Hardness of Finding Independent Sets in Almost 3-Colorable Graphs

Irit Dinur
Weizmann Institute
Rehovot, Israel
irit.dinur@weizmann.ac.il

Subhash Khot
New York University
New York, USA
khot@cs.nyu.edu

Will Perkins
New York University
New York, USA
perkins@cims.nyu.edu

Muli Safra
Tel-Aviv University
Tel-Aviv, Israel
safra@post.tau.ac.il

***Abstract*—For every $\epsilon > 0$, and integer $q \geq 3$, we show that given an N-vertex graph that has an induced q-colorable subgraph of size $(1-\epsilon)N$, it is NP-hard to find an independent set of size $\frac{N}{q^2}$.**

***Keywords*-graph coloring; hardness of approximation; PCPs**

I. INTRODUCTION

One of the most successful approaches to inapproximability results follows a paradigm initiated by [1], [2], [3], in which the hardness reduction is based on a so-called composition of an *outer verifier* with a *long code* based *inner verifier*. There are some notable examples, such as 3SAT and MAX-3LIN, for which this approach leads to *tight* NP-hardness results [3]. In these cases, the outer verifier is obtained rather directly from the PCP Theorem [4], [5], [6] together with the Parallel Repetition Theorem [7], referred henceforth as the Raz Verifier. However, it is often the case that the best NP-hardness result falls short of the best known algorithmic result. In the past few years, the Unique Games Conjecture [8] has filled in this gap and led to tight inapproximability results for an impressively large and diverse collection of problems (see [9], [10] for surveys). In these results, one uses a *Unique Game* as an outer verifier, whose hardness is, at this time, only conjectured. Unfortunately, if we replace the Unique Game with an outer verifier that is known to be NP-hard, say the Raz Verifier, the entire analysis breaks down.

One result that slightly deviates from this scheme is the hardness result for Vertex Cover due to Dinur and Safra [11]. In this result, a certain new outer verifier is constructed on top of the Raz verifier, and when combined with the long code, it leads to a 1.36 inapproximability result for Vertex Cover. This improves on a fairly long-standing $\frac{7}{6} - \varepsilon$ inapproximability result by Håstad [3], but still short of the possibly tight lower bound of $2 - \varepsilon$.

Dinur and Safra's result shows, roughly, that given an N-vertex graph that has an independent set of size $\frac{N}{3}$, it is NP-hard to find an independent set of size $\frac{N}{9}$. In this paper, we strengthen their result showing that even when the graph has three disjoint independent sets of size $\frac{(1-\varepsilon)N}{3}$ each, i.e. the graph is almost 3-colorable, it is NP-hard to find an independent set of size $\frac{N}{9}$. In particular, given an almost 3-colorable graph, it is NP-hard to almost color it with 8 colors. We follow a similar framework as in [11]; the outer verifier is identical to theirs, whereas the inner verifier requires some new ideas.

Almost Coloring

We now turn to describe the almost coloring problem. The (approximate) graph coloring problem is a classical optimization problem defined as follows. For an undirected graph $G = (V, E)$, let $\chi(G)$ be its chromatic number, i.e., the smallest number of colors needed to color the vertices of G without monochromatic edges. Then the graph coloring problem is defined as follows.

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

- $\chi(G) \leq q$
- $\chi(G) \geq Q$.

This problem also has a natural search variant, which can be stated as follows: Given a graph G with $\chi(G) \leq q$, color G with less than Q colors. It is easy to see that the search variant is not easier than the original decision variant, and hence for the purpose of showing hardness results it is enough to consider the decision variant.

Due to a technical limitation we do not prove NP-hardness for this problem, but rather for a close variant, called ALMOSTCOLORING$_\epsilon(q, Q)$.

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

- $\chi_\varepsilon(G) \leq q$
- $IS(G) \leq 1/Q$.

Here $IS(G)$ denotes the maximum fractional size of an independent set in the graph G. In this variant, the condition $\chi(G) \leq q$ is relaxed to $\chi_\varepsilon(G) \leq q$ which means that one can color all but an ϵ fraction of the vertices using q colors such that no edge is monochromatic. Note that the condition $IS(G) \leq 1/Q$ implies that $\chi_\varepsilon(G) \geq Q$ for a sufficiently small ε. Our main theorem is,

Theorem 1. *For any $\varepsilon > 0$ and any $q \geq 3$, the problem* ALMOSTCOLORING$_\epsilon(q, q^2)$ *is NP-hard.*

Known Results

It is easy to solve the problem COLORING$(2, Q)$ for any $Q \geq 3$ in polynomial time as it amounts to checking bipartiteness. The situation with COLORING(q, Q) for $3 \leq q < Q$

is much more interesting, as there is a huge gap between the value of Q for which an efficient algorithm is known and that for which a hardness result exists.

Indeed, even for $q = 3$, the best known algorithm colors a 3-colorable graph with n^α colors for a constant value of $\alpha \approx 0.207$ [12], [13], [14], [15], [16], [14]. The situation with COLORING(q, Q) for small values $q \geq 4$ is similar. The best known algorithm, due to Halperin, Nathaniel, and Zwick [17], solves COLORING(q, Q) for $Q = \tilde{O}(n^{\alpha_q})$ where $0 < \alpha_q < 1$ is some constant depending on q. For example, $\alpha_4 \approx 0.37$. As far as we know, the algorithmic results are the same even for the ALMOSTCOLORING$_\epsilon(q, Q)$ variant.

In contrast, the strongest known NP-hardness result for $q = 3$ shows that the problem is NP-hard for $Q = 5$ [18], [19], and in general for $q \geq 3$, the problem is NP-hard for $Q = q + 2\lfloor \frac{q}{3} \rfloor$ [18]. When q is sufficiently large, Khot [20] shows that for $Q = q^{\frac{\log q}{25}}$, COLORING$(q, Q)$ is NP-hard (see also Fürer [21] for a weaker result). Assuming variants of the Unique Games Conjecture, the problem COLORING(q, Q) was shown to be NP-hard for all $3 \leq q < Q$ in [22]. Assuming the UGC, the problem ALMOSTCOLORING$_\epsilon(2, Q)$ was shown to be NP-hard for all $Q \geq 3$ in [23].

Techniques

As we mentioned, our result follows the framework in [11], using the same outer verifier, but using a new long code based construction. Here we give an overview of the latter and state how it differs from that of [11].

The main combinatorial construction in [11] is a graph $G(\mathcal{P}, E)$, where $\mathcal{P} := \{F | F \subseteq [n]\}$ and $(F_1, F_2) \in E$ if and only of $F_1 \cap F_2 = \emptyset$. The vertices of the graph, i.e. all subsets of $[n]$, are equipped with a p-biased measure for some $p < \frac{1}{2}$:

$$\mu_p(F) := p^{|F|}(1-p)^{n-|F|}.$$

An independent set in this graph corresponds to a family $\mathcal{F} \subseteq \mathcal{P}$ that is pairwise intersecting, i.e. $\forall F_1, F_2 \in \mathcal{F}, F_1 \cap F_2 \neq \emptyset$. A maximal independent set corresponds to a monotone[1] family. There are two properties that are crucial: (1) The maximum weight of a pairwise intersecting family is p, achieved precisely by the *dictatorship* family $\mathcal{F}_{i_0} = \{F | i_0 \in F\}$ for any fixed $i_0 \in [n]$. (2) By an application of Russo's and Friedgut's theorems, every monotone family is essentially a *junta* under a slight perturbation of the the bias parameter. To be precise, the weight of a monotone family $\mu_p(\mathcal{F})$ is an increasing function of p and for any given range $(p_0, p_0 + \varepsilon)$, there exists $p' \in (p_0, p_0 + \varepsilon)$ such that w.r.t. the p'-biased measure, $\mathcal{F}$ is *close* to another family $\mathcal{F}'$ that depends only on a constant number $C = C(p_0, \varepsilon)$ of elements in $[n]$.

