A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem

Boaz Barak*, Samuel B. Hopkins†, Jonathan Kelner‡, Pravesh Kothari§, Ankur Moitra¶, and Aaron Potechin‖
* John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA, b@boazbarak.org
†Cornell University Department of Computer Science, Ithaca, NY, USA. samhop@cs.cornell.edu
‡MIT Department of Mathematics, Cambridge, MA, USA. kelner@mit.edu
§ UT Austin Department of Computer Science, Austin, TX, USA. kothari@cs.utexas.edu
¶ MIT Department of Mathematics, Cambridge, MA, USA. moitra@mit.edu
‖ Cornell University Department of Computer Science, Ithaca, NY, USA. aaronpotechin@gmail.com

***Abstract*—We prove that with high probability over the choice of a random graph G from the Erdős-Rényi distribution $G(n, 1/2)$, the $n^{O(d)}$-time degree d Sum-of-Squares semidefinite programming relaxation for the clique problem will give a value of at least $n^{1/2-c(d/\log n)^{1/2}}$ for some constant $c > 0$. This yields a nearly tight $n^{1/2-o(1)}$ bound on the value of this program for any degree $d = o(\log n)$. Moreover we introduce a new framework that we call *pseudo-calibration* to construct Sum-of-Squares lower bounds. This framework is inspired by taking a computational analogue of Bayesian probability theory. It yields a general recipe for constructing good pseudo-distributions (i.e., dual certificates for the Sum-of-Squares semidefinite program), and sheds further light on the ways in which this hierarchy differs from others.**

***Index Terms*—algorithm design and analysis; mathematical programming; computational complexity**

I. Introduction

The *planted clique* (also known as *hidden clique*) problem is a central question in average-case complexity. Arising from the 1976 work of Karp [1], the problem was formally defined by Jerrum [2] and Kucera [3] as follows: given a random Erdős-Rényi graph G from the distribution $G(n, 1/2)$ where every edge is chosen to be included with probability $1/2$ independently of all others in which we *plant* an additional clique (i.e., set of vertices that are all neighbors of one another) S of size ω, find S. It is not hard to see that the problem can be solved by brute force search which in this case takes quasipolynomial time whenever $\omega > c \log n$ for any constant $c > 2$. Despite considerable effort, the best polynomial-time algorithms only work when $\omega = \varepsilon\sqrt{n}$, for any constant $\varepsilon > 0$ [4].

Over the years planted clique and related problems have found applications to important questions in a variety of areas including community detection [5], finding signals in molecular biology [6], discovering motifs in biological networks [7], [8], computing Nash equilibrium [9], [10], property testing [11], sparse principal component analysis [12], compressed sensing [13], cryptography [14], [15] and even mathematical finance [16].

Thus, the question of whether the currently known algorithms can be improved is of great interest. Unfortunately, it is unlikely that lower bounds for planted clique can be derived from conjectured complexity class separations such as $\mathbf{P} \neq \mathbf{NP}$, precisely because it is an average-case problem [17], [18]. Our best evidence for its difficulty comes from showing limitations on powerful *classes* of algorithms. In particular, since many of the algorithmic approaches for this and related problems involve spectral techniques and convex programs, limitations for these types of algorithm are of significant interest. One such negative result was shown by Feige and Krauthgamer [19] who proved that the $n^{O(d)}$-time *degree d Lovász-Schrijver semidefinite programming hierarchy* (LS_+ for short) can only recover the clique if its size is at least $\sqrt{n/2^d}$.[1]

However, recently it was shown that in several cases, the *Sum-of-Squares (SOS) hierarchy* [20], [21], [22] — a stronger family of semidefinite programs which can be solved in time $n^{O(d)}$ for degree parameter d — can be significantly more powerful than other algorithms such as LS_+ [23], [24], [25]. In particular, it was conceivable that the SOS hierarchy might be able to find planted cliques that are much smaller than $\sqrt{n}$ in polynomial time, or at least be able to beat brute force search.

The first SOS lower bound for planted clique was shown by Meka, Potechin and Wigderson [26] who proved that the degree d SOS hierarchy cannot recover a clique of size $\tilde{O}(n^{1/d})$. This bound was later improved on by Deshpande and Montanari [27] and then Hopkins et al [28] to $\tilde{O}(n^{1/2})$ for degree $d = 4$ and $\tilde{O}(n^{1/(\lceil d/2 \rceil+1)})$ for general d. However,

Part of this work was done while BB, SBH, PK, and AP were at Microsoft Research New England. SBH was partially supported by an NSF GRFP under grant no. 1144153, and by NSF grants 1408673 and 1350196. JK was partially supported by NSF Award 1111109. PK was partially supported by grant 360724 from the Simons Foundation. AM was partially supported by NSF CAREER Award CCF-1453261, a grant from the MIT NEC Corporation and a Google Faculty Research Award. A full version of this paper is available at https://arxiv.org/abs/1604.03084

[1] As we discuss in Remark I.2 below, formally such results apply to the incomparable *refutation* problem, which is the task of certifying that there is no ω-sized clique in a random $G(n, 1/2)$ graph. However, our current knowledge is consistent with these variants having the same computational complexity.

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.53

this still left open the possibility that the constant degree (and hence polynomial time) SoS algorithm can significantly beat the $\sqrt{n}$ bound, perhaps even being able to find cliques of size n^{ε} for any fixed $\varepsilon > 0$. This paper answers this question negatively by proving the following theorem:

Theorem I.1 (Main Theorem)**.** *There is an absolute constant c so that for every $d = d(n)$ and large enough n, the SOS relaxation of the planted clique problem has integrality gap at least $n^{1/2 - c(d/\log n)^{1/2}}$.*

Beyond improving the previously known results, our proof is significantly more general and we believe provides a better intuition behind the limitations for SOS algorithms by viewing them from a "computational Bayesian probability" lens that is of its own interest. Moreover, there is some hope (as we elaborate below) that this view could be useful not just for more negative results but for SOS *upper bounds* as well. In particular our proof elucidates some aspects of the way in which the SOS algorithm is more powerful than the LS_+ algorithm.

Remark I.2 (*The different variants of the planted clique problem*)*.* Like other average-case problems in $\mathbf{NP}$, the planted clique problem with parameter ω has three variants of *search*, *refutation*, and *decision*. The *search* variant is the task of recovering the clique from a graph in which it was planted. The *refutation* variant is the task of *certifying* that a random graph in $G(n, 1/2)$ (where with high probability the largest clique has size $(2 + o(1)) \log n$) does not have a clique of size ω. The *decision* problem is to distinguish between a random graph from $G(n, 1/2)$ and a graph in which an ω-sized clique has been planted. The decision variant can be reduced to either the search or the refutation variant, but we know of no reduction between the latter two variants. Integrality gaps for mathematical relaxations such as the Sum-of-Squares hierarchy are most naturally stated as negative results for the *refutation* variant, as they show that such relaxations cannot certify that a random graph has no ω-sized clique by looking at the maximum value of the objective function. Our result can also be viewed as showing that the natural SoS-based algorithm for the *decision* problem (which attempts to distinguish on the objective value) also fails. Moreover, our result also rules out some types of SoS-based algorithms for the *search* problem as it shows that in a graph with a planted clique, there exists a solution with an objective value of ω based only on the random part, which means that it does not contain any information about which nodes participate in the clique and hence is not useful for rounding algorithms.

II. Planted Clique and Probabilistic Inference

We now discuss the ways in which planted clique differs from problems for which strong SOS lower bounds have been shown before, and how this relates to a "computational Bayesian" perspective. There have been several strong lower bounds for the SOS algorithm before, in particular for problems such as 3SAT, 3XOR and other constraint satisfaction problems as well as the knapsack problem [29], [30], [31]. However, obtaining strong lower bounds for the planted clique problem seems to have required different techniques. A high-level way to describe the difference is that lower bounds for planted clique require accounting for **weak global constraints** rather than **strong local** ones. In the random 3SAT/3XOR setting, the effect of one variable on another is either extremely strong (if they are "nearby" in the formula) or essentially zero. In contrast in planted clique each variable has a weak *global* effect on all of the other variables. We now explain this in more detail.

