58th Annual IEEE Symposium on Foundations of Computer Science

Efficient Bayesian estimation from few samples: community detection and related problems

Samuel B. Hopkins
Cornell University
Ithaca, USA
samhop@cs.cornell.edu

David Steurer
Cornell/IAS
Princeton, USA
dsteurer@cs.cornell.edu

***Abstract*—**
We propose an efficient meta-algorithm for Bayesian inference problems 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 partial recovery in the stochastic block model, a widely-studied class of inference problems for community detection in graphs. We obtain the first partial recovery guarantees for the mixed-membership stochastic block model (Airoldi et el.) for constant average degree—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.

***Keywords*-Bayesian inference, stochastic block-model, low-degree polynomials, tensor decomposition, semidefinite programming, sum of squares algorithms, average-case hardness, phase transitions**

I. Introduction

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

Many ubiquitous statistical inference and unsupervised learning problems fit this description: independent component analysis, variants of principal component analysis, and planted problems—for example, planted constraint satisfaction problems and planted coloring problems—are just some examples.

In this paper we develop general algorithmic techniques for such problems and apply them to the stochastic blockmodel. In the blockmodel, the parameter θ is a labeling of each of $[n]$ nodes with one or more *communities*. The samples $x_1, \dots, x_m$ are edges $(i, j) \in [n]^2$, generated with a bias towards pairs (i, j) from the same community (or sharing many communities). Taken together these inputs form a random graph x on $[n]$ which exhibits latent community structure; the goal is to estimate the community memberships θ. The problem becomes easier as more samples are generated, which is to say as the average degree of the graph x increases. The question is: how many samples (i.e. what average degree of x) is required for an efficient algorithm to estimate θ?

Our main contribution is a meta-algorithm for Bayesian estimation problems. The algorithm is based on low-degree polynomials and related to the method of moments. We apply the algorithm to recover the best-known polynomial time sample complexity guarantees for the stochastic blockmodel. We give the first sample complexity guarantees in the regime where x has constant degree and each node may participate in several communities simultaneously. Our work, and especially our meta-algorithm, unifies several previous lines of work on this problem: algorithms based on the *belief propagation* method from statistical physics,

0272-5428/17 $31.00 © 2017 IEEE
DOI 10.1109/FOCS.2017.42

algorithms based on *tensor decomposition*, and algorithms based on *semidefinite programming*. Our meta-algorithm gives a recipe to design algorithms which achieve the tight sample complexity guarantees of belief propagation, share the broad applicability of tensor decomposition methods, and leverage ideas from semidefinite programming and the sum-of-squares hierarchy to obtain provable bounds.

In this proceedings version we give an overview of our main results and proof techniques. Full proofs are deferred to the full version of this work.

Why focus on sample bounds?.: Our focus is on obtaining the tightest possible sample complexity bounds (ideally precise up to low-order additive terms) achievable by polynomial time algorithms. Unlike running times, which vary by constant factors (at least) with only subtle changes in underlying computational models, sample complexity bounds can be studied very precisely. In this respect, from the perspective of algorithms research, sample complexity bounds are more akin to approximation ratios than they are to running times. As with approximation ratios, even constant factor improvements in sample complexity bounds often involve substantial algorithmic insights, and much can be learned from studying the optimal sample complexity of an estimation problem.

Obtaining the best possible sample complexity bounds is also an important stepping stone to answering the question: how complex can unsupervised learning models be while permitting efficient learning from limited data? Modern unsupervised learning models, like deep neural networks, are specified by enormous numbers of parameters; to understand how sophisticated we can make these models and maintain polynomial-time learnability from reasonable numbers of samples we need a precise theory of sample complexity.

Relation to previous approaches and phase transitions in sample complexity..: The stochastic blockmodel has been studied for decades, and parameter estimation is among the oldest problems in statistics. (We defer a more thorough overview of previous work on the stochastic blockmodel till later in this paper.) However, precise sample complexity bounds for variants of the blockmodel are a relatively recent development. Algorithms achieving these bounds are sophisticated and often intricate to analyze. Using ideas from semidefinite programming and Fourier analysis, we are able to unify the analysis of preious algorithms using belief propagation, moment methods, and tensor decomposition.

