2012 IEEE 53rd Annual Symposium on Foundations of Computer Science

Hardness of Finding Independent Sets in Almost q-Colorable Graphs

Subhash Khot*
New York Univ., NY and Univ. of Chicago, IL USA
Email: khot@cs.nyu.edu

Rishi Saket
IBM T. J. Watson Research Center
Yorktown Heights, NY, USA
Email: rsaket@us.ibm.com

***Abstract*—We show that for any $\varepsilon > 0$, and positive integers k and q such that $q \geq 2^k + 1$, given a graph on N vertices that has a q-colorable induced subgraph of $(1-\varepsilon)N$ vertices, it is NP-hard to find an independent set of $\frac{N}{q^{k+1}}$ vertices. This substantially improves upon the work of Dinur et al. [1] who gave a corresponding bound of $\frac{N}{q^2}$.**

Our result implies that for any positive integer k, given a graph that has an independent set of $\approx (2^k+1)^{-1}$ fraction of vertices, it is NP-hard to find an independent set of $(2^k+1)^{-(k+1)}$ fraction of vertices. This improves on the previous work of Engebretsen and Holmerin [2] who proved a gap of $\approx 2^{-k}$ vs $2^{-\binom{k}{2}}$, which is best possible using techniques (including those of [2]) based on the query efficient PCP of Samorodnitsky and Trevisan [3].

***Keywords*-Coloring; Graphs; Independent-Set; Hardness; PCP;**

I. Introduction

Given a graph, an independent set is a subset of the vertices that does not contain both end points of any edge. Computing a maximum sized independent set is a very well studied problem in computer science and combinatorics. A related problem is that of graph coloring, where the goal is to color the vertices of a given graph using minimum number of colors such that no edge has both end points of the same color. It is easy to see that if a graph is colorable by q colors then it must contain an independent set of q^{-1} fraction of the vertices. Conversely, if a graph does not contain an independent set of q^{-1} fraction of vertices, then it cannot be colored using q colors. Thus, the absence of a large independent set certifies that the *chromatic number* of the graph – the minimum number of colors required to color the graph – is large.

Another closely related problem is that of computing the minimum vertex cover, i.e. a subset of vertices of minimum size that contains at least one end point of every edge in the graph. The complement of vertex cover is an independent set. Therefore, as in the case of coloring, the absence of a large independent set guarantees that every vertex cover is large. The best known inapproximability result for vertex cover is due to Dinur and Safra [4] who showed that it is NP-hard to approximate it within a factor of 1.36. The result of [4] showed in particular that it is NP-hard to decide whether a graph of N vertices has an independent set of size $\approx \frac{N}{3}$ or every independent set is of size at most $\approx \frac{N}{9}$.

Building upon the work of Dinur and Safra [4], Dinur, Khot, Perkins and Safra [1] proved that: for any positive integer $q \geq 3$ and small constant $\varepsilon > 0$, given a graph, it is NP-hard to decide whether (i) the graph contains an induced subgraph of $(1-\varepsilon)$-fraction of the vertices which is q-colorable, where each color class has $(1-\varepsilon)\frac{1}{q}$ fraction of the vertices in the graph, or (ii) the maximum independent set in the graph contains less than $\frac{1}{q^2}$ fraction of the vertices. In other words, it is NP-hard to find an independent set of $\frac{1}{q^2}$ fraction of the vertices in *almost q-colorable* graphs.

In this work, we generalize and substantially improve upon the result of Dinur, Khot, Perkins and Safra [1]. We show that for any positive positive integer k, any integer q such that $q \geq 2^k+1$, and an arbitrarily small constant $\varepsilon > 0$, given a graph, it is NP-hard to decide whether,

- The graph contains a q-colorable induced subgraph of $(1-\varepsilon)$-fraction of vertices, where each color class has $(1-\varepsilon)\frac{1}{q}$ fraction of vertices.
- Every independent set in the graph has less than $\frac{1}{q^{k+1}}$ fraction of the vertices.

The reduction used to prove this result builds upon the ideas of [4] and [1] and combines them with a new *outer verifier* based on the query-efficient *Probabilistically Checkable Proof* (PCP) system of Håstad and Khot [5]. We elaborate more on these techniques later in this section.

Note that for the setting of parameters $k = 1$ and $3 \leq q \leq 4$, our result is the same as that of [1], i.e. it is NP-hard to distinguish between an almost q-colorable graph and a graph where the maximum independent set contains at most $\frac{1}{q^2}$ fraction of vertices. For larger values of k and q, our result yields substantially better hardness factors. For example, setting $k = 2$ and $q = 2^2 + 1 = 5$ shows that it is NP-hard to find an independent set of $\frac{1}{125}$ fraction of the vertices in an almost 5-colorable graph. Setting $k = 3$ and $q = 2^3 + 1 = 9$ gives us that it is NP-hard to find an independent set of $\frac{1}{6561}$ fraction of vertices in an almost 9-colorable graph.

In terms of hardness of approximating maximum independent set, our result shows that: for any positive integer k

*Research supported by NSF CAREER grant CCF-0833228, NSF Expeditions grant CCF-0832795, NSF Waterman Award and BSF grant 2008059.

0272-5428/12 $26.00 © 2012 IEEE
DOI 10.1109/FOCS.2012.75

it is NP-hard to find an independent set of $(2^k+1)^{-(k+1)}$ fraction of vertices in a graph which has an independent set of $\approx (2^k+1)^{-1}$ fraction of vertices. This improves upon previous work of Engebretsen and Holmerin [2] who showed that, for any integer $k \geq 2$ it is NP-hard to compute an independent set of $\approx 2^{-\binom{k}{2}}$ fraction in a graph which has an independent set of $\approx 2^{-k}$ fraction of vertices. The result of Engebretsen and Holmerin [2] extended the work of Samorodnitsky and Trevisan [3] who gave a gap of $\approx 2^{-k}$ versus $2^{-k^2/4}$ (say for even k). The latter paper also demonstrated why $2^{-\binom{k}{2}}$ is a natural bottleneck for these proof techniques. Roughly, the reason is that there exist functions $f : \{0,1\}^n \mapsto \{0,1\}$ that have no non-negligible Fourier coefficients, but still pass the *full hyper-graph linearity test* with probability $2^{-\binom{k}{2}}$, i.e. with this much probability $f(\bigoplus_{i\in S} x_i) = \bigoplus_{i\in S} f(x_i),\ \forall S \subseteq \{1,2,\ldots,k\}$ for a random choice of $x_1,\ldots,x_k \in \{0,1\}^n$. On the other hand, as soon as the probability exceeds $2^{-\binom{k}{2}}$, the function must have a non-negligible Fourier coefficient, even when the test is carried out only for all $\binom{k}{2}$ sets $S, |S| = 2$. We find it quite interesting that this bottleneck is bypassed in our result via completely different techniques based on [4], [1].

It is pertinent to note that assuming the *Unique Games Conjecture* (UGC) of Khot [6], Bansal and Khot [7], [8] showed that for any constant $\delta > 0$ it is NP-hard to find an independent set of δ fraction of vertices in almost 2-colorable graphs. This bound is significantly stronger than those obtained unconditionally, including the results in this paper. However, since UGC remains unresolved we believe that our results – in addition to proving stronger unconditional lower bounds – shed new light on the applicability of some important techniques in PCP theory, especially on the methods developed in the work of Dinur and Safra [4]. In the rest of this section we shall define the problems studied and formally state our results. We shall also review the related previous work and informally describe the techniques used in this work.

A. Problem Definition

We begin by defining a decision problem for the size of the maximum independent set as follows.

INDEPENDENTSET(c,s): Given a graph $G(V,E)$, decide between the following cases.

- YES Case: $\mathsf{IS}(G) \geq c|V|$.
- NO Case: $\mathsf{IS}(G) < s|V|$.

where $\mathsf{IS}(G)$ is the size of the maximum independent set in G.

Given a graph G, let $\chi(G)$ be its chromatic number, i.e. the minimum number of colors required to color the graph such that every edge has distinctly colored end points. The graph coloring problem is defined as:

COLORING(q,Q) : Given a graph $G(V,E)$, decide between,

- YES Case: $\chi(G) \leq q$.
- NO Case: $\chi(G) \geq Q$.

It is easy to see that if COLORING(q,Q) is NP-hard for some parameters $q, Q \in \mathbb{Z}^+$ then it is NP-hard to color a q-colorable graph with $Q-1$ colors. In this paper we study a slight variant of graph coloring, which we refer to as *almost coloring*, which is defined, for positive integers q, Q and a constant $\varepsilon > 0$, as follows.