[1] $\mathcal{F}$ is monotone if $F \in \mathcal{F}$ and $F \subseteq F'$ implies that $F' \in \mathcal{F}$.

The analogous construction in this paper considers the graph $G(\mathcal{P}, E)$, where $\mathcal{P} := \{F | F \in \{*, 1, \ldots, q\}^n\}$ and $(F_1, F_2) \in E$ if and only if $\forall i \in [n], (F_1(i), F_2(i)) \notin \{(1,1), (2,2), \ldots, (q,q)\}$, where $F(i)$ denotes the i^{th} co-ordinate of F. The symbol $*$ is thought of as a dummy symbol. The vertices of the graph are equipped with a biased measure that assigns a probability mass of $1-p$ to $*$ and a probability mass of $\frac{p}{q}$ to each of the symbols $\{1, \ldots, q\}$.

$$\mu_p(F) := \prod_{i=1}^{n} (p/q)^{\mathbf{1}_{F(i) \neq *}} (1-p)^{\mathbf{1}_{F(i) = *}}.$$

The analogous properties are: (1) For each co-ordinate $i_0 \in [n]$, one can color the graph $\mathcal{P}$ with q colors except leaving out $1-p$ fraction of its vertices. The coloring is obtained by partitioning the vertices into $q+1$ sets according to their i_0 co-ordinate, and leaving out $1-p$ fraction of the vertices where the relevant co-ordinate equals $*$. The parameter $1-p$ is thought of as very small. (2) We again want to claim that every monotone family is essentially a *junta* under a slight perturbation of the the bias parameter p. Towards this end, we need to define an appropriate notion of monotonicity, and re-prove Russo's and Friedgut's Theorems. It suffices to define a partial order on the set $\{*, 1, \ldots, q\}$ where $\forall i \in [q], * \leq i$ and the symbols $\{1, \ldots, q\}$ are incomparable. Then for $F_1, F_2 \in \{*, 1, \ldots, q\}^n$, $F_1 \leq F_2$ if and only if $\forall i \in [n], F_1(i) \leq F_2(i)$. With respect to this definition, it turns out that every maximal independent set is monotone and Russo's and Friedgut's theorems can be re-proved appropriately.

Theorem 1 is proved by combining the long code based construction described above with the outer verifier, along similar lines as in [11], albeit with an even more complicated analysis.

Overview of the Paper

In Section II we give definitions and useful results in our setting of $\{*, 1, \ldots, q\}^n$. In Section III we give the reduction from an NP-hard problem, $hIS(r, \varepsilon, h)$, to ALMOSTCOLORING$_\epsilon(q, q^2)$. In Section IV we prove the completeness of the reduction. In Section V we give the proof of soundness.

II. PRELIMINARIES

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

Let $G_{q,p}[n]$ be the following weighted graph: The vertices are elements in $\{*, 1, \ldots q\}^n$ with measure μ_p, the product measure that on each coordinate assigns $1-p$ to $*$ and $\frac{p}{q}$ to each $i \in [q]$. The edges of $G_{q,p}[n]$ are defined as follows:

$$(F_1, F_2) \in E \text{ if } \forall i,$$

$$(F_1(i), F_2(i)) \notin \{(1,1), (2,2), \ldots (q,q)\}.$$

B. Definitions

- Monotonicity: A family $\mathcal{F} \subseteq \{*, 1, \ldots, q\}^n$ is *monotone* if $F \in \mathcal{F}$ implies $F' \in \mathcal{F}$ for any F' that can be obtained from F by changing a $*$ coordinate to some $k \in [q]$. Note that a maximal independent set $\mathcal{F} \subseteq G_{q,p}[n]$ is a monotone subset of $\{*, 1, \ldots, q\}^n$.
- We will call an element $F \in \{*, 1, \ldots, q\}^n$ a coloring of $[n]$.
- Two colorings F_1, F_2 of $[n]$ *agree* on $i \in [n]$ if $F_1(i) = F_2(i) \in [q]$ (in particular the common co-ordinate is not $*$).
- A family $\mathcal{F}$ of colorings of $[n]$ is *agreeing* if for all $F_1, F_2 \in \mathcal{F}$, there exists $i \in [n]$ so that F_1, F_2 agree on i.
- Two families of colorings $\mathcal{F}_1, \mathcal{F}_2$ are *cross-agreeing* if for all $F_1 \in \mathcal{F}_1, F_2 \in \mathcal{F}_2$, there exists $i \in [n]$ so that F_1, F_2 agree on i.
- A family $\mathcal{F}$ is 2-*agreeing* if for all $F_1, F_2 \in \mathcal{F}$ there exist $i \neq j \in [n]$ so that F_1, F_2 agree on both i and j.
- Let $\mathcal{F} \subseteq \{*, 1, \ldots, q\}^n$. A (δ, p)-core for $\mathcal{F}$ is set $C \subseteq [n]$ so that there is a family $\mathcal{F}'$ that depends only on C such that $\mu_p(\mathcal{F} \triangle \mathcal{F}') < \delta$. I.e., changing the coloring of coordinates in $[n] \backslash C$ does not affect whether a coloring is in $\mathcal{F}'$ or not.
- A core-family: for $C \subseteq [n]$ and $\mathcal{F} \subseteq \{*, 1, \ldots, q\}^n$, define $[\mathcal{F}]_C^t =$

$$\left\{F \in \{*, 1, \ldots, q\}^C : \mathbb{P}_{F' \in \mu_p^{\overline{C}}}[(F, F') \in \mathcal{F}] > t\right\}.$$

  where $0 < t < 1$ and (F, F') is the combined coloring of $[n]$ that equals F on C and equals F' on $\overline{C} = [n] \backslash C$. Note that $[\mathcal{F}]_C^{\frac{1}{2}}$ is the family that best approximates $\mathcal{F}$ on C.
- Define the *influence* of a coordinate i on $\mathcal{F}$ as follows:

$$Inf_i^p(\mathcal{F}) = \mu_p \left(\{F : F|_{i=*} \notin \mathcal{F} \text{ but } F|_{i=r} \in \mathcal{F} \text{ for some } r \in [q]\}\right).$$

  Here $F|_{i=*}$ refers to the coloring that agrees with F on all coordinates but assigns $*$ to coordinate i, and the same applies to the notation $F|_{i=r}$.
- The *average sensitivity* of $\mathcal{F}$ at p is

$$as_p(\mathcal{F}) = \sum_i Inf_i^p(\mathcal{F}).$$

C. Useful Results

Proposition 1. *Let $\mathcal{F} \subseteq G_{q,p}[n]$ be monotone. Then $\mu_p(\mathcal{F})$ is increasing in p.*

Proof: See the proof of Lemma 1 below and note that in our expression for $\frac{d\mu_p(\mathcal{F})}{dp}$ we can group together non-negative terms when $\mathcal{F}$ is monotone. ■