Furthermore, many precise sample complexity bounds or sample complexity "phase transitions"—a number of samples fewer than which estimation suddenly becomes impossible for efficient algorithms—were initially studied using ideas from statistical physics. These ideas are able to heuristically predict the locations of and give physical explanations for these phase transition phenomena. However, until now rigorous verification of these predictions—by designing efficient and provable algorithms—required innovation on a case-by-case basis.

We provide a recipe to design algorithms achieving the predicted optimal sample complexity guarantee and a recipe for their analysis. We give a physics-free explanation for the origin and locations of these phase transitions, based on the low-degree structure of the underlying probability distributions $p(x, \theta)$. In addition to an algorithm-design recipe, this explanation allows rigorous study of the other side of the phase transition, where physics methods predict that there are too few samples for computationally-efficient algorithms to perform parameter estimation. We prove impossibility results for a class of efficient algorithms based on low-degree polynomials in the stochastic blockmodel setting.

Relation to pseudocalibration and the sum-of-squares algorithm..: This work is related to recent work of Barak et al [BHK$^+$16] and concurrent work of Hopkins et al [HKP$^+$17] on the power of sum-of-squares algorithms for planted problems, closely related to Bayestian estimation. Those works also demonstrate that properties of the low-degree structure of distributions $p(x, \theta)$ play a role in the complexity of inferring hidden variables from samples. These works focus on structural theorems and lower bounds for estimation problems, while here our main focus is on algorithms.

Those works study sum-of-squares algorithms—a powerful family of semidefinite programs—which we also utilize here in some of our algorithms. The goal of the work [HKP$^+$17] is to obtain a very general characterization of sum-of-squares for many planted problems at once. As such, the results are much less fine-grained than those we present here: they can be interpreted only as understanding the numbers of samples required

for sum-of-squares algorithms up to subpolynomial factors, while here we focus on very precise bounds, which require substantially different algorithmic ideas.

A. Detecting overlapping communities

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 blockmodel graphs under (conjecturally) optimal sparsity conditions. For an excellent survey, see [Abbar].

Such sharp results remain limited to relatively simple versions of the blockmodel; 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. The mixed-membership stochastic block model [ABFX08] is one such natural extension of the stochastic block model that allows for communities to overlap. 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 I.1 (Mixed-membership stochastic block model)**.** The mixed-membership stochastic block model $\mathrm{SBM}(n, d, \varepsilon, k, \alpha)$ is the following distribution over n-vertex graphs G and k-dimensional probability vectors $\sigma_1, \dots, \sigma_n$ for the vertices:

- draw $\sigma_1, \dots, \sigma_n$ independently from $\mathrm{Dir}(\alpha)$ the symmetric k-dimensional Dirichlet distribution with parameter $\alpha \geqslant 0$,[1]
- for every potential edge $\{i, j\}$, add it to G with probability $\frac{d}{n} \cdot \left(1 + \left(\langle \sigma_i, \sigma_j \rangle - \frac{1}{k}\right)\varepsilon\right)$.[2]

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 $\mathbf{1}_1, \dots, \mathbf{1}_k$ and the model becomes $\mathrm{SBM}(n, d, \varepsilon, k)$, 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 $\mathrm{Dir}(\alpha)$ 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.[3] 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.: Our algorithm for stochastic block models returns a list of vectors $L \subseteq \mathbb{R}^n$. The intention is that L consists of the indicator vectors for the k communities. If we let $\sigma = (\sigma_1, \dots, \sigma_n)$ be a labeling of the vertices by k-dimensional probability vectors, we define the *correlation* $\mathrm{corr}(\sigma, L)$ to be the minimum of the following two quantities:

1)
$$\max_{\tilde{Y} \in M_k(L)} \frac{\langle Y_\sigma, \tilde{Y} \rangle}{\|Y_\sigma\|_F \cdot \|\tilde{Y}\|_F}$$

