Bayesian estimation from few samples: community detection and related problems

Samuel B. Hopkins* David Steurer†

September 9, 2019

Abstract

We propose an efficient meta-algorithm for Bayesian estimation problems that is based on low-degree polynomials, semidefinite programming, and tensor decomposition. The algorithm is inspired by recent lower bound constructions for sum-of-squares and related to the method of moments. Our focus is on sample complexity bounds that are as tight as possible (up to additive lower-order terms) and often achieve statistical thresholds or conjectured computational thresholds.

Our algorithm recovers the best known bounds for community detection in the sparse stochastic block model, a widely-studied class of estimation problems for community detection in graphs. We obtain the first recovery guarantees for the mixed-membership stochastic block model (Airoldi et el.) in constant average degree graphs—up to what we conjecture to be the computational threshold for this model. We show that our algorithm exhibits a sharp computational threshold for the stochastic block model with multiple communities beyond the Kesten–Stigum bound—giving evidence that this task may require exponential time.

The basic strategy of our algorithm is strikingly simple: we compute the best-possible low-degree approximation for the moments of the posterior distribution of the parameters and use a robust tensor decomposition algorithm to recover the parameters from these approximate posterior moments.

*Cornell University. Supported by an NSF graduate research fellowship, a Microsoft Research PhD fellowship, a Cornell University fellowship, and David Steurer's NSF CAREER award. Part of this work was accomplished while this author was an intern at Microsoft Research New England.

†ETH Zürich. Much of this work was done while at Cornell University and the Institute for Advanced Study. Supported by a Microsoft Research Fellowship, a Alfred P. Sloan Fellowship, NSF awards, and the Simons Collaboration for Algorithms and Geometry.

Contents

1 Introduction

Bayesian[1] *parameter estimation* [Wik17a] is a basic task in statistics with a wide range of applications, especially for machine learning. The estimation problems we study have the following form: For a known joint probability distribution $p(x, \theta)$ over data points x and parameters θ (typically both high-dimensional objects), nature draws a parameter $\theta \sim p(\theta)$ from its marginal distribution and we observe i.i.d. samples $x_1, \ldots, x_m \sim p(x \mid \theta)$ from the distribution conditioned on θ. The goal is to efficiently estimate the underlying parameter θ from the observed samples $x_1, \ldots, x_m$.

A large number of important problems in statistics, machine learning, and average-case complexity fit this description. Some examples are principal component analysis (and its many variants), independent component analysis, latent Dirichlet allocation, stochastic block models, planted constraint satisfaction problems, and planted graph coloring problems.

For example, in stochastic block models the parameter θ imposes a community structure on n nodes. In the simplest case, this structure is a partition into two communities. Richer models support more than two communities and allow nodes to participate in multiple communities. The samples $x_1, \ldots, x_m$ are edges between the nodes drawn from a distribution $p(x \mid \theta)$ that respects the community structure θ, which typically means that the edge distribution is biased toward endpoints with the same or similar communitity memberships. Taken together the samples $x_1, \ldots, x_m$ form a random graph x on n vertices that exhibits a latent community structure θ; the goal is to estimate this structure θ. This problem becomes easier the more samples (i.e., edges) we observe. The question is how many samples are required such that we can efficiently estimate the community structure θ? Phrased differently: how large an average degree of the random graph x do we require to be able to estimate θ?

In this work, we develop a conceptually simple meta-algorithm for Bayesian estimation problems. We focus on the regime that samples are scarce. In this regime, the goal is to efficiently compute an estimate $\hat{\theta}$ that is positively correlated with the underlying parameter θ given as few samples from the distribution as possible. In particular, we want to understand whether for particular estimation problems there is a difference between the sample size required for efficient and inefficient algorithms (say, exponential vs. polynomial time). In this regime, we show that our meta-algorithm recovers the best previous bounds for stochastic block models [Mas14, MNS15a, AS16a]. Moreover, for the case of richer community structures like multiple communities and especially overlapping communities, our algorithm achieves significantly stronger recovery guarantees.[2]

In order to achieve these improved guarantees, our meta-algorithm draws on several ideas from previous lines of work and combines them in a novel way. Concretely, we draw on ideas from recent analyses of belief propagation and their use of non-backtracking and self-avoiding random walks [Mas14, MNS15a, AS16a]. We also use ideas from recent works based on the method of moments and tensor decomposition [AGH+14, AGHK14, BKS15]. Our algorithm also employs

[1] Here, "Bayesian" refers to the fact that there is a prior distribution over the parameters.

[2] If we represent the community structure by k vectors $y_1, \ldots, y_k \in \{0, 1\}^n$ that indicate community memberships, then previous algorithms [AS16a] do not aim to recover these vectors but, roughly speaking, only a random linear combination of them. While for some settings it is in fact impossible to estimate the individual vectors, we show that in many settings it is possible to estimate them (in particular for symmetric block models).

convex-programming techniques, namely the sum-of-squares semidefinite programming hierarchy, and gives a new perspective on how these techniques can be used for estimation.[3]

Our meta-algorithm allows for a tuneable parameter which corresponds roughly to running time. Under mild assumptions on a Bayesian estimation problem $p(x, \theta)$ (that are in particular satisfied for discrete problems such as the stochastic block model), when this parameter is set to allow the meta-algorithm to run in exponential time, if there is any estimator $\hat{\theta}$ of θ obtaining correlation δ, the meta-algorithm offers one obtaining correlation at least $\delta^{O(1)}$. While this parameter does not correspond directly to the *degree* paramter used in convex hiararchies such as sum of squares, the effect is similar to the phenomenon that sum of squares convex programs of exponential size can solve any combinatorial optimization problem exactly. (Since Bayesian estimation problems do not always correspond to optimization problems, this guarantee would not be obtained by sum of squares in our settings.)

For many Bayesian estimation problems there is a critical number of samples n_0 such that when the number of samples n is less than n_0, computationally-efficient algorithms seem unable to compute good estimators for θ. This is in spite of the presence of sufficient information to identify θ (and therefore estimate it with computationally inefficient algorithms), even when $n < n_0$. Providing rigorous evidence for such *computational thresholds* has been a long-standing challenge. One popular approach is to prove impossibility of estimating θ from $n < n_0$ samples using algorithms from some restricted class. Such results are most convincing the chosen class captures the lowest-sample-complexity algorithms for many Bayesian inference problems, which our meta-algorithm does.[4] We prove that in the k-community block model, no algorithm captured by our meta-algorithm can tolerate smaller-degree graphs than the best known algorithms. This provides evidence for a computational phase transition at the *Kesten-Stigum threshold* for stochastic block models.

Organization In the remainder of this introduction we discuss our results and their relation to previous work in more detail. In Section 2 (Techniques) we describe the mathematical techniques involved in our meta-algorithm and its analysis, and we illustrate how to apply the meta-algorithm to recover a famous result in the theory of spiked random matrices with a much simplified proof. In Section 3 (Warmup) we re-prove (up to some loss in the running time) the result of Mossel-Neeman-Sly on the two-community block model as an application of our meta-algorithm, again with very simple proofs. In Section 4 (Matrix estimation) we re-interpret the best existing results on the block model, due to Abbe and Sandon, as applications of our meta-algorithm.

In Section 5 (Tensor estimation) we apply our meta-algorithm to the mixed-membership block model. Following that, in Section 6 (Lower bounds) we prove that no algorithm captured by our

[3]Previously, convex-programming techniques have been used in this context only as a way to obtain efficient relaxations for maximum-likelihood estimators. In contrast, our work uses convex programming to drive the method of moments approach and decompose tensors in an entropy maximizing way.

[4]Recent work in this area has focused on sum of squares lower bounds [HSS15, MW15, BHK+16]. While the sum of squares method is algorithmically powerful, it is not designed to achieve optimal sample guarantees for Bayesian estimation. Lower bounds against our meta-algorithm therefore serve better the purpose of explaining precise computational sample thresholds.

meta-algorithm can recover communities in the block model past the Kesten-Stigum threshold.

In Section 7 (Tensor decomposition), which can be read independently of much of the rest of the paper, we give a new algorithm for tensor decomposition and prove its correctness; this algorithm is used by our meta-algorithm as a black box.

1.1 Meta-algorithm and meta-theorems for Bayesian estimation

We first consider a version of the meta-algorithm that is enough to capture the best known algorithms for the stochastic block model with k disjoint communities, which we now define. Let $\varepsilon, d > 0$. Draw y uniformly from $[k]^n$. For each pair $i \neq j$, add the edge $\{i, j\}$ to a graph on n vertices with probability $(1 + (1 - \frac{1}{k})\varepsilon)\frac{d}{n}$ if $y_i = y_j$ and $(1 - \frac{\varepsilon}{k})\frac{d}{n}$ otherwise. The resulting graph has expected average degree d.

A series of recent works has explored the problem of estimating y in these models for the sparsest-possible graphs. The emerging picture, first conjectured via techniques from statistical physics in the work [DKMZ11], is that in the k-community block model it is possible to recover a nontrivial estimate of y via a polynomial time algorithm if and only if $d = (1 + \delta)\frac{k^2}{\varepsilon^2}$ for $\delta \geqslant \Omega(1)$. This is called the Kesten-Stigum threshold. The algorithmic side of this conjecture was confirmed by [Mas14, MNS15a] for $k = 2$ and [AS16a] for general k.

One of the goals of our meta-algorithm is that it apply in a straightforward way even to complex Bayesian estimation problems. A more complex model (yet more realistic for real-world networks) is the *mixed-membership* block model [ABFX08] which we now define informally. Let $\alpha \geqslant 0$ be an overlap parameter. Draw y from $\binom{k}{t}^n$, where $t = \frac{k(\alpha+1)}{k+\alpha} \approx \alpha + 1$; that is for each of n nodes pick a set S_j of roughly $\alpha + 1$ communities.[5] For each pair i, j, add an edge to the graph with probability $(1 + (\frac{|S_i \cap S_j|}{t^2} - \frac{1}{k})\varepsilon)\frac{d}{n}$. (That is, with probability which increases as i and j participate in more communities together.) In the limit $\alpha \to 0$ this becomes the k-community block model.

Returning to the meta-algorithm (but keeping in mind the block model), let $p(x, y)$ be a joint probability distribution over observable variables $x \in \mathbb{R}^n$ and hidden variables $y \in \mathbb{R}^m$. Nature draws (x, y) from the distribution p, we observe x and our goal is to provide an estimate $\hat{y}(x)$ for y. Often the mean square error $\mathbb{E}_{p(x,y)} \left\|\hat{y}(x) - y\right\|^2$ is a reasonable measure for the quality of the estimation. For this measure, the information-theoretically optimal estimate is the mean of the posterior distribution $\hat{y}(x) = \mathbb{E}_{p(y|x)} y$. This approach has two issues that we address in the current work.

The first issue is that naively computing the mean of the posterior distribution takes time exponential in the dimension of y. For example, if $y \in \{\pm 1\}^m$, then $\mathbb{E}_{p(y|x)} y = \sum_{y \in \{\pm 1\}^m} y \cdot p(y \mid x)$; there are 2^m terms in this sum. There are many well-known algorithmic approaches that aim to address this issue or related ones, for example, belief propagation [Gal62, Pea82] or expectation maximization [DLR77]. While these approaches appear to work well in practice, they are notoriously difficult to analyze.

[5] In actuality one draws for each node $i \in [n]$ a probability vector $\sigma_i \in \Delta_{k-1}$ from the Dirichlet distribution with parameter α; we describe a nearly-equivalent model here for the sake of simplicity—see Section 1.2 for details. Our guarantees for recovery in the mixed-membership model also apply to the model here because it has the same second moments as the Dirichlet distribution.

In this work, we can resolve this issue in a very simple way: We analytically determine a low-degree polynomial $f(x)$ so that $\mathbb{E}_{p(x,y)} \left\| f(x) - y \right\|^2$ is as small as possible and use the fact that low-degree polynomials can be evaluated efficiently (even for high dimensions n).[6] Because the maximum eigenvector of an n-dimensional linear operator with a spectral gap is an $O(\log n)$-degree polynomial of its entries, our meta-algorithm captures spectral properties of linear operators whose entries are low-degree polynomials of observable variables x. Examples of such operators include adjacency matrices (when x is a graph), empirical covariance matrices (when x is a list of vectors), as well as more sophisticated objects such as linearized belief propagation operators (e.g., [AS15]) and the Hashimoto non-backtracking operator.

The second issue is that even if we could compute the posterior mean exactly, it may not contain any information about the hidden variable y and the mean square error is not the right measure to assess the quality of the estimator. This situation typically arises if there are symmetries in the posterior distribution. For example, in the stochastic block model with two communities we have $\mathbb{E}_{p(y|x)} y = 0$ regardless of the observations x because $p(y \mid x) = p(-y|x)$. A simple way to resolve this issue is to estimate higher-order moments of the hidden variables. For stochastic block models with disjoint communities, the second moment $\mathbb{E}_{p(y|x)} yy^\mathsf{T}$ would suffice. (For overlapping communities, we need third moments $\mathbb{E}_{p(y|x)} y^{\otimes 3}$ due to more substantial symmetries.)

For now, we think of y as an m-dimensional vector and x as an n-dimensional vector (in the blockmodel on N nodes, this would correspond to $m \approx kN$ and $n = N^2$). Our algorithms follow a two-step strategy:

1. Given $x \sim p(x|y)$, evaluate a fixed, low-degree polynomial $P(x)$ taking values in $(\mathbb{R}^m)^{\otimes \ell}$. (Usually ℓ is 2 or 3.)
2. Apply a robust eigenvector or semidefinite-programming based algorithm (if $\ell = 2$), or a robust tensor decomposition algorithm (if $\ell = 3$ or higher) to P to obtain an estimator $\hat{y}$ for y.

The polynomial $P(x)$ should be an optimal low-degree approximation to $y^{\otimes \ell}$, in the following sense: if n is sufficiently large that some low-degree polynomial $Q(x)$ has constant correlation with $y^{\otimes \ell}$

$$\mathop{\mathbb{E}}_{x,y} \langle Q, y^{\otimes \ell} \rangle \geqslant \Omega(1) \cdot (\mathop{\mathbb{E}}_{x} \|Q\|^2)^{1/2} (\mathbb{E} \|y^{\otimes \ell}\|^2)^{1/2} ,$$

then P has this guarantee. (The inner products and norms are all Euclidean.)

A prerequisite for applying our meta-algorithm to a particular inference problem $p(x, y)$ is that it be possible to estimate y given $\mathbb{E}\left[y^{\otimes \ell} \mid x\right]$ for some constant ℓ. For such a problem, the optimal Bayesian inference procedure (ignoring computational constraints) can be captured by computing $F(x) = \mathbb{E}\left[y^{\otimes \ell} \mid x\right]$, then using it to estimate y. When $p(x, y)$ is such that it is information-theoretically possible to estimate y from x, these posterior moments will generally satisfy $\mathbb{E}\langle F(x), y^{\otimes \ell} \rangle \geqslant \Omega(1) \cdot (\mathbb{E} \|F(x)\|^2)^{1/2} (\mathbb{E} \|y^{\otimes \ell}\|^2)^{1/2}$, for some constant ℓ. Our observation is that when F is approximately a low-degree function, this estimation procedure can be carried out via an efficient algorithm.

[6] Our polynomials typically have logarithmic degree and naive evaluation takes time $n^{O(\log n)}$. However, we show that under mild conditions it is possible to approximately evaluate these polynomials in polynomial time using the idea of color coding [AYZ95].

Matrix estimation and prior results for block models In the case $\ell = 2$, where one uses the covariance $\mathbb{E}\left[yy^{\mathsf{T}} \mid x\right]$ to estimate y, the preceding discussion is captured by the following theorem.

Theorem 1.1 (Bayesian estimation meta-theorem—2nd moment)**.** *Let $\delta > 0$ and $p(x, y)$ be a distribution over vectors $x \in \{0,1\}^n$ and unit vectors $y \in \mathbb{R}^d$. Assume $p(x) \geqslant 2^{-n^{O(1)}}$ for all $x \in \{0,1\}^n$.*[7] *Suppose there exists a matrix-valued degree-D polynomial $P(x)$ such that*

$$\mathop{\mathbb{E}}_{p(x,y)} \langle P(x), yy^{\mathsf{T}} \rangle \geqslant \delta \cdot \left(\mathop{\mathbb{E}}_{p(x)} \|P(x)\|_F^2 \right)^{1/2} . \tag{1.1}$$

Then, there exists $\delta' \geqslant \delta^{O(1)} > 0$ and an estimator $\hat{y}(x)$ computable by a circuit of size $n^{O(D)}$ such that

$$\mathop{\mathbb{E}}_{p(x,y)} \langle \hat{y}(x), y \rangle^2 \geqslant \delta' . \tag{1.2}$$

To apply this theorem to the previously-discussed setting of samples $x_1, \ldots, x_N$ generated from $p(x \mid y)$, assume the samples $x_1, \ldots, x_N$ are in some fixed way packaged into a single n-length vector x.

One curious aspect of the theorem statement is that it yields a nonuniform algorithm—a family of circuits—rather than a uniform algorithm. If the coefficients of the polynomial P can themselves be computed in polynomial time, then the conclusion of the algorithm is that an $n^{O(D)}$-time algorithm exists with the same guarantees.

As previously mentioned, the meta-algorithm has a parameter D, the degree of the polynomial P. If $D = n$, then whenever it is information-theoretically possible to estimate y from $\mathbb{E}[yy^\top \mid x]$, the meta-algorithm can do so (in exponential time). This follows from the fact that every function in n Boolean variables is a polynomial of degree at most n. It is also notable that, while a degree D polynomial can be evaluated by an $n^{O(D)}$-size circuit, some degree-D polynomials can be evaluated by much smaller circuits. We exploit such polynomials for the block model (computable via *color coding*), obtaining $n^{O(1)}$-time algorithms from degree $\log n$ polynomials. By using very particular polynomials, which can be computed via powers of *non-backtracking operators*, previous works on the block model are able to give algorithms with near-linear running times [MNS15a, AS16a].[8]

Using the appropriate polynomial P, this theorem captures the best known guarantees for partial recovery in the k-community stochastic block model. Via the same polynomial, applied in the mixed-membership setting, it also yields our first nontrivial algorithm for the mixed-membership model. However, as we discuss later, the recovery guarantees are weak compared to our main theorem.

[7] This mild condition on the marginal distribution of x allows us to rule out pathological situations where a low-degree polynomial in x may be hard to evaluate accurately enough because of coefficients with super-polynomial bit-complexity.

[8] In this work we choose to work with *self-avoiding* walks rather than non-backtracking ones; while the corresponding polynomials cannot to our knowledge be evaluated in near-linear time, the analysis of these polynomials is much simpler than the analysis needed to understand non-backtracking walks. This helps to make the analysis of our algorithms much simpler than what is required by previous works, at the cost of large polynomial running times. It is an interesting question to reduce the running times of our algorithm for the mixed-membership block model to near-linear via non-backtracking walks, but since our aim here is to distinguish what is computable in polynomial time versus, say, exponential time, we do not pursue that improvement here.

Recalling the ε, d, k block model from the previous section, let $y \in \mathbb{R}^n$ be the centered indicator vector of, say, community 1.

Theorem 1.2 (Implicit in [Mas14, MNS15a, AS16a], special case of our main theorem, Theorem 1.4)**.** *Let $\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d}$. If x is sampled according to the n-node, k-community, ε-biased, α-mixed-membership block model with average degree d and y is the centered indicator vector of community 1, there is a $n \times n$-matrix valued polynomial P of degree $O(\log n)/\delta^{O(1)}$ such that*

$$\mathop{\mathbb{E}}_{x} \langle P(x), yy^\mathsf{T} \rangle \geqslant \left(\frac{\delta}{k(\alpha+1)}\right)^{O(1)} (\mathbb{E}\,\|P(x)\|^2)^{1/2} (\mathbb{E}\,\|yy^\mathsf{T}\|^2)^{1/2}\,.$$

Together with Theorem 1.1, up to questions of $n^{O(\log n)}$ versus $n^{O(1)}$ running times, when $\alpha \to 0$ this captures the previous best efficient algorithms for the k-community block model. (Once one has a unit vector correlated with y, it is not hard to approximately identify the vertices in community 1.) While the previous works [Mas14, MNS15a, AS16a] did not consider the mixed-membership blockmodel, this theorem is easily obtained using techniques present in those works (as we show when we rephrase those works in our meta-algorithm, in Section 4).[9]

Symmetries in the posterior, tensor estimation, and improved error guarantees We turn next to our main theorem on the mixed-membership model, which offers substantial improvement on the correlation which can be obtained via Theorem 1.2. The matrix-based algorithm discussed above, Theorem 1.2, contains a curious asymmetry; namely the arbitrary choice of community 1. The block model distributions are symmetric under relabeling of the communities, which means that any estimator $P(x)$ of yy^T is also an estimator of $y'y'^\mathsf{T}$, where y' is the centered indicator of community $j > 1$. Since one wants to estimate all the vectors $y_1, \ldots, y_k$ (with y_i corresponding to the i-th community), it is more natural to consider the polynomial P to be an estimator of the matrix $M = \sum_{i \in [k]} y_i y_i^\mathsf{T}$.[10] Unsurprisingly, P is a better estimator of M than it is of y_1. In fact, with the same notation as in the theorems,

$$\mathop{\mathbb{E}}_{x,y} \langle P(x), M(y) \rangle \geqslant \delta^{O(1)} (\mathbb{E}\,\|P(x)\|^2)^{1/2} (\mathbb{E}\,\|M(y)\|^2)^{1/2}\,,$$

removing the $k^{O(1)}$ factor in the denominator. This guarantee is stronger: now the error in the estimator depends only on the distance δ of the parameters $\varepsilon, d, k, \alpha$ from the critical threshold $\frac{k^2(\alpha+1)^2}{\varepsilon^2 d} = 1$ rather than additionally on k.

If given the matrix M exactly, one way to extract an estimator $\hat{y}_i$ for some y_i is just to sample a random unit vector in the span of the top k eigenvectors of M. Such an estimator $\hat{y}_i$ would have

[9] In fact, if one is willing to lose an additional 2^{-k} in the correlation obtained in this theorem, one can obtain a similar result for the mixed-membership model by reducing it to the disjoint-communities with $K \approx 2^k$ communities, one for each subset of k communities. This works when each node participates in a subset of communities; if one uses the Dirichlet version of the mixed-membership model then suitable discretization would be necessary.

[10] In more general versions of the blockmodel studied in [AS16a], where each pair i, j of communities may have a different edge probability Q_{ij} it is not always possible to estimate all of $y_1, \ldots, y_k$. We view it as an interesting open problem to extract as much information about $y_1, \ldots, y_k$ as possible in that setting; the guarantee of [AS16a] amounts, roughly, to finidng a single vector in the linear span of $y_1, \ldots, y_k$.

$\mathbb{E}\langle \hat{y}_i, y_i\rangle^2 \geqslant \frac{1}{k^{O(1)}}\|y_i\|$, recovering the guarantees of the theorems above but not offering an estimator $\hat{y}_i$ whose distance to y_i depends only on the distance δ above the critical threshold. Indeed, without exploiting additional structure of the vectors y_i is unclear how to remove this $1/k^{O(1)}$ factor. As a thought experiment, if one had the matrix $M' = \sum_{i\leqslant k} a_i a_i^\mathsf{T}$, where $a_1, \dots, a_k$ were random unit vectors, then $a_1, \dots, a_k$ would be nearly orthonormal and one could learn essentially only their linear span. (From the linear span it is only possible to find $\hat{a}_i$ with correlation $\langle \hat{a}_i, a_i\rangle^2 \geqslant 1/k^{O(1)}$.)

In the interest of generality we would like to avoid using such additional structure: while in the disjoint-community model the vectors y_i have disjoint support (after un-centering them), no such special structure is evident in the mixed-membership setting. Indeed, when α is comparable to k, the vectors y_i are similar to independent random vectors of the appropriate norm.

To address this issue we turn to tensor methods. To illustrate the main idea simply: if $a_1, \dots, a_k$ are orthonormal, then it is possible to recover $a_1, \dots, a_k$ exactly from the 3-tensor $T = \sum_{i\leqslant k} a_i^{\otimes 3}$. More abstractly, the meta-algorithm which uses 3rd moments is able to estimate hidden variables whose posterior distributions have a high degree of symmetry, without errors which worsen as the posteriors become more symmetric.

Theorem 1.3 (Bayesian estimation meta-theorem—3rd moment)**.** *Let $p(x, y_1, \dots, y_k)$ be a joint distribution over vectors $x \in \{0,1\}^n$ and exchangable,[11] orthonormal[12] vectors $y_1, \dots, y_k \in \mathbb{R}^d$. Assume the marginal distribution of x satisfies $p(x) \geqslant 2^{-n^{O(1)}}$ for all $x \in \{0,1\}^n$.[13] Suppose there exists a tensor-valued degree-D polynomial $P(x)$ such that*

$$\mathop{\mathbb{E}}_{p(x,y_1,\dots,y_k)} \langle P(x), \sum_{i=1}^k y_i^{\otimes 3}\rangle \geqslant \delta \cdot \left(\mathop{\mathbb{E}}_{p(x)} \|P(x)\|^2\right)^{1/2} \cdot \sqrt{k}\,. \tag{1.3}$$

(Here, $\|\cdot\|$ is the norm induced by the inner product $\langle\cdot,\cdot\rangle$. The factor $\sqrt{k}$ normalizes the inequality because $\|\sum_{i=1}^k y_i^{\otimes 3}\| = \sqrt{k}$ by orthonormality.) Then, there exists $\delta' \geqslant \delta^{O(1)} > 0$ and a circuit of size $n^{O(D)}$ that given $x \in \{0,1\}^n$ outputs a list of unit vectors $z_1, \dots, z_m$ with $m \leqslant n^{\mathrm{poly}(1/\delta)}$ so that

$$\mathop{\mathbb{E}}_{p(x,y_1,\dots,y_k)} \mathop{\mathbb{E}}_{i\sim[k]} \max_{j\in[m]} \langle y_i, z_j\rangle^2 \geqslant \delta'\,. \tag{1.4}$$

That the meta-algorithm captured by this theorem outputs a list of $n^{1/\mathrm{poly}(\delta)}$ vectors rather than just k vectors is an artifact of the algorithmic difficulty of multilinear algebra as compared to linear algebra. However, in most Bayesian estimation problems it is possible by using a very small number of additional samples (amounting to a low-order additive term in the total sample complexity) to cross-validate the vectors in the list $z_1, \dots, z_m$ and throw out those which are not correlated with some $y_1, \dots, y_k$. Our eventual algorithm for tensor decomposition (see Section 1.3 and Section 7) bakes this step in by assuming access to an oracle which evaluates the function $v \mapsto \sum_{i\in[k]} \langle v, y_i\rangle^4$.

[11] Here, exchangeable means that for every $x \in \{0,1\}^n$ and every permutation $\pi\colon [k] \to [k]$, we have $p(y_1, \dots, y_k \mid x) = p(y_{\pi(1)}, \dots, y_{\pi(k)} \mid x)$.

[12] Here, we say the vector-valued random variables $y_1, \dots, y_k$ are orthonormal if with probability 1 over the distribution p we have $\langle y_i, y_j\rangle = 0$ for all $i \neq j$ and $\|y_i\|^2 = 1$.

[13] As in the previous theorem, this mild condition on the marginal distribution of x allows us to rule out pathological situations where a low-degree polynomial in x may be hard to evaluate accurately enough because of coefficients with super-polynomial bit-complexity.

A key component of the algorithm underlying Theorem 1.3 is a new algorithm for very robust orthogonal tensor decomposition.[14] Previous algorithms for tensor decomposition require that the input tensor is close (in an appropriate norm) to only one orthogonal tensor. By contrast, our tensor decomposition algorithm is able to operate on a tensor T which is just $\delta \ll 1$ correlated to the orthogonal tensor $\sum y_i^{\otimes 3}$, and in particular might also be δ-correlated with $1/\delta$ other orthogonal tensors. If one views tensor decomposition as a *decoding* task, taking a tensor T and decoding it into its rank-one components, then our guarantees are analogous to list-decoding. Our algorithm in this setting involves a novel entropy-maximization program which, among other things, ensures that given a tensor T which for example is δ-correlated with two distinct orthogonal tensors A and B, the algorithm produces a list of vectors correlated with both the components of A and those of B.

Applying this meta-theorem (plus a simple cross-validation scheme to prune the vectors in the $n^{1/\mathrm{poly}(\delta)}$-length list) to the mixed-membership block model (and its special case, the k-disjoint-communities block model) yields the following theorem. (See Section 1.2 for formal statements.)

Theorem 1.4 (Main theorem on the mixed-membership block model, informal)**.** *Let $\varepsilon, d, k, \alpha$ be paramters of the mixed-membership block model, and let $\delta = 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} \geqslant \Omega(1)$. Let y_i be the centered indicator vector of the i-th community. There is an $n^{1/\mathrm{poly}(\delta)}$-time algorithm which, given a sample x from the $\varepsilon, d, k, \alpha$ block model, recovers vectors $\hat{y}_1(x), \dots, \hat{y}_k(x)$ such that there is a permutation $\pi : [k] \to [k]$ with*

$$\mathbb{E}\langle \hat{y}_{\pi(i)}, y_i \rangle^2 \geqslant \delta^{O(1)} (\mathbb{E} \|\hat{y}_{\pi(i)}\|^2)^{1/2} (\mathbb{E} \|y_i\|^2)^{1/2} .$$

The eventual goal, as we discuss in Section 1.2, is to label each vertex by a probability vector τ_i which is correlated with the underlying label σ_i, but given the $\hat{y}$ vectors from this theorem this is easily accomplished.

Comparison to the sum of squares method The sum of squares method has been recently been a popular approach for designing algorithms for Bayesian estimation [BKS15, HSS15, RRS16, GM15]. The technique works best in settings where the maximum-likelihood estimator can be phrased as a polynomial optimization problem (subject to semialgebraic constraints). Then the strategy is to use the sum of squares method to obtain a strong convex relaxation of the maximum-likelihood problem, solve this relaxation, and round the result.

This strategy has been quite successful, but thus far it does not seem to allow the sharp (up to low-order additive terms) sample-complexity guarantees we study here. (Indeed, for some problems, including the stochastic block model, it is not clear that maximum likelihood estimation recovers those guarantees, much less the SoS-relaxed version.)

One similarity between our algorithms and these applications of sum of squares is that the rounding procedures used at the end often involve tensor decomposition, which is itself often done via the sum of squares method. We do employ the SoS algorithm as a black box to solve tensor decomposition problems for versions of our algorithm which use higher moments.

[14] An orthogonal 3-tensor is $\sum_{i=1}^m a_i^{\otimes 3}$, where $a_1, \dots, a_m$ are orthonormal.

Recent works on SoS show that the low-degree polynomials computed by our meta-algorithm are closely connected to *lower bounds* for the SoS hierarchy, though this connection remains far from fully understood. The recent result [BHK+16] on the planted clique problem first discovered this connection. The work [HKP+17] (written concurrently with the present paper) shows that this connection extends far beyond the planted clique setting.

Comparison to the method of moments Another approach for designing statistical estimators for provable guarantees is the method of moments. Typically one considers parameters θ (which need not have a prior distribution $p(\theta)$) and iid samples $x_1, \dots, x_n \sim p(x|\theta)$. Generally one shows that the moments of the distribution $\{x|\theta\}$ are related to some function of θ: for example perhaps $\mathbb{E}[xx^\top \mid \theta] = f(\theta)$. Then one uses the samples x_i to estimate the moment $M = \mathbb{E}[xx^\top \mid \theta]$, and finally to estimate θ by $f^{-1}(M)$.

While the method of moments is quite flexible, for the high-noise problems we consider here it is not clear that it can achieve optimal sample complexity. For example, in our algorithms (and existing sample-optimal algorithms for the block model) it is important to exploit the flexibility to compute any polynomial of the samples jointly—given n samples our algorithms can evaluate a polynomial $P(x_1, \dots, x_n)$, and P often will not be an empirical average of some simpler function like $\sum_{i \leqslant n} q(x_i)$. The best algorithm for the mixed-membership block model before our work uses the method of moments and consequently requires much denser graphs than our method [AGHK14].

1.2 Detecting overlapping communities

We turn now to discuss our results for stochastic block models in more detail and compare them to the existing literature.

The stochastic block model is a widely studied (family of) model(s) of random graphs containing latent community structure. It is most common to study the block model in the sparse graph setting: many large real-world networks are sparse, and the sparse graph setting is nearly always more mathematically challenging than the dense setting. A series of recent works has for the first time obtained algorithms which recover communities in block model graphs under (conjecturally) optimal sparsity conditions. For an excellent survey, see [Abb17].

Such sharp results remain limited to relatively simple versions of the block model; where, in particular, each vertex is assigned a single community in an iid fashion. A separate line of work has developed more sophisticated and realistic random graph models with latent community structure, with the goal of greater applicability to real-life networks. The mixed-membership stochastic block model [ABFX08] is one such natural extension of the stochastic block model that allows for communities to overlap, as they do in large networks found in the wild.

In addition to the number of vertices n, the average degree d, the correlation parameter ε, and the number of communities k, this model has an overlap parameter $\alpha \geqslant 0$ that controls how many communities a typical vertex participates in. Roughly speaking, the model generates an n-vertex graph by choosing k communities as random vertex subsets of size $(1+\alpha)n/k$ and choosing $dn/2$ random edges, favoring pairs of vertices that have many communities in common.

Definition 1.5 (Mixed-membership stochastic block model)**.** The mixed-membership stochastic block model SBM$(n, d, \varepsilon, k, \alpha)$ is the following distribution over n-vertex graphs G and k-dimensional probability vectors $\sigma_1, \ldots, \sigma_n$ for the vertices:

- draw $\sigma_1, \ldots, \sigma_n$ independently from Dir(α) the symmetric k-dimensional Dirichlet distribution with parameter $\alpha \geqslant 0$,[15]
- for every potential edge $\{i, j\}$, add it to G with probability $\frac{d}{n} \cdot \left(1 + \left(\langle \sigma_i, \sigma_j \rangle - \frac{1}{k}\right)\varepsilon\right)$.

Due to symmetry, $\langle \sigma_i, \sigma_j \rangle$ has expected value $\frac{1}{k}$, which means that the expected degree of every vertex in this graph is d. In the limit $\alpha \to 0$, the Dirichlet distribution is equivalent to the uniform distribution over coordinate vectors $\mathbf{1}_1, \ldots, \mathbf{1}_k$ and the model becomes SBM(n, d, ε, k), the stochastic block model with k *disjoint* communities. For $\alpha = k$, the Dirichlet distribution is uniform over the open $(k-1)$-simplex [Wik17b]. For general values of α, a probability vector from Dir(α) turns out to have expected collision probability $(1 - \frac{1}{k})\frac{1}{\alpha+1} + \frac{1}{k}$, which means that we can think of the probability vector being concentrated on about $\alpha + 1$ coordinates.[16] This property of the Dirichlet distribution is what determines the threshold for our algorithm. Correspondingly, our algorithm and analysis extends to a large class of distributions over probability vectors that share this property.