Consider a random graph G in which a clique S of size ω has been planted. If someone tells us some simple statistics of G and then tells us that vertex 17 is not in S, this new informatoin makes it slightly less likely that 17's neighbors are in S and slightly more likely that 17's non-neighbors are in S. So, this information has a *weak global* effect. In contrast, when we have a random sparse 3SAT formula φ in which an assignment x has been planted, if someone tells us that $x_{17} = 0$ then it gives us a lot of information about the local neighborhood of the 17^{th} variable (the variables that are involved in constraints with 17 or one that have a short path of constraints to it) but there is an exponential decay of these correlations and so this information tells us almost nothing about the distribution of most of the variables x_i (that are far away from 17 in the sparse graph induced by φ)[2]. Thus, in the random 3SAT setting information about the assignments of individual variables has a *strong local effect*. Indeed, previous Sum-of-Squares lower bounds for random 3SAT and 3XOR [29], [30] could be interpreted as producing a "distribution-like" object in which, conditioned on the value of a small set of variables S, some of the variables "close" to S in the formula were fixed, and the rest were completely independent.

This difference between the random SAT and the planted clique problems means that some subtleties that can be ignored in setting of random constraint satisfaction problems need to be tackled head-on when dealing with planted cliques. However to make this clearer, we need to take a detour and discuss Bayesian probabilities and their relation to Sum-of-Squares.

A. Computational Bayesian Probabilities and Pseudo-distributions

Strictly speaking, if a graph G contains a unique clique S of size ω, then for every vertex i the probability that i is in S is either zero or one. But, a computationally bounded observer may not know whether i is in the clique or not, and we could try to quantify this ignorance using probabilities. These

[2] This exponential decay can be shown formally for the case of satisfiable random 3SAT or 3XOR formulas whose clause density is sufficiently smaller than the threshold. In our regime of overconstrainted random 3SAT/3XOR formulas there will not exist any satisfying assignments, and so to talk about "correlations" in the distributions of assignments we need to talk about the "Bayesian estimates" that arise from algorithms such as Sum-of-Squares or belief propagation. Both these algorithms exhibit this sort of exponential decay we talk about; see also Remark II.1

can be thought of as a computational analogue of *Bayesian probabilities*, that, rather than aiming to measure the frequency at which an event occurs in some sample space, attempts to capture the subjective beliefs of some observer.

That is, the Bayesian probability that an observer B assigns to an event E can be thought of as corresponding to the odds at which B would make the bet that E holds. Note that this probability could be strictly between zero and one even if the event E is fully determined, depending on the evidence available to B. While typically Bayesian analysis does not take into account computational limitations, one could imagine that even if B has access to information that fully determines whether E happened or not, he could still rationally assign a subjective probability to E that is strictly between zero and one if making the inferences from this information is computationally infeasible. In particular, in the example above, even if a computationally bounded observer has access to the graph G, which information-theoretically fully determines the planted ω-sized clique, he could still assign a probability strictly between zero and one to the event that vertex 17 is in the planted ω-sized clique, based on some simple to compute statistics such as how many neighbors 17 has, etc.

The Sum-of-Squares algorithm can be thought of as giving rise to an internally consistent set of such "computational probabilities". These probabilities may not capture *all* possible inferences that a computationally bounded observer could make, but they do capture all inferences that can be made via a powerful proof system.

a) Bayesian estimates for planted clique.: To get a sense for our results and techniques, it is instructive to consider the following scenario. Let $G(n,1/2,\omega)$ be the distribution over pairs (G,x) of an n-vertex graphs G and a vector $x \in \mathbb{R}^n$ which is obtained by sampling a random graph in $G(n,1/2)$, planting an ω-sized clique in it, and letting G be the resulting graph and x the $0/1$ characteristic vector of the planted clique. Let $f : \{0,1\}^{\binom{n}{2}} \times \mathbb{R}^n \to \mathbb{R}$ be some function that maps a graph G and a vector x into some real number $f_G(x)$. Now imagine two parties, Alice and Bob (where Bob can also stand for "Bayesian") that play the following game: Alice samples (G,x) from the distribution $G(n,1/2,\omega)$ and sends G to Bob, who wants to output the expected value of $f_G(x)$. We denote this value by $\tilde{\mathbb{E}}_G f_G$.

If we have no computational constraints then it is clear that Bob can simply output $\tilde{\mathbb{E}}_G f_G$ be equal to $\mathbb{E}_{x|G} f_G(x)$, by which we mean the expected value of $f_G(x)$ where x is chosen according to the conditional distribution on x given the graph G.[3] In particular, the value $\tilde{\mathbb{E}}_G f_G$ will be *calibrated* in the sense that

$$\mathop{\mathbb{E}}_{G \in_R G(n,1/2,\omega)} \tilde{\mathbb{E}}_G f_G = \mathop{\mathbb{E}}_{(G,x) \in_R G(n,1/2,\omega)} f_G(x) \tag{II.1}$$

Now if Bob is computationally bounded, then he will not necessarily be able to compute the value of $E_{x|G} f_G(x)$ even for a simple function such as $f_G(x) = x_{17}$. Indeed, as we mentioned, since with high probability the clique x is uniquely determined by G, $\mathbb{E}_{x|G}\, x_{17}$ will simply equal 1 if vertex 17 is in the clique and equal 0 otherwise. However, note that we don't need to compute the true conditional expectation to obtain a calibrated estimate. In the above example, if Bob simply outputs $\tilde{\mathbb{E}} x_{17} = \omega/n$ then his estimate will satisfy (II.1).

Our Sum-of-Squares lower bound amounts to coming up with some reasonable "pseudo-expectation" that can be efficiently computed, where $\tilde{\mathbb{E}}_G$ is meant to capture a "best effort" of a computationally bounded party of approximating the Bayesian conditional expectation $\mathbb{E}_{x|G}$. Our pseudo-expectation will be far from the true conditional expectations, but will be internally consistent in the sense that for all "simple" functions f it will satisfy (II.1). The key property is that our pseudo-expectation will not distinguish between a graph G drawn from $G(n,1/2,\omega)$ and a random G from $G(n,1/2)$. In particular, it will also satisfy the following *pseudo-calibration* condition:

$$\mathop{\mathbb{E}}_{G \in_R G(n,1/2)} \tilde{\mathbb{E}}_G f_G = \mathop{\mathbb{E}}_{(G,x) \in_R G(n,1/2,\omega)} f_G(x) \tag{II.2}$$

for all "simple" functions $f = f(G,x)$. Note that (II.2) does not make sense for the estimates of a truly Bayesian (i.e., computationally unbounded) Bob, since almost all graphs G in $G(n,1/2)$ are not even in the support of $G(n,1/2,\omega)$. Nevertheless, our pseudo-distributions will be well defined even for a random graph and hence will yield estimates for the probabilities over this hypothetical object (i.e., the ω-sized clique) that does not exist. The "pseudo-calibration" condition (II.2) might seem innocent, but it turns out to imply many useful properties. In particular is not hard to see that (II.2) implies that for every *simple strong constraint* of the clique problem — a function f such that $f(G,x) = 0$ for every x that is a characteristic vector of an ω-clique in G — it must hold that $\tilde{\mathbb{E}}_G f_G = 0$. But even beyond these "strong constraints", (II.2) implies that the pseudo-expectation satisfies many *weak constraints* as well, such as the fact that a vertex of high degree is more likely to be in the clique and that if i is not in the clique then its neighbors are less likely and non-neighbors are more likely to be in it.

Indeed, the key conceptual insight of this paper is to phrase the pseudo-calibration property (II.2) as a desiderata for our pseudo-distributions. Namely, we say that a function $f = f(G,x)$ is "simple" if it is a low degree polynomial in both the entries of G's adjacency matrix and the variables x, and then require (II.2) to hold for all simple functions. It turns out that once you do so, the choice for the pseudo-distribution is essentially determined, and hence proving the main result amounts to showing that it satisfies the constraints of the SOS algorithm. In the next section we will outline the main ideas of our proof.