ALMOSTCOLORING$_\varepsilon(q,Q)$: Given a graph $G(V,E)$, decide between,

- YES Case: There is a subset of $(1-\varepsilon)$ fraction of the vertices, such that the graph G' induced by it satisfies, $\chi(G') \leq q$. We also denote this by $\chi_\varepsilon(G) \leq q$.
- NO Case: $\mathsf{IS}(G) < \frac{|V|}{Q}$.

Note that the second property above, i.e. $\mathsf{IS}(G) < \frac{|V|}{Q}$, implies that $\chi_\varepsilon(G) \geq Q$ for sufficiently small $\varepsilon > 0$.

B. Our Results

The main theorem that we prove is stated below.

Theorem I.1. *For any constant $\varepsilon > 0$, and positive integers k and q such that $q \geq 2^k + 1$, given a graph $G(V,E)$, it is NP-hard to distinguish between the following two cases:*

- **YES Case:** *There are q disjoint independent sets $V_1,\ldots,V_q \subseteq V$, such that $|V_i| = \frac{(1-\varepsilon)}{q}|V|$ for $i = 1,\ldots,q$.*
- **NO Case:** *There is no independent set in G of size $\frac{1}{q^{k+1}}|V|$.*

The following theorems follow directly from Theorem I.1.

Theorem I.2. *For any constant $\varepsilon > 0$, a positive integer k and integer q such that $q \geq 2^k + 1$,* ALMOSTCOLORING$_\varepsilon(q, q^{k+1})$ *is NP-hard.*

Theorem I.3. *For any constant $\varepsilon > 0$, a positive integer k and integer q such that $q \geq 2^k + 1$,* INDEPENDENTSET$((1-\varepsilon)\frac{1}{q}, \frac{1}{q^{k+1}})$ *is NP-hard.*

C. Previous Work

In the results stated in this subsection, $\varepsilon > 0$ shall be taken to be an arbitrarily small constant. The independent set and graph coloring problems are NP-hard in general and much of the research on these problems has focused on understanding their approximability. For approximating maximum independent set, the best algorithm is due to Feige [9] who gave a $O\left(\frac{n(\log\log n)^2}{\log^3 n}\right)$-approximation for it, where n is the number of vertices in the graph. On the other hand, Håstad [10] gave a $n^{1-\varepsilon}$ hardness factor for maximum independent set which was improved by Engebretsen and Holmerin [11] to $n^{1-O(1/\sqrt{\log\log n})}$ and by Khot [12] to $n/2^{(\log n)^{1-\gamma}}$ for some fixed $\gamma > 0$. The current best inapproximability is by Khot and Ponnuswami [13], who showed that INDEPENDENTSET(c,s) is

NP-hard, assuming NP$\not\subseteq$ DTIME($2^{\text{poly}(\log n)}$), where $c/s \geq n/2^{(\log n)^{3/4+\varepsilon}}$. However, the parameter c is a subconstant, i.e. $c = o(1)$ in the result of [13] (as also in [10], [11] and [12]). For constant values of c, Khot and Regev [14] and subsequently Bansal and Khot [7], [8] showed, assuming UGC, that INDEPENDENTSET($\frac{1}{2} - \varepsilon, \varepsilon$) is NP-hard. Unconditionally however, the previous best result showing that INDEPENDENTSET($(1-\varepsilon)2^{-k}, (1+\varepsilon)2^{-\binom{k}{2}}$) is NP-hard for any integer $k \geq 2$ was proved by Engebretsen and Holmerin [2] (after applying the FGLSS reduction [15]). They extended the query efficient PCP of Samorodnitsky and Trevisan [3] who, as mentioned before, also demonstrated why $2^{-\binom{k}{2}}$ is a natural bottleneck for these techniques. In our work, we bypass this bottleneck by proving in Theorem I.3 that INDEPENDENTSET($(1-\varepsilon)(2^k+1)^{-1}, (2^k+1)^{-(k+1)}$) is NP-hard for any positive integer k.

The related problem of COLORING(q, Q) has also been well studied for various ranges of the parameters q, Q. COLORING$(2, Q)$ can be solved efficiently by checking whether a graph is bipartite. For $q = 3$, a long line of research ([16], [17], [18], [19], [20]) shows that COLORING$(3, n^\alpha)$ can be solved where the best value of $\alpha \approx 0.207$. For general values of q, Halperin, Nathaniel and Zwick [21] solve COLORING(q, n^{α_q}) for some constant $\alpha_q \in (0,1)$ depending on q. These algorithmic results also hold for ALMOSTCOLORING$_\varepsilon(q, Q)$, for small enough values of ε.

The hardness of approximation results for graph coloring are quite far from matching the algorithmic upper bounds. For $q = 3$ the best hardness result shows that COLORING(q, Q) is NP-hard for $Q = 5$, and for general q it is hard for $Q = q + 2\lceil \frac{q}{3} \rceil$ [22], [23]. For all sufficiently large (but unspecified) q, Khot [12] showed that the problem is NP-hard for $Q = q^{\frac{\log q}{25}}$. For the almost coloring variant, Dinur, Khot, Perkins and Safra [1] showed that ALMOSTCOLORING$_\varepsilon(q, q^2)$ is NP-hard for all values of $q \geq 3$. We note that our result in Theorem I.2 significantly improves on the work of [1] for all values of $q \geq 5$. It is incomparable to the result of Khot [12] since our lower bound is better and holds also for small values of q, though it is only for the almost coloring variant.

Assuming the UGC yields stronger results for graph coloring. Bansal and Khot [7], [8] showed, assuming UGC, that ALMOSTCOLORING$_\varepsilon(2, Q)$ is NP-hard for any positive integer $Q \geq 3$. Similarly, assuming a variant of UGC, Dinur, Mossel and Regev [24] showed that COLORING$(3, Q)$ is NP-hard for all positive integers $Q \geq 3$.

D. Our Techniques

The overall approach of our proof builds upon the ideas of [4] and [1]. The constructions in [4], [1] and in our paper consist of three main parts : (i) An initial *Label Cover Problem*, (ii) *Outer Verifier* on blocks of variables, and (iii) a final *Combinatorial Gadget*. In the remainder of this section we shall describe how our construction differs from and builds upon the work of [4] and [1] in each of these steps ([1] can be considered as the special case $k = 1$).

Label Cover: The constructions of [4] and [1] use the same Label Cover instance based on the PCP Theorem [25], [26] combined with Raz's Parallel Repetition Theorem [27]. Using this they prove the following theorem on independent sets in a special class of graphs. Before we state the theorem, let us define (m, r)-co-partite graphs: a graph $G(V, E)$ is (m, r)-co-partite if $V = M \times R$, where $|M| = m$, $|R| = r$, such that for each $i \in M$, the subset of vertices $\{i\} \times R$ is a clique.

The following theorem can be deduced from the PCP Theorem and Raz's Parallel Repetition Theorem.

Theorem I.4. *For any $\delta > 0$ and positive integer h, there exists a parameter r such that the following problem is NP-hard : Given an (m, r)-co-partite graph G, decide between the following two cases:*

- *YES Case: There is an independent set in G of size m.*
- *NO Case: Any subset of vertices of G of size at least δm contains a clique of size h.*

The above theorem is, however, not strong enough to be used in our reduction. We require a strengthening of the NO Case to lower bound the size of cliques in the *k-wise repeated graph* G^k, for a given positive integer k. This k-wise repeated graph has vertex set V^k and any two tuples $(u_1, \ldots, u_k), (v_1, \ldots, v_k) \in V^k$ are adjacent if there exist edges between u_i and v_i in G for all $i = 1, \ldots, k$. Note that $|V^k| = (mr)^k$. In this work we prove and use the following theorem.

Theorem I.5. *For any $\delta > 0$ and positive integers k, h, there exists a parameter r such that the following problem is NP-hard: Given a (m, r)-co-partite graph G and letting G^k be its k-wise repetition, decide between the following two cases:*

- *YES Case: There is an independent set in G of size m.*
- *NO Case: Any subset of V^k of size at least δm^k contains a clique (in G^k) of size h.*

We would like to emphasize that the above theorem does not automatically follow from Theorem I.4, even in the case $k = h = 2$. In the NO Case of Theorem I.4, the graph G may have an independent set I of size $\frac{m}{r}$ (since only independent set of size δm is ruled out and $r = (1/\delta)^C$ for some large constant C therein). Then $I \times V$ is an independent set of size m^2 (recall that $|V| = mr$) in G^2!

Indeed, we require a considerable effort to prove Theorem I.5. The proof proceeds via the query efficient PCP over a large alphabet constructed by Håstad and Khot [5] (which itself builds on [3], [28]). This PCP consists of a set of tests or predicates, each over a small subset of the variables. The number of satisfying patterns, say B, for any predicate is much smaller than the inverse of the *soundness* of the

PCP. Roughly speaking, this rules out the existence of even the not-so-large independent sets (of $\approx 1/B^k$ fraction) in the NO case of the co-partite graph G derived from this PCP. Combining this property with a careful analysis of the structure of G^k yields the desired hardness result. Theorem I.5 is restated as Theorem III.1, though its proof is omitted due to lack of space and appears in the full version of this paper.