2)
$$\min_{\tilde{Y} \in M_k(L), s \in [k]} \max_{t \in [k]} \frac{\langle \tilde{Y}\,\mathbf{1}_s, Y_\sigma\,\mathbf{1}_t \rangle}{\|\tilde{Y}\,\mathbf{1}_s\| \cdot \|Y_\sigma\,\mathbf{1}_t\|}$$

Here, $Y_\sigma \in \mathbb{R}^{n \times k}$ is the matrix with rows $\sigma_1 - \frac{1}{k}\mathbf{1}, \dots, \sigma_n - \frac{1}{k}\mathbf{1}$ (so that $\mathbb{E}_\sigma Y_\sigma = 0$ by symmetry) and $M_k(L) \subseteq \mathbb{R}^{n \times k}$ consists of all matrices with columns $u_1 - \frac{1}{k}\mathbf{1}, \dots, u_k - \frac{1}{k}\mathbf{1}$ such that $u_1, \dots, u_k \in L$. All norms are Euclidean and the inner products correspond to the norms.

The best-possible L consists of the columns of the matrix with rows $\sigma_1, \dots, \sigma_n$. This choice of L has correlation $\mathrm{corr}(\sigma, L) = 1$. To illustrate this notion of correlation, consider the case of disjoint communities (i.e., $\alpha = 0$). Suppose $\mathrm{corr}(\sigma, L) \geqslant \delta$.

[1] 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 $\mathrm{Dir}(0)$ to be the uniform distribution over the coordinate vectors $\mathbf{1}_1, \dots, \mathbf{1}_k$.

[2] An equivalent, more operational description of this process is that for every potential edge $\{i, j\}$, we draw labels $s \sim \sigma_i$, $t \sim \sigma_j$ and add the edge with probability $\frac{d}{n} \cdot (1 + (1 - \frac{1}{k})\varepsilon)$ if $s = t$ and with probability $\frac{d}{n} \cdot (1 - \frac{1}{k}\varepsilon)$ if $s \neq t$.

[3] 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.

Then, the condition $\langle Y_\sigma, \tilde{Y}\rangle \geqslant \delta \cdot \|Y_\sigma\| \cdot \|\tilde{Y}\|$ means that by looking at the large coordinates of $\tilde{Y}$ we can correctly identify the community membership of a $\delta^{O(1)} + \frac{1}{k}$ fraction of the vertices. The second term in the correlation definition means that every vector in L is δ-correlated with the indicator vector of one of the communities.

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 small list of vectors that has constant correlation with the true labeling (similar in spirit to list-decoding). Here, the parameters $d, \varepsilon, k, \alpha$ of the mixed-membership stochastic block model may even depend on the number of vertices n.

Theorem I.2 (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 $\varepsilon^2 d \geqslant (1+\delta)\cdot k^2(\alpha+1)^2$ for some $\delta > 0$. Then, there exists $\delta' \geqslant \min\{\delta^{O(1)}, 0.1\} > 0$ and a polynomial-time algorithm that given an n-vertex graph G outputs a list of vectors $L(G) \subseteq [0,1]^n$ of size $|L(G)| \leqslant k^{1/\delta'}$ satisfying*

$$\mathop{\mathbb{E}}_{(G,\sigma)\sim \mathrm{SBM}(n,d,\varepsilon,k,\alpha)} \operatorname{corr}(\sigma, L(G)) \geqslant \delta' . \tag{I.1}$$

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.

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. This condition is called the Kesten–Stigum threshold and is exactly the threshold achieved by previous best polynomial-time algorithms[4] (due to [Mas14], [MNS15] for $k = 2$ and [AS16] for general k). For $k = 2$, our notion of correlation is equivalent to the ones in previous works [Mas14], [MNS15]. For general k, our notion of correlation is stronger. Previous algorithms output a single vector $\tilde{y} \in \mathbb{R}^n$ such that $\langle Y_\sigma \mathbf{1}_S, \tilde{y}\rangle \geqslant \delta' \cdot \|Y_\sigma \mathbf{1}_S\| \cdot \|\tilde{y}\|$ for a subset $S \subseteq [k]$ of communities.[5] The idea is that y corresponds to a single community or a union of up to $k-1$ communities. The difference between these notions of correlation can be a multiplicative factor of k.

