Semirandom Planted Clique and the Restricted Isometry Property

Jarosław Błasiok
jaroslaw.blasiok@inf.ethz.ch
ETH Zürich

Rares-Darius Buhai
rares.buhai@inf.ethz.ch
ETH Zürich

Pravesh K. Kothari
kothari@cs.princeton.edu
IAS & Princeton University

David Steurer
dsteurer@inf.ethz.ch
ETH Zürich

October 10, 2024

Abstract

We give a simple, greedy $O(n^{\omega+0.5}) = O(n^{2.872})$-time algorithm to list-decode planted cliques in a semirandom model introduced in [CSV17] (following [FK01]) that succeeds whenever the size of the planted clique is $k \geq O(\sqrt{n}\log^2 n)$. In the model, the edges touching the vertices in the planted k-clique are drawn independently with probability $p = 1/2$ while the edges not touching the planted clique are chosen by an adversary in response to the random choices. Our result shows that the computational threshold in the semirandom setting is within a $O(\log^2 n)$ factor of the information-theoretic one [Ste17] thus resolving an open question of Steinhardt. This threshold also essentially matches the conjectured computational threshold for the well-studied special case of fully random planted clique.

All previous algorithms [CSV17, MMT20, BKS23] in this model are based on rather sophisticated rounding algorithms for entropy-constrained semidefinite programming relaxations and their sum-of-squares strengthenings and the best known guarantee is a $n^{O(1/\varepsilon)}$-time algorithm to list-decode planted cliques of size $k \geq \tilde{O}(n^{1/2+\varepsilon})$. In particular, the guarantee trivializes to quasi-polynomial time if the planted clique is of size $O(\sqrt{n}\operatorname{polylog} n)$. Our algorithm achieves an almost optimal guarantee with a surprisingly simple greedy algorithm.

The prior state-of-the-art algorithmic result above is based on a reduction to certifying bounds on the size of unbalanced bicliques in random graphs — closely related to certifying the restricted isometry property (RIP) of certain random matrices and known to be hard in the low-degree polynomial model. Our key idea is a new approach that relies on the truth of — but not efficient certificates for — RIP of a new class of matrices built from the input graphs.

1 Introduction

Finding planted cliques [Jer92, Kuc95] in Erdős-Rényi random graphs $G(n, 1/2)$ is a well-studied average-case variant of the notoriously hard clique problem. The added clique is uniquely identifiable with high probability (and recoverable via a brute-force quasi-polynomial time algorithm) whenever it has size $k \gg 2\log n$. However, the best known efficient algorithms require $k \gg \sqrt{n}$. The algorithms themselves are rather simple. The simple greedy degree heuristic of reporting the k largest-degree vertices works if $k \gg \sqrt{n \log n}$ and the $\sqrt{\log n}$ factor can be shaved off by a natural spectral algorithm [AKS98]. Lower bounds in restricted models [BHK⁺16, FGR⁺17] provide evidence that efficient algorithms cannot beat the $\sqrt{n}$ threshold and this *information-computation* gap is the fountainhead of several hardness results in average-case optimization [BBH18, BB20].

The simple algorithms above for finding planted cliques are rather brittle — modifying only $O(k^2)$ edges that do not even touch the planted clique is enough to completely break their guarantees. To disallow such brittle heuristics that tend to "overfit" to the specific idealized random model, Feige and Kilian [FK01] introduced *semirandom* models [Fei19] for the planted clique problem (following the seminal work of Blum and Spencer [BS95]). The input graph in their model is chosen by a combination of benign random and adaptive adversarial choices that preclude brittle heuristics while still hopefully steering clear of the worst-case hard instances. Such a semirandom model is the main focus of this work:

Definition 1.1 (Semirandom planted clique, $\mathsf{SRPC}(n, k, p)$ [CSV17])**.** To an empty graph $G([n], E)$:

1. *Plant a clique*: Plant a clique on an arbitrary subset of vertices $S^* \subseteq [n]$ with $|S^*| = k$.

2. *Include cut edges at random*: Add each edge in $\mathsf{cut}(S^*)$ independently with probability p.

3. *Choose rest of the edges adversarially*: Adaptively choose any induced graph on $[n] \setminus S^*$.

In the fully random planted clique setting, the edges in the third step are also chosen at random. Note further that the adversarial choice is *adaptive*, i.e., made in response to the random choice in the second step. Approximating the maximum clique in such a graph is clearly as hard as the worst-case variant [Hås99, Zuc07], since in the third step we can simply plant a worst-case hard instance on $[n] \setminus S^*$. The "right" goal, instead, is to find a k-clique in the input graph. In fact, all known algorithms focus on the formally stronger goal of finding a small ($\sim n/k$ in the strongest possible results) *list* of k-cliques guaranteed to contain S^*. This *list-decoding* goal relaxes the usual *unique recovery* goal that is clearly impossible in the semirandom model.

A new course of progress on the problem was begun in [CSV17] who gave a polynomial-time algorithm based on rounding an appropriate semidefinite programming relaxation that works whenever $k \geq O(n^{2/3} \log^{1/3} n)$. McKenzie, Mehta and Trevisan [MMT20] improved this threshold by logarithmic factors to $k \geq O(n^{2/3})$. Recently [BKS23] made further progress by finding an $n^{O(1/\varepsilon)}$-time algorithm that succeeds whenever the planted clique has size $k \geq n^{1/2+\varepsilon}$ for any $\varepsilon > 0$. The results in the latter two works also tolerate a monotone adversary that can delete an arbitrary subset of edges in $\mathsf{cut}(S^*)$. On the flip side, Steinhardt [Ste17] proved that if $k = o(\sqrt{n})$ it is *information-theoretically* impossible to recover a list of size $O(n/k)$ that is guaranteed to contain the planted clique with high probability.[1][2] That is, the information-theoretic threshold for the natural estimation task is *shifted* from $\lceil 2\log_2 n \rceil$ to $\sim \sqrt{n}$ in the semirandom model. If this threshold were also achievable by a polynomial-time algorithm, Steinhardt noted, then the widely conjectured information-computation gap in the fully random planted clique problem disappears for the natural *robust* version of the problem studied in this (and his) work. With this motivation, Steinhardt [Ste17] (and for a closely related variant, Feige [Fei19]) explicitly poses the natural question of whether semirandom planted clique admits an efficient algorithm at $k \sim \tilde{O}(\sqrt{n})$.

When $k = \sqrt{n}\,\mathrm{polylog}\, n$, the state-of-the-art algorithm above devolves into a near-brute-force running time of $n^{O(\log n/\log\log n)}$. Further, even for $k \geq n^{1/2+\varepsilon}$, all the known results above [CSV17, MMT20, BKS23] are based on rather sophisticated rounding algorithms for semidefinite programming relaxations (and their sum-of-squares strengthenings) with *entropy maximization* constraints. This raises the natural question of whether simple heuristics could succeed at $k = \tilde{O}(\sqrt{n})$ or at least match the best known results.

This work. In this work, we give a surprisingly simple greedy algorithm that finds a list of size $(1 + o(1))n/k$ containing the planted clique with high probability, whenever $k \geq O(\sqrt{n}\log^2 n)$, essentially resolving Steinhardt's question. Our algorithm admits an optimized implementation via black-box usage of fast matrix multiplication and runs in time $O(n^{\omega+0.5}) = O(n^{2.872})$.

Theorem 1.2 (Main result)**.** *There is an $O(n^{\omega+0.5})$-time algorithm, where $\omega \leq 2.372$ is the matrix multiplication exponent, that takes as input a graph $G \sim \mathsf{SRPC}(n, k, 1/2)$ and, if $k \geq O(\sqrt{n}\log^2 n)$, outputs a list of k-cliques of size $(1 + o(1))n/k$ such that, with probability at least 0.99 over the draw of G and the random choices of the algorithm, S^* is contained in the list.*

Observe that our result comes within a $O(\log^2 n)$ factor of the information-theoretic lower bound [Ste17].[3] Theorem 4.1 presents an appropriate generalization to $\mathsf{SRPC}(n, k, p)$ for arbitrary p that improves upon the bound on k in previous results when $p \leq 1 - n^{-0.01}$.

A new connection to RIP. Our algorithm relies on a *new* connection to the *restricted isometry property*[4] of a certain $n^{O(1)}$-size matrix built from the graph that relies only on the randomness of the edges in the cut defined by the planted clique. We show how to relate the success of a simple greedy algorithm to the RIP property of this matrix that, crucially, *does not need* any efficient certificates for the RIP itself. This is crucial because we need the RIP property in the near optimal (and conjecturally, hard to efficiently certify) parameter regimes.

[1] Observe that a disjoint union of k-cliques with all the remaining edges chosen independently with probability $1/2$ is a valid input instance in the model, with n/k indistinguishable k-cliques. Thus, the minimum list size possible is $\geq n/k$.

[2] The lower bound instance formally adds a clique that is slightly larger than k — hence, the input k is a *lower bound* on the size of the planted clique. All algorithmic results including ours easily generalize to only need a lower bound on size of the planted clique.

[3] Steinhardt suggests that the lower bound on the threshold can possibly be improved to $k \geq O(\sqrt{n\log n})$ — the threshold at which the natural quasi-polynomial time brute-force algorithm works. This would make the threshold of our algorithm off by a $\log^{1.5} n$ factor.