Proposition 2. *(Cross-agreeing families can't be too large). Let $\mathcal{F}_1, \mathcal{F}_2 \subseteq \{*, 1, \ldots, q\}^n$ be monotone with $\mu_p(\mathcal{F}_1) + \mu_p(\mathcal{F}_2) > 1$. Then there exist $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$ s.t. for all $i \in [n]$, F_1, F_2 do not agree on i. Moreover, F_1 and F_2 can be taken to be in $\{1, \ldots, q\}^n$.*

Proof: Since μ_p is an increasing function of p for monotone sets, we may as well take $p = 1$. Note that this amounts to restricting the families $\mathcal{F}_1$ and $\mathcal{F}_2$ to the colorings in $\{1, \ldots q\}^n$ under uniform probability measure $\mu(\cdot)$, since the $*$ symbol now receives zero probability mass. Assume such and F_1, F_2 do not exist, i.e. $\mathcal{F}_1, \mathcal{F}_2$ are cross-agreeing. Define a bijection $\phi : \{1, \ldots q\}^n \to \{1, \ldots q\}^n$ by $\phi(F)(i) = F(i) + 1 \pmod q \ \forall i \in [n]$, i.e. shift each coordinate of F by 1 modulo q. Since $\mathcal{F}_1$ and $\mathcal{F}_2$ are cross-agreeing, for every $F \in \mathcal{F}_1$, $\phi(F) \notin \mathcal{F}_2$. Thus the fraction of the colorings ruled out of $\mathcal{F}_2$ is $\geq \mu(\mathcal{F}_1)$ and so $\mu(\mathcal{F}_2) \leq 1 - \mu(\mathcal{F}_1)$. But this contradicts the assumption that $\mu(\mathcal{F}_1) + \mu(\mathcal{F}_2) > 1$. ■

Russo's Lemma [24]

Lemma 1. *Let $\mathcal{F} \subseteq \{*, 1, \ldots, q\}^n$ be monotone. Then*

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

Proof: For $F \in \{*, 1, \ldots, q\}^n$ we have

$$\mu_p(F) = \prod_{i \in [n]} \mu_p^i(F),$$

where $\mu_p^i(F) = p/q$ if $F(i) \in [q]$ and $1 - p$ if $F(i) = *$.

$$\frac{d\mu_p(F)}{dp} = \sum_{i \in [n]} (1/q)^{\mathbf{1}_{F(i) \in [q]}} \cdot (-1)^{\mathbf{1}_{F(i)=*}} \cdot \prod_{j \neq i} \mu_p^j(F),$$

and summing over $F \in \mathcal{F}$ and exchanging the order of summation

$$\frac{d\mu_p(\mathcal{F})}{dp} = \sum_{i \in [n]} \sum_{F \in \mathcal{F}} (1/q)^{\mathbf{1}_{F(i) \in [q]}} \cdot (-1)^{\mathbf{1}_{F(i)=*}} \cdot \prod_{j \neq i} \mu_p^j(F).$$

Now fix a coordinate i and a coloring $\bar{F}$ of $[n] \setminus \{i\}$. By monotonicity, the contribution to the second sum is non-zero only if $(\bar{F}, *) \notin \mathcal{F}$. In that case the contribution is

$$\mu_p^{[n] \backslash i}(\bar{F}) \cdot \frac{1}{q} \cdot |\{k \in [q] : (\bar{F}, k) \in \mathcal{F}\}|.$$

If each non-* extension of $\bar{F}$ is in $\mathcal{F}$, the contribution is 1, matching the contribution in the definition of $Inf_i^p(\mathcal{F})$. When the contribution in the influence is 1, the contribution here can be as low as $\frac{1}{q}$ leading to the first inequality in the Lemma. ■

Friedgut's Theorem [25]

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

Erdős-Rado Theorem [26]

Theorem 3. *There is a function $\Gamma_*(k,d)$ so that for any $\mathcal{F} \subseteq \binom{R}{k}$, if $|\mathcal{F}| \geq \Gamma_*(k,d)$, there are d distinct set $F_1, \ldots F_d \in \mathcal{F}$ so that if $\Delta = F_1 \cap \cdots \cap F_d$, then the sets $F_i \setminus \Delta$ are pairwise disjoint. This also holds when the subsets can be of size at most k.*

Bourgain-Kahn-Kalai-Katznelson-Linial [27]

Theorem 4.

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

Maximum Size of a 2-Agreeing Family

Lemma 2. *Let $\mathcal{F} \subseteq \{*, 1, \ldots, q\}^n$ be monotone with $\mu_p(\mathcal{F}) \geq \frac{1}{q^2}$, where $p < 1$. Then $\mathcal{F}$ is not 2-agreeing. Specifically, there are $F_1, F_2 \in \mathcal{F} \cap \{1, \ldots, q\}^n$ that agree on at most one co-ordinate.*

Proof: Since $\mathcal{F}$ is monotone, we can raise p to 1 and have $\mu_1(\mathcal{F}) > \frac{1}{q^2}$. The maximal size of a 2-agreeing family in $[q]^n$ with the uniform measure is $\frac{1}{q^2}$ by Ahlswede and Khachatrian [28], so $\mathcal{F}$ cannot be 2-agreeing. ■

Proposition 3. *Let $\mathcal{F} \subseteq \{*, 1, \ldots, q\}^n$, and $C \subseteq [n]$.*

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

Proof: (1) If $F_1 \in \{*, 1, \ldots, q\}^C$ is a monotonically above[2] $F \in [\mathcal{F}]^{\frac{3}{4}}_C$, then each coloring of $\overline{C}$ that extends F to a coloring in $\mathcal{F}$ also works for F_1. (2) By contradiction. Let $F_1^C, F_2^C \in [\mathcal{F}]^{\frac{3}{4}}_C$ so that they disagree on all $i \in C$. If $\mathcal{F}_1', \mathcal{F}_2'$ are the respective sets of colorings of $\overline{C}$ that extend F_1^C, F_2^C to colorings in $\mathcal{F}$, then we know $\mu_p^{[n]\setminus C}(\mathcal{F}_1'), \mu_p^{[n]\setminus C}(\mathcal{F}_2') > \frac{3}{4}$, and so by Proposition 2 they cannot be agreeing. Combining the non-agreeing pair with F_1^C and F_2^C gives a non-agreeing pair of colorings in $\mathcal{F}$, a contradiction. ■

Proposition 4. *[Lemma 5.1, [11]]*

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

III. The Reduction

In this section, we give a reduction that proves Theorem 1. The reduction is from a certain independent set problem on co-partite graphs, denoted as $hIS(r, \varepsilon, h)$.

A. The $hIS(r, \varepsilon, h)$ Problem

An (m, r)-co-partite graph $G(V, E)$ has the following structure:

- $V = M \times R$, $|M| = m$, $|R| = r$.
- For each $i \in M$, the set of of vertices $\{i\} \times R$ is a clique.

[2]Meaning F_1 agrees with F with the possible exception that $*$'s in F are changed to $\{1, \ldots q\}$.

Let $IS(G)$ denote the size of the maximum independent set in a graph G. Note that for an (m, r)-co-partite graph G, $IS(G) \leq m$. For an integer $h \geq 2$, let $IS_h(G)$ denote the maximum size of a set of vertices in G that contains no clique of size h (thus $IS(G) = IS_2(G)$). Given an (m, r)-co-partite graph G, the $hIS(r, \varepsilon, h)$ problem is to distinguish between:

- YES case: $IS(G) = m$.
- NO case: $IS_h(G) \leq \varepsilon m$.