Outer Verifier

The Outer Verifier is a new graph derived from the (m, r)-co-partite graph $G(V, E)$ given by Theorem I.5 stated above. A simplified overview of the Outer Verifier graph is as follows. Set a parameter $l \gg r$ and for every *block* $B \in \binom{V}{l}$ and true-false assignment $a : B \mapsto \{\mathsf{T}, \mathsf{F}\}$, there is a vertex (B, a). The T values are supposed to correspond to an independent set in G. For some technical reasons, only those assignments are considered which have a sufficient number of T values, and we call this set of assignments R_B. The vertices corresponding to assignments R_B for any particular block B form a clique. Our construction crucially differs from that of [4] and [1] in the manner in which edges across different blocks are defined.

In the previous constructions of [4], [1], there is an edge between (B_1, a_1) and (B_2, a_2) if (i) $B_1 = \hat{B} \cup \{u\}$ and $B_2 = \hat{B} \cup \{v\}$ where $\{u, v\} \in E$ and $\hat{B} \in \binom{V}{(l-1)}$ is a *sub-block*, and (ii) either the restrictions of a_1 and a_2 on the sub-block $\hat{B}$ are inconsistent or $a_1(u) = a_2(v) = \mathsf{T}$. The vertices u and v are referred to as *pivot* vertices for this edge. It can be shown that if there is a large independent set in G then there is a large independent set in the Outer Verifier graph as well.

In our construction, we define edges in a more general manner: (B_1, a_1) and (B_2, a_2) have an edge between them iff ([1] corresponds to the case $k = 1$):

(i) $B_1 = \hat{B} \cup \{u_1, \ldots, u_k\}$ and $B_2 = \hat{B} \cup \{v_1, \ldots, v_k\}$ where $\{u_i, v_i\} \in E$ for all $i = 1, \ldots, k$ and $\hat{B} \in \binom{V}{(l-k)}$ is a *sub-block*.

(ii) Either the restrictions of a_1 and a_2 on $\hat{B}$ are inconsistent or $a_1(u_i) = a_2(v_i) = \mathsf{T}$ for some $i \in \{1, \ldots, k\}$.

In our construction there are k pairs of pivot vertices. As before, it can be shown that the Outer Verifier graph in our construction, which we denote as $G_{\mathcal{B}}$, has a large independent set if G has a large independent set. Since we have k pairs of pivot vertices, for our analysis we require the property about the existence of large cliques in G^k given by Theorem I.5.

Combinatorial Gadget: The combinatorial gadget we use is essentially the same as the one used in [1]. For any integer q, probability parameter p and dimension n, consider the graph $G_{q,p}[n]$ whose vertex set is $\{*, 1, \ldots, q\}^n$, which we refer to as colorings. There are edges between two colorings F_1 and F_2 if for all $i = 1, \ldots, n$, $(F_1(i), F_2(i)) \neq \{(1,1), (2,2), \ldots, (q,q)\}$, i.e. no coordinate is colored with the same symbol in $[q]$ by both the colorings. The colorings are equipped with a product measure that assigns, in each coordinate, a weight $1 - p$ (which is set to be very small) for the '$*$' symbol and weight p/q to each of the symbols in $[q]$.

Every coordinate $i \in [n]$ yields $[q]$ disjoint independent sets of measure p/q each, by taking the q subsets of colorings which have a particular symbol from $[q]$ in the ith coordinate. It is also easy to see that maximal independent sets in $G_{q,p}[n]$ are monotone under the following partial order in each coordinate: $* < j$ for all $j \in [q]$. The results proved in [1], based on Friedgut's and Russo's theorems, imply that every monotone subset of the colorings is a *junta* under a small perturbation of the bias p. The Diametric Theorem of Ahlswede and Khachatarian [29] can be applied to show that when $p < 1$, any monotone subset of the colorings of measure at least $1/q^{k+1}$ contains two colorings F, F' such that there are at most k coordinates on which both F, F' have the same value in $[q]$. The work of [1] also uses a similar fact, albeit only for $k = 1$.

We use this gadget with the graph $G_{\mathcal{B}}$ in our construction as follows: in our final graph there is a copy of $G_{q,p}[|R_B|]$ for each block $B \in \binom{V}{l}$, consisting of the vertex set $\{*, 1, \ldots, q\}^{R_B}$, i.e. colorings of the block assignments R_B. In other words there is a vertex (B, F) for every coloring F to the block assignments R_B. There is an edge between (B_1, F_1) and (B_2, F_2) if both the following conditions are satisfied.

(1) $B_1 = \hat{B} \cup \{u_1, \ldots, u_k\}$ and $B_2 = \hat{B} \cup \{v_1, \ldots, v_k\}$ where $\{u_i, v_i\} \in E$ for all $i = 1, \ldots, k$ and $\hat{B} \in \binom{V}{(l-k)}$ is a *sub-block*.

(2) For all $a_1 \in R_{B_1}$ and $a_2 \in R_{B_2}$: $F_1(a_1) = F_2(a_2) \in [q] \implies$ there is an edge between (B_1, a_1) and (B_2, a_2) in $G_{\mathcal{B}}$.

A crucial ingredient in our construction is to upper bound the value of q for which the final graph is dense enough. For this we prove that for all values of $q \geq 2^k + 1$ the following holds : for any two blocks B_1 and B_2 satisfying property (1) above, there is a joint distribution on the q-colorings $F_1 \in \{1, \ldots, q\}^{R_{B_1}}$ and $F_2 \in \{1, \ldots, q\}^{R_{B_2}}$ such that F_1 and F_2 are uniformly distributed as marginals and (B_1, F_1) and (B_2, F_2) have an edge between them in the final graph. This follows from the following lemma that is restated as Lemma II.10 and proved in Section II-D.

Lemma. *For any positive integers k, q such that $q \geq 2^k + 1$, there exists a joint distribution over two q-colorings of $2^{[k]}$, viz. $f, g : 2^{[k]} \mapsto [q]$ such that : (i) the marginal distributions of both f and g are uniform over all the q-colorings of $2^{[k]}$, and (ii) for any $S, T \subseteq [k]$, if $S \cap T = \emptyset$, then $f(S) \neq g(T)$.*

The work of [1] uses a similar fact, but only for $k = 1$ and $q \geq 2^1 + 1 = 3$. For $k = 1$, the structural constraints

satisfied by the two colorings have been frequently referred to as "α-constraints" (for eg. in [4], [1] and [24]).

E. Organization of the paper

In the next section we describe some useful objects and state some relevant combinatorial results. In Section III we give the hardness reduction and state its completeness and soundness properties which are proved in Sections IV and V respectively. The rest of the paper contains the proof of a strengthened hardness result for finding independent sets in co-partite graphs, formally stated as Theorem III.1.

II. Preliminaries

As in [1] we first formally define and state properties of several objects before we describe the reduction. For the common objects we follow notation similar to that of [1].

A. Graph $G_{q,p}[n]$

For any positive integers n, q and probability parameter $p \in (0,1]$ the graph $G_{q,p}[n]$ consists of the weighted vertex set $\{*,1,\ldots,q\}^n$. The measure μ_p on the vertex set is a product measure assigning, in each coordinate, probability mass $1-p$ to $*$ and $\frac{p}{q}$ to each of the remaining q elements. There is an edge between vertices $F_1, F_2 \in \{*,1,\ldots,q\}^n$ iff for all $i \in [n]$ $(F_1(i), F_2(i)) \notin \{(1,1),(2,2),\ldots,(q,q)\}$.

