Robust recovery for stochastic block models, simplified and generalized

Sidhanth Mohanty*   Prasad Raghavendra†   David X. Wu‡

February 21, 2024

Abstract

We study the problem of *robust community recovery*: efficiently recovering communities in sparse stochastic block models in the presence of adversarial corruptions. In the absence of adversarial corruptions, there are efficient algorithms when the *signal-to-noise ratio* exceeds the *Kesten–Stigum (KS) threshold*, widely believed to be the computational threshold for this problem. The question we study is: *does the computational threshold for robust community recovery also lie at the KS threshold?* We answer this question affirmatively, providing an algorithm for robust community recovery for arbitrary stochastic block models on any constant number of communities, generalizing the work of Ding, d'Orsi, Nasser & Steurer [DdNS22] on an efficient algorithm above the KS threshold in the case of 2-community block models.

There are three main ingredients to our work:

1. The Bethe Hessian of the graph is defined as $H_G(t) \triangleq (D_G - I)t^2 - A_G t + I$ where D_G is the diagonal matrix of degrees and A_G is the adjacency matrix. Empirical work suggested that the Bethe Hessian for the stochastic block model has outlier eigenvectors corresponding to the communities right above the Kesten-Stigum threshold [KMM+13, SKZ14].

   We formally confirm the existence of outlier eigenvalues for the Bethe Hessian, by explicitly constructing outlier eigenvectors from the community vectors.

2. We develop an algorithm for a variant of robust PCA on sparse matrices. Specifically, an algorithm to partially recover top eigenspaces from adversarially corrupted sparse matrices under mild delocalization constraints.

3. A rounding algorithm to turn vector assignments of vertices into a community assignment, inspired by the algorithm of Charikar & Wirth [CW04] for 2XOR.

---

*MIT. sidhanth@csail.mit.edu. Much of this work was done while the author was a PhD student at UC Berkeley.

†UC Berkeley. raghavendra@berkeley.edu.

‡UC Berkeley. david_wu@berkeley.edu.

Contents

1 Introduction

The stochastic block model (SBM) has provided an enlightening lens into understanding a wide range of computational phenomena in Bayesian inference problems, such as computational phase transitions & information-computation gaps [DKMZ11, MNS18, Mas14, HS17], spectral methods for sparse matrices [KMM+13, SKZ14, BLM15], local message-passing algorithms [MNS14, YP23, GP23], and robustness [MPW16, BMR21, DdNS22, LM22].

The SBM is a model of random graphs where the vertices are partitioned into *communities*, denoted by x, and the probability of an edge existing is contingent on the communities that the two endpoints are part of. The algorithmic task is the *community recovery* problem: given an input graph G, estimate the posterior $x|G$ with an efficient algorithm.

Definition 1.1 (Informal). In the *stochastic block model*, we are given a $k \times k$ matrix M, a distribution π over $[k]$, and $d > 0$, and $\text{SBM}_n(\text{M}, \pi, d)$ denotes the distribution where an n-vertex graph G is sampled by:

1. drawing a *color* $x(u) \sim \pi$ for every $u \in [n]$,

2. for each pair of vertices u, v, the edge $\{u, v\}$ is chosen with probability $\text{M}_{x(u), x(v)} \cdot \frac{d}{n}$.

In the *community recovery* problem, the goal is to give an efficient algorithm that takes G as input and outputs a community assignment $\widehat{x}$ approximating $x|G$ (see Definition 4.4 for a formal definition).

Computational thresholds. For a given M and π, increasing d can only possibly make the problem easier. The main question is to understand the *computational threshold* for community recovery — i.e. the minimum value of d where the problem goes from being intractable to admitting efficient algorithms.

The first predictions for this computational threshold came from the *cavity method* in statistical physics in the work of Decelle, Krzakala, Moore & Zdeborova [DKMZ11]. They posited that the location of this transition is at the *Kesten–Stigum threshold* (henceforth KS threshold), a threshold for broadcast processes on trees studied in the works of Kesten & Stigum [KS66, KS67]. The algorithmic side of these predictions was confirmed in the case of the 2-community block model in the works of Mossel, Neeman & Sly [MNS18] and Massoulié [Mas14], and then for block models in increasing levels of generality by Bordenave, Lelarge & Massoulié [BLM15], Abbe & Sandon [AS15], and Hopkins & Steurer [HS17].

Robust algorithms. All of these algorithms utilize the knowledge of the distribution the input is sampled from quite strongly — they are based on $\Omega(\log n)$-length walk statistics in the stochastic block model. However, the full generative process in inference is not always known precisely. Thus, we would like algorithms that utilize but do not overfit to the distributional assumptions.

Demanding that our algorithm be *robust*, i.e. resilient to adversarial corruptions to the input, is often a useful way to design algorithms that are less sensitive to distributional assumptions. This leads one to wonder: can algorithms that don't strongly exploit the prior distribution approach the KS threshold?

Optimization vs. inference. Earlier approaches to robust recovery in 2-community block models were based on *optimization*: semidefinite programming relaxations of the minimum bisection problem, as in the work of Guedon & Vershynin [GV16]. These approaches have the advantage of

being naturally robust, since the algorithms are approximately Lipschitz around random inputs, but the minimum bisection relaxation is not known to achieve statistical optimality and only succeeds well above the KS threshold.

The following two results point to the suboptimality of optimization-based strategies. Moitra, Perry & Wein [MPW16] considered the *monotone* adversary in the 2-community setting, where the adversary is allowed to make an *unbounded* number of edge insertions within communities and edge deletions across communities. At an intuitive level, this is supposed to only make the problem easier and indeed does so for the minimum bisection approach, but to the contrary [MPW16] proves that the threshold for recovery *increases*. Dembo, Montanari & Sen [DMS17] exactly nailed the size of the minimum bisection in Erdős–Rényi graphs, which are complete noise and have no signal in the form of a planted bisection — and strikingly, it is actually *smaller* than the size of the planted bisection in the detectable regime! Thus, it is conceivable that there are bisections completely orthogonal to the planted bisection in a stochastic block model graph that nevertheless have the same size.

The problem of recovering communities is more related to the task of *Bayesian inference*, i.e., applying Bayes' rule and approximating $x|G$. Optimizing for the minimum bisection is akin to computing the *maximum likelihood estimate*, which does not necessarily produce samples representative of the posterior distribution of $x|G$.

SDPs for inference. The work of Banks, Mohanty & Raghavendra [BMR21] proposed a semidefinite programming-based algorithm for inference tasks that incorporates the prior distribution in the formulation, and illustrated that this algorithm can distinguish between G sampled from the stochastic block model from an Erdős–Rényi graph of equal average degree anywhere above the KS threshold while being resilient to $\Omega(n)$ arbitrary edge insertions and deletions.

A similar SDP formulation was later studied by Ding, d'Orsi, Nasser & Steurer [DdNS22] in the 2-community setting, and was used to give an algorithm to recover the communities with a constant advantage over random guessing in the presence of $\Omega(n)$ edge corruptions for all degrees above the KS threshold. They analyze the spectra of matrices associated with random graphs after deleting vertices with large neighborhoods, which introduces unfriendly correlations, and causes their analysis to be highly technical.

The main contribution of our work is an algorithm for robust recovery, which is amenable to a significantly simpler analysis. Our algorithm also succeeds at the recovery task for arbitrary block models with a constant number of communities.

Theorem 1.2 (Informal statement of main theorem). *Let (M, π, d) be SBM parameters such that d is above the KS threshold, and let $G, x \sim \mathrm{SBM}_n(M, \pi, d)$. There exists $\delta = \delta(M, \pi, d) > 0$ such that the following holds. There is a polynomial time algorithm that takes as input any graph $\widetilde{G}$ that can be obtained by performing arbitrary δn edge insertions and deletions to G and outputs a coloring $\widehat{x}$ that has "constant correlation" with x, with high probability over the randomness of G and x.*

Many of the ingredients in the above result are of independent interest. First, we exhibit a symmetric matrix closely related to the Bethe Hessian of the graph, such that its bottom eigenspace is correlated with the communities. Next, we design an efficient algorithm to robustly recover the bottom-r eigenspace of a *sparse* matrix in the presence of adversarial corruptions. Finally, we demonstrate a general rounding scheme to obtain community assignments from this eigenspace.

Remark 1.3 (Robustness against node corruptions). The node corruption model, introduced by Liu & Moitra [LM22], is a harsher generalization of the edge corruption model. In recent work, Ding, d'Orsi, Hua & Steurer [DdHS23] proved that in the setting of sparse SBM, any algorithm that is robust to edge corruptions can be turned into one robust to node corruptions in a blackbox manner. Hence, our results apply in this harsher setting too.

1.1 Related work

We refer the reader to the survey of Abbe [Abb17] for a detailed treatment of the rich history and literature on community detection in block models, its study in other disciplines, and the many information-theoretic and computational results in various parameter regimes.

Introducing an adversary into the picture provides a beacon towards algorithms that utilize but do not *overfit* to distributional assumptions. Over the years, a variety of adversarial models have been considered, some of which we survey below.

Corruption models for stochastic block model. Prior to the works of [BMR21, DdNS22], Stefan & Massoulié [SM19] considered the robust recovery problem, and gave a robust spectral algorithm to recover communities under $O(n^{\varepsilon})$ adversarial edge corruptions for some small enough $\varepsilon > 0$.

Liu & Moitra [LM22] introduced the *node corruption* model where an adversary gets to perform arbitrary edge corruptions incident to a constant fraction of corrupted vertices, and gave algorithms that achieved optimal accuracy in the presence of node corruptions and the monotone adversary sufficiently above the KS threshold. Soon after, Ding, d'Orsi, Hua & Steurer [DdHS23] gave algorithms achieving the Kesten–Stigum threshold using algorithms for the edge corruption model in the low-degree setting [DdNS22], and results on the optimization SDP in the high-degree setting [MS16] in a blackbox manner.

Semirandom & smoothed models. Some works have considered algorithm design under harsher adversarial models, where an adversarially chosen input undergoes some random perturbations.

Remarkably, at this point, the best algorithms for several graph and hypergraph problems match the performance of our best algorithms for their completely random counterparts. For example, at this point, the semirandom planted coloring and clique problems were introduced by Blum & Spencer [BS95], and Feige & Kilian [FK01], and a line of work [CSV17, MMT20] culminating in the work of Buhai, Kothari & Steurer [BKS23] showed that the size of the planted clique/coloring recoverable in the semirandom setting matches the famed $\sqrt{n}$ in the fully random setting.

Another example where algorithms for a semirandom version of a block model-like problem have been considered is semirandom CSPs with planted solutions, where the work of Guruswami, Hsieh, Kothari & Manohar [GHKM23] gives algorithms matching the guarantees of solving fully random planted CSPs.

1.2 Organization

In Section 2, we give an overview of our algorithm and proof. In Section 3, we give some technical preliminaries. In Section 4, we describe our algorithm and show how to analyze it. In Section 5, we prove results about the spectrum of the Bethe Hessian matrix and variants, key to our algorithm. In Section 6, we show how to use a trimming procedure and a spectral algorithm on a Bethe Hessian variant to recover a subspace which has good correlation with the true communities in the

presence of adversarial corruptions. Finally, in Section 7, we prove some technical claims relevant to analyzing our algorithm.

2   Technical overview

An n-vertex graph G is drawn from a stochastic block model and undergoes δn adversarial edge corruptions, and then the corrupted graph $\widetilde{G}$ is given to us as input. For simplicity of discussion, we restrict our attention to *assortative* symmetric k-community block models above the KS threshold, i.e. the connection probability between two vertices i and j only depends on whether they belong to the same community or different communities, and the intra-community probability is higher. Nevertheless, our approach generalizes to any arbitrarily specified k-community block model above the KS threshold.

Let us first informally outline the algorithm; see Section 4 for formal details.

1. First, we preprocess the corrupted graph $\widetilde{G}$ by truncating high degree vertices, which removes corruptions localized on small sets of vertices in the graph.

2. We then construct an appropriately defined graph-aware symmetric matrix $M_G \in \mathbb{R}^{n \times n}$ whose negative eigenvalues contains information about the true communities for the *uncorrupted* graph. We motivate this construction in Section 2.1.

3. We recursively trim the rows and columns of $M_{\widetilde{G}}$ to remove small negative eigenvalues in its spectrum. Then we use a spectral algorithm to robustly recover a subspace U which contains information about the communities. Both points are described in Section 2.2.

4. Finally, we round the subspace U into a community assignment, using a vertex embedding provided by U. This is detailed in Section 2.3.

2.1   Outlier eigenvectors for the Bethe Hessian

Bordenave, Lelarge & Massoulié [BLM15] analyzed the spectrum of the *nonbacktracking matrix* and rigorously established its connection to community detection. The asymmetric nonbacktracking matrix $B_G \in \{0, 1\}^{2|E(G)| \times 2|E(G)|}$ is indexed by directed edges, with

$$(B_G)_{(u_1 \to v_1),(u_2 \to v_2)} \triangleq \mathbf{1}[v_1 = u_2]\mathbf{1}[v_2 \neq u_1].$$

[BLM15] showed that above the KS threshold, the k outlier eigenvalues for B_G correspond to the k community vectors. More precisely, in the case of symmetric k-community stochastic block models above the KS threshold, [BLM15] proved that for the randomly drawn graph G, there is a small $\varepsilon > 0$ for which its nonbacktracking matrix B_G has exactly k eigenvalues larger than $(1 + \varepsilon)\sqrt{d}$ in magnitude (Theorem 5.3).

The *Bethe Hessian* matrix is a symmetric matrix associated with a graph, whose early appearances can be traced to the works of Ihara [Iha66] and Bass [Bas92]. The Bethe Hessian of a graph with parameter $t \in \mathbb{R}$ is defined as

$$H_G(t) \triangleq (D_G - I)t^2 - A_G t + I,$$

where D_G and A_G are the diagonal degree matrix and adjacency matrix of G, respectively. For t in the interval $[0, 1]$, it can be interpreted as a regularized version of the standard graph Laplacian. The Bethe Hessian for the stochastic block model was considered in the empirical works [KMM⁺13, SKZ14], where they observed that for some choice of t, the Bethe Hessian and the nonbacktracking matrix has outlier eigenvectors which can be used for finding communities in block models. Concretely, in [SKZ14] they observed that for G drawn from stochastic block models above the KS threshold, there is a choice of t such that $H_G(t)$ only has a small number of negative eigenvectors, all of which correlate with the hidden community assignment.