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 every polynomial-time algorithm that given a graph G outputs a polynomial-size list of vectors $L(G)$ satisfies

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{(G,\sigma)\sim \mathrm{SBM}(n,d,\varepsilon,k,\alpha)} \operatorname{corr}(\sigma, L(G)) = 0 . \tag{I.2}$$

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 e.g., [Moo17]. For large enough values of k, this conjecture predicts a computation-information gap because the condition $\varepsilon^2 d \geqslant O(k)$ is enough for achieving constant correlation information-theoretically (and in fact by a simple exponential-time algorithm).

Comparison to previous algorithms for mixed-membership models.: The best previous algorithms for stochastic block models with overlapping communities, i.e., α is (significantly) larger than 0, require $\varepsilon^2 d \geqslant O(\log n)^{O(1)} \cdot k^2(\alpha+1)^2$ [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.) We remark that there is a stark difference between the algorithmic techniques that go into detecting disjoint communities and overlapping ones. The former is based on matrices and pairwise correlations whereas the latter is based on tensors and higher-order correlations. In order to achieve the Kesten–Stigum threshold for disjoint communities, the key ingredients are spectral properties of matrices that correspond to non-standard random walks like self-avoiding or non-backtracking ones. Our algorithm for overlapping communities is based on higher-order tensors associated with such non-standard random walks. To analyze these tensors, we introduce new kinds of random walks called *colorful random walks* inspired by color coding [AYZ95], which greatly simplify the analysis (even for disjoint communities). The most involved part of our algorithm is a new

[4] Here, achieving the Kesten–Stigum threshold means that if $\varepsilon^2 d/k^2 - 1 > 0$ is lower bounded by any constant, then the algorithm achieves constant correlation with the true community structure (for a notion of correlation similar to ours).

[5] Algorithms in previous works typically output subsets of vertices as opposed to vectors in $\mathbb{R}^n$. For disjoint communities, there is little difference between the two. In particular, if we have a set of vectors L, we can convert it to a set of L' of 0/1 vectors by a randomized algorithm such that $\mathbb{E}_{L'} \operatorname{corr}(\sigma, L') \geqslant \Omega(1) \cdot \operatorname{corr}(\sigma, L)^{O(1)}$.

algorithm based on sum-of-squares to decompose tensors that have only constant correlation with an orthogonal tensor. Previous algorithms require correlation close to 1 [MSS16], [SS17] or even inverse-polynomial distance.

Recovering overlapping communities with higher accuracy.: Our above theorem focuses on providing an estimate that has constant correlation with the ground truth. The following theorem shows that if we are a constant multiplicative factor above the threshold, so that $\varepsilon^2 d \gg k^2(\alpha+1)^2$, we can achieve correlation close to 1 with the ground truth. In this case, our algorithm outputs exactly k vectors as opposed to $k^{O(1)}$ vectors in the previous theorem.

Theorem I.3 (Mixed-membership SBM—high accuracy)**.** *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 $\varepsilon^2 d \geqslant 1/\eta \cdot k^2(\alpha+1)^2$ for some η with $0 < \eta < 0.9$. Then, there exists $\eta' \leqslant \eta^{\Omega(1)}$ and a polynomial-time algorithm that given an n-vertex graph G outputs a list of vectors $L(G) \subseteq [0,1]^n$ of size $|L(G)| = k$ satisfying*

$$\mathop{\mathbb{E}}_{(G,\sigma)\sim \mathrm{SBM}(n,d,\varepsilon,k,\alpha)} \operatorname{corr}(\sigma, L(G)) \geqslant 1-\eta' . \qquad \text{(I.3)}$$

The above theorem is byproduct of our proof of Theorem I.2. We regard Theorem I.2 as our main result for the mixed-membership stochastic block model because it gives non-trivial guarantees all the way up to what we believe to be the threshold. Theorem I.3 by itself could potentially be proved in a more direct way and with a better trade-off between the multiplicative distance η from the threshold and the achieved accuracy η'.

B. Meta-theorems for Bayesian estimation

Our Theorem I.2 is an instantiation of a meta-algorithm for Bayesian estimation. In the following, we describe the guarantees of this meta-algorithm in greater generality.