[4] An $m \times n$ matrix H is said to satisfy (r, δ)-RIP if for every r-sparse vector v, $\|Hv\|_2 \in (1 \pm \delta)\,\|v\|_2$.

To put this discussion in proper context, recall that the fully-random planted clique problem is closely related to efficiently certifying sparse quadratic forms (i.e., certifying upper bounds on $v^\top Av$ for the adjacency matrix A of the input graph and sparse v) or sparse PCA but these reductions also produce instances in the conjectured hard regimes [BR13b, HKP$^+$17, KZ14a]. Indeed, such results are usually interpreted as hardness results for certifying the RIP and bounds on sparse quadratic forms (assuming the planted clique conjecture). To use such reductions algorithmically, we need efficient certificates of RIP in conjectured hard parameter regimes [BDMS13, KZ14b, DKWB21][5].

As an important special case, we note that the problem of certifying bounds on the biclique numbers of random graphs that forms the crux of the approach for semirandom planted clique in [BKS23] also reduces to certifying RIP in the conjectured hard regime.

Our reduction, in contrast, shows how to solve the semirandom planted clique problem as long as some matrices (rather different from the ones arising in the above reductions) built from the input graph satisfy RIP. Crucially, we *do not need* efficient certificates of RIP — this important difference allows our approach to obtain a significantly better guarantee in this work.

Monotone deletions. As we noted above, while the original work of Charikar, Steinhardt and Valiant [CSV17], the open question of Steinhardt [Ste17], and the lower bound [Ste17] are all phrased about the model studied in this paper, the results in [MMT20, BKS23] tolerate, in addition, an adversary that can delete an arbitrary subset of edges in $\mathsf{cut}(S^*)$ [FK01]. Our approach needs certificates of RIP (in the conjectured hard regime) to tolerate such a monotone adversary. A similar bottleneck prevents algorithms based on certifying biclique numbers to reach the $k = \sqrt{n}$ polylog n threshold. This is made formal via low-degree polynomial lower bounds in [BKS23].

Our result raises an intriguing possibility of a shift in the *computational threshold* for a natural estimation problem due to a monotone adversary. This is in contrast to the intriguing work of Moitra, Perry and Wein [MPW16] who showed that the *information-theoretic threshold* for community detection in the stochastic block model *does shift* under a monotone adversary.

2 Overview of our algorithm

In this section we give a brief overview of our approach for semirandom planted clique. As in the prior work [BKS23] (see Lemma B.3 in the Appendix), when $k \geq O(\sqrt{n \log n})$ we can prune any polynomial-size list containing S^* down to size $(1 + o(1))n/k$ without removing S^*. This can be achieved by removing from the list all cliques that have intersection larger than $\Omega(\log n)$ with any other clique in the list. Therefore we will focus on how to recover a list of k-cliques containing S^* of polynomial size.

Recalling the approach of [BKS23]. Before describing our approach, let us recall the key idea in the prior state-of-the-art [BKS23] that gave a polynomial-time algorithm that works whenever

[5] For example, while an $m \times n$ matrix with independent $\pm 1/\sqrt{m}$ entries is known to be $(r, 0.1)$-RIP for $r \sim \Omega(m/\log(n/m))$, efficiently certifying such a fact is conjectured to be hard [BDMS13, KZ14b, DKWB21].

$k \geq n^{1/2+\varepsilon}$ for an arbitrary constant $\varepsilon > 0$. Their main idea is a reduction (within the sum-of-squares framework) to efficiently "certifying" bounds on the biclique numbers of bipartite random graphs. In particular, in order for their algorithm to succeed at $k = \tilde{O}(\sqrt{n})$, they need a constant-degree sum-of-squares certificate (roughly, polynomial-time verifiable via an SDP relaxation) that a $k \times n$ bipartite random graph (each edge included independently with probability 1/2) has no $O(\log^C n) \times k$ bipartite clique in it for $C = O(1)$ (with high probability). Crucially, they need certificates even for the model *without monotone deletions*. This certification task is naturally related (and, in fact, reduces) to *certifying* the restricted isometry property of matrices built from the bipartite random graph in the conjectured-to-be-hard regime [BDMS13, KZ14b]. Unfortunately, they also show that certifying such biclique numbers likely suffers from information-computation gaps by establishing lower bounds in the low-degree polynomial model, and thus appear to run into an inherent barrier in obtaining algorithms that work when $k = \sqrt{n}\log^{O(1)}(n)$.

Our main idea is a new approach that circumvents the need for certificates on the biclique number of bipartite random graphs altogether and succeeds with just the truth of (but without efficient certificates for) the restricted isometry property. Our approach leads to an algorithm that does not need semidefinite programming and is based on a simple, efficient, greedy procedure, the correctness of which relies on the RIP of a natural matrix built from the input graph. In the following, we explain how to build up to this procedure by starting with a naive greedy procedure.

2.1 The naive greedy algorithm

We let G denote the ± 1-adjacency matrix of the graph G, i.e., $G(i,j) = 1$ iff $\{i,j\}$ is an edge in G. We also let G_i be the i-th row of G, $G_i^{\text{in}} \in \{\pm 1\}^k$ be the projection of G_i to coordinates in S^*, and $G_i^{\text{out}} \in \{\pm 1\}^{n-k}$ the projection of G_i to coordinates in $[n] \setminus S^*$.

Let us make the simple observation that G_i and G_j are non-trivially correlated if i and j both are in S^*. Indeed, we have $\langle G_i, G_j \rangle = \langle G_i^{\text{in}}, G_j^{\text{in}} \rangle + \langle G_i^{\text{out}}, G_j^{\text{out}} \rangle$. Since $i, j \in S^*$, the first term is clearly k. Further, since every edge in $\mathsf{cut}(S^*)$ is chosen independently to be in G with probability 1/2, G_i^{out} and G_j^{out} are uniformly random and independent elements of $\{\pm 1\}^{n-k}$ and thus, with $1 - 1/n^{O(1)}$ probability over the draw of edges in $\mathsf{cut}(S^*)$, $|\langle G_i^{\text{out}}, G_j^{\text{out}} \rangle| \leq O(\sqrt{n \log n})$. Thus, $\langle G_i, G_j \rangle \geq k \pm O(\sqrt{n \log n}) \geq k/2$ if $k \geq O(\sqrt{n \log n})$.

This observation naturally suggests the following simple algorithm:

Naive greedy procedure: For a uniformly random $i \in [n]$, add $S_i = \{j : \langle G_i, G_j \rangle \geq k/2\}$ to the list.

We choose an $i \in S^*$ with probability k/n and, thus, repeating the above procedure $O(n/k)$ times ensures that with probability at least 0.99 we pick some $i \in S^*$. From the above correlation computation, $S^* \subseteq S_i$ for any $i \in S^*$. Could such an S_i contain a $j \notin S^*$?

For an $i \in S^*$ and $j \notin S^*$, we can write $\langle G_i, G_j \rangle = \langle G_i^{\text{in}}, G_j^{\text{in}} \rangle + \langle G_i^{\text{out}}, G_j^{\text{out}} \rangle$. The first term is at most $O(\sqrt{k \log n})$ with high probability. However, we have little control on the second term since G_j^{out} is chosen by an adversary (in response to the random choice of edges in $\mathsf{cut}(S^*)$). In fact, it turns out that the adversary can arrange (multiple!) $j \notin S^*$ such that $\langle G_i, G_j \rangle \gg k$ *simultaneously* for

$\geq O(n^2/k^2)$ different $i \in S^*$ — see Lemma C.1. Notice that $n^2/k^2 \gg k$ if $k = \tilde{O}(\sqrt{n})$, in which case the adversary can ensure that *every* $i \in S^*$ fails to produce $S_i = S^*$ in the above greedy algorithm.

Analyzing naive greedy algorithm for $k \geq \tilde{O}(n^{3/4})$. While the naive greedy procedure fails for $k = \tilde{O}(\sqrt{n})$, we now argue that it does succeed for $k \geq \tilde{O}(n^{3/4})$. This will form the starting point of our new approach.

The following simple lemma uses standard bounds on random matrices to control the number of spurious $j \notin S^*$ that can be in S_i for $i \in S^*$.