Measuring correlation with community structures In the constant-average-degree regime of the block model, recovering the label of every vertex correctly is information-theoretically impossible. For example, no information is present in a typical sample about the label of any isolated vertex, and in a typical sample a constant fraction of the vertices are isolated. Instead, at least in the k-disjoint-community setting, normally one looks to label vertices by labels $1, \ldots, k$ so that (up to a global permutation), this labeling has positive correlation with the true community labels.

When the communities are disjoint, one can measure such correlation using the sizes of $|S_j \cap \widehat{S}_j|$, where $S_j \subseteq [n]$ is the set of nodes in community j and $\widehat{S}_j$ is an estimated set of nodes in community j. The original definition of *overlap*, the typical measure of labeling-accuracy in the constant-degree regime, takes this approach [DKMZ11].

For present purposes this definition must be somewhat adapted, since in the mixed-membership block model there is no longer a good notion of a discrete set of nodes S_j for each community $j \in [k]$. We define a smoother notion of correlation with underlying community labels to accommodate that the labels σ_i are vectors in Δ_{k-1}. In discrete settings, for example when $\alpha \to 0$ (in which case one recovers the k-disjoint-community model), or more generally when each σ_i is the uniform distribution over some number of communities, our correlation measure recovers the usual notion of overlap.

Let $\sigma = (\sigma_1, \ldots, \sigma_n)$ and $\tau = (\tau_1, \ldots, \tau_n)$ be labelings of the vertices $1, \ldots, n$ by by k-dimensional

[15] In the symmetric k-dimensional Dirichlet distribution with parameter $\alpha > 0$, the probability of a probability vector σ is proportional to $\prod_{t=1}^{k} \sigma(t)^{\alpha/k-1}$. By passing to the limit, we define Dir(0) to be the uniform distribution over the coordinate vectors $\mathbf{1}_1, \ldots, \mathbf{1}_k$.

[16] When k and α are comparable in magnitude, it is important to interpret this more accurately as $(\alpha + 1) \cdot \frac{k}{k+\alpha}$ coordinates.

probability vectors. We define the *correlation* $\mathrm{corr}(\sigma, \tau)$ as

$$\max_{\pi} \mathop{\mathbb{E}}_{i\sim n} \langle \sigma_i, \tau_{\pi(i)} \rangle - \frac{1}{k} \tag{1.5}$$

where π ranges over permutations of the k underlying communities. This notion of correlation is closely related to the *overlap* of the distributions σ_i, τ_i.

To illustrate this notion of correlation, consider the case of disjoint communities (i.e., $\alpha = 0$), where the ground-truth labels τ_i are indicator vectors in k dimensions. Then, if $\mathbb{E}_i \langle \sigma_i, \tau_{\pi(i)} \rangle - \frac{1}{k} > \delta$, by looking at the large coordinates of σ_i it is possible to correctly identify the community memberships of a $\delta^{O(1)} + \frac{1}{k}$ fraction of the vertices, which is a $\delta^{O(1)}$ fraction more than would be identified by randomly assigning labels to the vertices without looking at the graph.

When the ground truth labels τ_i are spread over more than one coordinate—say, for example, they are uniform over t coordinates—the best recovery algorithm cannot find σ's with correlation better than

$$\mathrm{corr}(\sigma, \tau) = \frac{1}{t} - \frac{1}{k}\,,$$

which is achieved by $\sigma = \tau$. This is because in this case τ has collision probability $\langle \tau, \tau \rangle = \frac{1}{t}$.

Main result for mixed-membership models The following theorem gives a precise bound on the number of edges that allows us to find in polynomial time a labeling of the vertices of an n-node mixed membership block model having nontrivial correlation with the true underlying labels. Here, the parameters $d, \varepsilon, k, \alpha$ of the mixed-membership stochastic block model may even depend on the number of vertices n.

Theorem 1.6 (Mixed-membership SBM—significant correlation)**.** *Let $d, \varepsilon, k, \alpha$ be such that $k \leqslant n^{o(1)}$, $\alpha \leqslant n^{o(1)}$, and $\varepsilon^2 d \leqslant n^{o(1)}$. Suppose $\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0$. (Equivalently for small δ, suppose $\varepsilon^2 d \geqslant (1+\delta) \cdot k^2(\alpha+1)^2$.) Then, there exists $\delta' \geqslant \delta^{O(1)} > 0$ and an $n^{1/\mathrm{poly}(\delta)}$-time algorithm that given an n-vertex graph G outputs $\tau_1, \ldots, \tau_n \in \Delta_{k-1}$ such that*

$$\mathop{\mathbb{E}}_{(G,\sigma)\sim \mathrm{SBM}(n,d,\varepsilon,k,\alpha)} \mathrm{corr}(\sigma, \tau) \geqslant \delta' \cdot \left(\frac{1}{t} - \frac{1}{k} \right) \tag{1.6}$$

where $t = (\alpha+1) \cdot \frac{k}{k+\alpha}$ (samples from the α, k Dirichlet distribution are roughly uniform over t out of k coordinates). In particular, as $\delta \to 1$ we have $\mathbb{E}\,\mathrm{corr}(\sigma, \tau) \to \frac{1}{t} - \frac{1}{k}$, while $\mathbb{E}\,\mathrm{corr}(\sigma, \sigma) = \frac{1}{t} - \frac{1}{k}$.

Note that in the above theorem, the correlation δ' that our algorithm achieves depends only on δ (the distance to the threshold) and in particular is independent of n and k (aside from, for the latter, the dependence on k via δ). For disjoint communities ($\alpha = 0$), our algorithm achieves constant correlation with the planted labeling if $\varepsilon^2 d / k^2$ is bounded away from 1 from below.

We conjecture that the threshold achieved by our algorithm is best-possible for polynomial-time algorithms. Concretely, if $d, \varepsilon, k, \alpha$ are constants such that $\varepsilon^2 d < k^2(\alpha+1)^2$, then we conjecture that for every polynomial-time algorithm that given a graph G outputs $\tau_1, \ldots, \tau_n \in \Delta_{k-1}$,

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{(G,\sigma)\sim \mathrm{SBM}(n,d,\varepsilon,k,\alpha)} \mathrm{corr}(\sigma, \tau) = 0\,. \tag{1.7}$$

This conjecture is a natural extension of a conjecture for disjoint communities ($\alpha = 0$), which says that beyond the Kesten–Stigum threshold, i.e., $\varepsilon^2 d < k^2$, no polynomial-time algorithm can achieve correlation bounded away from 0 with the true labeling [Moo17]. For large enough values of k, this conjecture predicts a computation-information gap because the condition $\varepsilon^2 d \geqslant \Omega(k \log k)$ is enough for achieving constant correlation information-theoretically (and in fact by a simple exponential-time algorithm). We discuss these ideas further in Section 1.4.

Comparison with previous matrix-based algorithms We offer a reinterpretation in our meta-algorithmic framework of the algorithms of Mossel-Neeman-Sly and Abbe-Sandon. This will permit us to compare our algorithm for the mixed-membership model with what could be achieved by the methods in these prior works, and to point out one respect in which our algorithm improves on previous ones even for the disjoint-communities block model. The result we discuss here is a slightly generalized version of Theorem 1.2.

Let $\mathcal{U}$ be a (possibly infinite or continuous) universe of labels, and let W assign to every $x, y \in \mathcal{U}$ a nonnegative real number $W(x, y) = W(y, x) \geqslant 0$. Let μ be a probability distribution on $\mathcal{U}$, which induces the inner product of functions $f, g : \mathcal{U} \to \mathbb{R}$ given by $\langle f, g \rangle = \mathbb{E}_{x \sim \mu} f(x) g(x)$. The function W can be considered as linear operator on $\{f : \mathcal{U} \to \mathbb{R}\}$, and under mild assumptions it has eigenvalues $\lambda_1, \lambda_2, \ldots$ with respect to the inner product $\langle \cdot, \cdot \rangle$.

The pair μ, W along with an average degree parameter d induce a generalized stochastic block model, where labels for nodes are drawn from μ and an edge between a pair of nodes with labels x and y is present with probability $\frac{d}{n} \cdot W(x, y)$. When $\mathcal{U}$ is Δ_{k-1} and μ is the Dirichlet distribution, this captures the mixed-membership block model.

Assume $\lambda_1 = 1$ and that μ and W are sufficiently *nice* (see Section 4 for all the details). Then one can rephrase results of Abbe and Sandon in this setting as follows.

Theorem 1.7 (Implicit in [AS16a])**.** *Suppose the operator W has eigenvalues $1 = \lambda_1 > \lambda_2 > \cdots > \lambda_r$ (each possibly with higher multiplicity) and $\delta \stackrel{\text{def}}{=} 1 - \frac{1}{d\lambda_2^2} > 0$. Let Π be the projector to the second eigenspace of the operator W. For types $x_1, \ldots, x_n \sim \mathcal{U}$, let $A \in \mathbb{R}^{n \times n}$ be the random matrix $A_{ij} = \Pi(x_i, x_j)$, where we abuse notation and think of $\Pi \colon \mathcal{U} \times \mathcal{U} \to \mathbb{R}$. There is an algorithm with running time $n^{\mathrm{poly}(1/\delta)}$ which outputs an $n \times n$ matrix P such that for $x, G \sim G(n, d, W, \mu)$,*

$$\mathop{\mathbb{E}}_{x,G} \operatorname{Tr} P \cdot A \geqslant \delta^{O(1)} \cdot (\mathop{\mathbb{E}}_{x,G} \|A\|^2)^{1/2} (\mathop{\mathbb{E}}_{x,G} \|P\|^2)^{1/2}.$$

In one way or another, existing algorithms for the block model in the constant-degree regime are all based on estimating the random matrix A from the above theorem, then extracting from an estimator for A some labeling of vertices by communities. In our mixed-membership setting, one may show that the matrix A is $\sum_{s \in [k]} v_s v_s^\mathsf{T}$, where $v_s \in \mathbb{R}^n$ has entries $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Furthermore, as we show in Section 4, the condition $d\lambda_2^2 > 1$ translates for the mixed-membership model to the condition $\varepsilon^2 d > k(\alpha + 1)^2$, which means that under the same hypotheses as our main theorem on the mixed-membership model it is possible in polynomial time to evaluate a constant-correlation estimator for $\sum_{s \in [k]} v_s v_s^\mathsf{T}$. As we discussed in Section 1.1, however, extracting estimates for $v_1, \ldots, v_k$ (or, almost equivalently, estimates for $\sigma_1, \ldots, \sigma_n$) from this matrix seems

to incur an inherent $1/k$ loss in the correlation. Thus, the final guarantee one could obtain for the mixed-membership block model using the techniques in previous work would be estimates $\tau_1, \ldots, \tau_n$ for $\sigma_1, \ldots, \sigma_n$ such that $\mathrm{corr}(\sigma, \tau) \geqslant \left(\frac{\delta}{k}\right)^{O(1)}$.[17] We avoid this loss in our main theorem via tensor methods.

Although this $1/k$ multiplicative loss in the correlation with the underlying labeling is not inherent in the disjoint-community setting (roughly speaking this is because the matrix A is a 0/1 block-diagonal matrix), previous algorithms nonetheless incur such loss. (In part this is related to the generality of the work of Abbe and Sandon: they aim to allow W where A might only have rank one, while in our settings A always has rank $k-1$. For low-rank A this $1/k$ loss is probably necessary for polynomial time algorithms.)

Thus our main theorem on the mixed membership model offers an improvement on the guarantees in the previous literature even for the disjoint-communities setting: when W only has entries $1-\varepsilon$ and ε we obtain a labeling of the vertices whose correlation with the underlying labeling depends only on δ. This allows the number k of communities to grow with n without incurring any loss in the correlation (so long as the average degree of the graph grows accordingly).

For further discussion of the these results and a proof of the above theorem, see Section 4.

Comparison to previous tensor algorithm for mixed-membership models Above we discussed a reinterpretation (allowing a continuous space $\mathcal{U}$ of labels) of existing algorithms for the constant-average-degree block model which would give an algorithm for the mixed-membership model, and discussed the advantages of our algorithm over this one. Now we turn to algorithms in the literature which are specifically designed for stochastic block models with overlapping communities.

The best such algorithm requires $\varepsilon^2 d \geqslant O(\log n)^{O(1)} \cdot k^2(\alpha+1)^2$ [AGHK13]. Our bound saves the $O(\log n)^{O(1)}$ factor. (This situation is analogous to the standard block model, where simpler algorithms based on eigenvectors of the adjacency matrix require the graph degree to be logarithmic.) Notably, like ours this algorithm is based on estimating the tensor $T = \sum_{s\in[k]} v_s^{\otimes 3}$, where $v_s \in \mathbb{R}^n$ has entries $v_s(i) = \sigma_i(s) - \frac{1}{k}$. However, the algorithm differs from ours in two key respects.

1. The algorithm [AGHK13] estimates the tensor T using a 3-tensor analogue of a high power of the adjacency matrix of an input graph, while we use self-avoiding walks (which are rather like a tensor analogue of the nonbacktracking operator).
2. The tensor decomposition algorithm used in [AGHK13] to decompose the (estimator for the) tensor T tolerates much less error than our tensor decomposition algorithm; the result is that a higher-degree graph is needed in order to obtain a better estimator for the tensor T.

The setting considered by [AGHK13] does allow a more sophisticated version of the Dirichlet distribution than we allow, in which different communities have different sizes. It is an interesting open problem to extend the guarantees of our algorithm to that setting.

[17] In fact, it is not clear one can obtain even this guarantee using strictly matrix methods. Strictly speaking, in estimating, say, v_1 using the above described matrix method, one obtains a unit vector v such that $\langle v, v_1\rangle^2 \geqslant \Omega(1)\cdot\|v_1\|^2$. Without knowing whether v or $-v$ is the correct vector it is not clear how to transform estimates for the v_s's to estimates for the σ's. However, matrix methods cannot distinguish between v_s and $-v_s$. In our main algorithm we avoid this issue because the 3rd moments $\sum v_s^{\otimes 3}$ are not sign-invariant.

1.3 Low-correlation tensor decomposition

Tensor decomposition is the following problem. For some unit vectors $a_1, \dots, a_m \in \mathbb{R}^n$ and a constant k (often $k = 3$ or 4), one is given the tensor $T = \sum_{i=1}^m a_i^{\otimes k} + E$, where E is some error tensor. The goal is to recover vectors $b_1, \dots, b_m \in \mathbb{R}^n$ which are as close as possible to $a_1, \dots, a_m$.

Tensor decomposition has become a common primitive used by algorithms for parameter learning and estimation problems [CJ10, AGH+14, GHK, GVX14, BKS15, MSS16, SS17]. In the simplest examples, the hidden variables are orthogonal vectors $a_1, \dots, a_m$ and there is a simple function of the observed variables which estimates the tensor $\sum_{i \leqslant m} a_i^{\otimes k}$ (often an empirical k-th moment of observed variables suffices). Applying a tensor decomposition algorithm to such an estimate yields estimates of the vectors $a_1, \dots, a_m$.

We focus on the case that $a_1, \dots, a_m$ are orthonormal. Algorithms for this case are already useful for a variety of learning problems, and it is often possible to reduce more complicated problems to the orthonormal case using a small amount of side information about $a_1, \dots, a_m$ (in particular their covariance $\sum_{i=1}^m a_i a_i^\mathsf{T}$). In this setting the critical question is: how much error E (and measured in what way) can the tensor decomposition algorithm tolerate and still produce useful outputs $b_1, \dots, b_m$?

When we use tensor decomposition in our meta-algorithm, the error E will be incurred when estimating $\sum_{i=1}^m a_i^{\otimes k}$ from observable samples. Using more samples would decrease the magnitude of $T - \sum_{i=1}^m a_i^{\otimes k}$, but because our goal is to obtain algorithms with optimal sample complexity we need a tensor decomposition algorithm which is robust to greater errors than those in the existing literature.

Our main theorem on tensor decomposition is the following.

Theorem 1.8 (Informal). *For every $\delta > 0$, there is a randomized algorithm with running time $n^{1/\mathrm{poly}(\delta)}$ that given a 3-tensor $T \in (\mathbb{R}^n)^{\otimes 3}$ and a parameter k outputs $n^{\mathrm{poly}(1/\delta)}$ unit vectors $b_1, \dots, b_m$ with the following property: if T satisfies $\langle T, \sum_{i=1}^k a_i^{\otimes 3} \rangle \geqslant \delta \cdot \|T\| \cdot \sqrt{k}$ for some orthonormal vectors $a_1, \dots, a_k$, then*

$$\mathop{\mathbb{E}}_{i \sim [k]} \max_{j \in [m]} \langle a_i, b_j \rangle^2 \geqslant \delta^{O(1)} .$$

Furthermore, if the algorithm is allowed to make $n^{1/\mathrm{poly}(\delta)}$ calls to an oracle $\mathcal{O}$ which correctly answers queries of the form "does the unit vector v satisfy $\sum_{i=1}^m \langle a_i, v \rangle^4 \geqslant \delta^{O(1)}$?", then it outputs just k orthonormal vectors, $b_1, \dots, b_k$ such that there is a permutation $\pi : [k] \to [k]$ with

$$\mathop{\mathbb{E}}_{i \in [k]} \langle a_i, b_{\pi(i)} \rangle^2 \geqslant \delta^{O(1)} .$$

(These guarantees hold in expectation over the randomness used in the decomposition algorithm.)

(For a more formal statement, and in particular the formal requirements for the oracle $\mathcal{O}$, see Section 7.)

Rescaling T as necessary, one may reinterpret the condition $\langle T, \sum_{i=1}^k a_i^{\otimes 3} \rangle \geqslant \delta \cdot \|T\| \cdot \sqrt{k}$ as $T = \sum_{i=1}^k a_i^{\otimes 3} + E$, where $\langle E, \sum_{i=1}^m a_i^{\otimes 3} \rangle = 0$ and $\|E\| \leqslant \sqrt{k}/\delta$ and $\|\cdot\|$ is the Euclidean norm. In particular, E may have Euclidean norm which is a large constant factor $1/\delta$ larger than the Euclidean

norm of the tensor $\sum_{i=1}^m a_i^{\otimes 3}$ that the algorithm is trying to decompose! (One way such error could arise is if T is actually correlated with $1/\delta$ unrelated orthogonal tensors; our algorithm in that case ensures that the list of outputs vectors is correlated with every one of these orthogonal tensors.)

In all previous algorithms of which we are aware (even for the case of orthogonal $a_1, \ldots, a_m$), the error E must have spectral norm (after flattening to an $n^2 \times n^2$ matrix) at most ε for $\varepsilon < \frac{1}{2}$,[18] or E must have Euclidean norm at most $\varepsilon\sqrt{m}$ [SS17]. The second requirement is strictly stronger than ours (thus our algorithm has weaker requirements and so stronger guarantees). The first, on the spectral norm of E when flattened to a matrix, is incomparable to the condition in our theorem. However, when E satisfies such a spectral bound it is possible to decompose T using (sophisticated) spectral methods [MSS16, SS17]. In our setting such methods seem unable to avoid producing only vectors b which are correlated with E but not with any $a_1, \ldots, a_m$. In other words, such methods would *overfit to the error* E. To avoid this, our algorithm uses a novel maximum-entropy convex program (see Section 7 for details).

One a priori unusual requirement of our tensor decomposition algorithm is access to the oracle $\mathcal{O}$. In any tensor decomposition setting where E satisfies $\|E\|_{inj} = \max_{\|x\|=1} \langle E, x^{\otimes 3}\rangle \leqslant o(1)$, the oracle $\mathcal{O}$ can be implemented just be evaluating $\langle T, v^{\otimes 3}\rangle = \sum_{i=1}^k \langle a_i, v\rangle^3 + o(1)$. All previous works on tensor decomposition of which we are aware either assume that the injective norm $\|E\|_{inj}$ is bounded as above, or (as in [SS17]) can accomplish this with a small amount of preprocessing on the tensor T. Our setting allows, for example, $E = 100 \cdot v^{\otimes 3}$ for some unit vector v, and does not appear to admit the possibility of such preprocessing, hence the need for an auxiliary implementation of $\mathcal{O}$. In our learning applications we are able to implement $\mathcal{O}$ by straightforward holdout set/cross-validation methods.

1.4 Information-computation gaps and concrete lower bounds

The meta-algorithm we offer in this paper is designed to achieve optimal sample complexity *among polynomial-time algorithms* for many Bayesian estimation problems. It is common, though, that computationally inefficient algorithms can obtain accurate estimates of hidden variables with fewer samples than seem to be tolerated by any polynomial-time algorithm. This appears to be true for the k-community stochastic block model we have used as our running example here: efficient algorithms seem to require graphs of average degree $d > k^2/\varepsilon^2$ to estimate communities, while inefficient algorithms are known which tolerate d of order $k \log k$ [AS16b, Moo17].

Such phenomena, sometimes called *information-computation gaps*, appear in many other Bayesian estimation problems. For example, in the classical *planted clique* problem [Jer92, Kuc95], a clique of size $k > (2+\varepsilon)\log n$ is randomly added to a sample $G \sim G(n, \frac{1}{2})$; the goal is to find the clique given the graph. Since the largest clique in $G(n, \frac{1}{2})$ has expected size $2\log n$, so long as $k > (2+\varepsilon)\log n$ it is information-theoretically possible, via brute-force search, to recover the planted clique. On the other hand, despite substantial effort, no polynomial-time algorithm is known which can find a clique of size $k \leqslant o(\sqrt{n})$, exponentially-larger than cliques which can be found by brute force.

For other examples one need not look far: sparse principal component analysis, planted

[18] Or, mildly more generally, E should have SoS norm less than ε [MSS16].

constraint satisfaction problems, and densest-k-subgraph are just a few more problems exhibiting information-computation gaps. This ubiquity leads to several questions:

1. What rigorous evidence can be provided for the claim: no polynomial algorithm tolerates $n < n_*$ samples and produces a constant-correlation estimator $\widehat{\theta}(x_1, \ldots, x_n)$ for particular a Bayesian estimation problem $p(x, \theta)$?
2. Can information-computation gaps of different problems be explained by similar underlying phenomena? That is, is there a structural feature of a Bayesian estimation problem which determines whether it exhibits an information-computation gap, and if so, what is the critical number of samples n_* required by polynomial-time algorithms?
3. Are there methods to easily predict the location of a critical number n_* of samples, without analyzing every polynomial-time algorithm one can think of?

Rigorous evidence for computational phase transitions The average-case nature of Bayesian estimation problems makes it unlikely that classical tools like (NP-hardness) reductions allow us to reason about the computational difficulty of such problems in the too-few-samples regime. Instead, to establish hardness of an estimation problem when $n < n_*$ for some critical n_*, one proves impossibility results for restricted classes of algorithms. Popular classes of algorithms for this purpose include Markov-Chain Monte Carlo algorithms and various classes of convex programs, especially arising from convex hierarchies such as the Sherali-Adams linear programming hierarchy and the sum of squares semidefinite programming hierarchy.

Results like this are meaningful only if the class of algorithms for which one proves an impossibility result captures the best known (i.e. lowest-sample-complexity) polynomial-time algorithm for the problem at hand. Better yet would be to use a class of algorithms which captures the lowest-sample-complexity polynomial-time algorithms for many Bayesian estimation problem simultaneously.

In the present work we study sample complexity up to low-order additive terms in the number n of samples. For example, in the k-community α-mixed-membership stochastic block model, we provide an algorithm which estimates communities in graphs of average degree $d \geqslant (1+\delta)k^2(\alpha+1)^2/\varepsilon^2$, for any constant $\delta > 0$. Such precise algorithmic guarantees suggest the pursuit of equally-precise lower bounds.

Proving a lower bound against the convex-programming-based algorithms most commonly considered in previous work on lower bounds for Bayesian estimation problems does not suit this purpose. While powerful, these algorithms are generally not designed to achieve optimal sample complexity up to low-order additive terms. Indeed, there is mounting evidence in the block model setting that sum of squares semidefinite programs actually require a constant multiplicative factor more samples than our meta-algorithm [MS16, BKM17].

Another approach to providing rigorous evidence for the impossibility side of a computational threshold is *statistical query* complexity, used to prove lower bounds against the class of *statistical query* algorithms [FGR+13, FGV15, FPV15]. To the best of our knowledge, similar to sum of squares algorithms, statistical query algorithms are not known to achieve optimal sample rates for problems such as community recovery in the block model up to low-order additive terms.

Such sample-optimal algorithms seem to intrinsically require the ability to compute a complicated function of many samples simultaneously (for example, the top eigenvector of the non-backtracking operator). But even the most powerful statistical query algorithms (using the 1-STAT oracle) can access one bit of information about each sample x (by learning the value of some function $h(x) \in \{0,1\}$). This makes it unclear how the class of statistical query algorithms can capture the kinds of sample-optimal algorithms we want to prove lower bounds against.

Our meta-algorithm offers an alternative. By showing that when n is less than a critical n_* there are no constant-correlation low-degree estimators for a hidden random variable, one rules out any efficient algorithm captured by our meta-algorithm. Concretely, Theorems 1.1 and 1.3 show that in order for an estimation problem to be intractable it is necessary that every low-degree polynomial fails to correlate with the second or third moment of the posterior distribution (in the sense of Eqs. (1.1) and (1.3)). This kind of fact about low-degree polynomial is something we can aim to prove unconditionally as a way to give evidence for the intractability of a Bayesian estimation problem. Next we discuss our example result of this form in the block model setting.

Concrete unconditional lower bound at the Kesten–Stigum threshold In this work, we show an unconditional lower bound about low-degree polynomials for the stochastic block model with k communities at the Kesten–Stigum threshold. For $k \geqslant 4$, this threshold is bounded away from the information-theoretic threshold [AS15]. In this way, our lower bounds gives evidence for an inherent gap between the information-theoretical and computational thresholds.

For technical reasons, our lower bound is for a notion of correlation mildly different from Eqs. (1.1) and (1.3). Our goal is to compare the stochastic block model distribution $\mathrm{SBM}(n, d, \varepsilon, k)$ graphs to the Erdős-Rényi distribution $G(n, \frac{d}{n})$ with respect to low-degree polynomials. As before we represent graphs as adjacency matrices $x \in \{0,1\}^{n \times n}$. Among all low-degree polynomials $p(x)$, we seek one so that the typical value of $p(x)$ for graphs x from the stochastic blocks model is as large as possible compared to its typical for Erdős-Rényi graphs. The following theorem shows that a suitable mathematical formalization of this question exhibits a sharp "phase transition" at the Kesten–Stigum threshold.

Theorem 1.9. *Let d, ε, k be constants. Then,*

$$\max_{p \in \mathbb{R}[x]_{\leqslant \ell}} \frac{\mathbb{E}_{x \sim \mathrm{SBM}(n,d,\varepsilon,k)} p(x)}{\left(\mathbb{E}_{x \sim G(n,d/n)} p(x)^2\right)^{1/2}} = \begin{cases} \geqslant n^{\Omega(1)} & \text{if } \varepsilon^2 d > k^2,\ \ell \geqslant O(\log n) \\ \leqslant n^{o(1)} & \text{if } \varepsilon^2 d < k^2,\ \ell \leqslant n^{o(1)} \end{cases} \tag{1.8}$$

Let $\mu \colon \{0,1\}^{n \times n} \to \mathbb{R}$ be the relative density of $\mathrm{SBM}(n, d, \varepsilon, k)$ with respect to $G(n, \frac{d}{n})$. Basic linear algebra shows that the left-hand side of Eq. (1.8) is equal to $\|\mu^{\leqslant \ell}\|_2$, where $\|\cdot\|_2$ is the Euclidean norm with respect to the measure $G(n, d/n)$ and $\mu^{\leqslant \ell}$ is the projection (with respect to this norm) of μ to the subspace of functions of degree at most ℓ. This is closely related to the χ^2-divergence of μ with respect to $G(n, d/n)$, which in the present notation would be given by $\|(\mu - 1)\|_2$. When the latter quantity is small, $\|(\mu - 1)\|_2 \leqslant o(1)$, one may conclude that the distribution μ is information-theoretically indistinguishable from $G(n, d/n)$. This technique is used in the best current bounds on the information-theoretic properties of the block model [MNS12, BMNN16].

The quantity in Theorem 1.9 is a low-degree analogue of the χ^2-divergence. If it were true that $\|(\mu^{\leqslant \ell} - 1)\|_2 \leqslant o(1)$, then by a straightforward application of Cauchy-Schwarz it would follow that no low-degree polynomial $p(x)$ distinguishes the block model from $G(n, d/n)$, since every such p would have (after setting $\mathbb{E}_{G(n,d/n)} p(x) = 0$) that $\mathbb{E}_{SBM} p(x) \leqslant o(\mathbb{E}_{G(n,d/n)} p(x)^2)^{1/2}$. This condition turns out to be quite powerful: [BHK+16, HKP+17] give evidence that for problems such as planted clique, for which distinguishing instances drawn from a null model from instances with hidden structure should be computationally intractable, the condition $\|(\mu^{\leqslant \ell} - 1)\| \leqslant o(1)$ is closely related to sum of squares lower bounds.[19]

The situation in the k-community block model is a bit more subtle. One has only that $\|\mu^{\leqslant \ell}\| \leqslant n^{o(1)}$ below the Kesten-Stigum threshold because even in the latter regime it remains possible to distinguish a block model graph from $G(n, d/n)$ via a low-degree polynomial (simply counting triangles will suffice). However, we can still hope to rule out algorithms which accurately estimate communities below the Kesten-Stigum threshold. For this we prove the following theorem.

Theorem 1.10. *Let $d, \varepsilon, k, \delta$ be constants such that $\varepsilon^2 d < (1-\delta)k^2$. Let $f : \{0,1\}^{n \times n} \to \mathbb{R}$ be any function, let $i, j \in [n]$ be distinct. Then if f satisfies $\mathbb{E}_{x \sim G(n, \frac{d}{n})} f(x) = 0$ and is correlated with the indicator $\mathbf{1}_{\sigma_i = \sigma_j}$ that i and j are in the same community in the following sense:*

$$\frac{\mathbb{E}_{x \sim SBM(n,d,\varepsilon,k)} f(x)(\mathbf{1}_{\sigma_i = \sigma_j} - \frac{1}{k})}{(\mathbb{E}_{x \sim G(n, \frac{d}{n})} f(x)^2)^{1/2}} \geqslant \Omega(1)$$

then $\deg f \geqslant n^{c(d,\varepsilon,k)}$ *for some* $c(d, \varepsilon, k) > 0$.

There is one subtle difference between the polynomials ruled out by this theorem and those which could be used by our meta-algorithm. Namely, this theorem rules out any f whose correlation with the indicator $\mathbf{1}_{\sigma_i = \sigma_j}$ is large *compared to f's standard deviation under* $G(n, d/n)$, whereas our meta-algorithm needs a polynomial f where this correlation is large compared to f's standard deviation under the block model. In implementing our meta-algorithm for the block model and for other problems, we have found that these two measures of standard deviation are always equal (up to low-order additive terms) for the polynomials which turn out to provide sample-optimal constant-correlation estimators of hidden variables.

Interesting open problems are to prove a version of the above theorem where standard deviation is measured according to the block model and to formalize the idea that $\mathbb{E}_{SBM} f(x)^2$ should be related to $\mathbb{E}_{G(n,d/n)} f(x)^2$ for good estimators f. It would also be quite interesting to see how large the function $c(d, \varepsilon, k)$ can be made: the above theorem shows that when $d < (1-\delta)k^2/\varepsilon^2$ the degree of any good estimator of $\mathbf{1}_{\sigma_i = \sigma_j}$ must be polynomial in n—perhaps it must be linear, or even quadratic in n.

General strategies to locate algorithmic thresholds The preceding theorems suggest a general strategy to locate critical a sample complexity n_* for almost any Bayesian estimation problem:

[19] In particular, the so-called *pseudocalibration* approach to sum of squares lower bounds works only when $\|(\mu^{\leqslant \ell} - 1)\| \leqslant o(1)$.

compute a Fourier transform of an appropriate relative density μ and examine the 2-norm of its low-degree projection. This strategy has several merits beyond its broad applicability. One advantage is that in showing $\|(\mu^{\geqslant \ell} - 1)\| \geqslant \Omega(1)$, one automatically has discovered a degree-ℓ polynomial and a proof that it distinguishes samples with hidden structure from an appropriate null model. Another is the mounting evidence (see [BHK+16, HKP+17]) that when, on the other hand $\|(\mu^{\leqslant \ell} - 1)\| \leqslant o(1)$ for large-enough ℓ, even very powerful convex programs cannot distinguish these cases. A final advantage is simplicity: generally computing $\|(\mu^{\geqslant \ell} - 1)\|$ is a simple exercise in Fourier analysis.

Finally, we compare this strategy to the only other one we know which shares its applicability across many Bayesian estimation problems, namely the replica and cavity methods (and their attendant algorithm, belief propagation) from statistical physics [MM09]. These methods were the first used to predict the sharp sample complexity thresholds we study here for the stochastic block model, and they have also been used to predict similar phenomena for many other hidden variable estimation problems [LBB+16, LBB+16, LBB+16]. Though remarkable, the predictions of these methods are much more difficult than ours make rigorous—in particular, it is notoriously challenging to rigorously analyze the belief propagation algorithm, and often when these predictions are made rigorous, only a modified version ("linearized BP") can be analyzed in the end. By contrast, our methods to predict critical sample complexities, design algorithms, and prove lower bounds all study essentially the same low-degree-polynomial algorithm.

We view it as a fascinating open problem to understand why predicted critical sample complexities offered by the replica and cavity methods are so often identical to the predictions of the low-degree-polynomials meta-algorithm we propose here.

2 Techniques

To illustrate the idea of low-degree estimators for posterior moments, let's first consider the most basic stochastic block model with $k = 2$ disjoint communities ($\alpha = 0$). (Our discussion will be similar to the analysis in [MNS15a].) Let $y \in \{\pm 1\}^n$ be chosen uniformly at random and let $x \in \{0,1\}^{n \times n}$ be the adjacency matrix of a graph such that for every pair $i < j \in [n]$, we have $x_{ij} = 1$ with probability $(1 + \varepsilon y_i y_j)\frac{d}{n}$. Our goal is to find a matrix-valued low-degree polynomial $P(x)$ that correlates with yy^T. It turns out to be sufficient to construct for every pair $i, j \in [n]$ a low-degree polynomial that correlates with $y_i y_j$.