We first consider a version of the meta-algorithm that is enough to capture the stochastic block model with two disjoint communities. Let $p(x,y)$ be a joint probability distribution over observable variables $x \in \mathbb{R}^n$ and hidden variables $y \in \mathbb{R}^d$. 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 $\mathbb{E}_{p(x,y)} \big(\hat{y}(x) - y\big)^2$ 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 dimensions n or d. There are many well-known algorithmic approach that aim to address this issue or related ones, for example, belief propagation or expectation maximization. While these approaches appear to work well in practice, they are notoriously difficult to analyze. In this work, we can resolve this issue in a very simple way: We analytically determine a low-degree polynomial $f(x)$ so that $\mathbb{E}_{f(x,y)} \big(f(x) - y\big)^2$ is as small as possible and use the simple fact that low-degree polynomials can be evaluated efficiently (even for high dimensions n).[6] In this way, we can capture linearized variants of belief propagation (e.g., [AS15]) and spectral properties of linear operators that have low-degree in the observable variables x (e.g., the Hashimoto non-backtracking operator).

The second issue is that even if we can compute the posterior mean, 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 would suffice $\mathbb{E}_{p(y|x)}\, yy^\intercal$. (For overlapping communities, we need third moments $\mathbb{E}_{p(y|x)}\, y^{\otimes 3}$ due to more substantial symmetries.)

Theorem I.4 (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-ℓ polyno-*

[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].

[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.

mial $P(x)$ such that

$$\mathop{\mathbb{E}}_{p(x,y)} \langle P(x), y y^{\intercal} \rangle \geqslant \delta \cdot \left(\mathop{\mathbb{E}}_{p(x)} \|P(x)\|_F^2 \right)^{1/2} . \tag{I.4}$$

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

$$\mathop{\mathbb{E}}_{p(x,y)} \mathop{\mathbb{E}}_{p(x,y)} \langle \hat{y}(x), y \rangle^2 \geqslant \delta' . \tag{I.5}$$

Using the appropriate polynomial P, this theorem captures the best known guarantees for partial recovery in the k-community stochastic block-model. 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 $n^{O(\ell)}$-time algorithm exists with the same guarantees.

The following theorem is an analogue of Theorem I.4 for 3rd order moments and the key ingredient for Theorem I.2. The advantage of this theorem compared to the previous one is that it can deal with the case that the posterior distribution exhibits more symmetries. When we apply this theorem for detecting k communities, the vector x is the adjacency matrix of a graph and $y_1, \dots, y_k$ are related to the indicator vectors of the k communities.[8]

Theorem I.5 (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,[9] orthonormal[10] 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 $x \in \{0,1\}^n$.[11] Suppose there exists a tensor-valued degree-ℓ polynomial $P(x)$ such that*

$$\mathop{\mathbb{E}}_{p(x,y_1,\dots,y_k)} \langle P(x), \sum_{i=1}^k y_i^{\otimes 3} \rangle \geqslant \delta \cdot \left(\mathop{\mathbb{E}}_{p(x)} \|P(x)\|^2 \right)^{1/2} \cdot \sqrt{k} . \tag{I.6}$$

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

$$\mathop{\mathbb{E}}_{p(x,y_1,\dots,y_k)} \mathop{\mathbb{E}}_{i \in [k]} \max_{\hat{y} \in L(x)} \langle \hat{y}, y_i \rangle \geqslant \delta' . \tag{I.7}$$

Previous algorithmic results for overlapping communities [AGHK13] can be viewed as instances of this meta-theorem. However, for this meta-theorem we use a more robust tensor decomposition algorithm than in previous algorithms, which allows us to tolerate more error in the estimate P of the tensor $\sum_{i=1}^k y_i^{\otimes 3}$. This improvement is not on its own sufficient: we also need to use a more sophisticated polynomial P (which will have degree roughy $\log n$) than the constant-degree polynomials used by previous work. The latter improvement is akin to the difference between the information carried by the adjacency matrix of a graph and that carried by the matrix whose (i, j)-th entry is the number of $\log n$-length simple paths from i to j.