Lemma 2.1. *Let $v_1, v_2, \ldots, v_k \in \{\pm 1\}^n$ be uniformly random and independent vectors for $k \geq O(\sqrt{n \log n})$. Then with probability at least $1 - 1/n$ over the draw of vectors v_i, for every $u \in \{\pm 1\}^n$ there are at most $O(n^2/k^2)$ vectors v_i such that $\langle u, v_i \rangle \geq k/3$.*

Proof of Lemma 2.1. Consider the $n \times k$ matrix H with columns $v_1, v_2, \ldots, v_k$. Then, by standard results in random matrix theory, $\|H\|_2 \leq O(\sqrt{n} + \sqrt{k})$ with probability $1 - 1/n$. Let $\mathcal{B} \subseteq [k]$ be such that, for every $i \in \mathcal{B}$, $\langle u, v_i \rangle \geq k/3$. Then, by the Cauchy-Schwartz inequality, we have

$$\langle u, H\mathbf{1}_{\mathcal{B}} \rangle \leq \|u\|_2 \|H\mathbf{1}_{\mathcal{B}}\|_2 \leq \sqrt{n}\sqrt{|\mathcal{B}|}\|H\|_2 \leq O(n\sqrt{\mathcal{B}})\,.$$

On the other hand, by the choice of set $\mathcal{B}$, we have

$$\langle u, H\mathbf{1}_{\mathcal{B}} \rangle = \sum_{i \in \mathcal{B}} \langle u, v_i \rangle \geq |\mathcal{B}|k/3\,.$$

Combining those two inequalities and rearranging, we get $|\mathcal{B}| \leq O(n^2/k^2)$. □

Using this lemma, it is easy to show that if $k \gg n^{3/4}$ then for a uniformly random $i \in S^*$ we have $|S_i \setminus S^*| \leq o(k)$: We apply Lemma 2.1 with vectors G_i^{out} for $i \in S^*$ — each a uniformly random and independent element of $\{\pm 1\}^{n-k}$. Then, taking $u = G_j^{\text{out}}$ for any $j \in [n]$, we know that there are at most $O(n^2/k^2)$ different $i \in S^*$ such that $\langle G_i^{\text{out}}, G_j^{\text{out}} \rangle \geq k/3$. Thus, there are in total $O(n^3/k^2)$ pairs (i, j) such that $i \in S^*$ and $j \notin S^*$ with $\langle G_i, G_j \rangle \geq k/2 \geq k/3 + O(\sqrt{k \log n})$. In particular, by averaging, a uniformly random $i \in S^*$ satisfies $\langle G_i, G_j \rangle \geq k/2$ for at most Cn^3/k^3 different $j \notin S^*$ (for some fixed constant $C > 0$) with probability at least 0.99. If $k \gg O(n^3/k^3)$ (e.g., if $k \geq \omega(n^{3/4})$), then $|S_i \setminus S^*| \leq o(k)$. Finally, from any such set $S_i \supseteq S^*$ that has at most $o(|S^*|)$ erroneous elements, one can obtain S^* by a simple pruning procedure — for instance by keeping only vertices that have at least $k - 1$ neighbors in S_i (see, e.g., Claim 5.9 on page 35 of [BKS23]).

2.2 Tensoring twice: greedy matches the guarantees of [BKS23]

The simple greedy procedure above 1) relies on standard spectral norm bounds on random matrices with independent entries and 2) gets stuck at $k \geq \tilde{O}(n^{3/4})$. We now show how to obtain an algorithm that works for any $k \geq n^{1/2+\varepsilon}$ in time $n^{O(1/\varepsilon)}$, matching the guarantees of [BKS23] via a simple greedy algorithm analogous to the one in previous subsection.

For $\alpha \subset [n]$ of size $|\alpha| = 2$, let $G_\alpha \in \{\pm 1\}^n$ be the vector so that $G_\alpha(i) = G_{\alpha_1,i} \cdot G_{\alpha_2,i}$. That is, $G_\alpha(i)$ is the product of the ±1-indicators of the two edges (α_1, i) and (α_2, i). Note that G_α is the Hadamard (and not the tensor) product of G_{α_1} and G_{α_2}.

Similarly to the previous section, it is easy to observe that for every $\alpha, \alpha' \subset S^*$, $\langle G_\alpha, G_{\alpha'} \rangle \geq k/2$ if $k \geq O(\sqrt{n \log n})$. This motivates the following generalization of the greedy procedure above:

"Tensored" greedy procedure: For a uniformly random $\alpha \in \binom{[n]}{2}$, add $S_\alpha = \{i : \langle G_i, G_\alpha \rangle \geq k/2\}$ to the list.

Let us analyze this algorithm. The same argument as before shows that $S_\alpha \supseteq S^*$ if $\alpha \subset S^*$. Further, a uniformly random α is in S^* with probability $(k/n)^2$, so repeating the above procedure $O(n/k)^2$ times includes a uniformly random sample of $\alpha \subset S^*$.

We will now provide a generalization of Lemma 2.1 with no change in parameters that allows reasoning about $|S_\alpha \setminus S^*|$. The proof, however, will rely on the RIP instead of the straightforward spectral norm argument above.

Lemma 2.2. *Let $v_1, v_2, \ldots, v_k \in \{\pm 1\}^n$ be uniformly random and independent vectors for $k \geq O(\sqrt{n \log n})$ and let $v_\alpha \in \{\pm 1\}^n$ for $\alpha \subset [k]$ of size $|\alpha| = 2$ be defined by $v_\alpha(i) = v_{\alpha_1}(i) v_{\alpha_2}(i)$ for every i. Then with probability at least $1 - 1/n$ over the draw of vectors v_i, for every $u \in \{\pm 1\}^n$ there are at most $O(n^2/k^2)$ vectors v_α such that $\langle u, v_\alpha \rangle \geq k/3$.*

Tensored greedy succeeds for $k \geq \tilde{O}(n^{0.6})$. Before proving Lemma 2.2, let us repeat the calculations we did above to see if we improve the range of k where our algorithm succeeds. By applying the lemma above to the vectors G_α^{out} for $\alpha \in S^*$, we obtain that the number of pairs (α, j) such that $\langle G_\alpha^{\text{out}}, G_j^{\text{out}} \rangle \geq k/3$ and $j \notin S^*$ is at most $O(n^3/k^2)$. Thus, by averaging, at least 1/2 of the $\sim k^2$ different sets α must satisfy $\langle G_\alpha^{\text{out}}, G_j^{\text{out}} \rangle \geq k/3$ for at most $O(n^3/k^4)$ different j. In particular, $|S_\alpha \setminus S^*| \leq O(n^3/k^4)$ for 1/2 fraction of $\alpha \subset S^*$. This bound is $\ll k$ if $n^3/k^4 \ll k$ or $k \gg n^{3/5} = n^{0.6}$.

Observe that this already improves on the $n^{3/4}$ bound above. In fact, this simple greedy algorithm already improves on the (more involved) algorithms in prior works [CSV17, MMT20] that work when $k \gg n^{2/3} (\gg n^{3/5})$!

Matching the guarantee of [BKS23]. We now show a simple modification of the greedy procedure above that works to give an $n^{O(1/\varepsilon)}$-size list when $k \sim n^{1/2+\varepsilon}$, allowing us to match the running-time vs k trade-off of the recent work [BKS23].

Procedure: Fix $t \in \mathbb{N}$. For independent and uniformly random $\alpha_1, \alpha_2, \ldots, \alpha_t \in \binom{[n]}{[2]}$, add $S_{\alpha_1,\alpha_2,\ldots,\alpha_t} = \{j : \langle G_j, G_{\alpha_i} \rangle \geq k/2 \text{ for every } 1 \leq i \leq t\}$ to the list.

Note that tensored greedy procedure from before corresponds to $t = 1$ in the above algorithm. Observe that, with probability $(k/n)^{2t}$, $\alpha_1, \alpha_2, \ldots, \alpha_t \subset S^*$, so repeating the procedure $O(n/k)^{2t}$ times includes a random subset of S^*. As before, for such tuples $S_{\alpha_1,\alpha_2,\ldots,\alpha_t} \supseteq S^*$. We will show that for a 0.99-fraction of tuples $\alpha_1, \alpha_2, \ldots, \alpha_t \subset S^*$, we have $S_{\alpha_1,\alpha_2,\ldots,\alpha_t} = S^*$ if $k \geq O(n^{1/2+1/2t})$. Taking $t = 1/\varepsilon$ yields the trade-off obtained in [BKS23].

Let us see why: from Lemma 2.2, for every $j \notin S^*$, the number of sets $\alpha \subset S^*$ such that $\langle G_\alpha, G_j \rangle \geq k/2$ is at most $O(n^2/k^2)$. Thus, if we pick t uniformly random subsets $\alpha_i \subset S^*$, the chance that for a given j we have $\langle G_{\alpha_i}, G_j \rangle \geq k/2$ for every $1 \leq i \leq t$ is at most $((n^2/k^2)/k^2)^t = (n^2/k^4)^t$. If $k \geq 100n^{1/2+1/2t}$, then, this chance is clearly $\ll 1/n$ and thus, by a union bound over the choice $j \notin S^*$, with probability at least 0.99 over the choice of the α_i we have $S_{\alpha_1,\alpha_2,\ldots,\alpha_t} = S^*$.

Proving Lemma 2.2 using RIP. Let us introduce the star of the show to prove Lemma 2.2: the restricted isometry property.

Definition 2.3 (Restricted isometry property)**.** An $m \times d$ matrix H is said to be (r, δ)-RIP if for every $v \in \mathbb{R}^d$ such that $\|v\|_0 \leq r$ we have

$$(1-\delta)\|v\|_2 \leq \|Hv\|_2 \leq (1+\delta)\|v\|_2 .$$

In fact, for the purpose of the analysis of the algorithm, we care only about the upper bound in the restricted isometry property. More concretely:

Definition 2.4 (Sparse operator norm)**.** An $m \times d$ matrix H is said to be r-sparse operator norm bounded by C if for every $v \in \mathbb{R}^d$ such that $\|v\|_2 \leq 1$, $\|v\|_0 \leq r$, we have $\|Hv\|_2 \leq C$.

We note that sparse operator norm behaves differently compared to the sparse quadratic form in our setting (i.e., maximizing $v^\top Hv$ over sparse vectors for square H). The latter occurs naturally in the context of sparse PCA and has connections to the planted clique problem [BR13a].