The linear polynomial $p_{ij}(x) = \frac{n}{\varepsilon d}\left(x_{ij} - \frac{d}{n}\right)$ is an unbiased estimator for $y_i y_j$ in the sense that $\mathbb{E}[p_{ij}(x) \mid y] = y_i y_j$. By itself, this estimator is not particular useful because its variance $\mathbb{E}\, p_{ij}(x)^2 \approx \frac{n}{\varepsilon^2 d}$ is much larger than the quantity $y_i y_j$ we are trying to estimate. However, if we let $\alpha \subseteq [n]^2$ be a length-ℓ path between i and j (in the complete graph), then we can combine the unbiased estimators along the path α and obtain a polynomial

$$p_\alpha(x) = \prod_{ab \in \alpha} p_{ab}(x) \tag{2.1}$$

that is still an unbiased estimator $\mathbb{E}[p_\alpha(x) \mid y_i, y_j] = \mathbb{E}\left[\prod_{ab \in \alpha} y_a y_b \mid y_i, y_j\right] = y_i y_j$. This estimator

has much higher variance $\mathbb{E}\, p_\alpha(x)^2 \approx (\frac{n}{\varepsilon^2 d})^\ell$. But we can hope to reduce this variance by averaging over all such paths. The number of such paths is roughly $n^{\ell-1}$ (because there are $\ell-1$ intermediate vertices to choose). Hence, if these estimators $\{p_\alpha(x)\}_\alpha$ were pairwise independent, this averaging would reduce the variance by a multiplicative factor $n^{\ell-1}$, giving us a final variance of $(\frac{n}{\varepsilon^2 d})^\ell \cdot n^{1-\ell} = (\frac{1}{\varepsilon^2 d})^\ell \cdot n$. We can see that above the Kesten–Stigum threshold, i.e., $\varepsilon^2 d \geqslant 1+\delta$ for $\delta > 0$, this heuristic variance bound $(\frac{1}{\varepsilon^2 d})^\ell \cdot n \leqslant 1$ is good enough for estimating the quantity $y_i \cdot y_j$ for paths of length $\ell \geqslant \log_{1+\delta} n$.

Two steps remain to turn this heuristic argument into a polynomial-time algorithm for estimating the matrix yy^T. First, it turns out to be important to consider only paths that are self-avoiding. As we will see next, estimators from such paths are pairwise independent enough to make our heuristic variance bound go through. Second, a naive evaluation of the final polynomial takes quasi-polynomial time because it has logarithmic degree (and a quasi-polynomial number of non-zero coefficients in the monomial basis). We describe the high-level ideas for avoiding quasi-polynomial running time later in this section (Section 2.5).

2.1 Approximately pairwise-independent estimators

Let $\mathrm{SAW}_\ell(i,j)$ be the set of self-avoiding walks $\alpha \subseteq [n]^2$ of length ℓ between i and j. Consider the unbiased estimator $p(x) = \frac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha \in \mathrm{SAW}_\ell(i,j)} p_\alpha(x)$ for $y_i y_j$. Above the Kesten–Stigum threshold and for $\ell \geqslant O(\log n)$, we can use the following lemma to show that $p(x)$ has variance $O(1)$ and achieves constant correlation with $z = y_i y_j$. We remark that the previous heuristic variance bound corresponds to the contribution of the terms with $\alpha = \beta$ in the left-hand side of Eq. (2.2).

Lemma 2.1 (Constant-correlation estimator)**.** *Let (x, z) be distributed over $\{0,1\}^n \times \mathbb{R}$. Let $\{p_\alpha\}_{\alpha \in \mathcal{I}}$ be a collection of real-valued n-variate polynomials with the following properties:*

1. *unbiased estimators: $\mathbb{E}[p_\alpha(x) \mid z] = z$ for every $\alpha \in \mathcal{I}$*
2. *approximate pairwise independence: for $\delta > 0$,*

$$\sum_{\alpha,\beta \in \mathcal{I}} \mathbb{E}\, p_\alpha(x) \cdot p_\beta(x) \leqslant \tfrac{1}{\delta^2} \cdot |\mathcal{I}|^2\, \mathbb{E}\, z^2 \tag{2.2}$$

Then, the polynomial $p = \frac{1}{|\mathcal{I}|} \sum_{\alpha \in \mathcal{I}} p_\alpha$ satisfies $\mathbb{E}\, p(x) \cdot z \geqslant \delta \cdot \left(\mathbb{E}\, p(x)^2 \cdot \mathbb{E}\, z^2\right)^{1/2}$.

Remark 2.2. In applying the lemma we often substitute for Eq. (2.2) the equivalent condition

$$\mathbb{E}\, z^2 \cdot \sum_{\alpha,\beta \in \mathcal{I}} \mathbb{E}\, p_\alpha(x) \cdot p_\beta(x) \leqslant \frac{1}{\delta^2} \cdot \sum_{\alpha,\beta \in \mathcal{I}} (\mathbb{E}\, p_\alpha(x) z) \cdot (\mathbb{E}\, p_\beta(x) z)$$

which is conveniently invariant to rescaling of the p_α's.

Proof. Since the polynomial p is an unbiased estimator for z, we have $\mathbb{E}\, p(x) z = \mathbb{E}\, z^2$. By Eq. (2.2), $\mathbb{E}\, p(x)^2 \leqslant (1/\delta^2) \cdot \mathbb{E}\, z^2$. Taken together, we obtain the desired conclusion. □

In Section 3.1, we present the short combinatorial argument that shows that above the Kesten–Stigum bound the estimators for self-avoiding walks satisfy the conditions Eq. (2.2) of the lemma.

We remark that if instead of self-avoiding walks we were to average over all length-ℓ walks between i and j, then the polynomial $p(x)$ computes up to scaling nothing but the (i,j)-entry of the ℓ-th power of the centered adjacency $x - \frac{d}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$. For $\ell \approx \log n$, the ℓ-th power of this matrix converges to vv^T, where v is the top eigenvector of the centered adjacency matrix. For constant degree $d = O(\log n)$, it is well-known that this eigenvector fails to provide a good approximation to the true labeling. In particular, the corresponding polynomial fails to satisfy the conditions of Lemma 2.1 close to the Kesten–Stigum threshold.

2.2 Low-degree estimators for higher-order moments

Let's turn to the general mixed-membership stochastic block model SBM$(n, d, \varepsilon, k, \alpha_0)$. Let (G, σ) be graph G and community structure $\sigma = (\sigma_1, \dots, \sigma_n)$ drawn from this model. Recall that $\sigma_1, \dots, \sigma_n$ are k-dimensional probability vectors, each roughly uniform over $\alpha_0 + 1$ of the coordinates. Let $x \in \{0,1\}^{n \times n}$ be the adjacency matrix of G and let $y_1, \dots, y_k \in \mathbb{R}^n$ be centered community indicator vectors, so that $(y_s)_i = (\sigma_i)_s - \frac{1}{k}$.

It's instructive to see that, unlike for disjoint communities, second moments are not that useful for overlapping communities. As a thought experiment suppose we are given the matrix $\sum_{s=1}^k (y_s)(y_s)^\mathsf{T}$ (which we can estimate using the path polynomials described earlier).

In case of disjoint communities, this matrix allows us to "read off" the community structure directly (because two vertices are in the same community if and only if the entry in the matrix is $1 - O(1/k)$).

For overlapping communities (say the extreme case $\alpha_0 \gg k$ for simplicity), we can think of each σ_i as a random perturbation of the uniform distribution so that $(\sigma_i)_s = (1 + \xi_{i,s})\frac{1}{k}$ for iid Gaussians $\{\xi_{i,s}\}$ with small variance. Then, the centered community indicator vectors $y_1, \dots, y_k$ are iid centered, spherical Gaussian vectors. In particular, the covariance matrix $\sum_{s=1}^k y_s y_s^\mathsf{T}$ essentially only determines the subspace spanned by the vectors $y_1, \dots, y_k$ but not the vectors themselves. (This phenomenon is sometimes called the "rotation problem" for matrix factorizations.)

In contrast, classical factor analysis results show that if we were given the third moment tensor $\sum_{s=1}^k y_s^{\otimes 3}$, we could efficiently reconstruct the vectors $y_1, \dots, y_k$ [Har70, LRA93]. This fact is the reason for aiming to estimate higher order moments in order to recover overlapping communities.

In the same way that a single edge $x_{i,j} - \frac{d}{n}$ gives an unbiased estimator for the (i,j)-entry of the second moment matrix, a 3-star $(x_{i,c} - \frac{d}{n})(x_{j,c} - \frac{d}{n})(x_{k,c} - \frac{d}{n})$ gives an unbiased estimator for the (i,j,k)-entry of the third moment tensor $\sum_{s=1}^k y_s^{\otimes 3}$. This observation is key for the previous best algorithm for mixed-membership community detection [AGHK13]. However, even after averaging over all possible centers c, the variance of this estimator is far too large for sparse graphs. In order to decrease this variance, previous algorithms [AGHK13] project the tensor to the top eigenspace of the centered adjacency matrix of the graph. In terms of polynomial estimators this projection corresponds to averaging over all length-ℓ-armed 3-stars[20] for $\ell = \log n$. Even for disjoint communities, this polynomial estimator would fail to achieve the Kesten–Stigum bound.

[20] A length-ℓ-armed 3-star between $i, j, k \in [n]$ consists of three length-ℓ walks between i, j, k and a common center $c \in [n]$

In order to improve the quality of this polynomial estimator, informed by the shape of threshold-achieving estimator for second moments, we average only over such long-armed 3-stars that are self-avoiding. We show that the resulting estimator achieves constant correlation with the desired third moment tensor precisely up to the Kesten–Stigum bound (Section 5.2).

2.3 Correlation-preserving projection

A recurring theme in our algorithms is that we can compute an approximation vector P that is correlated with some unknown ground-truth vector Y in the Euclidean sense $\langle P, Y\rangle \geqslant \delta \cdot \|P\| \cdot \|Y\|$, where the norm $\|\cdot\|$ is induced by the inner product $\langle \cdot, \cdot \rangle$. (Typically, we obtain P by evaluating a low-degree polynomial in the observable variables and Y is the second or third moment of the hidden variables.)

In this situation, we often seek to improve the quality of the approximation P—not in the sense of increasing the correlation, but in the sense of finding a new approximation Q that is "more similar" to Y while roughly preserving the correlation, so that $\langle Q, Y\rangle \geqslant \delta^{O(1)} \cdot \|Q\| \cdot \|Y\|$. As a concrete example, we may know that Y is a positive semidefinite matrix with all-ones on the diagonal and our goal is to take an arbitrary matrix P correlated with Y and compute a new matrix Q that is still correlated with Y but in addition is positive semidefinite and has all-ones on the diagonal. More generally, we may know that Y is contained in some convex set C and the goal is "project" P into the set C while preserving the correlation. We note that the perhaps most natural choice of Q as the vector closest to P in C does not work in general. (For example, if $Y = (1, 0)$, $C = \{(a, b) \mid a \leqslant 1\}$, and $P = (\delta \cdot M, M)$, then the closest vector to P in C is $(1, M)$, which has poor correlation with Y for large M.)

Theorem 2.3 (Correlation-preserving projection)**.** *Let C be a convex set and $Y \in C$. Let P be a vector with $\langle P, Y\rangle \geqslant \delta \cdot \|P\| \cdot \|Y\|$. Then, if we let Q be the vector that minimizes $\|Q\|$ subject to $Q \in C$ and $\langle P, Q\rangle \geqslant \delta \cdot \|P\| \cdot \|Y\|$, we have*

$$\langle Q, Y\rangle \geqslant \delta/2 \cdot \|Q\| \cdot \|Y\| \,. \tag{2.3}$$

Furthermore, Q satisfies $\|Q\| \geqslant \delta\|Y\|$.

Proof. By construction, Q is the Euclidean projection of 0 into the set $C' := \{Q \in C \mid \langle P, Q\rangle \geqslant \delta\|P\| \cdot \|Y\|\}$. It's a basic geometric fact (sometimes called Pythagorean inequality) that a Euclidean projection into a set decreases distances to points into the set. Therefore, $\|Y - Q\|^2 \leqslant \|Y - 0\|^2$ (using that $Y \in C'$). Thus, $\langle Y, Q\rangle \geqslant \|Q\|^2/2$. On the other hand, $\langle P, Q\rangle \geqslant \delta\|P\| \cdot \|Y\|$ means that $\|Q\| \geqslant \delta\|Y\|$ by Cauchy–Schwarz. We conclude $\langle Y, Q\rangle \geqslant \delta/2 \cdot \|Y\| \cdot \|Q\|$. □

In our applications the convex set C typically consists of probability distributions or similar objects (for example, quantum analogues like density matrices or pseudo-distributions—the sum-of-squares analogue of distributions). Then, the norm minimization in Theorem 2.3 can be viewed as maximizing the Rényi entropy of the distribution Q. From this perspective, maximizing the entropy within the set C' ensures that the correlation with Y is not lost.

2.4 Low-correlation tensor decomposition

Earlier we described how to efficiently compute a 3-tensor P that has correlation $\delta > 0$ with a 3-tensor $\sum_{i=1}^{k} y_i^{\otimes 3}$, where $y_1, \ldots, y_k$ are unknown orthonormal vectors we want to estimate (Section 2.2). Here, the correlation δ depends on how far we are from the threshold and may be minuscule (say 0.001).

It remains to decompose the tensor P into a short list of vectors L so as to ensure that $\mathbb{E}_{i \in [k]} \max_{\hat{y} \in L} \langle \hat{y}, y \rangle \geqslant \delta^{O(1)}$. (Ideally of course $|L| = k$. In the block model context this guarantee requires a small amount of additional work to cross-validate vectors in a larger list.) To the best of our knowledge, previous tensor decomposition algorithms do not achieve this kind of guarantee and require that the correlation of P with the orthogonal tensor $\sum_{i=1}^{k} y_i^{\otimes 3}$ is close to 1 (sometimes even within polynomial factors $1/n^{O(1)}$).

In the current work, we achieve this guarantee building on previous sum-of-squares based tensor decomposition algorithms [BKS15, MSS16]. These algorithms optimize over moments of pseudo-distributions (a generalization of probability distributions) and then apply Jennrich's classical tensor decomposition algorithms to these "pseudo-moments". The advantage of this approach is that it provably works even in situations where Jennrich's algorithm fails when applied to the original tensor.

As a thought experiment, suppose we are able to find pseudo-moments M that are correlated with the orthogonal tensor $\sum_{i=1}^{k} y_i^{\otimes 3}$. Extending previous techniques [MSS16], we show that Jennrich's algorithm applied to M is able to recover vectors that have constant correlation with a constant fraction of the vectors $y_1, \ldots, y_k$.

A priori it is not clear how to find such pseudo-moments M because we don't know the orthogonal tensor $\sum_{i=1}^{k} y_i^{\otimes 3}$, we only know a 3-tensor P that is slightly correlated with it. Here, the correlation-preserving projection discussed in the previous section comes in: by Theorem 2.3 we can efficiently project P into the set of pseudo-moments in a way that preserves correlation. In this way, we obtain pseudo-moments M that are correlated with the unknown orthogonal tensor $\sum_{i=1}^{k} y_i^{\otimes 3}$.

When P is a 3-tensor as above, we encounter technical difficulties inherent to odd-order tensors. (This is a common phenomenon in the tensor-algorithms literature.) To avoid these difficulties we give a simple algorithm, again using the correlation-preserving projection idea, to lift a 3-tensor P which is δ-correlated with an orthogonal tensor A to a 4-tensor P' which is $\delta^{O(1)}$-correlated with an appropriate orthogonal 4-tensor. See Section 7.2.

2.5 From quasi-polynomial time to polynomial time

In this section, we describe how to evaluate certain logarithmic-degree polynomials in polynomial-time (as opposed to quasi-polynomial time). The idea is to use color coding [AYZ95].[21]

For a coloring $c : [n] \to [\ell]$ and a subgraph $\alpha \subseteq [n]^2$ on ℓ vertices, let $F_{c,\alpha} = \frac{\ell^\ell}{\ell!} \cdot \mathbf{1}_{c(\alpha)=[\ell]}$ be a scaled indicator variable of the event that α is colorful.

[21] We thank Avi Wigderson for suggesting that color coding may be helpful in this context.

Theorem 2.4 (Evaluating colorful-path polynomials)**.** *There exists a $n^{O(1)} \cdot \exp(\ell)$-time algorithm that given vertices $i, j \in [n]$, a coloring $c\colon [n] \to [\ell]$ and an adjacency matrix $x \in \{0,1\}^{n\times n}$ evaluates the polynomial*

$$p_c(x) := \tfrac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha \in \mathrm{SAW}_\ell(i,j)} p_\alpha(x) \cdot F_{c,a}\,. \tag{2.4}$$

(Here, $p_\alpha \propto \prod_{ab\in\alpha}(x_{ab} - \frac{d}{n})$ is the polynomial in Eq. (2.1).)

Proof. We can reduce this problem to computing the ℓ-th power of the following $n \cdot 2^\ell$-by-$n \cdot 2^\ell$ matrix: The rows and columns are indexed by pairs (a, S) of vertices $a \in [n]$ and color sets $S \subseteq [\ell]$. The entry for column (a, S) and row (b, T) is equal to $x_{ab} - \frac{d}{n}$ if $T = S \cup \{c(a)\}$ and 0 otherwise. If we compute the ℓ-th power of this matrix, then the entry for column $(i, \emptyset)$ and row $(j, [\ell])$ is the sum over all colorful ℓ-paths from i to j. □

For a fixed coloring c, the polynomial p_c does not provide a good approximation for the polynomial $p(x) := \frac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha\in\mathrm{SAW}_\ell(i,j)} p_\alpha(x)$. In order to get a good approximation, we will choose random colorings and average over them.

If we let c be a random coloring, then by construction $\mathbb{E}_c F_{c,\alpha} = 1$ for every simple ℓ-path α. Therefore, $\mathbb{E}_c p_c(x) = p(x)$ for every $x \in \{0,1\}^{n\times n}$. We would like to estimate the variance of $p_c(x)$. Here, it turns out to be important to consider a typical x drawn from stochastic block model distribution SBM.

$$\mathop{\mathbb{E}}_{x\sim\mathrm{SBM}(n,d,\varepsilon)} \mathop{\mathbb{E}}_c p_c(x)^2 = \tfrac{1}{|\mathrm{SAW}_\ell(i,j)|^2} \sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \mathop{\mathbb{E}}_c F_{c,\alpha} \cdot F_{c,\beta} \cdot \mathop{\mathbb{E}}_{x\sim\mathrm{SBM}} p_\alpha(x) p_\beta(x) \tag{2.5}$$

$$\leqslant e^{2\ell} \cdot \tfrac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \left|\mathop{\mathbb{E}}_x p_\alpha(x) p_\beta(x)\right|. \tag{2.6}$$

For the last step, we use that $\mathbb{E}_c F_{c,\alpha}^2 \leqslant e^{2\ell}$ (because $\ell^\ell/\ell! \leqslant e^\ell$).

The right-hand side of Eq. (2.6) corresponds precisely to our notion of approximate pairwise independence in Lemma 2.1. Therefore, if we are within the Kesten–Stigum bound, $\varepsilon^2 d \geqslant 1 + \delta$, the right-hand side of Eq. (2.6) is bounded by $e^{2\ell} \cdot 1/\delta^{O(1)}$.

We conclude that with high probability over x, the variance of $p_c(x)$ for random c is bounded by $e^{O(\ell)}$. It follows that by averaging over $e^{O(\ell)}$ random colorings we obtain a low-variance estimator for $p(x)$.

2.6 Illustration: push-out effect in spiked Wigner matrices

We turn to a first demonstration of our meta-algorithm beyond the stochastic block model: deriving the critical signal-to-noise ratio for (Gaussian) Wigner matrices (i.e. symmetric matrices with iid entries) with rank-one spikes. This section demonstrates the use of Theorem 1.1; more sophisticated versions of the same ideas (for example our 3rd-moment meta-theorem, Theorem 1.3) will be used in the course of our block model algorithms.

Consider the following Bayesian estimation problem: We are given a spiked Wigner matrix $A = \lambda vv^\mathsf{T} + W$ so that W is a random symmetric matrix with Gaussian entries $W_{ij} \sim \mathcal{N}(0, \frac{1}{n})$ and

$v \sim \mathcal{N}(0, \frac{1}{n}\mathrm{Id})$. The goal is to estimate v, i.e., compute a unit vector $\hat{v}$ so that $\langle v, \hat{v}\rangle^2 \geqslant \Omega(1)$. Since the spectral norm of a Wigner matrix satisfies $\mathbb{E}\|W\| = \sqrt{2}$, it follows that for $\lambda > \sqrt{2}$, the top eigenvector $\hat{v}$ of A satisfies $\langle v, \hat{v}\rangle^2 \geqslant \Omega(1)$. However, it turns out that we can estimate the spike v even for smaller values of λ: a remarkable property of spiked Wigner matrices is that as soon as $\lambda > 1$, the top eigenvector $\hat{v}$ becomes correlated with the spike v [BBAP05]. (This property is sometimes called the "pushout effect".)

Unfortunately known proofs of this property a quite involved. In the following, we apply Theorem 1.1 to give an alternative proof of the fact that it is possible to efficiently estimate the spike v as soon as $\lambda > 1$. Our algorithm is more involved and less efficient than computing the top eigenvector of A. The advantage is that its analysis is substantially simpler compared to previous analyses.

Theorem 2.5 (implicit in [BBAP05])**.** *If $\lambda = 1 + \delta$ for some $1 > \delta > 0$, there is a degree $\delta^{-O(1)} \cdot \log n$ matrix-valued polynomial $f(A) = \{f_{ij}(A)\}_{ij \leqslant n}$ such that*

$$\frac{\mathbb{E}_{W,v} \operatorname{Tr} f(A) v v^\top}{(\mathbb{E}\,\|f(A)\|_F^2)^{1/2} \cdot (\mathbb{E}\,\|vv^\top\|_F^2)^{1/2}} \geqslant \delta^{O(1)}\,.$$

Together with Theorem 1.1, the above theorem gives an algorithm with running time $n^{\log n/\delta^{O(1)}}$ to find $\hat{v}$ with nontrivial $\mathbb{E}\langle \hat{v}, v\rangle^2$.[22]

The analysis of [BBAP05] establishes the above theorem for the polynomial $f(A) = A^\ell$ with $\ell = \delta^{-O(1)} \cdot \log n$. Our proof chooses a different polynomial, which affords a substantially simpler analysis.

Proof of Theorem 2.5. For $\alpha \subseteq \binom{n}{2}$, let $\chi_\alpha(A) = \prod_{\{i,j\}\in\alpha} A_{ij}$. Let $L = \log n/\delta^C$ for C a large enough constant. For $ij \in [n]$, let $SAW_{ij}(L)$ be the collection of all self-avoiding paths from i to j in the complete graph on n vertices. Observe that $\frac{n^{L-1}}{\lambda^L}\chi_\alpha$ for $\alpha \in SAW_{ij}(L)$ is an unbiased estimator of $v_i v_j$:

$$\mathbb{E}\left[\chi_\alpha(A) \mid v_i, v_j\right] = \mathop{\mathbb{E}}_{v}\left[\prod_{k\ell\in\alpha} \mathop{\mathbb{E}}_{W}(W_{k\ell} + \lambda v_k v_\ell) \mid v_i, v_j\right] = \lambda^L v_i v_j \,\mathbb{E}\prod_{k\in\alpha\setminus\{i,j\}} v_k^2 = \frac{\lambda^L}{n^{L-1}} \cdot v_i v_j\,.$$

We further claim that the collection $\{\frac{n^{L-1}}{\lambda^L}\chi_\alpha\}_{\alpha\in SAW_{ij}(L)}$ is approximately pairwise independent in the sense of Lemma 2.1. To show this we must check that

$$\frac{n^{2(L-1)}}{\lambda^{2L}} \sum_{\alpha,\beta} \mathbb{E}\,\chi_\alpha\chi_\beta \leqslant \frac{1}{\delta^2}|SAW_{ij}(L)|^2\,\mathbb{E}\,v_i^2 v_j^2 = \frac{1}{\delta^2}|SAW_{ij}(L)|^2 \cdot \frac{1}{n^2}\,.$$

[22] While this algorithm is much slower than the eigenvector-based algorithm—even after using color coding to improve the $n^{\log n/\delta^{O(1)}}$ running time to $n^{1/\delta^{O(1)}}$—the latter requires many sophisticated innovations and ideas from random matrix theory. This algorithm, by contrast, can be derived and analyzed with our meta-theorem, little innovation required.

The dominant contributers to the sum are α, β which intersect only on the vertices i and j. In that case,

$$\frac{n^{2(L-1)}}{\lambda^{2L}} \mathbb{E}\, \chi_\alpha \chi_\beta = n^{2(L-1)} \mathbb{E} \prod_{k \in \alpha \cup \beta} v_k^2 = \mathbb{E}\, v_i^2 v_j^2 \,.$$

The only other terms which might contribute to the same order are α, β such that $\alpha \cap \beta$ is a union of two paths, one starting at i and one at j. If the lengths of these paths are t and t', respectively, and $t' + t' < L$, then

$$\frac{n^{2(L-1)}}{\lambda^{2L}} \mathbb{E}\, \chi_\alpha \chi_\beta = \frac{n^{2(L-1)}}{\lambda^{2(t+t')}} \mathop{\mathbb{E}}_{v} \left[\prod_{(k,\ell) \in \alpha \cap \beta} (\mathop{\mathbb{E}}_{W} A_{k\ell}^2) \cdot \prod_{(k,\ell) \in \alpha \triangle \beta} v_k v_\ell \right] = \frac{n^{t+t'}}{\lambda^{t+t'}} \cdot (1 + O(\lambda^2/n))^{t+t'}$$

where we have used that $\mathbb{E}\left[A_{k\ell}^2 \,|\, v_k, v_\ell\right] = \frac{1}{n}(1 + O(\lambda^2/n)) \cdot \mathbb{E}\, v_i^2 v_j^2$.

There are at most $|SAW_{ij}(L)|^2 / n^{t+t'}$ choices for such pairs α, β, so long as $t + t' < L$. If $t + t' = L$, then there are n times more choices than the above bound. All together,

$$\frac{n^{2(L-1)}}{\lambda^{2L}} \sum_{\alpha,\beta \in SAW_{ij}(L)} \mathbb{E}\, \chi_\alpha \chi_\beta \leqslant |SAW_{ij}(L)| \cdot \left(\left(\sum_{t=0}^{L} \frac{1}{\lambda^t} \right)^2 + \frac{n}{\lambda^L} \right) \cdot \mathbb{E}\, v_i^2 v_j^2 \leqslant \frac{1 + o(1)}{1 - 1/\lambda} \cdot |SAW_{ij}(L)| \cdot \mathbb{E}\, v_i^2 v_j^2$$

where we have used that $\lambda = 1 + \delta > 1$ and chosen C large enough that $n/\lambda^L \leqslant 1/n$. Rewriting in terms of $\delta = \lambda - 1$ and applying Lemma 2.1 finishes the proof. □

3 Warmup: stochastic block model with two communities

We demonstrate our meta-algorithm by applying it to the two-community stochastic block model. The algorithm achieves here the same threshold for partial recovery as the best previous algorithms [MNS13, Mas13], which is also known to be the information-theoretic threshold [MNS15b].

While the original works involved a great deal of ingenuity, the merit of our techniques is to provide a simple and automatic way to discover and analyze an algorithm achieving the same guarantees.

Definition 3.1 (Two-community stochastic block model)**.** For parameters $\varepsilon, d > 0$, let $\mathrm{SBM}(n, d, \varepsilon)$ be the following distribution on pairs (x, y) where $x \in \{0,1\}^{\binom{n}{2}}$ is the adjacency matrix of an n-vertex graph and $y \in \{\pm 1\}^n$ is a labeling of the n vertices. First, sample $y \sim \{\pm 1\}^n$ uniformly. Then, independently for every pair $i < j$, add the edge $\{i, j\}$ with probability $(1 + \varepsilon)\frac{d}{n}$ if $y_i = y_j$ and with probability $(1 - \varepsilon)\frac{d}{n}$ if $y_i \neq y_j$.

The following theorem gives the best bounds for polynomial-time algorithms for partial recovery in this model. (We remark that the algorithms in [MNS13, Mas13] actually run in time close to linear. In this work, we content ourselves with coarser running time bounds.)

Theorem 3.2 ([MNS13, Mas13]). *Let $\varepsilon \in \mathbb{R}$, $d \in \mathbb{N}$ with $\delta := 1 - \frac{1}{\varepsilon^2 d}$ and $d \leqslant n^{o(1)}$. Then, there exists a randomized polynomial-time algorithm A that given a graph $x \in \{0,1\}^{\binom{n}{2}}$ outputs a labeling $\tilde{y}(x)$ such that for all sufficiently large $n \geqslant n_0(\varepsilon, d)$,*

$$\mathop{\mathbb{E}}_{(x,y)\sim \mathrm{SBM}(n,d,\varepsilon)} \langle \tilde{y}(x), y\rangle^2 \geqslant \delta^{O(1)} \cdot n^2 \,.$$

Here, the factor n^2 in the conclusion of the theorem normalizes the vectors $\tilde{y}(x)$ and y because $\|\tilde{y}(x)\|^2 \cdot \|y\|^2 = n^2$.

In the remainder of this section, we will prove the above theorem by specializing our meta-algorithm for two-community stochastic block model. For simplicity, we will here only analyze a version of algorithm that runs in quasi-polynomial time. See Section 2.5 for how to improve the running time to $n^{1/\mathrm{poly}(\delta)}$.

Algorithm 3.3. For a given n-vertex graph $x \in \{0,1\}^{\binom{n}{2}}$ with average degree d and some parameter $\delta > 0$, execute the following steps:[23]

1. evaluate the following matrix-valued polynomial $P(x) = (P_{ij}(x))$

$$P_{ij}(x) := \sum_{\alpha \in \mathrm{SAW}_\ell(i,j)} p_\alpha(x) \,. \tag{3.1}$$

   Here as in Section 2, $\mathrm{SAW}_\ell(i,j) \subseteq \binom{n}{2}^\ell$ consists of all sets of vertex pairs that form a simple (self-avoiding) path between i and j of length $\ell = \Theta(\log n)/\delta^{O(1)}$.[24] The polynomial p_α is a product of centered edge indicators, so that $p_\alpha(x) = \prod_{ab\in\alpha} \left(x_{ab} - \frac{d}{n}\right)$.[25]

2. compute a matrix Y with minimum Frobenius norm satisfying the constraints

$$\left\{ \begin{array}{r} \mathrm{diag}(Y) = \mathbf{1} \\ \frac{1}{\|P(x)\|_F \cdot n} \cdot \langle P(x), Y\rangle \geqslant \delta' \\ Y \succeq 0 \end{array} \right\} . \tag{3.2}$$

   and output a vector $\tilde{y} \in \{\pm 1\}^n$ obtained by taking coordinate-wise signs of a centered Gaussian vector with covariance Y.[26]

The matrix $P(x)$ is essentially the same as the matrix based on self-avoiding walks analyzed in [MNS13]. The main departure from previous algorithms lies in the second step of our algorithm.

As stated, the first step of the algorithm takes quasi-polynomial because it involves a sum over n^ℓ terms (for $\ell = \Theta(\log n)/\delta^{O(1)}$). In prior works this running time is improved by using non-backtracking paths instead of self-avoiding paths. Non-backtracking paths can be counted in $n^{O(1)}$ time using matrix multiplication, but relating the non-backtracking path polynomial to the self-avoiding path polynomial requires intensive moment-method calculations. An alternative,

[23] The right choice of δ' will depend in a simple way on the parameters ε and d.

[24] In particular, the paths in $\mathrm{SAW}_\ell(i,j)$ are not necessarily paths in the graph x but in the complete graph on n vertices.

[25] Up to scaling, this polynomial is a d/n-biased Fourier character of sparse Erdős-Rényi graph.

[26] In other words, we apply the hyperplane rounding algorithm of Goemans and Williamson.

described in Section 2.5, is to compute the self-avoiding path polynomial P using color-coding, requiring time $n^{O(1)+1/\delta^{O(1)}}$, still polynomial time for any constant $\delta > 0$.

The second step of the algorithm is a convex optimization problem over an explicitly represented spectrahedron. Therefore, this step can be carried out in polynomial time.

We break the analysis of the algorithm into two parts corresponding to the following lemmas. The first lemma shows that if $\varepsilon^2 d > 1$ then the matrix $P(x)$ has constant correlation with yy^T for $(x, y) \sim \mathrm{SBM}(n, d, \varepsilon)$ and n sufficiently large. (Notice that this the main preconditon to apply meta-Theorem 1.1.)

Lemma 3.4 (Low-degree estimator for posterior second moment)**.** *Let $\varepsilon \in \mathbb{R}$ and $d \in \mathbb{N}$, and assume $d = n^{o(1)}$. If $\delta \stackrel{\text{def}}{=} 1 - \frac{1}{\varepsilon^2 d} > 0$ and $n > n_0(\varepsilon, d, \delta)$ is sufficiently large, then the matrix-valued polynomial $P(x)$ in Eq. (3.1) satisfies*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathrm{SBM}(n,d,\varepsilon)} \langle P(x), yy^\mathsf{T}\rangle \geqslant \delta^{O(1)} \cdot \left(\mathop{\mathbb{E}}_{x\sim\mathrm{SBM}(n,d,\varepsilon)} \|P(x)\|_F^2\right)^{1/2} \cdot n \tag{3.3}$$

(Here, the factor n in the conclusion normalizes the matrix yy^T because $\|yy^\mathsf{T}\|_F = n$.)

By application of Markov's inequality to the conclusion of this theorem one shows that with P has $\Omega(1)$-correlation with $yy^\top$ with $\Omega(1)$-probability. As we have noted several times, the same theorem would hold if we replaced P, an average over self-avoiding walk polynomials, with an average over nonbacktracking walk polynomials. This would have the advantage that the resulting polynomial can be evaluated in $n^{O(1)}$ time (i.e. with running time independent of δ), rather than $n^{O(\log n)/\mathrm{poly}(\delta)}$ for P (which can be improved to $n^{\mathrm{poly}(1/\delta)}$ via color coding), but at the cost of complicating the moment-method analysis. Since we are aiming for the simplest possible proofs here we use P as is.

The second lemma shows that given a matrix P that has constant correlation with yy^T for an unknown labeling $y \in \{\pm 1\}^n$, we can efficiently compute a labeling $\tilde{y} \in \{\pm 1\}^n$ that has constant correlation with y. We remark that for this particular situation simpler and faster algorithms work (e.g., choose a random vector in the span of the top $1/\delta^{O(1)}$ eigenvectors of P); these are captured by the meta-Theorem 1.1, which we could use in place of the next lemma. (We are presenting this lemma, which involves a more complex and slower algorithm, in order to have a self-contained analysis in this warmup and because it illustrates a simple form of a semidefinite programming technique that is important for our tensor decomposition algorithm, which we use for overlapping communities.)