We confirm this empirical observation in the following proposition.

Proposition 2.1 (Bethe Hessian spectrum). *Let* (M, π, d) *be k-community SBM parameters such that d is above the KS threshold, and let* $G, x \sim \mathrm{SBM}_n(M, \pi, d)$. *Then there exists* $\varepsilon > 0$ *such that for* $t^* = \frac{1}{(1+\varepsilon)\sqrt{d}}$, *the Bethe Hessian* $H_G(t^*)$ *has at most k negative eigenvalues and at least k − 1 negative eigenvalues.*

Constructing the outlier eigenspace There are two assertions in Proposition 2.1. To show that $H_G(t^*)$ has at most k negative eigenvalues, one can relate these negative eigenvalues to the k outlier eigenvalues of B_G using an Ihara–Bass argument and use a continuity argument as outlined in Fan and Montanari [FM17, Theorem 5.1].

The more interesting direction is to exhibit at least $k − 1$ negative eigenvalues; we will explicitly construct a $k-1$ dimensional subspace starting with the community vectors to witness the negative eigenvalues for $H_G(t^*)$.

Let $\mathbf{1}_c$ denote the indicator vector for the vertices belonging to community c and $\mathbf{1}$ the all-ones vector. We show that every vector in the span of $\{A^{(\ell)}(\mathbf{1}_c - \frac{1}{k}\mathbf{1})\}_{c\in[k]}$ achieves a negative quadratic form against $H_G(t^*)$, where $A^{(\ell)}$ is the $n \times n$ matrix where the (i, j)-th entry encodes the number of length-ℓ nonbacktracking walks between i and j. This demonstrates a $(k-1)$-dimensional subspace on which the quadratic form is negative. Formally, we show the following:.

Proposition 2.2. *Under the same setting and notations as* Proposition 2.1, *for* $\ell \geqslant 0$ *define*

$$M_{G,\ell} \triangleq A^{(\ell)} H_G(t^*) A^{(\ell)}.$$

For $\ell = \Theta\left(\frac{\log(1/\varepsilon)}{\varepsilon}\right)$ *and every* $c \in [k]$, *we have*

$$\left\langle \mathbf{1}_c - \tfrac{1}{k}\mathbf{1}, M_{G,\ell}(\mathbf{1}_c - \tfrac{1}{k}\mathbf{1}) \right\rangle \leqslant -\Omega(n).$$

Hence, $M_{G,\ell}$ *has at most k negative eigenvalues and at least k − 1 negative eigenvalues.*

Nonbacktracking powers and related constructions were previously studied in [Mas14, MNS18], but there they take $\ell = \Theta(\log n)$, whereas we only consider constant ℓ. Besides simplifying the analysis of the quadratic form, using constant ℓ is also critical for tolerating up to $\Omega(n)$ corruptions.

As a consequence of Proposition 2.2, the negative eigenvectors of $M_{G,\ell}$ are correlated with the centered community indicators $\{\mathbf{1}_c - \frac{1}{k}\mathbf{1}\}_{c\in[k]}$, while the negative eigenvectors of $H_G(t^*)$ are correlated with $\{A^{(\ell)}(\mathbf{1}_c - \frac{1}{k}\mathbf{1})\}_{c\in[k]}$. The upshot is that we can directly use the negative eigenvectors of $M_{G,\ell}$ to recover the true communities in the absence of corruptions.

Remark 2.3. Based on the empirical observations in [KMM$^+$13, SKZ14], a natural hope is to directly use the Bethe Hessian for recovery. However, it turns out that the quadratic form of the centered true community indicators $\langle (\mathbf{1}_c - \frac{1}{k}\mathbf{1}), H_G(t^*)(\mathbf{1}_c - \frac{1}{k}\mathbf{1}) \rangle$ are actually *positive* close to the KS threshold, so the same approach does not establish that the negative eigenvectors of $H_G(t^*)$ correlate with the communities.

We will now discuss how to recover the outlier eigenspace in the presence of adversarial corruptions.

2.2 Robust PCA for sparse matrices

The discussion above naturally leads to the following algorithmic problem of robust recovery: Given as input a corrupted version $\widetilde{M}$ of a symmetric matrix M, can we recover the bottom/top r-dimensional eigenspace of M? Since the true communities are constantly correlated with the outlier eigenspace of $M = M_{G,\ell}$, recovering the outlier eigenspace of M from its corrupted version $\widetilde{M} = \widetilde{M}_{G,\ell}$ is a major step towards robustly recovering communities.

The problem of robustly recovering the top eigenspace, a.k.a. robust PCA has been extensively studied, and algorithms with provable guarantees have been designed (see [CLMW11]). However, the robust PCA problem in our work is distinct from those considered in the literature in a couple of ways. For us, the uncorrupted matrix M is sparse and both the magnitude and location of the noisy entries are adversarial. Furthermore, for our purposes, we need not recover the actual outlier eigenspace of M. Indeed, as we discuss below, it suffices to robustly recover a constant dimensional subspace which is constantly correlated with the true communities.

We design an efficient algorithm to robustly recover such a subspace under a natural set of sufficient conditions on M. Before we describe these conditions, let us fix some notation. We will call a vector $x \in \mathbb{R}^n$ to be C-delocalized if no coordinate is large relative to others, i.e., $|x_i|^2 \leqslant \frac{C}{n}\|x\|^2$ for all $i \in [n]$. Delocalization has previously been used in the robust PCA literature under the name "incoherence" [CLMW11].

Let M be a $n \times n$ matrix with at most r negative eigenvalues. In particular, the r-dimensional negative eigenspace V_M of M is the object of interest. Let $\widetilde{M}$ be a corrupted version of M, differing from M in δn coordinates.

Given the corrupted version $\widetilde{M}$, a natural goal would be to recover the r-dimensional negative eigenspace V_M. It is easy to see that it could be impossible to recover the space V_M. Instead, we will settle for a relaxed goal, namely, recover a slightly larger dimensional subspace U that non-trivially correlates with delocalized vectors in the true eigenspace V_M. More formally, we will solve the following problem.

Problem 2.4. Given the corrupted matrix $\widetilde{M}$ as input, give an efficient algorithm to output a subspace U with the following properties:

1. **Low dimensional.** The dimension of U is $O(r)$.

2. **Delocalized.** The diagonal entries of its projection matrix Π_U are bounded by $O\left(\frac{r}{n}\right)$.

3. **Preserves delocalized part of negative eigenspace.** For any C-delocalized unit vector y such that $\langle y, My \rangle < -\Omega(1)$, we have $\langle y, \Pi_U y \rangle \geqslant \Omega(1)$.

6

In fact, our algorithm will recover a principal submatrix of $\widetilde{M}$ whose eigenspace V for eigenvalues less than $-\eta$ is $O(r)$-dimensional. Moreover, the eigenspace V can be processed to another delocalized, $O(r)$-dimensional subspace U that satisfies the conditions outlined above.

Although the matrix M has a constant number of negative eigenvalues, its corruption $\widetilde{M}$ can have up to $\Omega(n)$ many. At first glance, it may be unclear how a constant dimensional subspace U can be extracted from $\widetilde{M}$. The crucial observation is that the large negative eigenvalues introduced by the corruptions are highly localized. Thus, we will design an iterative trimming algorithm that aims to delete rows and columns to clean up these localized corruptions. When the algorithm terminates, it yields the $O(r)$-dimensional subspace V.

Recovering a principal submatrix. We now describe the trimming algorithm informally and refer the reader to Section 6 for the formal details.

We first fix some small parameter $\eta > 0$ and execute the following procedure, which produces a series of principal submatrices $\widetilde{M}^{(t)}$ for $t \geqslant 0$, starting with $\widetilde{M}^{(0)} \triangleq \widetilde{M}$.

1. At step t, if the eigenspace V of eigenvalues of $\widetilde{M}^{(t)}$ less than $-\eta$ is $O(r)$-dimensional, we terminate the algorithm and output V.

2. Otherwise, compute the projection $\Pi^{(t)}$ corresponding to the $\leqslant -\eta$ eigenspace of $\widetilde{M}^{(t)}$.

3. Sample an index $i \in [n]$ of $\widetilde{M}^{(t)}$ with probability proportional to $\Pi_{i,i}^{(t)}$.

4. Zero out row and column i, and set this new principal submatrix as $\widetilde{M}^{(t+1)}$.

We now discuss the intuition behind the procedure and formally prove its correctness in Section 6.1. The main idea of step 3 is that one should prefer to delete highly localized eigenvectors which have relatively large negative eigenvalues. This is reasonable because the size of the diagonal entries of $\widetilde{M}^{(t)}$ serve as a rough proxy for the level of delocalization.

As a concrete illustration of this intuition, suppose that $\widetilde{M} = \Pi^{(0)} = -uu^\top - vv^\top$, where u, v are orthogonal unit vectors. Moreover, suppose u is C-delocalized whereas $v = e_1$. Then $\Pi_{1,1}^{(0)} = 1$ whereas $|\Pi_{i,i}^{(0)}| \leqslant C^2/n$ for $i > 1$. Hence, deleting the first row and column of $\widetilde{M}$ also deletes the localized eigenvector v. In general, whenever one of the eigenvectors of $\widetilde{M}^{(t)}$ is heavily localized on a subset of coordinates S, the diagonal entries in $\Pi_{S,S}^{(t)}$ are disproportionately large. This leads to a win-win scenario: either we reach the termination condition, or we are likely to mitigate the troublesome large localized eigenvectors.

We now discuss how we achieve the second and third guarantees in Problem 2.4.

Trimming the subspace. The final postprocessing step is simple. Let V denote the eigenspace with eigenvalues less than $-\eta$ for the matrix $\widetilde{M}^{(T)}$ obtained at end of iterative procedure.

To ensure delocalization (condition 2 in Problem 2.4), the idea is to take its projector Π_V and trim away the rows and columns with diagonal entry exceeding $\frac{\tau}{n}$ for some large parameter $\tau > 0$. The desired delocalized subspace U is obtained by taking the eigenspace of the trimmed Π_V corresponding to the eigenvalues exceeding a threshold that is $O(\eta)$. Since V is $O(r)$-dimensional, so too is U.

The more delicate part is condition 3 in Problem 2.4. Namely, we must show that despite corruptions and the repeated trimming steps, x remains a delocalized witness vector for Π_U, and thus has constant correlation with the subspace U. The key intuition for this is that delocalized

7

witnesses are naturally robust to adversarial corruptions, so long as the adversarial corruptions have bounded row-wise ℓ_1 norm. In particular, since delocalization is an ℓ_∞ constraint, Hölder's inequality bounds the difference in value of the quadratic form using M and $\widetilde{M}$. In Section 6.2, we prove that for sufficiently small constant levels of corruption, x is also a delocalized witness for $\widetilde{M}$ and Π_U.

Finally, we discuss how to round the recovered subspace U into a community assignment.

2.3 Rounding to communities

At this stage, we are presented with a constant-dimensional subspace U with the key feature that it is correlated with the community assignment vectors $\{\mathbf{1}_c\}_{c\in[k]}$. Our goal is to round U to a community assignment that is "well-correlated" with the ground truth. In order to discuss how we achieve this goal, we must make precise what it means to be "well-correlated" with the ground truth. Notice that a community assignment is just as plausible as the same assignment with the names of communities permuted, and thus counting the number of correctly labeled vertices is not a meaningful metric.

A more meaningful metric is the number of pairwise mistakes, i.e. the number of pairs of vertices in the same community assigned to different communities or in different communities assigned to the same community. A convenient way to express this metric is via the inner product of positive semidefinite matrices encoding whether pairs of vertices belong to the same community or not. Given a community assignment x, we assign it the matrix X, defined as

$$X[i,j] = \begin{cases} 1 & \text{if } x(i) = x(j) \\ -\frac{1}{k-1} & \text{if } x(i) \neq x(j). \end{cases}$$

For the ground truth assignment x and the output of our algorithm $\widehat{x}$, we measure the correlation with $\langle X, \widehat{X} \rangle$. Observe that for any guess $\widehat{X}$ that is oblivious to the input (for example, classifying all vertices to the same community, or blindly guessing), the value of $\langle X, \widehat{X} \rangle$ is concentrated below $\tilde{O}(n^{3/2})$. On the other hand, if $\widehat{X} = X$, then this correlation is $\Omega(n^2)$. See Definition 4.4 and Appendix A for how this notion generalizes to arbitrary block models, and subsumes other notions of weak-recovery defined in literature.

The projection matrix Π_U satisfies $\langle \Pi_U, X \rangle \geq \Omega(\|\Pi_U\|_F \cdot \|X\|_F) = \Omega(n)$. We give a randomized rounding strategy according to which $\mathbf{E}\,\widehat{X} \succeq c \cdot n \cdot \Pi_U$ for some constant $c > 0$. Consequently, $\mathbf{E}\langle X, \widehat{X} \rangle = cn \cdot \langle \Pi_U, X \rangle \geq \Omega(n^2)$.

Observe that for any community assignment x, its matrix representation X is rank-$(k-1)$, which lets us write it as $VV^\top$ for some $n \times (k-1)$ matrix V. Here, the i-th row of V is some vector $v_{x(i)}$ that only depends only on the community $x(i)$ where vertex i is assigned.

Our rounding scheme uses Π_U to produce an embedding of the n vertices as rows of a $n \times (k-1)$ matrix W whose rows are in $\{v_1, \ldots, v_k\}$. In the community assignment $\widehat{x}$ outputted by the algorithm, the i-th vertex is assigned to community j if the i-th row of W is equal to v_j. We then show that $\mathbf{E}\,WW^\top \succeq c \cdot n \cdot \Pi_U$. Since $\widehat{X} = \mathbf{E}\,WW^\top$, we can conclude $\mathbf{E}\langle X, \widehat{X} \rangle \geq \Omega(n^2)$.

Rounding scheme. Our first step is to obtain an embedding of the n vertices into $\mathbb{R}^{k-1}$ by choosing a $(k-1)$-dimensional random subspace U' of U, then writing its projector as $M'M'^\top$, and choosing the embedding as the rows of M': $u'_1, \ldots, u'_n$. Suppose this embedding has the property that for

8

some $c' > 0$, the rows of $\sqrt{c'n}U'$ lie inside the convex hull of $v_1, \ldots, v_k$, then we can express each u'_i as a convex combination $\sum_{j=1}^k p_j^{(i)} v_j$ and then independently sample w_i from $\{v_1, \ldots, v_k\}$ according to the probability distribution $(p_j^{(i)})_{j \in [k]}$. The resulting embedding W would satisfy the property that $\mathbf{E}\, WW^\top \succeq c' \cdot \frac{k-1}{\dim(U)} \cdot n \cdot \Pi_U$, where this inequality holds since the off-diagonal entries are equal, and the diagonal of $WW^\top$ is larger.

