $\boldsymbol{\alpha} = \alpha_0 \mathbf{1}_K / K$. However the guarantees only work in the dense regime where average degree grows faster than $\sqrt{n}$. In contrast we consider the general model where the only condition on $\mathbf{B}$ is full rank. We propose a different algorithm which works when the degree grows faster than poly-logarithm of n.

Zhang et al. [14] propose a provably consistent spectral algorithm (OCCAM) for a related but different model with degree correction. Similar to non-negative matrix factorization methods [15, 13], the authors assume that each community has some "pure" nodes (which only belong to that community). The authors also assume that $\mathbf{B}$ is positive semidefinite and full rank with equal diagonal entries. Other assumptions ensure that the k-medians loss function on $\boldsymbol{\theta}_i$ attains its minimum at the locations of the pure nodes and there is a curvature around this minimum. This condition is typically hard to check.

Concurrent work [16] studies the degree corrected MMSB model, which extends the MMSB model by allowing degree heterogeneity. The authors show an interesting fact that the top eigenvectors, normalized appropriately, still form a simplex. However, their proposed algorithm requires a combinatorial search step (SVS)[2], and has a complexity $O(n^{KL} + K^3 L^{K+1})$ for some tuning parameter $L \geq K$. This can be prohibitive for large K. SVS is analyzed under three separate settings, a) $\boldsymbol{\theta}_i$ are sampled from a distribution on the simplex such that every cluster has $\Theta(n)$ pure nodes, and the non pure nodes are sufficiently separated from the pure ones; b) the $\boldsymbol{\theta}_i$'s are fixed, but form a few clusters, or c) the $\boldsymbol{\theta}_i$'s are fixed, and most nodes are pure nodes.

Other notable examples of related but different models include [17, 18]. In [17], the authors show consistency when the overlap between clusters is small, whereas in [18], a combinatorial algorithm (SAAC) is proposed for detecting overlapping communities for a related model.

In this paper, our contributions are as follows.

Identifiability: We present both necessary and sufficient conditions for identifiability of the MMSB model in Sec 2. To our knowledge, we are the first to report both necessary and sufficient conditions for identifiability under the MMSB model.

Recovery algorithm: As shown by many authors [13, 16, 19], the population eigenvectors (i.e., eigenvectors of the matrix $\mathbf{P}$) form a rotated and scaled simplex. We present an algorithm called SPACL, which re-purposes an existing algorithm [20] for detecting corners in a rotated and scaled simplex to find pure nodes, and then uses these to infer $\boldsymbol{\Theta}$ and $\mathbf{B}$. It also includes a novel preprocessing step that improves performance in sparse settings. The main compute-intensive parts of the algorithm are a) top-K eigen-decomposition of $\mathbf{A}$,

[2] In the latest version of [16], the authors have added other methods, and proved node-wise error bounds. But they note that among these methods, SVS performs the best. We compare against the newer bounds later in our paper.

Table 1: Table of notations. K leading eigenvectors of a matrix correspond to K largest eigenvalues in magnitude.

n	Number of nodes	K	Number of communities
$\rho\mathbf{B} \in [0,1]^{K\times K}$	Community link probabilities ($\mathbf{B} = \mathbf{B}^T$)	$\boldsymbol{\alpha} \in \mathbb{R}_+^{K\times 1}$	Dirichlet prior parameters
$\mathbf{\Theta} \in \mathbb{R}_+^{n\times K}$	Fractional community memberships	α_0	$\sum_i \alpha_i$
$\alpha_{\min}$ ($\alpha_{\max}$)	$\min_{i\in[K]} \alpha_i$ ($\max_{i\in[K]} \alpha_i$)	ν	$\alpha_0/\alpha_{\min}$
$\mathbf{A}$	Adjacency matrix	$\mathbf{P}$	$\rho\mathbf{\Theta}\mathbf{B}\mathbf{\Theta}^T$
ρ	Upper bound on $\mathbf{P}_{ij}$	$\mathbf{I}_m$	$m \times m$ identity matrix
$\mathbf{E}$	Diagonal matrix of K largest eigenvalues in magnitude of $\mathbf{P}$	$\mathbf{V} \in \mathbb{R}^{n\times K}$	K leading eigenvectors of $\mathbf{P}$
$\hat{\mathbf{E}}$	Diagonal matrix of K largest eigenvalues in magnitude of $\mathbf{A}$	$\hat{\mathbf{V}} \in \mathbb{R}^{n\times K}$	K leading eigenvectors of $\mathbf{A}$
$\mathbf{V}_P \in \mathbb{R}^{K\times K}$	True K pure node index rows of $\mathbf{V}$	$\lambda_K(\mathbf{M})$	K^{th} largest eigenvalue of $\mathbf{M}$
$\mathbf{V}_p \in \mathbb{R}^{K\times K}$	Estimated K pure node index rows of $\mathbf{V}$	$\lambda^*(\mathbf{M})$	K^{th} largest singular value of $\mathbf{M}$
$\kappa(\mathbf{M})$	Condition number of matrix $\mathbf{M}$	λ_i	i^{th} largest eigenvalue of $\mathbf{P}$
$\mathbf{\Pi} \in \{0,1\}^{K\times K}$	Permutation matrix	$\hat{\lambda}_i$	i^{th} largest eigenvalue of $\mathbf{A}$
$\mathbf{1}_m$	All ones vector of length m	$\mathbf{e}_i$	$\mathbf{e}_i(j) = 1(i = j)$

b) calculating k-nearest neighbors of a point for preprocessing. There are highly optimized algorithms and data structures for both of these steps [21, 22, 23].

Node-wise error bound: Some of the existing works on MMSB type models show consistency in terms of the deviation or correlation of $\hat{\mathbf{\Theta}}$ as a whole with respect to the truth [14, 18, 12, 16]. Others establish consistency of the deviation of columns of $\hat{\mathbf{\Theta}}$ [11] (soft memberships of all nodes to a particular community) from their population counterpart. In contrast, we obtain a uniform rate of convergence of *each* cluster membership vector $\hat{\boldsymbol{\theta}}_i, i \in [n]$ to $\boldsymbol{\theta}_i$. To our knowledge this is the first work to establish uniform node-wise error bounds for an estimation algorithm for overlapping network models.

Empirical validation: In Sec 4, we compare SPACL with OCCAM, variational methods, SAAC and existing non-negative matrix factorization algorithms (GeoNMF, BSNMF) on both simulated and large real world networks with up-to 100,000 nodes.

2 Notations, Identifiability and Algorithms

Before presenting our results on identifiability we introduce some notations and assumptions. Let $[n] := \{1, 2, \cdots, n\}$. For any matrix $\mathbf{M}$, we use $\mathbf{M}(i, :)/\mathbf{M}(:, i)$, $\mathbf{M}(S, :)/\mathbf{M}(:, S)$ to denote the i^{th} row/column of matrix $\mathbf{M}$ and the submatrix formed by rows/columns in set S of matrix $\mathbf{M}$ respectively, and $S = i : j$ denotes the set of indices from i to j. We use $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_F$ to respectively denote the operator and Frobenius norms of a matrix $\mathbf{M}$, and $\|\mathbf{v}\|$ to denote the Euclidean norm of a vector $\mathbf{v}$. We denote $[\mathbf{X}|\mathbf{Y}]$ as the concatenation of columns of matrices $\mathbf{X}$ and $\mathbf{Y}$. We use $\tilde{O}$ and $\tilde{\Omega}$ to denote upper and lower bounds up to poly-logarithmic factors. Finally we present a consolidated list of notations in Table 1.

We shall now provide necessary and sufficient conditions for the identifiability of the MMSB model with respect to $\mathbf{\Theta}$ and $\mathbf{B}$.

2.1 Identifiability

In this section, we obtain necessary and sufficient conditions for the identifiability of MMSB. In contrast, prior work [14, 13, 18, 19] typically establishes sufficient conditions. We defer the proofs of the theorems in this section to the supplementary material (Sec I).

Define a pure node as a node which belongs to exactly one community. All nodes in a Stochastic Blockmodel are pure nodes, since every node belongs to exactly one community. Define a "completely mixed" node as a node m such that $\theta_{mj} > 0$ for all $j \in [K]$.

Theorem 2.1. *Suppose there are K communities, with at least one pure node for each community. Then,*

(a) If $\operatorname{rank}(\mathbf{B}) = K$, then the MMSB model is identifiable up to a permutation.

(b) If $\operatorname{rank}(\mathbf{B}) = K - 1$, and no row of $\mathbf{B}$ is an affine combination of the other rows of $\mathbf{B}$, then the MMSB model is identifiable up to a permutation.

(c) In any other case, if there exists a completely mixed node, then the model is not identifiable.

Theorem 2.2. *Suppose that $\rho\mathbf{B}_{ij} \in (0,1)$ for all $i, j \in [K]$. MMSB is identifiable up to a permutation only if there is at least one pure node for each of the K communities.*

The above theorems show that the existence of pure nodes is necessary in most practical scenarios.

2.2 Algorithm

We do inference for the MMSB model under the following assumption, which is sufficient for identifiability.

Assumption 2.1. $\mathbf{B} \in \mathbb{R}^{K\times K}$ is full rank, and there is at least one pure node for each of the K communities.

Since the Dirichlet distribution does not give rise to pure nodes, we assume that the set $\{\boldsymbol{\theta}_i, i \in [n]\}$ includes one pure nodes from each cluster in addition to $n - K$ vectors drawn from a Dirichlet. The addition of one pure node per cluster to the standard Dirichlet draws does not affect the analysis and we ignore this for ease of exposition.

We will now discuss our inference algorithm, whose consistency results are presented in Sec 3. Let $\mathbf{P} = \mathbf{V}\mathbf{E}\mathbf{V}^T$ be the top-K eigendecomposition of $\mathbf{P}$. We proceed from a simple observation that the population eigenvectors lie on a rotated and scaled simplex, as shown next. The following lemma is the starting point of most existing analysis for Stochastic Blockmodels, and different variants of this have been observed independently by a number of other researchers [19, 16, 13].