Remark II.1 (Planted Clique vs 3XOR). In the light of the discussion above, it is instructive to consider the case of

[3] The astute reader might note that this expectation is somewhat degenerate since with very high probability the graph G will uniquely determine the vector x, but please bear with us, as in the computational setting we will be able to treat x as "undetermined".

random 3XOR discussed before. Random 3XOR instances on n variables and $\Theta(n)$ constraints are easily seen to be maximally unsatisfiable (that is, at most $\approx 1/2$ the constraints can be satisfied by any assignment) with high probability. On the other hand, Grigorev [29] constructed a sum of squares pseudoexpectation that pretends that such instances instances are satisfiable with high probability, proving a sum of squares lower bound for refuting random 3XOR formulas.

Analogous to the planted distribution $G(n, 1/2, \omega)$, one can define a natural planted distribution over 3XOR instances - roughly speaking, this corresponds to first choosing a random Boolean assignment x^* to n variables and then sampling random 3XOR constraints conditioned on being consistent with x^*. It is not hard to show that pseudo-calibrating with respect to this planted distribution a la (II.2) produces precisely the pseudoexpectation that Grigoriev constructed. However, unlike in the planted clique case, in the case of 3XOR, the pseudo-calibration condition implies that for every low-degree monomial x_S, either the value of x_S is completely fixed (if it can be derived via low width resolution from the 3XOR equations of the instance) or it is completely unconstrained.

The pseudoexpectations considered in previous works [32], [26], [27]) are similar to Grigoriev's construction, in the sense that they essentially respect only strong constraints (e.g., that if A is not a clique in the graph, then the probability that it is contained in the planted clique is zero), but other than that assume that variables are independent. However, unlike the 3XOR case, in the planted clique problem respecting these strong constraints is not enough to achieve the pseudo-calibration condition (II.2) and the pseudoexpectation of [32], [26], [27] can be shown to violate weak probabilistic constraints imposed by (II.2) even at degree four. See Observation II.4 for an example.

B. From Calibrated Pseudo-distributions to Sum-of-Squares Lower Bounds

What does Bayesian inference and calibration have to do with Sum-of-Squares? In this section, we show how calibration is almost forced on any pseudodistribution feasible for the Sum-of-Squares algorithm. In order to show that the degree d SOS algorithm fails to certify that a random graph does not contain a clique of size ω, what we need is to show that for a random G, with high probability we can come up with an operator that maps a degree at most d, n-variate polynomial p to a real number $\tilde{\mathbb{E}}_G p$ satisfying the following constraints:

1) (Linearity) The map $p \mapsto \tilde{\mathbb{E}}_G p$ is linear.
2) (Normalization) $\tilde{\mathbb{E}}_G 1 = 1$.
3) (Booleanity constraint) $\tilde{\mathbb{E}}_G x_i^2 p = \tilde{\mathbb{E}} x_i p$ for every p of degree at most $d-2$ and $i \in [n]$.
4) (Clique constraint) $\tilde{\mathbb{E}}_G x_i x_j p = 0$ for every (i, j) that is not an edge and p of degree at most $d-2$.
5) (Size constraint) $\tilde{\mathbb{E}}_G \sum_{i=1}^{n} x_i = \omega$.
6) (Positivity) $\tilde{\mathbb{E}}_G p^2 \geqslant 0$ for every p of degree at most $d/2$.

Definition II.2. A map $p \mapsto \tilde{\mathbb{E}}_G p$ satisfying the above constraints 1–6 is called a *degree d pseudo-distribution* (w.r.t. the planted clique problem with parameter ω).

We can now restate our main result as follows:

Theorem II.3 (Theorem I.1, restated)**.** *There is some constant c such that if $\omega \leqslant n^{1/2-c(d/\log n)^{1/2}}$ then with high probability over G sampled from $G(n, 1/2)$, there is a degree d pseudodistribution $\tilde{\mathbb{E}}_G$ satisfying constraints 1–6 above.*

Note that all of these constraints would be satisfied if $\tilde{\mathbb{E}}_G p$ was obtained by taking the expectation of p over a distribution on ω-sized cliques in G. However, with high probability there is no $2.1 \log n$-sized clique in G (and let alone a roughly $\sqrt{n}$-sized one) so we will need a completely different mechanism to obtain such a pseudo-distribution.

Previously, the choice of the pseudo-distribution seemed to require a "creative guess" or an "ansatz". For problems such as random 3SAT this guess was fairly natural and almost "forced", while for planted clique as well as some related problems [33] the choice of the pseudo-distribution seemed to have more freedom, and more than one choice appeared in the literature.

For example, Feige and Krauthgamer [32] (henceforth FK) defined a very natural pseudo-distribution $\tilde{\mathbb{E}}^{FK}$ for a weaker hierarchy. For a graph G on n vertices, and subset $A \subseteq [n]$, $\tilde{\mathbb{E}}_G^{FK} x_A$ is equal to zero if A is not a clique in G and equal to $2^{\binom{|A|}{2}} \left(\frac{\omega}{n}\right)^{|A|}$ if A is a clique, and extended to degree d polynomials using linearity.[4] [32] showed that that for every d, and $\omega < O(\sqrt{n/2^d})$, this pseudo-distribution satisfies the constraints 1–5 as in Definition II.2 as well as a weaker version of positivity (this amounts to the so called "LovÃąsz-Schrijver+" SDP). Meka, Potechin and Wigderson [26] proved that the same pseudo-distribution satisfies all the constraints 1–6 (and hence is a valid degree d pseudo-distribution) as long as $\omega < \tilde{O}(n^{1/d})$. This bound on ω was later improved to $\tilde{O}(n^{1/3})$ for $d = 4$ by [27] and to $\tilde{O}(n^{(\lfloor d/2 \rfloor + 1)^{-1}})$ for a general d by [34].

Interestingly, the FK pseudo-distribution does *not* satisfy the full positivity constraint for larger values of ω. The issue is that while the FK pseudo-distribution satisfies the "strong" constraints that $\tilde{\mathbb{E}}_G^{FK} x_A = 0$ if A is not a clique, it does not satisfy weaker constraints that are implied by (II.2). For example, for every constant ℓ, if vertex i participates in $\sqrt{n}$ more ℓ-cliques than the expected number then one can compute that the conditional probability of i belonging in the clique should be a factor $1 + c\omega/\sqrt{n}$ larger for some constant $c > 0$. However, the FK pseudo-distribution does not make this correction. In particular, for every ℓ, there is a simple polynomial that shows that the FK pseudoexpectation is not calibrated.

Observation II.4. Fix $i \in [n]$ and let ℓ be some constant. If $p_G = (\sum_j G_{i,j} x_j)^\ell$ then **(i)** $\mathbb{E}_{G \sim G(n,1/2)} \tilde{\mathbb{E}}_G^{FK}[p_G^2] \leqslant \omega^\ell$ and **(ii)** $\mathbb{E}_{(G,x) \sim G(n,1/2,\omega)}[p_G(x)^2] \geqslant \frac{\omega^{2\ell+1}}{n}$. In par-

[4] The actual pseudo-distribution used by [32] (and the followup works [26], [27]) was slightly different so as to satisfy $\tilde{\mathbb{E}}_G(\sum_{i=1}^{m} x_i)^\ell = \omega^\ell$ for every $\ell \in \{1, \ldots, d\}$. This property is sometimes described as satisfying the constraint $\{\sum_i x_i = \omega\}$.

ticular, when $\omega \gg n^{\frac{1}{\ell+1}}$, $\mathbb{E}_{G\sim G(n,1/2)}\tilde{\mathbb{E}}_G^{FK}[p_G^2] \ll \mathbb{E}_{(G,x)\sim G(n,1/2,\omega)}\, p_G(x)$.