B. Definitions

- Monotonicity: A family $\mathcal{F} \subseteq \{*,1,\ldots,q\}^n$ is *monotone* if $F \in \mathcal{F}$ implies $F' \in \mathcal{F}$ where F' can be obtained by changing some $*$ in any coordinate of F to some element in $[q]$. It is easy to see that any maximal independent set in $G_{q,p}[n]$ is monotone.
- Every element of $\{*,1,\ldots,q\}^n$ is referred to as a *coloring* of $[n]$.
- Two colorings F_1 and F_2 *agree* on $i \in [n]$ if $F_1(i) = F_2(i) \in [q]$ (note that the common coordinate is not $*$).
- A family $\mathcal{F}$ of colorings of $[n]$ is *agreeing* if for all $F_1, F_2 \in \mathcal{F}$, there is an $i \in [n]$ such that F_1, F_2 agree on i.
- A family $\mathcal{F}$ is $(k+1)$-*agreeing* if for all $F_1, F_2 \in \mathcal{F}$ there exists a set of $k+1$ coordinates $\{i_1, i_2, \ldots, i_{k+1}\} \subseteq [n]$ so that F_1, F_2 agree on all of them, i.e. $F_1(i_j) = F_2(i_j) \in [q]$ for all $j \in [k+1]$.
- A set $C \subseteq [n]$ is a (δ, p)-*core* for a family $\mathcal{F}$, if there exists a family $\mathcal{F}'$ such that $\mu_p(\mathcal{F} \triangle \mathcal{F}') \leq \delta$ and $\mathcal{F}'$ depends only on the coordinates in C, i.e. for any $F \in \{*,1,\ldots,q\}^n$, changing the value of the coordinates in $[n] \setminus C$ does not affect whether F is in $\mathcal{F}'$.
- For $t \in (0,1)$ and a subset $C \subseteq [n]$ let a *core-family* $[\mathcal{F}]^t_C$ be defined as follows,

$$[\mathcal{F}]^t_C = \{F \in \{*,1,\ldots,q\}^C \mid \Pr_{F' \in \mu_p^{[n]\setminus C}}[(F,F') \in \mathcal{F}] > t\},$$

where (F, F') is a combined coloring that is obtained by choosing the coloring assigned by F on coordinates in C and by F' on coordinates in $[n] \setminus C$.

- The *influence* of a coordinate $i \in [n]$ for a family $\mathcal{F}$ is defined as follows:

$$\mathsf{Inf}_i^p(\mathcal{F}) := \mu_p(\{F : F|_{i=*} \notin \mathcal{F} \text{ and, } F|_{i=r} \in \mathcal{F} \text{ for some } r \in [q]\}),$$

where $F|_{i=*}$ is a coloring identical to F except on the ith coordinate where it is $*$, and similarly for $F|_{i=r}$ for any $r \in [q]$.

- The *average sensitivity* of $\mathcal{F}$ at p is,

$$\mathsf{as}_p(\mathcal{F}) := \sum_{i=1}^{n} \mathsf{Inf}_i^p(\mathcal{F}).$$

C. Useful Results

We begin by stating following simple lemma proved in [1] (as Lemma 1), and we omit the proof here.

Lemma II.1. *[Russo's Lemma [30]] Let $\mathcal{F} \subseteq \{*,1,\ldots,q\}^n$ be monotone, then $\mu_p(\mathcal{F})$ is increasing with p. In fact,*

$$\frac{1}{q} \cdot \mathsf{as}_p(\mathcal{F}) \leq \frac{d\mu_p(\mathcal{F})}{dp} \leq \mathsf{as}_p(\mathcal{F}).$$

For our analysis we shall require Friedgut's Theorem [31] which we state below.

Theorem II.2. *[Friedgut's Theorem [31]] Fix $\delta > 0$. Let $\mathcal{F} \subseteq \{*,1,\ldots,q\}^n$ with $a = \mathsf{as}_p(\mathcal{F})$. There exists a function $C_{Friedgut}(p,\delta,a) \leq c_p^{a/\delta}$, for a constant c_p depending only on p, so that $\mathcal{F}$ has a (δ,p)-core C of size $|C| \leq C_{Friedgut}(p,\delta,a)$.*

The following theorem of Bourgain, Kahn, Kalai, Katznelson and Linial [32] lower bounds the average sensitivity of monotone families.

Theorem II.3. *([32]) For any monotone family $\mathcal{F}$ such that $\mu_p(\mathcal{F}) \leq 1/2$,*

$$\mathsf{as}_p(\mathcal{F}) \geq \mu_p(\mathcal{F}) \log\left(\frac{1}{\mu_p(\mathcal{F})}\right).$$

We now bound the maximum size of a $(k+1)$-agreeing monotone family. We shall use the Diamteric Theorem of Ahlswede and Khachatarian [29] which we restate here.

Theorem II.4. *(Diametric Theorem [29]) Consider the family $\{1,\ldots,q\}^n$ for some $q \geq 2$, and fix a positive integer $t < n$. For any integer $i \geq 0$, define,*

$$\mathcal{K}_i := \Big\{F \in \{1,\ldots,q\}^n \mid |\{j \in [1, t+2i] \mid F(j) = 1\}| \geq t+i\Big\}.$$

Let $r \geq 0$ be the smallest non-negative integer satisfying,

$$t + 2r < \min\left\{n+1, t + 2\frac{t-1}{q-2}\right\}.$$

Then the maximum size of a t-agreeing subset of $\{1,\ldots,q\}^n$ is $|\mathcal{K}_r|$. Note that $|\mathcal{K}_0| = q^{n-t}$.

The following two lemmas follows from a simple application of the above theorem.

Lemma II.5. *Let k, q be positive integers with $q \geq 2^k + 1$, $p \in (0,1]$. Let $\mathcal{I} \subseteq \{*,1,\ldots,q\}^n$ be a monotone and agreeing family, where $n \gg k$. Then, $\mu_p(\mathcal{I}) \leq 1/q$.*

Proof: By monotonicity and using Lemma II.1 we have that $\mu_1(\mathcal{I}) \geq \mu_p(\mathcal{I})$. With our setting of parameters and applying Theorem II.4 we obtain that any 1-agreeing subset of $\{1,\ldots,q\}^n$ contains at most $1/q$ fraction of elements. Thus $\mu_p(\mathcal{I}) \leq \mu_1(\mathcal{I}) \leq 1/q$. ∎

Lemma II.6. *Let $k > 0$ be any integer such that $\mathcal{F} \subseteq \{*,1,\ldots,q\}^n$ be monotone with $n \gg k$ and $q \geq 2^k + 1$ and $\mu_p(\mathcal{F}) \geq \left(\frac{1}{q}\right)^{k+1}$ for some $0 < p < 1$. Then $\mathcal{F}$ is not $(k+1)$-agreeing. Specifically, there exist $F_1, F_2 \in \mathcal{F}$ such that F_1 and F_2 agree on at most k coordinates.*

Proof: We may assume that $\mu_p(\mathcal{F}) < 1$, otherwise the lemma is trivially true. Along with monotonicity, this implies that $\mathsf{as}_p(\mathcal{F}) > 0$. From Lemma II.1 and since $p < 1$, we have that $\mu_1(\mathcal{F}) > \mu_p(\mathcal{F}) \geq \left(\frac{1}{q}\right)^{k+1}$. Therefore, we can assume $\mathcal{F}$ to be a subset of $[q]^n$ – this clearly preserves the $(k+1)$-agreeing property if it existed. Since $n \gg k$, in Theorem II.4 using the setting of parameters $t = k+1$ and $q \geq 2^k+1$, we get that $r = 0$. This implies that since $\mu_1(\mathcal{F}) > \left(\frac{1}{q}\right)^{k+1}$, the subset $\mathcal{F}$ is not $(k+1)$-agreeing. ∎

We shall also require the following "Sunflower Theorem" of Erdős and Rado [33].

Theorem II.7. *There is a function $C_{ER}(\ell, d)$ so that for any $\mathcal{I} \subseteq \binom{R}{\ell}$, if $|\mathcal{I}| \geq C_{ER}(\ell, d)$ then there are d distinct sets $I_1, \ldots, I_d \in \mathcal{I}$ so that if $\Delta = I_1 \cap I_2 \cap \cdots \cap I_d$ then the sets $I_j \setminus \Delta$ are pairwise disjoint for $j = 1, \ldots, d$. This also holds when $\mathcal{I}$ contains subsets of size at most ℓ rather than exactly ℓ.*

Finally we state below Propositions II.8 and II.9 (proved as Proposition 3 in [1] and Lemma 3.1 in [4] respectively).

Proposition II.8. *Let $\mathcal{F} \subseteq \{*,1,\ldots,q\}^n$, and $C \subseteq [n]$.*

- *If $\mathcal{F}$ is monotone, then so is $[\mathcal{F}]_C^{\frac{3}{4}}$.*
- *If $\mathcal{F}$ is agreeing, then so is $[\mathcal{F}]_C^{\frac{3}{4}}$.*

Proposition II.9. *If C is a (δ, p)-core of $\mathcal{F}$, then $\mu_p^C\left([\mathcal{F}]_C^{\frac{3}{4}}\right) \geq \mu_p(\mathcal{F}) - 4\delta$.*

D. A Joint Distribution of Two Colorings of $2^{[k]}$

Before we describe the reduction we shall show the existence of a joint distribution over two colorings of subsets of $[k]$ (for any integer $k \geq 1$) satisfying some desired properties. Let q be an integer such that $q \geq 2^k+1$. Consider the set $2^{[k]}$. A coloring f of $2^{[k]}$ with q colors is a mapping $f : 2^{[k]} \mapsto [q]$. We prove the following lemma.