Lemma 2.3. *Let $\mathbf{V}$ be the top K eigenvectors of $\mathbf{P}$. Then, under Assumption 2.1, $\mathbf{V} = \boldsymbol{\Theta}\mathbf{V}_P$, where $\mathbf{V}_P = \mathbf{V}(\mathcal{I}, :)$ is full rank and $\mathcal{I}$ is the indices of rows corresponding to K pure nodes, one from each community.*

Proof. W.L.O.G., reorder the nodes so that $\boldsymbol{\Theta}(\mathcal{I}, :) = \mathbf{I}$. Then, $\mathbf{V}_P\mathbf{E}\mathbf{V}_P^T = \mathbf{P}(\mathcal{I}, \mathcal{I}) = \rho\mathbf{B}$, so

Algorithm 1 SPACL

```
Input: Adjacency matrix $\mathbf{A}$, number of clusters $K$
Output: $\hat{\mathbf{\Theta}}$, $\hat{\mathbf{B}}$, $\hat{\rho}$.
Get the top-$K$ eigen-decomposition of $\mathbf{A}$ as $\hat{\mathbf{V}}\hat{\mathbf{E}}\hat{\mathbf{V}}^T$.
$S = \text{Prune}(\hat{\mathbf{V}}, 10, .75, .95)$
$\boldsymbol{X} = \hat{\mathbf{V}}([n] \setminus S, :)$
$\mathcal{S}_p = \text{SPA}(\boldsymbol{X}^T)$
$\boldsymbol{X}_p = \boldsymbol{X}(\mathcal{S}_p, :)$
$\hat{\mathbf{\Theta}} = \hat{\mathbf{V}}\boldsymbol{X}_p^{-1}$.
$\hat{\mathbf{\Theta}} = \text{diag}(\hat{\mathbf{\Theta}}_+\mathbf{1}_K)^{-1}\hat{\mathbf{\Theta}}_+$
$\hat{\mathbf{B}} = \boldsymbol{X}_p\hat{\mathbf{E}}\boldsymbol{X}_p^T$
$\hat{\rho} = \max_{i,j}\hat{\mathbf{B}}_{ij}$. $\hat{\mathbf{B}} = \hat{\mathbf{B}}/\hat{\rho}$
```

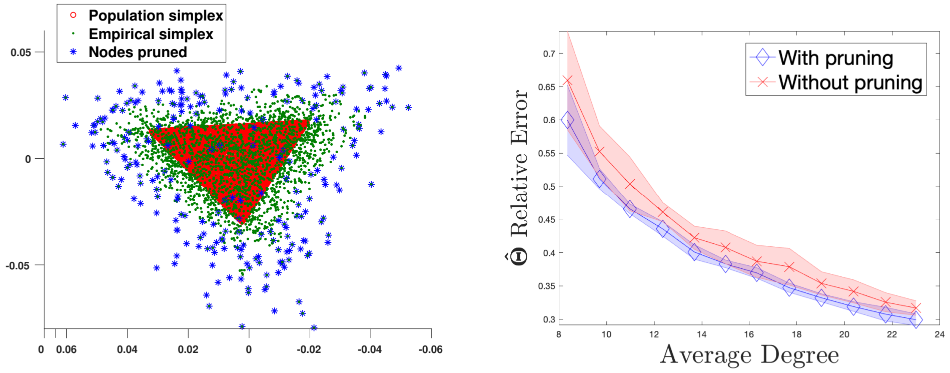
Algorithm 2 Prune

```
Input: Empirical eigenvectors $\hat{\mathbf{V}} \in \mathbb{R}^{n\times K}$, an integer $r$, and two numbers $q, \varepsilon \in (0,1)$.
Output: Set $S$ of nodes to be pruned.
for $i \in n$ do
    $v_i = \|\mathbf{e}_i^T\hat{\mathbf{V}}\|$
end for
$S_0 = \{i : \|\mathbf{e}_i^T\hat{\mathbf{V}}\| \geq \text{quantile}(\mathbf{v}, q)\}$
for $i \in S_0$ do
    $d_i$ :={Dist. to $r$ nearest neighbors}
    $x_i = \sum_j d_{ij}/r$
end for
$S = \{i : x_i \geq \text{quantile}(x, 1-\varepsilon)\}$
```

$\mathbf{V}_P \in \mathbb{R}^{K\times K}$ is full rank. Now, observe that $\mathbf{V}_P\mathbf{E}\mathbf{V}^T = \mathbf{P}(\mathcal{I}, :) = \rho\mathbf{\Theta}(\mathcal{I}, :\;)\mathbf{B}\mathbf{\Theta}^T = \rho\mathbf{B}\mathbf{\Theta}^T$. Hence, $\mathbf{V} = \mathbf{P}\mathbf{V}\mathbf{E}^{-1} = \rho\mathbf{\Theta}\mathbf{B}\mathbf{\Theta}^T\mathbf{V}\mathbf{E}^{-1} = \mathbf{\Theta}\mathbf{V}_P\mathbf{E}\mathbf{V}^T\mathbf{V}\mathbf{E}^{-1} = \mathbf{\Theta}\mathbf{V}_P$. □

Lemma 2.3 establishes that the corners of the simplex have the highest norm. This allows us to find the pure nodes using existing corner-finding methods such as the successive projection algorithm (SPA) [20].

Our algorithm, called "Sequential Projection After CLeaning" (SPACL, Algorithm 1) applies SPA after a preprocessing step that prunes away noisy high-norm points. SPA first finds the node with the maximum row norm of empirical eigenvector matrix $\hat{\mathbf{V}}$. This node is added to the set of pure nodes. Then, all remaining rows of $\hat{\mathbf{V}}$ are projected on to the subspace that is orthogonal to the span of the pure nodes. The process is repeated for K iterations, and yields a set of K pure nodes, one from each community. With the pure nodes



(A) (B)

Figure 1: MMSB model with $n = 5000$, $\boldsymbol{\alpha} = (0.4, 0.4, 0.4)$, $\mathbf{B} = (1 - q)\mathbf{I}_3 + q\mathbf{1}_3\mathbf{1}_3^T$ with $q = 0.001$. (A) Nodes picked out by Pruning with $\rho = 0.007$. (B) Effect of pruning on estimating $\hat{\boldsymbol{\Theta}}$ (relative error defined in Sec 4.1).

in hand, SPACL estimates $\boldsymbol{\Theta}$ and $\mathbf{B}$ using Lemma 2.3. We will show that these estimates are consistent up to a permutation (Theorem 3.5).

If we had access to the population eigenvectors $\mathbf{V}$, SPA would return the true pure nodes. However, in reality we only observe the empirical eigenvectors, which are noisy versions of the population eigenvectors. So there can be spurious nodes with row norm larger than those of the "pure" nodes. As the graph gets sparser, the empirical points deviate more from the population simplex. This motivates the pruning step of SPACL. The main idea of pruning is to identify and remove the nodes which are far away from the population simplex. Algorithm 2 finds these by first finding contenders of pure nodes, i.e., nodes i whose eigenvector rows $\hat{\mathbf{V}}_i := \mathbf{e}_i^T\hat{\mathbf{V}}$ have large norm. Among these, it prunes nodes which do not have too many nearest neighbors, or in other words, have larger average distance to their nearest neighbors in comparison to others. The removal of these nodes improves the performance of SPA on sparse networks.

Fig 1 (A) shows the benefits of pruning on a simulated network. After pruning, the remaining nodes are closer to the population simplex. This leads to better estimation. Fig 1

(B) varies ρ from 0.0050 to 0.0138 leading to average degrees increasing from 8 to 23, and shows the effect of pruning (blue $\Diamond$) over not pruning (red $\times$) on the relative estimation error of $\mathbf{\Theta}$. A more detailed discussion on pruning can be found in the supplementary material (Sec X).

3 Main results

We want to prove that the sample-based estimates $\hat{\mathbf{\Theta}}$, $\hat{\mathbf{B}}$ and $\hat{\rho}$ concentrate around their population counterparts, respectively, $\mathbf{\Theta}$, $\mathbf{B}$, and ρ. By Lemma 2.3, this requires concentration of the rows of the empirical eigenvector matrix $\hat{\mathbf{V}}$ to the population counterpart $\mathbf{V}$. Existing techniques like the Davis-Kahan Theorem [24] only provide convergence in the Frobenius norm $\|\mathbf{V} - \hat{\mathbf{V}}\mathbf{O}\|_F$ (for some rotation matrix $\mathbf{O}$) or the operator norm $\|\mathbf{V}\mathbf{V}^T - \hat{\mathbf{V}}\hat{\mathbf{V}}^T\|$. These lead to loose bounds on the rows of $\hat{\mathbf{V}}$. Other existing techniques [25, 26, 13] can be applied to show that rows of $\hat{\mathbf{V}}$ have $\tilde{O}_P(1/\sqrt{n\rho^2})$ relative error, but this is only meaningful when the degree grows faster than square root of n, i.e. the dense degree regime. We show that, under a broad parameter regime, the suitably defined relative deviation of any row of $\hat{\mathbf{V}}$ from its population counterpart, converges to zero when average degree only grows faster than the poly-logarithm of n.

In Section 3.1, we show the row-wise eigenspace error bound in terms of eigenvalues of $\mathbf{\Theta}^T\mathbf{\Theta}$. In Section 3.2, we translate the eigenspace bounds into error bounds on estimated $\hat{\mathbf{\Theta}}$ and $\hat{\mathbf{B}}$ matrices. Then, in Section 3.3, we provide detailed results when the rows of $\mathbf{\Theta}$ are drawn i.i.d from a Dirichlet distribution. Throughout, we compare our bounds to other bounds in concurrent results. We also discuss the implications of our results for specific models like the Stochastic Blockmodel.

3.1 Row-wise eigenvector error bounds

Assumption 3.1. Assume $\rho n = \Omega(\log n)$, $\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta}) \geq 1/\rho$, and $\lambda^*(\mathbf{P}) \geq 4\sqrt{n\rho}(\log n)^\xi$ for some constant $\xi > 1$.

Theorem 3.1 (Row-wise eigenspace error)**.** *If Assumptions 2.1 and 3.1 are satisfied, then with probability at least* $1 - O(Kn^{-2})$,

$$\max_{i\in[n]} \left\|\mathbf{e}_i^T(\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T)\right\| = \tilde{O}\left(\frac{\psi(\mathbf{P})\sqrt{Kn}}{\sqrt{\rho}\lambda^*(\mathbf{B})(\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta}))^{1.5}}\right), \tag{2}$$

where $\psi(\mathbf{P})$ *measures how well the eigenvalues of* $\mathbf{P}$ *can be packed into bins. The precise definition is deferred to Eq* (7)*, Sec 5 for ease of exposition.*

Later, we will show that $\psi(\mathbf{P}) \leq 2\min\{K, \kappa(\mathbf{P})\}^2$ in the worst case. But $\psi(\mathbf{P}) = O(1)$ if the eigenvalues of $\mathbf{P}$ can be divided into a constant number of bins where each bin has eigenvalues of the same order.