Clearly, a bound on the sparse operator norm is implied by the RIP — if a matrix H is (r, δ)-RIP, then its r-sparse operator norm is bounded by $1+\delta$. Now, it is well-know that an $m \times d$ ($m \ll d$) matrix of independent, uniform ±1 entries satisfies (r, C)-RIP for $C = O(1)$ and $r = \Omega(m/\log(d/m))$. We will actually establish and use the RIP of a random matrix with significantly correlated columns.

Fact 2.5 (See Lemma 3.4)**.** *Let $H \in \mathbb{R}^{(n-k)\times\binom{k}{2}}$ be a matrix with columns G_α^{out} for $\alpha \subset S^*$ of size $|\alpha| = 2$. Then with high probability $H/\sqrt{n-k}$ satisfies $(r, O(1))$-RIP for some $r \geq \Omega(n/\log^{O(1)}(n))$.*

Observe that H is a function of at most kn random bits even as it has nk^2 entries. Despite this highly dependent setting, it turns out that H still satisfies RIP. This follows from a powerful result of [RV08] that allows us to conclude that a matrix with independent rows chosen from an isotropic, ℓ_∞-bounded distribution satisfies strong RIP properties. The following theorem is usually stated in terms of *bounded orthonormal systems*. In Appendix A, we discuss how this standard formulation is equivalent to the following version that is directly interpretable and useful for us.

Theorem 2.6 (Corollary of [RV08, FR13])**.** *Let $A \in \mathbb{R}^{m\times N}$ be a random matrix with rows sampled i.i.d. according to the distribution of some random variable $\mathbf{X} \in \mathbb{R}^N$ that satisfies $\mathbb{E}\mathbf{X}\mathbf{X}^\top = I_N$ and $\|\mathbf{X}\|_\infty \leq K$ almost surely. Then, for any $m \geq O(rK^2 \log^3(r)\log(N)/\delta^2)$, we have with probability at least $1 - N^{-\log^3(r)}$ that the matrix $A/\sqrt{m}$ satisfies (r, δ)-RIP.*

To finish this section, let us see how the sparse operator norm bound helps settle Lemma 2.2.

Proof of Lemma 2.2. Let H be the matrix from Fact 2.5, with r-sparse operator norm of $H/\sqrt{n-k}$ bounded by $O(1)$ for $r \geq \Omega(n/\log^{O(1)}(n))$. Suppose (toward a contradiction) that there is a vector $u \in \{\pm 1\}^{n-k}$ such that $\langle G_\alpha, u \rangle \geq k/3$ for all $\alpha \in \mathcal{B}$, where $\mathcal{B}$ is a set of size r (if the set is larger than r in size, just choose any r-size subset). Then, using Cauchy-Schwarz and the sparse operator norm bound on H,

$$\langle u/\|u\|_2\,, H\mathbf{1}_{\mathcal{B}}/\sqrt{|\mathcal{B}|}\rangle \leq \left\| H\mathbf{1}_{\mathcal{B}}/\sqrt{|\mathcal{B}|} \right\|_2 \leq O(\sqrt{n-k})\,,$$

while by the choice of $\mathcal{B}$ we have

$$\langle u/\|u\|_2, H\mathbf{1}_{\mathcal{B}}/\sqrt{|\mathcal{B}|}\rangle \geq \frac{\sqrt{|\mathcal{B}|}k}{3\sqrt{n-k}}\,.$$

Rearranging yields that $|\mathcal{B}| \leq O(n^2/k^2)$. □

Tensoring thrice: our algorithm. At this point, it is intuitive to try the greedy procedure with tensoring thrice. Indeed, this is precisely our algorithm, with the crux of the analysis being the RIP of a matrix analogous to the one above. Our proof is short and simple and is presented in full in the following section.

3 Semirandom planted clique with $p = 1/2$

We will prove the following result in this section.

Theorem 3.1. *There exists an $O(n^{\omega+0.5})$-time algorithm, where $\omega \leq 2.372$ is the matrix multiplication exponent, that takes input a graph G generated according to* $\mathsf{SRPC}(n, k, 1/2)$ *for any $k \geq O(\sqrt{n}\log^2 n)$ and with probability at least* 0.99 *outputs a list of size $(1 + o(1))n/k$ of k-cliques containing the planted clique in G.*

We will first analyze the following simple algorithm:

Algorithm 1:

1. For $O((n/k)^3)$ rounds, sample $\alpha \subset [n]$ with $|\alpha| = 3$ and construct $S_\alpha := \{j : \langle G_\alpha, G_j \rangle \geq k/2\}$.
2. Construct a refined set $S'_\alpha \subseteq S_\alpha$ by removing all vertices from S_α that are connected to less than $k-1$ of the vertices in S_α, and then add S'_α to the list if it forms a k-clique.
3. Apply pruning (see Appendix B) to reduce the list size to $(1 + o(1))n/k$.

This algorithm naively runs in time $O(n^{3.5})$. Later we will discuss how to implement computationally expensive steps in this algorithm by black-box calls to a matrix multiplication oracle, improving the running time to $O(n^{\omega+0.5})$.

We will prove two claims: First, that for all $\alpha \subset S^*$ with $|\alpha| = 3$ we have $S_\alpha \supseteq S^*$, and second, that for any fixed $j \notin S^*$, there are at most $O(n^2/k^2)$ sets $\alpha \subset S^*$ with $|\alpha| = 3$ such that $j \in S_\alpha$. The first claim:

Lemma 3.2. *Let $k \geq O(\sqrt{n \log n})$. With high probability, for all $\alpha \subset S^*$ with $|\alpha| = 3$ and all $j \in S^*$ with $j \notin \alpha$, we have that $\langle G_\alpha, G_j \rangle \geq k/2$.*

Proof. We have that $\langle G_\alpha^{\text{in}}, G_j^{\text{in}} \rangle = k$. On the other hand, G_α^{out} and G_j^{out} have independent Rademacher entries, so by Hoeffding's inequality

$$\mathbb{P}\left(|\langle G_\alpha^{\text{out}}, G_j^{\text{out}} \rangle| \geq t\right) \leq 2 \exp\left(-\Omega(t^2/n)\right) .$$

Taking $t = O(\sqrt{n \log n})$ large enough, we have that $|\langle G_\alpha^{\text{out}}, G_j^{\text{out}} \rangle| \leq t$ with probability at least $1 - 1/n^{10}$. We choose k such that $t \leq k/2$, so $\langle G_\alpha, G_j \rangle \geq k - t \geq k/2$. Finally, an application of the union bound shows that this holds simultaneously for all choices of α and j with probability at least $1 - o(1)$. □

Next, we want to prove that each $j \notin S^*$ satisfies $\langle G_\alpha, G_j \rangle \geq k/2$ for at most $O(n^2/k^2)$ sets $\alpha \subset S^*$ with $|\alpha| = 3$. It is easy to bound $|\langle G_\alpha^{\text{in}}, G_j^{\text{in}} \rangle| \leq O(\sqrt{k \log n})$ using the fact that G_j^{in} has independent Rademacher entries. To talk about $\langle G_\alpha^{\text{out}}, G_j^{\text{out}} \rangle$, we define the matrix $H \in \{\pm 1\}^{(n-k) \times \Theta(k^3)}$ with columns G_α^{out} for all $\alpha \subset S^*$ with $|\alpha| = 3$. A simple observation shows that a bound on the r-sparse operator norm of the matrix $H/\sqrt{n-k}$ ensures the desired bound.

Lemma 3.3. *Let $H \in \mathbb{R}^{q \times m}$ be a matrix satisfying*

$$\sup_{\|v\|_2 \leq 1, \|v\|_0 \leq r} \|Hv\| \leq C \,, \tag{1}$$

with $r \geq C^2/\tau^2 + 1$. Then any vector $w \in \mathbb{R}^q$ with $\|w\|_2 \leq 1$ has inner product greater than or equal to τ with at most C^2/τ^2 columns of H.

Proof. Let $\mathcal{B} \subset [m]$ be any set of size $|\mathcal{B}| \leq r$, such that for all $t \in \mathcal{B}$ we have $\langle H_{\cdot,t}, w \rangle \geq \tau$. Then on one hand we have by Cauchy-Schwarz and (1) that

$$\langle H\mathbf{1}_{\mathcal{B}}, w \rangle \leq \|H\mathbf{1}_{\mathcal{B}}\| \cdot \|w\| \leq C\sqrt{|\mathcal{B}|} \,,$$

and on the other hand we have by the definition of $\mathcal{B}$ that

$$\langle H\mathbf{1}_{\mathcal{B}}, w \rangle = \sum_{t \in \mathcal{B}} \langle H_{\cdot,t}, w \rangle \geq \tau |\mathcal{B}| \,.$$

This leads, after rearranging, to $|\mathcal{B}| \leq C^2/\tau^2$. □

If $H \in \mathbb{R}^{q \times m}$ was indeed a matrix with independent Rademacher entries, it is well-known that $H/\sqrt{q}$ would satisfy RIP of order $r = \Omega(q/\log(m/q))$. For the matrix H generated by taking the products of 3-tuples of random signs that we care about, the columns are heavily dependent. It turns out, however, that we can still establish that $H/\sqrt{q}$ satisfies the restricted isometry property (with only a slight degradation in the sparsity parameter r).