Lemma 3.5 (Partial recovery from posterior moment estimate)**.** *Let $P \in \mathbb{R}^{n\times n}$ be a matrix and $y \in \{\pm 1\}^n$ be a vector with $\delta' \coloneqq \frac{1}{\|P\|\cdot n}\langle P, yy^\mathsf{T}\rangle$. Let Y be the matrix of minimum Frobenius such that $Y \succeq 0$, $\mathrm{diag} Y = \mathbf{1}$, and $\frac{1}{\|P\|\cdot n}\langle Y, P\rangle \geqslant \delta'$ (i.e., the constraints Eq. (3.2)). Then, the vector $\tilde{y}$ obtained by taking coordinate-wise signs of a Gaussian vector with mean 0 and covariance Y satisfies*

$$\mathbb{E}\langle \tilde{y}, y\rangle^2 \geqslant \Omega(\delta')^2 \cdot n^2 \,.$$

(Here, the factor n^2 in the conclusion normalizes the vectors $\tilde{y}, y$ because $\|\tilde{y}\|^2 \cdot \|y\|^2 = n^2$.)

Proof. By Theorem 2.3, the matrix Y satisfis $\langle Y, yy^\mathsf{T}\rangle \geqslant (\delta'/2)\|Y\| \cdot \|y\|^2$ and $\|Y\| \geqslant \delta \cdot \|y\|^2$. In particular, $\langle Y, yy^\mathsf{T}\rangle \geqslant \delta^2 n^2/2$. The analysis of rounding algorithm for the Grothendieck problem on psd matrices [AN04], shows that $\mathbb{E}\langle \tilde{y}, y\rangle^2 \geqslant \frac{2}{\pi}\langle Y, yy^\mathsf{T}\rangle \geqslant \Omega(\delta^2)\cdot n^2$. (Here, we use that yy^T is a psd matrix.) □

Taken together, the above lemmas imply a quasi-polynomial time algorithm for partial recovery in $\mathrm{SBM}(n, d, \varepsilon)$ when $\varepsilon^2 d > 1$.

Proof of Theorem 3.2 (quasi-polynomial time version). Let $(x, y) \sim \mathrm{SBM}(n, d, \varepsilon)$ with $\delta \coloneqq 1 - 1/\varepsilon^2 d > 0$. Run Algorithm 3.3 on x with the parameter δ' chosen as $\frac{1}{10}$ times the correlation factor in the conclusion of Lemma 3.4.

Then, by Lemma 3.4, $\mathbb{E}_{(x,y)\sim\mathrm{SBM}(n,d,\varepsilon)}\langle P(x), yy^\mathsf{T}\rangle \geqslant 10\delta' \cdot \mathbb{E}_{x\sim\mathrm{SBM}(n,d,\varepsilon)}\|P(x)\| \cdot n$. By a variant of Markov inequality Theorem A.1, the matrix $P(x)$ satisfies with constant probability $\langle P(x), yy^\mathsf{T}\rangle \geqslant \delta' \cdot \|P(x)\| \cdot n$. In this event, by Lemma 3.5, the final labeling $\tilde{y}$ satisfies $\mathbb{E}_{\tilde{y}}\langle \tilde{y}, y\rangle^2 \geqslant \Omega(\delta')^2 \cdot n^2$. Since this event has constant probability, the total expected correlation satisfies $\mathbb{E}_{(x,y)\sim\mathrm{SBM}(n,d,\varepsilon)}\langle \tilde{y}(x), y\rangle^2 \geqslant \Omega(\delta')^2 \cdot n^2$ as desired. □

It remains to prove Lemma 3.4.

3.1 Low-degree estimate for posterior second moment

We will apply Lemma 2.1 to prove Lemma 3.4. The next two lemmas verify that the conditions of that lemma hold; they immediately imply Lemma 3.4.

Lemma 3.6 (Unbiased estimators for $y_i y_j$)**.** *For $i, j \in [n]$ distinct, let $\mathrm{SAW}_\ell(i, j)$ be the set of all simple paths from i to j in the complete graph on n vertices of length ℓ. Let x_{ij} be the ij-th entry of the adjacency matrix of $G \sim \mathrm{SBM}(n, d, \varepsilon)$, and for $\alpha \in \mathrm{SAW}_\ell(i, j)$, let $p_\alpha(x) = \prod_{ab\in\alpha}(x_{ab} - \frac{d}{n})$. Then for any $y_i, y_j \in \{\pm 1\}$ and $\alpha \in \mathrm{SAW}_\ell(i, j)$,*

$$\left(\frac{n}{\varepsilon d}\right)^\ell \mathbb{E}\left[p_\alpha(x) \mid y_i y_j\right] = y_i y_j\,.$$

Thus, each simple path α from i to j in the complete graph provides an unbiased estimator $(n/\varepsilon d)^\ell p_\alpha(x)$ of $y_i y_j$. It is straightforward to compute that each has variance $\left(\frac{n}{\varepsilon^2 d}\right)^\ell$. If they were pairwise independent, they could be averaged to give an estimator with variance $\frac{1}{|\mathrm{SAW}_\ell(i,j)|}\cdot\left(\frac{n}{\varepsilon^2 d}\right)^\ell = n(\varepsilon^2 d)^{-\ell}$, since there are $n^{\ell-1}$ simple paths from i to j. If ℓ is logarithmic in n, this becomes small. The estimators are not strictly pairwise independent, but they do satisfy an approximate pairwise independence property which will be enough for us.

Lemma 3.7 (Approximate conditional independence)**.** *Suppose $\delta \overset{\text{def}}{=} 1 - \frac{1}{\varepsilon^2 d} \geqslant \Omega(1)$ and $d = n^{o(1)}$. For $i, j \in [n]$ distinct, let $\mathrm{SAW}_\ell(i, j)$ be the set of all simple paths from i to j in the complete graph on n vertices of length $\ell = \Theta(\log n)/\delta^C$ for a large-enough constant C. Let x_{ij} be the ij-th entry of the adjacency matrix of $G \sim \mathrm{SBM}(n, d, \varepsilon)$. Let $p_\alpha(x) = \prod_{ab\in\alpha}(x_{ab} - \frac{d}{n})$. Then*

$$\mathbb{E}\, y_i^2 y_j^2 \sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \mathbb{E}\, p_\alpha(x) p_\beta(x) \leqslant \delta^{-O(1)} \cdot \sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \left(\mathbb{E}\, p_\alpha(x) y_i y_j\right)\left(\mathbb{E}\, p_\beta(x) y_i y_j\right)\,.$$

To prove the lemmas we will use the following fact; the proof is straightforward.

Fact 3.8. *For $x, y \sim$ SBM, the entries of x are all independent conditioned on y, and a, b distinct,*

$$\mathbb{E}\left[x_{ab} - \tfrac{d}{n} \mid y_a, y_b\right] = \frac{\varepsilon d}{n} \cdot y_a y_b \quad \text{and} \quad \mathbb{E}\left[\left(x_{ab} - \tfrac{d}{n}\right)^2 \mid y_a, y_b\right] = \frac{d}{n}\left(1 + \varepsilon y_a y_b + O(d/n)\right).$$

We can prove both of the lemmas.

Proof of Lemma 3.6. We condition on y and expand the expectation.

$$\mathbb{E}\left[p_\alpha(x) \mid y_i y_j\right] = \mathop{\mathbb{E}}_{y}\left[\prod_{ab\in\alpha} \mathbb{E}[x_{ab} - \tfrac{d}{n} \mid y]\right] = \left(\frac{\varepsilon d}{n}\right)^{\ell} \mathop{\mathbb{E}}_{y}\left[\prod_{ab\in\alpha} y_a y_b\right] \quad \text{by Fact 3.8.}$$

Because α is a path from i to j, every index $a \in [n]$ except for i and j appears exactly twice in the product. So, removing the conditioning on y_a for all $a \neq i, j$, we obtain $\mathbb{E}\left[p_\alpha(x) \mid y_i y_j\right] = \left(\frac{\varepsilon d}{n}\right)^{\ell} \cdot y_i y_j$ as desired. □

The proof of Lemma 3.7 is the heart of the proof, and will use the crucial assumption $\varepsilon^2 d > 1$.

Proof of Lemma 3.7. Let $\alpha, \beta \in \mathrm{SAW}_\ell(i, j)$, and suppose that α and β share r edges. Let $\alpha \triangle \beta$ denote the symmetric difference of α and β. Then

$$\begin{aligned}\mathbb{E}\, p_\alpha(x) p_\beta(x) &= \mathop{\mathbb{E}}_{y}\left[\prod_{ab\in\alpha\cap\beta} \mathop{\mathbb{E}}_{x}\left[(x_{ab} - \tfrac{d}{n})^2 \,|\, y_a, y_b\right] \cdot \prod_{ab\in\alpha\triangle\beta} \mathop{\mathbb{E}}_{x}\left[x_{ab} - \tfrac{d}{n} \,|\, y_a, y_b\right]\right] \\ &= \left(\frac{d}{n}\right)^{2\ell-r} \varepsilon^{2\ell-2r} \mathop{\mathbb{E}}_{y}\left[\prod_{ab\in\alpha\cap\beta} (1 + \varepsilon y_a y_b + O(d/n)) \cdot \prod_{ab\in\alpha\triangle\beta} y_a y_b\right]\end{aligned}$$

using Fact 3.8 in the second step. Since α and β are paths, the graph $\alpha \triangle \beta$ has all even degrees, so $\prod_{ab\in\alpha\triangle\beta} y_a y_b = 1$. Furthermore, any subgraph of $\alpha \cap \beta$ contains some odd-degree vertex. So $\mathbb{E}_y \prod_{ab\in\alpha\cap\beta}(1 + \varepsilon y_a y_b + O(d/n)) = (1 + O(d/n))^r$. All in all, we obtain

$$\mathbb{E}\, p_\alpha(x) p_\beta(x) = \left(\frac{d}{n}\right)^{2\ell-r} \varepsilon^{2\ell-2r}(1 + O(d/n))^r \tag{3.4}$$

Suppose $r < \ell$. Paths α, β sharing r edges must share at least r vertices. If they share exactly r vertices, then the shared vertices must form paths in α and β beginning at i and j. Since each path has length ℓ and therefore contains $\ell - 1$ vertices in addition to i and j, there are at most $r \cdot n^{2(\ell-1)-r}$ such pairs α, β (the multiplicative factor r comes because the shared paths starting from i and j could have lengths between 0 and r). Other pairs α, β share r edges but s vertices for some $s > r$. For each s and r, there are at most $n^{2(\ell-1)-s}\ell^{O(s-r)}$ such pairs, because the shared edges must occur as at most $s - r$ paths. Furthermore, $\ell^{O(s-r)} n^{-(s-r)} \leqslant n^{-\Omega(1)}$ when $s > r$. Putting all of this together,

$$\sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \mathbb{E}\, p_\alpha(x) p_\beta(x) \leqslant n^{-2} \cdot \left[\sum_{r=0}^{\ell-1} d^{2\ell-r}\varepsilon^{2\ell-2r}(1 + O(d/n))^r \left(r + n^{-\Omega(1)}\right) + (\varepsilon^2 d)^\ell \cdot n\right]$$

$$= n^{-2} \cdot (1 + n^{-\Omega(1)}) \cdot (\varepsilon d)^{2\ell} \cdot \left(\sum_{r=0}^{\ell} r \cdot (\varepsilon^2 d)^{-r} + (\varepsilon^2 d)^{-\ell} \cdot n \right),$$

The additive factor of $(\varepsilon^2 d)^\ell n$ in the first line comes from the case $r = \ell$ (i.e., $\alpha = \beta$), where there are $n^{\ell-1}$ paths. In the second line we have used the assumption that $d \ll n$ to simplify the expression. Finally, by convergence of the series $\sum_{m=0}^{\infty} m \cdot z^m$ for $|z| < 1$, and the choice of ℓ logarithmic in n, this is at most

$$(1 + n^{-\Omega(1)}) \cdot (\varepsilon d)^{2\ell} \cdot \left(\frac{1}{1 - \frac{1}{\varepsilon^2 d}} \right)^{O(1)}.$$

So, now our goal is to show that

$$\sum_{\alpha,\beta \in \mathrm{SAW}_\ell(i,j)} (\mathbb{E}\, p_\alpha(x) y_i y_j)(\mathbb{E}\, p_\beta(x) y_i y_j) \geqslant n^{-2} \cdot (1 + n^{-\Omega(1)}) \cdot (\varepsilon d)^{2\ell} \cdot \left(\frac{1}{1 - \frac{1}{\varepsilon^2 d}} \right)^{O(1)}.$$

Each term in the left-hand sum is $(\varepsilon d/n)^{2\ell}$ (by Lemma 3.6) and there are $\Omega(n^{2\ell-2})$ such terms, so the left-hand side of the above is at least $\Omega((\varepsilon d)^{2\ell}/n^2)$. This proves the Lemma. □

4 Matrix estimation for generalized block models

In this section we phrase a result essentially due to Abbe and Sandon [AS16a] (and closely related to results by Bordenave et al [BLM15]) in somewhat more general terms. This turns out to be enough to capture an algorithm to estimate a pairwise-vertex-similarity matrix in the $d, k, \alpha, \varepsilon$ mixed-membership block model when $\varepsilon^2 d > k^2(\alpha + 1)^2$.

Let $\mathcal{U}$ be a universe of labels, endowed with some base measure ν, such that $\int 1 \cdot d\nu = 1$. Let μ be a probability distribution on $\mathcal{U}$, with a density relative to ν. (We abuse notation by conflating μ and its associated density). Let $W : \mathcal{U} \times \mathcal{U} \to \mathbb{R}_+$ be a bounded nonnegative function with $W(x, y) = W(y, x)$ for every $x, y \in \mathcal{U}$. Consider a random graph model $G(n, d, W, \mu)$ sampled as follows. For each of n vertices, draw a label $x_i \sim \mu$ independently. Then for each pair $ij \in [n]^2$, independently add the edge (i, j) to the graph with probability $\frac{d}{n} W(x_i, x_j)$. (This captures the W-random graph models used in literature on graphons.)

Let $\mathcal{F}$ denote the space of square-integrable functions $f : \mathcal{U} \to \mathbb{R}$, endowed with the inner product $\langle f, g \rangle = \mathbb{E}_{x \sim \mu} f(x) g(x)$. That is, $f \in \mathcal{F}$ if $\mathbb{E}_{x \sim \mu} f(x)^2$ exists.

We assume throughout that

1. (Stochasticity) For every $x \in \mathcal{U}$, the average $\mathbb{E}_{y \sim \mu} W(x, y) = 1$.
2. (Finite rank) W has a finite-rank decomposition $W(x, y) = \sum_{i \leqslant r} \lambda_i f_i(x) f_i(y)$ where $\lambda_i \in \mathbb{R}$ and $f_i : \mathcal{U} \to \mathbb{R}$. The values λ_i are the eigenvalues of W with respect to the inner product generated by μ. The eigenfunctions are orthonormal with respect to the μ inner product. Notice that the assumptions on W imply that its top eigenfunction $f_1(x)$ is the constant function, with eigenvalue $\lambda_1 = 1$.

3. (Niceness I) Certain rational moments of μ^{-1} exist; that is $\mathbb{E}_{x\sim\mu}\,\mu(x)^{-t}$ exists for $t = -3/2, -2$.

4. (Niceness II) W and μ are nice enough that $W(x,y) \leqslant 1/\sqrt{\mu(x)\mu(y)}$ and $|\overline{W}(x,y)| \leqslant \lambda_2/\sqrt{\mu(x)\mu(y)}$ for every $x, y \in \mathcal{U}$, where $\overline{W}(x,y) = W(x,y) - 1$. (Notice that in the case of discrete W and μ this is always true, and for smooth enough W and μ it is true via a δ-function argument.)

The function W induces a Markov operator $W\colon \mathcal{F} \to \mathcal{F}$. If $f \in \mathcal{F}$, then

$$(Wf)(x) = \mathop{\mathbb{E}}_{y\sim\mu} W(x,y)f(y)\,.$$

(We abuse notation by conflating the function W and the Markov operator W.)

Theorem 4.1 (Implicit in [AS16a])**.** *Suppose the operator W has eigenvalues $1 = \lambda_1 > \lambda_2 > \cdots > \lambda_r$ (each possibly with higher multiplicity) and $\delta \stackrel{\text{def}}{=} 1 - \frac{1}{d\lambda_2^2} > 0$. Let Π be the projector to the second eigenspace of the operator W. For types $x_1, \ldots, x_n \sim \mu$, let $A \in \mathbb{R}^{n\times n}$ be the random matrix $A_{ij} = \Pi(x_i, x_j)$, where we abuse notation and think of $\Pi\colon \mathcal{U}\times\mathcal{U} \to \mathbb{R}$. There is an algorithm with running time $n^{\mathrm{poly}(1/\delta)}$ which outputs an $n \times n$ matrix P such that for $x, G \sim G(n, d, W, \mu)$,*

$$\mathop{\mathbb{E}}_{x,G} \operatorname{Tr} P\cdot A \geqslant \delta^{O(1)}\cdot (\mathop{\mathbb{E}}_{x,G}\|A\|^2)^{1/2}(\mathop{\mathbb{E}}_{x,G}\|P\|^2)^{1/2}\,.$$

When $\mathcal{U}$ is discrete with k elements one recovers the usual k-community stochastic block model, and the condition $\lambda_2^2 > 1$ matches the Kesten-Stigum condition in that setting. When $\lambda_2^2 > 1 + \delta$, the guarantees of Abbe and Sandon can be obtained by applying the above theorem to obtain an estimator P for the matrix $M = \sum_{s\in[k]} v_s v_s^\top$, where v_s is the centered indicator vector of community s. The estimator P will have at least $\delta^{O(1)}/k$ correlation with M, and a random vector in the span of the top $k/\delta^{O(1)}$ eigenvectors of M will have correlation $(\delta/k)^{O(1)}$ with some v_s. Thresholding that vector leads to the guarantees of Abbe and Sandon for the k-community block model, with one difference: Abbe and Sandon's algorithm runs in $O(n\log n)$ time, much faster than the $n^{\mathrm{poly}(1/\delta)}$ running time outlined above. In essence, they achieve this by computing an estimator P' for M which counts only non-backtracking paths in G (the estimator P counts *self-avoiding* paths).

In Section 4.1 we prove a corollary of Theorem 4.1. This yields the algorithm discussed Theorem 1.2 for the mixed-membership blockmodel. As discussed before, the quantitative recovery guarantees of this algorithm are weaker than those of our final algorithm, whose recovery accuracy depends only on the distance δ of the signal-to-noise ratio of the mixed-membership blockmodel to 1. In Section 4.2 we prove Theorem 4.1.

4.1 Matrix estimation for the mixed-membership model

We turn to the mixed-membership model and show that Theorem 4.1 yields an algorithm for partial recovery in the mixed-membership block model. However, the correlation of the vectors output by this algorithm with the underlying community memberships depends both on the signal-to-noise ratio and the number k of communiteis. (In particular, when k is super-constant this algorithm no longer solves the partial recovery task.)

Definition 4.2 (Mixed-Membership Block Model)**.** Let $G(n, d, \varepsilon, \alpha, k)$ be the following random graph ensemble. For each node $i \in [n]$, sample a probability vector $\sigma_i \in \mathbb{R}^k_{\geqslant 0}$ with $\sum_{t\in[k]} \sigma_i(t) = 1$ according to the following (simplified) Dirichlet distribution.

$$\mathbb{P}(\sigma) \propto \prod_{t\in[k]} \sigma_i(t)^{\alpha/k-1}$$

For each pair of vertices $i, i' \in [n]$, sample communities $t \sim \sigma_i$ and $t' \sim \sigma_{i'}$. If $t = t'$, add the edge $\{i, i'\}$ to G with probability $\frac{d}{n}(1 + (1 - \frac{1}{k})\varepsilon)$. If $t \neq t'$, add the edge $\{i, i'\}$ to G with probability $\frac{d}{n}(1 - \frac{\varepsilon}{k})$. (For simplicity, throughout this paper we consider only the case that the communities have equal sizes and the connectivity matrix has just two unique entries.)

Theorem 4.3 (Constant-degree partial recovery for mixed-membership block model, k-dependent error)**.** *For every $\delta > 0$ and $d(n), \varepsilon(n), k(n), \alpha(n)$, there is an algorithm with running time $n^{O(1)+1/\delta^{O(1)}}$ with the following guarantees when*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad k, \alpha \leqslant n^{o(1)} \text{ and } \varepsilon^2 d \leqslant n^{o(1)}\,.$$

Let $\sigma, G \sim G(n, d, \varepsilon, k, \alpha)$ and for $s \in [k]$ let $v_s \in \mathbb{R}^n$ be given by $v_s(i) = \sigma_i(s) - \frac{1}{k}$.

The algorithm outputs a vector x such that $\mathbb{E}\langle x, v_1\rangle^2 \geqslant \delta'\|x\|^2\|v_1\|^2$, for some $\delta' \geqslant (\delta/k)^{O(1)}$.[27]

Ideally one would prefer an algorithm which outputs $\tau_1, \ldots, \tau_n \in \Delta_{k-1}$ with $\mathrm{corr}(\sigma, \tau) \geqslant \delta'/(\alpha+1)$. If one knew that $\langle x, v_1\rangle \geqslant \delta'\|x\|\|v\|$ rather than merely the guarantee on $\langle x, v_1\rangle^2$ (which does not include a guarantee on the sign of x), then this could be accomplished by correlation-preserving projection, Theorem 2.3. The tensor methods we use in our final algorithm for the mixed-membership model are able to obtain a guarantee on $\langle x, v_1\rangle$ and hence can output probability vectors $\tau_1, \ldots, \tau_n$.[28]

To prove Theorem 4.3 we will apply Theorem 4.1 and then a simple spectral rounding algorithm; the next two lemmas capture these two steps.

Lemma 4.4 (Mixed-membership block model, matrix estimation)**.** *If $\mathcal{U}$ is the $(k-1)$-simplex, μ is the α, k Dirichlet distribution, and $W(\sigma, \sigma') = 1 - \frac{\varepsilon}{k} + \varepsilon\langle\sigma, \sigma'\rangle$, then $G(n, d, W, \mu)$ is the mixed-membership block model with parameters $k, d, \alpha, \varepsilon$. In this case, the second eigenvalue of W has multiplicity $k-1$ and has value $\lambda_2 = \frac{\varepsilon}{k(\alpha+1)}$.*

Proof. The first part of the claim follows from the definitions. For the second part, note that W has the following decomposition

$$W(\sigma, \tau) = 1 + \sum_{i\leqslant k} \varepsilon(\sigma_i - \tfrac{1}{k})(\tau_i - \tfrac{1}{k})\,.$$

[27] The requirement $\varepsilon^2 d \leqslant n^{o(1)}$ is for technical convenience only; as $\varepsilon^2 d$ increases the recovery problem only becomes easier.

[28] Such a guarantee could be obtained here by using a cross-validation scheme on x to choose between x and $-x$. Since we are focused on what can be accomplished by matrix estimation methods generally we leave this to the reader.

The functions $\sigma \mapsto \sigma_i - \frac{1}{k}$ are all orthogonal to the constant function $\sigma \mapsto 1$ with respect to μ; i.e.

$$\mathop{\mathbb{E}}_{\sigma\sim\mu} 1 \cdot (\sigma_i - \tfrac{1}{k}) = 0$$

because $\mathbb{E}\,\sigma_i = \frac{1}{k}$.

It will be enough to test the above Rayleigh quotient

$$\frac{\mathbb{E}_{\sigma\sim\mu} f(\sigma) \cdot (Wf)(\sigma)}{\mathbb{E}_{\sigma\sim\mu} f(\sigma)^2}$$

with any function $f(\sigma)$ in the span of the functions $\sigma \mapsto \sigma_i - \frac{1}{k}$. If we pick $f(\sigma) = \sigma_1 - \frac{1}{k}$ the remaining calculation is routine, using only the second moments of the Dirichlet distribution (see Fact 4.5 below). □

Fact 4.5 (Special case of Fact A.3)**.** *Let $\sigma \in \mathbb{R}^k$ be distributed according to the α, k Dirichlet distribution. Let $\tilde{\sigma} = \sigma - \frac{1}{k} \cdot 1$ be centered. Then $\mathbb{E}(\tilde{\sigma})(\tilde{\sigma})^\top = \frac{1}{k(\alpha+1)} \cdot \Pi$ where Π is the projector to the complement of the all-1s vector in $\mathbb{R}^k$.*

We analyze a simple rounding algorithm.

Lemma 4.6. *Let $M = \sum_{i=1}^k v_i v_i^\top$ be an $n \times n$ symmetric rank-k PSD matrix. Let $P \in \mathbb{R}^{n\times n}$ be another symmetric matrix such that $\langle P, M\rangle \geqslant \delta\|P\|\|M\|$ (where $\|\cdot\|$ is the Frobenious norm). Then for at least one vector v among $v_1, \ldots, v_k$, a random unit vector x in the span of the top $(k/\delta)^{O(1)}$ eigenvectors of P satisfies*

$$\mathbb{E}\langle x, v\rangle^2 \geqslant (\delta/k)^{O(1)}\|v\|^2 .$$

Now we can prove Theorem 4.3.

Proof of Theorem 4.3. Lemma 4.4 shows that the conditions of Theorem 4.1 hold, and hence (via color coding) there is an $n^{\mathrm{poly}(1/\delta)}$ time algorithm to compute a matrix P such that $\langle P, M\rangle \geqslant \delta^{O(1)}\|P\|\|M\|$ with probability at least $\delta^{O(1)}$, where $M = \sum_{s\in[k]} v_s v_s^\top$. (The reader may check that the matrix A of Theorem 4.1 is in this case the matrix M described here.)

Applying Lemma 4.6 shows that a random unit vector x in the span of the top $(k/\delta)^{O(1)}$ eigenvectors of P satisfies $\langle x, v\rangle^2 \geqslant (\delta/k)^{O(1)}\|v\|^2$, where $v \in \mathbb{R}^n$ has entries $v_i = \sigma_i(1)$. (The choice of 1 is without loss of generality.) □

4.2 Proof of Theorem 4.1

Definition 4.7. For a pair of functions $A, B : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$, we denote by AB their product, whose entries are $(AB)(x, y) = \mathbb{E}_{z\sim\mu} A(x, z)B(z, y)$.

The strategy to prove Theorem 4.1 will as usual be to apply Lemma 2.1. We check the conditions of that Lemma in the following Lemmas, deferring their proofs till the end of this section.

Lemma 4.8. *Let G_{ij} be the 0/1 indicator for the presence of edge $i \sim j$ in a graph G. As usual, let $\mathrm{SAW}_\ell(i,j)$ be the collection of simple paths of length ℓ in the complete graph on n vertices from i to j.*

Let $x, G \sim G(n,d,W,\mu)$. Let $\alpha \in \mathrm{SAW}_\ell(i,j)$. Let $p_\alpha(G) = \prod_{ab\in\alpha}(G_{ab} - \frac{d}{n})$. Let $\overline{W}(x,y) = W(x,y) - 1$. Then

$$\mathbb{E}\left[p_\alpha(G) \mid x_i, x_j\right] = \left(\frac{d}{n}\right)^\ell \overline{W}^{\ell-1}(x_i,x_j)$$

Lemma 4.9. *With the same notation as in Lemma 4.8, as long as $\ell \geqslant C\log n/\delta^{O(1)}$ for a large-enough constant C,*

$$\left(\frac{n}{d}\right)^{2\ell} \sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \mathbb{E}\, p_\alpha(G)p_\beta(G) \leqslant \delta^{-O(1)} \cdot |\mathrm{SAW}_\ell(i,j)|^2 \cdot \mathbb{E}\,\overline{W}^{\ell-1}(x_i,x_j)^2\,.$$

(The constant C depends on W and the moments of μ.)

Proof of Theorem 4.1. Let $B_{ij} = \lambda_2^{-(\ell-1)}\overline{W}^{\ell-1}(x_i,x_j)$. By Lemma 4.8, Lemma 4.9, and Lemma 2.1, there is matrix polynomial $P(G)$, computable to $1/\mathrm{poly}(n)$-accuracy in time $n^{\mathrm{poly}(1/\delta)}$ by color coding, such that

$$\mathbb{E}\,\mathrm{Tr}\,PB^T \geqslant \delta^{O(1)}(\mathbb{E}\,\|P\|^2)^{1/2}(\mathbb{E}\,\|B\|^2)^{1/2}\,.$$

At the same time, $B - A$ has entries

$$(B-A)_{ij} = \sum_{3\leqslant t\leqslant r}\left(\frac{\lambda_t}{\lambda_2}\right)^{\ell-1}\Pi_t(x_i,x_j)$$

where the Π_t projects to the t-th eigenspace of W. Since W is bounded, choosing ℓ a large enough multiple of $\log n$ ensures that $\mathbb{E}\,\|B-A\|^2 \leqslant n^{-100}\,\mathbb{E}\,\|B\|^2$, so the theorem now follows by standard manipulations. □

4.3 Proofs of Lemmas

Proof of Lemma 4.8. As usual, we simply expand p, obtaining

$$\begin{aligned}\mathbb{E}\left[p_\alpha(G) \mid x_i, x_j\right] &= \mathop{\mathbb{E}}_x\left[\prod_{ab\in\alpha}\frac{d}{n}\cdot(W_{x_a,x_b}-1) \mid x_i,x_j\right]\\ &= \left(\frac{d}{n}\right)^\ell \cdot \overline{W}^{\ell-1}(x_i,x_j)\,.\end{aligned}$$

□

We will need some small facts to help in proving Lemma 4.9.

Fact 4.10. *If $\ell - t \geqslant C\log n$ for large enough $C = C(W)$, then*

$$\lambda_2^{2t}\mathop{\mathbb{E}}_{x,y\sim\mu}\overline{W}^{\ell-t}(x,y)^2 \leqslant (1+o(1))\cdot\mathop{\mathbb{E}}_{x,y\sim\mu}\overline{W}^{\ell}(x,y)^2\,.$$

Also, for any $t \leqslant \ell$,

$$\lambda_2^{2t}\mathop{\mathbb{E}}_{x,y\sim\mu}\overline{W}^{\ell-t}(x,y)^2 \leqslant r\cdot\mathop{\mathbb{E}}_{x,y\sim\mu}\overline{W}^{\ell}(x,y)^2\,.$$

where r is the rank of W.

Proof. Using the eigendecomposition of $\overline{W}$, we have that $\mathbb{E}_{x,y\sim\mu}\,\overline{W}^{\ell-t}(x,y)^2 = \sum_{2\leqslant i\leqslant r}\lambda_i^{2(\ell-t)}$ and similarly $\mathbb{E}_{x,y\sim\mu}\,\overline{W}^{\ell}(x,y)^2 = \sum_{2\leqslant i\leqslant r}\lambda_i^{2\ell}$. If $i>2$, then

$$\lambda_2^{2t}\lambda_i^{2(\ell-t)} = \lambda_2^{2\ell}(\lambda_i/\lambda_2)^{2(\ell-t)} \leqslant \lambda_2^{2\ell}/n$$

by our assumption that $\ell - t \geqslant C\log n$ for large enough C. This finishes the proof of the first claim; the second one is similar. □

Proof of Lemma 4.9. Pairs α,β which share only the vertices i,j each contribute exactly $\mathbb{E}\,\overline{W}^{\ell-1}(x_i,x_j)^2$ to the left-hand side, by Lemma 4.8. Consider next the contribution of α,β whose shared edges form paths originating at i and j. Suppose there are t such shared edges. Then

$$\begin{aligned}\mathbb{E}\,p_\alpha p_\beta &= \left(\frac{d}{n}\right)^{2\ell-t}\mathop{\mathbb{E}}_{x}\prod_{ab\in\alpha\triangle\beta}\overline{W}(x_a,x_b)\cdot\prod_{ab\in\alpha\cap\beta}(W(x_a,x_b)+O(d/n))\\ &= \left(\frac{d}{n}\right)^{2\ell-t}(1+O(d/n))^t\,\mathbb{E}\,\overline{W}^{2(\ell-t-1)}(x,y)^2\,,\end{aligned}$$

where for the second equality we used the assumption $\mathbb{E}_{x\sim\mu}\,W(x,y)=1$ for every y.

If $\ell - t > C\log n$ for the constant in Fact 4.10, then this is at most $(1+o(1))\left(\frac{d}{n}\right)^{2\ell-t}\lambda_2^{-2t}\,\mathbb{E}\,\overline{W}^{2(\ell-1)}(x,y)^2$, and for every $t\leqslant\ell$ it is at most $r\cdot\left(\frac{d}{n}\right)^{2\ell-t}\lambda_2^{-2t}\,\mathbb{E}\,\overline{W}^{2(\ell-1)}(x,y)^2$.