Lemma II.10. *For any integer $k \geq 1$ and integer $q \geq 2^k+1$, there exists a joint distribution $\mathcal{D}$ over two colorings $f, g : 2^{[k]} \mapsto [q]$ satisfying the following two properties.*

1) *The marginal distributions on f and g are identical and same as that of a random mapping of $2^{[k]}$ to $[q]$.*
2) *For any pair of colorings such that $\mathcal{D}(f,g) > 0$, if $S, T \subseteq [k]$ such that $S \cap T = \emptyset$, then $f(S) \neq g(T)$.*

Proof: The distribution $\mathcal{D}$ is constructed as follows. Choose a random coloring f. Based on this choice of f we shall construct a coloring g which satisfies the properties in the lemma. Let the set $A_f := \{j \in [q] \mid |f^{-1}(j)| \geq 1\}$, i.e. A_f is the set of colors that are used at least once in the coloring f. Also, let $B_f := \{j \in [q] \mid |f^{-1}(j)| > 1\}$ be the set of colors used at least twice in the coloring f. Let $j^* := f(\emptyset)$, and let $B'_f = B_f \cup \{j^*\}$. By a simple counting argument it is easy to see that,

$$|[q] \setminus A_f| \geq |B'_f|. \tag{1}$$

We construct the coloring g in a random way as follows:

1) Choose a random injective (one-to-one) map $h : B'_f \mapsto ([q] \setminus A_f)$. Such a map exists by Equation (1).
2) Define g as follows:

$$g(S) = \begin{cases} h(f(S)) & \text{if } f(S) \in B'_f. \\ f(S) & \text{otherwise.} \end{cases}$$

Observe that in the above construction, the color classes in f corresponding to the set of colors B'_f are colored in g with randomly chosen distinct colors from $[q] \setminus A_f$. This preserves the size of the color classes and therefore the marginal distribution of g is identical to the uniform distribution. It is easy to see this also satisfies property 2 of the lemma. ∎

III. Reduction

In this section we give the hardness reduction for proving Theorem I.1. As in previous works [4], [1], we begin with the problem of finding independent sets in *co-partite* graphs. However, we need a stronger hardness of finding nearly independent sets in what we refer to as the *k-wise repeated graph*.

A graph $G(V,E)$ is (m,r)-co-partite if $V = M \times R$, where $|M| = m$, $|R| = r$, such that for each $i \in M$, the subset of vertices $\{i\} \times R$ is a clique. Let $\mathsf{IS}(G)$ be the size of maximum independent set in G.

For an (m,r)-co-partite graph $G(V,E)$, define the *k-wise repeated graph* G^k consisting of vertex set V^k. Note that $|V^k| = (mr)^k$. There is an edge between $\overline{u} = (u_1, \ldots u_k), \overline{v} = (v_1, \ldots, v_k) \in V^k$ if there is an edge in G between u_i and v_i for all $i \in [k]$. For any integer $h \geq 2$, let $\mathsf{IS}_h(G^k)$ be the maximum size of a subset of vertices of

G^k which do not contain a clique of size h. For parameters h, γ, r, k the $\mathsf{khIS}(r, \gamma, h, k)$ problem is: given a (m, r)-co-partite graph G, distinguish between the cases:

- YES case: $\mathsf{IS}(G) = m$.
- NO case: $\mathsf{IS}_h(G^k) < \gamma m^k$.

We prove the following theorem whose proof is omitted due to lack of space and appears in the full version of this paper.

Theorem III.1. *For any $k, h, \gamma > 0$, there exists a constant $r = r(k, h, \gamma)$ such that the problem $\mathsf{khIS}(r, \gamma, h, k)$ is NP-hard.*

The theorem below states our reduction starting from a (m, r)-co-partite graph G to a weighted graph $G^q_{\mathcal{B}}$ and, along with Theorem III.1, proves Theorem I.1.

Theorem III.2. *For any $\varepsilon > 0$ and positive integers k, q such that $q \geq 2^k + 1$, there exists a small enough $\varepsilon_0 > 0$ and large enough $h > 0$ (both depending only on k, q and ε) such that: for a constant integer r, given an (m, r)-co-partite graph $G(V, E)$, there is a polynomial time reduction to a weighted graph $G^q_{\mathcal{B}}$, such that:*

- *Total weight of all the vertices in $G^q_{\mathcal{B}}$ is 1.*
- *(Completeness) $\mathsf{IS}(G) = m$ implies that there are q disjoint independent sets $I_1, \ldots, I_k$ in $G^q_{\mathcal{B}}$, each of weight $\frac{1-2\varepsilon}{q}$. In particular, there is a q-colorable subset of vertices V' in $G^q_{\mathcal{B}}$ with weight $1 - 2\varepsilon$.*
- *(Soundness) $\mathsf{IS}_h(G^k) < \varepsilon_0 m^k$ implies that $\mathsf{IS}(G^q_{\mathcal{B}}) < 1/q^{k+1}$ where $\mathsf{IS}(G^q_{\mathcal{B}})$ is the maximum weight of an independent set in $G^q_{\mathcal{B}}$.*

The rest of this section and Sections IV and V are devoted to proving the above theorem. We begin with the setting of certain parameters based on the values of ε, k and q given in the statement of Theorem III.2.

A. Setting of parameters

From the statement of Theorem III.2 we are given $\varepsilon > 0$ and positive integers k and q satisfying $q \geq 2^k + 1$. Additionally, we define the following list of parameters.

- $p = 1 - \varepsilon$.
- $\varepsilon_1 = \frac{\varepsilon}{8q^{2k+4}}$.
- $h_0 = \sup_{p' \in [1-\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{4}]} C_{Friedgut}(p', \frac{1}{16}\varepsilon_1, \frac{8q}{\varepsilon})$, where $C_{Friedgut}$ is the function from Theorem II.2.
- $\eta = \frac{1}{h_0 2^{k+3}} \left(\frac{p}{q}\right)^{(2^{k+1}+1)h_0}$.
- $h_1 = h_0 + \left\lceil \frac{8q}{\varepsilon\eta} \right\rceil$.
- $h_s = 1 + \sum_{j=0}^{h_0} \binom{h_1}{j} \cdot q^{h_0} \cdot q^{h_0}$.
- $h = C_{ER}(h_1, h_s)$, where C_{ER} is the function given by Theorem II.7.
- $\varepsilon_0 = \varepsilon_1 \cdot \frac{1}{2^{k+5}} \cdot \frac{1}{k^k}$.
- $l_{\mathsf{T}} = \max\{4 \ln \frac{2}{\varepsilon}, 2^{k+4} \cdot (h_1)^2 \cdot k!\} \cdot (10^6 \cdot k^3)$.

B. Construction of $G^q_{\mathcal{B}}$

We begin with a graph $G(V, E)$ which is (m, r)-co-partite with $V = M \times R$ where $|M| = m$ and $|R| = r$. Let $l = 2l_{\mathsf{T}} \cdot r^k$, and let $\mathcal{B}$ be the family of all subsets of V of size exactly l, i.e. $\mathcal{B} = \binom{V}{l}$. Each element $B \in \mathcal{B}$ is called a *block*. As in the constructions of [4], [1] we think of any independent set in G as an assignment of $\{\mathsf{T}, \mathsf{F}\}$ to each vertex in V, where T means the vertex is in the independent set. In the YES case, G has an independent set of size m, and therefore the expected number of T values in a randomly chosen block is $2l_{\mathsf{T}} \cdot r^{k-1}$. Therefore, w.h.p a random block has at least $l_{\mathsf{T}} \cdot r^{k-1}$ many T values in them. Using this fact, we let the set of *block assignments* R_B to the block B be the set of all $\{\mathsf{T}, \mathsf{F}\}$ assignments to the vertices in B such that at least $l_{\mathsf{T}} \cdot r^{k-1}$ of them are assigned value T. Formally, for any block $B \in \mathcal{B}$,

$$R_B = \{a : B \mapsto \{\mathsf{T}, \mathsf{F}\} \;\mid\; |a^{-1}(\mathsf{T})| \geq l_{\mathsf{T}} \cdot r^{k-1}\}.$$

We now construct an ***intermediate graph*** $G_{\mathcal{B}} = (V_{\mathcal{B}}, E_{\mathcal{B}})$. The vertex set $V_{\mathcal{B}} := \bigcup_{B \in \mathcal{B}} \{B\} \times R_B$, i.e. for each block B, there is a cluster of vertices, one for every block assignment in R_B. We let each cluster be a clique. We add edges between two clusters in the following manner. Let B_1 and B_2 be two blocks satisfying the following properties:

(i) B_1 and B_2 are $(l-k)$-wise intersecting, i.e. $\hat{B} := B_1 \cap B_2$ and $|\hat{B}| = l - k$. Let $B_1 = \{u_1, \ldots, u_k\} \cup \hat{B}$ and $B_2 = \{v_1, \ldots, v_k\} \cup \hat{B}$.