Proof sketch. For **(ii)** note that with probability (ω/n) vertex i is in the clique, in which case $\sum_j G_{i,j}x_j = \omega$, and hence the expectation of p_G^2 is at least $(\omega/n)\omega^{2\ell}$. For **(i)**, we open up the expectation and the definition to get (up to a constant depending on ℓ)

$$\sum_{j_1,\ldots,j_{2\ell}} G_{i,j_1}\ldots G_{i,j_{2\ell}}(\omega/n)^{2\ell} \mathop{\mathbb{E}}_{G\sim G(n,1/2)} 1_{\{i_1,\ldots,i_{2\ell}\}\text{ is clique}}$$

Since this expectation is zero unless every variable $G_{i,j}$ is squared, in which case the number of distinct j's is at most ℓ, we can bound the sum by $n^{\ell}(\omega/n)^{\ell} = \omega^{\ell}$. This completes the proof sketch. □

Observation II.4 captures the failure of calibration for a specific polynomial $p_G(x)$ where the coefficients are low-degree functions of the graph G. The polynomial p_G above can be used to show that degree d $\tilde{\mathbb{E}}^{FK}$ does not satisfy the positivity constraint for $\omega \gg n^{1/(\frac{d}{2}+1)}$. This observation is originally due to Kelner, see [34]

Fact II.5. *Let p_G be as in the Observation II.4. Then, there exists a C such that for $q = q_G = (C\omega^{\ell}x_S - p_G)$ with high probability over the graph $G \sim G(n, 1/2)$, $\tilde{\mathbb{E}}^{FK}[q_G^2] < 0$ for $\omega \gg n^{\frac{1}{\ell+1}}$.*

For the case $d = 4$, Hopkins et al [28] proposed an "ad hoc" fix for the FK pseudo-distribution that satisfies positivity up to $\omega = \tilde{O}(\sqrt{n})$, by explicitly adding a correction term to essentially calibrate for the low-degree polynomials q_G from Fact II.5. However, their method did not extend even for $d = 6$, because of the sheer number of corrections that would need to be added and analyzed. Specifically, there are multiple families of polynomials such that their $\tilde{\mathbb{E}}^{FK}$ value departs significantly from their calibrated value in expectation and gives multiple points of failure of positivity in a manner similar to Observation II.4 and Fact II.5. Moreover, "fixing" these families by the correction as in case of degree four leads to new families of polynomials that fail to achieve their calibrated value and exhibit negative pseudoexpectation for their squares and so on.

The *coefficients* of the polynomial p_G of Observation II.4 are themselves low degree polynomials in the adjacency matrix of G. This turns out to be a common feature in all the families of polynomials one encounters in the above works. Thus our approach is to fix all these polynomials *by fiat*, by placing the constraint that the pseudo-distribution must satisfy (II.2) for every such polynomial, and using that as our implicit definition of the pseudo-distribution. Indeed it turns our that once we do so, the pseudo-distribution is essentially determined. Moreover, (II.2) guarantees that it satisfies many of the "weak global constraints" that can be shown using Bayesian calculations.

Ultimately we will construct the map $G \mapsto \tilde{\mathbb{E}}_G$ as a low degree polynomial in G. Why is it OK to make such a restriction? One justification is the heuristic that the pseudo-distribution itself must be simple since we know that it is efficiently computable (via the SOS algorithm) from the graph G. Another justification is that by forcing the pseudo-distribution to be low-degree we are essentially making it *smooth* or "high entropy", which is consistent with the Jaynes *maximum entropy principle* [35], [36]. Most importantly — and this is the bulk of the technical work of this paper and the subject of the next subsection — this pseudo-distribution can be shown to satisfy *all* the constraints 1–6 of Definition II.2 including the positivity constraint.

We believe that this principled approach to designing pseudo-distributions elucidates the power and limitations of the SOS algorithm in cases such as planted clique, where accounting for weak global correlations is a crucial aspect of the problem.

Remark II.6 (*Where does the planted distribution arise from?*). Theorem II.3 (as well as Theorem I.1) makes no mention of the planted distribution $G(n, 1/2, \omega)$ and only refers to an actual random graph. Thus it might seem strange that we base our pseudo-distribution on the planted distribution via (II.2). One way to think about the planted distribution is that it corresponds to a *Bayesian prior* distribution on the clique. Note that this is the *maximum entropy* distribution on cliques of size ω, and so it is a natural choice for a prior per Jaynes's principle of maximum entropy. Our actual pseudo-distribution can be viewed as correcting this planted distribution to a posterior that respects simple inferences from the observed graph G.

C. Towards Proving Positivity: Structure vs. Randomness

We have seen that pseudo-calibration is desirable both *a priori* and in light of the failure of previous lower-bound attempts. Now we turn to the question: How do we formally define a pseudo-calibrated linear map $\tilde{\mathbb{E}}_G$, and show that it satisfies constraints 1–6 with high probability, to yield Theorem II.3?

We will require (II.2) to hold with respect to every function $f = f(G, x)$ that has degree at most τ in the entries of the adjacency matrix G and degree at most d in the variables x, and in addition we require that the map $G \mapsto \tilde{\mathbb{E}}_G$ is itself of degree at most τ in G, then this completely determines $\tilde{\mathbb{E}}_G$. For any $S \subseteq [n]$, $|S| \leqslant d$, using the Fourier transform we can write $\tilde{\mathbb{E}}_G[x_S]$ as an explicit low degree polynomial in G_e:

$$\tilde{\mathbb{E}}_G[x_S] = \sum_{\substack{T\subseteq\binom{[n]}{2}\\ |\mathcal{V}(T)\cup S|\leqslant\tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(T)\cup S|} \chi_T(G), \tag{II.3}$$

where $\mathcal{V}(T)$ is the set of nodes incident to the subset of edges (i.e., graph) T and $\chi_T(G) = \prod_{e\in T} G_e$. (This calculation is carried out in the full version.) For $\omega \approx n^{0.5-\varepsilon}$, we will need to choose the truncation threshold $\tau \gtrapprox d/\varepsilon$. It turns out that constraints 1–5 are easy to verify and thus we are left with proving the *positivity constraint*. Indeed this is not surprising as verifying this constraint is always the hardest part of a Sum-of-Squares lower bound.

As is standard, to analyze this positivity requirement we work with the *moment matrix* of $\tilde{\mathbb{E}}_G$. Namely, let $\mathcal{M}$ be the $\binom{n}{\leqslant d/2} \times \binom{n}{\leqslant d/2}$ matrix where $\mathcal{M}(I,J) = \tilde{\mathbb{E}}_G \prod_{i \in I} x_i \prod_{j \in J} x_j$ for every pair of subsets $I, J \subseteq [n]$ of size at most $d/2$. Our goal can be rephrased as showing that $\mathcal{M} \succeq 0$ (i.e., $\mathcal{M}$ is positive semidefinite).

Given a (symmetric) matrix N, to show that $N \succeq 0$ our first hope might be to diagonalize N. That is, we would hope to find a matrix V and a diagonal matrix D so that $N = VDV^\dagger$. Then as long as every entry of D is nonnegative, we would obtain $N \succeq 0$. Unfortunately, carrying this out directly can be far too complicated. Even the eigenvectors of simple random matrices are not completely understood, let alone matrices like ours with intricate dependencies among the entries. However, as the next example demonstrates, it is sometimes possible to prove positivity for a random matrix using what we call *approximate* diagonalization.

a) Example: Planted Clique Lower Bound for $d = 2$ (a.k.a. Basic SDP).: Consider the problem of producing a pseudo-distribution $\tilde{\mathbb{E}}$ satisfying constraints 1–6 of Definition II.2, with $d = 2$. In this simple case, many subtleties can be safely be ignored, but can still provide some intuition. For $d = 2$, it is enough to define $\tilde{\mathbb{E}}x_i$ and $\tilde{\mathbb{E}}x_i x_j$ for every $i \in [n]$ and $\{i, j\} \subseteq [n]$. Let $\tilde{\mathbb{E}}x_i = (\omega/n)$ for every i, and set $\tilde{\mathbb{E}}x_i x_j$ to be $\left(\frac{\omega}{n}\right)^2$ if (i, j) is an edge in G and zero otherwise. It is not hard to show that positivity reduces to showing that $\mathcal{N} \succeq 0$ where $\mathcal{N}$ is the $n \times n$ matrix with $\mathcal{N}_{i,j} = \tilde{\mathbb{E}}x_i x_j$. Using standard results on random matrices, $\mathcal{N}$ has one eigenvalue (whose corresponding eigenvector is close to the vector $u = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$) of value ω^2/n, while all others are distributed in the interval $\frac{\omega}{n} \pm O\left(\frac{\omega^2}{n^2}\sqrt{n}\right)$ which is strictly positive as long as $\omega \ll \sqrt{n}$. Thus, while we cannot explicitly diagonalize $\mathcal{N}$, we have enough information to conclude that it is positive semidefinite. In other words, it was enough for us to get an *approximate diagonalization* for $\mathcal{N}$ of the form $\mathcal{N} \approx \frac{\omega^2}{n} uu^\dagger + \frac{\omega}{n} Id + E$ for some sufficiently small (in spectral norm) "error matrix" E. Ultimately we will need to do something similar, but with many eigenvalues and many error matrices that are inter-dependent.