The reason an appropriate scaling c' exists follows from the facts that the convex hull of $v_1, \ldots, v_k$ is full-dimensional and contains the origin, which we prove in Section 7.

3 Preliminaries

Stochastic block model notation. We write $\mathbf{1}$ to denote the all-ones vector and e_i to denote the ith standard basis vector, with the dimensions implicit. For a k-community block model, let $\pi \in \mathbb{R}^k$ denote the prior community probabilities, and $\Pi = \mathrm{diag}(\pi)$, so that $\pi = \Pi\mathbf{1}$. The edge probabilities are parameterized by a symmetric matrix $\mathrm{M} \in \mathbb{R}^{k \times k}$, the block probability matrix. A true community assignment $x : [n] \to [k]$ is sampled i.i.d. from π. Conditioned on x, an edge between i and j is sampled with probability $\frac{\mathrm{M}_{x(i),x(j)}d}{n}$. To ensure that the average degree is d, we stipulate that $\mathrm{M}\pi = \mathbf{1}$.

We will also use $X \in \mathbb{R}^{n \times k}$ to denote the *one-hot encoding* of x, i.e., the matrix where the t-th row is equal to $e_{x(t)}$. We will sometimes find it convenient to access the columns of X, which are the indicator vectors for the k different communities; we denote these by $\mathbf{1}_c$ for any community $c \in [k]$. For any $f : [k] \to \mathbb{R}$, define the lift of f with respect to the true community assignment by $f^{(n)} \triangleq \sum_{c \in [k]} f(c) \cdot \mathbf{1}_c$.

Another natural matrix that appears throughout the analysis is the Markov transition matrix $T \triangleq \mathrm{M}\Pi$, which by detailed balance evidently has stationary distribution π. This is an asymmetric matrix, but since T defines a time-reversible Markov chain with respect to π, T is self-adjoint with respect to the inner product $\langle \cdot, \cdot \rangle_\pi$ in $\mathbb{R}^k$ induced by π. Hence T is diagonalizable with real eigenvalues and its eigenvalues are $1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_k|$, with ties broken by placing positive eigenvalues before the negative ones. Note that the normalization condition $\mathrm{M}\pi = \mathbf{1}$ translates into $T\mathbf{1} = \mathbf{1}$.

Matrix notation. We use $\preceq$ and $\succeq$ to denote inequalities on matrices in the Loewner order. For any $n \times n$ matrix X, we use $\Pi_{\leq a}(X)$ and $\Pi_{\geq a}(X)$ to denote the projectors onto the spaces spanned by eigenvectors of X with eigenvalue at most and at least a respectively. We also define $X_{\leq a} \triangleq \Pi_{\leq a}(X)X\Pi_{\leq a}(X)$ and $X_{\geq a} \triangleq \Pi_{\geq a}(X)X\Pi_{\geq a}(X)$, the corresponding truncations of the eigendecomposition of X.

For $S \subseteq [n]$, we use $X_{S,S}$ to denote the matrix obtained by taking X and zeroing out all rows and columns with indices outside S.

Nonbacktracking matrix and Bethe Hessian. For a graph G, let B_G be its nonbacktracking matrix, A_G be its adjacency matrix, D_G be its diagonal matrix of degrees, $A_G^{(\ell)}$ be its ℓ-th nonbacktracking power of A_G, and $H_G(t) \triangleq (D_G - I)t^2 - A_G t + I$ be its Bethe Hessian matrix. The matrix we use for our algorithm is $M_{G,\ell}(t) \triangleq A_G^{(\ell)} H_G(t) A_G^{(\ell)}$. We will drop the G from the subscript when the graph G is clear from context.

Determinants. Below, we collect some standard linear algebraic facts that will prove useful.

Fact 3.1. *Suppose a matrix X has a kernel of dimension k, then every $(n-j) \times (n-j)$ submatrix of X for $j < k$ is singular.*

Fact 3.2 (Jacobi's formula). *For any differentiable function $X : \mathbb{R} \to \mathbb{R}^{n \times n}$,*

$$\frac{d}{du} \det(X(u)) = \sum_{i=1}^{n} \sum_{j=1}^{n} \det\left(X(u)_{[n]\setminus\{i\},[n]\setminus\{j\}}\right) \cdot (-1)^{i+j} \cdot \frac{d}{du}(X(u))_{i,j}.$$

Lemma 3.3. *Let $X : \mathbb{R} \to \mathbb{R}^{n \times n}$ and $f : \mathbb{R} \to \mathbb{R}$ be any pair of smooth functions. For any $j \geqslant 0$, there exist functions $(q_{S,T} : \mathbb{R} \to \mathbb{R})_{S,T \subseteq [n], |S|=|T| \geqslant n-j}$ such that:*

$$\left(\frac{d}{du}\right)^j [\det(X(u)) \cdot f(u)] = \sum_{S,T \subseteq [n], |S|=|T| \geqslant n-j} \det(X(u)_{S,T}) q_{S,T}(u).$$

Proof. We prove this by induction. This is clearly true when $j = 0$, and the induction step is a consequence of Jacobi's formula. $\qquad\square$

Kesten-Stigum threshold. We say that a stochastic block model is above the Kesten–Stigum (KS) threshold if $\lambda_2(T)^2 d > 1$, where recall that λ_2 is the second largest eigenvalue in absolute value. We use r to denote the number of eigenvalues of T equal to $\lambda_2(T)$.

4   Recovery algorithm

Let G be the graph drawn from $\mathrm{SBM}_n(M, \pi, d)$, and let $\widetilde{G}$ denote the input graph which is G along with an arbitrary δn adversarial edge corruptions. Our algorithm for clustering the vertices into communities proceeds in multiple phases, described formally below.

The first phase preprocesses the graph by making it bounded degree and constructs an appropriate matrix M associated to the graph. The second phase cleans up M and uses a spectral algorithm to robustly recover a subspace containing nontrivial information about the true communities. Finally, the third phase rounds the subspace to an actual community assignment.

Algorithm 4.1. $\widetilde{G}$ is given as input, and a community assignment to the vertices is produced as output.

Phase 1: Deletion of high-degree vertices. For some large constant $B > 0$ to be specified later, we perform the following truncation step: delete all edges incident on vertices with degree larger than B in $\widetilde{G}$. This forms a graph $\widetilde{G}_B$, with corresponding adjacency matrix $A_{\widetilde{G}_B} \in \mathbb{R}^{|V(G)| \times |V(G)|}$. To avoid confusion, we preserve the vertex set $V(G)$, but it should be understood that the truncated vertices do not contribute to the graph.

For technical considerations, we also define a (nonstandard) truncated diagonal matrix

$$\overline{D}_{\widetilde{G}_B} \triangleq \mathrm{diag}\left(\deg(v)\mathbf{1}[\deg(v) \leqslant B]\right)_{v \in V(G)} \tag{1}$$

With this, we can then define the truncated Bethe Hessian matrix

$$\overline{H}_{\widetilde{G}_B}(t) \triangleq I - t A_{\widetilde{G}_B} + t^2(\overline{D}_{\widetilde{G}_B} - I). \tag{2}$$

Finally, the input matrix to the next phase is

$$\overline{M}_{\widetilde{G}_B,\ell}(t) \triangleq A^{(\ell)}_{\widetilde{G}_B} \overline{H}_{\widetilde{G}_B}(t) A^{(\ell)}_{\widetilde{G}_B}, \tag{3}$$

where we also choose the value of t later.

Remark 4.2. To reduce any chance of confusion with the notation, we reiterate our conventions for distinguishing between different versions of various matrices. If a graph is truncated at level B, then we add a subscript B. We use tilde to denote that we are working with a corrupted graph. Finally, we use overline to denote that we are working with the nonstandard version of the Bethe Hessian after truncation.

For example, the matrix $\overline{D}_{G_B}$ no longer corresponds to the degree matrix of G_B, since as stated it still counts edges from truncated vertices. This is done to simplify the analysis of the spectrum of $\overline{M}_{G_B,\ell}(t)$ but we do not believe it to be essential.

Phase 2: Recovering a subspace with planted signal. Define $M \triangleq \overline{M}_{\widetilde{G}_B,\ell}$. We give an iterative procedure to "clean up" M by deleting a few rows and columns. We then run a spectral algorithm on the cleaned up version of M.

Let $\eta > 0$ be a small constant we choose later, and let $K \triangleq B^{2\ell+3}$.

1. Define $M^{(0)}$ as M. Let t be a counter initialized at 0, and $\Phi(X)$ as the number of eigenvalues of X smaller than $-\eta$.

2. While $\Phi(M^{(t)}) > \frac{2K}{\eta}r$: compute the projection matrix $\Pi^{(t)} \triangleq \Pi_{\leqslant -\eta}(M^{(t)})$, choose a random $i \in [n]$ with probability $\frac{\Pi^{(t)}_{i,i}}{\text{Tr}(\Pi^{(t)})}$, and define $M^{(t+1)}$ as the matrix obtained by zeroing out the i-th row and column of $M^{(t)}$. Then increment t.

Let T be the time of termination and $\tau > 0$ be a large enough constant we choose later. We compute $\Pi^{(T)}$, and then compute as the set S of all indices i where $\Pi^{(T)}_{i,i} \leqslant \frac{\tau}{n}$. Define $\widetilde{\Pi}$ as $\left(\Pi^{(T)}_{S,S}\right)_{\geqslant \eta/K}$, and compute its span U, where we recall that $(X)_{\geqslant a}$ denotes the truncation of the eigendecomposition of X for eigenvalues at least a. This subspace U is passed to the next phase.

Phase 3: Rounding to a community assignment. Define r' as $r - 1$ when $\lambda_2(T) > 0$ and as r when $\lambda_2(T) < 0$. We first obtain an r'-dimensional embedding of the vertices into $\mathbb{R}^{r'}$. Compute a random r'-dimensional subspace U' of U, and take an orthogonal basis $u'_1, \ldots, u'_{r'}$. Place these vectors as a column of a matrix M' in $\mathbb{R}^{n \times r'}$. The rows of M' gives us the desired embedding.

On the other hand, we use the natural embedding of the k communities into $\mathbb{R}^{r'}$ induced by the r' nontrivial right eigenvectors corresponding to the eigenvalue $\lambda_2(T)$: $(\psi_i)_{1 \leqslant i \leqslant r'}$ of T. In particular, let $\Psi_{r'} \triangleq \begin{bmatrix} \psi_1 & \cdots & \psi_{r'} \end{bmatrix} \in \mathbb{R}^{k \times r'}$ be the matrix of these r' nontrivial eigenvectors of T. Then the row vectors $\phi_1, \ldots, \phi_k \in \mathbb{R}^{r'}$ form the desired embedding of communities.

In the rounding algorithm, we first find the largest c such that all the rows of $c \cdot M'$ lie in the convex hull of $\phi_1, \ldots, \phi_k$. We can find such a value of c if it exists by solving a linear program, and this $c > 0$ is guaranteed to exist by Lemma 7.1. Then, for each $i \in [n]$ we express each row of $c \cdot M'$ as a convex combination $\sum_{j=1}^k w_i^{(j)} \phi_j$ for nonnegative $w_i^{(j)}$ such that $\sum_{j=1}^k w_i^{(j)} = 1$. Finally, we assign vertex i to community j with probability $w_i^{(j)}$, and output the resulting community assignment $\widehat{x}$.

Remark 4.3. Scaling the rows of M' so as to lie in the convex hull of $\{\phi_j\}_{j\in[k]}$, is reminiscent of the rounding algorithm of Charikar & Wirth [CW04] to find a cut of size $\frac{1}{2} + \Omega(\frac{\varepsilon}{\log(1/\varepsilon)})$ in a graph with maximum cut of size $\frac{1}{2} + \varepsilon$: in their algorithm, they scale n scalars to lie in the interval $[-1, 1]$.

4.1 Analysis of algorithm

Our goal is to prove that the output $\widehat{x}$ of our algorithm is well-correlated with the true community assignment x. We begin by defining a notion of weak recovery for k-community stochastic block models.

Definition 4.4 (Weak recovery). Let $\Psi \triangleq \begin{bmatrix} \psi_2 & \cdots & \psi_k \end{bmatrix} \in \mathbb{R}^{k\times(k-1)}$ be the matrix of the top-$(k-1)$ nontrivial eigenvectors of the transition matrix T of a stochastic block model.

For $\rho > 0$, we say that a (randomized) algorithm for producing community assignments $\widehat{X} \in \mathbb{R}^{n\times k}$ achieves ρ-weak recovery if

$$\left\langle \mathbf{E}\,\widehat{X}_\Psi, X_\Psi \right\rangle \geqslant \rho \|\mathbf{E}\,\widehat{X}_\Psi\|_F \|X_\Psi\|_F,$$

where $B_\Psi \triangleq (B\Psi)(B\Psi)^\top$ for a matrix $B \in \mathbb{R}^{n\times k}$.

Remark 4.5. Intuitively, this notion is capturing the "advantage" of the algorithm over random guessing, or simply outputting the most likely community. See Appendix A for a more detailed discussion of this notion, how it recovers other previously considered measures of correlation in the case of the symmetric block model, and why it is meaningful. In particular, it implies the notion of weak recovery used in [DdNS22].

Our main guarantee is stated below.

Theorem 4.6. *For any SBM parameters* (M, π, d) *above the KS threshold, there is a constant* $\rho(\mathrm{M}, \pi, d) > 0$ *such that the above algorithm takes in the corrupted graph* $\widetilde{G}$ *as input and outputs* $\widehat{x}$ *achieving* $\rho(\mathrm{M}, \pi, d)$- *weak recovery with probability* $1 - o_n(1)$ *over the randomness of* $G \sim \mathrm{SBM}_n(\mathrm{M}, \pi, d)$.

To prove the above theorem it suffices to analyze $\left\langle (\mathbf{E}\,\widehat{X})_\Psi, X_\Psi \right\rangle$. To see why, let us first set up some notation. For each vertex i, we obtain a simplex vector $w_i \in \mathbb{R}^k$, which we can stack as rows into a weight matrix $W \in \mathbb{R}^{n\times k}$. We then independently round each vertex so that $\mathbf{E}\,\widehat{X} = W$.