This meta-theorem also captures algorithms for other estimation problems based on tensor methods, e.g., latent Dirichlet allocation [AFH+12], learning spherical mixtures of Gaussians [HK13], and independent component analysis [VX15]. As for detecting overlapping communities, it is plausible that more careful choices of polynomial estimators give improved sample bounds for these problems as well.

C. Concrete unconditional lower bounds for Bayesian estimation

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. More recently, sum-of-squares lower bounds emerged as a promising tool to understand the computational complexity of certain classes of average-case problems [Gri01], [Sch08], [HSS15], [MW15], [BHK+16]. However, before our work, it was not clear if or in what sense sum-of-squares algorithms can achieve the kind of precise sample bounds we seek for stochastic block models and related problems.

Our meta-algorithm for Bayesian estimation opens up a new avenue of research to give evidence for the computational difficulty of estimation problems and inherent gaps between the

[8] Concretely, for community detection the vectors $y_1, \dots, y_k$ are obtained by orthogonalizing the centered indicator vectors of the k communities.

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

[10] 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 $\|y_i\|^2 = 1$.

[11] 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.

information-theoretic threshold for estimation and the threshold for efficient estimation algorithms. Concretely, Theorems I.4 and I.5 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. (I.4) and (I.6)). This kind of fact about low-degree polynomial is something we can aim to prove unconditionally as a way to give evidence for the intracability of a Bayesian estimation problem.

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 slightly notion of correlation different than we have yet discussed. 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 I.6. *Let d, ε, k be constants. Then,*

$$\max_{p \in \mathbb{R}[x]_{\leqslant \ell}} \frac{\mathbb{E}_{x \sim \mathrm{SBM}(n,d,\varepsilon,k)} p(x)}{\left(\mathbb{E}_{x \sim G(n,d/n)} p(x)^2\right)^{1/2}} \text{ is} \tag{I.8}$$

- $\geqslant n^{\Omega(1)}$ *if* $\varepsilon^2 d > k^2$, $\ell \geqslant O(\log n)$
- $\leqslant n^{o(1)}$ *if* $\varepsilon^2 d < k^2$, $\ell \leqslant n^{o(1)}$

In the full version of this paper, we show that this theorem also implies a sharp threshold for polynomials which estimate, say, $\mathbf{1}_{\text{vertices 1 and 2 in same community}}$; nontrivial estimation of this random variable by degree $\ell \leqslant n^{o(1)}$ polynomials is possible only when $d > k^2/\varepsilon^2$.

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. (I.8) is equal to $\|\mu^{\leqslant \ell}\|_2$, where $\|\cdot\|$ 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. It follows that we can analyze $\|\mu^{\leqslant \ell}\|_2$ using Fourier analysis over the $\frac{d}{n}$-biased hypercube. We defer the proof of Theorem I.6 to the full version of this paper.

D. Low-correlation tensor decomposition

The algorithmically most involved part of our tensor-based estimation is to decompose tensors that have small correlation with orthogonal tensors.

Theorem I.7. *There exists a polynomial-time algorithm that given a 3-tensor $T \in (\mathbb{R}^n)^{\otimes 3}$ and a parameter δ outputs a list of unit vectors $L(T)$ of cardinality $|L(t)| \leqslant n^{1/\delta'}$ for $\delta' \geqslant \delta^{O(1)}$ with the following property: if T satisfies $\langle T, \sum_{i=1}^k a_i^{\otimes 3} \rangle$ for some orthonormal vectors $a_1, \ldots, a_k$, then*

$$\mathop{\mathbb{E}}_{i \in [k]} \max_{\hat{a} \in L(T)} \langle \hat{a}, a_i \rangle \geqslant \delta' .$$

In words, thinking of δ as a constant, the algorithm finds a polynomial-length list of unit vectors that have constant correlation with a constant fraction of the components of the orthogonal tensor $\sum_{i=1}^k a_i^{\otimes 3}$.