There are at most $|\mathrm{SAW}_\ell(i,j)|^2/n^t\cdot t$ choices for such pairs α,β, except when $t=\ell$, in which case there are $|\mathrm{SAW}_\ell(i,j)|^2/n^{t-1}$ choices. So the total contribution from such α,β is at most

$$\begin{aligned}&|\mathrm{SAW}_\ell(i,j)|^2\cdot\mathop{\mathbb{E}}_{x,y}\overline{W}^{\ell-1}(x,y)^2\cdot\left(\sum_{t\leqslant\ell/2}td^{-t}\lambda_2^{-2t}+nr\cdot\sum_{\ell\geqslant t>\ell/2}td^{-t}\lambda_2^{-2t}\right)\\ &\leqslant\delta^{-O(1)}|\mathrm{SAW}_\ell(i,j)|^2\mathop{\mathbb{E}}_{x,y}\overline{W}^{\ell-1}(x,y)^2\,.\end{aligned}$$

It remains to handle pairs α,β which share t vertices and s edges for $t>s$. If $t,s\leqslant\ell-2$, then there are only $n^{2(\ell-1)-s}\ell^{O(t-s)}$ choices for such a pair α,β. The contribution of each such pair we bound as follows

$$\mathbb{E}\,p_\alpha p_\beta = \left(\frac{d}{n}\right)^{2\ell-s}\mathbb{E}\prod_{ab\in\alpha\cap\beta}\mathbb{E}(G_{ab}-\tfrac{d}{n})^2\cdot\prod_{ab\in\alpha\triangle\beta}\overline{W}_{x_a,x_b}\,.$$

Now, $\mathbb{E}\left[(G_{ab}-\frac{d}{n})^2\,|\,x\right] = \frac{d}{n}(W(x_a,x_b)+O(d/n))$ by straightforward calculations, so the above is

$$\begin{aligned}&(1+O(d/n))^s\left(\frac{d}{n}\right)^{2\ell-s}\mathop{\mathbb{E}}_{x}\prod_{ab\in\alpha\cap\beta}W(x_a,x_b)\cdot\prod_{ab\in\alpha\triangle\beta}\overline{W}(x_a,x_b)\\ &\leqslant(1+O(d/n))^s\left(\frac{d}{n}\right)^{2\ell-s}\lambda_2^{2\ell-s}\prod_{a\in\alpha\cup\beta}\mu(x_a)^{-\deg_{\alpha,\beta}(a)/2}\end{aligned}$$

where $\deg_{\alpha,\beta}(a)$ is the degree of the vertex a in the graph $\alpha \cup \beta$. Any degree-2 vertices simply contribute 1 in the above, since $\mathbb{E}_{x\sim\mu} 1/\mu(x) = 1$. There are at most $t - s$ vertices of higher degree; they may have degree at most 4. They each contribute at most some number $C = C(\mu)$ by the niceness assumptions on μ. So the above is at most

$$(1 + o(1))\left(\frac{d}{n}\right)^{2\ell-s} \lambda_2^{2\ell-s} \exp\{O(t-s)\}\,.$$

Putting things together as in Lemma 3.7 finishes the proof. □

Proof of Lemma 4.6. By averaging, there is some v among $v_1, \dots, v_k$ such that

$$\langle P, vv^\top\rangle \geqslant \frac{\delta}{k} \cdot \|P\| \cdot \|M\| \geqslant \frac{\delta}{k} \cdot \|P\| \cdot \|vv^\top\|$$

where the second inequality uses $M \succeq 0$. Renormalizing, we may assume $\|P\|$ has Frobenious norm 1 and v is a unit vector; in this case we obtain $\langle v, Pv\rangle \geqslant \delta/k$. Writing out the eigendecomposition of P, let $P = \sum_{i=1}^n \lambda_i u_i u_i^\top$ and we get

$$\sum_{i=1}^n \lambda_i \langle v, u_i\rangle^2 \geqslant \delta/k$$

By Cauchy-Schwarz,

$$\sum_{i=1}^n \lambda_i \langle v, u_i\rangle^2 \leqslant \left(\sum_{i=1}^n \lambda_i^2 \langle v, u_i\rangle^2\right)^{1/2}$$

and hence $\sum_{i=1}^n \lambda_i^2 \langle v, u_i\rangle^2 \geqslant (\delta/k)^2$, while $\sum_{i=1}^n \lambda_i^2 = 1$. Now the Lemma follows from Markov's inequality. □

5 Tensor estimation for mixed-membership block models

5.1 Main theorem and algorithm

Theorem 5.1 (Constant-degree partial recovery for mixed-membership block model)**.** *There is a constant C such that the following holds. Let $G(n, d, \varepsilon, k, \alpha)$ be the mixed-membership block model. For every $\delta \in (0, 1)$ and $d(n), \varepsilon(n), k(n), \alpha(n)$, there is an algorithm with running time $n^{O(1)+1/\delta^{O(1)}}$ with the following guarantees when*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad k, \alpha \leqslant n^{o(1)} \text{ and } \varepsilon^2 d \leqslant n^{o(1)}\,.$$

Let $\sigma, G \sim G(n, d, \varepsilon, k, \alpha)$. Let $t = (\alpha+1)\cdot\frac{k}{k+\alpha}$ (samples from the α, k Dirichlet distribution are approximately uniform over t coordinates). Given G, the algorithm outputs probability vectors $\tau_1, \dots, \tau_n \in \Delta_{k-1}$ such that

$$\mathbb{E}\,\mathrm{corr}(\sigma, \tau) \geqslant \delta^C\left(\frac{1}{t} - \frac{1}{k}\right)\,.$$

(Recall the definition of correlation from (1.5).)[29]

[29]The requirement $\varepsilon^2 d \leqslant n^{o(1)}$ is for technical convenience only; as $\varepsilon^2 d$ increases the recovery problem only becomes easier.

Let $c \in (0,1)$ be a small-enough constant. Let $C(c) \in \mathbb{N}$ be a large-enough constant (different from the constant in the theorem statement above). There are three important parameter regimes:

1. Large δ, when $\delta \in [1-c, 1)$.
2. Small δ, when $\delta \in (1-c, 1/k^{1/C})$. This is the main regime of interest. In particular when $k(n) \to \infty$ this contains most values of δ.
3. Tiny δ, when $\delta \in (0, 1/k^{1/C}]$. (This regime only makes sense when $k(n) \leqslant O(1)$.)

Let G_{input} be an n-node graph.

Algorithm 5.2 (Main algorithm for mixed-membership model)**.** Let $\eta > 0$ be chosen so that $1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d(1-\eta)} \geqslant \delta^2$ and $o(1) \geqslant \eta \geqslant n^{-\gamma}$ for every constant $\gamma > 0$. (This guarantees that enough edges remain in the input after choosing a holdout set.)

1. Select a partition of $[n]$ into A and $\overline{A}$ at random with $|\overline{A}| = \eta n$. Let $G = A \cap G_{\text{input}}$
2. If δ is large, run Algorithm 5.5 on (G_{input}, G, A).
3. If δ is small, run Algorithm 5.4 on (G_{input}, G, A).
4. If δ is tiny, run Algorithm 5.3 on (G_{input}, G, A).

Algorithm 5.3 (Tiny δ)**.**

1. Run the algorithm from Theorem 4.3 on G with parameters $(1-\eta)d, k, \varepsilon, \alpha$ to obtain a vector $x \in \mathbb{R}^{n-\eta n}$.
2. Evaluate the quantities $s_x^{(3)} = S_3(G_{\text{input}} \setminus G, x)$ and $s_x^{(4)} = S_4(G_{\text{input}} \setminus G, x)$, the polynomials from Lemma 5.8. If $s_x^{(4)} < C(n, \alpha, k, \varepsilon, d, \eta)$, output random labels $\tau_1, \dots, \tau_n$. (The scalar $C(n, \alpha, k, \varepsilon, d, \eta)$ depends in a simple way on the parameters.)
3. If $s_x^{(3)} < 0$, replace x by $-x$.
4. Run the cleanup algorithm from Lemma 5.11 on the vector x, padded with zeros to make a length n vector. Output the resulting $\tau_1, \dots, \tau_n$.

Algorithm 5.4 (Small δ)**.**

1. Using color coding, evaluate the degree-$\log n/\text{poly}(\delta)$ polynomial $P(G) = (P_{ijk}(G))$ from Lemma 5.6. (This takes time $n^{\text{poly}(1/\delta)}$.)
2. Run the 3-tensor to 4-tensor lifting algorithm (Theorem 7.14) on $P(G)$ to obtain a 4-tensor T.
3. Run the low-correlation tensor decomposition algorithm (Corollary 7.3) on T, implementing the cross-validation oracle O as follows. For each query $x \in \mathbb{R}^{n-\eta n}$, compute $s_x^{(4)} = S_4(G_{\text{input}} \setminus G, x)$, the quantity from Lemma 5.9. If $s_x^{(4)} > C(n, d, k, \varepsilon, \alpha, \eta)$ (distinct from the C above, again depending in a simple way on the parameters), output YES, otherwise output NO. The tensor decomposition algorithm returns unit vectors $x_1, \dots, x_k \in \mathbb{R}^{n-\eta n}$.
4. For each $x_1, \dots, x_k$, compute $s_i^{(3)} = S_3(G_{\text{input}} \setminus G, x_i)$ and $s_i^{(4)} = S_4(G_{\text{input}} \setminus G, x_i)$. For any x_i for which $s_i^{(4)} < C(n, d, k, \varepsilon, \alpha, \eta)$, replace x_i with a uniformly random unit vector. For any x_i for which $s_i^{(3)} < 0$, replace x_i with $-x_i$.
5. Run the algorithm from Lemma 5.10 on $(x_1, \dots, x_k)$ (padded with zeros to make an $n \times k$ matrix) and output the resulting $\tau_1, \dots, \tau_n$.

Algorithm 5.5 (Large δ)**.**

1. Using color coding, evaluate the degree-$\log n/\mathrm{poly}(\delta)$ polynomial $P(G) = (P_{ijk}(G))$ from Lemma 5.6. (This takes time $n^{\mathrm{poly}(1/\delta)}$.)
2. Run the 3-tensor to 4-tensor lifting algorithm (Theorem 7.14) on $P(G)$ to obtain a 4-tensor T.
3. Run the low-correlation tensor decomposition algorithm on T, obtaining unit vectors $x_1, \ldots, x_k$.
4. For each x_i, compute the quantity $s_i^{(4)} = S_4(G_{\text{input}} \setminus G, x_i)$ from Lemma 5.9. If $s_i^{(4)} < C(n, d, k, \varepsilon, \alpha, \eta)$, replace x_i with a uniformly random unit vector. (The scalar threshold $C(n, d, k, \varepsilon, \alpha, \eta)$ depends in a simple way on the parameters.)
5. For each x_i, compute the quantity $s_i^{(3)} = S_3(G_{\text{input}} \setminus G, x_i)$ from Lemma 5.9. If $s_i^{(3)} < 0$, replace x_i with $-x_i$.
6. Run the algorithm from Lemma 5.10 on the matrix $x = (x_1, \ldots, x_k)$ and output the resulting $\tau_1, \ldots, \tau_n$.

We will analyze each of these algorithms separately, but we state the main lemmas together because many are shared among tiny, small, and large δ cases. Two of the algorithms use the low-correlation tensor decomposition algorithm as a black box; Corollary 7.3 in Section 7 captures the guarantees of that algorithm.

The first thing we need is Theorem 4.3, which describes a second-moment based algorithm used as a subroutine by Algorithm 5.3. (This subroutine was already analyzed in Section 4.)

Theorem (Restatement of Theorem 4.3)**.** *For every $\delta > 0$ and $d(n), \varepsilon(n), k(n), \alpha(n)$, there is an algorithm with running time $n^{O(1)+1/\delta^{O(1)}}$ with the following guarantees when*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad k, \alpha \leqslant n^{o(1)} \text{ and } \varepsilon^2 d \leqslant n^{o(1)}.$$

Let $\sigma, G \sim G(n, d, \varepsilon, k, \alpha)$ and for $s \in [k]$ let $v_s \in \mathbb{R}^n$ be given by $v_s(i) = \sigma_i(s) - \frac{1}{k}$.

The algorithm outputs a vector x such that $\mathbb{E}\langle x, v_1\rangle^2 \geqslant \delta' \|x\|^2 \|v_1\|^2$, for some $\delta' \geqslant (\delta/k)^{O(1)}$.[30]

The proofs of all the lemmas that follow can be found later in this section. Next, we state the tensor estimation lemma used to analyze the tensor P computed in Algorithm 5.4 and Algorithm 5.5.

Lemma 5.6. *Suppose*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad \varepsilon^2 d \leqslant n^{1-\Omega(1)} \text{ and } k, \alpha \leqslant n^{o(1)}.$$

For a collection $\sigma_1, \ldots, \sigma_n$ of probability vectors, let $V(\sigma) = \sum_{s\in[k]} v_s^{\otimes 3}$, where the vectors $v_s \in \mathbb{R}^n$ have entries $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $w_s \in \mathbb{R}^n$ have entries $w_s(i) = v_s(i) + \frac{1}{k\sqrt{\alpha+1}}$. (Note that $\mathbb{E}\langle w_s, w_t\rangle = 0$ for $s \neq t$.) Let $W(\sigma) = \sum_{s\in[k]} w_s^{\otimes 3}$.

[30] The requirement $\varepsilon^2 d \leqslant n^{o(1)}$ is for technical convenience only; as $\varepsilon^2 d$ increases the recovery problem only becomes easier.

If $G \sim G(n, d, \varepsilon, \alpha, k)$, there is a degree $O(\log n/\delta^{O(1)})$ polynomial $P(G) \in (\mathbb{R}^n)^{\otimes 3}$ such that

$$\frac{\mathbb{E}_{\sigma,G}\langle P(G), W(\sigma)\rangle}{\left(\mathbb{E}_{\sigma,G}\|P(G)\|^2\right)^{1/2} \cdot \left(\mathbb{E}_{\sigma,G}\|W(\sigma)\|^2\right)^{1/2}} \geqslant \delta^{O(1)}$$

Furthermore, P can be evaluated up to $(1 + 1/\mathrm{poly}(n))$ multiplicative error (whp) in time $n^{\mathrm{poly}(1/\delta)}$.

Two of our algorithms use the low-correlation tensor decomposition algorithm of Corollary 7.3. That corollary describes an algorithm which recovers an underlying orthogonal tensor, but the tensor W is not quite orthogonal. The following lemma, proved via standard matrix concentration, captures the notion that W is close to orthogonal.

Lemma 5.7. *Let $\sigma_1, \ldots, \sigma_n$ be iid draws from the α, k Dirichlet distribution. Let $w_s \in \mathbb{R}^n$ be given by $w_s(i) = \sigma_i(s) - \frac{1}{k}(1 - 1/\sqrt{\alpha + 1})$. Then as long as $k, \alpha \leqslant n^{o(1)}$, with high probability*

$$(1 + n^{-\Omega(1)}) \cdot \mathrm{Id} \preceq \frac{1}{k}\sum_{s=1}^{k} \frac{w_s w_s^\top}{\mathbb{E}\|w_s\|^2} \preceq (1 + n^{-\Omega(1)}) \cdot \mathrm{Id}\,.$$

All of the algorithms perform some cross-validation using the holdout set $\overline{A}$. The next two lemmas offer what we need to analyze the cross-validations.

Lemma 5.8. *Let n_0, n_1 satisfy $n_0 + n_1 = n$. Let $A \subseteq [n]$ have size $|A| = n_1 \geqslant n^{\Omega(1)}$. Let $k = k(n), d = d(n), \varepsilon = \varepsilon(n), \alpha = \alpha(n) > 0$ and $\alpha, k, \varepsilon^2 d \leqslant n^{o(1)}$. Let $\sigma \in \Delta_{k-1}^{n_0}$. Let $v_s \in \mathbb{R}^{n_0}$ have entries $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $\tau_1, \ldots, \tau_{n_1}$ be iid from the α, k Dirichlet distribution.*

Let G be a random bipartite graph on vertex sets $A, [n] \setminus A$, with edges distributed according to the $n, d, \varepsilon, k, \alpha$ mixed-membership model with labels σ, τ. Let $x \in \mathbb{R}^{n_0}$. For $a \in A$, let $P_a(G, x)$ be the expression

$$P_a(G, x) = \sum_{ijk \in \overline{A} \text{ distinct}} (G_{ai} - \tfrac{d}{n})(G_{aj} - \tfrac{d}{n})(G_{ak} - \tfrac{d}{n}) x_i x_j x_k\,.$$

Let $S_3(G, x)$ be

$$S_3(G, x) = \sum_{a \in A} P_a(G, x)\,.$$

There is a number $C = C(n, d, k, \varepsilon, \alpha, n_1)$ such that

$$\mathop{\mathbb{P}}_{G,\tau}\left\{\left|C \cdot S_3(G, x) - \sum_{s \in [k]} \frac{\langle v_s, x\rangle^3}{\|v_s\|^3}\right| > n^{-\Omega(1)}\right\} \leqslant \exp(-n^{\Omega(1)})\,.$$

Similarly, there are scalars $C(n, d, k, \varepsilon, \alpha, n_1), C'(n, d, k, \varepsilon, \alpha, n_1)$ such that the following holds. For $a \in A$, let

$$Q_a(G, x) = \sum_{ijk\ell \in \overline{A} \text{ distinct}} (G_{ai} - \tfrac{d}{n})(G_{aj} - \tfrac{d}{n})(G_{ak} - \tfrac{d}{n})(G_{a\ell} - \tfrac{d}{n}) x_i x_j x_k x_\ell\,.$$

and let

$$R_a(G, x) = \sum_{ij \in \overline{A} \text{ distinct}} (G_{ai} - \tfrac{d}{n})(G_{aj} - \tfrac{d}{n}) x_i x_j\,.$$

Finally let

$$S_4(G,x) = C \cdot \sum_{a\in A} Q_a(G,x) - C' \cdot \left(\sum_{a\in A} R_a(G,x)\right)^2 .$$

Then

$$\mathop{\mathbb{P}}_{G,\tau}\left\{\left|S_4(G,x) - \sum_{s\in[k]} \frac{\langle v_s,x\rangle^4}{\|v_s\|^4}\right| > n^{-\Omega(1)}\right\} \leqslant \exp(-n^{\Omega(1)}) .$$

Lemma 5.9. *Under the same hypotheses as Lemma 5.8, there are $S_3(G,x)$, $S_4(G,x)$, polynomials of degree 3 and 4, respectively, in x and in the edge indicators of G, such that*

$$\mathop{\mathbb{P}}_{G,\tau}\left\{\left|S_4(G,x) - \sum_{s\in[k]} \frac{\langle w_s,x\rangle^4}{\|w_s\|^4}\right| > n^{-\Omega(1)}\right\} \leqslant \exp(-n^{\Omega(1)}) ,$$

and

$$\mathop{\mathbb{P}}_{G,\tau}\left\{\left|C\cdot S_3(G,x) - \sum_{s\in[k]} \frac{\langle w_s,x\rangle^3}{\|w_s\|^3}\right| > n^{-\Omega(1)}\right\} \leqslant \exp(-n^{\Omega(1)}) ,$$

where $w_1,\dots,w_k$ are the vectors $w_s(i) = v_s(i) + \frac{1}{k\sqrt{\alpha+1}}$.

Finally, all of the algorithms have a cleanup phase to transform n-length vectors to probability vectors $\tau_1,\dots,\tau_n \in \Delta_{k-1}$. The following lemma describes the guarantees of the cleanup algorithm used by the small and large δ algorithms, which takes as input vectors x correlated with the vectors w.

Lemma 5.10. *Let $\delta \in (0,1)$ and $k = k(n) \in \mathbb{N}$ and $\alpha = \alpha(n) \geqslant 0$, with $\alpha, k \leqslant n^{o(1)}$. Suppose $\delta \geqslant 1/k^{1/C}$ for a big-enough constant C. There is a $\mathrm{poly}(n)$-time algorithm with the following guarantees.*

Let $\sigma_1,\dots,\sigma_n$ be iid draws from the α,k Dirichlet distribution. Let $v_1,\dots,v_k \in \mathbb{R}^n$ be the vectors given by $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $w_1,\dots,w_k \in \mathbb{R}^n$ be the vectors given by $w_s(i) = v_s(i) + \frac{1}{k\sqrt{\alpha+1}}$, so that $\mathbb{E}\langle w_s,w_t\rangle = 0$ for $s \neq t$. Let $M = \sum_s w_s w_s^\mathsf{T}$. Let E be the event that

1. $\left\|M^{-1/2}w_s - \frac{w_s}{(\mathbb{E}\,\|w_s\|^2)^{1/2}}\right\| \leqslant \frac{1}{\mathrm{poly}\, n}$ *for every $s \in [k]$.*
2. $\|w_s\| = (1 \pm 1/\mathrm{poly}(n))(\mathbb{E}\,\|w_s\|^2)^{1/2}$ *for every $s \in [k]$.*
3. $\|v_s\| = (1 \pm 1/\mathrm{poly}(n))(\mathbb{E}\,\|v_s\|^2)^{1/2}$ *for every $s \in [k]$.*

Suppose $x_1,\dots,x_k \in \mathbb{R}^n$ are unit vectors such that for at least δk vectors $w_1,\dots,w_m$ there exists $t \in [k]$ such that $\langle w_s,x_t\rangle \geqslant \delta\|w_s\|$.

The algorithm takes input $x_1,\dots,x_k$ and when E happens returns probability vectors $\tau_1,\dots,\tau_n \in \Delta_{k-1}$ such that

$$\mathrm{corr}(\sigma,\tau) \geqslant \delta^{O(1)}\,\mathbb{E}\,\|v\|^2 = \delta^{O(1)}\left(\frac{1}{\alpha+1}\cdot\frac{k+\alpha}{k} - \frac{1}{k}\right) .$$

Finally, the last lemma captures the cleanup algorithm used by the tiny-δ algorithm, which takes a single vector x correlated with v_1.

Lemma 5.11. *Let $\delta \in (0,1)$ and $k = k(n) \in \mathbb{N}$ and $\alpha = \alpha(n) \geqslant 0$, with $\alpha, k \leqslant n^{o(1)}$. Suppose $\delta \leqslant k^{1/C}$ for any constant C. There is a $\mathrm{poly}(n)$-time algorithm with the following guarantees.*

Let $\sigma_1, \dots, \sigma_n$ be iid draws from the α, k Dirichlet distribution. Let $v_1, \dots, v_k \in \mathbb{R}^n$ be the vectors given by $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $x \in \mathbb{R}^n$ be a unit vector satisfying $\langle x, v_s \rangle \geqslant \delta \|v_s\|$ for some $s \in [k]$. On input x, the algorithm produces $\tau_1, \dots, \tau_n \in \Delta_{k-1}$ such that

$$\mathrm{corr}(\sigma, \tau) \geqslant \left(\frac{\delta}{k}\right)^{O(1)} \cdot \mathbb{E}\,\|v\|^2 = \delta^{O(1)} \left(\frac{1}{\alpha+1} \cdot \frac{k+\alpha}{k} - \frac{1}{k}\right).$$

so long as the event E from Lemma 5.10 occurs.

Analysis for tiny δ (Algorithm 5.3)

Proof of Theorem 5.1, tiny-δ case. Let $C \in \mathbb{N}$ and $1 \geqslant \delta > 0$ be any fixed constants. We will prove that if $k \leqslant \delta^C$ then the output of Algorithm 5.2 satisfies the conclusion of Theorem 5.1. Let $x \in \mathbb{R}^{(1-\eta)n}$ be the output of the matrix estimation algorithm of Theorem 4.3. By Markov's inequality, with probability $(\delta/k)^{O(1)}$ over G and $\sigma_1, \dots, \sigma_{(1-\eta)n}$, the vector x satisfies $\langle v, x \rangle^2 \geqslant (\delta/k)^{O(1)} \|v\|^2 \|x\|^2$, where $v \in \mathbb{R}^{(1-\eta)n}$ is the vector $v(i) = \sigma_i(1) - \frac{1}{k}$. By our assumption $k \leqslant \delta^C$, this means that with probability $\delta^{O(1)}$ the vector x satisfies $\langle x, v \rangle^2 \geqslant \delta^{O(1)} \|x\|^2 \|v\|^2$.

Now, the labels $\sigma_{(1-\eta)n}, \dots, \sigma_n$ and the edges from nodes $1, \dots, (1-\eta)n$ to nodes $(1-\eta)n, \dots, n$ are independent of everything above. So, invoking Lemma 5.8, we can assume that the quantity $s_x^{(4)}$ computed by Algorithm 5.3 satisfies

$$\left| s_x^{(4)} - \sum_{s \in [k]} \frac{\langle v_s, x \rangle^4}{\|v_s\|^4} \right| \leqslant n^{-\Omega(1)}.$$

Now, if x satisfies $\langle v_s, x \rangle^2 \geqslant \delta^{O(1)} \|v_s\|^2$ for some v_s, then also $s_x^{(4)} \geqslant \delta^{O(1)}$. On the other hand, if $s_x^{(4)} \geqslant \delta^{O(1)}$ then there is some s such that $\langle x, v_s \rangle^2 \geqslant \frac{\delta^{O(1)}}{k} \|v_s\|^2$. So choosing the threshold C in Algorithm 5.3 appropriately, we have obtained that with probability $\delta^{O(1)}$ the algorithm reaches step 3 with a vector x which satisfies $\langle x, v_s \rangle^2 \geqslant \delta^{O(1)} \|v_s\|^2$, and otherwise the algorithm outputs random $\tau_1, \dots, \tau_n$.

Step 3 is designed to check the sign of $\langle x, v_s \rangle$. Call x good if there is $s \in [k]$ such that $\langle x, v_s \rangle \geqslant \delta^{O(1)} \|v_s\|$. If $|s_x^{(3)}| \leqslant \delta^{O(1)}$ then x there are v_s, v_t such that $\langle v_s, x \rangle \geqslant \delta^{O(1)} \|v_s\|$ and $\langle v_t, x \rangle \leqslant -\delta^{O(1)} \|v_t\|$, so both x and $-x$ are good If $|s_x^{(3)}| > \delta^{O(1)}$ then clearly step 3 outputs a good vector. Since after step 3 the vector x is good, applying Lemma 5.11 finishes the proof in the tiny δ case. □

Analysis for small and large δ (Algorithm 5.4, Algorithm 5.5)

Proof of Theorem 5.1, small δ case. Let $n_0 = (1-\eta)n$ and $n_1 = \eta n$ with η as in Algorithm 5.2.

By Markov's inequality applied to Lemma 5.6, with probability $\delta^{O(1)}$ the tensor P satisfies $\langle P, W\rangle \geqslant \delta^{O(1)}\|P\|\|W\|$, where $W \in (\mathbb{R}^{n_0})^{\otimes 3}$ is as in Lemma 5.6. Let $M = \sum_{s\in[k]} w_s w_s^\top$, where w_s is as in Lemma 5.6. The vectors $M^{-1/2}w_s$ are orthonormal, and Lemma 5.7 guarantees that $\|\frac{w_s}{\|w_s\|} - M^{-1/2}w_s\| \leqslant n^{-\Omega(1)}$ with high probability. Let $W' = \sum_{s\in[k]}(M^{-1/2}w_s)^{\otimes 3}$ and let $W'_4 = \sum_{s\in[k]}(M^{-1/2}w_s)^{\otimes 4}$. Then also $\langle P, W'\rangle \geqslant \delta^{O(1)}\|P\|\|W'\|$. By the guarantees of the 3-to-4 lifting algorithm (Theorem 7.14), finally we get $\langle T, W'_4\rangle \geqslant \delta^{O(1)}\|T\|\|W'_4\|$.

In order to conclude that Algorithm 5.4 successfully runs the low-correlation tensor decomposition algorithm, we have to check correctness of its implementation of the cross-validation oracle. This follows from Lemma 5.7, Lemma 5.9, the size of η, and a union bound over the $\exp(k/\mathrm{poly}(\delta)) \leqslant \exp(n^{o(1)})$ queries made by the nonadaptive implementation of the low-correlation tensor decomposition algorithm, and independence of the randomness in the holdout set.

We conclude that with probability at least $\delta^{O(1)}$, the tensor decomposition algorithm returns unit vectors $x_1, \dots, x_k \in \mathbb{R}^{n_0}$ such that a $\delta^{O(1)}$ fraction of w_s among $w_1, \dots, w_k$ have x_t such that $\langle w_s, x_t\rangle^2 \geqslant \delta^{O(1)}\|w_s\|^2$. By the same reasoning as in the tiny δ case, using Lemma 5.9 after the sign-checking step the same guarantee holds with the strengthened conclusion $\langle w_s, x_t\rangle \geqslant \delta^{O(1)}\|w_s\|$. Finally, we apply Lemma 5.10 (along with elementary concentration arguments to show that the event E occurs with high probability) to conclude that the last step of Algorithm 5.4 gives $\tau_1, \dots, \tau_n$ such that (in expectation) $\mathrm{corr}(\sigma, \tau) \geqslant \delta^{O(1)}\left(\frac{1}{\alpha+1}\cdot\frac{k}{k+\alpha} - \frac{1}{k}\right)$ as desired. □

5.2 Low-degree estimate for posterior third moment

In this section we prove Lemma 5.6. The strategy is to apply Lemma 2.1 to find an estimator for the 3-tensor $\sum_{s\in[k]} v_s^{\otimes 3}$. With that in hand, combining with the estimators in Section 4 for the second moments $\sum_{s\in[k]} v_s v_s^\top$ is enough to obtain an estimator for W, since

$$\sum_{s\in[k]} w_s^{\otimes 3} = \sum_{s\in[k]} (v_s + c\cdot 1)^{\otimes 3} \tag{5.1}$$

$$= \sum_{s\in[k]} v_s^{\otimes 3} + c(v_s\otimes v_s\otimes 1 + v_s\otimes 1\otimes v_s + 1\otimes v_s\otimes v_s) + c^3\cdot 1^{\otimes 3} \tag{5.2}$$

where 1 is the all-1s vector, $c = \frac{1}{k\sqrt{\alpha+1}}$, and we have used that $\sum_{s\in[k]} v_s = 0$. Thus if R is a degree $n^{\mathrm{poly}(1/\delta)}$ polynomial such that

$$\langle R, \sum_{s\in[k]} v_s^{\otimes 3}\rangle \geqslant \delta^{O(1)}(\mathbb{E}\,\|R\|^2)^{1/2}(\mathbb{E}\left\|\sum_{s\in[k]} v_s^{\otimes 3}\right\|^2)^{1/2}$$

and Q is similar but estimates $\sum_{s\in[k]} v_s v_s^\top$, then R and Q can be combined according to (5.2) to obtain the polynomial P from the lemma statement.

Thus in the remainder of this section we focus on obtaining such a polynomial R; we change notation to call this polynomial P. The first step will be to define a collection of polynomials $\{G^\alpha\}_\alpha$ for all distinct $i, j, k \in [n]$.

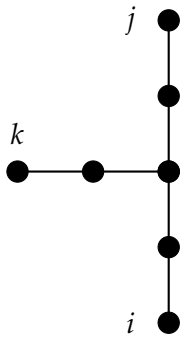

Figure 1: A
3-armed star with arms of length 2. We will eventually use arms of length $t \approx \log n$.

Definition 5.12. Any $\alpha \subseteq \binom{n}{2}$ can be interpreted as a graph on some nodes in $[n]$. Such an α is a long-armed star if it consists of three self-avoiding paths, each with ℓ edges, joined at one end at a single central vertex, at the other end terminating at distinct nodes $i, j, k \in [n]$. (See figure.) Let $\mathrm{STAR}_\ell(i, j, k)$ be the set of 3-armed stars with arms of length ℓ and terminal vertices i, j, k. For any $\alpha \subseteq \binom{n}{2}$ let $G^\alpha = \prod_{ab \in \alpha}(x_{ab} - \frac{d}{n})$ be the product of centered edge indicators.

The next two lemmas check the conditions to apply Lemma 2.1 to the sets $\{G^\alpha\}_{\alpha \in \mathrm{STAR}_\ell(i,j,k)}$.

Lemma 5.13 (Unbiased Estimator)**.** *Let $i, j, k \in [n]$ all be distinct. Let $\alpha \in \mathrm{STAR}_\ell(i, j, k)$.*

For a collection of probability vectors $\sigma_1, \dots, \sigma_k$, let $V(\sigma) = \sum_{s \in [k]} v_s^{\otimes 3}$ where $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $G \sim G(n, d, \varepsilon, \alpha_0, k)$.

$$\mathbb{E}\left[G^\alpha \mid \sigma_i, \sigma_j, \sigma_k\right] = \left(\frac{\varepsilon d}{n}\right)^{3\ell} \left(\frac{1}{k(\alpha_0 + 1)}\right)^{3(\ell-1)} \cdot C_3 \cdot V(\sigma)_{ijk}\,.$$

Here $\alpha_0 \geqslant 0$ is the Dirichlet concentration paramter, unrelated to the graph α, and $C_3 = 1/(k^{O(1)}\alpha_0^{O(1)})$ is a constant related to third moments of the Dirichlet distribution.

Lemma 5.14 (Approximate conditional independence)**.** *If*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha_0 + 1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad k, \alpha_0 \leqslant n^{o(1)} \text{ and } \varepsilon^2 d \leqslant n^{o(1)}\,.$$

and $\ell \geqslant C \log n / \delta^{O(1)}$ for a large enough constant C, then for $G \sim G(n, d, \varepsilon, k, \alpha_0)$,

$$\mathbb{E}\left[V(\sigma)_{ijk}^2\right] \cdot \sum_{\alpha,\beta \in \mathrm{STAR}_\ell(i,j,k)} \mathbb{E}\, G^\alpha G^\beta \leqslant 1/\delta^{O(1)} \cdot \sum_{\alpha,\beta \in \mathrm{STAR}_\ell(i,j,k)} \mathbb{E}\left[G^\alpha V(\sigma)_{i,j,k}\right] \cdot \mathbb{E}\left[G^\beta V(\sigma)_{i,j,k}\right]\,.$$

Now we can prove Lemma 5.6.

Proof of Lemma 5.6. As discussed at the beginning of this section, it is enough to find an estimator for the tensor $V(\sigma)$. Lemma 5.13 and Lemma 5.14 show that Lemma 2.1 applies to each set of

polynomials $\mathrm{STAR}_\ell(i,j,k)$. The conclusion is that for every distinct $i,j,k \in [n]$ there is a degree $\log n \operatorname{poly}(1/\delta)$ polynomial $P(G)_{ijk}$ so that

$$\frac{\mathbb{E}\, P(G)_{ijk} V(\sigma)_{ijk}}{(\mathbb{E}\, P(G)_{ijk}^2)^{1/2} \cdot (\mathbb{E}\, V(\sigma)_{ijk}^2)^{1/2}} \geqslant \Omega(1)\,.$$

One may check that the entries i,j,k for i,j,k all distinct of the tensor $V(\sigma)$ comprise nearly all of its 2-norm. That is,

$$\sum_{i,j,k \text{ distinct}} \mathbb{E}\, V(\sigma)_{i,j,k}^2 \geqslant (1-o(1))\, \mathbb{E}\, \|V(\sigma)\|^2\,.$$

This is sufficient to conclude that the tensor-valued polynomial $P(G)$ whose (i,j,k)-th entry is $P_{i,j,k}(G)$ when i,j,k are all distinct and is 0 otherwise is a good estimator of $V(\sigma)$ (see Fact A.2). Thus,

$$\frac{\mathbb{E}_{\sigma,G} \langle P(G), V(\sigma)\rangle}{\left(\mathbb{E}_{\sigma,G} \|P(G)\|^2\right)^{1/2} \cdot \left(\mathbb{E}_{\sigma,G} \|V(\sigma)\|^2\right)^{1/2}} \geqslant \Omega(1)\,. \qquad \square$$

5.2.1 Details of unbiased estimator

We work towards proving Lemma 5.13. We will need to assemble a few facts. The first will help us control moment tensors of the Dirichlet distribution. The proof can be found in the appendix.

Fact 5.15 (Special case of Fact A.3)**.** *Let σ be distributed according to the α, k Dirichlet distribution. Let $\tilde{\sigma} = \sigma - \frac{1}{k}1$. There are numbers C_2, C_3 depending on α, k so that for every x_1, x_2, x_3 in $\mathbb{R}^k$ with $\sum_{s\in[k]} x_i(s) = 0$,*

$$\mathop{\mathbb{E}}_{\sigma} \langle \tilde{\sigma}, x_1\rangle\langle \tilde{\sigma}, x_2\rangle = C_2 \langle x_1, x_2\rangle$$

and

$$\mathop{\mathbb{E}}_{\sigma} \langle \tilde{\sigma}, x_1\rangle\langle \tilde{\sigma}, x_2\rangle\langle \tilde{\sigma}, x_3\rangle = C_3 \sum_{s\in[k]} x_1(s)x_2(s)x_3(s)\,.$$

Furthermore,

$$C_2 = \frac{1}{k(\alpha+1)} \quad \text{and} \quad C_3 = \frac{1}{k^{O(1)}\alpha^{O(1)}}\,.$$

Now we can prove Lemma 5.13.