Lemma 3.4. *Let $\mathcal{D}$ be any distribution over $\mathbb{R}$ such that $\mathbf{x} \sim \mathcal{D}$ satisfies $\mathbb{E}\mathbf{x} = 0, \mathbb{E}\mathbf{x}^2 = 1$, and $|\mathbf{x}| \leq B$ almost surely.*

Consider a matrix $H \in \mathbb{R}^{q \times \Theta(k^t)}$ with independent rows, where the columns are indexed by subsets $\alpha \subset [k]$ of size t for constant t, and the rows are generated by first drawing $\mathbf{X}_i \sim \mathcal{D}^k$ and setting $H_{i,\alpha} = \prod_{j \in \alpha} X_{i,j}$. Then with high probability $H/\sqrt{q}$ satisfies $(r, O(1))$-RIP with $r = \Omega(q/(B^{2t}\log^4(n)))$.

Proof. We want to apply Theorem 2.6 to conclude that $H/\sqrt{q}$ satisfies RIP. To this end, all we need to show is that entries within any row of H are uncorrelated and have variance one.

For any pair of distinct tuples α, β, we have

$$\begin{aligned} \mathbb{E}[H_{i,\alpha}H_{i,\beta}] &= \mathbb{E}\left[\left(\prod_{j \in \alpha} \mathbf{X}_{i,j}\right)\left(\prod_{j \in \beta} \mathbf{X}_{i,j}\right)\right] \\ &= \prod_{j \in \alpha\Delta\beta} \mathbb{E}[\mathbf{X}_{i,j}] \\ &= 0\,, \end{aligned}$$

where $\alpha\Delta\beta$ is the symmetric difference between sets α and β (since for $j \in \alpha \cap \beta$ the term $\mathbf{X}_{i,j}^2$ appears in the monomial and $\mathbb{E}[\mathbf{X}_{i,j}^2] = 1$). Similarly $\mathbb{E}[H_{i,\alpha}^2] = \prod_{j \in \alpha} \mathbb{E}[\mathbf{X}_{i,j}^2] = 1$. Finally, since $|\mathbf{X}_{i,j}| \leq B$ for all j, we have $|H_{i,\alpha}| = \prod_{j \in \alpha} |\mathbf{X}_{i,j}| \leq B^t$. Therefore the conditions of Theorem 2.6 are satisfied for the matrix H, and the conclusion coincides with the lemma statement. □

We prove now the second claim.

Lemma 3.5. *Let $k \geq O(\sqrt{n}\log^2 n)$. With high probability, for any $j \notin S^*$, there exist at most $O(n^2/k^2)$ sets $\alpha \subset S^*$ with $|\alpha| = 3$ such that $\langle G_\alpha, G_j \rangle \geq k/2$.*

Proof. Let $H \in \{\pm 1\}^{(n-k) \times \Theta(k^3)}$ be the matrix with columns G_α^{out} for all $\alpha \subset S^*$ with $|\alpha| = 3$. By Lemma 3.4, the matrix $H/\sqrt{n-k}$ satisfies $(r, O(1))$-RIP for $r = \Omega((n-k)/\log^4(n))$. Then we apply Lemma 3.3 for $H/\sqrt{n-k}$ and $\tau = k/(3(n-k))$. Then no unit vector has correlation at least $\tau = k/(3(n-k))$ with more than $O(1/\tau^2) \leq O(n^2/k^2)$ columns of $H/\sqrt{n-k}$. Then also no vector in $\{\pm 1\}^{n-k}$ (of norm $\sqrt{n-k}$) has correlation at least $k/3$ with more than $O(n^2/k^2)$ columns of H. We need to ensure that the conditions of Lemma 3.3 are satisfied, namely that $r \gtrsim n^2/k^2$. This holds when $k \geq O(\sqrt{n}\log^2 n)$.

We can now finish the argument: If a vector G_j^{out} has correlation at most $k/3$ with G_α^{out}, then also $\langle G_\alpha, G_j \rangle \leq k/3 + O(\sqrt{k\log n}) < k/2$. Hence each vector G_j with $j \notin S^*$ has correlation greater than or equal to $k/2$ with at most $O(n^2/k^2)$ vectors G_α. □

We are now ready to prove the correctness of Algorithm 1.

Lemma 3.6. *Given a graph G generated according to* $\mathsf{SRPC}(n, k, 1/2)$ *for any $k \geq O(\sqrt{n}\log^2 n)$, Algorithm 1 outputs with probability at least* 0.99 *a list of size* $(1 + o(1))n/k$ *of k-cliques containing the planted clique in G.*

Proof. Consider a random $\alpha \subset S^*$ with $|\alpha| = 3$. Lemma 3.2 implies that with high probability $S_\alpha \supseteq S^*$. Furthermore, Lemma 3.5 says that, for each $j \notin S^*$, with high probability G_j has large inner product with at most $O(n^2/k^2)$ vectors G_α with $\alpha \subset S^*$ and $|\alpha| = 3$. Then G_j has large inner product with G_α for random $\alpha \subset S^*$ with probability at most $O((n^2/k^2)/k^3) = O(1/k)$, using that there are $\Omega(k^3)$ choices for α. Hence the expected number of G_j (for $j \notin S^*$) that have large inner product with G_α is at most $O(n/k)$. Therefore, by Markov's inequality, with probability at least 0.99 over the choice $\alpha \subset S^*$ we have that $|S_\alpha \setminus S^*| \leq O(n/k) = o(k)$.

Once this is the case, note that removing vertices of degree smaller than $k - 1$ in the induced graph given by S_α clearly does not affect vertices in S^*. On the other hand, for any $j \notin S^*$, denoting by $N(j)$ the neighborhood of the vertex j in G, we have

$$\begin{aligned} |N(j) \cap S_\alpha| &\leq |N(j) \cap S^*| + |N(j) \cap (S_\alpha \setminus S^*)| \\ &\leq k/2 + O(\sqrt{k \log n}) + o(k) \\ &< k - 1\,. \end{aligned}$$

Then the pruning algorithm described in Appendix B keeps S^* in the final list, while reducing the size of the list to $(1 + o(1))n/k$. □

We discuss now how, using fast matrix multiplication, Algorithm 1 can be implemented in time $O(n^{\omega+0.5})$.

Lemma 3.7. *Algorithm 1 can be equivalently implemented in time $O(Tn^\omega)$ where $T = \max((n/k)^3/n, 1)$ by invoking $O(T)$ calls to the $n \times n$ matrix-multiplication oracle. In particular, when $k \geq \sqrt{n}$, this yields $O(\sqrt{n})$ calls to the matrix multiplication oracle, and a total running time $O(n^{\omega+0.5})$.*

Proof. We can compute all inner products $\langle G_\alpha, G_j \rangle$ simultaneously for all $O((n/k)^3)$ sampled α: If we arrange the vectors G_α as the rows of a $O((n/k)^3) \times n$ matrix, and we multiply this matrix by the matrix $G \in \{\pm 1\}^{n \times n}$ — the ±1-adjacency matrix of the graph — then the (α, j)-th entry of the product is exactly $\langle G_\alpha, G_j \rangle$. This matrix multiplication can be computed in time $(O((n/k)^3)/n) \cdot n^\omega$, by partitioning the set of all sampled triples α into $m = O((n/k)^3)/n$ parts of size n, and applying the $n \times n$ matrix multiplication oracle for each part of the partition. This allows us to compute the sets S_α.

We can similarly use matrix multiplication to compute an $O((n/k)^3) \times n$ matrix M where $M_{\alpha,j}$ represents the number of vertices in S_α that vertex j is connected to — this is by multiplying a $\{0, 1\}$-valued matrix with rows being indicator vectors of sets S_α, by the $\{0, 1\}$-adjacency matrix of the graph G. Again, this can be done by using $O((n/k)^3/n)$ calls to the matrix multiplication oracle. This allows us to also obtain the sets S'_α.

Applying the exact same trick again, we can check which of the elements of S'_α are cliques — we now construct a matrix $M'_{\alpha,j}$ denoting the number of neighbors vertex j has in S'_α and keep only the sets S'_α such that for each $j \in S'_\alpha$ we have $M'_{\alpha,j} = |S'_\alpha| - 1$.

Therefore constructing the initial list involves $T = O((n/k)^3)/n \leq O(\sqrt{n})$ calls to the matrix multiplication oracle, and the size of the list is at most $O((n/k)^3)$.

Finally, by Lemma B.4 the pruning algorithm can be implemented by $O((n/k)^3/n)$ calls to the matrix multiplication oracle, leading to the overall time complexity bounded by $O(n^{\omega+0.5})$. □

The main theorem now follows immediately as a corollary of those two lemmas.

Proof of Theorem 3.1. The correctness of Algorithm 1 is given by Lemma 3.6 and the time complexity $O(n^{\omega+0.5})$ is given by Lemma 3.7. □

4 Semirandom planted clique with general p

In this section we generalize our algorithm for $\mathsf{SRPC}(n,k,p)$ to arbitrary $1 > p > 0$. Our bound on k is better than in previous results [MMT20, BKS23] when $p \le 1 - n^{-0.01}$. We note that the special case of $p = 1/2$ already contains all the necessary ideas.