(ii) There is an edge between u_i and v_i in G, for all $i = 1, \ldots, k$. We refer to u_i (v_i) as the ith *pivot vertex for* B_1 (B_2).

Add an edge in $G_{\mathcal{B}}$ between (B_1, a_1) and (B_2, a_2) where $a_1 \in R_{B_1}$ and $a_2 \in R_{B_2}$ if and only if either of the following two conditions are satisfied,

- $a_1|_{\hat{B}} \neq a_2|_{\hat{B}}$, where $a_i|_{\hat{B}}$ is the restriction of a_i onto the sub-block $\hat{B}$ for $i = 1, 2$.
- $a_1(u_i) = a_2(v_i) = \mathsf{T}$ for some $i \in \{1, \ldots, k\}$.

This reduction preserves large independent sets as the following proposition shows.

Proposition III.3. $\mathsf{IS}(G) = m \implies \mathsf{IS}(G_{\mathcal{B}}) = m'(1 - \varepsilon)$ *where $m' = |\mathcal{B}|$.*

Proof: Let $I \subseteq V$ be an independent set of size m in G. Let $\sigma : V \mapsto \{\mathsf{T}, \mathsf{F}\}$ be a map which assigns T iff the vertex is in I. For each block $B \in \mathcal{B}$ let σ_B be the restriction of σ to the vertices in B. The subset $I_{\mathcal{B}} = \{(B, \sigma_B) \mid B \in \mathcal{B}\}$ is an independent set in $G_{\mathcal{B}}$ as the following argument shows. Consider two vertices (B_1, σ_{B_1}) and (B_2, σ_{B_2}) such that B_1 and B_2 are $(l-k)$-wise intersecting and there are edges in G between u_i and v_i for all $i = 1, \ldots, k$, where u_i (v_i) are pivot vertices for B_1 (B_2). Since I is an independent set in G, at most one of u_i or v_i is assigned T by the assignment σ, for $i = 1, \ldots, k$. Therefore, there cannot be an edge between

(B_1, σ_{B_1}) and (B_2, σ_{B_2}) in $G_{\mathcal{B}}$. We lose a small number of elements in $I_{\mathcal{B}}$ which might not have the required number $(= l_{\mathsf{T}} \cdot r^{k-1})$ of T-values assigned by σ, and this loss is of at most ε fraction as per the setting of the parameters. Thus, $G_{\mathcal{B}}$ contains an independent set of size $m'(1-\varepsilon)$. ∎

The ***final graph*** $G^q_{\mathcal{B}}$ is constructed as follows. Recall that $q \geq 2^k + 1$. The graph $G^q_{\mathcal{B}}$ has vertex set $V^q_{\mathcal{B}}$, edge set $E^q_{\mathcal{B}}$ and a weight function Λ. For every cluster (B, R_B) in $G_{\mathcal{B}}$, there is a copy of $G_{q,p}[n]$ in $G^q_{\mathcal{B}}$ where $n = |R_B|$. The set of vertices $V^q_{\mathcal{B}}[B]$ of the cluster corresponding to block B is the set of all the colorings of R_B. Formally,

$$V^q_{\mathcal{B}}[B] := \left\{(B, F) \;\middle|\; F \in \{*, 1, \ldots, q\}^{R_B}\right\},$$

$$V^q_{\mathcal{B}} := \bigcup_B V^q_{\mathcal{B}}[B].$$

We shall frequently abuse notation to denote a vertex (B, F) by F when the block B is clear from the context.

The *weight function* Λ is defined as follows. Let μ_p be the distribution on $V^q_{\mathcal{B}}[B] = \{F \in \{*, 1, \ldots, q\}^{R_B}\}$. The weight function Λ assigns equal measure to each cluster, and within each cluster the vertices are weighted according to μ_p. Formally, for $F \in V^q_{\mathcal{B}}[B]$,

$$\Lambda(F) = |\mathcal{B}|^{-1} \mu_p(F).$$

The edges in the edge set $E^q_{\mathcal{B}}$ within each cluster are already determined by the edges in $G_{q,p}[n]$. The edges across clusters in $E^q_{\mathcal{B}}$ are as follows.

$$E^q_{\mathcal{B}} = \{((B_1, F_1), (B_2, F_2)) \in V^q_{\mathcal{B}}[B_1] \times V^q_{\mathcal{B}}[B_2] \;\mid\; F_1^{-1}(i) \times F_2^{-1}(i) \subseteq E_{\mathcal{B}}, \forall i \in [q]\}.$$

In other words, there is an edge between colorings F_1 in $V^q_{\mathcal{B}}[B_1]$ and F_2 in $V^q_{\mathcal{B}}[B_2]$ if and only if for all colors $i \in [q]$, if $a_1 \in R_{B_1}$ and $a_2 \in R_{B_2}$ such that $F_1(a_1) = F_2(a_2) = i$, then $((B_1, a_1), (B_2, a_2)) \in E_{\mathcal{B}}$. Stated in the contrapositive, this means that there is no edge between F_1 and F_2 if and only if there exist $a_1 \in R_{B_1}$ and $a_2 \in R_{B_2}$ such that $F_1(a_1) = F_2(a_2) \in [q]$ and $((B_1, a_1), (B_2, a_2)) \notin E_{\mathcal{B}}$.

The following proposition follows from Proposition 6 of [1].

Proposition III.4. *(Maximal independent sets in $G^q_{\mathcal{B}}$ are monotone) Let $\mathcal{I}$ be an independent set in $G^q_{\mathcal{B}}$. If $F \in \mathcal{I} \cap V^q_{\mathcal{B}}[B]$, and F' is monotonically above F, then $\mathcal{I} \cup \{F'\}$ is also an independent set.*

IV. Completeness

Lemma IV.1. *If $\mathsf{IS}(G) = m$ then there exist disjoint independent sets $I_1, \ldots, I_q$ in $G^q_{\mathcal{B}}$ such that $\Lambda(I_j) \geq \frac{1-2\varepsilon}{q}$, for $j = 1, \ldots, q$.*

Proof: By Proposition III.3, if $\mathsf{IS}(G) \geq m$ then $\mathsf{IS}(G_{\mathcal{B}}) \geq (1-\varepsilon)m'$. Let $\mathcal{I}_{\mathcal{B}}$ be an independent set in $G_{\mathcal{B}}$ of size $(1-\varepsilon)m'$. Due to the fact that each cluster in $G_{\mathcal{B}}$ is a clique, the set $\mathcal{I}_{\mathcal{B}}$ contains at most one vertex, say (B, a), from the cluster corresponding to $(1-\varepsilon)m'$ of the blocks B, where $a \in R_B$. For each $j \in [q]$ define,

$$I_j = \{(B, F) \in G^q_{\mathcal{B}} \mid \exists a \in R_B \text{ s.t. } (B, a) \in \mathcal{I}_{\mathcal{B}} \text{ and, } F(a) = j\}.$$

In other words, for each block B such that there is some $(B, a) \in \mathcal{I}_{\mathcal{B}}$, the set I_j contains all colorings that assign the color j to the assignment a. Since $\mathcal{I}_{\mathcal{B}}$ contains at most one vertex (B, a) from from each block, it is easy to see that the sets I_j are disjoint for $j = 1, \ldots, q$. To see that I_j is an independent set observe the following two cases:

- For a block B, let $(B, F_1), (B, F_2) \in I_j$. Then there is an assignment $a \in R_B$ such that $F_1(a) = F_2(a) = j$ and therefore from the structure of $G_{q,p}[n]$, there is no edge between the colorings F_1 and F_2 in the copy of $G_{q,p}[n]$ corresponding to block B.
- For $B_1 \neq B_2$, let $(B_1, F_1), (B_2, F_2) \in I_j$. Then there must exist $(B_1, a_1), (B_2, a_2) \in \mathcal{I}_{\mathcal{B}}$ so that $((B_1, a_1), (B_2, a_2)) \notin E_{\mathcal{B}}$ along with the property that $F_1(a_1) = F_2(a_2) = j$. This implies that there is no edge in $G^q_{\mathcal{B}}$ between (B_1, F_1) and (B_2, F_2).