Remark 3.1 (Generalizing to low rank population matrices)**.** *In the supplementary material (Sec VI), we also establish a similar eigenvector deviation result for networks generated from general low rank population matrices.*

Remark 3.2 (Row-wise eigenvector error)**.** *Note that the above row-wise error immediately gives us an error bound on rows of* $\hat{\mathbf{V}}$,

$$\|\mathbf{e}_i^T(\hat{\mathbf{V}} - \mathbf{V}(\mathbf{V}^T\hat{\mathbf{V}}))\| = \|\mathbf{e}_i^T(\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T)\hat{\mathbf{V}}\| \leq \|\mathbf{e}_i^T(\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T)\|.$$

The $K \times K$ *matrix* $\mathbf{V}^T\hat{\mathbf{V}}$ *takes out the projection of* $\mathbf{V}$ *on* $\hat{\mathbf{V}}$ *from* $\hat{\mathbf{V}}$*. Note that while we use* $\mathbf{V}^T\hat{\mathbf{V}}$ *to align* $\mathbf{V}$ *and* $\hat{\mathbf{V}}$*, most existing literature uses its matrix sign function [27, 28]. An detailed example can be found in Lemma IV.1 in the supplementary material.*

The proof of Theorem 3.1 can be found in Sec 5. A key element in the proof is the delocalization of population eigenvectors.

Lemma 3.2 (Delocalization of population eigenvectors)**.** *We have that,* $\max_i \|\mathbf{e}_i^T\mathbf{V}\|^2 \leq 1/\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta})$ *and* $\min_i \|\mathbf{e}_i^T\mathbf{V}\|^2 \geq 1/(K\lambda_1(\mathbf{\Theta}^T\mathbf{\Theta}))$*.*

We defer the proof to the supplementary material (Sec II). Using this, we can prove the following.

Corollary 3.3 (Row-wise relative convergence)**.** *If Assumption 3.1 is satisfied, and furthermore,* $\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta}) = \Omega(n/K)$*,* $K = \Theta(1)$ *and* $\lambda^*(\mathbf{B}) = \Omega(1)$*, then*

$$\max_{i\in[n]} \frac{\left\|\mathbf{e}_i^T(\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T)\right\|}{\|\mathbf{e}_i^T\mathbf{V}\mathbf{V}^T\|} = \tilde{O}\left(\frac{1}{\sqrt{n\rho}}\right)$$

with probability at least $1 - O(Kn^{-2})$*.*

In concurrent work on MMSB models [19], analysis of empirical eigenvectors yields a suboptimal $\tilde{O}_P(1/\sqrt{n\rho^2})$ rate on the Frobenius norm of the overall deviation of the whole community membership matrix from its population counterpart, thereby proving consistency only in the regime where average degree grows faster than square root of n, not poly-logarithm of n. While concurrent developments on entry-wise eigenvector analysis [27, 29, 30] obtain the better $\tilde{O}_P(1/\sqrt{n\rho})$ rate, they either have a relatively worse dependence on $\lambda^*(\mathbf{B})$ or implicitly assume that the population eigenvalues are of the same order. In [30], the authors assume that K grows slower than poly-log of n. We show that the row-wise eigenvector bounds in [27] yield a worse dependence of $\lambda^*(\mathbf{B})$ than ours in the supplementary material (Sec IV). We achieve this better dependence on $\lambda^*(\mathbf{B})$ by a new construction in which we consider groups of population eigenvalues lying within specially constructed intervals, such that the ratio of the largest and smallest eigenvalues within any interval is controlled. Note that, if the population eigenvalues are of the same order, average expected degree in [27] can be a constant times $\log n$, whereas we require it to grow faster than $\log^2 n$.

We can show that our bound is tighter by an order of $1/\sqrt{n\rho}$ than a direct application of the concentration bounds for general singular subspaces established in [28] to the MMSB model. While it is possible to improve this bound by using our theoretical results and careful analysis similar to that of the ρ-correlated SBM graphs in [28], even then, our row-wise eigenspace error bound is tighter by a factor of $\sqrt{\rho}$ under a broad parameter regime, a detailed discussion of which is deferred to Sec V of the supplementary material along with derivations.

So far we have talked about row-wise bounds on empirical eigenspaces. But it seems cumbersome to apply our algorithms on the $n \times n$ $\hat{\mathbf{V}}\hat{\mathbf{V}}^T$ matrix. The following simple result shows that our algorithms return the same set of pure nodes using $\hat{\mathbf{V}}$ and $\hat{\mathbf{V}}\hat{\mathbf{V}}^T$ (proof in Sec VII of the supplementary material). Thus, for the algorithm we simply use $\hat{\mathbf{V}}$.

Lemma 3.4. *The pruning algorithm (Algorithm 2) and the SPA algorithm will return the same node indices on both* $\hat{\mathbf{V}}$ *(*$\hat{\mathbf{V}}^T$ *for SPA) and* $\hat{\mathbf{V}}\hat{\mathbf{V}}^T$*.*

3.2 Consistency of estimated quantities

We now use our row-wise eigenspace error bounds to analyze Algorithm 1. We do not analyze the pruning algorithm (Algorithm 2), since that requires distributional assumptions on the row-wise eigenvector errors. We need the following assumption.

Assumption 3.2. Assume $\lambda^*(\mathbf{B}) = \tilde{\Omega}\left(\frac{\psi(\mathbf{P})(\kappa(\mathbf{\Theta}^T\mathbf{\Theta}))^{1.5}K\sqrt{n}}{\sqrt{\rho}\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta})}\right)$.

Theorem 3.5. *Let* $\hat{\mathbf{\Theta}}$ *be obtained from Step 6 of Algorithm 1. We denote the row-wise eigenspace error from Theorem 3.1 as follows:*

$$\epsilon = \tilde{O}\left(\frac{\psi(\mathbf{P})\sqrt{Kn}}{\sqrt{\rho}\lambda^*(\mathbf{B})(\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta}))^{1.5}}\right).$$

If Assumptions 2.1, 3.1, and 3.2 hold, there exists a permutation matrix $\mathbf{\Pi}$ *such that with probability at least* $1 - O(K/n^2)$*,*

$$\max_{i \in [n]} \left\| \mathbf{e}_i^T \left(\hat{\mathbf{\Theta}} - \mathbf{\Theta}\mathbf{\Pi}\right)\right\| = O\left(\sqrt{\lambda_1(\mathbf{\Theta}^T\mathbf{\Theta})}\kappa(\mathbf{\Theta}^T\mathbf{\Theta})\epsilon\right), \tag{3}$$

$$\frac{1}{\rho}\|\hat{\rho}\hat{\mathbf{B}} - \rho\mathbf{\Pi}^T\mathbf{B}\mathbf{\Pi}\|_F = O\left(\frac{\kappa(\mathbf{\Theta}^T\mathbf{\Theta})\sqrt{K}n}{\sqrt{\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta})}}\epsilon\right). \quad (4)$$

The proof can be found in the supplementary material (Sec VII).

Under the conditions of Corollary 3.3, our row-wise eigenvector bound leads to $\tilde{O}(1/\sqrt{n\rho})$ rates of convergence of $\hat{\boldsymbol{\theta}}_i$ to $\boldsymbol{\theta}_i$. To our knowledge, this is the first such result for detecting mixed memberships in networks.

Remark 3.3 (Application to Stochastic Blockmodels)**.** *Theorem 3.1 can be used to establish strong consistency for Spectral Clustering for Stochastic Blockmodels. Here $\mathbf{\Theta}$ is a binary membership matrix with exactly one "1" on each row representing the cluster that node belongs to. So, $\mathbf{\Theta}^T\mathbf{\Theta}$ is a diagonal matrix whose diagonal elements (and eigenvalues) represent the sizes of the clusters. Consider the standard settings of $K = 2$ equal-sized clusters: $\rho\mathbf{B} = (p_n - q_n)\mathbf{I}_2 + q_n\mathbf{1}_2\mathbf{1}_2^T$, and $\lambda_1(\mathbf{\Theta}^T\mathbf{\Theta}) = \lambda_K(\mathbf{\Theta}^T\mathbf{\Theta}) = n/2$. By definition, $\max_{ij}\mathbf{B}_{ij} = 1$, so $\rho = p_n$ and $\lambda^*(\mathbf{B}) = (p_n - q_n)/p_n$. Our results imply exact recovery with probability greater than $1 - O(K/n^2)$, as long as $(p_n - q_n)/\sqrt{p_n} = \tilde{\Omega}(1/\sqrt{n})$. This matches the separation condition in existing literature [4, 31] up-to logarithmic factors. Note that, existing work on sharp threshold for exact recovery [32] assumes $p_n = a\log n/n$ and $q_n = b\log n/n$, where a, b are some constants. This implies $\lambda^*(\mathbf{B}) = (a-b)/a$. But we also allow $\lambda^*(\mathbf{B}) \ll 1$ in the regime that the average expected degree grows as poly-log of n.*

Remark 3.4 (Comparison to [16])**.** *In the latest version of [16] (updated Sep. 4th, 2019), the authors have added row-wise concentration results for eigenspaces. Their assumptions translate to $\kappa(\mathbf{\Theta}^T\mathbf{\Theta}) = \Theta(1)$. Furthermore, their assumption on the eigenvalues of $\mathbf{P}$ translates to $\psi(\mathbf{P}) = O(1)$ in our terminology. Thus, in this regime, our error bound on estimating $\boldsymbol{\theta}_i$ (converted to ℓ_1 norm by multiplying $\sqrt{K}$) is $\sqrt{K}$ worse than theirs up-to logarithmic factors. A detailed discussion can be found in Sec VIII of the supplementary material.*

3.3 Application to Dirichlet Prior

Now we consider the case where the $\{\boldsymbol{\theta}_i\}$ vectors are drawn from a Dirichlet distribution. We cannot directly use the bound in Theorem 3.5 since that bound depends on $\boldsymbol{\Theta}$. However, we can probabilistically bound the relevant functions of $\boldsymbol{\Theta}$.