b) Approximate Factorization for $\mathcal{M}$.: We return now to the moment matrix $\mathcal{M}$ for our (pseudo)calibrated pseudodistribution. Our goal is to give an approximate diagonalization of $\mathcal{M}$. There are several obstacles to doing so:

1) In the case $d = 2$ there was just one rank-1 approximate eigenspace to be handled. The number of these approximate eigenspaces will grow with d, so we will need a more generic way to handle them.
2) Each approximate eigenspace corresponds to a family of polynomials $\{p\}$ whose calibrated pseudoexpectations are all roughly equal. (In the case $d = 2$, the only interesting polynomial was the polynomial $\sum_j x_j$ whose coefficients are proportional to the vector $u = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$.) As we saw in Observation II.4, if p_G is a polynomial whose coefficients depend on the graph G, even in simple ways, the calibrated value $\tilde{\mathbb{E}}_G p_G$ may also depend substantially on the graph. Thus, when we write $\mathcal{M} \approx \mathcal{L}\,\mathcal{Q}\,\mathcal{L}^\dagger$ for some approximately-diagonal matrix $\mathcal{Q}$, we will need the structured part $\mathcal{L} = \mathcal{L}(G)$ to itself be graph-dependent.
3) The errors in our diagonalization of $\mathcal{M}$ — corresponding in our $d = 2$ example to the matrix E — will not be small enough to ignore as we did above. Instead, each error matrix will itself have to be approximately diagonalized, recursively until these errors are driven down sufficiently far in magnitude.

We now discuss at a high level our strategy to address items (1) and (2). The resolution to item (3) is the most technical element of our proof, and we leave it for later. Consider the vector space of all polynomials $f : \{0,1\}^{\binom{n}{2}} \times \mathbb{R}^n \to \mathbb{R}$ which take a graph and an n-dimensional real vector and yield a real number. (We write $f_G(x)$, where G is the graph and $x \in \mathbb{R}^n$.) If we restrict attention to the subspace of those of degree at most d in x, we obtain the polynomials in the domain of our operator $\tilde{\mathbb{E}}_G$. If we additionally restrict to the subspace of polynomials which are low degree in G, we obtain the family of polynomials so that $\mathbb{E}_G \tilde{\mathbb{E}}_G f_G(x)$ is calibrated. Call this subspace $\mathcal{V}$.

Our goal would to be find an approximate diagonalization for all the non-trivial eigenvalues of $\mathcal{M}$ using only elements from $\mathcal{V}$. The advantage of doing so is that for every $f \in \mathcal{V}$, we can calculate $\mathbb{E}_G \tilde{\mathbb{E}}_G f_G^2$ using the pseudo-calibration condition (II.2). In particular it means that if we find a function f such that f_G is with high probability an approximate eigenvector of G, then we can compute the corresponding expected eigenvalue $\lambda(f)$.

A crucial tool in finding such an approximate eigenbasis is the notion of *symmetry*. For every f, if f' is obtained from f via a permutation of the variables $x_1, \ldots, x_n$, then $\mathbb{E}_G \tilde{\mathbb{E}}_G f_G^2 = \mathbb{E}_G \tilde{\mathbb{E}}_G f_G'^2$. The result of this symmetry, for us, is that our approximate diagonalization requires only a constant (depending on d) number of eigenspaces. This argument allows us to restrict our attention to a constant number of classes of polynomials, where each class is determined by some finite graph U that we call its *shape*. For every polynomial f with shape U, we compute (approximately) the value of $\mathbb{E}_G \tilde{\mathbb{E}}_G f_G^2$ as a function of a simple combinatorial property of U, and our approximate eigenspaces correspond to polynomials with different shapes.

We can show that that in expectation our approximate eigenspaces will have non-negative eigenvalues since the pseudo-calibration condition (II.2) in particular implies that for every f that is low degree in both G and x, $\mathbb{E}_G \tilde{\mathbb{E}}_G f_G^2 \geqslant 0$. However, the key issue is to deal with the error terms that arise from the fact that these are only approximate eigenspaces. One could hope that, like in other "structure vs. randomness" partitions, this error term is small enough to ignore. Alas, this is not the case, and we need to handle it recursively, which is the crux of item (3) and the cause of much of the technical complications of our paper.

Remark II.7 (*Structure vs. randomness*). At a high level our approach can be viewed as falling into the general paradigm of "structure vs. randomness" as discussed by Tao [37]. The general idea of this paradigm is to separate an object O into a "structured" part that is simple and predictable, and a "random" part that is unpredictable but has small magnitude or has some global statistical properties.

One example of this is the Szemerédi regularity lemma [38] as well variants such as [39] that partition a matrix into a sum of a low rank and pseudorandom components. Another example arises from the random models for the *primes* (e.g., see [40], [41]). These can be thought of positing that, as far as certain simple statistics are concerned, (large enough) primes can be thought of as being selected randomly conditioned on not being divisible by $2, 3, 5$ etc.. up to some bound w.

All these examples can be viewed from a computationally bounded Bayesian perspective. For every object O we can consider the part of O that can be inferred by a computationally bounded observer to be O's *structured* component, while the remaining uncertainty can be treated as if it is *random*, even if in actuality it is fully determined. Thus in our case, even though for almost every particular graph G from $G(n, 1/2, \omega)$, the clique x is fully determined by G, we still think of x as having a "structured" part which consists of all the inferences a "simple" observer can make from G (e.g., that if i and j are non-neighbors then $x_i x_j = 0$), and a "random" part that consists of the remaining uncertainty. As in other cases of applying this paradigm, part of the technical work is bounding the magnitude (in our case in spectral norm) that arises from the "random" part, though as mentioned above in our case we need a particularly delicate control of the error terms which ends up causing much of the technical difficulty.

III. Proving Positivity: A Technical Overview

We now discuss in more detail how we prove that the *moment matrix* $\mathcal{M}$ corresponding to our pseudo-distribution is positive semidefinite. (For full details, see the full version at https://arxiv.org/abs/1604.03084) Recall that this is the $\binom{n}{\leqslant d/2} \times \binom{n}{\leqslant d/2}$ matrix $\mathcal{M}$ such that $\mathcal{M}(I, J) = \tilde{\mathbb{E}}_G \prod_{i\in I} x_i \prod_{j\in J} x_j$ for every pair of subsets $I, J \subseteq [n]$ of size at most $d/2$, and that it is defined via (II.3) as

$$\mathcal{M}(I,J) = \sum_{\substack{T\subseteq \binom{[n]}{2}\\ |\mathcal{V}(T)\cup I\cup J|\leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(T)\cup I\cup J|} \chi_T(G) . \qquad \text{(III.1)}$$

The matrix $\mathcal{M}$ is generated from the random graph G, but its entries are *not* independent. Rather, each entry is a polynomial in G_e, and there are some fairly complex dependencies between different them. Indeed, these dependencies will create a spectral structure for $\mathcal{M}$ that is very different from the spectrum of standard random matrices with independent entries and makes proving that $\mathcal{M}$ is positive semidefinite challenging. Our approach to showing that $\mathcal{M}$ is positive semidefinite is through a type of "symbolic factorization" or "approximate diagonalization," which we explain next.

A. Warm Up

It is instructive to begin with the tight analysis presented in [34] of the moments constructed in [26][5]. These moments can in fact obtained by using truncation threshold $\tau = |S|$ in (II.3). This choice of τ is the smallest possible for which the resulting construction satisfies the hard clique constraints. [34] show that this construction satisfies positivity for $\omega \lessapprox n^{1/(\frac{d}{2}+1)}$.