Theorem 4.1. *There exists an $O(n^{\omega+0.5})$-time algorithm, where $\omega \le 2.372$ is the matrix multiplication exponent, that takes input a graph G generated according to $\mathsf{SRPC}(n,k,p)$ for any $k \ge O(\sqrt{n}\log^2 n \cdot \max(1, (p/(1-p))^4))$ and with probability at least 0.99 outputs a list of size $(1+o(1))n/k$ of k-cliques containing the planted clique in G.*

The first step is to consider a matrix $\bar{G}$ which is given by a construction analogous to p-biased characters: we want to shift and rescale the entries of the adjacency matrix so that in the random part the entries have mean zero and variance one. Specifically, let

$$\bar{G}(i,j) = \begin{cases} \sqrt{\frac{1-p}{p}} & \text{if } (i,j) \in E(G)\,, \\ -\sqrt{\frac{p}{1-p}} & \text{otherwise}\,. \end{cases}$$

An easy calculation confirms that indeed, if $\Pr((i,j) \in E(G)) = p$, then $\mathbb{E}[\bar{G}(i,j)] = 0$ and $\mathbb{E}[\bar{G}(i,j)^2] = 1$. To simplify notation in the following argument we introduce $p_+ = \sqrt{\frac{1-p}{p}}$, $p_- = \sqrt{\frac{p}{1-p}}$, and $B = \max(p_+, p_-)$ — so that all entries of $\bar{G}$ are bounded in absolute value by B (this property will turn out to be crucial later).

We also let $\bar{G}_i$ be the i-th row of $\bar{G}$, $\bar{G}_i^{\text{in}} \in \mathbb{R}^k$ be the projection of G_i to coordinates in S^*, and $\bar{G}_i^{\text{out}} \in \mathbb{R}^{n-k}$ the projection of $\bar{G}_i$ to coordinates in $[n] \setminus S^*$. Finally, for a subset of rows $\alpha \subset [n]$, we write $\bar{G}_\alpha$ to denote the element-wise product of rows $\bar{G}_j$ for $j \in \alpha$.

We prove the same sequence of results as in the $p = 1/2$ case: First, that for all $\alpha \subset S^*$ with $|\alpha| = 3$ we have $S_\alpha \supseteq S^*$, and second, that for any fixed $j \notin S^*$, there are at most $O(n^2/k^2)$ sets $\alpha \subset S^*$ with $|\alpha| = 3$ such that $j \in S_\alpha$. The first claim:

Lemma 4.2. *Let $k \ge O(\sqrt{n \log n} + B^4 \log n)/p_+^4$. With high probability, for all $\alpha \subset S^*$ with $|\alpha| = 3$ and all $j \in S^*$ with $j \notin \alpha$, we have that $\langle \bar{G}_\alpha, \bar{G}_j \rangle \ge kp_+^4/2$.*

Proof. We have that $\langle \bar{G}_\alpha^{\text{in}}, \bar{G}_j^{\text{in}} \rangle = kp_+^4$. On the other hand, the element-wise product of $\bar{G}_\alpha^{\text{out}}$ and $\bar{G}_j^{\text{out}}$ has independent mean-zero and variance-one entries with maximum absolute value B^4, so by Bernstein's inequality

$$\mathbb{P}\left(|\langle \bar{G}_\alpha^{\text{out}}, \bar{G}_j^{\text{out}} \rangle| \ge t\right) \le 2\exp\left(-\Omega(t^2/(n + tB^4))\right)\,.$$

Taking $t = O(\sqrt{n \log n} + B^4 \log n)$ large enough, we have that $|\langle \bar{G}_\alpha^{\text{out}}, \bar{G}_j^{\text{out}} \rangle| \le t$ with probability at least $1 - 1/n^{10}$. We choose k such that $t \le kp_+^4/2$, so we have $\langle \bar{G}_\alpha, \bar{G}_j \rangle \ge kp_+^4/2$. Finally, an application of the union bound shows that this holds simultaneously for all choices of α and j with probability at least $1 - o(1)$. □

Next, we want to prove that each $j \notin S^*$ satisfies $\langle G_\alpha, G_j \rangle \ge kp_+^4/2$ for few sets $\alpha \subset S^*$ with $|\alpha| = 3$. A similar application of Bernstein and union bound can be used to control the maximum of the quantity $\langle \bar{G}_\alpha^{\text{in}}, \bar{G}_j^{\text{in}} \rangle$ over all $\alpha \subset S^*$ with $|\alpha| = 3$ and $j \notin S^*$.

Lemma 4.3. *With high probability, for all $\alpha \subset S^*$ with $|\alpha| = 3$ and all $j \notin S^*$ we have that*

$$|\langle \bar{G}_\alpha^{\text{in}}, \bar{G}_j^{\text{in}} \rangle| \le O(\sqrt{k \log n} + B^4 \log n)\,.$$

Proof. The element-wise product of $\bar{G}_\alpha^{\text{in}}$ and G_j^{in} has independent mean-zero and variance-one entries with maximum absolute value B^4. The bound follows by Bernstein's inequality and a union bound as in Lemma 4.2. □

In order to ensure that, as in the $p = 1/2$ case, for each element $j \notin S^*$ we can bound the number of sets $\alpha \subset S^*$ that lead to large $\langle \bar{G}_\alpha^{\text{out}}, \bar{G}_j^{\text{out}} \rangle$, we want to appeal again to Lemma 3.3. To this end, we define the matrix $H \in \mathbb{R}^{(n-k)\times\Theta(k^3)}$ with columns G_α^{out} for all $\alpha \subset S^*$ with $|\alpha| = 3$.

As it turns out, this matrix H also satisfies the conditions of Lemma 3.4, with a bound B on the size of the entries that generate the matrix. Then, we can conclude:

Lemma 4.4. *Let $k \ge O(p_+^{-4} B^4 \sqrt{n} \log^2 n)$. With high probability, for any $j \notin S^*$, there exist at most $O(p_+^{-8} B^2 n^2/k^2)$ sets $\alpha \subset S^*$ with $|\alpha| = 3$ such that $|\langle G_\alpha, G_j \rangle| \ge kp_+^4/2$.*

Proof. Let $H \in \mathbb{R}^{(n-k)\times\Theta(k^3)}$ be the matrix with columns G_α^{out} for all $\alpha \subset S^*$ with $|\alpha| = 3$. By Lemma 3.4, the matrix $H/\sqrt{n-k}$ satisfies $(r, O(1))$-RIP for $r = \Omega((n-k)/(B^6 \log^4(n)))$. Then we apply Lemma 3.3 for $H/\sqrt{n-k}$ and $\tau = kp_+^4/(3B(n-k))$. Then no unit vector has correlation at least $\tau = kp_+^4/(3B(n-k))$ with more than $O(1/\tau^2) \le O(p_+^{-8} B^2 n^2/k^2)$ columns of $H/\sqrt{n-k}$. Because the norm of each G_j^{out} is bounded by $B\sqrt{n-k}$, we get that no vector G_j^{out} has correlation at least $kp_+^4/3$ with more than $O(p_+^{-8} B^2 n^2/k^2)$ columns of H. We need to ensure that the conditions of Lemma 3.3 are satisfied, namely that $r \gtrsim p_+^{-8} B^2 n^2/k^2$. This holds when $k \ge O(p_+^{-4} B^4 \sqrt{n} \log^2 n)$.

We can now finish the argument: If a vector G_j^{out} has correlation at most $kp_+^4/3$ with G_α^{out}, then also $\langle G_\alpha, G_j \rangle \le kp_+^4/3 + O(\sqrt{k \log n} + B^4 \log n) < kp_+^4/2$. Hence each vector G_j with $j \notin S^*$ has correlation greater than or equal to $kp_+^4/2$ with at most $O(p_+^{-8} B^2 n^2/k^2)$ vectors G_α. □

We are now ready to finish the proof of the theorem.

Proof of Theorem 4.1. We use an identical algorithm as in the $p = 1/2$ case. For $O((n/k)^3)$ turns, we take a random triple of vertices $\alpha \subset [n]$ with $|\alpha| = 3$ and construct a set $S_\alpha := \{j : \langle \bar{G}_\alpha, \bar{G}_j \rangle \ge kp_+^4/2\}$. Then we construct a refined set $S'_\alpha \subseteq S_\alpha$ by removing all vertices from S_α that are connected to less than $k - 1$ of the vertices in S_α, and then add S'_α to the list if it forms a k-clique. Finally, we apply pruning (see Appendix B) to reduce the list size to $(1 + o(1))n/k$.

To conclude that this algorithm works, we want to argue that conditioned on $\alpha \subset S^*$, the procedure succeeds in recovering S^* with probability at least $1/2$.

Indeed, by Lemma 4.2, once $\alpha \subset S^*$ we have with high probability $S^* \subseteq S_\alpha$. Moreover, for any $j \notin S^*$, by Lemma 4.4, $\bar{G}_j$ has large inner product with at most $T = O(p_+^{-8} B^2 n^2/k^2)$ vectors $\bar{G}_\alpha$ over $\alpha \subset S^*$ with $|\alpha| = 3$. There are $\Omega(k^3)$ such sets α, so the probability that $\bar{G}_j$ has large inner product with a $\bar{G}_\alpha$ chosen at random is at most $O(T/k^3)$. Then, the expected number of $\bar{G}_j$ that have large inner product with a vector $\bar{G}_\alpha$ for a random $\alpha \subset S^*$ with $|\alpha| = 3$ is $O(nT/k^3)$. Hence, by Markov's inequality, with probability 0.99 over the choice of random $\alpha \subset S^*$ we have that $|S_\alpha \setminus S^*| \leq O(nT/k^3) \leq O(p_+^{-8} B^2 n^3/k^5) \leq o(k(1-p))$.