Lemma 3.6. *If $\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$ with $\alpha_{\max} = \max_a \alpha_a$, $\alpha_{\min} = \min_a \alpha_a$ and $\nu := \alpha_0/\alpha_{\min}$,*

$$\mathrm{P}\left(\lambda_1(\boldsymbol{\Theta}^T\boldsymbol{\Theta}) \leq \frac{3n\left(\alpha_{\max} + \|\boldsymbol{\alpha}\|^2\right)}{2\alpha_0(1+\alpha_0)}\right) \geq 1 - K\exp\left(-\frac{n}{36\nu^2(1+\alpha_0)^2}\right)$$
$$\mathrm{P}\left(\lambda_K(\boldsymbol{\Theta}^T\boldsymbol{\Theta}) \geq \frac{n}{2\nu(1+\alpha_0)}\right) \geq 1 - K\exp\left(-\frac{n}{36\nu^2(1+\alpha_0)^2}\right)$$
$$\mathrm{P}\left(\kappa(\boldsymbol{\Theta}^T\boldsymbol{\Theta}) \leq 3\frac{\alpha_{\max} + \|\boldsymbol{\alpha}\|^2}{\alpha_{\min}}\right) \geq 1 - 2K\exp\left(-\frac{n}{36\nu^2(1+\alpha_0)^2}\right)$$

where $\kappa(.)$ is the condition number of a matrix.

Assumption 3.3 (Parameters of Dirichlet)**.** Assume for some constant $\xi > 1$, we have,

$$\nu := \frac{\alpha_0}{\alpha_{\min}} \leq \frac{\min(\sqrt{\frac{n}{27\log n}}, n\rho)}{2(1+\alpha_0)}, \frac{\lambda^*(\mathbf{B})}{\nu} \geq \frac{8(1+\alpha_0)(\log n)^{\xi}}{\sqrt{n\rho}}.$$

One can easily check that under Assumption 3.3, by Lemma 3.6, Assumption 3.1 is satisfied with probability at least $1 - O(Kn^{-3})$. When α_0 is a constant, the condition on $\lambda^*(\mathbf{B})/\nu$ immediately implies $\rho n = \Omega((\log n)^{2\xi})$, since $\lambda^*(\mathbf{B}) \leq \|\mathbf{B}\| \leq K \leq \nu$. Since the expected average degree is $O(n\rho)$, these conditions mean that the average degree must grow faster than poly-log of n. This is the most common regime where most consistency results on network clustering are shown [11, 2, 4]. The magnitude of α_0 limits the amount of overlap between communities. As noted also by [11], in many real world applications nodes belong a few communities – so a constant or slowly growing α_0 is a reasonable assumption. For example, the conditions imposed by [16] on $\boldsymbol{\Theta}$ can be translated to $\alpha_0 = O(1)$ in the context of MMSB models. Note that, our results can handle large α_0, but at the cost of a worse error bound.

Our conditions also allow K to grow with n. If $\rho = O(1)$, $\alpha_0 = O(1)$, and $\lambda^*(\mathbf{B}) = \Theta(1)$, then K can grow with $\sqrt{n}$, up to poly-log terms (using the fact that $\nu \geq K$). Now, consider the common case of a simple MMSB model with K communities: $\rho\mathbf{B} = (p_n - q_n)\mathbf{I}_K + q_n\mathbf{1}_K\mathbf{1}_K^T$ and $\boldsymbol{\alpha} = \alpha_0\mathbf{1}_K/K$. Since the largest element of $\mathbf{B}$ is one by definition, we have $\rho = p_n$. This yields $\lambda^*(\mathbf{B}) = (p_n - q_n)/p_n$. We also have $\nu = K$. Hence the second condition can be interpreted as a lower bound on cluster separation: $(p_n - q_n)/\sqrt{p_n} = \tilde{\Omega}\left(K/\sqrt{n}\right)$. This matches the separation condition in existing literature [11].

We now show error bounds on $\hat{\mathbf{\Theta}}$ and $\hat{\mathbf{B}}$, when $\boldsymbol{\theta}_i$ is drawn from a Dirichlet distribution. For ease of exposition, we focus on the case with similar α_i and $\alpha_0 = O(1)$. This corresponds to roughly-balanced communities with limited overlap.

Corollary 3.7. *Let $\boldsymbol{\theta}_i \sim \text{Dirichlet}(\alpha)$ with $\max_a \alpha_a \leq C \min_a \alpha_a$ for some constant $C \geq 1$, $\alpha_0 = O(1)$. If Assumptions 2.1 and 3.3 hold, and $\lambda^*(\mathbf{B}) = \tilde{\Omega}(\frac{\min\{K, \kappa(\mathbf{B})\}^2 K^2}{\sqrt{n\rho}})$, there exists a permutation matrix $\mathbf{\Pi}$ such that with probability at least $1 - O(K/n^2)$,*

$$\max_{i \in [n]} \left\| \mathbf{e}_i^T \left(\hat{\mathbf{\Theta}} - \mathbf{\Theta}\mathbf{\Pi} \right) \right\| = \tilde{O}\left(\frac{\min\{K, \kappa(\mathbf{B})\}^2 K^{1.5}}{\sqrt{\rho n}\lambda^*(\mathbf{B})} \right), \tag{5}$$

$$\frac{1}{\rho}\|\hat{\rho}\hat{\mathbf{B}} - \rho\mathbf{\Pi}^T\mathbf{B}\mathbf{\Pi}\|_F = \tilde{O}\left(\frac{\min\{K, \kappa(\mathbf{B})\}^2 K^3}{\sqrt{\rho n}\lambda^*(\mathbf{B})} \right). \tag{6}$$

Remark 3.5 (Error bound on $\hat{\mathbf{\Theta}}$ as a whole)**.** *Note that we can get the Frobenius norm of the error for the whole matrix by directly accumulating the row-wise error bounds. With all other hyperparameters and parameters like α_0, ν, K and $\lambda^*(\mathbf{B})$ held constant, our Frobenius-norm bound on $\hat{\mathbf{\Theta}}$ is tighter by a factor of $\sqrt{\rho}$ than that in [13, 19], which allows the analysis to work on networks with average degree $\tilde{\Omega}(\log n)$ rather than $\tilde{\Omega}(\sqrt{n})$.*

Anandkumar et al. [11] have the same degree regime as ours, but their algorithm assumes prior knowledge of α_0. Our bound has a worse dependence on K, α_0 and ν compared to them. To be concrete, when $\kappa(\mathbf{B}) = \Theta(1)$ and the clusters are balanced with mild overlap, i.e. $\max_a \alpha_a / \min_a \alpha_a = \Theta(1)$ and $\alpha_0 = O(1)$, we have an additional $\sqrt{K}$ factor (after converting our Frobenius norm bound to ℓ_1 norm by multiplying $\sqrt{Kn}$ and theirs by K

to get the error of the whole $\hat{\mathbf{\Theta}}$ matrix). In the worst case, our bound has an additional $K^2\sqrt{\nu}(1+\alpha_0)$ factor. We provide more details in the supplementary material (Sec IX).

4 Experimental results

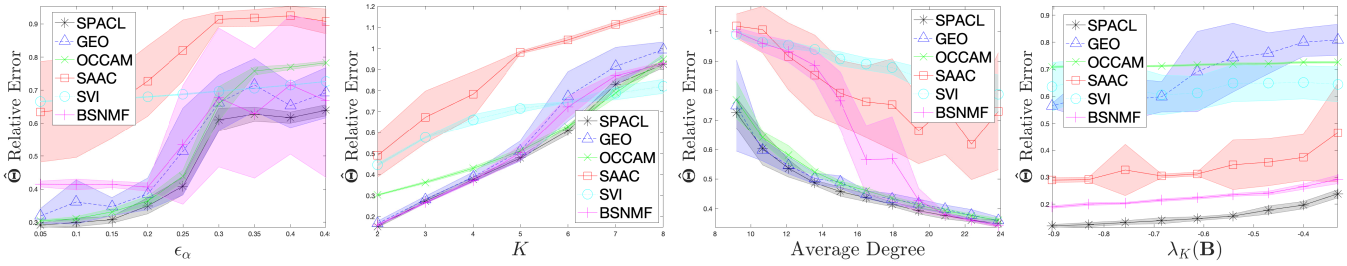
We present both simulation results and real data experiments to compare SPACL with existing algorithms for overlapping network models. We compare with the Stochastic Variational Inference algorithm (SVI) [6], a geometric algorithm for non-negative matrix factorization for MMSB models with equal Dirichlet parameters (GeoNMF) [13], Bayesian SNMF (BSNMF) [10], the OCCAM algorithm [14] for recovering mixed memberships, and the SAAC algorithm [18].[3] For real data experiments we use two large datasets (with up to 100,000 nodes) from the DBLP corpus. One of these is assortative ($\mathbf{B}$ has positive eigenvalues) and one which is disassortative ($\mathbf{B}$ has negative eigenvalues). We show that for the disassortative setting, SPACL significantly outperforms other methods.

4.1 Simulations

In this section, we investigate the sensitivity of SPACL and competing algorithms to the Dirichlet parameter $\boldsymbol{\alpha}$, the number of communities K, the sparsity control parameter ρ, and to the eigenvalues of $\mathbf{B}$. Our simulated graphs have $n = 5000$, unless specified otherwise. We show the relative error ($\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F / \|\mathbf{\Theta}\|_F$) of different methods, averaged over 10 random runs in a range of parameter settings. The largest row-wise relative error has similar trends. Further results for varying $\mathbf{B}$ are presented in the supplementary material (Sec XI).

Some algorithms have an underlying model that is slightly different from MMSB. We handle these as follows. For OCCAM, we normalize each row of $\mathbf{\Theta}$ by its ℓ_2 norm, thereby

[3] We were unable to run the GPU implementation of [11] since a required library CULA is no longer open source. We could not get good results with the CPU implementation with default settings.



(A) (B) (C) (D)

Figure 2: (A) Error against ϵ_α: $\boldsymbol{\alpha} = (0.5 - \epsilon_\alpha, 0.5, 0.5 + \epsilon_\alpha)$. (B) Error against increasing K. (C) Error against increasing ρ (D) Error against $\lambda_K(\mathbf{B})$.

absorbing the ℓ_2 norm in the degree parameter. For SAAC, we threshold elements of $\boldsymbol{\Theta}$ by $1/K$ to get a binary matrix. For BSNMF, no adjustment is necessary. However, note that BSNMF assumes $\mathbf{B}$ is identity.