To analyze our rounding scheme, first note that $\mathbf{E}[\widehat{X}_\Psi]$ is equal to W_Ψ off of the diagonal and is larger than W_Ψ on the diagonal, and thus $\mathbf{E}[\widehat{X}_\Psi] \succeq W_\Psi$. Since X_Ψ is positive semidefinite, $\left\langle \mathbf{E}[\widehat{X}_\Psi], X_\Psi \right\rangle \geqslant \langle W_\Psi, X_\Psi \rangle$. Thus, it suffices to lower bound $\langle W_\Psi, X_\Psi \rangle$. By construction, W_Ψ is equal to $c^2 \cdot \Pi_{U'}$, where recall that U' was a random r'-dimensional subspace of U, the output of Phase 2 of the algorithm. Thus,

$$\mathbf{E}_{U'}\, W_\Psi = c^2 \cdot \mathbf{E}_{U'}\, \Pi_{U'} = c^2 \cdot \frac{r'}{\dim(U)}\Pi_U.$$

In Lemma 6.10, we prove that when (M, π, d) are above the KS threshold, $\langle \Pi_U, X_\Psi \rangle \geqslant \Omega(1) \cdot \|\Pi_U\|_F \cdot \|X_\Psi\|_F$ and $\dim(U) = O(1)$. In Lemma 7.1, we show that when $\mathrm{diag}(\Pi_U) = O(1/n)$, we can take $c = \Omega(\sqrt{n})$; this delocalization condition is guaranteed by phase 2 of the algorithm. Combined with the fact that $\|\widehat{X}_\Psi\|_F = O(n)$, it follows that $\left\langle \mathbf{E}\,\widehat{X}_\Psi, X_\Psi \right\rangle \geqslant \Omega(1) \cdot \|\widehat{X}_\Psi\|_F \cdot \|X_\Psi\|_F$, which establishes Theorem 4.6.

5 The number of outlier eigenvalues

In this section, we pinpoint the number of outlier eigenvalues of $M_{G,\ell}(t) \triangleq A_G^{(\ell)} H_G(t) A_G^{(\ell)}$. Specifically, we will show the following theorem.

Theorem 5.1. *Let (M, π, d) be a model and let T be its corresponding Markov transition matrix. Let r be multiplicity of $\lambda_2(T)$. There exists constants ℓ, δ and α, such that for $t = \frac{1+\delta}{\lambda_2 d}$, with high probability over $G \sim \mathrm{SBM}_n(M, \pi, d)$, $M_{G,\ell}(t)$ has at least r negative eigenvalues of magnitude at most $-\alpha$, and at most $r + 1$ negative eigenvalues.*

We first establish the upper bound on the number of outlier eigenvalues, and then the lower bound. To establish the theorem, we will harness the Ihara-Bass formula that connects $H_G(t)$ to the non-backtracking walk matrix, and the work of Bordenave et al. [BLM15] on the spectrum of non-backtracking walk matrix B in stochastic block models. We begin by recalling these results from the literature.

Theorem 5.2 (Ihara–Bass formula). *For any simple, undirected, unweighted graph G on n vertices and m edges, we have*

$$\det(I - tB_G) = \det(H_G(t))(1 - t^2)^{m-n}$$

For a matrix X, we use $\lambda_i(X)$ to denote its i-th largest eigenvalue when sorted by absolute value.

Theorem 5.3 (Nonbacktracking spectrum of SBM [BLM15]). *For $G \sim \mathrm{SBM}_n(M, \pi, d)$:*

$$\lambda_i(B_G) = \lambda_i(T) \cdot d \pm o_n(1)$$

for all $1 \leqslant i \leqslant k$ such that $|\lambda_i(T)| > \frac{1}{\sqrt{d}}$ with high probability. All the remaining eigenvalues of B_G are at most $\sqrt{d} \cdot (1 + o_n(1))$ in magnitude with high probability.

The following is a consequence of the Ihara–Bass formula. Along with Theorem 5.3, this implies an upper bound on the number of outlier eigenvalues in Theorem 5.1.

Lemma 5.4. *For any $t^* > 0$, the number of negative eigenvalues of $H_G(t^*)$ is at most the number of real eigenvalues of B_G larger than $1/t^*$, and for $t^* < 0$, the number of negative eigenvalues of $H_G(t^*)$ is at most the number of real eigenvalues of B_G smaller than $1/t^*$.*

Proof. We present the proof assuming $t^* > 0$, since the case when $t^* < 0$ follows an identical proof. First, observe that the number of negative eigenvalues of $H_G(t^*)$ is at most:

$$\sum_{0 \leqslant t < t^* : H_G(t) \text{ is singular}} \dim \ker(H_G(t)),$$

since $H_G(0)$ has no negative eigenvalues, and the eigenvalues are a continuous function of t.

By the Ihara–Bass formula, if $H_G(t)$ is singular, then $\det(I - tB_G)$ must be 0, and in particular, $1/t$ must be an eigenvalue of B_G. We now prove that $\dim \ker(H_G(t))$ is at most the multiplicity of the eigenvalue $1/t$, from which the desired statement follows.

13

Let r be the dimension of the kernel of $H_G(t)$. We will show $\left(\frac{d}{du}\right)^j \det(I - uB_G)\Big|_{u=t} = 0$ for $0 \leqslant j \leqslant r - 1$, thereby implying that t occurs as a root of $\det(I - uB_G)$, and hence $1/t$ as an eigenvalue, with multiplicity at least r. Now, by the Ihara–Bass formula,

$$\left(\frac{d}{du}\right)^j \det(I - uB_G) = \left(\frac{d}{du}\right)^j \left[\det(H(u)) \cdot (1 - u^2)^{m-n}\right],$$

which by Lemma 3.3 is equal to

$$\sum_{S,T \subseteq [n],\, |S|=|T| \geqslant n-j} \det(H(u)_{S,T}) \cdot q_{S,T}(u)$$

for some collection of scalar functions $q_{S,T}$. Since $H(t)$ has a kernel of dimension r, by Fact 3.1 $H(t)_{S,T}$ is singular for any choice of $S, T \subseteq [n]$ with $|S| = |T| \geqslant n - r + 1$, and hence the above quantity is 0 for all $0 \leqslant j \leqslant r - 1$, which completes the proof. $\qquad \square$

5.1 Lower bound on the number of outlier eigenvalues

In this section, we lower bound the number of negative eigenvalues of $M_{G,\ell}(t)$, where $t > 0$. The same arguments apply to the case where $t < 0$, so we suppress the subscript in the dimension r of the outlier eigenspace for ease of notation.

To lower bound the number of outliers, we will show that with high probability the quadratic form of $M_{G,\ell}(t)$ is negative on an r-dimensional subspace spanned by vectors corresponding to the true community assignment. To do so, we will calculate the first and second moments of this random quadratic form, but the bulk of the work comes in calculating the first moment. The key insight here is that since these vectors are constructed using the true communities, the expectations can be tractably computed.

It is fruitful to give a combinatorial interpretation of the matrix $M_{G,\ell}(t)$. As is standard in trace moment calculations, the quadratic form can be decomposed into sums over various different graph shapes. Some standard accounting for these shapes then leads us to the final answer. To concretely articulate what we mean, we first define the notion of a *partially labeled shape* and its associated matrix.

Definition 5.5 (Shape matrix). A *shape* is a graph H along with two distinguished vertices: a *left vertex* u_L and a *right vertex* u_R. Given an n-vertex graph G, its shape matrix of G, denoted $M_H(G)$, is an $n \times n$ matrix where the (i, j) entry is the number of copies of H in G such that u_L can be identified with i and u_R with j. Equivalently,

$$M_H(G)_{i,j} := \sum_{\substack{\tau:V(H)\to[n] \\ \tau \text{ injective} \\ \tau(u_L)=i,\ \tau(u_R)=j}} \prod_{\{a,b\}\in E(H)} G_{\tau(a),\tau(b)}$$

Remark 5.6. We only consider shapes H whose size does not grow with n.

Let us explain at this point how we explicitly construct the r-dimensional subspace. The idea is to lift the r right eigenvectors $\psi_1, \ldots, \psi_r \in \mathbb{R}^k$ of T above the KS threshold using the indicator vectors $\mathbf{1}_c$ for $c \in [k]$. This makes the computation tractable because edges in the SBM are

conditionally independent given the true community assignment. Accordingly, we are interested in polynomials of the form:

$$\mathbf{1}_c^\top M_H(G)\mathbf{1}_{c'},$$

where c and c' are colors in $[k]$ and $\mathbf{1}_c$ is the indicator vector of vertices with color c. The above quadratic form counts the number of copies of H inside G where the vertex identified with u_L has color c and the vertex identified with u_R has color c'.

We use the following special case of [BMR21, Theorem 6.6].

Lemma 5.7. *Suppose H is a connected graph, then for $G \sim \mathrm{SBM}_n(\mathrm{M}, \pi, d)$, we have:*

$$\mathbf{1}_c^\top M_H(G)\mathbf{1}_{c'} = \sum_{\substack{\sigma:V(H)\to[k] \\ \sigma(u_L)=c,\sigma(u_R)=c'}} \prod_{v\in V(H)} \pi_{\sigma(v)} \prod_{ij\in H} M_{\sigma(i)\sigma(j)} \cdot d^{|E(H)|} \cdot n^{|E(H)|-|V(H)|} \pm o(n)$$

with probability $1 - o_n(1)$, where the $o(\cdot)$ hides factors depending on $|E(H)|$.

Remark 5.8. Observe that in the above if H is not a tree, then $|\mathbf{1}_c^\top M_H(G)\mathbf{1}_{c'}| = o(n)$. If H is a tree, then it can be expressed as a path $P = p_0, \ldots, p_\ell$ where $p_0 = u_L$ and $p_\ell = u_R$, along with a subtree per vertex on P. In this case, with high probability:

$$\mathbf{1}_c^\top M_H(G)\mathbf{1}_{c'} = \left\langle e_c, T^\ell e_{c'} \right\rangle_\pi \cdot d^{|V(H)|-1} \cdot n \pm o(n).$$

More generally, we are interested in understanding the quadratic form of vectors that are constant on each color class with $M_G(H)$, i.e. lifts of vectors in $\mathbb{R}^k$.

Corollary 5.9. *Let H be a tree such that u_L and u_R are distance-ℓ apart. For any $f : [k] \to \mathbb{R}$, define $f^{(n)} := \sum_{c\in[k]} f(c) \cdot \mathbf{1}_c$. Then with high probability, for all $f, g : [k] \to \mathbb{R}$ simultaneously:*

$$\langle f^{(n)}, M_H(G)g^{(n)} \rangle = \left\langle f, T^\ell g \right\rangle_\pi \cdot d^{|V(H)|-1} \cdot n \pm o(n).$$

The bulk of the contribution to the value of the quadratic form comes from tree shapes.

Corollary 5.10. *With high probability over $G \sim \mathrm{SBM}_n(\mathrm{M}, \pi, d)$, for any collection of shapes $\mathcal{H}$ along with coefficients $(c_H)_{H\in\mathcal{H}}$, and any functions $f, g : [k] \to \mathbb{R}$, we have:*

$$\left\langle f^{(n)}, \sum_{H\in\mathcal{H}} c_H M_H g^{(n)} \right\rangle = \left\langle f^{(n)}, \sum_{\substack{H\in\mathcal{H} \\ H\ \text{tree}}} c_H M_H g^{(n)} \right\rangle \pm o(n).$$

To lighten the notation, for the rest of this section we make the dependence on the original graph G implicit. So for our algorithm, we would like to illustrate negative eigenvalues in the spectrum of $M_\ell(t) = A^{(\ell)}H(t)A^{(\ell)}$.

Let $\psi_1, \ldots, \psi_r$ be the right eigenvectors of T corresponding to the largest nontrivial eigenvalue $\lambda_2(T) = \sqrt{\frac{1+\varepsilon}{d}}$ for some $\varepsilon > 0$. We will show that with high probability for all v in $\mathrm{span}\{\psi_1, \ldots, \psi_r\}$, the quadratic form $\langle v^{(n)}, M_\ell(t)v^{(n)} \rangle$ is substantially negative, where $t = \frac{1+\delta}{\sqrt{(1+\varepsilon)d}}$ for some small enough δ we choose in posterity.

Lemma 5.11. *For all ℓ sufficiently large and all $v \in \mathrm{span}\{\psi_1, \ldots, \psi_r\}$, $\langle v^{(n)}, M_\ell(t)v^{(n)} \rangle < -\alpha n$, where $\alpha > 0$ is an absolute constant. Consequently, $M_\ell(t)$ has at least r eigenvalues less than $-\alpha$.*

15

Proof. We can write $M_\ell(t)$ as a weighted combination of a collection of shape matrices $\mathcal{H}$. The plan is to extract out all the tree shapes that occur in $\mathcal{H}$ along with their weights and obtain a formula for the quadratic form using Corollary 5.9.

We expand $M_\ell(t) = A^{(\ell)}A^{(\ell)} - tA^{(\ell)}AA^{(\ell)} + t^2 A^{(\ell)}(D - I)A^{(\ell)}$. Each term counts walks following certain simple rules.

Observation 5.12. We can characterize the walks counted by each term as follows.

1. $A^{(\ell)}A^{(\ell)}[i, j]$ counts walks starting at i and ending at j with the phases (i) walk ℓ nonbacktracking steps, (ii) walk ℓ nonbacktracking steps.

2. $A^{(\ell)}AA^{(\ell)}[i, j]$ counts walks with phases (i) walk ℓ nonbacktracking steps, (ii) walk one step, (iii) walk ℓ nonbacktracking steps.

3. $A^{(\ell)}(D - I)A^{(\ell)}[i, j]$ counts walks with phases (i) walk $\ell + 1$ nonbacktracking steps and then backtrack one step, (ii) walk ℓ nonbacktracking steps.

Observation 5.13. For a walk to give rise to a tree shape, each step must either (i) visit a vertex not visited in the walk so far, or (ii) use an edge that has already been stepped on.

Using this rule, we can deduce all the tree shapes that come from each term. The tree shapes arising in $\mathcal{H}$ turn out to have quite simple structure. We use $H_{s,i,r}$ to denote the shape obtained by taking a length-s path where both endpoints are distinguished, and attaching a length-r path at vertex i.