For the purpose of this overview, let us work with the principal submatrix F indexed by subsets I and J of size exactly d. The analysis in [34] proceeds by first splitting F into $d+1$ components $F = F_0 + F_1 + \cdots + F_d$ where $F_i(I, J) = F(I, J)$ if $|I \cap J| = i$ and 0 otherwise. Below, we discuss two of the key ideas involved that will serve as an inspiration for us.

As discussed before, we must approximately diagonalize the matrix F in the sense that the off diagonals blocks must be "small enough" to be charged to the on diagonal block. Thus the main question before us is obtain an (approximate) understanding of the spectrum of F that allows us to come up with a "change of basis" in which the off diagonal blocks are small enough to be charged to the positive eigenmass in the on-diagonal blocks.

Let us consider the piece F_0 for our discussion here. As alluded to in Section III, we want to break F into minimal pieces so that each piece is symmetric under the permutation of vertices. We can hope that each piece will then essentially have a single dominating eigenvalue that can be determined relatively easily. Below, we will essentially implement this plan.

First, we need to decide what kind of "pieces" we will need. These are the *graphical matrices* that we define next.

Definition III.1 (Graphical Matrices (see full version for formal definition))**.** Let U be a graph on $[2d]$ with specially identified subsets left and right subsets $[d]$ and $[2d] \setminus [d]$. For any $I, J \in \binom{[n]}{d}$, $I \cap J = \varnothing$, let $\pi_{I,J}$ be an injective map that takes $[d]$ into I and $[2d] \setminus [d]$ into J using a fixed convention. The graphical matrix M_U with graph U is then defined by $M_U(I, J) = \chi_{\pi_{I,J}(U)}(G)$.

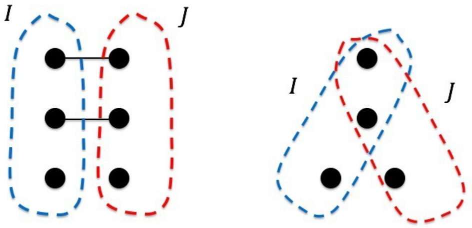
The starting point of the analysis is to decompose $F_0 = \sum_U \left(\frac{\omega}{n}\right)^{2d} M_U$, where M_U is the graphical matrix with shape U. Graphical matrices as above turn out to be the right building blocks for spectral analysis of our moment matrix. This is because a key observation in [34] shows that a simple combinatorial parameter, the size of the maximum bipartite matching between the left and right index in U (i.e. between $[d]$ and $[2d] \setminus [d]$), determines the spectral norm of M_U. Specifically, when U has a maximum matching of size $t < d$, the spectral norm of M_U is $\tilde{O}(n^{d-\frac{t}{2}})$, with high probability. Observe that when $d = 2$ and U is a single edge connecting the left vertex with the right, M_U is just the $\{-1, +1\}$-adjacency matrix of the underlying random graph and it is well known

[5] The construction in [26] actually also satisfies $\sum x_i = \omega$ as a constraint which causes the precise form to differ. We ignore this distinction here.

that the spectral norm in this case is $\Theta(\sqrt{n})$ matching the more general claim above.

In particular, this implies that when U has a perfect matching, M_U is pseudorandom in the sense that F_U essentially has the spectral norm $\approx n^{d/2}$, the same as that of an independent $\{-1,+1\}$ random matrix of the same dimensions. This allows M_U to be bounded against the positive eigenvalue $\left(\frac{\omega}{n}\right)^d$ of the diagonal matrix F_d as $\left(\frac{\omega}{n}\right)^d \gg \left(\frac{\omega}{n}\right)^{2d} n^{d/2}$ (even for ω approaching $\sqrt{n}$!). However for M_U when U has a maximum matching of size $t < d$, one can't bound against the diagonal matrix F_d anymore.

The next main idea is to note that for every M_U there's an appropriate "diagonal" against which we must charge the negative eigenvalues of M_U. When U has a perfect matching, this is literally the diagonal matrix F_d as done above. However, when, say, U is a (bipartite) matching of size $t < d$, we should instead charge against the "diagonal" matrix that can thought of as obtained by "collapsing" each matching edge into a vertex in U. In particular, this collapsing produces a matrix that lies in the decomposition of F_t.



the graph U

corresponding diagonal

There are a two main takeaways from this analysis that would serve as inspiration in the analysis of our actual construction. First is the decomposition into graphical matrices in order to have a coarse handle on the spectrum of the moment matrix. Second, the "charging" of negative eigenvalues against appropriate "diagonals" is essentially governed by the combinatorics of matchings in U.

B. *The Main Analysis*

We can now try to use the lessons from the warm up analysis to inspire our actual analysis. To begin with, we recall that each graphical matrix was obtained by choosing an appropriate (set of) Fourier monomials for any entry indexed by I, J. However, since for our actual construction we have monomials of much higher degree, we need to extend the notion of graphical matrices with *shapes* corresponding to larger graphs U. See full version for a formal definition.

It turns out that the right combinatorial idea to generalize the size of the maximum matching and control the spectral norm of the graphical matrices $\mathcal{M}_U$ is the maximum number of *vertex disjoint paths* between specially designated left and right endpoints of U (themselves the generalization of the bipartition we had in the warmup). Using Menger's theorem, this is equal to the size of a minimal collection of vertices that separates the left and right sets in the graph U, which we call the *separator size* of U.

Finally, we need a "charging" argument to work with the approximate diagonalization we end up with. Generalizing the idea in the warm up here is the hardest part of our proof, but relates again to the notion of vertex separators defined above. In the warm up, we used a naive charging scheme, breaking the moment matrix into simpler (graphical) matrices, each of which was either a “positive diagonal” mass or a “negative off-diagonal mass”, and pairing up the terms. Such a crude association doesn't work out immediately in the general setting. Instead, large groups of graphical matrices must be treated all at once. In each subspace of our approximate diagonalization of the moment matrix $\mathcal{M}$, we collect the "positive diagonal mass" and the "negative off digonal mass" that needs to be charged to it together and build an approximately PSD matrix out of it. As alluded to before, the error in this approximation is not negligible and thus we must further recurse on the error terms. In what follows, we discuss the factorization process that accomplishes the charging scheme implicitly and the recursive factorization for the error terms in some more detail. Consider some graph $T \subseteq \binom{[n]}{2}$, that corresponds to one term in the sum in (III.1) above, and let q be the minimum size of a set that separates I from J in T. Such a set is not necessarily unique but we can define the *leftmost* separator $\mathsf{left}-\mathsf{sep}(T) = S_\ell$ to be the q-sized separator that is closest to I and the *rightmost* separator $\mathsf{right}-\mathsf{sep}(T) = S_r$ to be the q-sized separator that is closest to J.

We can rewrite the (I, J) entry moment matrix $\mathcal{M}$ (III.1) by collecting monomials T with a fixed choice of the leftmost and rightmost separators S_ℓ and S_r. This step corresponds to collecting terms with similar spectral norms together accomplishing the goal of collecting together into a term, the "positive diagonal mass" and the "negative off diagonal mass" that are implicitly charged to each other in the intended approximate diagonalization.

$$\mathcal{M}(I,J) = \sum_{1 \leqslant q \leqslant |I|,|J|} \sum_{S_\ell, S_R : |S_\ell| = |S_r| = q} \tag{III.2}$$

$$\sum_{\substack{T \subseteq \binom{[n]}{2} \\ |\mathcal{V}(T) \cup I \cup J| \leqslant \tau \\ \mathsf{left}-\mathsf{sep}(T) = S_\ell, \mathsf{right}-\mathsf{sep}(T) = S_r}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(T) \cup I \cup J|} \chi_T(G) \tag{III.3}$$

We can then partition T into three subsets $\mathcal{R}_\ell$, $\mathcal{R}_m$ and $\mathcal{R}_r$ that represent the part of the graph T between I and S_ℓ, the part between S_ℓ and S_r and the part between S_r and J respectively (where edges within S_ℓ and edges within S_r are all placed in $\mathcal{R}_m$, see full version for details). We thus immediately obtain that

$$\chi_T(G) = \chi_{\mathcal{R}_\ell}(G) \chi_{\mathcal{R}_m}(G) \chi_{\mathcal{R}_r}(G) .$$