Changing $\boldsymbol{\alpha}$: In Fig 2 (A) we use $\boldsymbol{\alpha} = (0.5 - \epsilon_\alpha, 0.5, 0.5 + \epsilon_\alpha)$ and plot the relative error against ϵ_α. We set $K = 3$, $\rho = 0.15$, $\mathbf{B}_{ii} = 1$, $i \in [K]$, $\mathbf{B}_{ij} = 0.5$ for $i \neq j$. Recall that for skewed $\boldsymbol{\alpha}$ we get unbalanced cluster sizes. SPACL is better than SAAC, SVI, BSNMF and GeoNMF, and also more stable (small variance). For imbalanced clusters (large ϵ_α), SPACL also outperforms OCCAM.

Changing K: In Fig 2 (B) we plot relative error against increasing K. We use $\rho = 0.1$, $\boldsymbol{\alpha}_i = 3/K = 1$, $\mathbf{B}_{ii} = 1$, $i \in [K]$, $\mathbf{B}_{ij} = 0.2$ for $i \neq j$. We can see that SPACL outperforms SAAC, and is more stable than BSNMF and GeoNMF. When K is very large (>7), everyone performs poorly. When K is small (<5), SPACL works much better than OCCAM and SVI. However, when K is moderately large, OCCAM is slightly better than SPACL. This is because in those cases, the eigenspaces do not concentrate very well, and estimating $\hat{\boldsymbol{\Theta}}$ with cluster centroids (as in OCCAM) seems to reduce the noise.

Changing sparsity: We set $\boldsymbol{\alpha} = (0.4, 0.4, 0.4)$, $\mathbf{B}_{ii} = 1$, $i \in [K]$, $\mathbf{B}_{ij} = 0.05$ for $i \neq j$. We increase ρ from 0.005 to 0.013, Fig 2 (C) shows the result. We see that, the error of SPACL

Table 2: Statistics for author-author (Mono) and bipartite paper-author (Bi) graphs.

Dataset	DBLP1		DBLP2		DBLP3		DBLP4		DBLP5	
	Mono	Bi	Mono	Bi	Mono	Bi	Mono	Bi	Mono	Bi
# nodes n	30,566	103,660	16,817	50,699	13,315	42,288	25,481	53,369	42,351	81,245
# communities K	6	12	3	6	3	6	3	6	4	8
Average Degree	8.9	3.4	7.6	3.4	8.5	3.6	5.2	2.6	6.8	3.0
Overlap %	18.2	6.3	14.9	5.6	21.1	5.7	14.4	6.9	18.5	9.7

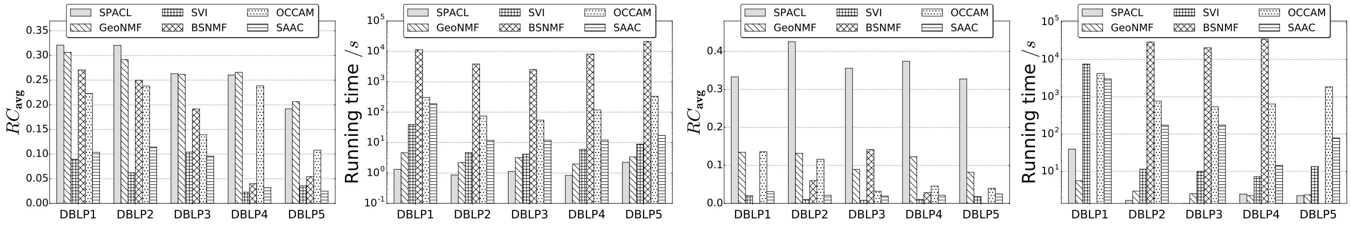
is smaller than or similar to that of the best performing algorithm among the others. In addition, it also has smaller variance.

Changing $\lambda_K(\mathbf{B})$: We conclude the simulations with experiments on $\mathbf{B}$ with negative eigenvalues. We generate $\mathbf{B}$ so that the smallest eigenvalue $\lambda_K(\mathbf{B})$ of $\mathbf{B}$ is negative. We set $\mathbf{B} = \begin{bmatrix} 1 & 0.2 & 0.1 \\ 0.2 & 0.5 & 0.075 \cdot i \\ 0.1 & 0.075 \cdot i & 0 \end{bmatrix}$ and vary $i \in [15]$. As i grows, $\lambda_K(\mathbf{B})$ becomes more negative. We set $K = 3$, $\rho = 0.15$, $\boldsymbol{\alpha} = (1/3, 1/3, 1/3)$. In the plot of relative error against $\lambda_K(\mathbf{B})$ (Fig 2 (D)), we see that SPACL is much better than others over the entire parameter range.

4.2 Real Data

We use the two types of DBLP networks obtained from the DBLP dataset[4], where each ground truth community is a group of conferences on one topic. The author-author networks were used in [13]; in this paper we also conduct experiments on the bipartite networks by using both papers and authors as nodes. Each community is split into two, the paper community and the author community. The papers are pure nodes since they belong to one conference and hence one community, whereas the authors may belong to more than one community, since they often publish in many conferences. The details of the subfields

[4] `http://dblp.uni-trier.de/xml/`



(A) (B) (C) (D)

Figure 3: RC_{avg} and running time (log scale) on DBLP Mono (A, B) and DBLP Bi (C, D). BSNMF was out of memory for bipartite versions of DBLP1 and DBLP5.

can be found in [13]. We have two simple preprocessing steps for the adjacency matrix: 1) delete nodes that do not belong to any community; 2) delete nodes with zero degree. The statistics of the network are in Table 2, which show that despite being sparse, the networks have large overlaps between communities. The amount of overlap is measured by the number of overlapping nodes divided by n.

Implementation details: For real world networks, specially the bipartite networks, when average degree of graphs with 100,000 nodes is smaller than four, some nodes may have extremely small values of $\hat{\boldsymbol{\Theta}}$ and the corresponding rows may in fact become zero after thresholding. For those we essentially cannot make any prediction. This is why for Step 7 of Algorithm 1, we threshold all values smaller than 10^{-12} to zero and we do not normalize rows which are all zeros. This does not make any difference for simulations, but for the real world networks, this stabilizes the results.

Evaluation Metric: For author nodes, we construct the corresponding row of $\boldsymbol{\Theta}$ by normalizing the number of papers an author has in different ground truth communities. We present the averaged Spearman rank correlation coefficients (RC) between $\boldsymbol{\Theta}(:,a)$, $a \in [K]$ and $\hat{\boldsymbol{\Theta}}(:,\sigma(a))$, where σ is a permutation of $[K]$. The formal definition is:

$$\text{RC}_{\text{avg}}(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) = \frac{1}{K} \max_{\sigma} \sum_{i=1}^{K} \text{RC}(\hat{\boldsymbol{\Theta}}(:,i), \boldsymbol{\Theta}(:,\sigma(i))).$$

Note that $\text{RC}_{\text{avg}}(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \in [-1, 1]$, and higher is better. Since SAAC returns binary assignment, we compute its $\mathrm{RC}_{\mathrm{avg}}$ against the binary ground truth.

Performance: We report the $\mathrm{RC}_{\mathrm{avg}}$ score in Fig 3. The superior performance of SPACL on the paper-author networks over the author-author networks can be explained by the fact that the bipartite network retains information that is lost when the author-author networks are constructed. Also SPACL outperforms all other methods on bipartite networks, since these are disassortative and the corresponding $\mathbf{B}$ will have negative eigenvalues. On co-authorship graphs, SPACL performs comparably to GeoNMF, while the other methods are worse. Both SPACL and GeoNMF are much faster than the competing algorithms.

5 Analysis

Here we present the main proof idea of Theorem 3.1. We equate the difference in empirical and population eigenspaces with the Cauchy integral of a matrix resolvent. To bound the row-wise difference in eigenspaces, we have to specify the contours for the complex integration and then bound a matrix series expansion. Our contours are carefully chosen by a discretization of the eigenvalues of $\mathbf{P}$. This yields an error bound with the proper dependence on $\lambda^*(\mathbf{B})$ and $\kappa(\mathbf{P})$. The matrix series expansion is controlled by upper-bounding the first $\log n$ terms and the rest separately, where the partial sum for the first $\log n$ terms is controlled by applying the union bound. This is a common technique in perturbation analysis [33]. A similar strategy is also used in concurrent work [29]. We defer proofs of some of the technical lemmas to the supplementary material (Sec III).

5.1 Eigenspace Row-wise Concentration

Before presenting the analysis of the row-wise error-bounds of empirical eigenvectors, we present a discretization scheme of the population eigenvalues, which later helps in getting a better dependence of the overall row-wise error on the smallest singular value of $\mathbf{P}$, which can also be thought of as the separation between blocks.

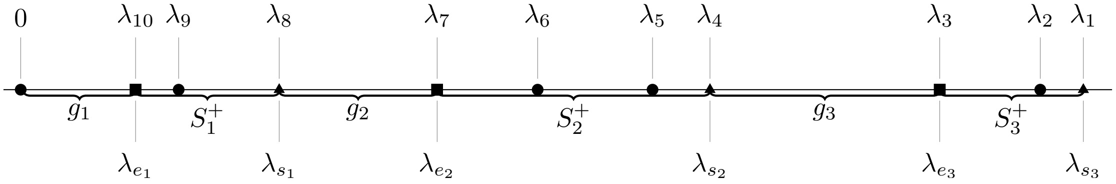


Figure 4: An illustration of Definition 5.1.

Definition 5.1 (A discretization of eigenvalues)**.** Let us divide the eigenvalues of $\mathbf{P}$ into the positive ones (S^+) and negative ones (S^-). We start with the smallest eigenvalue in S^+. Denote this by λ_+^*. We set the gap $g_1 = \lambda_+^*$ and keep moving through the eigenvalues in S^+ in increasing order until we find two consecutive eigenvalues which have gap $g_2 > g_1$. We repeat this until all eigenvalues in S^+ are covered. Then every pair of consecutive eigenvalues in the k^{th} interval is within gap g_k, and g_k grows with k. We define s_k and e_k as the starting and ending index of eigenvalues of the k^{th} interval. Formally, the k^{th} interval of positive eigenvalues is the set

$$S_k^+ = \{\lambda_{s_k}, \dots \lambda_{e_k} \in S^+ : \lambda_i - \lambda_{i+1} \le g_k \text{ for } s_k \le i \le e_k \text{ , } \lambda_{e_{k+1}} - \lambda_{s_k} > g_k\}.$$

Let $n_k := |S_k^+|$ be the number of eigenvalues in the k^{th} interval. Fig 4 shows an example.