It follows from the PCP Theorem and Raz's Parallel Repetition Theorem that the following problem is NP-hard (this is essentially the so-called Raz Verifier):

Theorem 5. *For any $h, \varepsilon > 0$, the problem $hIS(r, \varepsilon, h)$ is NP-hard, as long as $r \geq \left(\frac{h}{\varepsilon}\right)^c$ for some constant c.*

Given an (m, r)-co-partite graph G, we construct a graph $G^q_{\mathcal{B}}$ below, thus proving Theorem 1. Note that the graph we construct is weighted, but using a standard transformation, the same result holds for un-weighted graphs.

Theorem 6. *For any $\varepsilon > 0$ and an integer $q \geq 3$, for large enough h and small enough ε_0 (see parameter setting below), given an (m, r)-co-partite graph $G = (V = M \times R, E)$, one can construct a weighted graph $G^q_{\mathcal{B}}$ in polynomial time such that:*

- *Total weight of all the vertices in $G^q_{\mathcal{B}}$ is 1.*
- *(Completeness:) $IS(G) = m$ implies that there exists a q-colorable subset of vertices $V' \subseteq G^q_{\mathcal{B}}$ with weight $1 - 2\varepsilon$.*
- *(Soundness:) $IS_h(G) < \varepsilon_0 m \Rightarrow IS(G^q_{\mathcal{B}}) < \frac{1}{q^2}$, where $IS(G^q_{\mathcal{B}})$ is the maximum weight of an independent set in $G^q_{\mathcal{B}}$.*

B. Parameter Setting

Given $\varepsilon > 0$ and $q \geq 3$ we define the following (long list of) parameters. It is perhaps best to skip reading the list for now and refer to it whenever the parameters are first encountered.

- $p = 1 - \varepsilon$
- $\varepsilon_1 = \frac{\varepsilon}{2q^3}$.
- $h_0 = \sup_{p' \in [1-\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{4}]} \Gamma(p', \frac{1}{16}\varepsilon_1, \frac{8q}{\varepsilon})$, where Γ is the function from Theorem 2.
- $\eta = \frac{1}{16h_0}(p/q)^{5h_0}$.
- $h_1 = h_0 + \lceil \frac{4q}{\varepsilon\eta} \rceil$
- $h_s = 1 + (q+1)^{2h_0} \cdot \sum_{k=0}^{h_0} \binom{h_1}{k}$.
- $h = \Gamma_*(h_1, h_s)$ where Γ_* is the function from Theorem 3.
- $\varepsilon_0 = \frac{1}{32}\varepsilon_1$.
- $l_T = \max(4 \ln \frac{2}{\varepsilon}, (h_1)^2)$.

C. Constructing $G^q_{\mathcal{B}}$

The initial graph $G(V, E)$: : The vertex set of $G(V, E)$ is $V = M \times R$, where $|M| = m$ and $|R| = r$. Let $l = 2l_T \cdot r$. Consider the family $\mathcal{B}$ of all subsets of V of size l:

$$\mathcal{B} = \binom{V}{l} = \{B \subseteq V : |B| = l\}.$$

Each $B \in \mathcal{B}$ will be called a *block*. An independent set in G is thought of as an assignment $\sigma : V \mapsto \{T, F\}$ that assigns T to a vertex if and only if the vertex is in the independent set. We will deal with restrictions of this (supposed) global assignment to the blocks. In the completeness case, there is an independent set of size m, and thus for the corresponding assignment σ, $|\sigma^{-1}(T)| = m = \frac{1}{r}|V|$. For a *random* block B of size $l = 2l_T \cdot r$, the restriction $\sigma|_B$ takes $2l_T$ T-values in expectation, and hence w.h.p. at least l_T T-values. We therefore let the set of *block-assignments* R_B to the block B as the the set of all $\{T, F\}$-assignments that have at least l_T many T values.

$$R_B := \{a : B \to \{T, F\} : |a^{-1}(T)| \geq l_T\}.$$

The intermediate graph $G_{\mathcal{B}}(V_{\mathcal{B}}, E_{\mathcal{B}})$: : We now construct an intermediate graph $G_{\mathcal{B}}(V_{\mathcal{B}}, E_{\mathcal{B}})$ whose vertex set is defined as

$$V_{\mathcal{B}} = \bigcup_{B \in \mathcal{B}} \{B\} \times R_B.$$

In words, for each block $B \in \mathcal{B}$, there is a cluster of vertices $\{B\} \times R_B$. We let this cluster be a clique. For two $(l-1)$-wise intersecting blocks B_1, B_2, $\hat{B} := B_1 \cap B_2, |\hat{B}| = l - 1$, we introduce edges between the two corresponding clusters in the following manner. Suppose $B_1 = \hat{B} \cup \{v_1\}$ and $B_2 = \hat{B} \cup \{v_2\}$ (v_1, v_2 will be referred to as "pivots"). Then $((B_1, a_1), (B_2, a_2)) \in E_{\mathcal{B}}$ if and only if

- $(v_1, v_2) \in E$ and,
- Either $a_1|_{\hat{B}} \neq a_2|_{\hat{B}}$ or $a_1(v_1) = a_2(v_2) = T$.

In words, there is an edge between vertices $(B_1, a_1), (B_2, a_2) \in V_{\mathcal{B}}$ if and only if B_1 and B_2 are $(l-1)$-wise intersecting, there is an edge in G between the two pivot vertices, and either the assignments a_1, a_2 disagree on the intersection, or assign T to both v_1, v_2 respectively.

Proposition 5. *$IS(G) = m \Rightarrow IS(G_{\mathcal{B}}) \geq m'(1-\varepsilon)$ where $m' = |\mathcal{B}|$.*

Proof: Let $I \subseteq V$ be an independent set of size m in $G(V, E)$. Let $\sigma : V \mapsto \{T, F\}$ be an assignment where $\sigma(v) = T$ if and only if $v \in I$. For each block $B \in \mathcal{B}$, let $\sigma(B)$ denote the restriction of assignment σ to the block B. We claim that the set $I_{\mathcal{B}} := \{(B, \sigma(B)) : B \in \mathcal{B}\}$ is an independent set in $G_{\mathcal{B}}(V_{\mathcal{B}}, E_{\mathcal{B}})$. This is because, for any two vertices $(B_1, \sigma(B_1)), (B_2, \sigma(B_2)) \in I_{\mathcal{B}}$, there cannot be an edge between them, since the assignments $\sigma(B_1)$ and $\sigma(B_2)$ are consistent on the intersection of the two blocks, and moreover, if v_1, v_2 are the pivot vertices, $(v_1, v_2) \in E$, and since I is an independent set, it cannot be the case that both $\sigma(v_1) = \sigma(v_2) = T$. The only caveat is that for some of the blocks $B \in \mathcal{B}$, the assignment $\sigma(B)$ might not have the required number (= l_T) of T's. The fraction of such blocks is at most ε by the choice of parameters; these *bad blocks* may be ignored, giving an independent set of size $(1-\varepsilon)|\mathcal{B}|$ in the graph $G_{\mathcal{B}}$. ∎

The final graph $G^q_{\mathcal{B}}$:: We are now ready to construct the final graph $G^q_{\mathcal{B}} \langle V^q_{\mathcal{B}}, E^q_{\mathcal{B}}, \Lambda \rangle$. $G^q_{\mathcal{B}}$ has the same clusters as $G_{\mathcal{B}}$ but each cluster instead of being a clique, is a copy of $G_{q,p}[n]$, with $n = |R_B|$. The cluster of vertices corresponding to a block $B \in \mathcal{B}$ is denoted $V^q_{\mathcal{B}}[B]$; its vertices are all colorings of R_B, i.e.

$$V^q_{\mathcal{B}}[B] := \{(B, F) \mid F \in \{*, 1, \ldots, q\}^{R_B}\}, \quad \text{and}$$

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

We will often abuse the notation by denoting a vertex (B, F) by F when the corresponding block B is clear from the context.