Analysis of $A^{(\ell)}A^{(\ell)}$. By Observation 5.12 and Observation 5.13, any tree shape for this term must arise as a result of first walking ℓ self-avoiding steps, then backtracking s steps, and then walking $\ell - s$ self-avoiding steps for $0 \leqslant s \leqslant \ell$. The shape this gives rise to is $H_{2\ell-2s,\ell-s,s}$. Therefore, by Corollary 5.10, with high probability, for all $v : [k] \to \mathbb{R}$ we have:

$$\left\langle v^{(n)}, A^{(\ell)}A^{(\ell)}v^{(n)} \right\rangle = \sum_{s=0}^{\ell} \left\langle v^{(n)}, M_{H_{2\ell-2s,\ell-s,s}} v^{(n)} \right\rangle \pm o(n) = \sum_{s=0}^{\ell} \left\langle v, T^{2\ell-2s} v \right\rangle_\pi \cdot d^{2\ell-s} \cdot n \pm o(n).$$

Analysis of $A^{(\ell)}AA^{(\ell)}$. By Observation 5.12 and Observation 5.13, the tree shapes arising here must fall into one of the following cases.

1. Length-ℓ self-avoiding walk, then s backtracking edges, and then a length-$(\ell - s + 1)$ self-avoiding walk for $0 \leqslant s \leqslant \ell$. The shapes arising in this case are $H_{2\ell-2s+1,\ell-s,s}$.

2. Length-$(\ell + 1)$ self-avoiding walk, then s backtracking edges, and then a length-$(\ell - s)$ self-avoiding walk for $1 \leqslant s \leqslant \ell$. The shapes arising in this case are $H_{2\ell-2s+1,\ell+1-s,s}$.

3. Length-ℓ self-avoiding walk using edge e for its final step, then a backtracking step that uses e, and then a backtracking step that uses e for a third time, followed by a self-avoiding walk of length-$(\ell - 1)$. The shape arising in this case is simply a length-$(2\ell - 1)$ path with both endpoints distinguished.

16

By Corollary 5.10, with high probability, for all $v : [k] \to \mathbb{R}$:

$$\left\langle v^{(n)}, A^{(\ell)} A A^{(\ell)} v^{(n)} \right\rangle = \left(\langle v, T^{2\ell-1} v \rangle \cdot d^{2\ell-1} - \langle v, T^{2\ell+1} v \rangle \cdot d^{2\ell+1} \right) \cdot n +$$

$$\sum_{s=0}^{\ell} 2 \langle v, T^{2\ell-2s+1} v \rangle \cdot d^{2\ell-s+1} \cdot n \pm o(n).$$

Analysis of $A^{(\ell)}(D-I)A^{(\ell)}$. By Observation 5.12 and Observation 5.13, the tree shapes fall into one of the cases below.

1. Length-$(\ell+1)$ self-avoiding walk, then s nonbacktracking steps, followed by a length-$(\ell-s+1)$ self-avoiding walk for $1 \leqslant s \leqslant \ell + 1$. The shapes arising in this case are $H_{2\ell-2s+2, \ell-s+1, s}$.

2. Length-$(\ell+1)$ self-avoiding walk where the last used edge is e, a backtracking step that uses e, and then a backtracking step that uses e for a third time, followed by a length-$(\ell-1)$ self-avoiding walk. This shape is simply a length-2ℓ path with both endpoints distinguished.

By Corollary 5.10, with high probability, for all $v : [k] \to \mathbb{R}$:

$$\left\langle v^{(n)}, A^{(\ell)}(D-I)A^{(\ell)} v^{(n)} \right\rangle = \langle v, T^{2\ell} v \rangle \cdot d^{2\ell} \cdot n + \sum_{s=1}^{\ell+1} \langle v, T^{2\ell-2s+2} v \rangle \cdot d^{2\ell-s+2} \cdot n \pm o(n).$$

Estimating $\langle v^{(n)}, M_\ell(t) v^{(n)} \rangle$. By combining the above estimates, the value of $\langle v^{(n)}, M_\ell(t) v^{(n)} \rangle$ is: $n \cdot \langle v, p(T) v \rangle \pm o(n)$ where p is the univariate polynomial

$$p(\lambda) \triangleq -t \cdot d^{2\ell-1} \lambda^{2\ell-1} \cdot (1 - \lambda^2 d^2) + t^2 d^{2\ell} \lambda^{2\ell} + \sum_{s=0}^{\ell} d^{2\ell-s} \lambda^{2\ell-2s} - 2t d^{2\ell-s+1} \lambda^{2\ell-2s+1} + t^2 d^{2\ell-s+1} \lambda^{2\ell-2s}.$$

To prove that $\langle v, p(T) v \rangle$ is negative for all v in $\text{span}\{\psi_1, \ldots, \psi_r\}$, it suffices to show it is negative for $v = \psi_1, \ldots, \psi_r$ since those are also eigenvectors of $p(T)$. In particular, it suffices to show that $p(\lambda_2)$ is negative. Towards doing so, we first simplify the expression for $p(\lambda)$.

$$p(\lambda) = (\lambda d)^{2\ell} \cdot \left(t^2 - \frac{t}{\lambda d} \cdot (1 - \lambda^2 d^2) + \left(1 - 2\lambda dt + dt^2\right) \cdot \left(\frac{1 - (\lambda^2 d)^{-\ell-1}}{1 - (\lambda^2 d)^{-1}} \right) \right).$$

We plug in $\lambda = \sqrt{\frac{1+\varepsilon}{d}}$, $\ell = K \cdot \left(\frac{\log(1/\varepsilon)}{\varepsilon} \vee 1 \right)$ for some large $K > 0$, and $t = \frac{1+\delta}{\sqrt{(1+\varepsilon)d}}$ for some small $\delta > 0$ we choose in posterity, which yields:

$$(\lambda d)^{2\ell} \cdot \left(\frac{(1+\delta)^2}{(1+\varepsilon)d} - \frac{1+\delta}{(1+\varepsilon)d} + (1+\delta) + \left(1 - 2(1+\delta) + \frac{(1+\delta)^2}{1+\varepsilon} \right) \cdot \frac{1+\varepsilon}{\varepsilon} + o_K(1) \right)$$

$$= (\lambda d)^{2\ell} \cdot \left(\frac{\delta + \delta^2}{(1+\varepsilon)d} + (1+\delta) + \frac{1 + \varepsilon - 2(1+\delta)(1+\varepsilon) + (1+\delta)^2}{\varepsilon} + o_K(1) \right)$$

$$= (\lambda d)^{2\ell} \cdot \left(\frac{\delta + \delta^2}{(1+\varepsilon)d} + (1+\delta) + \left(-1 - 2\delta + \frac{\delta^2}{\varepsilon} \right) + o_K(1) \right)$$

$$= (\lambda d)^{2\ell} \cdot \left(-\delta + \frac{\delta}{(1+\varepsilon)d} + \frac{\delta^2}{(1+\varepsilon)d} + \frac{\delta^2}{\varepsilon} + o_K(1) \right),$$

where $o_K(1)$ denotes a quantity which vanishes as $K \to \infty$ regardless of the other parameters. The above expression can be made negative by choosing δ small enough, K large enough, and using that $d \geqslant 1$. Since $\lambda_2 = \sqrt{\frac{1+\varepsilon}{d}}$ for some $\varepsilon > 0$, for $\ell = O\left(\frac{\log(1/\varepsilon)}{\varepsilon}\right)$ there exists a choice of δ small enough and an absolute constant $\alpha > 0$ such that $p(\lambda_2) \leqslant -\alpha$. Consequently, for any v in $\mathrm{span}\{\psi_1, \ldots, \psi_r\}$, $\langle v, p(T)v \rangle \leqslant -\alpha$. $\qquad\square$

5.2 Outlier eigenspace after degree truncation

In this section we show that by picking the truncation threshold B to be large enough, the r-dimensional subspace which witnesses the negative eigenvalues for $M_{G,\ell}(t)$ also does the same for the truncated matrix $\overline{M}_{G_B,\ell}(t)$. In this section, we fix t as specified by Theorem 5.1 and suppress dependence on t, as the same arguments apply regardless of whether we consider positive or negative t.

Lemma 5.14 (Truncation doesn't affect negative witnesses). *Let $G \sim \mathrm{SBM}_n(M, \pi, d)$. For $B > 0$ sufficiently large, the following holds true with probability $1 - o_n(1)$. For all unit $x \in \mathrm{span}\{\psi_1, \ldots, \psi_r\}$, we have $\left\langle x^{(n)}, \overline{M}_{G_B,\ell} x^{(n)} \right\rangle < -\alpha_{5.14} n$, where $\alpha_{5.14} > 0$ is an absolute constant.*

To facilitate the discussion, we introduce the following notion of an *affected* vertex.

Definition 5.15 (Affected vertices). Let G be a graph. We say that a vertex $v \in V(G)$ is *affected* if its $2\ell + 1$ neighborhood differs in G and G_B, otherwise say that v is *unaffected*.

We also need the following concentration inequality, which follows from [BLM15, Lemma 29].

Lemma 5.16. *Let $G \sim \mathrm{SBM}_n(M, \pi, d)$. For any vertex v, let $\mathcal{N}_\ell(v)$ denote the distance-ℓ neighborhood of v in G. Then for any $s \geqslant 0$, there exists universal constants $C, c > 0$ such that*

$$\mathbf{Pr}[\exists \ell : |\mathcal{N}_\ell(v)| \geqslant s d^\ell] \leqslant C \exp(-cs).$$

In other words, the neighborhoods in SBMs look approximately regular. With these ingredients assembled, we can now prove Lemma 5.14.

Proof of Lemma 5.14. The desired statement follows from Lemma 5.11 once we prove $\left|\left\langle uu^\top, M_{G,\ell} - \overline{M}_{G_B,\ell} \right\rangle\right| \leqslant o_B(1)$, and choose B as a large enough constant.

Throughout the proof, we use the following crude estimates: if $|\mathcal{N}_{2\ell+1}(v)| \leqslant L$, then $\|M_{G,\ell}[v]\|_1 \leqslant 2L$, and $\left\|\overline{M}_{G_B,\ell}[v]\right\|_1 \leqslant B^{2\ell+1}$.

By Hölder's inequality we have

$$\left|\left\langle xx^\top, M_{G,\ell} - \overline{M}_{G_B,\ell} \right\rangle\right| \leqslant \left\|xx^\top\right\|_{\max} \left\|M_{G,\ell} - \overline{M}_{G_B,\ell}\right\|_1,$$

where $\left\|M_{G,\ell} - \overline{M}_{G_B,\ell}\right\|_1$ is the entrywise ℓ_1 norm. Let v be any vertex in G; we first upper bound the expected contribution of the vth row of $M_{G,\ell} - \overline{M}_{G_B,\ell}$ to the norm.

By our earlier observation and triangle inequality, we deduce

$$\left\|M_{G,\ell} - \overline{M}_{G_B,\ell}\right\|_1 \leqslant 2 \sum_v \mathbf{1}[v \text{ is affected}](|\mathcal{N}_{2\ell+1}(v)| + B^{2\ell+1})$$

18

$$= 2 \sum_v \mathbf{1}[\deg(v) > B] \sum_{w \in \mathcal{N}_{2\ell+1}(v)} (|\mathcal{N}_{2\ell+1}(w)| + B^{2\ell+1})$$

$$\leqslant 2 \sum_v \mathbf{1}[\deg(v) > B] |\mathcal{N}_{4\ell+2}(v)|(|\mathcal{N}_{4\ell+2}(v)| + B^{2\ell+1}),$$

since $\mathcal{N}_{4\ell+2}(v) \supseteq \mathcal{N}_{2\ell+1}(w)$. For ease of notation, introduce the quantity $f(v) \triangleq |\mathcal{N}_{4\ell+2}(v)|(|\mathcal{N}_{4\ell+2}(v)| + B^{2\ell+1})$. Taking expectations, we see that

$$\mathbf{E}\left[\left\|M_{G,\ell} - \overline{M}_{G_B,\ell}\right\|_1\right] \leqslant 2 \sum_v \mathbf{Pr}[\deg(v) > B]\,\mathbf{E}[f(v)|\deg(v) > B]$$

To proceed, we use the general property that if $\mathcal{E}$ is an event, X is a random variable with CDF Φ, and $f : \mathbb{R} \to \mathbb{R}$ is increasing, then $\mathbf{E}[f(X)|\mathcal{E}] \leqslant \mathbf{E}[f(X)|X \geqslant \Phi^{-1}(\mathbf{Pr}[\mathcal{E}])]$. Take $\mathcal{E}$ to be the event $\{\deg(v) > B\}$, and set $p \triangleq \mathbf{Pr}[\deg(v) > B]$. Then as $\mathbf{E}[X|X \geqslant \Phi^{-1}(p)] = \frac{\mathbf{E}[X\mathbf{1}[X \geqslant \Phi^{-1}(p)]]}{p}$, we can upper bound the conditional expectation by inverting Lemma 5.16. We see that

$$\mathbf{Pr}[\deg(v) > B]\,\mathbf{E}[f(v)|\deg(v) > B] \leqslant \int_{c\log(C/p)}^{\infty} y d^{4\ell+2}(y d^{4\ell+2} + B^{2\ell+1})\mathbf{Pr}[|\mathcal{N}_{4\ell+2}(v)| \geqslant y d^{4\ell+2}]dy$$

$$\leqslant C \int_{c\log(C/p)}^{\infty} y d^{4\ell+2}(y d^{4\ell+2} + B^{2\ell+1})\exp(-cy)dy$$

$$\leqslant O(B^{8\ell+4}) \cdot p(\log(1/p)^2 + \log(1/p)),$$

where the last line follows from from an explicit integral computation. Note that $p(\log(1/p)^2 + \log(1/p))$ is monotonically increasing for $p < 0.1$, so for sufficiently large B we can simply set p to be the upper bound furnished by Lemma 5.16. Indeed, since $\ell = O(1)$, we conclude that

$$\mathbf{Pr}[\deg(v) > B]\,\mathbf{E}[f(v)|\deg(v) > B] \leqslant O(B^{8\ell+6})\exp(-cB/d)$$

$$= o_B(1),$$

for $B = \Omega(\ell \log \ell)$. To summarize, we have just shown that

$$\mathbf{E}\left[\left\|M_{G,\ell} - \overline{M}_{G_B,\ell}\right\|_1\right] = o_B(1)n$$

Next, we bound the variance. Since $\ell = O(1)$, an Efron-Stein argument as before shows that the variance of $\left\|M_{G,\ell} - \overline{M}_{G_B,\ell}\right\|_1$ is $\widetilde{O}(n)$. More precisely, consider rerandomizing edge $(i, j) \in G$. With probability $1 - O(1/n)$, the rerandomized edge agrees with the original edge, in which case the ℓ_1 norm doesn't change. Otherwise, with probability $O(1/n)$, we can crudely bound the effect on the ℓ_1 norm by $(\log n)^\ell$. Hence $\mathbf{Var}\left(\left\|M_{G,\ell} - \overline{M}_{G_B,\ell}\right\|_1\right) \leqslant \widetilde{O}(n)$, and the desired concentration follows. $\qquad\square$

As a corollary, we obtain the analogous version of Theorem 5.1 for the truncated matrix $\overline{M}_{G_B,\ell}$.