Thus:

$$\mathcal{M}(I,J) = \sum_{1\leqslant q\leqslant |I|,|J|} \sum_{S_\ell, S_R : |S_\ell|=|S_r|=q} \sum_{\substack{T\subseteq \binom{[n]}{2} \\ |\mathcal{V}(T)\cup I\cup J|\leqslant \tau \\ \mathsf{left-sep}(T)=S_\ell \\ \mathsf{right-sep}(T)=S_r}} \tag{III.4}$$

$$\left(\left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|} \chi_{\mathcal{R}_\ell}(G)\right) \cdot \left(\left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_m)|-2q} \chi_{\mathcal{R}_m}(G)\right) \tag{III.5}$$

$$\cdot \left(\left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_r)|} \chi_{\mathcal{R}_r}(G)\right) \tag{III.6}$$

One could hope that we could replace the RHS of (III.4) by

$$\sum_{\substack{1\leqslant q\leqslant |I|,|J| \\ \tau_1+\tau_2+\tau_3\leqslant\tau}} \sum_{\substack{S_\ell\subseteq\binom{[n]}{q} \\ S_r\subseteq\binom{[n]}{q}}} \tag{III.7}$$

$$\left(\sum_{\substack{\mathcal{R}_\ell \\ \mathcal{V}(\mathcal{R}_\ell)\supseteq I\cup S_\ell \\ |\mathcal{V}(\mathcal{R}_\ell)|=\tau_1}} \left(\tfrac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|} \chi_{\mathcal{R}_\ell}(G) \right) \tag{III.8}$$

$$\cdot \left(\sum_{\substack{\mathcal{R}_m \\ \mathcal{V}(\mathcal{R}_m)\supseteq S_\ell\cup S_r \\ |\mathcal{V}(\mathcal{R}_m)|=\tau_2}} \left(\tfrac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_m)|-2q} \chi_{\mathcal{R}_m}(G) \right) \tag{III.9}$$

$$\cdot \left(\sum_{\substack{\mathcal{R}_r \\ \mathcal{V}(\mathcal{R}_r)\supseteq S_r\cup J \\ |\mathcal{V}(\mathcal{R}_r)|=\tau_3}} \left(\tfrac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_r)|} \chi_{\mathcal{R}_r}(G) \right) \tag{III.10}$$

In fact, it turns out we can focus attention (up to sufficiently small error in the spectral norm) to the case $\tau_1 \leqslant \tau/3$, $\tau_2 \leqslant \tau/3$, $\tau_3 \leqslant \tau/3$ in which case if $M(I,J)$ was equal to (III.10) we could simply write

$$\mathcal{M} = \sum_q \mathcal{L}_q \, \mathcal{Q}_q \, \mathcal{L}_q^\dagger$$

where for $I, S \subseteq [n]$ with $|I| \leqslant d$ and $|S| = q$, we let $\mathcal{L}_q(I,S)$ be the sum of $(\omega/n)^{|\mathcal{V}(\mathcal{R}_\ell)|}\chi_{\mathcal{R}_\ell}(G)$ over all graphs $\mathcal{R}_\ell$ of at most $\tau/3$ vertices connecting I to S, and for S, S' of size q, we let $\mathcal{Q}_q(S,S')$ be the sum of $(\omega/n)^{|\mathcal{R}_m|-2q}\chi_{\mathcal{R}_m}(G)$ over all graphs $\mathcal{R}_m$ of at most $\tau/3$ vertices connecting S to S'.

Thus, in this case, this reduces our task of showing that $\mathcal{M}$ is positive semidefinite to showing that for every q, the matrix $\mathcal{Q} = \mathcal{Q}_q$ is positive semidefinite. However the main complication is that there are cross terms in the product $\mathcal{L}_q \, \mathcal{Q}_q \, \mathcal{L}_q^\dagger$ that correspond to repeating the same vertex (not in S_ℓ and S_r) in more than one of $\mathcal{R}_\ell$, $\mathcal{R}_m$ and $\mathcal{R}_r$. There is no matching term in the Fourier decomposition of $\mathcal{M}(I,J)$. So at best, for every fixed q, we can write the part of $\mathcal{M}$ corresponding to indices I, J with minimal vertex separator equal to q as

$$\mathcal{L} \, \mathcal{Q}_0 \, \mathcal{L}^\dagger - \mathcal{E}_1$$

for some error matrix $\mathcal{E}_1$ that exactly cancels out the extra terms contributed by cross terms with repeated vertices. Unfortunately, the spectral norm of this error matrix $\mathcal{E}_1$ is not small enough that we could simply ignore it. Luckily however, we can recurse and factorize $\mathcal{E}_1$ approximately as well. We can form a new graph T' by taking the parity of the edge sets in $\mathcal{R}_\ell$, $\mathcal{R}_m$ and $\mathcal{R}_r$. Now we find the leftmost and rightmost separators that separate I and J from each other, and from all repeated vertices. This gives us another decomposition of a graph into three pieces, from which we can write

$$\mathcal{E}_1 = \mathcal{L} \, \mathcal{Q}_1 \, \mathcal{L}^\dagger - \mathcal{E}_2$$

for some other matrix $\mathcal{Q}_1$. Continuing this argument gives us for every q a factorization of $\mathcal{M}_q$ as

$$\mathcal{L}(\mathcal{Q}_0 - \mathcal{Q}_1 + \mathcal{Q}_2 - \ldots - \mathcal{Q}_{2d-1} + \mathcal{Q}_{2d}) \, \mathcal{L}^\dagger \\ - (\xi_0 - \xi_1 + \xi_2 - \ldots - \xi_{2d-1} + \xi_{2d})$$

The error matrices $\xi_0, \xi_1, \ldots, \xi_{2d}$ arise from truncation issues, which we have ignored in the argument above and turn out to be negligible.

It is not hard to show that $\mathcal{Q}_0 \succeq D$ for some positive semidefinite matrix D that we define later. What remains is to bound the remaining matrices $\mathcal{Q}_1, \ldots \mathcal{Q}_{2d-1}$ in order to conclude that $\mathcal{M}$ is positive semidefinite. Next, we elaborate on the structure of these matrices. It turns out that we can define the "shape" of a graph $\mathcal{R}_m$ in an appropriate way so that

$$\mathcal{Q}_i^U(S_\ell, S_r) = \sum_{\text{shape}(\mathcal{R}_m)=U} c_i(\mathcal{R}_m)\chi_{\mathcal{R}_m}$$

where U is a finite (for constant d) sized graph with vertex set $A \cup B \cup C$, where we call A the "left" side of U and B the "right" side of U. Moreover $\mathcal{Q}_i = \sum_U \mathcal{Q}_i^U$. Now $\mathcal{Q}_i^U$ is a random matrix and special cases of this general family of matrices (for particular choices of U) arise in several earlier works on lower bounds for planted clique. Medarametla and Potechin [42] showed that the spectral norm of $\mathcal{Q}^U$ can be controlled by a bound on its coefficients and a few combinatorial parameters of U — namely $|\mathcal{V}(U)|$, $|A \cap B|$ and the number of vertex disjoint paths between A/B and B/A.

A major challenge in our work is to understand and analyze the coefficients c_i. In the course of decomposing $\mathcal{M}$, we are able to characterize $c_i(\mathcal{R}_m)$ as an appropriately weighted sum over $c_{i-1}(\mathcal{R}'_m)$ where $\mathcal{R}'_m$ ranges over the middle piece of all graphs with leftmost and rightmost separators S_ℓ and S_r that could have resulted in $\mathcal{R}_m$ due to repeated vertices. Recall that when there are repeated vertices, we take the parity of the edge sets of the three pieces and compute a new set of left and rightmost vertex separators. The set of $\mathcal{R}'_m$'s that could result in $\mathcal{R}_m$ is complicated. Instead, our approach is to show that the various combinatorial parameters of $\mathcal{R}'_m$

(which affect the spectral norm bounds) tradeoff against each other when accounting for the effect of repeated vertices. This allows us to bound their contribution and ultimately show that the coefficients c_i decay quickly enough for all values of $\omega < n^{1/2-\varepsilon}$ that we can bound each $\mathcal{Q}_i$ for $i > 1$ as $-\frac{D}{8d} \succeq \mathcal{Q}_i \succeq \frac{D}{8d}$, and this completes our proof.