Assigning a Weight Function: Let μ_p be the distribution on $V^q_{\mathcal{B}}[B] = \{F \in \{*, 1, \ldots, q\}^{R_B}\}$. The probability measure Λ on $V^q_{\mathcal{B}}$ assigns equal measure to each cluster, and within a cluster, the vertices are weighted according to μ_p. Specifically, for $F \in V^q_{\mathcal{B}}[B]$,

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

Edges: The edges within a cluster $V^q_{\mathcal{B}}[B]$ are already determined by the structure of $G_{q,p}[n]$. We now describe[3] edges between different clusters $V^q_{\mathcal{B}}[B_1]$ and $V^q_{\mathcal{B}}[B_2]$.

$$\begin{aligned} E^q_{\mathcal{B}} = \quad & \{(F_1, F_2) \in V^q_{\mathcal{B}}[B_1] \times V^q_{\mathcal{B}}[B_2] : \\ & F_1^{-1}(i) \times F_2^{-1}(i) \subseteq E_{\mathcal{B}}, \forall i \in [q]\}, \end{aligned}$$

where $E_{\mathcal{B}}$ is the edge set of the intermediate graph $G_{\mathcal{B}}$.

In words, (F_1, F_2) is an edge if and only if for any pair of block assignments $a_1 \in R_{B_1}, a_2 \in R_{B_2}$, $F_1(a_1) = F_2(a_2) \in [q] \implies (a_1, a_2) \in E_{\mathcal{B}}$. Since this definition is somewhat complicated, let us state its contra-positive as well. (F_1, F_2) is *not* an edge if and only if there exist $a_1 \in R_{B_1}, a_2 \in R_{B_2}$ such that $(a_1, a_2) \notin E_{\mathcal{B}}$ and $F_1(a_1) = F_2(a_2) \in [q]$. Recall that the first condition, i.e. $(a_1, a_2) \notin E_{\mathcal{B}}$, is the same as saying:

- a_1 and a_2 are consistent on the $(l-1)$-wise intersection of the two blocks B_1 and B_2.
- If v_1, v_2 are the pivot vertices with $(v_1, v_2) \in E$, then $(a_1(v_1), a_2(v_2)) \neq (T, T)$.

Proposition 6. *(Maximal Independent sets in $G^q_{\mathcal{B}}$ are monotone). Let $\mathcal{I}$ be an independent set in $G^q_{\mathcal{B}}$. If $F \in \mathcal{I} \cap V^q_{\mathcal{B}}[B]$,*

[3]It is implicit here that B_1 and B_2 are $(l-1)$-wise intersecting and there is an edge in G between the pivot vertices.

and F' is monotonically above F, then $\mathcal{I} \cup \{F'\}$ is also an independent set.

Proof: For all $F_2 \in \mathcal{I} \cap V^q_{\mathcal{B}}[B_2]$, since (F, F_2) is not an edge, there must exist $a \in R_B$, $a_2 \in R_{B_2}$, with $(a, a_2) \notin E_{\mathcal{B}}$ s.t. $F(a) = F_2(a_2) \in [q]$. By our definition of monotonicity, $F'(a) = F(a)$, hence the previous condition applies to F' as well, implying that (F', F_2) is not an edge either, and we can safely add F' to the independent set $\mathcal{I}$. ∎

IV. Completeness

Lemma 3. *If $IS(G) = m$, then there exist disjoint independent sets $I_1, \ldots I_q$ in $G^q_{\mathcal{B}}$ such that*

$$\Lambda(I_1 \cup \cdots \cup I_q) \geq 1 - 2\varepsilon.$$

Proof: By Proposition 5, if $IS(G) = m$, then $IS(G_{\mathcal{B}}) \geq m'(1-\varepsilon)$ where $m' = |\mathcal{B}|$. Let $\mathcal{I}_{\mathcal{B}}$ be an independent set in $G_{\mathcal{B}}$ of size $m'(1-\varepsilon)$. It necessarily holds that $\mathcal{I}_{\mathcal{B}}$ is of the form:

$$\mathcal{I}_{\mathcal{B}} = \{(B, a) | a \in R_B\},$$

containing exactly one vertex from $(1-\varepsilon)m'$ of the clusters. For each $j \in [q]$, let

$$\begin{aligned} \mathcal{I}_j = \quad & \{(B, F) \in G^q_{\mathcal{B}} : \exists a \in R_B \text{ s.t.} \\ & (B, a) \in \mathcal{I}_{\mathcal{B}} \text{ and } F(a) = j\}. \end{aligned}$$

We claim that $\mathcal{I}_j$ is an independent set in $G^q_{\mathcal{B}}$. Indeed, for any two vertices $(B_1, F_1), (B_2, F_2) \in \mathcal{I}_j$, there are vertices $(B_1, a_1), (B_2, a_2) \in \mathcal{I}_{\mathcal{B}}$ with $(a_1, a_2) \notin E_{\mathcal{B}}$, and $F_1(a_1) = F_2(a_2) = j$. Hence (F_1, F_2) is not an edge. $\mathcal{I}_j$ and $\mathcal{I}_k$ are disjoint for $j \neq k$ since they assign a different labels to block assignments $(B, a) \in I_{\mathcal{B}}$. Finally, $\Lambda(\mathcal{I}_j) \geq (1-\varepsilon)\frac{1-\varepsilon}{q} \geq \frac{1-2\varepsilon}{q}$ since the weight in one cluster of all F's that assign $j \in [q]$ to a chosen block assignment (B, a) is $\frac{1-\varepsilon}{q}$. ∎

V. Proof of Soundness

Lemma 4. *If $IS(G^q_{\mathcal{B}}) \geq \frac{1}{q^2} \Rightarrow IS_h(G) \geq \varepsilon_0 m$.*

Let $\mathcal{I}$ be an independent set in $G^q_{\mathcal{B}}$ with $\Lambda(\mathcal{I}) \geq \frac{1}{q^2}$ and let $\mathcal{I}[B]$ denote $\mathcal{I} \cap V^q_{\mathcal{B}}[B]$. We assume without loss of generality that $\mathcal{I}$ is maximal and hence each $\mathcal{I}[B]$ is monotone.

First we raise p from $1-\varepsilon$ to $1-\frac{\varepsilon}{2}$. By Theorem 4 and Lemma 1, $\mu(\mathcal{I})$ must increase by $\frac{\varepsilon}{2} \cdot \frac{1}{q} \cdot \frac{1}{q^2}$. So we have $\Lambda_{1-\frac{\varepsilon}{2}}(\mathcal{I}) \geq \frac{1}{q^2} + \frac{\varepsilon}{2q^3} = \frac{1}{q^2} + \varepsilon_1$.