Lemma 5.17 (Spectral gap for truncated matrix). *For sufficiently large $B > 0$, with high probability over $G \sim \mathrm{SBM}_n(\mathrm{M}, \pi, d)$, the matrix $\overline{M}_{G_B,\ell}$ has at most $r + 1$ negative eigenvalues, at least r of which are upper bounded by $-\alpha_{5.14}$.*

Proof. By construction (see Section 4), $\overline{H}_\ell$ is a principal submatrix of H_ℓ. This is where we use the non-standard way in which we truncated the degree matrix to get $\overline{H}_\ell$. It would not have been true if we had naively truncated D_{G_B}.

Since H_ℓ has at most $r+1$ negative eigenvalues, by the Cauchy interlacing theorem, its principal submatrix $\overline{H}_\ell$ has at most $r+1$ negative eigenvalues. The number of negative eigenvalues of $\overline{M}_{G_B,\ell} = A_{G_B}^{(\ell)} \overline{H}_\ell A_{G_B}^{(\ell)}$ is the same as that of $\overline{H}_\ell$. Hence $\overline{M}_{G_B,\ell}$ has at most $r+1$ negative eigenvalues, and Lemma 5.14 implies that $\overline{M}_{G_B,\ell}$ has at least r negative eigenvalues. Now using $\alpha_{5.14}$ guaranteed by the second part of Lemma 5.14, we conclude. $\qquad\square$

6 Robust recovery of a subspace

In this section, we consider the following algorithmic problem, which we recall from Section 2.2.

Assumption 6.1. Let M be a $n \times n$ matrix with at most r negative eigenvalues. Let $\widetilde{M}$ be a corrupted version of M. Formally, let $\widetilde{M} = M + \Delta$ where Δ is supported on a $\gamma n \times \gamma n$ submatrix, indexed by $Q \subseteq [n]$. Further, assume that the ℓ_1-norm of every row and column of M and Δ are bounded by K. Observe that this implies that $\|M\| \leqslant K$ and $\|\Delta\| \leqslant K$.

Problem 2.4. Given the corrupted matrix $\widetilde{M}$ as input, give an efficient algorithm to output a subspace U with the following properties:

1. **Low dimensional.** The dimension of U is $O(r)$.

2. **Delocalized.** The diagonal entries of its projection matrix Π_U are bounded by $O\left(\frac{r}{n}\right)$.

3. **Preserves delocalized part of negative eigenspace.** For any C-delocalized unit vector y such that $\langle y, My \rangle < -\Omega(1)$, we have $\langle y, \Pi_U y \rangle \geqslant \Omega(1)$.

We show that Phase 2 of the algorithm described in Section 4 with $\widetilde{M}$ as input indeed solves Problem 2.4, where a more precise quantitative dependence is articulated below.

Theorem 6.2. *For every $C, v, K > 0$, the following holds for all sufficiently small $\gamma > 0$: There is an efficient algorithm that takes in $\widetilde{M}$ from Assumption 6.1 along with C and v as input, and with high probability over the randomness of the algorithm, outputs a subspace U with the following properties.*

1. *The dimension of U is at most $O\left(\frac{K}{v} \cdot r\right)$.*

2. *The diagonal entries of the projection matrix Π_U are at most $O\left(\frac{C^2 K^3}{v^3} \cdot \frac{r}{n}\right)$.*

3. *For any C-delocalized unit vector y such that $\langle y, My \rangle < -v$, we have $\langle y, \Pi_U y \rangle \geqslant \Omega\left(\frac{v}{K}\right)$.*

To prove Theorem 6.2, we analyze Phase 2 of the algorithm in Section 4. The precise version of the algorithm we analyze is stated below.

Algorithm 6.3. The provided input is $\widetilde{M}$ according to Assumption 6.1, and real numbers v and C.
  Define $\eta \triangleq \frac{v}{48}$.

1. Define $\widetilde{M}^{(0)}$ as $\widetilde{M}$. Let t as a counter initialized at 0, and let $\Phi(X)$ be the number of eigenvalues of X smaller than $-\eta$.

2. While $\Phi(\widetilde{M}^{(t)}) > \frac{2K}{\eta}r$: compute the projection matrix $\Pi^{(t)} \triangleq \Pi_{\leqslant -\eta}(\widetilde{M}^{(t)})$, choose a random $i \in [n]$ with probability $\frac{\Pi^{(t)}_{i,i}}{\mathrm{Tr}(\Pi^{(t)})}$, and define $\widetilde{M}^{(t+1)}$ as the matrix obtained by zeroing out the i-th row and column of $\widetilde{M}^{(t)}$. Then increment t.

3. Let T be the time of termination and let $\tau = \frac{2C^2K^2r}{\eta^2}$. Compute $\Pi^{(T)}$, and define the set of indices S as,
$$S = \left\{ i \middle| \Pi^{(T)}_{i,i} \leqslant \frac{\tau}{n} \right\} .$$
Define $\widetilde{\Pi}$ as $\left(\Pi^{(T)}_{S,S} \right)_{\geqslant \eta/K}$, and compute its span U.[1]

4. Output U.

We first bound on the number of eigenvalues of $\widetilde{M}^{(T)}$ smaller than $-\eta$, thus establishing Item 1 of Theorem 6.2.

6.1   Proof of Item 1: spectrum of cleaned-up matrix

Lemma 6.4. *The matrix $\widetilde{M}^{(T)}$ has the following properties:*

1. *$\widetilde{M}^{(T)}$ has at most $\frac{2K}{\eta}r$ eigenvalues smaller than $-\eta$.*

2. *$\widetilde{M}^{(T)}$ is a submatrix $\widetilde{M}_{R,R}$ where $|R| \geqslant \left(1 - \frac{4K}{\eta}\gamma\right)n$ with high probability over the randomness of the algorithm.*

In service of proving Lemma 6.4, we prove the following statement about the localization of outlier eigenvectors on the indices of the corruptions Q.

Lemma 6.5. *Let $R^{(t)} \subseteq [n]$ denote the non-zero rows and columns of $\widetilde{M}^{(t)}$ and let $Q^{(t)} = Q \cap R^{(t)}$. Then the following holds:*
$$\frac{\mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right)_{Q^{(t)},Q^{(t)}} \right)}{\mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right) \right)} \geqslant \frac{\eta}{K} - \frac{r}{\mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right) \right)}$$

Proof. The desired statement follows from the chain of inequalities below.

$$-\eta \cdot \mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right) \right) \geqslant \mathrm{Tr}\left(\widetilde{M}^{(t)}_{\leqslant -\eta} \right)$$
$$= \left\langle \widetilde{M}^{(t)}, \Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right) \right\rangle$$
$$= \left\langle M_{R^{(t)},R^{(t)}} + \Delta_{R^{(t)},R^{(t)}}, \Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right) \right\rangle$$

---
[1]Recall that $(X)_{\geqslant a}$ denotes the truncation of the eigendecomposition of X for eigenvalues at least a.

$$\geq \mathrm{Tr}\left(\left(M_{R^{(t)},R^{(t)}}\right)_{\leqslant 0}\right) - \|\Delta\| \cdot \mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right)_{Q^{(t)},Q^{(t)}}\right)$$

$$\geq -Kr - K \cdot \mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right)_{Q^{(t)},Q^{(t)}}\right).$$

In the last step, we used the fact that $\left\|M_{R^{(t)},R^{(t)}}\right\| \leqslant \|M\| \leqslant K$ and the number of negative eigenvalues of $(M)_{R^{(t)},R^{(t)}}$ is at most the number of negative eigenvalues of M, namely r. $\qquad\square$

We now prove Lemma 6.4.

Proof of Lemma 6.4. By definition of the iterative process, Item 1 is satisfied on termination, and the process has to terminate since it cannot run for more than n steps.

It remains to prove Item 2. Towards doing so, first observe that the process terminates if all rows and columns in Q are zeroed out. Indeed, M itself satisfies the termination condition, and if all the rows and columns in Q are zeroed out, then the support is a principal submatrix of M, which satisfies the termination condition by Cauchy's interlacing theorem.

Suppose the termination condition is not satisfied at iteration t, then the i-th row/column of the matrix is zeroed out where i is chosen with probability $\frac{\Pi_{\leqslant -\eta}(\widetilde{M}^{(t)})_{i,i}}{\mathrm{Tr}\left(\Pi_{\leqslant -\eta}(\widetilde{M}^{(t)})\right)}$. Consequently, the probability that i is in Q is equal to:

$$\frac{\mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right)_{Q^{(t)},Q^{(t)}}\right)}{\mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right)\right)},$$

which by Lemma 6.5 is at least

$$\frac{\eta}{K} - \frac{r}{\mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right)\right)}.$$

Since $\mathrm{Tr}\left(\Pi_{\leqslant -\eta}\left(\widetilde{M}^{(t)}\right)\right)$ counts the number of eigenvalues smaller than $-\eta$ and the termination condition is not satisfied, the above probability is at least $\frac{\eta}{2K}$.

Since the process terminates if all of Q is zeroed out (in which case $R = \overline{Q}$), a standard martingale argument shows that the process can last for at most $\frac{4K}{\eta}\gamma n$ steps with high probability. $\qquad\square$

We now proceed to proving Items 2 and 3 of Theorem 6.2.

6.2 Proofs of Items 2 and 3: delocalization and correlation with recovered subspace

The proof of Item 2 is fairly straightforward.

Lemma 6.6. *All the diagonal entries of Π_U, the projection matrix onto U, are bounded by $\frac{2C^2K^3}{\eta^3} \cdot \frac{r}{n}$.*

Proof. Let S be the set of indices whose diagonal entries are at most $\frac{\tau}{n}$, where $\tau = \frac{2C^2K^2}{\eta^2} \cdot r$. By definition, $\Pi_{S,S}$ has entries bounded by $\frac{\tau}{n}$, and since $\Pi_{S,S} \succeq (\Pi_{S,S})_{\geqslant \frac{\eta}{K}} = \widetilde{\Pi}$, the diagonal entries of $\widetilde{\Pi}$ are at most $\frac{\tau}{n}$. All eigenvalues of $\widetilde{\Pi}$ are at least $\frac{\eta}{K}$, and hence $\frac{K}{\eta} \cdot \widetilde{\Pi} \succeq \Pi_U$, which gives us a bound of $\frac{K\tau}{\eta n}$ on the diagonal entries of Π_U, from which the statement follows. $\qquad\square$

Now, we turn our attention to proving Item 3. We first show that $\widetilde{M}_{R,R} = \widetilde{M}^{(T)}$ preserves negative correlations with delocalized vectors. Henceforth, we assume that the high probability guarantee from Lemma 6.4 that $|R| \geqslant \left(1 - \frac{4K}{\eta}\gamma\right)n$ holds.

Lemma 6.7. *For unit C-delocalized vector y such that $\langle y, My \rangle < -v$, we have:* $\left\langle y, \widetilde{M}_{R,R}y \right\rangle < -\frac{v}{2}$.

Proof. By Assumption 6.1 and Lemma 6.4, we can write $\widetilde{M}_{R,R}$ as $M + E_1 + E_2$ where E_1 is supported on at most $\left(\frac{4K}{\eta} + 1\right)\gamma n$ rows and E_2 is supported on at most $\left(\frac{4K}{\eta} + 1\right)\gamma n$ columns. Consequently, the entrywise ℓ_1 norm of $E \triangleq E_1 + E_2$ is at most $\frac{10K^2}{\eta}\gamma n$. The statement follows from the inequality below for small enough γ:

$$\langle y, Ey \rangle \leqslant \|yy^\top\|_{\max} \cdot \|E\|_1 \leqslant O\left(\frac{10K^2C^2}{\eta}\gamma\right) < \frac{v}{2}. \quad \square$$

We have the following immediate corollary of Lemma 6.7.

Corollary 6.8. *For any C-delocalized unit vector y such that $\langle y, My \rangle < -v$, we have:*

$$\left\langle y, \Pi_{\leqslant -\eta}\left(\widetilde{M}_{R,R}\right)y \right\rangle \geqslant \frac{v/2 - \eta}{2K - \eta}.$$

Proof. Starting with Lemma 6.7, we get

$$-\frac{v}{2} > \left\langle y, \widetilde{M}_{R,R}y \right\rangle$$
$$\geqslant -\left\|\widetilde{M}_{R,R}\right\| \cdot \left\langle y, \Pi_{\leqslant -\eta}\left(\widetilde{M}_{R,R}\right)y \right\rangle + (-\eta) \cdot \left(1 - \left\langle y, \Pi_{\leqslant -\eta}\left(\widetilde{M}_{R,R}\right)y \right\rangle\right).$$

Since the spectral norm of a matrix is at most its maximum ℓ_1 norm of a row, $\|\widetilde{M}_{R,R}\| \leqslant 2K$. Substituting in the above inequality and simplifying yields the desired result. $\quad \square$

We are finally ready to establish Item 3.