Acknowledgment

We thank Raghu Meka, Ryan O'Donnell, Prasad Raghavendra, Tselil Schramm, David Steurer, and Avi Wigderson for many useful discussions related to this paper.

References

[1] R. M. Karp, "Probabilistic analysis of some combinatorial search problems," *Algorithms and Complexity: New Directions and Recent Results*, 1976.

[2] M. Jerrum, "Large cliques elude the metropolis process," *Random Struct. Algorithms*, vol. 3, no. 4, pp. 347–360, 1992. [Online]. Available: http://dx.doi.org/10.1002/rsa.3240030402

[3] L. Kucera, "Expected complexity of graph partitioning problems," *Discrete Applied Mathematics*, vol. 57, no. 2-3, pp. 193–212, 1995. [Online]. Available: http://dx.doi.org/10.1016/0166-218X(94)00103-K

[4] N. Alon, M. Krivelevich, and B. Sudakov, "Finding a large hidden clique in a random graph," in *SODA*, 1998, pp. 594–598.

[5] B. E. Hajek, Y. Wu, and J. Xu, "Computational lower bounds for community detection on random graphs," in *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, 2015, pp. 899–928. [Online]. Available: http://jmlr.org/proceedings/papers/v40/Hajek15.html

[6] *Combinatorial approaches to finding subtle signals in DNA sequences.*, vol. 8, 2000.

[7] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002. [Online]. Available: http://www.sciencemag.org/cgi/content/abstract/298/5594/824

[8] H. Javadi and A. Montanari, "The hidden subgraph problem," *arXiv preprint arXiv:1511.05254*, 2015.

[9] E. Hazan and R. Krauthgamer, "How hard is it to approximate the best nash equilibrium?" *SIAM J. Comput.*, vol. 40, no. 1, pp. 79–91, 2011. [Online]. Available: http://dx.doi.org/10.1137/090766991

[10] P. Austrin, M. Braverman, and E. Chlamtac, "Inapproximability of np-complete variants of nash equilibrium," *Theory of Computing*, vol. 9, pp. 117–142, 2013. [Online]. Available: http://dx.doi.org/10.4086/toc.2013.v009a003

[11] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie, "Testing k-wise and almost k-wise independence," in *STOC*, 2007, pp. 496–505.

[12] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," in *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, 2013, pp. 1046–1066. [Online]. Available: http://jmlr.org/proceedings/papers/v30/Berthet13.html

[13] P. Koiran and A. Zouzias, "Hidden cliques and the certification of the restricted isometry property," *IEEE Trans. Information Theory*, vol. 60, no. 8, pp. 4999–5006, 2014. [Online]. Available: http://dx.doi.org/10.1109/TIT.2014.2331341

[14] A. Juels and M. Peinado, "Hiding cliques for cryptographic security," *Des. Codes Cryptography*, vol. 20, no. 3, pp. 269–280, Jul. 2000. [Online]. Available: http://dx.doi.org/10.1023/A:1008374125234

[15] B. Applebaum, B. Barak, and A. Wigderson, "Public-key cryptography from different assumptions," in *STOC*, 2010, pp. 171–180.

[16] *Computational Complexity and Information Asymmetry in Financial Products (Extended Abstract)*, 2010.

[17] J. Feigenbaum and L. Fortnow, "Random-self-reducibility of complete sets," *SIAM J. Comput.*, vol. 22, no. 5, pp. 994–1005, 1993. [Online]. Available: http://dx.doi.org/10.1137/0222061

[18] A. Bogdanov and L. Trevisan, "On worst-case to average-case reductions for NP problems," *SIAM J. Comput.*, vol. 36, no. 4, pp. 1119–1159, 2006. [Online]. Available: http://dx.doi.org/10.1137/S0097539705446974

[19] U. Feige and R. Krauthgamer, "The probable value of the lovász–schrijver relaxations for maximum independent set," *SIAM J. Comput.*, vol. 32, no. 2, pp. 345–370, 2003.

[20] N. Z. Shor, "Class of global minimum bounds of polynomial functions," *Cybernetics*, vol. 23, no. 6, pp. 731–734, 1987, (Russian orig.: Kibernetika, No. 6, (1987), 9–11).

[21] P. A. Parrilo, "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization," Ph.D. dissertation, California Institute of Technology, May 2000.

[22] J. B. Lasserre, "An explicit exact sdp relaxation for nonlinear 0-1 programs," in *IPCO*, 2001, pp. 293–303.

[23] B. Barak, F. G. Brandao, A. W. Harrow, J. Kelner, D. Steurer, and Y. Zhou, "Hypercontractivity, sum-of-squares proofs, and their applications," in *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM, 2012, pp. 307–326.

[24] B. Barak, J. A. Kelner, and D. Steurer, "Rounding sum-of-squares relaxations," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. ACM, 2014, pp. 31–40.

[25] ——, "Dictionary learning and tensor decomposition via the sum-of-squares method," 2015.

[26] R. Meka, A. Potechin, and A. Wigderson, "Sum-of-squares lower bounds for planted clique," pp. 87–96, 2015. [Online]. Available: http://doi.acm.org/10.1145/2746539.2746600

[27] Y. Deshpande and A. Montanari, "Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems," *COLT*, 2015.

[28] S. B. Hopkins, P. Kothari, A. H. Potechin, P. Raghavendra, and T. Schramm, "On the integrality gap of degree-4 sum of squares for planted clique," in *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, 2016, pp. 1079–1095. [Online]. Available: http://dx.doi.org/10.1137/1.9781611974331.ch76

[29] D. Grigoriev, "Complexity of positivstellensatz proofs for the knapsack," *Computational Complexity*, vol. 10, no. 2, pp. 139–154, 2001.

[30] G. Schoenebeck, "Linear level lasserre lower bounds for certain k-csps," in *FOCS*, 2008, pp. 593–602.

[31] B. Barak, S. O. Chan, and P. K. Kothari, "Sum of squares lower bounds from pairwise independence," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, 2015, pp. 97–106. [Online]. Available: http://doi.acm.org/10.1145/2746539.2746625

[32] U. Feige and R. Krauthgamer, "The probable value of the lovász–schrijver relaxations for maximum independent set," *SIAM J. Comput.*, vol. 32, no. 2, pp. 345–370, 2003. [Online]. Available: http://dx.doi.org/10.1137/S009753970240118X

[33] T. Ma and A. Wigderson, "Sum-of-squares lower bounds for sparse pca," in *Advances in Neural Information Processing Systems*, 2015, pp. 1603–1611.

[34] S. B. Hopkins, P. K. Kothari, and A. Potechin, "Sos and planted clique: Tight analysis of MPW moments at all degrees and an optimal lower bound at degree four," *CoRR*, vol. abs/1507.05230, 2015. [Online]. Available: http://arxiv.org/abs/1507.05230

[35] E. T. Jaynes, "Information theory and statistical mechanics," *Physical review*, vol. 106, no. 4, p. 620, 1957.

[36] ——, "Information theory and statistical mechanics. ii," *Phys. Rev.*, vol. 108, pp. 171–190, Oct 1957. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRev.108.171

[37] T. Tao, "The dichotomy between structure and randomness, arithmetic progressions, and the primes," *arXiv preprint math/0512114*, 2005.

[38] E. Szemerédi, "Regular partitions of graphs," *ProblÃĺmes combinatoires et thÃl'orie des graphes*, 1978.

[39] A. Frieze and R. Kannan, "The regularity lemma and approximation schemes for dense problems," in *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, Oct 1996, pp. 12–20.

[40] "Probabilistic models and heuristics for the primes," 2015, available at https://terrytao.wordpress.com/2015/01/04/254a-supplement-4-probabilistic-models-and-heuristics-for-the-primes-optional/.

[41] A. Granville, "Harald cramér and the distribution of prime numbers," *Scandinavian Actuarial Journal*, vol. 1995, no. 1, pp. 12–28, 1995.

[42] D. Medarametla and A. Potechin, "Bounds on the norms of uniform low degree graph matrices," *Preprint*.