Once this is the case, note that removing vertices of degree smaller than $k-1$ in the induced graph given by S_α clearly does not affect vertices in S^*. On the other hand, for any $j \notin S^*$, denoting by $N(j)$ the neighborhood of the vertex j in G, we have

$$\begin{aligned}|N(j) \cap S_\alpha| &\leq |N(j) \cap S^*| + |N(j) \cap (S_\alpha \setminus S^*)| \\ &\leq kp + O\left(\sqrt{kp(1-p)\log n}\right) + o(k(1-p)) \\ &< k-1\,.\end{aligned}$$

Then the pruning algorithm in Appendix B keeps S^* in the final list.

The time complexity is the same as in the $p = 1/2$ case, using matrix multiplication. □

5 Discussion

We include some commentary and open questions that naturally arise given our new approach.

1) Handling a monotone adversary. In this work, we gave a simple algorithm that solves the semirandom planted clique recovery problem at nearly the right threshold of $k = \tilde{O}(\sqrt{n})$. Unlike the prior works [BKS23, MMT20] (we note that the algorithm in [CSV17, Ste17] does not tolerate monotone deletions), our algorithm does not handle a monotone adversary that can delete an arbitrary subset of edges in the cut defined by the planted clique.

In prior works, approaches that handle a monotone adversary [FK00, MPW16] have been naturally based on semidefinite programming. In fact, the analyses of such semidefinite programs can be naturally interpreted as yielding an efficient certificate of some natural property that controls the uniqueness of the relevant solution concept (e.g., clique, community, coloring). In our context, such a certificate of uniqueness naturally corresponds to the biclique numbers for random bipartite graphs studied in [BKS23]. Indeed, they provide some evidence of non-existence of efficient certificates for this property by means of lower bounds in the low-degree polynomial model. Our key conceptual contribution is finding an algorithmic approach that circumvents the need for such certificates.

We are thus left with an intriguing question: is there an algorithm for semirandom planted clique that succeeds at $k = \tilde{O}(\sqrt{n})$ *and* tolerates a monotone adversary? Can such an algorithm yield an efficient (e.g., low-degree sum-of-squares) certificate for biclique bounds on random bipartite graphs and circumvent the low-degree polynomial lower bounds?

On the flip side, could there be a computational–statistical complexity gap for semirandom planted clique that arises entirely due to a monotone adversary? This would run counter to the central thesis of [Ste17] that suggests that robustness (i.e., success in the semirandom model) might obliterate the computational–statistical complexity gap for the fully random planted clique problem. We *do not* know of any problem for which a monotone adversary "creates" such a gap. We note that for community detection in stochastic block models, the work of Moitra, Perry and Wein [MPW16] showed that monotone deletions shift the *information-theoretic* threshold.

2) Can our algorithm be captured by semidefinite programming? In light of the discussion above, it is also natural to ask: can our algorithm be "captured" (i.e., suggest an analysis with the same guarantees) by a natural semidefinite programming relaxation for the semirandom planted clique problem? We do not know how to do this so far. As discussed above, analyses of semidefinite programming relaxations usually yield certificates for an underlying property (biclique number bounds for random bipartite graphs in our setting) that implies uniqueness of the relevant solution concept. We are aware of two exceptions to this "rule" where algorithms based on SDPs succeed despite the potential impossibility of the associated certification problem: 1) the SDP relaxations for low-rank matrix sensing [RFP10] without a certificate of RIP, 2) robust mean and covariance estimation for Gaussian distributions [KMZ22] without a certificate of "resilience". Is there a similar approach that manages to recover our guarantees for semirandom planted clique without the need for improved certificates for biclique numbers in random bipartite graphs?

3) Achieving $k = O(\sqrt{n \log n})$? Our algorithm currently needs $k \geq O(\sqrt{n}\log^2 n)$, which is $\log^{1.5}(n)$ off of the likely "right" bound (i.e., matching the guarantee of the best-known brute-force algorithm). This $\log^{1.5}(n)$-factor loss arises from the loss in the sparsity parameter in Lemma 3.4. We believe that for our matrix H this loss can likely be removed and consequently our algorithm could succeed at $k \geq O(\sqrt{n \log n})$. We note that the general result (Theorem 2.6) we rely on cannot be improved to provide such a guarantee: there exists an isotropic distribution over bounded vectors from which it is necessary to sample at least $\Omega(\sqrt{n}\log^2 n)$ rows in order to obtain RIP at sparsity $\sqrt{n}$ [BLL+23]. However, our application only needs the weaker *sparse operator norm* bounds to which the lower bounds in [BLL+23] do not apply. While this gap has not been consequential in known results, it does leave open the possibility of a stronger, general variant of Theorem 2.6 for sparse operator norm bounds.

Acknowledgements

JB, RB, and DS received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 815464). PK was supported by NSF CAREER Award #2047933, NSF #2211971, an Alfred P. Sloan Fellowship, and a Google Research Scholar Award.

References

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. In *Proceedings of the Eighth International Conference "Random Structures and Algorithms" (Poznan, 1997)*, volume 13, pages 457–466, 1998.

[BB20] Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.

[BBH18] Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Conference On Learning Theory*, pages 48–166. PMLR, 2018.

[BDMS13] Afonso S. Bandeira, Edgar Dobriban, Dustin G. Mixon, and William F. Sawin. Certifying the restricted isometry property is hard. *IEEE Trans. Information Theory*, 59(6):3448–3450, 2013.

[BHK$^+$16] Boaz Barak, Samuel B. Hopkins, Jonathan Kelner, Pravesh K. Kothari, Ankur Moitra, and Aaron Potechin. A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, 2016.

[BKS23] Rares-Darius Buhai, Pravesh K. Kothari, and David Steurer. Algorithms approaching the threshold for semi-random planted clique. In *STOC'23—Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1918–1926. ACM, New York, [2023] ©2023.

[BLL$^+$23] Jaroslaw Blasiok, Kyle Luh, Patrick Lopatto, Jake Marcinek, and Shravas Rao. An improved lower bound for sparse reconstruction from subsampled walsh matrices. *Discrete Analysis*, 5 2023.

[BR13a] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 1046–1066. JMLR.org, 2013.

[BR13b] Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828*, 2013.

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

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *STOC*, pages 47–60. ACM, 2017.

[DKWB21] Yunzi Ding, Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. The average-case time complexity of certifying the restricted isometry property. *IEEE Trans. Inform. Theory*, 67(11):7355–7361, 2021.

[Fei19] Uriel Feige. Introduction to semirandom models. In Tim Roughgarden, editor, *Beyond Worst-case Analysis of Algorithms*, chapter 10, pages 266–290. Oxford, 2019.

[FGR+17] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S. Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM*, 64(2):Art. 8, 37, 2017.

[FK00] Uriel Feige and Robert Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Struct. Algorithms*, 16(2):195–208, 2000.

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

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

[Hås99] Johan Håstad. Clique is hard to approximate within $n^{1-\varepsilon}$. *Acta Math.*, 182(1):105–142, 1999.

[HKP+17] Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

[Jer92] Mark Jerrum. Large cliques elude the metropolis process. *Random Struct. Algorithms*, 3(4):347–360, 1992.

[KMZ22] Pravesh K Kothari, Peter Manohar, and Brian Hu Zhang. Polynomial-time sum-of-squares can robustly estimate mean and covariance of gaussians optimally. In *International Conference on Algorithmic Learning Theory*, pages 638–667. PMLR, 2022.

[Kuc95] Ludek Kucera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57(2-3):193–212, 1995.

[KZ14a] Pascal Koiran and Anastasios Zouzias. Hidden cliques and the certification of the restricted isometry property. *IEEE Trans. Information Theory*, 60(8):4999–5006, 2014.

[KZ14b] Pascal Koiran and Anastasios Zouzias. Hidden cliques and the certification of the restricted isometry property. *IEEE Trans. Inform. Theory*, 60(8):4999–5006, 2014.

[Luk69] Yudell L. Luke. Chapter II. The Gamma Function and Related Functions. In *The Special Functions and Their Approximations*, volume 53 of *Mathematics in Science and Engineering*, pages 8–37. Elsevier, 1969.

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

[MPW16] Ankur Moitra, William Perry, and Alexander S. Wein. How robust are reconstruction thresholds for community detection? In *STOC*, pages 828–841. ACM, 2016.

[Rau10] Holger Rauhut. Compressive sensing and structured random matrices. In *Theoretical foundations and numerical methods for sparse recovery*, volume 9 of *Radon Ser. Comput. Appl. Math.*, pages 1–92. Walter de Gruyter, Berlin, 2010.

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[RV08] Mark Rudelson and Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[Ste17] Jacob Steinhardt. Does robustness imply tractability? a lower bound for planted clique in the semi-random model. *arXiv preprint arXiv:1704.05120*, 2017.

[Zuc07] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory Comput.*, 3:103–128, 2007.

A Restricted isometry property

One of the main tools used in the analysis of our algorithm is a strong general theorem guaranteeing RIP for any matrix with independent rows drawn from any isotropic-distribution vectors with bounded entries.

Theorem 2.6, although known, is usually stated in terms of so-called *bounded orthonormal systems*, a formulation that makes its direct applicability in situations like Lemma 3.4 less transparent. For completeness, we introduce the notion of bounded orthonormal system, the formulation of Theorem 2.6 as it is stated in [Rau10], and we discuss why this is equivalent to the statement we are using in this paper.