Let the number of intervals with positive eigenvalues be I^+. Note that $\lambda^*(\mathbf{P}) \le \lambda_+^* \le g_1 < g_2 \cdots < g_{I^+}$. By a similar splitting process for the negative eigenvalues in S^-, we can define I^-, s_{-k}, e_{-k}, and g_{-k}. Let $\lambda_{s_0} = 0$ and define

$$\psi(\mathbf{P}) := \sum_{k=1}^{I^+} \frac{\lambda_{s_k}(\lambda_{s_k} - \lambda_{s_{k-1}})}{g_k^2} + \sum_{k=1}^{I^-} \frac{\lambda_{s_{-k}}(\lambda_{s_{-k}} - \lambda_{s_{-k+1}})}{g_{-k}^2}. \tag{7}$$

$\psi(\mathbf{P})$ measures how tightly the eigenvalues of $\mathbf{P}$ can be packed together.

The above discretization lets us control the ratio of the largest eigenvalue in each interval and the gap between an interval and the next. This in turn helps bound $\psi(\mathbf{P})$.

Lemma 5.1. *In general, $\psi(\mathbf{P}) \leq 2\min\{K, \kappa(\mathbf{P})\}^2$. If the eigenvalues of $\mathbf{P}$ can be divided into a constant number of bins where eigenvalues in each bin are of the same order, $\psi(\mathbf{P}) = O(1)$.*

For ease of exposition, we shall henceforth work with just the positive eigenvalues in our proofs, and use I for the number of intervals. The proofs go through for negative eigenvalues using a nearly identical argument. We emphasize that the statement of Theorem 3.1 considers both positive and negative eigenvalues.

In order to prove Theorem 3.1, we will first introduce the notion of matrix resolvents and useful identities on resolvents.

Definition 5.2. A resolvent of a matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ is defined as $\mathbf{G}_{\mathbf{M}}(z) = (\mathbf{M} - z\mathbf{I})^{-1}$, where $z \notin \{\lambda_i(\mathbf{M})\}_{i=1}^n$. We can also write the resolvent as $\sum_{i=1}^n \frac{\mathbf{v}_i(\mathbf{M})\mathbf{v}_i(\mathbf{M})^T}{\lambda_i(\mathbf{M})-z}$, where $\mathbf{v}_i(\mathbf{M})$ is the i^{th} eigenvector of $\mathbf{M}$.

Let us define:

$$\mathbf{E}_z = \text{diag}\left(\left\{\frac{\lambda_i}{z(\lambda_i - z)}\right\}_{i=1}^K\right), \qquad \mathbf{M}_z = \mathbf{V}\mathbf{E}_z\mathbf{V}^T. \tag{8}$$

As we see below this matrix is an integral part of the resolvent of the expectation matrix $\mathbf{P}$.

$$\mathbf{G}_{\mathbf{P}}(z) = \sum_{i=1}^n \frac{\mathbf{v}_i\mathbf{v}_i^T}{\lambda_i - z} = \sum_{i=1}^K \mathbf{v}_i\mathbf{v}_i^T\left(\frac{1}{\lambda_i - z} + \frac{1}{z}\right) - \frac{\mathbf{I}}{z} = \mathbf{M}_z - \frac{\mathbf{I}}{z} \tag{9}$$

We will use a standard technique to compute eigenspaces of matrices (also used in [34] Lemma A.2). Consider an interval (a, b) such that no eigenvalue of a symmetric matrix $\mathbf{M}$ equals a or b. Now consider a rectangular contour $\mathcal{C}$ in the complex plane which passes through $a + \gamma\sqrt{-1}, a - \gamma\sqrt{-1}, b - \gamma\sqrt{-1}, b + \gamma\sqrt{-1}$ in counter clockwise direction, where $\gamma > 0$. From the Cauchy integration formula, we know that

$$\frac{1}{2\pi\sqrt{-1}}\oint_{\mathcal{C}} \mathbf{G}_{\mathbf{M}}(z)dz = -\sum_{i:\lambda_i(\mathbf{M})\in(a,b)} \mathbf{v}_i(\mathbf{M})\mathbf{v}_i(\mathbf{M})^T \tag{10}$$

Definition 5.3. We consider a sequence of non-overlapping contours $\mathcal{C}_k, k \in [I]$ $(I \leq K)$ created using a_k, b_k, γ_k, where $\|\mathbf{A} - \mathbf{P}\| < a_k < b_k$, and none of the eigenvalues of $\mathbf{A}$ or $\mathbf{P}$ equal a_k, b_k for $k \in [I]$.

Let $\mathbf{V}_k$ denote the $n \times n_k$ matrix with the eigenvectors of $\mathbf{P}$ corresponding to eigenvalues in (a_k, b_k). Similarly let $\hat{\mathbf{V}}_k$ denote the eigenvectors of $\mathbf{A}$ corresponding to eigenvalues in (a_k, b_k). Hence, using the Cauchy integration formula (10), we have:

$$\mathbf{V}_k \mathbf{V}_k^T - \hat{\mathbf{V}}_k \hat{\mathbf{V}}_k^T = \frac{1}{2\pi\sqrt{-1}} \oint_{\mathcal{C}_k} \left(\mathbf{G}_\mathbf{A}(z) - \mathbf{G}_\mathbf{P}(z)\right) dz \tag{11}$$

Furthermore, it is not hard to check that, $\forall x \in [n]$,

$$\mathbf{e}_x^T \left(\mathbf{V}_k \mathbf{V}_k^T - \hat{\mathbf{V}}_k \hat{\mathbf{V}}_k^T\right) = \frac{1}{2\pi\sqrt{-1}} \oint_{\mathcal{C}_k} \mathbf{e}_x^T \left(\mathbf{G}_\mathbf{A}(z) - \mathbf{G}_\mathbf{P}(z)\right) dz \tag{12}$$

We bound the Frobenius norm of the above quantities using Lemma 5.2 below.

Lemma 5.2. *For contours in Definition 5.3, we have:*

$$\left\|\mathbf{e}_x^T \sum_{k=1}^{I} (\mathbf{V}_k \mathbf{V}_k^T - \hat{\mathbf{V}}_k \hat{\mathbf{V}}_k^T)\right\| \leq \sum_{k=1}^{I} \frac{b_k - a_k + 2\gamma_k}{\pi} \max_{z \in \mathcal{C}_k} (P_1(z) + P_2(z)), \tag{13}$$

$$\begin{aligned} \textit{where} \quad & P_1(z) = |z| \|\mathbf{G}_\mathbf{A}(z)\| \|\mathbf{A} - \mathbf{P}\| \|\mathbf{E}_z\| \|\mathbf{e}_x^T \mathbf{G}_{\mathbf{A}-\mathbf{P}}(z) \mathbf{V}\|, \\ & P_2(z) = \|\mathbf{e}_x^T \mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)(\mathbf{A} - \mathbf{P})\mathbf{V}\|_F \|\mathbf{E}_z\|. \end{aligned}$$

Now we need to:

1. Define contours and events so that the LHS of Eq (13) covers the whole eigenspace.

2. Bound $P_1(z)$ and $P_2(z)$ over each contour, under these events. This requires bounds on $\|\mathbf{e}_x^T \mathbf{G}_{\mathbf{A}-\mathbf{P}}(z) \mathbf{v}_i\|$, and $\|\mathbf{e}_x^T \mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)(\mathbf{A} - \mathbf{P})\mathbf{v}_i\|$, where $\mathbf{v}_i$ denotes the i^{th} column of $\mathbf{V}$. We also need $\|\mathbf{E}_z\|$, $\|\mathbf{G}_\mathbf{A}(z)\|$, $\|\mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)\|$, etc. This requires us to bound $|\mathbf{e}_i^T \mathbf{H}^t \mathbf{v}_i|$ for $t \leq \log n$, where $\mathbf{H} := (\mathbf{A} - \mathbf{P})/\sqrt{n\rho}$. For $t = 1$, we prove the following lemma, which uses the fact that $\mathbf{V}$ is delocalized with high probability (see Lemma 3.2).

Lemma 5.3. *Let $\mathbf{v}_k$ denote the k^{th} population eigenvector of $\mathbf{P}$. If Assumption 3.1 is satisfied, for a fixed i,* $\mathrm{P}\left(\exists k \in [K], |\mathbf{e}_i^T\mathbf{H}\mathbf{v}_k| \geq 4\log n\|\mathbf{v}_k\|_\infty\right) = O(K/n^3)$.

For $1 < t \leq \log n$, we adapt a crucial result from [33].

Lemma 5.4. *Let $\mathbf{H} := (\mathbf{A} - \mathbf{P})/\sqrt{n\rho}$. As long as Assumption 3.1 is satisfied for some constant ξ, for any fixed vector $\mathbf{v}$, for a fixed i and for $1 < t \leq \log n$,*

$$\mathrm{P}\left(|\mathbf{e}_i^T\mathbf{H}^t\mathbf{v}| \leq (\log n)^{t\xi}\|\mathbf{v}\|_\infty\right) \geq 1 - \exp(-(\log n)^{\xi}/3).$$

Proofs of Lemmas 5.2, 5.3, and 5.4 are in the supplementary material (Secs III.2, III.3, and III.4 respectively).