To the best of our knowledge, all previous algorithms for tensor decomposition (even in the orthogonal) require that the correlation between the input tensor and the orthogonal tensor is close to 1 [GVX14], [AGH+14], [BCMV14], [BKS15], [MSS16]

II. Techniques

To illustrate the idea of low-degree estimators for posterior moments, let's first consider the most basic stochastic block model $k = 2$ disjoint communities ($\alpha = 0$). (Our discussion will be similar to the analysis in [MNS15].) 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 $yy^\intercal$. 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

$$p_\alpha(x) = \prod_{ab \in \alpha} p_{ab}(x) \tag{II.1}$$

that is still an unbiased estimator $\mathbb{E}[p_\alpha(x) \mid y] = \prod_{ab\in\alpha} y_a y_b = y_i y_j$. 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 $yy^\intercal$. 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 later in this section (Section II-E).

A. 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. (II.2).

Lemma II.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 $\delta > 0$,*

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

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

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

In the full version of this work, we present the short combinatorial argument that shows that above the Kesten–Stigum bound the estimators for self-avoiding walks satisfy the conditions Eq. (II.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 $x - \frac{d}{n}\mathbf{1}\mathbf{1}^\intercal$. For $\ell \approx \log n$, the ℓ-th power of this matrix converges to $vv^\intercal$, 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 II.1 close to the Kesten–Stigum threshold.

B. Low-degree estimators for higher-order moments

Let's turn to the general mixed-membership stochastic block model $\mathrm{SBM}(n, d, \varepsilon, k, \alpha_0)$. Let (G, σ) be graph G and community structure $\sigma = (\sigma_1, \ldots, \sigma_n)$ drawn from this model. Recall that $\sigma_1, \ldots, \sigma_n$ are k-dimensional probability vectors, each roughly uniform over α_0+1 of the coordinates. Let $x \in \{0,1\}^{n\times n}$ be the adjacency matrix of G and let $y_1, \ldots, y_k \in \mathbb{R}^n$ 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 $\sum_{s=1}^k (y_s)(y_s)^\intercal$ (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 $\sum_{s=1}^k y_s y_s^\intercal$ 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 third 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[12] for $\ell = \log n$. Even for disjoint communities, this polynomial estimator would fail to achieve the Kesten–Stigum bound.

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 in the full version of this paper that the resulting estimator achieves constant correlation with the desired third moment tensor precisely up to the Kesten–Stigum bound.

[12] 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 $c \in [n]$

C. 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 II.2 (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*

$$\langle Q, Y\rangle \geqslant \delta/2 \cdot \|Q\| \cdot \|Y\| \,. \tag{II.3}$$

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 ob-

jects (for example, quantum analogues like density matrices or pseudo-distributions—the sum-of-squares analogue of distributions). Then, the norm minimization in Theorem II.2 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.

D. 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 II-B). 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 $\mathbb{E}_{i\in[k]} \max_{\hat{y}\in L} \langle \hat{y}, y\rangle \geqslant \delta^{O(1)}$. 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 II.2 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 $\sum_{i=1}^k y_i^{\otimes 3}$.

E. 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].[13]

For a coloring $c\colon [n] \to [\ell]$ 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.

Theorem II.3 (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*

$$p_c(x) := \frac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha\in \mathrm{SAW}_\ell(i,j)} p_\alpha(x) \cdot F_{c,a}\,. \tag{II.4}$$

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

$$\mathop{\mathbb{E}}_{x\sim \mathrm{SBM}(n,d,\varepsilon)} \mathop{\mathbb{E}}_c p_c(x)^2 \tag{II.5}$$

$$= \frac{1}{|\mathrm{SAW}_\ell(i,j)|^2} \cdot \tag{II.6}$$

$$\sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \mathop{\mathbb{E}}_c F_{c,\alpha} \cdot F_{c,\beta} \cdot \mathop{\mathbb{E}}_{x\sim\mathrm{SBM}} p_\alpha(x) p_\beta(x) \tag{II.7}$$

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

$$\leqslant e^{2\ell} \cdot \tfrac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha,\beta \in \mathrm{SAW}_\ell(i,j)} |\mathop{\mathbb{E}}_x p_\alpha(x) p_\beta(x)| \,. \quad \text{(II.8)}$$

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. (II.8) corresponds precisely to our notion of approximate pairwise independence in Lemma II.1. Therefore, if we are within the Kesten–Stigum bound, $\varepsilon^2 d \geqslant 1 + \delta$, the right-hand side of Eq. (II.8) 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 $e^{O(k)}$. It follows that by averaging over $e^{O(k)}$ random colorings we obtain a low-variance estimator for $p(x)$.