Some terminology we will use (much of it already defined):

- $F \in V^q_{\mathcal{B}}[B]$ is referred to as a *coloring* of the block assignments in R_B, i.e. $F \in \{*, 1, \ldots, q\}^{R_B}$ or alternately $F : R_B \to \{*, 1, \ldots, q\}$.
- A family of colorings $\mathcal{F}$ is *monotone* if it is a monotone as a subset of $\{*, 1, \ldots, q\}^n$.
- Two colorings F_1, F_2 *agree* on a block assignment a if $F_1(a) = F_2(a) \in [q]$.
- A block assignment $a \in R_B$ is *distinguished* by a family $\mathcal{F}$ of colorings of R_B if there exist $F_1, F_2 \in \mathcal{F}$ such that F_1 and F_2 agree on a and do not agree on any other block assignment in R_B. We also say that $\mathcal{F}$ has a distinguished element a.

Proposition 7. *For all $B \in \mathcal{B}$, $\mathcal{I}[B]$ is monotone and agreeing.*

Proof: $\mathcal{I}[B]$ is monotone by maximality. For any F_1, F_2, there exists an $a \in R_B$ so that F_1, F_2 agree on a since there are edges in $G^q_{\mathcal{B}}[B]$ between colorings which do not agree on any block assignment. ∎

A. Selecting $\mathcal{B}' \subseteq \mathcal{B}$ and $\hat{B} \in \binom{V}{l-1}$

We want to find a subset of blocks $\mathcal{B}' \subseteq \mathcal{B}$ and an $(l-1)$ sub-block $\hat{B}$ that satisfy certain properties.

Selection of $\mathcal{B}'$:

Lemma 5. *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, $Core[B] \subseteq R_B$, of size $|Core[B]| \leq h_0$.*
2) *The core-family $\mathcal{CF}_B := [\mathcal{I}[B]]^{\frac{3}{4}}_{Core[B]}$ has a distinguished element $\dot{a}[B] \in Core[B]$.*

Proof: First we define a set of blocks for which $\mathcal{I}$ has large weight:

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

By averaging, $|\tilde{\mathcal{B}}| \geq \frac{1}{2}\varepsilon_1 \cdot |\mathcal{B}|$, and since $\mu_p(\mathcal{I}[B])$ is an increasing function of p, we have $\mu_{p'}(\mathcal{I}[B]) > \frac{1}{q^2} + \frac{1}{2}\varepsilon_1$ for all $B \in \tilde{\mathcal{B}}$, $p' \in [1-\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{4}]$. Now define ($\mathcal{B}'$ depends on p' which will be chosen judiciously):

$$\mathcal{B}' = \left\{ B \in \tilde{\mathcal{B}} : as_{p'}(\mathcal{I}[B]) \leq \frac{8q}{\varepsilon} \right\}.$$

Proposition 8. *There exists $p' \in [1-\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{4}]$ such that*

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

Proof: We use Lemma 1. Consider the sum,

$$|\tilde{\mathcal{B}}|^{-1} \cdot \sum_{B \in \tilde{\mathcal{B}}} \frac{d\mu_{p'}}{dp'}(\mathcal{I}[B]) \geq |\tilde{\mathcal{B}}|^{-1} \cdot \frac{1}{q} \cdot \sum_{B \in \tilde{\mathcal{B}}} as_{p'}(\mathcal{I}[B]).$$

There must be some $p' \in [1-\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{4}]$ for which the left hand side above is at most $\frac{4}{\varepsilon}$. Otherwise, integrating over p' from $1-\frac{\varepsilon}{2}$ to $1-\frac{\varepsilon}{4}$ would give that

$|\tilde{\mathcal{B}}|^{-1} \cdot \sum_{B\in\tilde{\mathcal{B}}} \mu_1(\mathcal{I}[B']) > 1$, which is impossible. By averaging, at least half of the $B \in \tilde{\mathcal{B}}$ have $as_{p'}(\mathcal{I}[B]) \leq \frac{8q}{\varepsilon}$ and so $|\mathcal{B}'| \geq \frac{1}{4}\varepsilon_1 \cdot |\mathcal{B}|$. ∎

Now we fix p' as above. Claim 1 of Lemma 5 follows directly from Theorem 2. To prove Claim 2, define for $B \in \mathcal{B}'$

$$\mathcal{CF}_B = [\mathcal{I}[B]]^{\frac{3}{4}}_{Core[B]}.$$

Note that by Proposition 4 and Proposition 3, we have $\mathcal{CF}_B$ is monotone, pairwise agreeing and

$$\mu_{p'}(\mathcal{CF}_B) > \mu_{p'}(\mathcal{I}[B]) - 4 \cdot \frac{\varepsilon_1}{16} > \frac{1}{q^2}.$$

But the maximal size of a 2-agreeing family is $\frac{1}{q^2}$, so there must be a distinguished element $\dot{a} \in Core[B]$: i.e. there are two colorings $F^\flat, F^\sharp \in \mathcal{CF}_B$ so that $F^\flat, F^\sharp$ agree only on $\dot{a}$. Moreover, both $F^\flat, F^\sharp$ may be taken to be in $\{1, \ldots, q\}^{Core[B]}$. ∎

Definition 1. *For $p', \mathcal{B}'$ as above and $B \in \mathcal{B}'$, the extended core is defined as:*

$$ECore[B] = Core[B] \cup \{a \in R_B : Inf_a^{p'}(\mathcal{I}[B]) \geq \eta\}.$$

Proposition 9. *(The extended core is not too large):*

$$|ECore[B]| \leq h_0 + \frac{as_{p'}(\mathcal{I}[B])}{\eta} \leq h_0 + \lceil \frac{4q}{\varepsilon\eta} \rceil = h_1.$$

Selection of $\hat{B}$:

Definition 2. *(Preservation:) Let $B \in \mathcal{B}'$ and $\hat{B} \subseteq B$, $|\hat{B}| = l-1$. Let $a|_{\hat{B}}$ be the restriction of a block assignment $a \in R_B$ to $\hat{B}$. $\hat{B}$ preserves B if there is no pair of block assignments $a_1 \neq a_2 \in ECore[B]$ with $a_1|_{\hat{B}} = a_2|_{\hat{B}}$.*

Proposition 10. *For all $B \in \mathcal{B}'$,*

$$|\{v \in B | B \setminus \{v\} \text{ does not preserve } B\}| < \frac{(h_1)^2}{2}.$$

Proof: Each pair of block assignments in $ECore[B]$ can cause at most one $\hat{B}$ to not preserve B, and for any $B \in \mathcal{B}'$, $|ECore[B]| \leq h_1$. So the number of $\hat{B}$ not preserving B is $\leq \binom{h_1}{2} < \frac{h_1^2}{2}$. ∎

Definition 3. *For any $(l-1)$-block $\hat{B}$, let $V_{\hat{B}} \subseteq V$ be*

$$V_{\hat{B}} = \Big\{v \in V \setminus \hat{B} : B = \hat{B} \cup \{v\} \in \mathcal{B}' \text{ and } \hat{B} \text{ preserves } B \text{ and } \dot{a}[B](v) = T\Big\}.$$

Proposition 11. *There exists $\hat{B} \in \binom{V}{l-1}$ with $|V_{\hat{B}}| \geq \varepsilon_0 \cdot m$.*

Proof: In the following calculation, the choice of $B, \hat{B}, v \in V \setminus \hat{B}$ is such that $\hat{B} \cup \{v\} = B$.

$$\mathbb{P}_{\hat{B},v}\left[v \in V_{\hat{B}}\right] \geq \frac{1}{4}\varepsilon_1 \cdot \mathbb{P}_{B,v}\left[v \in V_{\hat{B}} | B \in \mathcal{B}'\right] \geq \frac{1}{4}\varepsilon_1 \cdot \frac{1}{4r}.$$

The first inequality is from Proposition 8 and the second holds since for the distinguished assignment $\dot{a}[B] \in R_B$ there are at least $l_T = \frac{l}{2r}$ elements $v \in B$ so that $\dot{a}[B](v) = T$. Moreover, at most $\frac{h_1^2}{2}$ many v's may not preserve B and $\frac{l}{2r} - \frac{h_1^2}{2} \geq \frac{l}{4r}$. ∎

B. Main Soundness Lemma

We fix an $(l-1)$-block $\hat{B}$ as in Proposition 11. We are ready to prove the main soundness lemma, thereby completing the proof of Theorem 6.