We will now define some events, which will be used extensively to show that the contours cover all population and empirical eigenvalues, and to bound $P_1(z)$ and $P_2(z)$ in Eq (13). We will use $\mathcal{E}$ to denote an event and $\bar{\mathcal{E}}$ to denote its compliment. Let $\mathbf{v}_k$ be the k^{th} population eigenvector. Under Assumption 3.1, for $t \leq \log n$,

$$\begin{aligned}
&\mathcal{E}' := \{\|\mathbf{A} - \mathbf{P}\| \leq C\sqrt{n\rho}\} && \mathrm{P}(\bar{\mathcal{E}'}) \overset{(i)}{\leq} n^{-3} \\
&\mathcal{E}_1 := \left\{\left|\mathbf{e}_i^T\mathbf{H}\mathbf{v}_k\right| \leq 4\log n\|\mathbf{v}_k\|_\infty, \forall k \in [K]\right\} && \mathrm{P}(\bar{\mathcal{E}}_1) \overset{(ii)}{\leq} O\left(K/n^3\right) \\
&\mathcal{E}_t := \left\{\left|\mathbf{e}_i^T\mathbf{H}^t\mathbf{v}_k\right| \leq (\log n)^{t\xi}\|\mathbf{v}_k\|_\infty, \forall k \in [K]\right\} && \mathrm{P}(\bar{\mathcal{E}}_t) \overset{(iii)}{\leq} K\exp(-(\log n)^{\xi}/3), 1 < t \leq \log n
\end{aligned} \tag{14}$$

For any community membership matrix $\mathbf{\Theta}$, $\mathrm{P}(\bar{\mathcal{E}}'|\mathbf{\Theta})$ can be bounded directly using Theorem 5.2 of [2], since Assumption 3.1 requires that $n\rho = \Omega(\log n)$. Hence step (i) follows. Steps (ii) and (iii) follow from Lemmas 5.3 and 5.4 respectively. To denote order notation conditioned on event $\mathcal{E}'$, we will use, $X \overset{\mathcal{E}'}{=} O(.)$ to denote, $\mathrm{P}(X = O(.)) = \mathrm{P}(\mathcal{E}')$.

Picking the contours $\mathcal{C}_k$: Consider the discretization in Definition 5.1. For the k^{th} interval, use $\gamma_k = g_k/4$, $a_k = \max(\lambda_{e_k} - g_k/2, (1+c)\|\mathbf{A} - \mathbf{P}\|)$, for some $c > 0$ and $b_k = \lambda_{s_k} + g_k/2$. If $b_k \leq a_k$, we ignore the contour. If either a_k or b_k equal an eigenvalue of $\mathbf{A}$ or $\mathbf{P}$, for any $\epsilon > 0$, they can be perturbed by at most ϵ to guarantee that they do not coincide with eigenvalues of $\mathbf{A}$ or $\mathbf{P}$. This is possible because for a given n, the set $\{\mathbf{A} \mid \mathbf{A} \in \{0,1\}^{n\times n}\}$ is finite.

Now we bound $\|\mathbf{G_A}(z)\|$, $\|\mathbf{G_P}(z)\|$, $\|\mathbf{E}_z\|$ and $\|\mathbf{G_{A-P}}(z)\|$. Since the gap between the smallest eigenvalue (in magnitude) of the k^{th} interval and the largest eigenvalue in the $(k-1)^{th}$ interval is g_k, and by construction (Definition 5.1) $\lambda^*(\mathbf{P}) \leq g_1 < g_2 < \dots$, and $\lambda_{e_k} \geq g_k$, we note that for each contour $\mathcal{C}_k$, conditioned on $\mathcal{E}'$, $|z|$ can be upper and lower bounded as follows.

$$|z| \leq \sqrt{b_k^2 + \gamma_k^2} \leq b_k + \gamma_k = \lambda_{s_k} + 3g_k/4 \tag{15}$$

$$|z| \geq \max((1+c)\|\mathbf{A}-\mathbf{P}\|, |\lambda_{e_k} - g_k/2|) \geq |\lambda_{e_k} - g_k/2| \geq g_k/2 \tag{16}$$

$$|z - \lambda_i| \geq g_k/2, \qquad |z - \hat{\lambda}_i| \overset{(i)}{\geq} g_k/2 - O(\sqrt{n\rho}) \tag{17}$$

$$\|\mathbf{M}_z\| = \|\mathbf{E}_z\| \leq \max_i \left|\frac{1}{\lambda_i - z} + \frac{1}{z}\right| = O\left(\frac{1}{g_k}\right) \tag{18}$$

For all $i \in [n]$ and for all $z \in \mathcal{C}_k$, Eq (18) follows from Eqs (8), (16) and (17) and Assumption 3.1.

Step (i) in Eq (17), uses $|\hat{\lambda}_i - \lambda_i| \overset{\mathcal{E}'}{=} O(\sqrt{n\rho})$ via Weyl's inequality. Finally using Eqs (16), (17) and (18) we also have for all $z \in \mathcal{C}_k$, conditioned on $\mathcal{E}'$,

$$\|\mathbf{G_P}(z)\| = O\left(\frac{1}{g_k}\right) \quad \|\mathbf{G_A}(z)\| \leq \left\|\sum_i \frac{\hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T}{\hat{\lambda}_i - z}\right\| = O\left(\frac{1}{g_k - O(\sqrt{n\rho})}\right) \tag{19}$$

Now conditioned on $\mathcal{E}'$, Eq (19) gives:

$$\|\mathbf{G_A}(z) - \mathbf{G_P}(z)\| \leq \|\mathbf{G_P}(z)\|\|\mathbf{P}-\mathbf{A}\|\|\mathbf{G_A}(z)\| = O\left(\frac{\sqrt{n\rho}}{g_k}\right) O\left(\frac{1}{g_k - O(\sqrt{n\rho})}\right) \tag{20}$$

Now we will bound the RHS of Eq (13) in Lemma 5.2.

Lemma 5.5. *Let $\mathbf{v}_i$ denote the i^{th} column of $\mathbf{V}$. Let Assumption 3.1 be satisfied for some constant ξ. Consider the events defined in Eq (14). Conditioned on $\bigcap_{t=1}^{\log n} \mathcal{E}_t \cap \mathcal{E}'$,*

$$|\mathbf{e}_x^T \mathbf{G_{A-P}}(z) \mathbf{v}_i| = \frac{O\left(\|\mathbf{v}_i\|_\infty + n^{-2\xi}\right)}{\lambda^*(\mathbf{P})}.$$

$$|\mathbf{e}_x^T \mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)(\mathbf{A}-\mathbf{P})\mathbf{v}_i| = O\left(\frac{\sqrt{n\rho}\left((\log n)^{\xi}\|\mathbf{v}_i\|_\infty + n^{-2\xi}\right)}{\lambda^*(\mathbf{P})}\right).$$

Proof. First note that by construction $\forall z \in \mathcal{C}_k, \forall k,\ |z| \geq a_k > \|\mathbf{A}-\mathbf{P}\|$, we have the following series expansion for $\mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)$,

$$\mathbf{G}_{\mathbf{A}-\mathbf{P}}(z) = -\frac{1}{z}\sum_{t\geq 0}\left(\frac{\mathbf{A}-\mathbf{P}}{z}\right)^t. \tag{21}$$

For $\mathbf{H}$ defined in Lemma 5.4, for $1 \leq t \leq \log n$, conditioned on $\mathcal{E}_t$, $t \geq 1$,

$$\left|\frac{\mathbf{e}_x^T(\mathbf{A}-\mathbf{P})^t\mathbf{v}_i}{z^t}\right| = \left|\mathbf{e}_x^T\mathbf{H}^t\mathbf{v}_i\frac{(\sqrt{n\rho})^t}{z^t}\right| \leq \begin{cases}\left(\frac{\sqrt{n\rho}(\log n)^{\xi}}{|z|}\right)^t \|\mathbf{v}_i\|_\infty & t \leq \log n\\ \left(\frac{\|\mathbf{A}-\mathbf{P}\|}{|z|}\right)^t & t > \log n\end{cases}, \tag{22}$$

where we use Lemmas 5.3 and 5.4. It is easy to verify that the above holds for $t = 0$. As Assumption 3.1 gives:

$$\lambda^*(\mathbf{P}) \overset{\mathcal{E}'}{\geq} 4\sqrt{n\rho}(\log n)^{\xi} \quad \Rightarrow \quad \max_{k, z\in\mathcal{C}_k}\frac{\sqrt{n\rho}(\log n)^{\xi}}{|z|} \overset{\mathcal{E}'}{\leq} \frac{1}{2} \tag{23}$$

Conditioned on $\bigcap_{t=1}^{\log n}\mathcal{E}_t \cap \mathcal{E}'$, Eqs (21) and (22) give:

$$\begin{aligned}
\max_{k,z\in\mathcal{C}_k}|\mathbf{e}_x^T\mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)\mathbf{v}_i| &\leq \max_{k,z\in\mathcal{C}_k}\frac{1}{|z|}\left|\sum_{t=0}^{\infty}\frac{\mathbf{e}_x^T(\mathbf{A}-\mathbf{P})^t}{z^t}\mathbf{v}_i\right|\\
&\leq \max_{k,z\in\mathcal{C}_k}\frac{1}{|z|}\sum_{t=0}^{\log n}\left|\frac{\mathbf{e}_x^T(\mathbf{A}-\mathbf{P})^t\mathbf{v}_i}{z^t}\right| + \max_{k,z\in\mathcal{C}_k}\frac{1}{|z|}\sum_{t>\log n}\left|\frac{\mathbf{e}_x^T(\mathbf{A}-\mathbf{P})^t\mathbf{v}_i}{z^t}\right|\\
\text{(Eqs (22) and (23))}\quad &\leq \max_{k,z\in\mathcal{C}_k}\frac{\|\mathbf{v}_i\|_\infty}{|z|-\sqrt{n\rho}(\log n)^{\xi}} + \max_{k,z\in\mathcal{C}_k}\frac{(\|\mathbf{A}-\mathbf{P}\|/|z|)^{\log n+1}}{|z|-\|\mathbf{A}-\mathbf{P}\|}\\
&= O\left(\frac{\|\mathbf{v}_i\|_\infty}{\lambda^*(\mathbf{P})/2-\sqrt{n\rho}(\log n)^{\xi}} + \frac{(2C\sqrt{n\rho}/\lambda^*(\mathbf{P}))^{\log n+1}}{\lambda^*(\mathbf{P})/2 - C\sqrt{n\rho}}\right)\\
\text{(Eq (23))}\quad &= \frac{O\left(\|\mathbf{v}_i\|_\infty + n^{-2\xi}\right)}{\lambda^*(\mathbf{P})}
\end{aligned}$$

We also have, for large enough n, $\left(2C\sqrt{n\rho}/\lambda^*(\mathbf{P})\right)^{\log n+1} \leq \left(C/(2(\log n)^{\xi})\right)^{\log n+1} \leq \exp(O(\log n) - \xi(\log n+1)\log\log n) = O\left(1/(n^{2\xi})\right)$. Furthermore, using the same argument as before,

$$\max_{k,z\in\mathcal{C}_k}|\mathbf{e}_x^T\mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)(\mathbf{A}-\mathbf{P})\mathbf{v}_i| = \max_{k,z\in\mathcal{C}_k}\left|\sum_{t=1}^{\infty}\frac{\mathbf{e}_x^T(\mathbf{A}-\mathbf{P})^t}{z^t}\mathbf{v}_i\right| = O\left(\frac{\sqrt{n\rho}\left((\log n)^{\xi}\|\mathbf{v}_i\|_\infty + n^{-2\xi}\right)}{\lambda^*(\mathbf{P})}\right)$$

□

Now we are ready to finish the proof of Theorem 3.1.