Lemma 6.9. *For any unit C-delocalized vector y such that $\langle y, My \rangle < -v$,*

$$\langle y, \Pi_U y \rangle \geqslant \Omega\left(\frac{v}{K}\right).$$

Proof. We abbreviate $\Pi_{\leqslant -\eta}\left(\widetilde{M}_{R,R}\right)$ to Π. Recall $\tau = \frac{2C^2K^2r}{\eta^2}$ from Algorithm 6.3, and let S denote the set of indices where the diagonal entries of Π are at most $\frac{\tau}{n}$. By Lemma 6.4, $\text{Tr}(\Pi) \leqslant \frac{2K}{\eta}r$, and hence $|\overline{S}| \leqslant \frac{2Kr}{\eta\tau}n$. By positivity of Π and Corollary 6.8,

$$\langle y, \Pi_{S,S}y \rangle + \left\langle y, \Pi_{\overline{S},\overline{S}}y \right\rangle \geqslant \frac{1}{2}\langle y, \Pi y \rangle \geqslant \frac{1}{2} \cdot \frac{v/2 - \eta}{2K - \eta}.$$

The second term of the LHS is equal to

$$\left\langle yy_{\overline{S},\overline{S}}^\top, \Pi \right\rangle \leqslant \text{Tr}\left(yy_{\overline{S},\overline{S}}^\top\right) \cdot \|\Pi\| \leqslant \frac{C^2|\overline{S}|}{n} \leqslant \frac{\eta}{K},$$

23

and consequently,

$$\langle y, \Pi_{S,S} y \rangle \geqslant \frac{1}{2} \cdot \frac{v/2 - \eta}{2K - \eta} - \frac{\eta}{K}.$$

Finally, since $\widetilde{\Pi}$ is obtained by discarding eigenvalues of $\Pi_{S,S}$ of magnitude at most $\frac{\eta}{K}$:

$$\left\langle y, \widetilde{\Pi} y \right\rangle \geqslant \frac{1}{2} \cdot \frac{v/2 - \eta}{2K - \eta} - \frac{2\eta}{K} \geqslant \frac{v/8 - 3\eta}{K} \geqslant \frac{v}{16K}.$$

The statement follows since $\widetilde{\Pi} \preceq \Pi_U$. □

6.3 Correlation of subspace with communities

In the model of adversarial corruptions we are considering, a graph G is sampled from $\mathrm{SBM}_n(\mathrm{M}, \pi, d)$. The graph then undergoes δn adversarial edge insertions and deletions, and the resulting graph $\widetilde{G}$ is handed to us.

In this section, we prove a guarantee on the subspace U produced by Phase 2 of the algorithm in Section 4 as a special case of Theorem 6.2.

Lemma 6.10. *The subspace U produced as output at the end of phase 2 of the algorithm described in Section 4 has the following guarantees with high probability over the randomness of the stochastic block model.*

1. *The dimension of U is at most $O\left(\frac{B^{2\ell+3}}{\alpha_{5.14}} r\right)$.*

2. *All the diagonal entries of Π_U are bounded by $O\left(\frac{B^{6\ell+9}}{\alpha_{5.14}^3 \pi_{\min}} \cdot \frac{r}{n}\right)$.*

3. *For any unit x in $\mathrm{span}\{\psi_1, \ldots, \psi_r\}$,*

$$\left\langle x^{(n)}, \Pi_U x^{(n)} \right\rangle \geqslant \Omega\left(\frac{\alpha_{5.14}}{B^{2\ell+3}} \cdot n\right).$$

To prove Lemma 6.10, we plug in the following into the setup of Theorem 6.2.

$$M = \overline{M}_{G_B,\ell}, \Delta \triangleq \overline{M}_{\widetilde{G}_B,\ell} - \overline{M}_{G_B,\ell}, K = 2B^{2\ell+3}, C = \frac{1}{\sqrt{\pi_{\min}}}, \gamma = 8B^{2\ell+2}\delta, \text{ and } v = \alpha_{5.14}.$$

Our proof of Lemma 6.10 is complete once we verify that:

1. $\frac{1}{\sqrt{n}} x^{(n)}$ is C-delocalized and achieves quadratic form less than $-v$ with our choice of M.

2. The choice of γ is indeed a sufficiently small constant depending on v, C, and K.

3. M, Δ, and K satisfy Assumption 6.1.

Verifying the assumption on $x^{(n)}$. Note that Lemma 5.14 establishes that the quadratic form of $x^{(n)}$ is sufficiently negative, and the delocalization follows from the fact that $\|x^{(n)}\|_\infty \leqslant \|x\|_\infty \leqslant \frac{1}{\sqrt{\pi_{\min}}}$ since x is a unit vector, along with the fact that $\|x^{(n)}\|$ concentrates around $\sqrt{n}$.

Verifying the assumption on γ. The magnitude of γ can be made sufficiently small by choosing δ as a sufficient small constant.

Verifying the assumption on $M, \Delta,$ and K. The below lemma proves why our choice of M, Δ and K satisfy Assumption 6.1.

Lemma 6.11. *The matrix* $\Delta \triangleq \overline{M}_{G_B,\ell} - \overline{M}_{\widetilde{G}_B,\ell}$ *is supported on a* $8B^{2\ell+2}\delta n \times 8B^{2\ell+2}\delta n$ *principal submatrix. Further, each row and column of* $\overline{M}_{G_B,\ell}$ *and* $\overline{M}_{\widetilde{G}_B,\ell}$ *have* ℓ_1 *norm bounded by* $B^{2\ell+3}$, *and consequently each row and column of* Δ *has* ℓ_1 *norm bounded by* $2B^{2\ell+3}$.

Proof of Lemma 6.11. First, observe that $\overline{D}_{G_B}$ differs from $\overline{D}_{\widetilde{G}_B}$ only on rows and columns corresponding to vertices incident to a corrupted edge, of which there are at most $2\delta n$.

Next, we bound the number of vertices whose neighborhoods differ between G_B and $\widetilde{G}_B$. Observe that if a vertex v has differing neighborhoods between G_B and $\widetilde{G}_B$, then either v is incident to an edge corruption, or v has a neighbor in $G_B \cup \widetilde{G}_B$ that is truncated in one of $\widetilde{G}_B$ or G_B but not the other.

The number of vertices incident to an edge corruption is at most $2\delta n$. We can bound the number of vertices of the second kind by:

$$\sum_{v \in V(G)} \sum_{u \in N_{G_B \cup \widetilde{G}_B}(v)} |\mathbf{1}[\deg_G(u) > B] - \mathbf{1}[\deg_{\widetilde{G}}(u) > B]|$$

$$= \sum_{u \in V(G)} \sum_{v \in N_{G_B \cup \widetilde{G}_B}(u)} |\mathbf{1}[\deg_G(u) > B] - \mathbf{1}[\deg_{\widetilde{G}}(u) > B]|$$

$$\leqslant \sum_{u \in V(G)} B \cdot |\mathbf{1}[\deg_G(u) > B] - \mathbf{1}[\deg_{\widetilde{G}}(u) > B]|$$

$$\leqslant 2\delta B n$$

where the last inequality is because $\deg_G(u)$ can differ from $\deg_{\widetilde{G}}(u)$ only when u is incident to an edge corruption. Thus, the number of vertices with differing neighborhoods is at most $2\delta(B+1)n \leqslant 4\delta B n$.

Finally, $\overline{M}_{\widetilde{G}_B,\ell}$ and $\overline{M}_{G_B,\ell}$ can differ only on rows and columns corresponding to vertices that are at most $2\ell + 1$ away in either $\widetilde{G}_B$ or G_B from a vertex with differing neighborhoods. Since the degrees in both graphs are bounded by B, there are at most $8\delta B^{2\ell+3}n$ such vertices.

The ℓ_1 norm of any row v of either $\overline{M}_{\widetilde{G}_B,\ell}$ or $\overline{M}_{G_B,\ell}$ is bounded by the total number of walks leaving v of length at most $2\ell + 2$, which is at most $B^{2\ell+3}$. Thus, the ℓ_1 norm of any row of Δ is at most $2B^{2\ell+3}$, which completes the proof. $\qquad\square$

7 Rounding algorithm

In this section we fill in the details for Section 4.1, which analyzed the rounding guarantees. First, we review the rounding procedure described in Section 4. From phase 2 of the algorithm, we obtain a constant-dimensional delocalized subspace U which has constant correlation with the true communities. We process U into a new matrix $M' \in \mathbb{R}^{n \times r'}$ which is directly used in the rounding scheme. Consider the matrix $\Psi_{r'}$ of the r' right eigenvectors of T corresponding to λ_2, assumed to be above the KS threshold:

$$\Psi_{r'} \triangleq \begin{bmatrix} \psi_1 & \cdots & \psi_{r'} \end{bmatrix}.$$

The rows of $\Psi_{r'}$ induce row embeddings $\{\phi_i\}_{i \in [k]}$ into $\mathbb{R}^{r'}$. We then rescale the rows of M' by a uniform factor c such that we can express each row of $c \cdot M'$ as a convex combination of the row embeddings $\{\phi_i\}_{i \in [k]}$. These weights are then used to sample a valid community assignment.

In [Lemma 7.1](#) we prove that the scaling c used to obtain the rounding weights can be taken to be $\Omega(\sqrt{n})$ so long as the subspace U from phase 2 of the algorithm is delocalized.

Lemma 7.1. *Suppose that* $\mathrm{diag}(\Pi_U) \leqslant O(\frac{1}{n})$. *There exists* $c = \Theta(\sqrt{n})$ *such that the scaled rows of* $c \cdot M'$ *lie within the convex hull of the* ϕ_i's.

Proof. We first show that we can pick some $c = \Omega(1)$. Since $\psi_1 = \mathbf{1}$ is the trivial eigenvector of T, by spectral theory we have that $\langle \mathbf{1}, \psi_j \rangle_\pi = 0$ for $j \neq 1$. This implies that the convex combination $\sum_{i=1}^k \pi(i)\phi_i = 0 \in \mathbb{R}^{r-1}$, so the origin is in the convex hull of the row embeddings ϕ_i. Note that this alone does not suffice to show that we can take $c > 0$ because the origin might be on the boundary of the hull.

However, as we'll show in [Lemma 7.2](#) below, the origin in fact lies in the interior of the convex hull of the ϕ_i's, which permits us to take $c > 0$. Furthermore, the minimum distance from the origin to the boundary is bounded below by a positive constant (possibly depending on k), because 0 lies in the interior and the ϕ_i's are constant-sized objects. So we can take $c = \Omega(1)$. To attain $c = \Theta(\sqrt{n})$, first recall that $\mathrm{diag}(\Pi_U) \leqslant O(\frac{1}{n})$ by assumption, which implies that $\mathrm{diag}(\Pi'_U) \leqslant O(\frac{1}{n})$ since $U' \subseteq U$. By definition, $\Pi_{U'} = M'M'^{\top}$, the ℓ_2 norm of each row of M' is $O(\frac{1}{\sqrt{n}})$, from which the desired claim holds. $\square$

Lemma 7.2. *Assuming that* $\pi > 0$, *we have* $0 \in \mathrm{int}(\mathrm{conv}(\phi_1, \dots, \phi_k))$.

Proof. First, note that since the column eigenvectors ψ_j are orthogonal with respect to $\langle \cdot, \cdot \rangle_\pi$, and $\pi > 0$, the matrix of column eigenvectors has full row rank. On the other hand, we know that $0 \in \mathrm{conv}(\phi_1, \dots, \phi_k)$ because $\sum_i \pi(i)\phi_i = 0$. Since the ϕ_i are linearly independent, they in fact lie on the boundary of the convex hull and the interior of the convex hull is a nonempty open set in $\mathbb{R}^r$. As $\pi > 0$, we have just shown that 0 lies in the interior of the convex hull. $\square$

Acknowledgments

References

[Abb17]   Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017. 3, 29

[AS15]   Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015. 1

[Bas92]   Hyman Bass. The Ihara-Selberg zeta function of a tree lattice. *International Journal of Mathematics*, 3(06):717–797, 1992. 4

[BKS23]   Rares-Darius Buhai, Pravesh K Kothari, and David Steurer. Algorithms approaching the threshold for semi-random planted clique. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1918–1926, 2023. 3

[BLM15]   Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015. 1, 4, 13, 18

[BMR21]   Jess Banks, Sidhanth Mohanty, and Prasad Raghavendra. Local Statistics, Semidefinite Programming, and Community Detection. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1298–1316. SIAM, 2021. 1, 2, 3, 15

[BS95]   Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *Journal of Algorithms*, 19(2):204–234, 1995. 3

[CLMW11]   Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011. 6

[CSV17]   Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017. 3

[CW04]   Moses Charikar and Anthony Wirth. Maximizing quadratic programs: Extending grothendieck's inequality. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 54–60. IEEE, 2004. 0, 12

[DdHS23]   Jingqiu Ding, Tommaso d'Orsi, Yiding Hua, and David Steurer. Reaching kesten-stigum threshold in the stochastic block model under node corruptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4044–4071. PMLR, 2023. 3

[DdNS22]   Jingqiu Ding, Tommaso d'Orsi, Rajai Nasser, and David Steurer. Robust recovery for stochastic block models. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 387–394. IEEE, 2022. 0, 1, 2, 3, 12, 29

[DG06]   Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, 2006. 30

[DKMZ11]   Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical review E*, 84(6):066106, 2011. 1

[DMS17]   Amit Dembo, Andrea Montanari, and Subhabrata Sen. Extremal cuts of sparse random graphs. *The Annals of Probability*, 45(2):1190–1217, 2017. 2

[FK01]   Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001. 3

[FM17]   Zhou Fan and Andrea Montanari. How well do local algorithms solve semidefinite programs? In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 604–614, 2017. 5

[GHKM23] Venkatesan Guruswami, Jun-Ting Hsieh, Pravesh K Kothari, and Peter Manohar. Efficient Algorithms for Semirandom Planted CSPs at the Refutation Threshold. *arXiv preprint arXiv:2309.16897*, 2023. 3

[GP23]   Yuzhou Gu and Yury Polyanskiy. Uniqueness of BP fixed point for the potts model and applications to community detection. *arXiv preprint arXiv:2303.14688*, 2023. 1

[GV16]   Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016. 1

[HS17]   Samuel B Hopkins and David Steurer. Efficient Bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017. 1

[Iha66]   Yasutaka Ihara. On discrete subgroups of the two by two projective linear group over p-adic fields. *Journal of the Mathematical Society of Japan*, 18(3):219–235, 1966. 4

[KMM⁺13] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013. 0, 1, 5, 6

[KS66]   Harry Kesten and Bernt P Stigum. Additional limit theorems for indecomposable multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, 37(6):1463–1481, 1966. 1

[KS67]   Harry Kesten and Bernt P Stigum. Limit theorems for decomposable multidimensional galton-watson processes. *Journal of Mathematical Analysis and Applications*, 17(2):309–338, 1967. 1

[LM22]   Allen Liu and Ankur Moitra. Minimax rates for robust community detection. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 823–831. IEEE, 2022. 1, 3

[Mas14]   Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703, 2014. 1, 5

[MMT20]   Theo McKenzie, Hermish Mehta, and Luca Trevisan. A new algorithm for the robust semi-random independent set problem. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 738–746. SIAM, 2020. 3

[MNS14]   Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Conference on Learning Theory*, pages 356–370. PMLR, 2014. 1

[MNS18]   Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018. 1, 5

[MPW16]   Ankur Moitra, William Perry, and Alexander S Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 828–841, 2016. 1, 2

[MS16]   Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 814–827, 2016. 3

[SKZ14]   Alaa Saade, Florent Krzakala, and Lenka Zdeborová. Spectral clustering of graphs with the bethe hessian. *Advances in neural information processing systems*, 27, 2014. 0, 1, 5, 6

[SM19]   Ludovic Stephan and Laurent Massoulié. Robustness of spectral methods for community detection. In *Conference on Learning Theory*, pages 2831–2860. PMLR, 2019. 3

[YP23]   Qian Yu and Yury Polyanskiy. Ising model on locally tree-like graphs: Uniqueness of solutions to cavity equations. *IEEE Transactions on Information Theory*, 2023. 1

A   On defining weak-recovery

One might wonder whether our definition of weak recovery (Definition 4.4) is meaningful. In Appendix A.1, we argue that our definition is natural and interpret it in various natural settings. In Appendix A.2, we show that our definition implies the other definitions of weak recovery found in the literature.