Lemma 6. *The set $V_{\hat{B}}$ contains no clique of size h. In particular, since $|V_{\hat{B}}| \geq \varepsilon_0 \cdot m$, we have $IS_h(G) \geq \varepsilon_0 \cdot m$.*

Proof: Assume there is a clique, $v_1, \ldots v_h \in V_{\hat{B}}$. Let $B_i = \hat{B} \cup \{v_i\}$ and only these h blocks will be relevant henceforth. We will show that the set $\cup_{i=1}^h \mathcal{I}[B_i]$ is not an independent set, a contradiction. In particular, we will find $i_1, i_2 \in [h]$ s.t. $\mathcal{I}[B_{i_1}] \cup \mathcal{I}[B_{i_2}]$ is not an independent set. We begin with a few definitions:

- $R_{\hat{B}} = \{a : \hat{B} \to \{T, F\}\}$.
- Let $a|_{\hat{B}}$ be the restriction of a block assignment $a \in R_B$ to $R_{\hat{B}}$.
- For each block B_i, $i \in [h]$, we name the following important entities:
  - $\dot{a}_i = \dot{a}[B_i]$ is B_i's distinguished block assignment, with $\hat{a}_i = \dot{a}_i|_{\hat{B}}$.
  - $C_i = Core[B_i]$ and $\hat{C}_i = C_i|_{\hat{B}}$ (recall preservation property).
  - $E_i = ECore[B_i]$ and $\hat{E}_i = E_i|_{\hat{B}}$ (recall preservation property).
- Since $\dot{a}_i$ is distinguished, there are two colorings, $F_i^\flat, F_i^\sharp \in \mathcal{CF}_{B_i}$ with the only agreement between $F_i^\flat$ and $F_i^\sharp$ on $\dot{a}_i$. Let $\hat{F}_i^\flat, \hat{F}_i^\sharp$ be their restrictions to $\hat{B}$ which is well-defined because of preservation. Since $\hat{B}$ preserves each B_i, we have that $\hat{F}_i^\flat$ and $\hat{F}_i^\sharp$ agree only on $\dot{a}_i|_{\hat{B}}$.

C. Sunflower and Pigeonhole Principle

Proposition 12. *There exist $i_1 \neq i_2 \in [h]$ such that, for $\Delta = \hat{E}_{i_1} \cap \hat{E}_{i_2}$,*

1) $\hat{C}_{i_1} \cap \Delta = \hat{C}_{i_2} \cap \Delta$.
2) $\hat{F}_{i_1}^\flat$ *and* $\hat{F}_{i_2}^\flat$ *agree on* $\hat{C}_{i_1} \cap \Delta = \hat{C}_{i_2} \cap \Delta$.
3) $\hat{F}_{i_1}^\sharp$ *and* $\hat{F}_{i_2}^\sharp$ *agree on* $\hat{C}_{i_1} \cap \Delta = \hat{C}_{i_2} \cap \Delta$.

Proof: We apply Theorem 3. Here we set $R = R_{\hat{B}}$, $\mathcal{F} = \{\hat{E}_1, \ldots \hat{E}_h\}$. Since we set $h = \Gamma_*(h_1, h_s)$ the theorem says that there exists $J \subseteq [h]$, $|J| = h_s$, such that $\{\hat{E}_i \setminus \Delta\}_{i\in J}$ are pairwise disjoint for $\Delta = \cap_{i\in J}\hat{E}_i$.

Now we use the pigeonhole principle. Consider the triplets $\left\langle \hat{C}_i \cap \Delta, \hat{F}_i^\flat|_{\Delta\cap\hat{C}_i}, F_i^\sharp|_{\Delta\cap\hat{C}_i} \right\rangle$ over all $i \in J$ and note that the total number of possible triplets is at most

$$\sum_{k=0}^{h_0} \binom{h_1}{k} \cdot (q+1)^{h_0} \cdot (q+1)^{h_0} < h_s = |J|$$

by our choice of h_s. So there must be $i_1, i_2 \in J$ that satisfy the Proposition. ■

D. Finding an edge between $\mathcal{I}[B_{i_1}]$ and $\mathcal{I}[B_{i_2}]$

From now on let $i_1 = 1$ and $i_2 = 2$ without loss of generality.

Proposition 13. *For all $a_1 \in C_1$ and $a_2 \in C_2$ so that $F_1^\flat(a_1) = F_2^\sharp(a_2) \in [q]$, $(a_1, a_2) \in E_\mathcal{B}$.*

Proof: Assume $(a_1, a_2) \notin E_\mathcal{B}$. Then $a_1|_{\hat{B}} = a_2|_{\hat{B}}$ and since $a_1|_{\hat{B}} \in \hat{C}_1$ and $a_2|_{\hat{B}} \in \hat{C}_2$ we must have $a_1|_{\hat{B}} = a_2|_{\hat{B}} \in \Delta$. And if $F_1^\flat(a_1) = F_2^\sharp(a_2) \in [q]$ we have

$$\hat{F}_1^\flat(a_1|_{\hat{B}}) = \hat{F}_2^\sharp(a_2|_{\hat{B}}) = \hat{F}_1^\sharp(a_1|_{\hat{B}}), \text{ and}$$
$$\hat{F}_1^\flat(a_1|_{\hat{B}}) = \hat{F}_2^\flat(a_2|_{\hat{B}})$$

where the first equality is by the definition of a restricted coloring, the second is from the agreement of $\hat{F}_1^\sharp$ and $\hat{F}_2^\sharp$ on $\hat{C}_{i_1} \cap \Delta = \hat{C}_{i_2} \cap \Delta$ and the third is from the agreement of $\hat{F}_1^\flat$ and $\hat{F}_2^\flat$ on $\hat{C}_{i_1} \cap \Delta = \hat{C}_{i_2} \cap \Delta$. In particular, all these terms are equal. But the colorings $\hat{F}_1^\flat$ and $\hat{F}_1^\sharp$ agree only on $\dot{a}_1|_{\hat{B}}$, and similarly $\hat{F}_2^\flat$ and $\hat{F}_2^\sharp$ agree only on $\dot{a}_2|_{\hat{B}}$. So we have $\dot{a}_1|_{\hat{B}} = a_1|_{\hat{B}} = a_2|_{\hat{B}} = \dot{a}_2|_{\hat{B}}$. By preservation we must have $a_1 = \dot{a}_1$ and $a_2 = \dot{a}_2$, but $(\dot{a}_1, \dot{a}_2) \in E_\mathcal{B}$ since they assign T to v_1, v_2 respectively and v_1, v_2 are assumed to belong to a clique. ■

Now we separate the block assignments of $R_{\hat{B}}$. We let

$$\hat{D} = \hat{C}_1 \cup \hat{C}_2 \text{ and } \hat{R} = R_{\hat{B}} \setminus \hat{D},$$

and this partition separates the block assignments of R_{B_1} and R_{B_2}:

$$D_1 = \left\{a \in R_{B_1} : a|_{\hat{B}} \in \hat{D}\right\} \text{ and } R_1 = R_{B_1} \setminus D_1.$$

$$D_2 = \left\{a \in R_{B_2} : a|_{\hat{B}} \in \hat{D}\right\} \text{ and } R_2 = R_{B_2} \setminus D_2.$$

Proposition 14. *$|D_1| \leq 4h_0$ and $|D_2| \leq 4h_0$.*

Proof: $|D_1|, |D_2| \leq 2|\hat{D}| \leq 2(|\hat{C}_1| + |\hat{C}_2|) \leq 2(|C_1| + |C_2|) \leq 4h_0$. ■

Proposition 15. *$(D_1 \setminus C_1) \cap E_1 = \emptyset$.*

Proof: Suppose $a \in D_1 \cap E_1$, $a \notin C_1$ and we will reach a contradiction. Note that $C_1 \subseteq E_1$. Since $a \in D_1$, $a|_{\hat{B}} \in \hat{D} = \hat{C}_1 \cup \hat{C}_2$.