Proof of Theorem 3.1. Our goal is to bound the row norm of $\mathbf{V}\mathbf{V}^T - \hat{\mathbf{V}}\hat{\mathbf{V}}^T$ using Lemma 5.2. The first step is to show:

$$\|\mathbf{e}_x^T(\mathbf{V}\mathbf{V}^T - \hat{\mathbf{V}}\hat{\mathbf{V}}^T)\| \overset{\mathcal{E}'}{=\!=} \left\|\sum_{k=1}^{I} \mathbf{e}_x^T(\mathbf{V}_k\mathbf{V}_k^T - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^T)\right\|. \tag{24}$$

Recall that $a_k = \max(\lambda_{e_k} - g_k/2, (1+c)\|\mathbf{A}-\mathbf{P}\|)$. Conditioned on $\mathcal{E}'$, and using Assumption 3.1 and Lemma II.4 in the supplementary material, we have $\lambda_{e_k} - g_k/2 \geq \lambda^*(\mathbf{P})/2 = \omega(\|\mathbf{A}-\mathbf{P}\|)$. This gives $a_k = \lambda_{e_k} - g_k/2$. Hence the intervals are mutually exclusive and cover all the population eigenvalues, proving Eq (24). By triangle inequality, conditioned on $\bigcap_{t=1}^{\log n} \mathcal{E}_t \cap \mathcal{E}'$, from Lemma 5.2, we have:

$$\begin{aligned}
&\|\mathbf{e}_x^T(\mathbf{V}\mathbf{V}^T - \hat{\mathbf{V}}\hat{\mathbf{V}}^T)\| \leq \sum_{k=1}^{I} \|\mathbf{e}_x^T(\mathbf{V}_k\mathbf{V}_k^T - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^T)\| \\
&\overset{(i)}{=} \sum_{k=1}^{I} O\left(\frac{\lambda_{s_k} - \lambda_{e_k} + 2g_k}{g_k}\right) \max_{z\in\mathcal{C}_k}\left(O\left(\frac{\sqrt{n\rho}(b_k+\gamma_k)}{g_k}\right)\|\mathbf{e}_x^T\mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)\mathbf{V}\| + \|\mathbf{e}_x^T\mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)(\mathbf{A}-\mathbf{P})\mathbf{V}\|\right) \\
&\overset{(ii)}{=} O(\psi(\mathbf{P})) \max_{k,z\in\mathcal{C}_k}\left(O\left(\sqrt{n\rho}\right)\|\mathbf{e}_x^T\mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)\mathbf{V}\| + \|\mathbf{e}_x^T\mathbf{G}_{\mathbf{A}-\mathbf{P}}(z)(\mathbf{A}-\mathbf{P})\mathbf{V}\|\right) \\
&\overset{(iii)}{=} O\left(\frac{\psi(\mathbf{P})\sqrt{Kn\rho}}{\lambda^*(\mathbf{P})}\right)\left((1+(\log n)^{\xi})\max_i \|\mathbf{v}_i\|_\infty + 2n^{-2\xi}\right) \\
&\overset{(iv)}{=} O\left(\frac{\psi(\mathbf{P})\sqrt{Kn\rho}}{\rho\lambda^*(\mathbf{B})\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta})}\right)\left(\frac{1+(\log n)^{\xi}}{\sqrt{\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta})}} + 2n^{-2\xi}\right) \\
&\overset{(v)}{=} \tilde{O}\left(\frac{\psi(\mathbf{P})\sqrt{Kn}}{\sqrt{\rho}\lambda^*(\mathbf{B})(\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta}))^{1.5}}\right)
\end{aligned} \tag{25}$$

Step (i) uses Eq (15). Step (ii) uses the fact that $\lambda_{e_k} - \lambda_{s_{k-1}} = g_k$ and $\lambda_{s_k}/g_k = \Omega(1)$. Step (iii) follows from Lemma 5.5. Step (iv) uses $\lambda^*(\mathbf{P}) \geq \rho\lambda^*(\mathbf{B})\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta})$ (Lemma II.4 in the supplementary material) and Lemma 3.2. Step (v) uses $\lambda_1(\mathbf{\Theta}^T\mathbf{\Theta}) \leq n$ (Lemma II.2 in the supplementary material) and $1/\sqrt{\lambda_K(\mathbf{\Theta}^T\mathbf{\Theta})} \geq 1/\sqrt{\lambda_1(\mathbf{\Theta}^T\mathbf{\Theta})} = \Omega(1/\sqrt{n}) = \Omega(n^{-2\xi})$. To bound the failure probability, for some constant $\xi > 1$ and large enough n, Eq (14) gives:

$$\mathrm{P}(\bigcap_{t=1}^{\log n} \mathcal{E}_t \cap \mathcal{E}') \geq 1 - \mathrm{P}(\bar{\mathcal{E}}') - \sum_{t=1}^{\log n} \mathrm{P}(\bar{\mathcal{E}}_t) \geq 1 - O(Kn^{-3}).$$

Now the theorem statement follows by using a union bound.

□

6 Conclusion

In this paper, we propose a fast and provably consistent algorithm called SPACL for inferring community memberships of nodes in a network generated by a Mixed Membership Stochastic Blockmodel (MMSB). Our proof has several new aspects, including a sharp row-wise eigenvector bound using complex contour integration, a new grouping of the eigenvalues to yield better dependence on the smallest singular value of $\mathbf{B}$. Our eigenvector deviation results can be easily generalized to low rank population matrices arising from models other than MMSB. It also helps us establish the convergence of inferred soft community memberships of each node to its population counterpart, which is to our knowledge, the first such result for overlapping network models. In contrast to prior work, we only assume that each community has at least one pure node, and we prove both necessary and sufficient conditions for identifiability under MMSB. We demonstrate the empirical performance of SPACL on simulated and real-world networks of up-to 100,000 nodes. Our experiments show that SPACL has smaller error as well as lower variability than other competing methods. In terms of scalability, we can obtain overlapping cluster memberships of large 100,000 node networks in tens of seconds.

References

[1] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, June 1983. ISSN 0378-8733.

[2] J. Lei, A. Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[3] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

[4] F. McSherry. Spectral partitioning of random graphs. In *FOCS*, pages 529–537, 2001.

[5] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *JMLR*, 9:1981–2014, 2008.

[6] P. K. Gopalan and D. M. Blei. Efficient discovery of overlapping communities in massive networks. *PNAS*, 110(36):14534–14539, 2013.

[7] B. Ball, B. Karrer, and M. E. Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103, 2011.

[8] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding. Community discovery using nonnegative matrix factorization. *Data Mining and Knowl. Disc.*, 22(3):493–521, 2011.

[9] X. Wang, X. Cao, D. Jin, Y. Cao, and D. He. The (un) supervised nmf methods for discovering overlapping communities as well as hubs and outliers in networks. *Physica A: Statistical Mechanics and its Applications*, 446:22–34, 2016.

[10] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon. Overlapping community detection using bayesian non-negative matrix factorization. *Phys. Rev. E*, 83(6):066114, 2011.

[11] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A tensor approach to learning mixed membership community models. *JMLR*, 15(1):2239–2312, 2014.

[12] S. B. Hopkins and D. Steurer. Bayesian estimation from few samples: community detection and related problems. In *FOCS*, pages 379–390. IEEE, 2017.

[13] X. Mao, P. Sarkar, and D. Chakrabarti. On mixed memberships and symmetric

nonnegative matrix factorizations. In *ICML*, pages 2324–2333, 2017.

[14] Y. Zhang, E. Levina, and J. Zhu. Detecting overlapping communities in networks using spectral methods. *arXiv preprint arXiv:1412.3432*, 2014.

[15] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization–provably. In *STOC*, pages 145–162. ACM, 2012.

[16] J. Jin, Z. T. Ke, and S. Luo. Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*, 2017.

[17] A. Ray, J. Ghaderi, S. Sanghavi, and S. Shakkottai. Overlap graph clustering via successive removal. In *52nd Annual Allerton Conference*, pages 278–285. IEEE, 2014.

[18] E. Kaufmann, T. Bonald, and M. Lelarge. A spectral algorithm with additive clustering for the recovery of overlapping communities in networks. In *ALT*, pages 355–370, 2016.

[19] M. Panov, K. Slavnov, and R. Ushakov. Consistent estimation of mixed memberships with successive projections. In *COMPLEX NETWORKS*, pages 53–64. Springer, 2017.

[20] N. Gillis and S. A. Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *PAMI*, 36(4):698–714, 2014.

[21] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[23] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *International Conference on Machine Learning*, pages 97–104, 2006.

[24] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[25] S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh. Noise thresholds for spectral clustering. In *NIPS*, pages 954–962. 2011.

[26] A. Athreya, C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78

(1):1–18, 2016.

[27] E. Abbe, J. Fan, K. Wang, and Y. Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

[28] J. Cape, M. Tang, C. E. Priebe, et al. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47 (5):2405–2439, 2019.

[29] J. Eldridge, M. Belkin, and Y. Wang. Unperturbed: spectral analysis beyond davis-kahan. In *Algorithmic Learning Theory*, volume 83, pages 321–358. PMLR, 2018.

[30] J. Cape, M. Tang, and C. E. Priebe. Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *arXiv preprint arXiv:1802.00381*, 2018.

[31] Y. Chen, S. Sanghavi, and H. Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, 2014.

[32] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

[33] L. Erdős, A. Knowles, H.-T. Yau, J. Yin, et al. Spectral statistics of erdős–rényi graphs i: local semicircle law. *The Annals of Probability*, 41(3B):2279–2375, 2013.

[34] R. I. Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.