Also, the weight of I_j is $\Lambda(I_j) = (1-\varepsilon)\frac{1-\varepsilon}{q} \geq \frac{1-2\varepsilon}{q}$ since the weight of I_j in each cluster is of all colorings that assign color j to a chosen block assignment (B, a). ∎

V. Soundness

Lemma V.1. *If $\mathsf{IS}(G^q_{\mathcal{B}}) \geq \left(\frac{1}{q}\right)^{k+1}$, then $\mathsf{IS}_h(G^k) \geq \varepsilon_0 m^k$.*

Let $\mathcal{I}$ be an independent set in $G^q_{\mathcal{B}}$ with $\Lambda(\mathcal{I}) \geq \frac{1}{q^{k+1}}$. Denote by $\mathcal{I}[B]$ the intersection $\mathcal{I} \cap V^q_{\mathcal{B}}$. WLOG we may assume that $\mathcal{I}$ is maximal. The following proposition is a restatement of Proposition 7 of [1].

Proposition V.2. *For all $B \in \mathcal{B}$, $\mathcal{I}[B]$ is monotone and agreeing.*

Thus, by Lemma II.5, $\mu_p(\mathcal{I}[B]) \leq \mu_1(\mathcal{I}[B]) \leq 1/q$. By averaging, it must be that for at least $\frac{1}{2} \cdot \frac{1}{q^{k+1}}$ fraction of the blocks B, the measure $\mu_p(\mathcal{I}[B])$ is at least $\frac{1}{2} \cdot \frac{1}{q^{k+1}}$. Therefore, by Theorem II.3 and Lemma II.1, if p is increased from $1-\varepsilon$ to $1-\frac{\varepsilon}{2}$, then the new measure $\mu(\mathcal{I})$ increases by at least, $\frac{\varepsilon}{2} \cdot \frac{1}{2q^{k+1}} \cdot \frac{1}{q} \cdot \frac{1}{2q^{k+1}} \geq \frac{\varepsilon}{8q^{2k+4}}$. So, the new measure is,

$$\Lambda_{1-\frac{\varepsilon}{2}}(\mathcal{I}) \geq \frac{1}{q^{k+1}} + \frac{\varepsilon}{8q^{2k+4}} = \frac{1}{q^{k+1}} + \varepsilon_1. \tag{2}$$

We borrow and generalize the following terminology from [1].

- $F \in V^q_{\mathcal{B}}$ is referred to as a *coloring* of the block assignments in R_B, i.e. $F \in \{*, 1, \ldots, q\}^{R_B}$ or $F : R_B \mapsto \{*, 1, \ldots, q\}$.
- A family of colorings $\mathcal{F} \subseteq \{*, 1, \ldots, q\}^{R_B}$ is *monotone* if it is a monotone subset of $\{*, 1, \ldots, q\}^{R_B}$.

- Two colorings $F_1, F_2 \in \{*, 1, \ldots, q\}^{R_B}$ *agree* on a block assignment $a \in R_B$ if $F_1(a) = F_2(a) \in [q]$.
- A set of block assignments $\mathcal{A} = \{a_1, \ldots, a_s\} \subseteq R_B$ is *distinguished* by a family $\mathcal{F}$ of colorings of R_B if there exist $F_1, F_2 \in \mathcal{F}$ which agree on the block assignments in $\mathcal{A}$ and do not agree on any block assignment in $R_B \setminus \mathcal{A}$. We also say that $\mathcal{F}$ has a distinguished set $\mathcal{A}$.

In the next subsection we construct a large enough subset $\mathcal{B}' \subseteq \mathcal{B}$ of the blocks such that the families $\mathcal{I}[B]$ for $B \in \mathcal{B}'$ satisfy certain properties.

A. Selection of $\mathcal{B}' \subseteq \mathcal{B}$.

Lemma V.3. *There exists some $p' \in [1 - \frac{\varepsilon}{2}, 1 - \frac{\varepsilon}{4}]$ and a set of blocks $\mathcal{B}' \subseteq \mathcal{B}$ whose size is $|\mathcal{B}'| \geq \frac{1}{4}\varepsilon_1 \cdot |\mathcal{B}|$, such that for all $B \in \mathcal{B}'$:*

1) *$\mathcal{I}[B]$ has a $(\frac{1}{16}\varepsilon_1, p')$-core, $\mathsf{Core}[B] \subseteq R_B$, of size $|\mathsf{Core}[B]| \leq h_0$.*
2) *The core family $\mathcal{CF}_B := [\mathcal{I}[B]]^{\frac{3}{4}}_{\mathsf{Core}[B]}$ has a distinguished subset $\mathcal{A}_B \subseteq \mathsf{Core}[B]$ of block assignments such that $1 \leq |\mathcal{A}_B| \leq k$.*

Proof: First we define a subset of blocks in which $\mathcal{I}$ has significantly large weight.

$$\tilde{\mathcal{B}} = \left\{ B \in \mathcal{B} \;\mid\; \mu_{1-\frac{\varepsilon}{2}}(\mathcal{I}[B]) \geq \frac{1}{q^{k+1}} + \frac{\varepsilon_1}{2} \right\}.$$

Using Equation (2) and averaging we obtain, $|\tilde{\mathcal{B}}| \geq \frac{1}{2}\varepsilon_1 \cdot |\mathcal{B}|$. Since the measure of the monotone families $\mu_p(\mathcal{I}[B])$ increases with p, we have that $\mu_{p'}(\mathcal{I}[B]) \geq \frac{1}{q^{k+1}} + \frac{1}{2}\varepsilon_1$ for all $B \in \mathcal{B}'$ and $p' \in [1 - \frac{\varepsilon}{2}, 1 - \frac{\varepsilon}{4}]$ (we shall fix p' later). Define the subset $\mathcal{B}'$ as follows,

$$\mathcal{B}' = \left\{ B \in \tilde{\mathcal{B}} \;\mid\; \mathsf{as}_{p'}(\mathcal{I}[B]) \leq \frac{8q}{\varepsilon} \right\}. \tag{3}$$

The following proposition is identical to Proposition 8 in [1] and we state it without proof.

Proposition V.4. *There exists $p' \in [1 - \frac{\varepsilon}{2}, 1 - \frac{\varepsilon}{4}]$ such that*

$$|\mathcal{B}'| \geq \frac{1}{4}\varepsilon_1 \cdot |\mathcal{B}|.$$

We fix p' as given by the above proposition. Property 1 of Lemma V.3 follows directly from Theorem II.2 and the setting of the parameter h_0. To prove Property 2, observe that from the definition of $\mathcal{CF}_B$ in the statement of the lemma, and by Propositions V.2, II.8 and II.9, $\mathcal{CF}_B$ is monotone, pairwise agreeing and satisfies,

$$\mu_{p'}(\mathcal{CF}_B) \geq \mu_{p'}(\mathcal{I}[B]) - 4 \cdot \frac{\varepsilon_1}{16} \geq \frac{1}{q^{k+1}}. \tag{4}$$

Applying Lemma II.6 to the above lower bound we obtain that there are two colorings $F^\flat, F^\sharp \in \mathcal{CF}_B$ that agree only on a subset $\mathcal{A}_B \subseteq \mathsf{Core}[B]$ of the block assignments such that $1 \leq |\mathcal{A}_B| \leq k$. Moreover, since $\mathcal{CF}_B$ is monotone, both $F^\flat$ and $F^\sharp$ can be assumed to be in $\{1, \ldots, q\}^{\mathsf{Core}[B]}$. ∎

Definition V.5. *For $p', \mathcal{B}'$ as defined above and $B \in \mathcal{B}'$, the extended core $\mathsf{ECore}[B]$ is defined as,*

$$\mathsf{Ecore}[B] := \mathsf{Core}[B] \bigcup \left\{ a \in R_B \;\mid\; \mathsf{Inf}_a^{p'}(\mathcal{I}[B]) \geq \eta \right\}.$$

The next proposition follows from the definition of influence and average sensitivity.

Proposition V.6. *(The extended core is small): For $B \in \mathcal{B}'$,*

$$|\mathsf{Ecore}[B]| \leq h_0 + \frac{\mathsf{as}_{p'}(\mathcal{I}[B])}{\eta} \leq h_0 + \left\lceil \frac{8q}{\varepsilon\eta} \right\rceil = h_1.$$

Definition V.7. *(Preservation:) Let $B \in \mathcal{B}'$ and $\tilde{B} \subseteq B$ such that $|\tilde{B}| = (l - k)$. For any block assignment $a \in R_B$ let $a|_{\tilde{B}}$ be the restriction of a to $\tilde{B}$. We say that $\tilde{B}$ preserves B if there is no pair of block assignments $a_1 \neq a_2 \in \mathsf{ECore}[B]$ with $a_1|_{\tilde{B}} = a_2|_{\tilde{B}}$.*

Lemma V.8. *For all $B \in \mathcal{B}'$,*

$$|\{X \subseteq B \;\mid\; |X| = k \text{ and } B \setminus X \text{ does not preserve } B\}| \leq \frac{h_1^2 (l-1)^{k-1}}{2}.$$

Proof: Each pair of block assignments in $\mathsf{ECore}[B]$ can cause at most $\binom{l-1}{l-k} \leq (l-1)^{k-1}$ of sub-blocks $\tilde{B}$ to not preserve B. Since, for any $B \in \mathcal{B}'$, $|\mathsf{ECore}[B]| \leq h_1$, the lemma follows. ∎

B. Selection of $\hat{B}$

In this subsection we shall use the subset $\mathcal{B}'$ of the blocks to select a $(l-k)$-sized sub-block $\hat{B} \in \binom{V}{l-k}$ which satisfies some desired properties.