Proof of Lemma 5.13. For any collection of σ's and $\alpha \in \mathrm{STAR}_\ell(i,j,k)$,

$$\mathop{\mathbb{E}}_{G}[G^\alpha \mid \sigma] = \left(\frac{\varepsilon d}{n}\right)^{3\ell} \prod_{(a,b)\in\alpha} \langle \tilde{\sigma}_a, \tilde{\sigma}_b\rangle$$

Let a be the central vertex of the star α. Taking expectations over all the vertices in the arms of the star,

$$\mathbb{E}\left[G^\alpha \mid \sigma_i, \sigma_j, \sigma_k\right] = \left(\frac{\varepsilon d}{n}\right)^{3\ell} \left(\frac{1}{k(\alpha_0+1)}\right)^{3(\ell-1)} \mathop{\mathbb{E}}_{\sigma_a} \langle \tilde{\sigma}_i, \tilde{\sigma}_a\rangle\langle \tilde{\sigma}_j, \tilde{\sigma}_a\rangle\langle \tilde{\sigma}_k, \tilde{\sigma}_a\rangle\,.$$

Finally, using the second part of Fact 5.15 completes the proof. $\square$

5.2.2 Details of approximate conditional independence

We prove Lemma 5.14, first gathering some facts. In the sum $\sum_{\alpha,\beta\in \mathrm{STAR}_\ell(i,j,k)} G^\alpha G^\beta$, the terms α, β which (as graphs) share only the vertices i, j, k will not cause us any trouble, because such G^α and G^β are independent conditioned on $\sigma_i, \sigma_j, \sigma_k$.

Fact 5.16. *If $\alpha, \beta \in \mathrm{STAR}_\ell(i, j, k)$ share only the vertices i, j, k, then for any collection σ of probability vectors,*

$$\mathbb{E}\left[G^\alpha G^\beta \mid \sigma_i, \sigma_j, \sigma_k\right] = \mathbb{E}\left[G^\alpha \mid \sigma_i, \sigma_j, \sigma_k\right] \cdot \mathbb{E}\left[G^\beta \mid \sigma_i, \sigma_j, \sigma_k\right] .$$

Proof. To sample G^α, one needs to know σ_a for any $a \in [n]$ with nonzero degree in α, and similar for $b \in [n]$ and G^β. The only overlap is $\sigma_i, \sigma_j, \sigma_k$. □

The next fact is the key one. Pairs α, β which share vertices forming paths originating at i, j, and k make the next-largest contribution (after α, β sharing only i, j, k) to $\sum_{\alpha,\beta} \mathbb{E}\, G^\alpha G^\beta$.

Fact 5.17. *Let $i, j, k \in [n]$ be distinct. Let $V(\sigma)_{ijk}$ be as in the Lemma 5.14. Let $C_2 \in \mathbb{R}$ be as in Fact 5.15.*

Let $\alpha, \beta \in \mathrm{STAR}_\ell(i, j, k)$ share s vertices (in addition to i, j, k) for some $s \leqslant \frac{t}{2}$, and suppose the shared vertices form paths in α and β starting at i, j, and k. Then

$$\mathbb{E}\, V(\sigma)_{ijk}^2 \cdot \mathbb{E}\, G^\alpha G^\beta \leqslant \varepsilon^{-2s} \left(\frac{d}{n}\right)^{-s} (1 + O(d/n))^{-s} \cdot \left(\frac{1}{k(\alpha_0 + 1)}\right)^{-2s} \cdot \mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right] \cdot \mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right] .$$

Proof. Let $\sigma_{\alpha\cap\beta}$ be the σ's corresponding to vertices sharerd by α, β. Let i', j', k' be the last shared vertices along the paths beginning at i, j, k respectively. We expand $G^\alpha G^\beta$ and use conditional independence of the G_e's given the σ's:

$$\mathbb{E}\, G^\alpha G^\beta = \mathop{\mathbb{E}}_{\sigma_{i',j',k'}} \left[\mathbb{E}\left[(G^{\alpha\cap\beta})^2 \,|\sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] \cdot \mathbb{E}\left[G^{\alpha\setminus\beta} \mid \sigma_{i'}\sigma_{j'}\sigma_{k'}\right] \cdot \mathbb{E}\left[G^{\beta\setminus\alpha} \mid \sigma_{i'}\sigma_{j'}\sigma_{k'}\right]\right] .$$

Both $G^{\alpha\setminus\beta}$ and $G^{\beta\setminus\alpha}$ are long-armed stars with terminal vertices i', j', k'. The arm lengths of $G^{\alpha\setminus\beta}$ total $3\ell - s$. By a similar argument to Lemma 5.13, $G^{\alpha\setminus\beta}$ is an unbiased estimator of $V(\sigma)_{i'j'k'}$ with

$$\mathbb{E}\left[G^{\alpha\setminus\beta} \mid \sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] = \left(\frac{\varepsilon d}{n}\right)^{3\ell-s} \left(\frac{1}{k(\alpha_0 + 1)}\right)^{3(\ell-1)-s} \cdot C_3 \cdot V(\sigma)_{i',j',k'}$$

and the same goes for $G^{\beta\setminus\alpha}$. Furthermore,

$$\mathbb{E}\left[(G^{\alpha\cap\beta})^2 \mid \sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] = \left(\frac{d}{n}\right)^{|\alpha\cap\beta|} \mathbb{E}\left[\prod_{(a,b)\in\alpha\cap\beta} (1 + \varepsilon\langle\tilde{\sigma}_a, \tilde{\sigma}_b\rangle + O(d/n)) \,\middle|\, \sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] .$$

By our assumption that $\alpha \cap \beta$ consists just of paths, every subset of edges in the graph $\alpha \cap \beta$ contains a vertex of degree 1. Hence, $\mathbb{E}\left[(G^{\alpha\cap\beta})^2 \mid \sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] = (1 + O(d/n))^{|\alpha\cap\beta|}(d/n)^{|\alpha\cap\beta|}$. Putting these together,

$$\mathbb{E}\, G^\alpha G^\beta = (1 + O(d/n))^s\, \varepsilon^{6\ell-2s} \left(\frac{d}{n}\right)^{6\ell-s} \left(\frac{1}{k(\alpha_0 + 1)}\right)^{6(\ell-1)-2s} C_3^2\, \mathbb{E}\, V(\sigma)_{ijk}^2$$

At the same time, one may apply Lemma 5.13 to $\mathbb{E}\, G^\alpha V(\sigma)_{ijk}$ to obtain

$$\mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right]\cdot\mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right]=\left(\frac{\varepsilon d}{n}\right)^{6\ell}\left(\frac{1}{k(\alpha_0+1)}\right)^{6(\ell-1)}C_3^2\cdot\left(\mathop{\mathbb{E}}_{\sigma_i,\sigma_j,\sigma_k}V(\sigma)_{ijk}^2\right)^2.$$

The lemma follows. □

The last fact will allow us to control α,β which intersect in some way other than paths starting at i,j,k. The key idea will be that such pairs α,β must share more vertices than they do edges.

Fact 5.18. *Let $i,j,k\in[n]$ be distinct. Let $V(\sigma)_{ijk}$ be as in the Lemma 5.14. Let $C_2\in\mathbb{R}$ be as in Fact 5.15.* $C_2=\frac{1}{k(\alpha_0+1)}$.

Let $\alpha,\beta\in\mathrm{STAR}_\ell(i,j,k)$ share s vertices (in addition to i,j,k) and r edges. Then

$$\mathbb{E}\, V(\sigma)_{ijk}^2\cdot\mathbb{E}\, G^\alpha G^\beta\leqslant\varepsilon^{-2r}\left(\frac{d}{n}\right)^{-r}\cdot C_2^{-2s}\cdot k^{O(s-r)}(1+\alpha_0)^{O(s-r)}\cdot\mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right]\cdot\mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right].$$

Proof. Expanding as usual,

$$\mathbb{E}\, G^\alpha G^\beta=\left(\frac{d}{n}\right)^{6\ell-r}\mathop{\mathbb{E}}_{\sigma}\prod_{ab\in\alpha\triangle\beta}\langle\tilde\sigma_a,\tilde\sigma_b\rangle\cdot\prod_{ab\in\alpha\cap\beta}(1+\varepsilon\langle\tilde\sigma_a,\tilde\sigma_b\rangle+O(d/n)).$$

Any nontrivial edge-induced subgraph of $\alpha\cap\beta$ contains a degree-1 vertex; using this to expand the second product and simplifying with $\mathbb{E}\,\tilde\sigma_a=0$, the above is

$$\left(\frac{d}{n}\right)^{6\ell-r}\mathop{\mathbb{E}}_{\sigma}\prod_{ab\in\alpha\triangle\beta}\langle\tilde\sigma_a,\tilde\sigma_b\rangle\cdot(1+O(d/n))^r.$$

For every degree-2 vertex in $\alpha\triangle\beta$ we can use Fact A.3 to take the expectation. Each such vertex contributes a factor of C_2 and there are at least $3\ell-O(s-r)$ such vertices. The remaining expression will be bounded by 1. The fact follows. □

Now we can prove Lemma 5.14.

Proof of Lemma 5.14. Let us recall that our goal is to show

$$\mathbb{E}\left[V(\sigma)_{ijk}^2\right]\cdot\sum_{\alpha,\beta\in\mathrm{STAR}_\ell(i,j,k)}\mathbb{E}\, G^\alpha G^\beta\leqslant\delta^{O(1)}\cdot\sum_{\alpha,\beta\in\mathrm{STAR}_\ell(i,j,k)}\mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right]\cdot\mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right]$$

where $\delta=1-\frac{k^2(\alpha_0+1)^2}{\varepsilon^2 d}$. Let $c=\mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right]\cdot\mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right]$. (Notice this number does not depend on α or β.) The right-hand side above simplifies to $|\mathrm{STAR}_\ell(i,j,k)|^2\cdot c$.

On the left-hand side, what is the contribution from α,β sharing s vertices? First consider what happens with $s\leqslant t/2$ and the intersecting vertices form paths in α and β starting at i,j,k. Choosing a random pair α,β from $\mathrm{STAR}_\ell(i,j,k)$, the probability that they intersect along paths of length s_1,s_2,s_3 starting at i,j,k respectively is at most $n^{-s_1-s_2-s_3}$. There are at most $(1+s^2)$ choices for

nonnegative integers s_1, s_2, s_3 with $s_1 + s_2 + s_3 = s$. By Fact 5.17, such terms therefore contribute at most

$$c \cdot \frac{|\mathrm{STAR}_\ell(i,j,k)|^2}{n^{-s}} \cdot \left(\varepsilon\sqrt{\tfrac{d}{n}}(1+O(d/n))\right)^{-2s} C_2^{-2s} \cdot s^2 = c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 \cdot (\varepsilon^2 d C_2^2(1+O(d/n)))^{-s} \cdot s^2$$

where $C_2 = \frac{1}{k(\alpha_0+1)}$. By hypothesis, $\delta > 0$. Consider the sum of all such contributions for $s \leqslant t/2$; this is at most

$$c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 \cdot \sum_{s=0}^{t/2}(1+s^2) \cdot \left(\tfrac{k^2(\alpha_0+1)^2}{\varepsilon^2 d}\right)^s \leqslant \delta^{O(1)} \cdot c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 .$$

Next, consider the contribution from α, β which share s vertices in some pattern other than those considered above. Unless $\alpha = \beta$, this means α, β share at least one more vertex than the number r of edges that they share. Suppose $\alpha \neq \beta$ and let $s - r = q$. There are $t^{O(q)}$ patterns in which such an intersection might occur, and each occurs for a random pair $\alpha, \beta \in \mathrm{STAR}_\ell(i,j,k)$ with probabilty n^{-s}. So using Fact 5.18, the contribution is at most

$$c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 \cdot \sum_{q=1}^{t}\left(\frac{\varepsilon^2 d}{n}\right)^q \cdot k^{O(q)}(1+\alpha_0)^{O(q)} t^{O(q)}$$

By the hypotheses $k, \alpha = n^{o(1)}$ and $\varepsilon^2 d = n^{1-\Omega(1)}$, this is all $o(c|\mathrm{STAR}_\ell(i,j,k)|^2)$.

Finally, consider the case $\alpha = \beta$. Then, using Fact 5.18 again, the contribution is at most

$$c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 \left(\frac{\varepsilon^2 d}{k^2(\alpha_0+1)^2}\right)^{-t} k^{O(1)}\alpha^{O(1)}$$

which is $o(c|\mathrm{STAR}_\ell(i,j,k)|^2)$ because $t \gg \log(n)$. Putting these things together gives the lemma. □

5.3 Cross validation

In this section we show how to use a holdout set of vertices to cross-validate candidate community membership vectors. The arguments are all standard, using straightforward concentration inequalities. At the end we prove the first part of Lemma 5.8, on the estimator S_3. The proof of the second part, on S_4 is similar, using standard facts about moments of the Dirichlet distribution (see Fact A.3). The proof of Lemma 5.9 is also similar, using the discussion in Section 5.2 to turn estimators for moments of the v vectors into estimators for moments of the w vectors—we leave it to the reader.

We will need a few facts to prove the lemma.

Fact 5.19. *Let $n_0, n_1, A, k, d, \varepsilon, \alpha, \sigma, v, \tau, G, x, P$ be as in Lemma 5.8. Let $a \in A$. There is a number $C = C(k,\alpha) \leqslant \mathrm{poly}(k,\alpha)$ such that*

$$\mathop{\mathbb{E}}_{G,\tau} P_a(G,x) = \left(\frac{\varepsilon d}{n}\right)^3 \cdot C \cdot \sum_{ijk \in \overline{A} \text{ distinct}} \sum_{s\in[k]} \sigma_i(s)\sigma_j(s)\sigma_k(s) x_i x_j x_k .$$

Proof. Immediate from Fact 5.15. □

Fact 5.20. *Let $n_0, n_1, A, k, d, \varepsilon, \alpha, \sigma, v, \tau, G, x, P$ be as in Lemma 5.8. Let $a \in A$. The following variance bound holds.*

$$\mathop{\mathbb{E}}_{G,\tau} P_a(G,x)^2 - \left(\mathop{\mathbb{E}}_{G,\tau} P_a(G,x)\right)^2 \leqslant \frac{\mathrm{poly}(k,\alpha,\varepsilon,d)}{n^3}\,.$$

Proof. Expanding $P_a(G,x)$ and using that $|\langle \sigma, \sigma'\rangle| \leqslant 1$ for any $\sigma, \sigma' \in \Delta_{k-1}$ we get

$$\mathop{\mathbb{E}}_{G,\tau} P_a(G,x)^2 \leqslant \left(\frac{d}{n}\right)^6 \sum_{\substack{ijk \text{ distinct} \\ i'j'k' \text{ distinct}}} \left|x_i x_j x_k x_{i'} x_{j'} x_{k'}\right| \leqslant \left(\frac{d}{n}\right)^6 \cdot n^3 \cdot \|x\|^{12}\,.$$

□

Fact 5.21. *Let $n_0, n_1, A, k, d, \varepsilon, \alpha, \sigma, v, \tau, G, x, P$ be as in Lemma 5.8. Let $a \in A$. For some constant $\gamma_*(\varepsilon, d, k, \alpha)$ and every $\gamma_* > \gamma > 0$,*

$$\mathop{\mathbb{P}}_{G,\tau}\{|P_a(G,x)| > n^\gamma\} \leqslant \exp(-n^{\Omega(\gamma)})$$

Proof. The fact follows from a standard exponential tail bound on the degree of vertex a. □

We can put these facts together to prove the S_3 portion of Lemma 5.8 (as we discussed above, the S_4 portion and Lemma 5.9 are similar). The strategy will be to use the following version of Bernstein's inequality, applied to the random variables $\langle G_a, v^{\otimes 3}\rangle$. The proof of the inequality is in the appendix.

Proposition 5.22 (Bernstein wth tails)**.** *Let X be a random variable satisfying $\mathbb{E}\,X = 0$ and, for some numbers $R, \delta, \delta' \in \mathbb{R}$,*

$$\mathbb{P}\{|X| > R\} \leqslant \delta \text{ and } \mathbb{E}\,|X| \cdot \mathbf{1}_{|X|>R} \leqslant \delta'\,.$$

Let $X_1, \ldots, X_m$ be independent realizations of X. Then

$$\mathbb{P}\left\{\left|\tfrac{1}{m}\sum_{i\leqslant m} X_i\right| \geqslant t + \delta'\right\} \leqslant \exp\left(\frac{-\Omega(1)\cdot m \cdot t^2}{\mathbb{E}\,X^2 + t\cdot R}\right) + m\delta\,.$$

Now we can prove Lemma 5.8.

Proof of Lemma 5.8. We apply Proposition 5.22 to the n_1 random variables $X_a = \left(\frac{\varepsilon d}{n}\right)^{-3} C^{-1} P_a(G,x)$ for $a \in A$, where $C = C(k,\alpha)$ is the number from Fact 5.20. (For each $a \in A$ these are iid over G, τ.) Take $t = n^{3/2-\gamma'}$ for a small-enough constant γ' so that $n_1 t^2/n^3 \geqslant n^\gamma$ for some constant γ, using the assumption $n_1 \geqslant n^{\Omega(1)}$. All together, we get

$$\mathop{\mathbb{P}}_{G,\tau}\left\{\left|\frac{1}{n_1}\sum_{a\in A} X_a - \sum_{s\in[k]} \sum_{ijk\in\overline{A} \text{ distinct}} \sigma_s(i)\sigma_s(j)\sigma_s(k) x_i x_j x_k\right| \geqslant n^{3/2-\gamma'}\right\} \leqslant \exp(n^{-\gamma'})$$

for some constants γ, γ' (possibly different from γ, γ' above) and large-enough n. For any unit $x \in \mathbb{R}^{n_0}$ and $\sigma \in \Delta_{k-1}^{n_0}$, using that $k \leqslant n^{o(1)}$ it is not hard to show via Cauchy-Schwarz that

$$\left| \sum_{s \in [k]} \langle v_s, x \rangle^3 - \sum_{s \in [k]} \sum_{ijk \in \overline{A} \text{ distinct}} \sigma_s(i)\sigma_s(j)\sigma_s(k) x_i x_j x_k \right| \leqslant n^{1+o(1)} .$$

The lemma follows. □

5.4 Producing probability vectors

In this section we prove Lemma 5.10. The proof of Lemma 5.11 is very similar (in fact it is somewhat easier) so we leave it to the reader.

Lemma (Restatement of Theorem 5.10)**.** *Let $\delta \in (0,1)$ and $k = k(n) \in \mathbb{N}$ and $\alpha = \alpha(n) \geqslant 0$, with $\alpha, k \leqslant n^{o(1)}$. Suppose $\delta \geqslant 1/k^{1/C}$ for a big-enough constant C. There is a* poly(n)*-time algorithm with the following guarantees.*

Let $\sigma_1, \dots, \sigma_n$ be iid draws from the α, k Dirichlet distribution. Let $v_1, \dots, v_k \in \mathbb{R}^n$ be the vectors given by $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $w_1, \dots, w_k \in \mathbb{R}^n$ be the vectors given by $w_s(i) = v_s(i) + \frac{1}{k\sqrt{\alpha+1}}$, so that $\mathbb{E}\langle w_s, w_t \rangle = 0$ for $s \neq t$. Let $M = \sum_s w_s w_s^{\mathsf{T}}$. Let E be the event that

1. $\left\| M^{-1/2} w_s - \frac{w_s}{(\mathbb{E}\,\|w_s\|^2)^{1/2}} \right\| \leqslant \frac{1}{\text{poly}\, n}$ *for every $s \in [k]$.*
2. $\|w_s\| = (1 \pm 1/\text{poly}(n))(\mathbb{E}\,\|w_s\|^2)^{1/2}$ *for every $s \in [k]$.*
3. $\|v_s\| = (1 \pm 1/\text{poly}(n))(\mathbb{E}\,\|v_s\|^2)^{1/2}$ *for every $s \in [k]$.*

Suppose $x_1, \dots, x_k \in \mathbb{R}^n$ are unit vectors such that for at least δk vectors $w_1, \dots, w_m$ there exists $t \in [k]$ such that $\langle w_s, x_t \rangle \geqslant \delta \|w_s\|$.

The algorithm takes input $x_1, \dots, x_k$ and when E happens returns probability vectors $\tau_1, \dots, \tau_n \in \Delta_{k-1}$ such that

$$\text{corr}(\sigma, \tau) \geqslant \delta^{O(1)}\, \mathbb{E}\,\|v\|^2 = \delta^{O(1)} \left(\frac{1}{\alpha+1} \cdot \frac{k+\alpha}{k} - \frac{1}{k} \right) .$$

First some preliminaries. Let $\sigma_1, \dots, \sigma_n$ be iid from the α, k Dirichlet distribution. There are two important families of vectors in $\mathbb{R}^n$. Let

$$v_s(i) = \sigma_i(s) - \frac{1}{k} \qquad w_s(i) = \sigma_i(s) - \frac{1}{k}\left(1 - \frac{1}{\sqrt{\alpha+1}}\right) .$$

We will also work with a normalized version of the v vectors:

$$\overline{v}_s = \frac{v_s}{(\mathbb{E}\,\|v_s\|^2)^{1/2}} .$$

By construction, $\mathbb{E}\,\|\overline{v}_s\|^2 = 1$. Also by definition, $\sum_s v_s = \sum_s \overline{v}_s = 0$. Thus $\mathbb{E}\langle \sum_s \overline{v}_s, \sum_s \overline{v}_s \rangle = k + \sum_{s \neq t} \mathbb{E}\langle \overline{v}_s, \overline{v}_t \rangle = 0$ and so by symmetry $\mathbb{E}\langle \overline{v}_s, \overline{v}_t \rangle = \frac{-1}{k-1}$. We let

$$\overline{w}_s = \overline{v}_s + \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{1}{k-1}}$$

so that $\mathbb{E}\langle \overline{w}_s, \overline{w}_t \rangle = 0$ for $s \neq t$. (In the facts which follow we sometimes write $\overline{v}$ as v when both normalizations are not needed; this is always noted.)

We will want the following fact; the proof is elementary.

Fact 5.23. *Let σ, u, v, w as above, and suppose y is an $n \times k$ matrix whose rows are in $\Delta_{k-1} - \frac{1}{k}$ (that is they are shifted probability vectors). Then $\tau = y + \frac{1}{k}$ is a matrix whose rows are probability vectors, and τ satisfies*

$$\langle \tau, \sigma \rangle \geqslant \langle y, v \rangle + \frac{n}{k}.$$

The following fact will be useful when δ is small but not tiny; i.e. $\delta < 1 - c$ for some fixed constant c but $\delta \gg 1/\sqrt{k}$.

Fact 5.24. *Suppose that $x_1, \ldots, x_k$ are unit vectors and $w_1, \ldots, w_k$ are orthonormal. Also suppose that there is $1 > \delta > 0$ such that for at least δk vectors w_s among $w_1, \ldots, w_k$ there exists a vector x_t among $x_1, \ldots, x_k$ such that $\langle w_s, x_t \rangle \geqslant \delta$. Then there is a permutation $\pi : [k] \to [k]$ such that if $x = (x_1, \ldots, x_k)$ is an $n \times k$ matrix and similarly for w,*

$$\langle x, \pi \cdot w \rangle \geqslant \left(\delta^5 - \frac{1}{\sqrt{k}} \left(\frac{1}{1 - \delta^4} \right)^{1/2} \right) \|x\| \|w\|,$$

where $x = (x_1, \ldots, x_k)$ is an $n \times k$ matrix and similarly for w.

Proof. We will think of π as a matching of $w_1, \ldots, w_k$ to $x_1, \ldots, x_k$. Call x_t *good* for w_s if $\langle w_s, x_t \rangle \geqslant \delta$. First of all, by orthogonality of vectors $w_1, \ldots, w_k$, any particular vector x_t is good for at most $1/\delta^2$ vectors w_s. Hence, there is a set S of $\delta^4 k$ vectors w_s such that for each w_s there exists a good x_t and all the good x_t's are distinct.

Begin by matching each $w_s \in S$ to its good x_t. Let π be the result of extending that matching randomly to a perfect matching of k to k.

We need to lower bound $\mathbb{E} \sum_{s \notin S} \langle w_s, x_{\pi(s)} \rangle$. Consider that for a particular t,

$$\mathbb{E} - \langle x_t, w_{\pi^{-1}(t)} \rangle \leqslant (\mathbb{E} \langle x_t, w_{\pi^{-1}(t)} \rangle^2)^{1/2}.$$

The distribution of $\pi^{-1}(t)$ is uniform among all $s \notin S$. So

$$\mathbb{E} \langle x_t, w_{\pi^{-1}(t)} \rangle^2 = \frac{1}{k - |S|} \sum_{s \notin S} \langle w_s, x_t \rangle^2 \leqslant \frac{1}{k} \left(\frac{1}{1 - \delta^4} \right)$$

since $\sum_{s \in [k]} \langle w_s, x_t \rangle^2 \leqslant 1$. It follows that

$$\mathbb{E} \langle x_t, w_{\pi^{-1}(t)} \rangle \geqslant -\frac{1}{\sqrt{k}} \left(\frac{1}{1 - \delta^4} \right)^{1/2}.$$

Therefore, $\mathbb{E} \sum_{s \notin S} \langle w_s, x_{\pi(s)} \rangle \geqslant -\sqrt{k} \left(\frac{1}{1-\delta^4} \right)^{1/2}$. Thus there is some choice of π such that $\sum_{s \notin S} \langle w_s, x_{\pi(s)} \rangle \geqslant -\sqrt{k} \left(\frac{1}{1-\delta^4} \right)^{1/2}$. Hence for this π one gets

$$\sum_{s \in [k]} \langle w_s, x_{\pi(s)} \rangle \geqslant \delta^5 k - \sqrt{k} \left(\frac{1}{1 - \delta^4} \right)^{1/2} = \left(\delta^5 - \frac{1}{\sqrt{k}} \left(\frac{1}{1 - \delta^4} \right)^{1/2} \right) \|x\| \|w\|. \qquad \square$$

The next fact serves the same purpose as the previous one but in the large δ case (i.e. δ close to 1).

Fact 5.25. *Under the same hypotheses as Fact 5.24, letting $\delta = 1 - \varepsilon$ for some $\varepsilon > 0$, there is a permutation $\pi : [k] \to [k]$ such that $\langle x, \pi \cdot w \rangle \geqslant (1 - 9\varepsilon)\|x\|\|w\|$.*

Proof. As in the proof of Fact 5.24, we construct a matching π by first matching a set S of at least $\delta^4 k \geqslant (1 - 4\varepsilon)k$ vectors w_s to corresponding x_t. Then we match the remaining vectors arbitrarily. For any s, t we know $\langle w_s, x_t \rangle \geqslant -1$. So the result is

$$\langle x, \pi \cdot w \rangle \geqslant (1 - 5\varepsilon)k - 4\varepsilon k = (1 - 9\varepsilon)k = (1 - 9\varepsilon)\|x\|\|w\| \,. \qquad \square$$

We will also want a way to translate a matrix correlated with w to one correlated with v, so that we can apply Fact 5.23.

Fact 5.26. *Suppose v is an $n \times k$ matrix whose rows are centered probability vectors and $w = v + c$ is a coordinate-wise additive shift of v. Suppose y is also an $n \times k$ matrix whose rows are centered probability vectors shifted by c in each coordinate (so $y - c$ is a matrix of centered probability vectors). Then the shifted matrix $y - c$ satisfies*

$$\langle y - c, v \rangle \geqslant \langle y, w \rangle - c^2 nk \,.$$

Proof. By definition, $\langle y - c, v \rangle = \langle y, v \rangle$. Since $v = w - c$, we get

$$\langle y - c, v \rangle = \langle y, v \rangle = \langle y, w \rangle - c\langle y, 1 \rangle = \langle y, w \rangle - c^2 nk \,. \qquad \square$$

Proof of Lemma 5.10. First assume $\delta < 1 - c$ for any small constant c. Let π be the permutation guaranteed by Fact 5.24 applied to the vectors $x_1, \ldots, x_k$ and $M^{-1/2}w_1, \ldots, M^{-1/2}w_k$. (Without loss of generality reorder the vectors so that π is the identity permutation.) Since $1 - c \geqslant \delta \geqslant 1/k^{1/C}$ for big-enough C and small-enough c (which are independent of n, k) and the guarantee of Fact 5.24, by event E we get that

$$\langle x, w \rangle \geqslant \delta^{O(1)}\|x\|\|w\| \,.$$

So by taking a correlation-preserving projection of x into the set of matrices whose rows are shifted probability vectors, we get a matrix y with the guarantee

$$\langle y, w \rangle \geqslant \delta^{O(1)}\|y\|\|w\| \quad \text{and} \quad \|y\| \geqslant \delta^{O(1)}\|w\| \,.$$

Applying Fact 5.26, we obtain

$$\langle y - c, v \rangle \geqslant \langle y, w \rangle - c^2 nk = \langle y, w \rangle - \frac{\mathbb{E}\,\|w\|^2}{k}$$

where $c = \frac{1}{k\sqrt{\alpha+1}}$. Putting things together and using $\mathbb{E}\,\|v\|^2 \leqslant \mathbb{E}\,\|w\|^2$ and the event E, we get

$$\langle y - c, v \rangle \geqslant \delta^{O(1)}\,\mathbb{E}\,\|v\|^2 \,.$$

So applying Fact 5.23 finishes the proof in this case.

Now suppose $\delta \geqslant 1-c$ for a small-enough constant c. Then using event E and Fact 5.25, there is π such that $\langle x, w\rangle \geqslant (1-O(c))\|x\|(\mathbb{E}\|w\|^2)$ (where again we have without loss of generality reordered the vectors so that π is the identity permutation). Now taking the Euclidean projection of $x \cdot \frac{(\mathbb{E}\|w\|^2)^{1/2}}{\|x\|}$ into the $n \times k$ matrices whose rows are centered probability vectors shifted entrywise by $c = \frac{1}{k\sqrt{\alpha+1}}$, we get a matrix y which again satisfies $\langle y, w\rangle \geqslant (1-O(c))\|y\|\|w\|$ and $\|y\| \geqslant (1-O(c))\|w\|$, so (using event E), $\langle y, w\rangle \geqslant (1-O(c))\,\mathbb{E}\|w\|^2$. Removing the contribution from $\langle y, 1\rangle$, this implies that $\langle y - c, v\rangle \geqslant (1-O(c))\,\mathbb{E}\|v\|^2$. For c small enough, this is at least $\delta^{O(1)}\,\mathbb{E}\|v\|^2$. Applying Fact 5.23 finishes the proof. □

5.5 Remaining lemmas

We provide sketches of the proofs of Lemma 5.7 and Lemma 5.10, since the proofs of these lemmas use only standard techniques.

Proof sketch of Lemma 5.7. For $\sigma \in \mathbb{R}^k$, let $\tilde{\sigma} = \sigma - (1 - 1/\sqrt{\alpha+1})/k$. Standard calculations show that if σ is drawn from the α, k Dirichlet distribution then $\mathbb{E}\,\tilde{\sigma}\tilde{\sigma}^\top = \frac{1}{k(\alpha+1)}\mathrm{Id}$. It follows by standard matrix concentration and the assumption $k, \alpha \leqslant n^{o(1)}$ that the eigenvalues of $\frac{1}{n}\sum_{i\leqslant n}\tilde{\sigma}_i\tilde{\sigma}_i^\top$ are all $1 \pm n^{-\Omega(1)}$, where $\sigma_1, \ldots, \sigma_n$ are iid draws from the α, k Dirichlet distribution.

For the second part of the Lemma, use the first part to show that $\left\|\frac{v_s}{\|v_s\|} - w_s'\right\| \leqslant 1/\mathrm{poly}(k)$. Then when $k \geqslant \delta^{-C}$ for large-enough C, if $\langle x, v_s\rangle^3 \geqslant \delta^{O(1)}\|v_s\|^3$ it follows that also $\langle x, w_s\rangle \geqslant \delta^{O(1)} - 1/\mathrm{poly}(k) \geqslant \delta^{O(1)}$. The lemma follows. □

Proof sketch of Lemma 5.10. If $\delta < 1 - \Omega(1)$, then $\delta^2/2 \geqslant \delta^{O(1)}$, so the Lemma follows from standard concentration and Theorem 2.3 on correlation-preserving projection. On the other hand, if $\delta \geqslant 1 - o(1)$, then $\|v' - \tilde{\sigma}\| \leqslant o(1)\cdot\|\tilde{\sigma}\|$, so the same is also true for the projection of v' into $(\tilde{\Delta}_{k-1})^n$ by convexity and the lemma follows. □

6 Lower bounds against low-degree polynomials at the Kesten-Stigum threshold

In this section we prove two lower bounds for k-community partial recovery algorithms based on low-degree polynomials.

6.1 Low-degree Fourier spectrum of the k-community block model

Theorem 6.1. *Let d, ε, k be constants. Let $\mu : \{0,1\}^{n\times n} \to \mathbb{R}$ be the relative density of $SBM(n, d, \varepsilon, k)$ with respect to $G(n, \frac{d}{n})$. Let $\mu^{\leqslant \ell}$ be the projection of μ to the degree-ℓ polynomials with respect to the norm induced by $G(n, \frac{d}{n})$.*[31] *For any constant $\delta > 0$ and $\xi > 0$ (allowing $\xi \leqslant o(1)$),*

$$\|\mu^{\leqslant \ell}\| \text{ is } \begin{cases} \geqslant n^{\Omega(1)} \text{ if } \varepsilon^2 d > (1+\delta)k^2, \ell \geqslant O(\log n) \\ \leqslant n^{2\xi} \text{ if } \varepsilon^2 d < (1-\delta)k^2, \ell < n^{\xi} \end{cases}.$$

[31] That is, $\|f\| = (\mathbb{E}_{G\sim G(n,\frac{d}{n})} f(G)^2)^{1/2}$.

This proves Theorem 1.9 (see discussion following statement of that theorem). To prove the theorem we need the following lemmas.

Lemma 6.2. *Let $\chi_\alpha : \{0,1\}^{n\times n} \to \mathbb{R}$ be the $\frac{d}{n}$-biased Fourier character. If $\alpha \subseteq \binom{n}{2}$, considered as a graph on n vertices, has any degree-one vertex, then*

$$\mathop{\mathbb{E}}_{G\sim SBM(n,d,\varepsilon,k)} \chi_\alpha(G) = 0$$

The proof follows from calculations very similar to those in Section 5, so we omit it.

Proof of Theorem 6.1. The bound $\|\mu^{\leqslant \ell}\| \geqslant n^{\Omega(1)}$ when $\varepsilon^2 d > (1+\delta)k^2$ and $\ell \gg \log(n)$, follows from almost identical calculations to Section 5,[32] so we omit this argument and focus on the regime $\varepsilon^2 d < (1-\delta)k^2$.

By definition and elementary Fourier analysis,

$$\|\mu^{\leqslant \ell}\|^2 = \sum_{\alpha\subseteq\binom{n}{2}, |\alpha|\leqslant \ell} \widehat{\mu}(\alpha)^2 \tag{6.1}$$

Also by definition,

$$\widehat{\mu}(\alpha) = \mathop{\mathbb{E}}_{G\sim G(n,\frac{d}{n})} \mu(G)\chi_\alpha(G) = \mathop{\mathbb{E}}_{G\sim SBM(n,d,\varepsilon,k)} \chi_\alpha$$

where $\{\chi_\alpha\}$ are the $\frac{d}{n}$-biased Fourier characters. Thus, using Lemma 6.2 we may restrict attion to the contribution of those $\alpha \subseteq \binom{n}{2}$ with $|\alpha| \leqslant \ell$ and containing no degree-1 vertices.