Acknowledgments

We are indebted to Avi Wigderson who suggested that color coding is a technique that could help with evaluating 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 the Rényi entropy. We thank the anonymous reviewers for many suggested improvements to this paper.

SBH was supported by an NSF graduate research fellowship, a Microsoft Research PhD fellowship, a Cornell University fellowship, and David Steurer's NSF Career award. DS was supported by a Microsoft Research Fellowship, a Alfred P. Sloan Fellowship, a NSF awards, and the Simons Collaboration for Algorithms and Geometry.

References

[Abbar] Emmanuel Abbe, *Community detection and stochastic block models: recent developments*, Journal of Machine Learning Research, Special Issue (To appear). 3

[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

[AFH+12] Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham Kakade, and Yi-Kai Liu, *A spectral algorithm for latent dirichlet allocation*, NIPS, 2012, pp. 926–934. 6

[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. 7

[AGHK13] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, and Sham Kakade, *A tensor spectral approach to learning mixed membership community models*, COLT, JMLR Workshop and Conference Proceedings, vol. 30, JMLR.org, 2013, pp. 867–881. 4, 6, 9

[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. 5, 7

[AS16] ———, *Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation*, NIPS, 2016, pp. 1334–1342. 4

[AYZ95] Noga Alon, Raphael Yuster, and Uri Zwick, *Color-coding*, J. ACM **42** (1995), no. 4, 844–856. 4, 5, 10

[BCMV14] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan, *Smoothed analysis of tensor decompositions*, STOC, ACM, 2014, pp. 594–603. 7

[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, 6

[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. 7, 10

[Gri01] Dima Grigoriev, *Linear lower bound on degrees of positivstellensatz calculus proofs for the parity*, Theor. Comput. Sci. **259** (2001), no. 1-2, 613–622. 6

[GVX14] Navin Goyal, Santosh Vempala, and Ying Xiao, *Fourier PCA and robust tensor decomposition*, STOC, ACM, 2014, pp. 584–593. 7

[Har70] Richard A Harshman, *Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis*. 9

[HK13] Daniel Hsu and Sham M. Kakade, *Learning mixtures of spherical Gaussians: moment methods and spectral decompositions*, ITCS'13—Proceedings of the 2013 ACM Conference on Innovations in Theoretical Computer Science, ACM, New York, 2013, pp. 11–19. MR 3385380 6

[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). 2

[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. 6

[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 9

[Mas14] Laurent Massoulié, *Community detection thresholds and the weak ramanujan property*, STOC, ACM, 2014, pp. 694–703. 4

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly, *Consistency thresholds for the planted bisection model*, STOC, ACM, 2015, pp. 69–75. 4, 7

[Moo17] Cristopher Moore, *The computer science and physics of community detection: Landscapes, phase transitions, and hardness*, CoRR **abs/1702.00467** (2017). 4

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer, *Polynomial-time tensor decompositions with sum-of-squares*, FOCS, IEEE Computer Society, 2016, pp. 438–446. 5, 7, 10

[MW15] Tengyu Ma and Avi Wigderson, *Sum-of-squares lower bounds for sparse PCA*, NIPS, 2015, pp. 1612–1620. 6

[Sch08] Grant Schoenebeck, *Linear level lasserre lower bounds for certain k-csps*, FOCS, IEEE Computer Society, 2008, pp. 593–602. 6

[SS17] Tselil Schramm and David Steurer, *Fast and robust tensor decomposition with applications to dictionary learning*, Conference on Learning Theory (COLT) (2017). 5

[VX15] Santosh Vempala and Ying Xiao, *Max vs min: Tensor decomposition and ICA with nearly linear sample complexity*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 1710–1723. 6

[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]. 3