For self-containedness, we reproduce our and previous definitions of weak recovery below.

Definition 4.4 (Weak recovery). Let $\Psi \triangleq \begin{bmatrix} \psi_2 & \cdots & \psi_k \end{bmatrix} \in \mathbb{R}^{k \times (k-1)}$ be the matrix of the top-$(k-1)$ nontrivial eigenvectors of the transition matrix T of a stochastic block model.

For $\rho > 0$, we say that a (randomized) algorithm for producing community assignments $\widehat{X} \in \mathbb{R}^{n \times k}$ achieves ρ-weak recovery if

$$\left\langle \mathbf{E}\,\widehat{X}_\Psi, X_\Psi \right\rangle \geqslant \rho \|\mathbf{E}\,\widehat{X}_\Psi\|_F \|X_\Psi\|_F,$$

where $B_\Psi \triangleq (B\Psi)(B\Psi)^\top$ for a matrix $B \in \mathbb{R}^{n \times k}$.

Next, we recall the notion of weak recovery used in [DdNS22]. Note that their metric is specialized to the symmetric 2-community block model.

Definition A.1 (Weak recovery — symmetric 2-community SBM). We say that an algorithm achieves weak recovery for the symmetric 2-community SBM if it outputs a labeling $\widehat{x} \in \{\pm 1\}^n$ such that

$$\mathbf{E}_{x,G}\left[|\langle x, \widehat{x}\rangle|\right] \geqslant \Omega(n).$$

Moving beyond the symmetric 2-community setting, we recall a more general notion of weak recovery present in the literature, which applies to arbitrary k-community SBMs [Abb17]. To define it, it is helpful to introduce the shorthand $\Omega_i = \{v \in [n] : X_v = i\}$ for $i \in [k]$.

Definition A.2 (Weak recovery — nontrivial partition). We say that an algorithm achieves weak recovery in the sense of nontrivial partition if it outputs with high probability a subset of vertices $S \subseteq V$ such that there exists communities $i, j \in [k]$ with

$$\frac{|\Omega_i \cap S|}{|\Omega_i|} - \frac{|\Omega_j \cap S|}{|\Omega_j|} \geqslant \Omega(1).$$

Note that Definition A.2 implies Definition A.1, by setting $\widehat{x}_v = +1$ for each $v \in S$ and $\widehat{x}_v = -1$ for each $v \in \overline{S}$.

Another reasonable definition would require that the mutual information between X and $\widehat{X}$ is nontrivially large.

Definition A.3 (Weak recovery — mutual information). We say that an algorithm achieves weak recovery in the sense of mutual information if

$$\sum_{v \in [n]} \mathrm{I}(X_v; \widehat{X}_v) \geqslant \Omega(n).$$

By chain rule and monotonicity of entropy, the above guarantee also implies that $\mathrm{I}(X; \widehat{X}) \geqslant \Omega(n)$.

A.1 Interpreting our new notion of weak recovery

In this section, we show that if the correlation goes to 1 then we achieve exact recovery on r communities, and if the algorithm uses no information about the graph then we do not achieve ρ-weak recovery for any $\rho > 0$ as $n \to \infty$.

First, let us establish some elementary formulas which will prove useful. Since $\Psi \in \mathbb{R}^{k \times (k-1)}$ is the matrix of right eigenvectors of T with the all-ones vector removed, we have $\Psi \Psi^\top = \Pi^{-1} - \mathbf{1}_k \mathbf{1}_k^\top$. We therefore have

$$\langle W_\Psi, X_\Psi \rangle = \mathrm{Tr}\big(\Psi^\top W^\top X \Psi \Psi^\top X^\top W \Psi\big)$$
$$= \big\| \Psi^\top W^\top X \Psi \big\|_F^2.$$

It's not hard to see that conditioned on X and W, we have

$$(W^\top X)_{ij} = n \mathbf{Pr}_{v \sim [n], \widehat{X}}[\widehat{X}_v = i | X_v = j],$$

and below we suppress the subscript in the probabilities for sake of conciseness. Notice that the above formula connects our recovery metric to the *confusion matrix*, a popular metric for multiclass classification in machine learning [DG06]. In particular, the confusion matrix $P \in \mathbb{R}^{k \times k}$ is defined by $P_{ij} \triangleq \mathbf{Pr}[\widehat{X}_v = i | X_v = j]$, which is exactly the expression on the RHS.

On the other hand, we can calculate

$$\langle W_\Psi, X_\Psi \rangle = \big\langle W(\Pi^{-1} - \mathbf{1}_k \mathbf{1}_k^\top)W^\top, X(\Pi^{-1} - \mathbf{1}_k \mathbf{1}_k^\top)X^\top \big\rangle$$
$$= \big\langle W\Pi^{-1}W^\top, X\Pi^{-1}X^\top \big\rangle - \big\langle \mathbf{1}_n \mathbf{1}_n^\top, X\Pi^{-1}X^\top + W\Pi^{-1}W^\top \big\rangle + n^2.$$

Now using the fact that $\mathbf{1}_n^\top X = n\pi^\top(1 + \tilde{O}(1/\sqrt{n}))$, the fact that π is a distribution, and writing $\widehat{\pi} \triangleq \frac{1}{n}\mathbf{1}_n^\top W$, we obtain

$$\langle W_\Psi, X_\Psi \rangle = \big\| \Pi^{-1/2}W^\top X\Pi^{-1/2} \big\|_F^2 - n^2 \big\| \Pi^{-1/2}\widehat{\pi} \big\|_2^2 + \tilde{O}(n^{3/2})$$

$$= n^2 \left(\sum_{i,j} (\pi(i)\pi(j))^{-1} \mathbf{Pr}[\widehat{X}_v = i, X_v = j]^2 - \left\| \Pi^{-1/2}\widehat{\pi} \right\|_2^2 \right) + \tilde{O}(n^{3/2})$$

$$= n^2 \left(\sum_{i,j} \pi(i)^{-1}\pi(j) \mathbf{Pr}[\widehat{X}_v = i | X_v = j]^2 - \sum_i \pi(i)^{-1} \mathbf{Pr}[\widehat{X}_v = i]^2 \right) + \tilde{O}(n^{3/2}) \tag{4}$$

with high probability over X.

Exact recovery for r communities Suppose there are r eigenvalues of T above the KS threshold. Let U be the subspace spanned by $\{\psi_1^{(n)}, \ldots \psi_r^{(n)}\}$. If we achieve perfect correlation with U, then any orthonormal basis B of U with $U = BB^\top$ exactly recovers at least r of the communities. This follows from the fact that B has full row rank, so there are r rows which are pairwise distinct. From these rows we can exactly recover r communities.

Interpretation of the metric in the symmetric k-community block model Suppose we are in the symmetric k-community block model, with parameter λ, so that $\pi = \frac{1}{k}\mathbf{1}$ and $T = k\lambda I + (1 - \lambda)\mathbf{1}\mathbf{1}^\top$ with $a \neq b$. The trivial top eigenvector is $\mathbf{1}$, and for the remaining eigenvectors we can pick any orthonormal basis in $\mathbf{1}^\perp$. Then $\Psi\Psi^\top = kI - \mathbf{1}\mathbf{1}^\top$, and from Eq. (4), when we achieve weak recovery we have

$$\frac{1}{n^2} \langle W_\Psi, X_\Psi \rangle = \sum_{i,j} \left(\mathbf{Pr}[\widehat{X}_v = i | X_v = j]^2 - \mathbf{Pr}[\widehat{X}_v = i]^2 \right) + \tilde{O}(1/\sqrt{n})$$

$$\geqslant \Omega(1).$$

Therefore, when we succeed at weak recovery in this setting, there exists some j such that

$$\sum_i \mathbf{Pr}[\widehat{X}_v = i | X_v = j]^2 \geqslant \sum_i \mathbf{Pr}[\widehat{X}_v = i]^2 + \Omega(1).$$

For example, this correlation is maximized if after permuting the columns of W we have $W = X$.

Comparing to natural baselines There are two obvious baselines to sanity check our metric for: labeling vertices randomly according to the prior π, and deterministically guessing the largest community. In fact, we will characterize the correlation for any randomized rounding scheme which is independent of the true assignment. In this case, we have $W = \mathbf{1}w^\top$, where w is a simplex vector which is independent of X. Then we have

$$\langle W_\Psi, X_\Psi \rangle = \left\| \Psi^\top w \mathbf{1}^\top X\Psi \right\|_F^2$$
$$= n^2 \left\| \Psi^\top w\pi^\top \Psi \right\|_F^2 (1 + \tilde{O}(1/\sqrt{n}))$$
$$= \tilde{O}(n^{3/2}),$$

where the second line holds with high probability over X, and the third line holds because we removed $\mathbf{1}$ from Ψ. On the other hand, we know that $\|W_\Psi\|_F \|X_\Psi\|_F = \Theta(n^2)$, so indeed these baselines fail to achieve weak recovery.

Indistinguishability and the KS threshold Finally, suppose that two communities are actually indistinguishable, in the precise sense that their corresponding rows in T are identical. In our rounding scheme, notice that the rows for Ψ are identical so indeed these two communities are "merged" from the perspective of the correlation guarantee. In other words, we could have just reduced the weights $W \in \mathbb{R}^{n \times k}$ into weights $W' \in \mathbb{R}^{n \times (k-1)}$ in the obvious way.

This is an appealing property because it is a modest step towards understanding the more general setting where two communities are *computationally distinguishable*. For example, one could consider a 4-community setup where communities 1 and 2 are computationally indistinguishable (as well as communities 3 and 4), but the pairs of communities $(1, 2)$ and $(3, 4)$ are computationally distinguishable from each other.

A.2 Comparing the different notions of weak recovery

In this section, we show that our definition of weak recovery Definition 4.4 implies Definitions A.2 and A.3.

Lemma A.4 (Weak recovery implies nontrivial partition). *If a recovery algorithm achieves weak recovery in the sense of Definition 4.4, then it also satisfies Definition A.2.*

Proof. If $\langle W_\Psi, X_\Psi \rangle \geqslant \Omega(n^2)$, then it follows that

$$\sum_{c,c' \in [k]} \pi(c)^{-1} \pi(c') \mathbf{Pr}[\widehat{X}_v = c | X_v = c']^2 - \sum_c \pi(c)^{-1} \mathbf{Pr}[\widehat{X}_v = c]^2 \geqslant \Omega(1),$$

and we conclude that there must exist some c' such that

$$\sum_c \pi(c)^{-1} (\mathbf{Pr}[\widehat{X}_v = c | X_v = c']^2 - \mathbf{Pr}[\widehat{X}_v = c]^2) \geqslant \Omega(1).$$

It follows that there must exist some $i \in [k]$ for which $\mathbf{Pr}[\widehat{X}_v = i | X_v = c'] \geqslant \mathbf{Pr}[\widehat{X}_v = i] + \Omega(1)$. After relabeling the communities for X, using any $\sigma \in S_k$ which swaps c' and i, we can assume that in fact

$$\mathbf{Pr}[\widehat{X}_v = i | X_v = i] \geqslant \mathbf{Pr}[\widehat{X}_v = i] + \Omega(1) \tag{5}$$

Moreover, with i as defined above, since $\mathbf{Pr}[\widehat{X}_v = i] = \sum_c \mathbf{Pr}[\widehat{X}_v = i | X_v = c] \pi(c)$, by averaging there must exist some $j \neq i$ such that

$$\mathbf{Pr}[\widehat{X}_v = i | X_v = i] \geqslant \mathbf{Pr}[\widehat{X}_v = i | \widehat{X}_v = j] + \Omega(1). \tag{6}$$

With Eq. (6) in hand, it is easy to see how to define the set S for the partition in Definition A.2. In particular, set $S = \{v : \hat{X}_v = i\}$, and pick the same i and j as in the above guarantee. □

Lemma A.5 (Weak correlation implies nontrivial mutual information). *If $\langle W_\Psi, X_\Psi \rangle \geqslant \Omega(n^2)$, then $\sum_{v \in [n]} I(X_v; \widehat{X}_v) \geqslant \Omega(n)$.*

Proof. Notice that Eq. (5) and another averaging argument implies that there exists a subset S of $\Omega(n)$ vertices such that for each $v \in S$,

$$\mathbf{Pr}_{\widehat{X}_v, X, G}[\widehat{X}_v = i | X_v = i] \geqslant \mathbf{Pr}_{\widehat{X}_v, X, G}[\widehat{X}_v = i] + \Omega(1),$$

where we have used the subscript to emphasize that the probability no longer samples a random $v \sim [n]$. Now by Pinsker's inequality it suffices to lower bound $d_{\mathsf{TV}}(p_{X_v, \widehat{X}_v}, p_{X_v} p_{\widehat{X}_v})$. Since $\pi(i) \geqslant \Omega(1)$, the above inequality implies that $d_{\mathsf{TV}}(p_{X_v, \widehat{X}_v}, p_{X_v} p_{\widehat{X}_v}) \geqslant \Omega(1)$, so we conclude that

$$I(X_v; \widehat{X}_v) \geqslant \Omega(1).$$

Summing up over all v and using the fact that mutual information is nonnegative, we conclude that $I(X; \widehat{X}) \geqslant \Omega(n)$. $\qquad\square$