Fix such an α, and suppose it has $C(\alpha)$ connected components and $V_2(\alpha)$ vertices of degree 2 (considered again as a graph on $[n]$). Fact 6.3 (following this proof) together with routine computations shows that

$$\left(\mathop{\mathbb{E}}_{G\sim SBM(n,d,\varepsilon,k)} \chi_\alpha(G)\right)^2 \leqslant \left((1+O(\tfrac{d}{n}))\varepsilon^2\tfrac{d}{n}\right)^{|\alpha|} k^{-2(V(\alpha)-C(\alpha))} \leqslant \left(1+O(\tfrac{d}{n})\right)^{|\alpha|}\cdot n^{-|\alpha|}\cdot(1-\delta)^{|\alpha|}\cdot k^{2(|\alpha|-V(\alpha)+C(\alpha))}.$$

Let $c(\alpha) = \left(1+O(\frac{d}{n})\right)^{|\alpha|}\cdot n^{-|\alpha|}\cdot(1-\delta)^{|\alpha|}\cdot k^{2(|\alpha|-V(\alpha)+C(\alpha))}$ be this upper bound on the contribution of α to the right-hand side of (6.1). It will be enough to bound

$$(*) \quad \stackrel{\text{def}}{=} \sum_{\substack{\alpha\subseteq\binom{n}{2}\\ |\alpha|\leqslant\ell \\ \alpha \text{ has no degree 1 nodes}}} c(\alpha)$$

Given any α as in the sum, we may partition it into two vertex-disjoint subgraphs, α_0 and α_1, where α_0 is a union of cycles and no connected component of α_1 is a cycle, such that $\alpha = \alpha_0 \cup \alpha_1$. Thus,

$$(*) \leqslant \left(\sum_{\alpha_0} c(\alpha_0)\right)\left(\sum_{\alpha_1} c(\alpha_1)\right)$$

[32] The calculations in Section 5 are performed for long-armed stars; to prove the present result the analogous calculations should be performed for cycles of logarithmic lengh. Similar calculations also appear in many previous works.

where α_0 ranges over unions of cycles with $|\alpha_0| \leqslant \ell$ and α_1 ranges over graphs on $[n]$ with at most ℓ where all degrees are at least 2 and containing no connected component which is a cycle. Lemmas 6.4 and 6.5, which follow, the terms above as $O(1)$ and $n^{2\xi}$, respectively, which finishes the proof. □

Fact 6.3. *Let U be a connected graph on t vertices where all degrees are at least 2. For each vertex v of U let $\sigma_v \in \mathbb{R}^k$ be a uniformly random standard basis vector. Let $\tilde{\sigma}_v = \sigma_v - \frac{1}{k} \cdot 1$. Then*

$$\left| \mathbb{E} \prod_{(u,v)\in U} \langle \tilde{\sigma}_v, \tilde{\sigma}_u \rangle \right| \leqslant t k^{-t+1}$$

Proof. Consider a particular realization of $\sigma_1, \ldots, \sigma_t$. Suppose all but m vertices v in U are adjacent to at least 2 vertices u_1, u_2 such that $\sigma_{u_1} \neq \sigma_v$ and $\sigma_{u_2} \neq \sigma_v$. In this case,

$$\left| \prod_{(u,v)\in U} \langle \tilde{\sigma}_v, \tilde{\sigma}_u \rangle \right| \leqslant k^{-(t-m)} .$$

The probability of such a pattern of disagreements is at most k^{-m}, unless $m = t$, in which case the probability is at most k^{-t+1}. The fact follows. □

Lemma 6.4. *For $\alpha \subseteq \binom{n}{2}$, let $V(\alpha)$ be the number of vertices in α, let $C(\alpha)$ be the number of connected components in α. For constants ε, d, k, let $c(\alpha) \stackrel{\text{def}}{=} \left(1 + O(\frac{d}{n})\right)^{|\alpha|} \cdot n^{-|\alpha|} \cdot (1-\delta)^{|\alpha|} \cdot k^{2(|\alpha| - V(\alpha) + C(\alpha))}$ Let $\ell \leqslant n^{0.01}$ and*

$$U = \left\{ \alpha \subseteq \binom{n}{2} \,:\, \alpha \text{ has all degrees } \geqslant 2, \text{ has no connected components which are cycles, } |\alpha| \leqslant \ell \right\} .$$

Then

$$\sum_{\alpha \in U} c(\alpha) \leqslant O(1) .$$

Proof. We will use a coding argument to bound the number of $\alpha \in U$ with V vertices, E edges, and C connected components. We claim that any such α is uniquely specified by the following encoding.

To encode α, start by picking an arbitrary vertex v_1 in α. List the vertices $v_1, \ldots, v_{|V|}$ of α, each requiring $\log n$ bits, starting from v_1, using the following rules to pick v_i.

1. If v_{i-1} has a neighbor not yet appearing in the list $v_1, \ldots, v_{i-1}$, let v_i be any such neighbor.
2. Otherwise, if v_{i-1} has a neighbor v_j which
   (a) appears in the list $v_1, \ldots, v_{i-1}$ and
   (b) for which either $j = 1$ or v_{j-1} is not adjacent to v_j in α, and
   (c) for which if $j \neq i'$ for $i' \leqslant i-1$ being the minimal index such that $v_{i'}, \ldots, v_{i-1}$ is a path in α (i.e. $v_j, \ldots, v_{i-1}$ are not a cycle in α)

then reorder the list as follows. Remove vertices $v_j, \ldots, v_{j'}$ where j' is the greatest index so that all edges $v_\ell, v_{\ell+1}$ exist in α for $j \leqslant \ell \leqslant j'$. Also remove vertices $v_{i'}, \ldots, v_{i-1}$ where i' is analogously the minimal index such that edes $v_\ell, v_{\ell+1}$ exist in α for $i' \leqslant \ell \leqslant i-1$. Then, append the list $v_{j'}, v_{j'-1}, \ldots, v_j, v_{i-1}, \ldots, v_{i'}$. By construction, all of these vertices appear in a path in α. The new list retains the invariant that every vertex either preceeds a neighbor in α or has no neighbors in α which have not previous appeared in the list.

3. Otherwise, let v_i be an arbitrary vertex in α in the same connected component as v_{i-1}, if some such vertices has not yet appeared in the list.

4. Otherwise, let v_i be an arbitrary vertex of α not yet appearing among $v_1, \ldots, v_{i-1}$.

After the list of vertices, append to the encoding the following information. First, a list of the R (for removed) pairs v_i, v_{i+1} for which there is not an edge (v_i, v_{i+1}) in α. This uses $2R \log V$ bits. Last, a list of the edges in α which are not among the pairs v_i, v_{i+1} (each edge encoded using $2 \log V$ bits).

We argue that the number R of removed pairs (and hence the length of their list in the encoding) is not too great. In particular, we claim $R \leqslant 2(E - V)$. In fact, this is true connected-component-wise in α. To see it, proceed as follows.

Fix a connected component β of α. Let v_t be the first vertex in β to appear in the list $v_1, \ldots, v_{|V|}$. Proceeding in increasing order down the list from v_t, let $(v_{r_1}, v_{r_1+1}), (v_{r_2}, v_{r_2+1}), \ldots$ be the pairs encountered (before leaving β) which do not correspond to edges in α (and hence will later appear in the list of removed pairs).

Construct a sequence of subgraphs β_j of β as follows. The graph β_1 is the line on vertices $v_t, \ldots, v_{r_1}$. To construct the graph β_j, start from β_{j-1} and add the line from $v_{r_{j-1}+1}$ to v_{r_j} (by definition all these edges appear in β). Since v_{r_j} must have at least degree 2, it has a neighbor u_j in β among the vertices v_a for $a < r_j$ aside from v_{r_j-1}. (If v_{r_j} had a neighbor not yet appearing in the list, then v_{r_j+1} would have been that neighbor, contrary to assumption.) Choose any such neighbor and add it to β_j; this finishes construction of the graph β_j. For later use, note that either adding the edge to u_j turns $\beta_j \setminus b_{j-1}$ into a cycle or u_j is not itself among the v_r's, since otherwise in constructing the list we would have done a reordering operation.

In each of the graphs β_j, the number of edges is equal to the number of vertices. To obtain β, we must add $E_\beta - V_\beta$ edges (where E_β is the number of edges and β and V_β is the number of vertices). We claim that in so doing at least one half of a distinct such edge must be added per β_j; we prove this via a charging scheme. As noted above, each graph $\beta_j \setminus \beta_{j-1}$ either contains $v_{r_{j-1}}$ as a degree-1 vertex or it forms cycle. If it contains a degree-1 vertex, by construction this vertex is not $u_{j'}$ for any $j' > j$, otherwise we would have reordered. So charge β_j to the edge which must be added to fix the degree-1 vertex.

In the cycle case, either some edge among the $E_\beta - V_\beta$ additional edges is added incident to the cycle (in which case we charge β_j to this edge), or some $u_{j'}$ for $j' > j$ is in $\beta_j \setminus \beta_{j-1}$. If the latter, then $\beta_{j'} \setminus \beta_{j'-1}$ contains a degree-1 vertex and $\beta_j \setminus \beta j - 1$ can be charged to the edge which fixes that degree 1 vertex. Every additional edge was charged at most twice. Thus, $R \leqslant 2(E - V)$

It is not hard to check that α can be uniquely decoded from the encoding previously described. The final result of this encoding scheme is that each α can be encoded with at most $V \log n + 6(E -$

$V)\log V$ bits, and so there are at most $n^V \cdot V^{6(E-V)}$ choices for α. The contribution of such α to $\sum_{\alpha\in U} c(\alpha)$ is thus at most

$$n^{-(E-V)}V^{6(E-V)}(1-\delta/2)^E k^{2(E-V+C)}$$

We know that $C \leqslant E - V$. So as long as $k, V \leqslant n^{0.01}$, we obtain that this contributes at most $n^{(E-V)/2}(1-\delta/2)^E$. Summing across all $V, E \leqslant n^{0.01}$, the lemma follows. □

Lemma 6.5. *For $\alpha \subseteq \binom{n}{2}$, let $V(\alpha)$ be the number of vertices in α, let $C(\alpha)$ be the number of connected components in α. For constants $1 > \delta > 0$ and k, let $c(\alpha) \stackrel{\text{def}}{=} \left(1 + O(\frac{d}{n})\right)^{|\alpha|} \cdot n^{-|\alpha|} \cdot (1-\delta)^{|\alpha|} \cdot k^{2(|\alpha|-V(\alpha)+C(\alpha))}$ Let $\ell \leqslant n^{\xi/k^2}$ for some $\xi > 0$ (allowing $\xi \leqslant o(1)$) and*

$$U = \left\{\alpha \subseteq \binom{n}{2} \,:\, \alpha \text{ is a union of cycles}\right\}.$$

Then

$$\sum_{\alpha\in U} c(\alpha) \leqslant n^{2\xi}$$

Proof. Let U_t be the set of α which are unions of t-cycles (we exclude the empty α). Let $c_t = \sum_{\alpha\in U_t} c_\alpha$. Then

$$\sum_{\alpha\in U} c(\alpha) \leqslant \prod_{t\leqslant \ell}(1 + c_t).$$

Count the $\alpha \in U_t$ which contain exactly p cycles of length t by first choosing a list of pt vertices—there are n^{pt} choices. In doing so we will count each alpha $p!t^p$ times, since each of the p cycles can be rotated and the cycles can themselves be exchanged. All in all, there are at most $n^{pt}/(p!t^p)$ such α, and they contribute at most

$$\frac{c(\alpha)n^{pt}}{p!t^p} \leqslant \frac{(1-\delta/2)^{pt}k^{2p}}{p!t^p} \leqslant k^{2p}/(p!t^p).$$

for large enough n. Thus, summing over all $\alpha \in U_t$, we get

$$(1 + c_t) \leqslant \sum_{p=0}^{\ell} \frac{(1-\delta/2)^p k^{2p}}{p!t^p} \leqslant \exp(k^2/t).$$

So,

$$\prod_{t\leqslant \ell}(1 + c_t) \leqslant \exp(k^2 \sum_{t=1}^{\ell} 1/t) \leqslant \exp(k^2 \log 2\ell) \leqslant (2\ell)^{k^2} \leqslant n^{2\xi}.$$

□

6.2 Lower bound for estimating communities

Theorem 6.6. *Let $d, \varepsilon, k, \delta$ be constants such that $\varepsilon^2 d < (1-\delta)k^2$. Let $f : \{0,1\}^{n\times n} \to \mathbb{R}$ be any function, let $i, j \in [n]$ be distinct. Then if f satisfies $\mathbb{E}_{G\sim G(n,\frac{d}{n})} f(G) = 0$ and is correlated with the indicator $\mathbf{1}_{\sigma_i=\sigma_j}$ that i and j are in the same community in the following sense:*

$$\frac{\mathbb{E}_{G\sim SBM(n,d,\varepsilon,k)} f(G)(\mathbf{1}_{\sigma_i=\sigma_j} - \frac{1}{k})}{(\mathbb{E}_{G\sim G(n,\frac{d}{n})} f(G)^2)^{1/2}} \geqslant \Omega(1)$$

then $\deg f \geqslant n^{c(d,\varepsilon,k)}$ *for some* $c(d,\varepsilon,k) > 0$.

Proof. Let $g(G) = \mu(G)\,\mathbb{E}[\mathbf{1}_{\sigma_i=\sigma_j} - \frac{1}{k} \mid G]$, where $\mu(G)$ is the relative density of $SBM(n,d,\varepsilon,k)$. Standard Fourier analysis shows that the optimal degree-ℓ choice for such f to maximize the above correlation is $g^{\leqslant \ell}$, the orthogonal projection of g to the degree-ℓ polynomials with respect to the measure $G(n,\frac{d}{n})$, and the correlation is at most $\|g^{\leqslant\ell}\|$. It suffices to show that for some constant $c(d,\varepsilon,k)$, if $\ell < n^{c(d,\varepsilon,k)}$ then $\|g^{\leqslant\ell}\| \leqslant o(1)$.

For this we expand g in the Fourier basis, noting that

$$\widehat{g}(\alpha) = \mathop{\mathbb{E}}_{\sigma,G\sim SBM(n,d,\varepsilon,k)} \langle \tilde{\sigma}_i, \tilde{\sigma}_j\rangle \chi_\alpha(G)$$

where as usual $\tilde{\sigma}_i = \sigma_i - \frac{1}{k}\cdot 1$ is the centered indicator of i's community. By-now routine computations show that

$$\widehat{g}(\alpha)^2 \leqslant \left((1+O(d/n))\varepsilon^2\tfrac{d}{n}\right)^{|\alpha|} \cdot \left(\mathbb{E}\langle\tilde{\sigma}_i,\tilde{\sigma}_j\rangle \cdot \prod_{(k,\ell)\in\alpha}\langle\tilde{\sigma}_i,\tilde{\sigma}_j\rangle\right)^2$$

We assume that $(i,j)\notin\alpha$; it is not hard to check that such α's dominate the norm $\|g^{\leqslant\ell}\|$. If some vertex aside from i,j in α has degree 1 then this is zero. Similarly, if i or j does not appear in α then this is zero. Otherwise,

$$\widehat{g}(\alpha)^2 \leqslant ((1+O(d/n)))^{|\alpha|}\, n^{-|\alpha|}(1-\delta)^{|\alpha|}k^{2(|\alpha|-V(\alpha)+C(\alpha)}$$

where as usual $V(\alpha)$ is the number of vertices in α and $C(\alpha)$ is the number of connected components in α. Let $\beta(\alpha)$ be the connected component of α containing i and j (if they are not in the same component the arguments are mostly unchanged). Then we can bound

$$\|g^{\leqslant\ell}\|^2 = \sum_{|\alpha|\leqslant\ell}\widehat{g}(\alpha)^2 \leqslant \|\mu^{\leqslant\ell}\|^2 \cdot \sum_{\beta}((1+O(d/n)))^{|\beta|}\, n^{-|\beta|}(1-\delta)^{|\beta|}k^{2(|\beta|-V(\beta)+1}$$

where β ranges over connected graphs with vertices from $[n]$, at most ℓ edges, every vertex except i and j having degree at least 2, and containing i and j with degree at least 1. There are at most $n^{V-2}V^{O(E-V)}$ such graphs containing at V vertices aside from i and j and E edges (by an analogous argument as in Lemma 6.4). The total contribution from such β is therefore at most

$$\frac{k^{2(E-V+1)}V^{O(E-V)}}{n^{E-V+2}}$$

Summing over V and E, we get

$$\sum_{\beta} ((1 + O(d/n)))^{|\beta|} n^{-|\beta|} (1-\delta)^{|\beta|} k^{2(|\beta|-V(\beta)+1} \leqslant n^{-\Omega(1)}$$

so long as $\ell \leqslant n^c$ for small enough c. Using Theorem 6.1 to bound $\|\mu^{\leqslant \ell}\|$ finishes the proof. □

7 Tensor decomposition from constant correlation

Problem 7.1 (Orthogonal n-dimensional 4-tensor decomposition from constant correlation)**.** Let $a_1, \ldots, a_m \in \mathbb{R}^n$ be orthonormal, and let $A = \sum_{i=1}^m a_i^{\otimes 4}$. Let $B \in (\mathbb{R}^n)^{\otimes 4}$ satisfy $\frac{\langle A, B\rangle}{\|A\|\|B\|} \geqslant \delta = \Omega(1)$. Let $\mathcal{O}$ be an oracle such that for any unit $v \in \mathbb{R}^n$,

$$\mathcal{O}(v) = \begin{cases} \textbf{YES} \text{ if } \sum_{i=1}^m \langle a_i, v\rangle^4 \geqslant \delta^{O(1)} \\ \textbf{NO} \text{ otherwise} \end{cases}$$

Input: The tensor B, and if $\delta < 0.01$, access to the oracle $\mathcal{O}$.
Goal: Output orthonormal vectors $b_1, \ldots, b_m$ so that there is a set $S \subseteq [m]$ of size $|S| \geqslant \delta^{O(1)} \cdot m$ where for every $i \in S$ there is $j \leqslant m$ with $\langle b_j, a_i\rangle^2 \geqslant \delta^{O(1)}$.

We will give an $n^{1/\delta^{O(1)}}$-time algorithm (hence using at most $n^{1/\delta^{O(1)}}$ oracle calls) for this problem based on a maximum-entropy Sum-of-Squares relaxation. The main theorem is the following; the subsequent corollary arrives at the final algorithm.

Theorem 7.2. *Let A, B and $a_1, \ldots, a_m$ and $\delta \leqslant 0.01$ be as in Problem 7.1. Let $v_1, \ldots, v_r$ for $r \leqslant \delta^4 m$ be orthonormal vectors. There is a randomized algorithm ALG with running time $n^{O(1)}$ which takes input $B, v_1, \ldots, v_r$ and outputs a unit vector v, orthogonal to $v_1, \ldots, v_r$, with the following guarantee. There is a set $S \subseteq [m]$ of size $|S| \geqslant \delta^{O(1)} \cdot m$ so that for $i \in S$,*

$$\mathbb{P}\left\{\langle v, a_i\rangle^2 \geqslant \delta^{O(1)}\right\} \geqslant n^{-1/\mathrm{poly}(\delta)}.$$

The following corollary captures the overall algorithm for tensor decomposition, using the oracle $\mathcal{O}$ to filter the output of the algorithm of Theorem 7.2.

Corollary 7.3. *Let $a_1, \ldots, a_n, A, B, \delta$ be as in Theorem 7.2 and $\mathcal{O}$ as in Problem 7.1. There is a $n^{\mathrm{poly}(1/\delta)}$-time algorithm which takes the tensor B as input and returns $b_1, \ldots, b_m$ such that with high probability there is a set $S \subseteq [m]$ of size $|S| \geqslant \delta^{O(1)} m$ which has the guarantee that for all $i \in S$ there is $j \leqslant m$ with $\langle a_i, b_j\rangle^2 \geqslant \delta^{O(1)}$. If $\delta \leqslant 1 - \Omega(1)$, the algorithm makes $n^{1/\mathrm{poly}(\delta)}$ adaptive queries to the oracle $\mathcal{O}$.*

The algorithm can also be implemented with nonadaptive queries as follows. Once the input B and the random coins of the algorithm are fixed, there is a list of at most $n^{\mathrm{poly}(k/\delta)}$. Query the oracle $\mathcal{O}$ nonadaptively on all these vectors and assemble the answers into a lookup table; then the decomposition algorithm can be run using access only to the lookup table.

Proof of Corollary 7.3. If $\delta \geqslant 1 - \varepsilon^*$ for a small enough constant ε^* then the tensor decomposition algorithm of Schramm and Steurer has the appropriate guarantees. (See Theorem 4.4 and Lemma 4.9 in [SS17]. This algorithm has several advantages, including that it does not need to solve any semidefinite program, but it cannot handle the high-error regime we need to address here.)

From here on we assume $\delta \leqslant 0.01 < 1 - \varepsilon^*$. (Otherwise, we can replace δ with $\delta^C \leqslant 0.01$ for large enough C.) Our algorithm is as follows.

Algorithm 7.4 (Constant-correlation tensor decomposition)**.**

1. Let V be an empty set of vectors.
2. For rounds $1, \dots, T = \delta^{O(1)}m$, do:
   (a) Use the algorithm of Theorem 7.2 on the tensor B to generate $w_1, \dots, w_t$, where $t = n^{1/\delta^{O(1)}}$.
   (b) Call $\mathcal{O}$ on successive vectors $w_1, \dots, w_t$, and let w be the first for which it outputs **YES**. (If no such vector exists, the algorithm halts and outputs random orthonormal vectors $b_1, \dots, b_m$.)
   (c) Add w to V.
3. Let $b_1, \dots, b_{m-|V|}$ be random orthonormal vectors, orthogonal to each $v \in V$.
4. Output $\{b_1, \dots, b_{m-|V|}\} \cup V$.

Choosing $t = n^{1/\delta^{O(1)}}$ large enough, and $T = \delta^{O(1)}m$ small enough, by Theorem 7.2 with high probability in every round $1, \dots, T$ there is some w among $w_1, \dots, w_t$ for which $\mathcal{O}$ outputs **YES**. Suppose that occurs. In this case, the algorithm outputs (along with some random vectors b_i) a set of vectors V which are orthonormal, and each $v \in V$ satisfies $\langle v, a_i \rangle \geqslant \delta^{O(1)}$ for some a_i; say that this a_i is *covered* by v. Each a_i can be covered at most $1/\delta^{O(1)}$ times, by orthonormality of the set V. So, at least $\delta^{O(1)}|V| = \delta^{O(1)}m$ vectors are covered at least once, which proves the corollary. □

We turn to the proof of Theorem 7.2. We will use the following lemmas, whose proofs are later in this section. The problem is already interesting when the list $v_1, \dots, v_r$ is empty, and we encourange the reader to understand this case first.

The first lemma says that a pseudodistribution of high entropy (in the 2-norm sense[33]) which is correlated with the tensor B must also be nontrivially correlated with A.

Lemma 7.5. *Let A, B be as in Problem 7.1. Let $v_1, \dots, v_r \in \mathbb{R}^n$ be orthonormal, with $r \leqslant \delta^4 m$. Suppose $\tilde{\mathbb{E}}$ is the degree-4 pseudodistribution solving*

$$\min \| \tilde{\mathbb{E}}\, x^{\otimes 4} \|_F \tag{7.1}$$

$$\text{s.t. } \tilde{\mathbb{E}} \text{ satisfies } \{\|x\|^2 \leqslant 1, \langle x, v_1 \rangle = 0, \dots, \langle x, v_r \rangle = 0\}$$

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 4}, B \rangle \geqslant \frac{\delta}{2m}$$

$$\left\| \tilde{\mathbb{E}}\, xx^\top \right\| \leqslant \tfrac{1}{m} \tag{7.2}$$

$$\left\| \tilde{\mathbb{E}}\, xx^\top \otimes xx^\top \right\| \leqslant \tfrac{1}{m} \tag{7.3}$$

[33] For a distribution μ finitely-supported on a family of orthonormal vectors, the Frobenious norm $\| \mathbb{E}_{x \sim \mu} x^{\otimes k} \|$ is closely related to the collision probability of μ, itself closely related to the order-2 case of Rényi entropy.

Then $\tilde{\mathbb{E}} \sum_{i \leqslant m} \langle x, a_i \rangle^4 \geqslant \delta^2/8$. *Furthermore, it is possible to find* $\tilde{\mathbb{E}}$ *in polynomial time.*[34]

The second lemma says that given a high-entropy (in the spectral sense of [MSS16]) pseudodistribution $\tilde{\mathbb{E}}$ having nontrivial correlation with some $a \in \mathbb{R}^n$, contracting $\tilde{\mathbb{E}}$ with a yields a matrix whose quadratic form is large at a and which does not have too many large eigenvalues.

Lemma 7.6. *Let* $a_1, \ldots, a_m \in \mathbb{R}^n$ *be orthonormal.*
Let $\tilde{\mathbb{E}}$ *be a degree-4 pseudoexpectation such that*

1. $\tilde{\mathbb{E}}$ *satisfies* $\{\|x\|^2 \leqslant 1\}$
2. $\tilde{\mathbb{E}} \sum_{i \leqslant m} \langle x, a_i \rangle^4 \geqslant \delta$.
3. $\|\tilde{\mathbb{E}} xx^\top]\|_{op}, \|\tilde{\mathbb{E}} xx^\top \otimes xx^\top\|_{op} \leqslant \frac{1}{m}$.[35]

Let $M_i \in \mathbb{R}^{n \times n}$ *be the matrix* $\tilde{\mathbb{E}} \langle x, a_i \rangle^2 xx^\top$. *For every* $i \in [m]$, *the matrix* M_i *has at most* $4/\delta$ *eigenvalues larger than* $\frac{\delta}{4m}$. *Furthermore,*

$$\mathop{\mathbb{P}}_{i \sim [m]} \left\{ \langle a_i, M_i a_i \rangle \geqslant \tfrac{\delta}{2m} \right\} \geqslant \tfrac{\delta}{2} .$$

The last lemma will help show that a random contraction of a high-entropy pseudodistribution behaves like one of the contractions from Lemma 7.6, with at least inverse-polynomial probability.

Lemma 7.7. *Let* $g \sim \mathcal{N}(0, \Sigma)$ *for some* $0 \preceq \Sigma \preceq \mathrm{Id}$ *and let* $\tilde{\mathbb{E}}$ *be a degree-4 pseudoexpectation where*

- $\tilde{\mathbb{E}}$ *satisfies* $\{\|x\|^2 \leqslant 1\}$.
- $\left\|\tilde{\mathbb{E}} xx^\top\right\| \leqslant c$.
- $\left\|\tilde{\mathbb{E}} xx^\top \otimes xx^\top\right\| \leqslant c$

Then

$$\mathop{\mathbb{E}}_{g} \left\|\tilde{\mathbb{E}} \langle g, x \rangle^2 xx^\top\right\| \leqslant O(c \cdot \log n) .$$

Now we can prove Theorem 7.2.

Proof of Theorem 7.2. The algorithm is as follows:

Algorithm 7.8 (Low-correlation tensor decomposition)**.**

1. Use the first part of Lemma 7.5 to obtain a degree-4 pseudoexpectation with $\tilde{\mathbb{E}} \sum_{i \in [m]} \langle a_i, x \rangle^4 \geqslant \delta^2/4$ satisfying $\{\|x\|^2 \leqslant 1, \langle x, v_1 \rangle = 0, \ldots, \langle x, v_r \rangle = 0\}$.
2. Sample a random $g \sim \mathcal{N}(0, \mathrm{Id})$ and compute the contraction $M = \tilde{\mathbb{E}} \langle g, x \rangle^2 xx^\top$.
3. Output a random unit vector b in the span of the top $\frac{32}{\delta^2}$ eigenvectors of M.

[34] Up to inverse-polynomial error, which we ignore here. See [MSS16] for the ideas needed to show polynomial-time solvability.

[35] Recall that $\|\cdot\|$ denotes the operator norm, or maximum singular value, of a matrix.

First note that for any $v \in \mathrm{Span}\{v_1, \ldots, v_r\}$, we must have $\langle v, Mv\rangle = \tilde{\mathbb{E}}\langle g, x\rangle^2\langle v, x\rangle^2 = 0$, so v lies in the kernel of M. Hence, the ouput of the algorithm will always be orthogonal to $v_1, \ldots, v_r$.

Let Π_{32/δ^2} be the projector to the top $32/\delta^2$ eigenvectors of M. For any unit vector a with $\|\Pi_{32/\delta^2} a\| \geqslant \delta^{O(1)}$, the algorithm will output b with nontrivial correlation with a. Formally, for any such a,

$$\mathop{\mathbb{E}}_{b} \langle b, a\rangle^2 \geqslant \delta^{O(1)} .$$

So, our goal is to show that for a $\delta^{O(1)}$-fraction of the vectors $a_1, \ldots, a_m$,

$$\mathop{\mathbb{P}}_{g}\{\|\Pi_{32/\delta^2} a_i\| \geqslant \delta^{O(1)}\} \geqslant n^{-1/\delta^{O(1)}} .$$

For $i \in [m]$, let $M_i = \tilde{\mathbb{E}}\langle a_i, x\rangle^2 xx^\top$. Let i be the index of some a_i so that

$$\langle a_i, M_i a_i\rangle \geqslant \tfrac{\delta^2}{16m} \text{ and } \operatorname{rank} M_i \overset{\geqslant}{} \tfrac{\delta^2}{32m} \leqslant \tfrac{32}{\delta^2}$$

as in Lemma 7.6. (There are $\Omega(\delta^2 m)$ possible choices for a_i, according to the Lemma.)

We expand the Gaussian vector g from the algorithm as

$$g = g_0 \cdot a_i + g'$$

where $g_0 \sim \mathcal{N}(0,1)$ and $\langle g', a_i\rangle = 0$. We note for later use that g' is a Gaussian vector independent of g_0 and that $\mathbb{E}(g')(g')^\top \preceq \mathrm{Id}$. Using this expansion,

$$M = g_0^2\, \tilde{\mathbb{E}}\langle a_i, x\rangle^2 xx^\top + 2 \cdot g_0\, \tilde{\mathbb{E}}\langle g', x\rangle\langle a_i, x\rangle xx^\top + \tilde{\mathbb{E}}\langle g', x\rangle^2 xx^\top .$$

We will show that all but the first term have small spectral norm. Addressing the middle term first, by Cauchy-Schwarz, for any unit $v \in \mathbb{R}^n$,

$$\tilde{\mathbb{E}}\langle g', x\rangle\langle a_i, x\rangle\langle v, x\rangle^2 \leqslant \left(\tilde{\mathbb{E}}\langle g', x\rangle^2\langle x, v\rangle^2\right)^{1/2} \left(\tilde{\mathbb{E}}\langle a_i, x\rangle^2\langle v, x\rangle^2\right)^{1/2} \leqslant \left\|\tilde{\mathbb{E}}\langle g', x\rangle^2 xx^\top\right\|^{1/2} \cdot \left(\tfrac{1}{m}\right)^{1/2} ,$$

where in the last step we have used that $\left\|\tilde{\mathbb{E}}\, xx^\top \otimes xx^\top\right\| \leqslant \frac{1}{m}$.

By Markov's inequality and Lemma 7.7,

$$\mathop{\mathbb{P}}_{g'}\left\{\left\|\tilde{\mathbb{E}}\langle g', x\rangle^2 xx^\top\right\| > \tfrac{t \log n}{m}\right\} \leqslant O\left(\tfrac{1}{t}\right) .$$

Let t be a large enough constant so that

$$\mathop{\mathbb{P}}_{g'}\left\{\left\|\tilde{\mathbb{E}}\langle g', x\rangle^2 xx^\top\right\| \leqslant \tfrac{t \log n}{m}\right\} \geqslant 0.9 .$$

For any constant c, with probability $n^{-1/\mathrm{poly}(\delta)}$, the foregoing occurs and g_0 (which is independent of g') is large enough that

$$g_0^2 \cdot \tfrac{c\delta^2}{m} > \tfrac{1}{\delta^4}\left\|M - g_0^2 M_i\right\| .$$

Choosing c large enough, in this case

$$M' \overset{\text{def}}{=} \tfrac{1}{g_0^2} M = M_i + O(\delta^6/m) .$$

Hence the vector a_i satisfies

$$\tfrac{1}{g_0^2}\langle a_i, Ma_i\rangle \geqslant \tfrac{\delta^2}{33m}$$

This means that the projection b of a_i into the span of eigenvectors of M' with eigenvalue at least $\delta^2/60m$ has $\|b\|^2 \geqslant \delta^{O(1)}$. This finishes the proof. □

7.1 Proofs of Lemmas

These lemmas and their proofs use many ideas from [MSS16]. The main difference here is that we want to contract the tensor $\tilde{\mathbb{E}}\, x^{\otimes 4}$ in 2 modes, to obtain the matrix $\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top$. For us this is useful because $\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top \succeq 0$. By contrast, the tools in [MSS16] would only allow us to analyze the contraction $\tilde{\mathbb{E}}\langle h, x\otimes x\rangle xx^\top$ for $h \sim \mathcal{N}(0, \mathrm{Id}_{n^2})$.

We start with an elementary fact.

Fact 7.9. *Let $a_1, \dots, a_m \in \mathbb{R}^n$ be orthonormal. Let Π be the projector to a subspace of codimension at most δm. Let $A = \sum_{i=1}^m a_i^{\otimes 4}$ and $\Pi A = \sum_{i=1}^m (\Pi a_i)^{\otimes 4}$. Then $\langle A, \Pi A\rangle \geqslant (1 - O(\sqrt{\delta}))\|A\| \cdot \|\Pi A\|$.*

A useful corollary of Fact 7.9 is that if T is any 4-tensor satisfying $\langle T, \Pi A\rangle \geqslant \delta\|T\|\|\Pi A\|$ and Π has codimension $\ll \delta^2 m$, then $\langle T, A\rangle \geqslant \Omega(\delta)\|T\|\|A\|$.

Proof of Fact 7.9. We expand

$$\langle A, \Pi A\rangle = \sum_{i,j\leqslant m} \langle a_i, \Pi a_j\rangle^4 \geqslant \sum_{i,j\leqslant m} \|\Pi a_i\|^8$$

Writing Π in the a_i basis, we think of $\|\Pi a_i\|^4 = \Pi_{ii}^2$, the square of the i-th diagonal entry of Π. Since Π has codimension at most δm,

$$\operatorname{rank}\Pi = \operatorname{Tr}\Pi = \sum_{i\leqslant n} \Pi_{ii} \geqslant n - \delta m\,.$$

Furthermore, for each i, it must be that $0 \leqslant \Pi_{ii} \leqslant 1$. By Markov's inequality, at most $\sqrt{\delta}m$ diagonal entries of Π can be less than $1 - \sqrt{\delta}$ in magnitude. Hence, $\sum_{i\leqslant m} \Pi_{ii}^4 \geqslant (1 - 4\sqrt{\delta})m$. On the other hand, $\|A\|^2 = m$; this proves the fact. □

Now we can prove Lemma 7.5.