Definition A.1. A bounded orthonormal system is given by a region $\mathcal{D} \subset \mathbb{R}^k$, together with a probability measure μ on $\mathcal{D}$ and N functions $\phi_1, \phi_2, \ldots \phi_d : \mathcal{D} \to \mathbb{C}$ satisfying the following properties:

- For all $i \neq j$ in $[N]$, $\mathbb{E}_{t\sim\mu}\phi_i(t)\overline{\phi_j(t)} = 0$,
- For all $i \in [N]$, $\mathbb{E}_{t\sim\mu}\phi_i(t) = 0, \mathbb{E}_{t\sim\mu}|\phi_i(t)|^2 = 1$,
- For all $i \in [N]$, $|\phi_i(t)| \leq K$ almost surely.

Then Theorem 12.31 in [FR13] (originally proved in [RV08]) states:

Theorem A.2 (Theorem 12.31 in [FR13])**.** *If $\phi_1, \ldots, \phi_N$ together with a distribution μ is a bounded orthonormal system, then a matrix $H \in \mathbb{C}^{m\times N}$ obtained by sampling $t_1, t_2 \ldots t_m$ independently at random according to μ and setting $H_{i,j} = \phi_j(t_i)/\sqrt{m}$ satisfies (r, δ)-RIP with probability $1 - N^{-\log^3(r)}$ whenever*

$$m \gtrsim K^2 r \log^3 r \log(N)/\delta^2 .$$

To see how this theorem implies Theorem 2.6, let us consider an arbitrary distribution μ over $\mathbb{R}^n$ satisfying $\mathbb{E}_{\mathbf{X}\sim\mu}\mathbf{X}\mathbf{X}^\top = I$. We can take $\phi_1, \dots \phi_N : \mathbb{R}^n \to \mathbb{R}$ such that ϕ_j is a projection on the j-th coordinate. Then $\phi_1, \dots \phi_N$ together with the distribution μ forms a bounded orthonormal system according to Definition A.1, and the matrix $H_{i,j} = \phi_j(t_i)$ is exactly obtained by drawing m independent random rows according to the law of $\mathbf{X}$.

The other direction of the equivalence (in the case of real-valued bounded orthonormal system) is just as immediate: if $\phi_1, \dots, \phi_N$ together with μ form a bounded orthonormal system, then a random vector $\mathbf{X}$ obtained by drawing $t \sim \mu$ and setting $\mathbf{X} = (\phi_1(t), \dots \phi_N(t))$ is a bounded vector in isotropic position.

B Pruning a list of k-cliques

We start by recalling the following two lemmas from [BKS23]:

Lemma B.1 (Lemma 5.7 in [BKS23]). *Let $G \sim \mathsf{SRPC}(n, k, p)$. Let S^* be the planted clique in G. Then, with probability at least $1 - \frac{k}{n^2}$, any other clique S of size at least k satisfies $|S \cap S^*| \leq 3\frac{\log n}{\log 1/p}$.*

Lemma B.2 (Lemma 5.8 in [BKS23]). *Let $S_1, ..., S_m \subseteq [n]$ with $|S_i| = k$ and $|S_i \cap S_j| \leq \Delta$. Then, if $k \geq \sqrt{2n\Delta}$, we have $m \leq \frac{n}{k}\left(1 + \frac{2n\Delta}{k^2}\right)$.*

Now we prove the pruning result:

Lemma B.3 (Pruning, implicit in [BKS23]). *Let G be a graph on n vertices generated according to $\mathsf{SRPC}(n, k, p)$ where $k \geq C\sqrt{n \log n \cdot \max(1, p/(1-p))}$ for some large enough absolute constant C. Let L be a list of size m of k-cliques of G.*

Then taking L' to be the list of all cliques from L which have intersection at most $3\log n/\log p^{-1}$ with all other cliques in L, we have with high probability

- *If $S^* \in L$, then also $S^* \in L'$, and*
- *$|L'| \leq (1 + O(1/C^2))n/k$.*

Proof. First, we have from Lemma B.1 that with high probability the planted clique has intersection at most $3\frac{\log n}{\log 1/p}$ with any other clique in G, so we are guaranteed that we do not remove it from the list.

Second, all the k-cliques in the final list have intersection at most $3\frac{\log n}{\log 1/p}$, so by Lemma B.2 the size of the list is at most $\frac{n}{k}\left(1 + \frac{6n\frac{\log n}{\log 1/p}}{k^2}\right)$. It remains to show that $\frac{6n\frac{\log n}{\log 1/p}}{k^2} = O(1/C^2)$ for our choice of k. We use that $\log 1/p \geq (1-p)/4$ for $p \leq 1$ to get

$$\begin{aligned}\frac{6n\frac{\log n}{\log 1/p}}{k^2} &\leq \frac{24n\frac{\log n}{1-p}}{C^2 n \log n \cdot \max(1, p/(1-p))} \\ &= \frac{24}{C^2(1-p)\max(1, p/(1-p))}\end{aligned}$$

$$= \frac{24}{C^2 \max(1-p,p)} \leq \frac{48}{C^2},$$

which gives the desired result. □

Lemma B.4 (Fast pruning)**.** *The pruning procedure described in Lemma B.3 can be implemented in time $O(\lceil m/n \rceil n^{\omega} \frac{\log m}{\log k})$ by using $O(\lceil m/n \rceil \cdot \frac{\log m}{\log k})$ calls to the $n \times n$ matrix multiplication oracle.*

Proof. We implement the algorithm using matrix multiplication as follows:

- While $m > (n/k)(1 + o(1))$:
  - Split the list of m k-cliques arbitrarily into $\lceil m/n \rceil$ lists of at most n k-cliques,
  - For each of the $\lceil m/n \rceil$ lists of at most n k-cliques:
    - Compute $U \in \{0,1\}^{n \times n}$ such that $U_{i,j} = 1$ if and only if the i-th clique in the list contains vertex j,
    - Compute $M = UU^{\top}$ in time $O(n^{\omega})$,
    - For each clique i in the list, iterate over all cliques $j > i$ in the list, and if $M_{i,j} > 3\frac{\log n}{\log 1/p}$, mark clique j as removed,
  - Construct a list of all cliques that are not removed in any of the $\lceil m/n \rceil$ lists. By Lemma B.3, this list has size at most $\lceil m/n \rceil \cdot (n/k)(1 + o(1)) \leq (m/k)(1 + o(1))$. Update m to be the size of this new list.

The while loop performs at most $\log_k m = \frac{\log m}{\log k}$ iterations, and the inner loop has time complexity $O(\lceil m/n \rceil n^{\omega} + mn) \leq O(\lceil m/n \rceil n^{\omega})$. Therefore the total time complexity is $O(\lceil m/n \rceil n^{\omega} \frac{\log m}{\log k})$. □

C Adversarial correlation with independent Rademacher vectors

Lemma C.1. *Let $v_1, ..., v_m \in \{\pm 1\}^n$ be independent Rademacher vectors. Then with high probability there exists some vector $w \in \{\pm 1\}^n$ such that $\langle v_i, w \rangle \geq \Omega(n/\sqrt{m})$ for all $i \in [m]$ as long as $n \gg m \log m$.*

In other words, for any $k = \Omega(n/\sqrt{m})$ small enough (i.e., $m = \Omega(n^2/k^2)$ small enough), with high probability there exists some vector $w \in \{\pm 1\}^n$ such that $\langle v_i, w \rangle \geq k$ for all $i \in [m]$ as long as $n \gg m \log m$.

Proof. We consider without loss of generality the case when m is odd. Let $w_j = \mathrm{Maj}(v_{1,j}, \ldots, v_{m,j})$ for all $j \in [n]$, i.e., the majority element. Note that $\langle v_i, w \rangle = 2\sum_{j=1}^{n} \mathbf{1}_{w_j = v_{i,j}} - n$, where $\mathbf{1}_{w_j = v_{i,j}}$ are independent $\mathrm{Ber}(p)$ random variables, with p the probability that $v_{i,j} = \mathrm{Maj}(v_{1,j}, \ldots, v_{m,j})$. Note that if $v_{1,j}, \ldots, v_{i-1,j}, v_{i+1,j}, \ldots, v_{m,j}$ already forms a majority of $\geq \frac{m+1}{2}$ elements, then $v_{i,j}$ cannot influence the majority and it will be in the majority with probability 1/2. Otherwise, $v_{1,j}, \ldots, v_{i-1,j}, v_{i+1,j}, \ldots, v_{m,j}$ must have an equal number of 1 and -1 elements, and $v_{i,j}$ will always be in the majority. The latter case happens with probability $2\binom{m-1}{(m-1)/2}/2^{m-1}$, which grows asymptotically as $\frac{2}{\sqrt{\pi m/2}}(1 \pm o(1))$ (e.g., see the growth of the central binomial coefficient in [Luk69], page 35). Therefore $p = 1/2 + \Omega(1/\sqrt{m})$. From this, by Hoeffding's inequality, we get that $\mathbb{P}(\langle v_i, w\rangle - \Omega(n/\sqrt{m}) \geq t) \leq \exp(-\Omega(t^2/n))$, so with high probability $\langle w, v_i\rangle \geq \Omega(n/\sqrt{m}) - O(\sqrt{n \log m})$ for all $v_1, \ldots, v_m$. The first term dominates as long as $n \gg m \log m$.

□