First we need some more notation. Let $B \in \mathcal{B}'$, and $\mathcal{A}_B$ be the distinguished set of assignments given by Lemma V.3. Let T_B be the size of $\mathcal{A}_B$, so that $1 \leq T_B \leq k$. We fix an arbitrary numbering of the assignments in $\mathcal{A}_B$ and denote them by $\dot{a}_1[B], \ldots, \dot{a}_{T_B}[B]$, referring to $\dot{a}_i[B]$ as the ith distinguished assignment for block B.

Also, consider an element $\overline{v} \in V^k$, i.e. $\overline{v}$ is an ordered tuple $(v_1, \ldots, v_k)$ such that $v_i \in V$ for $i = 1, \ldots, k$. By abuse of notation, for any $\tilde{B} \in \binom{V}{l-k}$, we shall denote by $\tilde{B} \cup \overline{v}$ the set $\tilde{B} \cup \{v_1, \ldots, v_k\}$ where duplicates, if any, are removed. Similarly, for $B \in \binom{V}{l}$, $B \setminus \overline{v}$ shall be used to denote $B \setminus \{v_1, \ldots, v_k\}$.

Definition V.9. *For any $(l-k)$ sub-block $\tilde{B}$, let $V^k_{\tilde{B}} \subseteq V^k$ be,*

$$V^k_{\tilde{B}} = \Big\{ \overline{v} = (v_1, \ldots, v_k) \in (V \setminus \tilde{B})^k \;\mid\; B = \tilde{B} \cup \overline{v} \in \mathcal{B}', \\ \tilde{B} \text{ preserves } B,\ \dot{a}_i[B](v_i) = \mathsf{T}, \forall i \in \{1, \ldots, T_B\} \Big\}.$$

Proposition V.10. *There exists $\hat{B} \in \binom{V}{l-k}$ such that $|V^k_{\hat{B}}| \geq \varepsilon_1 \cdot \left(\frac{1}{2^{k+5}}\right) \cdot m^k$.*

The set $V^k_{\hat{B}}$ in the above proposition is used to complete the proof of Lemma V.1 and Theorem III.2. Due to lack of

space, the rest of the proof, including that of Proposition V.10 is omitted and appears in the full version of the paper.

REFERENCES

[1] I. Dinur, S. Khot, W. Perkins, and M. Safra, "Hardness of finding independent sets in almost 3-colorable graphs," in *Proc. IEEE FOCS*, 2010, pp. 212–221.

[2] L. Engebretsen and J. Holmerin, "More efficient queries in PCPs for NP and improved approximation hardness of maximum CSP," *Random Struct. Algorithms*, vol. 33, no. 4, pp. 497–514, 2008.

[3] A. Samorodnitsky and L. Trevisan, "A PCP characterization of NP with optimal amortized query complexity," in *Proc. ACM STOC*, 2000, pp. 191–199.

[4] I. Dinur and S. Safra, "On the hardness of approximating minimum vertex cover," *Annals of Mathematics*, vol. 165, no. 1, pp. 439–485, 2005.

[5] J. Håstad and S. Khot, "Query efficient PCPs with perfect completeness," *Theory of Computing*, vol. 1, no. 1, pp. 119–148, 2005.

[6] S. Khot, "On the power of unique 2-prover 1-round games," in *Proc. ACM STOC*, 2002, pp. 767–775.

[7] N. Bansal and S. Khot, "Optimal long code test with one free bit," in *Proc. IEEE FOCS*, 2009, pp. 453–462.

[8] ——, "Inapproximability of hypergraph vertex cover and applications to scheduling problems," in *Proc. ICALP*, 2010, pp. 250–261.

[9] U. Feige, "Approximating maximum clique by removing subgraphs," *SIAM Journal of Discrete Mathematics*, vol. 18, no. 2, pp. 219–225, 2004.

[10] J. Håstad, "Clique is hard to approximate within n to the power 1-epsilon," *Acta Mathematica*, vol. 182, pp. 105–142, 1999.

[11] L. Engebretsen and J. Holmerin, "Clique is hard to approximate within $n^{1\text{-}o(1)}$," in *Proc. ICALP*, 2000, pp. 2–12.

[12] S. Khot, "Improved inaproximability results for maxclique, chromatic number and approximate graph coloring," in *Proc. IEEE FOCS*, 2001, pp. 600–609.

[13] S. Khot and A. K. Ponnuswami, "Better inapproximability results for maxclique, chromatic number and min-3lin-deletion," in *Proc. ICALP*, 2006, pp. 226–237.

[14] S. Khot and O. Regev, "Vertex cover might be hard to approximate to within 2-epsilon," *J. Comput. Sys. Sci.*, vol. 74, no. 3, pp. 335–349, 2008.

[15] U. Feige, S. Goldwasser, L. Lovász, S. Safra, and M. Szegedy, "Interactive proofs and the hardness of approximating cliques," *J. ACM*, vol. 43, no. 2, pp. 268–292, 1996.

[16] A. Wigderson, "Improving the performance guarantee for approximate graph coloring," *J. ACM*, vol. 30, no. 4, pp. 729–735, 1983.

[17] A. Blum, "New approximation algorithms for graph coloring," *J. ACM*, vol. 41, no. 3, pp. 470–516, 1994.

[18] D. R. Karger, R. Motwani, and M. Sudan, "Approximate graph coloring by semidefinite programming," *J. ACM*, vol. 45, no. 2, pp. 246–265, 1998.

[19] A. Blum and D. R. Karger, "An $\tilde{O}(n^{3/14})$-coloring algorithm for 3-colorable graphs," *Information Processing Letters*, vol. 61, no. 1, pp. 49–53, 1997.

[20] S. Arora, E. Chlamtac, and M. Charikar, "New approximation guarantee for chromatic number," in *Proc. ACM STOC*, 2006, pp. 215–224.

[21] E. Halperin, R. Nathaniel, and U. Zwick, "Coloring k-colorable graphs using relatively small palettes," *J. Algorithms*, vol. 45, no. 1, pp. 72–90, 2002.

[22] S. Khanna, N. Linial, and S. Safra, "On the hardness of approximating the chromatic number," *Combinatorica*, vol. 20, no. 3, pp. 393–415, 2000.

[23] V. Guruswami and S. Khanna, "On the hardness of 4-coloring a 3-colorable graph," *SIAM Journal of Discrete Mathematics*, vol. 18, no. 1, pp. 30–40, 2004.

[24] I. Dinur, E. Mossel, and O. Regev, "Conditional hardness for approximate coloring," *SIAM Journ. of Comput.*, vol. 39, no. 3, pp. 843–873, 2009.

[25] S. Arora and S. Safra, "Probabilistic checking of proofs: A new characterization of NP," *J. ACM*, vol. 45, no. 1, pp. 70–122, 1998.

[26] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, "Proof verification and the hardness of approximation problems," *J. ACM*, vol. 45, no. 3, pp. 501–555, 1998.

[27] R. Raz, "A parallel repetition theorem," *SIAM Journ. of Comput.*, vol. 27, no. 3, pp. 763–803, 1998.

[28] J. Håstad and A. Wigderson, "Simple analysis of graph tests for linearity and PCP," *Random Struct. Algorithms*, vol. 22, no. 2, pp. 139–160, 2003.

[29] R. Ahlswede and L. H. Khachatarian, "The Diametric Theorem in Hamming spaces – optimal anticodes," *Advances in Applied Mathematics*, vol. 20, no. 4, 1998.

[30] L. Russo, "An approximate zero-one law," *Z. Wahrsch. Verw. Gebeite*, vol. 61, no. 1, pp. 129–139, 1982.

[31] E. Friedgut, "Boolean functions with low average sensitivity depend on few coordinates," *Combinatorica*, vol. 18, no. 1, pp. 27–35, 1998.

[32] J. Bourgain, J. Kahn, G. Kalai, Y. Katznelson, and N. Linial, "The influence of variables in product spaces," *Israel Journal of Mathematics*, vol. 77, no. 1, pp. 55–64, 1992.

[33] P. Erdős and R. Rado, "Intersection theorems for systems of sets," *London Math. Soc.*, vol. 35, pp. 85–90, 1960.