Proof of Lemma 7.5. We will appeal to Theorem 2.3. Let C be the convex set of all pseudo-moments $\tilde{\mathbb{E}}\, x^{\otimes 4}$ such that $\tilde{\mathbb{E}}$ is a deg-4 pseudo-distribution that satisfies the polynomial constraints $\{\|x\|^2 \leqslant 1, \langle x, v_i\rangle = 0\}$ and the operator norm conditions

$$\left\|\tilde{\mathbb{E}}\, xx^\top\right\| \leqslant \tfrac{1}{m},$$
$$\left\|\tilde{\mathbb{E}}\, xx^\top \otimes xx^\top\right\| \leqslant \tfrac{1}{m}\,.$$

Let Π be the projector to the orthogonal space of $v_1, \dots, v_r$. Notice that $\frac{1}{m}\Pi A \in C$. Furthermore, $\langle B, \Pi A\rangle \geqslant \delta/2$ by Fact 7.9, the assumption that $r \leqslant \delta^4 m$, and the assumption $\delta \leqslant 0.01$. By Theorem 2.3, and Fact 7.9 again, the optimizer of the convex program in the Lemma satisfies $\langle \tilde{\mathbb{E}}\, x^{\otimes 4}, \frac{1}{m}A\rangle \geqslant \frac{\delta^2}{8m}$) and the result follows. □

Proof of Lemma 7.6. By the assumption $\|\tilde{\mathbb{E}} xx^\top \otimes xx^\top\| \leqslant \frac{1}{m}$, for every a_i it must be that $\tilde{\mathbb{E}}\langle x, a_i\rangle^4 \leqslant \frac{1}{m}$. Since $\tilde{\mathbb{E}} \sum_{i=1}^m \langle x, a_i\rangle^4 \geqslant \delta$, at least $\delta m/2$ of the a_i's must satisfy $\tilde{\mathbb{E}}\langle x, a_i\rangle^4 \geqslant \frac{\delta}{2m}$. Rewritten, for any such a_i we obtain $\langle a_i, M_i a_i\rangle \geqslant \frac{\delta}{2m}$.

For any M_i,

$$\operatorname{Tr} M_i = \tilde{\mathbb{E}}\langle x, a_i\rangle^2 \|x\|^2 = \tilde{\mathbb{E}}\langle x, a_i\rangle^2 \leqslant \tfrac{1}{m}$$

because $\|\tilde{\mathbb{E}} xx^\top\| \leqslant \frac{1}{m}$. Also, $M_i \succeq 0$. Hence, M_i can have no more than $\frac{4}{\delta}$ eigenvalues larger than $\frac{\delta}{4m}$. □

Now we turn to the proof of Lemma 7.7. We will need spectral norm bounds on certain random matrices associated to the random contraction $\tilde{\mathbb{E}}\langle g, x\rangle xx^\top$. The following are closely related to Theorem 6.5 and Corollary 6.6 in [MSS16].

Lemma 7.10. *Let $g \sim \mathcal{N}(0, \mathrm{Id})$ and let $\tilde{\mathbb{E}}$ be a degree-4 pseudoexpectation where*

- *$\tilde{\mathbb{E}}$ satisfies $\{\|x\|^2 = 1\}$.*
- *$\left\|\tilde{\mathbb{E}} xx^\top\right\| \leqslant c$.*
- *$\left\|\tilde{\mathbb{E}} xx^\top \otimes xx^\top\right\| \leqslant c$*

Then

$$\mathop{\mathbb{E}}_g \left\|\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top\right\| \leqslant O(c \cdot \log n)\,.$$

Before proving the lemma, we will need a classical decoupling inequality.

Fact 7.11 (Special case of Theorem 1 in [dlPnMS94])**.** *Let $g, h \sim \mathcal{N}(0, \mathrm{Id}_n)$ be independent. Let M_{ij} for $i, j \in [n]$ be a family of matrices. There is a universal constant C so that*

$$\mathop{\mathbb{E}}_g \left\|\sum_{i\neq j} g_i g_j \cdot M_{ij}\right\| \leqslant C \cdot \mathop{\mathbb{E}}_{g,h} \left\|\sum_{i\neq j} g_i h_j \cdot M_{ij}\right\|\,.$$

We will also need a theorem from [MSS16].

Fact 7.12 (Corollary 6.6 in [MSS16])**.** *Let $T \in \mathbb{R}^p \otimes \mathbb{R}^q \otimes \mathbb{R}^r$ be an order-3 tensor. Let $g \sim \mathcal{N}(0, \Sigma)$ for some $0 \preceq \Sigma \preceq \mathrm{Id}_r$. Then for any $t \geqslant 0$,*

$$\mathop{\mathbb{P}}_g \left\{\left\|(\mathrm{Id} \otimes \mathrm{Id} \otimes g)^\top T\right\|_{\{1\},\{2\}} \geqslant t \cdot \max\left\{\|T\|_{\{1\},\{2,3\}}, \|T\|_{\{2\},\{1,3\}}\right\}\right\} \leqslant 2(p+q) \cdot e^{-t^2/2}\,,$$

and consequently,

$$\mathop{\mathbb{E}}_g \left[\left\|(\mathrm{Id} \otimes \mathrm{Id} \otimes g)^\top T\right\|_{\{1\},\{2\}}\right] \leqslant O(\log(p+q))^{1/2} \cdot \max\left\{\|T\|_{\{1\},\{2,3\}}, \|T\|_{\{2\},\{1,3\}}\right\}$$

Proof of Lemma 7.10. We expand the matrix $\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top$ as

$$\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top = \sum_{i\in[n]} g_i^2\, \tilde{\mathbb{E}}\, x_i^2 xx^\top + \sum_{i\neq j\in[n]} g_i g_j \cdot \tilde{\mathbb{E}}\, x_i x_j xx^\top .$$

Addressing the first term, by standard concentration, $\mathbb{E}\max_{i\in[n]} g_i^2 = O(\log n)$. So,

$$\mathop{\mathbb{E}}_{g}\left\|\sum_{i\in[n]} g_i^2\, \tilde{\mathbb{E}}\, x_i^2 xx^\top\right\| \leqslant \mathop{\mathbb{E}}_{g}\left[\max_{i\in[n]} g_i^2 \cdot \left\|\tilde{\mathbb{E}}\, \|x\|^2 xx^\top\right\|\right] = O(\log n)\cdot\left\|\tilde{\mathbb{E}}\, xx^\top\right\| = O(c\cdot\log n)\,.$$

The second term we will decouple using Fact 7.11.

$$\mathop{\mathbb{E}}_{g}\left\|\sum_{i\neq j} g_i g_j \cdot \tilde{\mathbb{E}}\, x_i x_j xx^\top\right\| \leqslant O(1)\cdot \mathop{\mathbb{E}}_{g,h}\left\|\sum_{i\neq j} g_i h_j \cdot \tilde{\mathbb{E}}\, x_i x_j xx^\top\right\| .$$

We add some aditional terms to the sum; by similar reasoning to our bound on the first term they do not contribute too much to the norm.

$$\mathop{\mathbb{E}}_{g,h}\left\|\sum_{i\neq j} g_i h_j \cdot \tilde{\mathbb{E}}\, x_i x_j xx^\top\right\| \leqslant O(1)\cdot \mathop{\mathbb{E}}_{g,h}\left\|\sum_{i,j\in[n]} g_i h_j \cdot \tilde{\mathbb{E}}\, x_i x_j xx^\top\right\| + O(c\cdot\log n)\,.$$

We can rewrite the matrix in the first term on the right-hand side as

$$\sum_{i,j\in[n]} g_i h_j \cdot \tilde{\mathbb{E}}\, x_i x_j xx^\top = \tilde{\mathbb{E}}\langle g, x\rangle\langle h, x\rangle xx^\top .$$

Now we can apply Fact 7.12 twice in a row; first to g and then to h, which together with our norm bound on $\mathbb{E}\, xx^\top \otimes xx^\top$, gives

$$\mathop{\mathbb{E}}_{g,h}\left\|\tilde{\mathbb{E}}\langle g, x\rangle\langle h, x\rangle xx^\top\right\| \leqslant O(c\cdot\log n)\,.$$

Putting all of the above together, we get the lemma. □

Next we prove Lemma 7.7 as a corollary of Lemma 7.7 which applies to random contractions which are non-spherical. The proof technique is very similar to that for Fact 7.12.

Proof of Lemma 7.7. Let $h \sim \mathcal{N}(0, \mathrm{Id} - \Sigma)$ be independent of g, and define $g' = g + h$ and $g'' = g - h$, so that $g = \frac{1}{2}(g' + g'')$. It is sufficient to bound $\mathbb{E}_{g,h}\left\|\tilde{\mathbb{E}}\langle g' + g'', x\rangle^2 xx^\top\right\|$. Expanding and applying triangle inequality,

$$\mathop{\mathbb{E}}_{g,h}\left\|\tilde{\mathbb{E}}\langle g' + g'', x\rangle^2 xx^\top\right\| \leqslant \mathop{\mathbb{E}}_{g,h}\left\|\tilde{\mathbb{E}}\langle g', x\rangle^2 xx^\top\right\| + 2\mathop{\mathbb{E}}_{g,h}\left\|\tilde{\mathbb{E}}\langle g', x\rangle\langle g'', x\rangle xx^\top\right\| + \mathop{\mathbb{E}}_{g,h}\left\|\tilde{\mathbb{E}}\langle g'', x\rangle^2 xx^\top\right\| .$$

The first and last terms are $O(c\cdot\log n)$ by Lemma 7.10. For the middle term, consider the quadratic form of the matrix $\tilde{\mathbb{E}}\langle g', x\rangle\langle g'', x\rangle xx^\top$ on a vector $v \in \mathbb{R}^n$:

$$\tilde{\mathbb{E}}\langle g', x\rangle\langle g'', x\rangle\langle x, v\rangle^2 \leqslant \tilde{\mathbb{E}}\langle g', x\rangle^2\langle x, v\rangle^2 + \tilde{\mathbb{E}}\langle g'', x\rangle^2\langle x, v\rangle^2$$

by pseudoexpectation Cauchy-Schwarz. Thus for every g', g'',

$$\left\|\tilde{\mathbb{E}}\langle g', x\rangle\langle g'', x\rangle xx^\top\right\| \leqslant \left\|\tilde{\mathbb{E}}\langle g', x\rangle^2 xx^\top\right\| + \left\|\tilde{\mathbb{E}}\langle g'', x\rangle^2 xx^\top\right\| .$$

Together with Lemma 7.10 this concludes the proof. □

7.2 Lifting 3-tensors to 4-tensors

Problem 7.13 (3-to-4 lifting)**.** Let $a_1, \dots, a_m \in \mathbb{R}^n$ be orthonormal. Let $A_3 = \sum_{i=1}^m a_i^{\otimes 3}$ and $A_4 = \sum_{i=1}^m a_i^{\otimes 4}$. Let $B \in \mathbb{R}^{n \times n \times n}$ satisfy $\langle B, A_3 \rangle \geqslant \delta \cdot \|A_3\| \cdot \|B\|$.
Input: The tensor B.
Goal: Output B' satisfying $\langle B', A_4 \rangle \geqslant \delta^{O(1)} \cdot \|A_4\| \cdot \|B'\|$.

Theorem 7.14. *There is a polynomial time algorithm, using the sum of squares method, which solves the 3-to-4 lifting problem.*

Proof. **Small δ regime: $\delta < 1 - \Omega(1)$:** The algorithm is to output the fourth moments of the optimizer of the following convex program.

$$\begin{aligned} \min_{\tilde{\mathbb{E}}} \quad & \|\tilde{\mathbb{E}}\, x^{\otimes 3}\| \\ \text{s.t.} \quad & \tilde{\mathbb{E}} \text{ is degree-4} \\ & \tilde{\mathbb{E}} \text{ satisfies } \{\|x\|^2 = 1\} \\ & \langle \tilde{\mathbb{E}}\, x^{\otimes 3}, B \rangle \geqslant \frac{\delta \|B\|}{\sqrt{m}} \\ & \|\tilde{\mathbb{E}}\, x^{\otimes 4}\| \leqslant \frac{1}{\sqrt{m}} . \end{aligned}$$

To analyze the algorithm we apply Theorem 2.3. Let C be the set of degree-4 pseudodistributions satisfying $\{\|x\|^2 = 1\}$ and having $\|\tilde{\mathbb{E}}\, x^{\otimes 4}\| \leqslant 1/\sqrt{m}$. The uniform distribution over $a_1, \dots, a_m$, whose third and fourth moments are $\frac{1}{m} A_3$ and $\frac{1}{m} A_4$, respectively, is in C.

Let $\tilde{\mathbb{E}}$ be the pseudoexpectation solving the convex program. By Theorem 2.3,

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 3}, \tfrac{1}{m} A_3 \rangle \geqslant \frac{\delta}{2} \cdot \frac{1}{\sqrt{m}} \cdot \|\tilde{\mathbb{E}}\, x^{\otimes 3}\| \geqslant \frac{\delta^2}{2m}$$

At the same time,

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 3}, \tfrac{1}{m} A_3 \rangle = \frac{1}{m} \sum_{i=1}^m \tilde{\mathbb{E}} \langle x, a_i \rangle^3 \leqslant \frac{1}{m} \left(\tilde{\mathbb{E}} \sum_{i=1}^m \langle x, a_i \rangle^4 \right)^{1/2}$$

by Cauchy-Schwarz. Putting these together, we obtain

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 4}, A_4 \rangle = \tilde{\mathbb{E}} \sum_{i=1}^m \langle x, a_i \rangle^4 \geqslant \delta^4 / 4 .$$

Finally, $\|A_4\| \cdot \|\tilde{\mathbb{E}}\, x^{\otimes 4}\| \leqslant 1$ (since we constrained $\|\tilde{\mathbb{E}}\, x^{\otimes 4}\| \leqslant 1/\sqrt{m}$), which finishes the proof.

Large δ regime: $\delta \geqslant 1 - o(1)$: Modify the convex program from the small-δ regime to project $(B/\|B\|) \cdot 1/\sqrt{m}$ to same convex set C. The normalization is so that

$$\left\| (B/\|B\|) \cdot 1/\sqrt{m} \right\| = \left\| \tfrac{1}{m} \cdot A_3 \right\| .$$

The analysis is similar. □

Acknowledgments

We are indebted to Avi Wigderson who suggested color coding as a technique to evaluate the kinds of polynomials we study in this work. We thank Moses Charikar for pointing out the relationship between our SOS program for low correlation tensor decomposition and Rényi entropy. We thank Christian Borgs, Jennifer Chayes, and Yash Deshpande for helpful conversations, especially relating to Section 4. We thank anonymous reviewers for many suggested improvements to this paper.

References

[Abb17] Emmanuel Abbe, *Community detection and stochastic block models: recent developments*, CoRR **abs/1703.10146** (2017). 9

[ABFX08] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing, *Mixed membership stochastic blockmodels*, NIPS, Curran Associates, Inc., 2008, pp. 33–40. 3, 9

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky, *Tensor decompositions for learning latent variable models*, Journal of Machine Learning Research **15** (2014), no. 1, 2773–2832. 1, 14

[AGHK13] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, and Sham Kakade, *A tensor spectral approach to learning mixed membership community models*, COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA (Shai Shalev-Shwartz and Ingo Steinwart, eds.), JMLR Workshop and Conference Proceedings, vol. 30, JMLR.org, 2013, pp. 867–881. 13, 21

[AGHK14] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, and Sham M. Kakade, *A tensor approach to learning mixed membership community models*, Journal of Machine Learning Research **15** (2014), no. 1, 2239–2312. 1, 9

[AN04] Noga Alon and Assaf Naor, *Approximating the cut-norm via grothendieck's inequality*, STOC, ACM, 2004, pp. 72–80. 29

[AS15] Emmanuel Abbe and Colin Sandon, *Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery*, FOCS, IEEE Computer Society, 2015, pp. 670–688. 4, 17

[AS16a] _____, *Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation*, NIPS, 2016, pp. 1334–1342. 1, 3, 5, 6, 12, 31, 32

[AS16b] _____, *Crossing the KS threshold in the stochastic block model with information theory*, ISIT, IEEE, 2016, pp. 840–844. 15

[AYZ95] Noga Alon, Raphael Yuster, and Uri Zwick, *Color-coding*, J. ACM **42** (1995), no. 4, 844–856. 4, 23

[BBAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, The Annals of Probability **33** (2005), no. 5, 1643–1697. 25

[BHK+16] Boaz Barak, Samuel B. Hopkins, Jonathan A. Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin, *A nearly tight sum-of-squares lower bound for the planted clique problem*, FOCS, IEEE Computer Society, 2016, pp. 428–437. 2, 9, 18, 19

[BKM17] Jess Banks, Robert Kleinberg, and Cristopher Moore, *The lovász theta function for random regular graphs and community detection in the hard regime*, APPROX-RANDOM, LIPIcs, vol. 81, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017, pp. 28:1–28:22. 16

[BKS15] Boaz Barak, Jonathan A. Kelner, and David Steurer, *Dictionary learning and tensor decomposition via the sum-of-squares method*, STOC, ACM, 2015, pp. 143–151. 1, 8, 14, 23

[BLM15] Charles Bordenave, Marc Lelarge, and Laurent Massoulié, *Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs*, FOCS, IEEE Computer Society, 2015, pp. 1347–1357. 31

[BMNN16] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli, *Information-theoretic thresholds for community detection in sparse networks*, COLT, JMLR Workshop and Conference Proceedings, vol. 49, JMLR.org, 2016, pp. 383–416. 17

[CJ10] Pierre Comon and Christian Jutten, *Handbook of blind source separation: Independent component analysis and applications*, Academic press, 2010. 14

[DKMZ11] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, CoRR **abs/1109.3041** (2011). 3, 10

[dlPnMS94] Victor H. de la Peña and S. J. Montgomery-Smith, *Bounds on the tail probability of U-statistics and quadratic forms*, Bull. Amer. Math. Soc. (N.S.) **31** (1994), no. 2, 223–227. MR 1261237 64

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the royal statistical society. Series B (methodological) (1977), 1–38. 3

[DM15] Yash Deshpande and Andrea Montanari, *Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 523–562. 72

[FGR+13] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao, *Statistical algorithms and a lower bound for detecting planted cliques*, STOC, ACM, 2013, pp. 655–664. 16

[FGV15] Vitaly Feldman, Cristobal Guzman, and Santosh Vempala, *Statistical query algorithms for stochastic convex optimization*, CoRR **abs/1512.09170** (2015). 16

[FPV15] Vitaly Feldman, Will Perkins, and Santosh Vempala, *On the complexity of random satisfiability problems with planted solutions*, STOC, ACM, 2015, pp. 77–86. 16

[Gal62] Robert Gallager, *Low-density parity-check codes*, IRE Transactions on information theory **8** (1962), no. 1, 21–28. 3

[GHK] Rong Ge, Qingqing Huang, and Sham Kakade, *Learning mixtures of gaussians in high dimensions, stoc 2015*, 761–770. 14

[GM15] Rong Ge and Tengyu Ma, *Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms*, APPROX-RANDOM, LIPIcs, vol. 40, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2015, pp. 829–849. 8

[GVX14] Navin Goyal, Santosh Vempala, and Ying Xiao, *Fourier PCA and robust tensor decomposition*, STOC, ACM, 2014, pp. 584–593. 14

[Har70] Richard A Harshman, *Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis*. 21

[HKP+17] Samuel B Hopkins, Pravesh Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer, *The power of sum-of-squares for detecting hidden structures*, Symposium on Foundations of Computer Science (2017). 9, 18, 19

[HSS15] Samuel B. Hopkins, Jonathan Shi, and David Steurer, *Tensor principal component analysis via sum-of-square proofs*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 956–1006. 2, 8

[Jer92] Mark Jerrum, *Large cliques elude the metropolis process*, Random Struct. Algorithms **3** (1992), no. 4, 347–360. 15

[Kuc95] Ludek Kucera, *Expected complexity of graph partitioning problems*, Discrete Applied Mathematics **57** (1995), no. 2-3, 193–212. 15

[LBB+16] Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová, *Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering*, Allerton, IEEE, 2016, pp. 601–608. 19

[LRA93] S. E. Leurgans, R. T. Ross, and R. B. Abel, *A decomposition for three-way arrays*, SIAM J. Matrix Anal. Appl. **14** (1993), no. 4, 1064–1083. MR 1238921 21

[Mas13] Laurent Massoulié, *Community detection thresholds and the weak ramanujan property*, CoRR **abs/1311.3085** (2013). 26, 27

[Mas14] _____, *Community detection thresholds and the weak ramanujan property*, STOC, ACM, 2014, pp. 694–703. 1, 3, 6

[MM09] Marc Mezard and Andrea Montanari, *Information, physics, and computation*, Oxford University Press, 2009. 19

[MNS12] Elchanan Mossel, Joe Neeman, and Allan Sly, *Stochastic block models and reconstruction*, arXiv preprint arXiv:1202.1499 (2012). 17

[MNS13] Elchanan Mossel, Joe Neeman, and Allan Sly, *A proof of the block model threshold conjecture*, CoRR **abs/1311.4115** (2013). 26, 27

[MNS15a] _____, *Consistency thresholds for the planted bisection model*, STOC, ACM, 2015, pp. 69–75. 1, 3, 5, 6, 19

[MNS15b] Elchanan Mossel, Joe Neeman, and Allan Sly, *Reconstruction and estimation in the planted partition model*, Probab. Theory Related Fields **162** (2015), no. 3-4, 431–461. MR 3383334 26

[Moo17] Cristopher Moore, *The computer science and physics of community detection: Landscapes, phase transitions, and hardness*, CoRR **abs/1702.00467** (2017). 12, 15

[MS16] Andrea Montanari and Subhabrata Sen, *Semidefinite programs on sparse random graphs and their application to community detection*, STOC, ACM, 2016, pp. 814–827. 16

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer, *Polynomial-time tensor decompositions with sum-of-squares*, FOCS, IEEE Computer Society, 2016, pp. 438–446. 14, 15, 23, 61, 63, 64

[MW15] Tengyu Ma and Avi Wigderson, *Sum-of-squares lower bounds for sparse PCA*, NIPS, 2015, pp. 1612–1620. 2

[Pea82] Judea Pearl, *Reverend bayes on inference engines: A distributed hierarchical approach*, Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California, Los Angeles, 1982. 3

[RRS16] Prasad Raghavendra, Satish Rao, and Tselil Schramm, *Strongly refuting random csps below the spectral threshold*, CoRR **abs/1605.00058** (2016). 8

[SS17] Tselil Schramm and David Steurer, *Fast and robust tensor decomposition with applications to dictionary learning*, COLT, Proceedings of Machine Learning Research, vol. 65, PMLR, 2017, pp. 1760–1793. 14, 15, 60

[Wik17a] Wikipedia, *Bayes estimator — Wikipedia, the free encyclopedia*, http://en.wikipedia.org/w/index.php?title=Bayes%20estimator&oldid=754605088, 2017, [Online; accessed 30-March-2017]. 1

[Wik17b] _____, *Dirichlet distribution — Wikipedia, the free encyclopedia*, http://en.wikipedia.org/w/index.php?title=Dirichlet%20distribution&oldid=762020989, 2017, [Online; accessed 30-March-2017]. 10

A Toolkit and Omitted Proofs

A.1 Probability and linear algebra tools

Fact A.1. *Consider any inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$ with associated norm $\|\cdot\|$. Let X and Y be jointly-distributed $\mathbb{R}^n$-valued random variables. Suppose that $\|X\|^2 \leqslant C\,\mathbb{E}\,\|X\|^2$ with probability 1, and that*

$$\frac{\mathbb{E}\langle X, Y\rangle}{(\mathbb{E}\,\|X\|^2)^{1/2}(\mathbb{E}\,\|Y\|^2)^{1/2}} \geqslant \delta\,.$$

Then

$$\mathbb{P}\left\{\frac{\langle X, Y\rangle}{\|X\| \cdot \|Y\|} \geqslant \frac{\delta}{2}\right\} \geqslant \frac{\delta^2}{4C^2}\,.$$

Proof of Fact A.1. Let $\mathbf{1}_E$ be the 0/1 indicator of an event E. Note that

$$\mathbb{E}\left[\langle X, Y\rangle\,\mathbf{1}_{\langle X,Y\rangle \leqslant \frac{\delta}{2}\cdot\|X\|\cdot\|Y\|}\right] \leqslant \tfrac{\delta}{2}\,\mathbb{E}\,\|X\| \cdot \|Y\| \leqslant \tfrac{\delta}{2}(\mathbb{E}\,\|X\|^2)^{1/2}(\mathbb{E}\,\|Y\|^2)^{1/2}\,.$$

Hence,

$$\mathbb{E}\left[\langle X, Y\rangle\,\mathbf{1}_{\langle X,Y\rangle > \frac{\delta}{2}\cdot\|X\|\cdot\|Y\|}\right] \geqslant \tfrac{\delta}{2}\,\mathbb{E}\,\|X\| \cdot \|Y\| \leqslant \tfrac{\delta}{2}(\mathbb{E}\,\|X\|^2)^{1/2}(\mathbb{E}\,\|Y\|^2)^{1/2}\,.$$

At the same time,

$$\begin{aligned}
\mathbb{E}\left[\langle X, Y\rangle\,\mathbf{1}_{\langle X,Y\rangle > \frac{\delta}{2}\cdot\|X\|\cdot\|Y\|}\right] &\leqslant \left(\mathbb{E}\,\|X\|^2 \cdot \|Y\|^2\right)^{1/2} \cdot \left(\mathbb{E}\,\mathbf{1}_{\langle X,Y\rangle > \frac{\delta}{2}\cdot\|X\|\cdot\|Y\|}\right)^{1/2} \\
&= \left(\mathbb{E}\,\|X\|^2 \cdot \|Y\|^2\right)^{1/2} \cdot \left(\mathbb{P}\{\langle X, Y\rangle > \tfrac{\delta}{2} \cdot \|X\| \cdot \|Y\|\}\right)^{1/2} \\
&\leqslant C(\mathbb{E}\,\|X\|^2)^{1/2}(\mathbb{E}\,\|Y\|^2)^{1/2} \cdot \left(\mathbb{P}\{\langle X, Y\rangle > \tfrac{\delta}{2} \cdot \|X\| \cdot \|Y\|\}\right)^{1/2}\,.
\end{aligned}$$

Putting the inequalities together and rearranging finishes the proof. □

Proof of Proposition 5.22. We decompose X_i as

$$X_i = X_i\,\mathbf{1}_{|X_i| \leqslant R} + X_i\,\mathbf{1}_{|X_i| > R}\,.$$

Let $Y_i = X_i\,\mathbf{1}_{|X_i| \leqslant R}$. Then

$$|\,\mathbb{E}\,Y_i| = |\,\mathbb{E}\,X_i - \mathbb{E}\,X_i\,\mathbf{1}_{|X_i| > R}\,| \leqslant \delta'$$

and

$$\mathbb{V}\,Y_i \leqslant \mathbb{E}\,Y_i^2 \leqslant \mathbb{E}\,X_i^2\,.$$

So we can apply Bernstein's inequality to $\frac{1}{m}\sum_{i \leqslant m} Y_i$ to obtain that

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i \leqslant m} Y_i\right| \geqslant t + \delta'\right\} \leqslant \exp\left(\frac{-\Omega(1) \cdot m \cdot t^2}{\mathbb{E}\,X^2 + t \cdot R}\right)\,.$$

Now, with probability at least $1 - \delta$ we know $X_i = Y_i$, so by a union bound,

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i \leqslant m} X_i\right| \geqslant t + \delta'\right\} \leqslant \exp\left(\frac{-\Omega(1) \cdot m \cdot t^2}{\mathbb{E}\,X^2 + t \cdot R}\right) + m\delta\,.$$ □

Fact A.2. *Let $\{X_1, \dots, X_n, Y_1, \dots, Y_m\}$ are jointly distributed real-valued random variables. Suppose there is $S \subseteq [m]$ with $|S| \geqslant (1 - o_m(1)) \cdot m$ such that for each $i \in S$ there a degree-D polynomials p_i satisfying*

$$\frac{\mathbb{E}\, p_i(X) Y_i}{(\mathbb{E}\, Y^2)^{1/2} (\mathbb{E}\, p_i(X)^2)^{1/2}} \geqslant \delta\,.$$

Furthermore, suppose $\sum_{i \in S} \mathbb{E}\, Y_i^2 \geqslant (1 - o(1)) \sum_{i \in [m]} \mathbb{E}\, Y_i^2$. Let $Y \in \mathbb{R}^m$ be the vector-valued random variable with i-th coordinate Y_i, and similarly let $P(X)$ have i-th coordinate $p_i(X)$. Then

$$\frac{\mathbb{E}\langle P(X), Y\rangle}{(\mathbb{E}\, \|Y\|^2)^{1/2} \cdot (\mathbb{E}\, \|P(X)\|^2)^{1/2}} \geqslant (1 - o(1)) \cdot \delta$$

Proof. The proof is by Cauchy-Schwarz.

$$\begin{aligned}
\mathbb{E}\langle P(X), Y\rangle &= \sum_{i \in S} \mathbb{E}\, p_i(X) Y \\
&\geqslant \delta \sum_{i \in S} (\mathbb{E}\, p_i(X)^2)^{1/2} (\mathbb{E}\, Y_i^2)^{1/2} \\
&\geqslant \delta \left(\mathbb{E} \sum_{i \in S} p_i(x)^2 \right)^{1/2} \cdot (1 - o(1)) \left(\sum_{i \in [m]} Y_i^2 \right)^{1/2}.
\end{aligned}$$

□

A.2 Tools for symmetric and Dirichlet priors

Proof of Fact 5.15. Let X be any $\mathbb{R}^k$-valued random variable which is symmetric in distribution with respect to permutations of coordinates and satisfies $\sum_{s \in [k]} X(s) = 0$ with probability 1. (The variable $\tilde{\sigma}$ is one example.)

We prove the claim about $\mathbb{E}\langle X, x\rangle\langle X, y\rangle\langle X, z\rangle\langle X, w\rangle$; the other proofs are similar. Consider the matrix $M = \mathbb{E}(X \otimes X)(X \otimes X)^\top$. Since x, y, z, w are orthogonal to the all-1's vector, we may add $1 \otimes v$, for any $v \in \mathbb{R}^n$, to any row or column of M without affecting the statement to be proved. Adding multiples of $1 \otimes e_i$ to rows and columns appropriately makes M a block diagonal matrix, with the top block indexed by coordinates (i, i) for $i \in [k]$ and the bottom block indexed by pairs (i, j) for $i \neq j$.

The resulting top block takes the form $c\mathrm{Id} + c' J$, where J is the all-1's matrix. The bottom block will be a matrix from the Johnson scheme. Standard results on eigenvectors of the Johnson scheme (see e.g. [DM15] and references therein) finish the proof. The values of constants C for the Dirichlet distribution follow from the next fact. □

Fact A.3. *Let $\sigma \in \mathbb{R}^k$ be distributed according to a (symmetric) Dirichlet distribution with parameter α. That is, $\mathbb{P}(\sigma) \propto \prod_{j \in [k]} \sigma^{\alpha - 1}$.*

Let $\gamma \in \mathbb{N}^k$ be a k-tuple, and let $\sigma^\gamma = \prod_{j \leqslant k} \sigma_j^{\gamma_j}$. Let $|\gamma| = \sum_{j \leqslant k} \gamma_j$. Then

$$\mathbb{E}\, \sigma^\gamma = \frac{\Gamma(k\alpha)}{\Gamma(k\alpha + |\gamma|)} \cdot \frac{\prod_{j \leqslant k} \Gamma(\alpha + \gamma_j)}{\Gamma(\alpha)^k}\,.$$

Furthermore, let $\tilde{\sigma} \in \mathbb{R}^k$ *be given by* $\tilde{\sigma}_i = \sigma_i - \frac{1}{k}$*. Then*

$$\mathbb{E}\,\tilde{\sigma}\tilde{\sigma}^\top = \frac{\Gamma(k\alpha)}{\Gamma(k\alpha+2)}\left(\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} - \frac{\Gamma(\alpha+1)^2}{\Gamma(\alpha)^2}\right)\cdot\Pi = \frac{1}{k(k\alpha+1)}\cdot\Pi\,,$$

where $\Pi \in \mathbb{R}^{k\times k}$ *is the projector to the subspace orthogonal to the all-*1*s vector.*

Proof. We recall the density of the k-dimensional Dirichlet distribution with parameter vector $\alpha_1, \ldots, \alpha_k$. Here Γ denotes the usual Gamma function.

$$\mathbb{P}\{\sigma\} = \frac{\Gamma(\sum_{j\leqslant k}\alpha_j)}{\prod_{j\leqslant k}\Gamma(\alpha_j)}\cdot\prod_{j\leqslant k}\sigma_j^{\alpha_j-1}\,.$$

In particular,

$$\frac{\Gamma(\sum_{j\leqslant k}\alpha_j)}{\prod_{j\leqslant k}\Gamma(\alpha_j)}\cdot\int\prod_{j\leqslant k}\sigma_j^{\alpha_j-1}\,d\sigma = 1$$

where the integral is taken with respect to Lebesgue measure on $\{\sigma : \sum_{j\leqslant k}\sigma_j = 1\}$.

Using this fact we can compute the moments of the symmetric Dirichlet distribution with parameter α. We show for example how to compute second moments; the general formula can be proved along the same lines. For $s \neq t \in [k]$,

$$\begin{aligned}
\mathbb{E}\,\sigma_s\sigma_t &= \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k}\cdot\int\sigma_s\sigma_t\prod_{j\leqslant k}\sigma_j^{\alpha-1}\\
&= \frac{\Gamma(k\alpha)}{\Gamma(k\alpha+2)}\cdot\frac{\Gamma(\alpha+1)^2}{\Gamma(\alpha)^2}\cdot\frac{\Gamma(k\alpha+2)}{\Gamma(\alpha)^{k-2}\Gamma(\alpha+1)^2}\cdot\int\sigma_s^{(\alpha+1)-1}\sigma_t^{(\alpha+1)-1}\prod_{j\neq s,t}\sigma_j^{\alpha-1}\\
&= \frac{\Gamma(k\alpha)}{\Gamma(k\alpha+2)}\cdot\frac{\Gamma(\alpha+1)^2}{\Gamma(\alpha)^2}\,.
\end{aligned}$$

Similarly,

$$\mathbb{E}\,\sigma_s^2 = \frac{\Gamma(k\alpha)}{\Gamma(k\alpha+2)}\cdot\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)}\,.$$

The formula for $\mathbb{E}\,\tilde{\sigma}\tilde{\sigma}^\top$ follows immediately. □