Case $a|_{\hat{B}} \in \hat{C}_1$: In this case, there exists $b \in C_1$ such that $b|_{\hat{B}} = a|_{\hat{B}}$. This contradicts preservation property, since no two distinct assignments in E_1 can have the same restriction to $\hat{B}$. Note that a and b are distinct since $a \in E_1 \setminus C_1$ and $b \in C_1$.

Case $a|_{\hat{B}} \in \hat{C}_2 \setminus \hat{C}_1$: In this case, $a|_{\hat{B}} \in \hat{C}_2 \setminus \Delta \subseteq \hat{E}_2 \setminus \Delta$ and the latter is disjoint from $\hat{E}_1 \setminus \Delta$. Thus $a \notin E_1$, a contradiction. ■

Proposition 16. *There is a coloring F_1' of every $a_1 \in D_1 \setminus C_1$ and a coloring F_2' of every $a_2 \in D_2 \setminus C_2$, using only colors in $[q]$, such that if we combine the colorings F_1' and $F_1^\flat$ to form F_1^* a coloring of D_1 and similarly combine F_2' and $F_2^\sharp$ to form a coloring F_2^* of D_2, we have, for all $a_1 \in D_1$, $a_2 \in D_2$*

$$F_1^*(a_1) = F_2^*(a_2) \in [q] \Rightarrow (a_1, a_2) \in E_\mathcal{B}.$$

Proof: We need to ensure that whenever $(a_1, a_2) \notin E_\mathcal{B}$, a_1 and a_2 receive different colors. Edges potentially not in $E_\mathcal{B}$ must agree on $\hat{B}$. For any fixed assignment on $\hat{B}$, consider the four block assignments $a_1^T, a_1^F \in D_1$ and $a_2^T, a_2^F \in D_2$ that extend it:

$$a_1^T(v_1) = a_2^T(v_2) = T, \quad a_1^F(v_1) = a_2^F(v_2) = F \quad \text{and,}$$
$$a_1^T|_{\hat{B}} = a_1^F|_{\hat{B}} = a_2^T|_{\hat{B}} = a_2^F|_{\hat{B}}.$$

Note that $(a_1^T, a_2^T) \in E_\mathcal{B}$. We will use the fact that we have $q \geq 3$ colors available. We do a case analysis ignoring cases that are symmetric to ones already covered. Proposition 13 guarantees that among the assignments that are already colored, only the pairs $(a_1^T, a_2^T), (a_1^T, a_1^F), (a_2^T, a_2^F)$, can have same colors.

Case 1: Neither of a_1^T, a_1^F is already colored: In this case, we first color a_2^T, a_2^F using at most two colors (preserving a color if already colored), say '1' and '2', and then color both a_1^T, a_1^F with a color '3'.

Case 2: Both a_1^T, a_1^F are already colored: If they are colored with the same color, say '1', then a_2^T, a_2^F can be colored using at most two other colors (preserving a color if already colored), say '2' and '3'. If a_1^T, a_1^F are already colored with different colors, say '1' and '2', then color a_2^T, a_2^F, if not already colored, with a color '3'.

Case 3: Exactly one of a_1^T, a_1^F and exactly one of a_2^T, a_2^F is colored already. If the colors are the same, say '1', the two remaining assignments can be colored with colors '2' and '3'. If the colors are different, then an uncolored assignment can be colored with the same color as its *mate*, where a_1^T, a_1^F are mates of each other and so are a_2^T, a_2^F.

This construction gives the desired colorings F_1^* and F_2^*. ■

Now define

$$\mathcal{I}_1 = \left\{F \in \{*, 1, \ldots q\}^{R_1} : (F, F_1^*) \in \mathcal{I}[B_1], \right\}$$

and

$$\mathcal{I}_2 = \left\{F \in \{*, 1, \ldots q\}^{R_2} : (F, F_2^*) \in \mathcal{I}[B_2]. \right\}$$

Proposition 17.

$$\mu_{p'}^{R_1}(\mathcal{I}_1) > \frac{1}{2} \text{ and } \mu_{p'}^{R_2}(\mathcal{I}_2) > \frac{1}{2}$$

Proof: We prove the proposition for $\mathcal{I}_1$; the proof for $\mathcal{I}_2$ is identical. We know that $> \frac{3}{4}$ fraction of extensions of $F_1^\flat$ outside of C_1 are in $\mathcal{I}[B_1]$. We will show that by setting the coloring of the block assignments in $D_1 \setminus C_1$ to F_1' as above, the coloring F_1^* still has $> \frac{1}{2}$ of its extensions in $\overline{D}_1 = R_1$ in $\mathcal{I}[B_1]$. We use the fact that block assignments in $D_1 \setminus C_1$ have influence $< \eta$ and $|D_1 \setminus C_1| < 4h_0$.

Let

$$\mathcal{I}[B_1]' = \{F \in \mathcal{I}[B_1] : F|_{D_1 \setminus C_1 \leftarrow F_1'} \notin \mathcal{I}[B_1]\}$$

where F_1' is the coloring of $D_1 \setminus C_1$ from Proposition 16. We claim that

$$\mu_{p'}(\mathcal{I}[B_1]') < \frac{1}{4}\left(\frac{p}{q}\right)^{h_0}$$

Let

$$\begin{aligned}\mathcal{F}'' = \quad & \{F \in \{*, 1, \dots, q\}^{(D_1 \setminus C_1)^c} : (F, F_1') \notin \mathcal{I}[B_1] \text{ but} \\ & (F, H) \in \mathcal{I}[B_1] \text{ for some } H \in \{1, \dots q\}^{D_1 \setminus C_1}\}.\end{aligned}$$

A coloring $F \in \mathcal{F}''$ contributes at least $\mu_{p'}^{(D_1 \setminus C_1)^c}(F) \cdot \left(\frac{p'}{q}\right)^{|D_1 \setminus C_1|}$ to the influence of at least one element $e \in D_1 \setminus C_1$. The sum of the influences in $D_1 \setminus C_1$ is less than $|D_1 \setminus C_1| \cdot \eta$, so

$$\begin{aligned}\mu_p^{(D_1 \setminus C_1)^c}(\mathcal{F}'') < \quad & |D_1 \setminus C_1| \cdot \eta \cdot \left(\tfrac{p'}{q}\right)^{-|D_1 \setminus C_1|} \\ < \quad & 4h_0 \cdot \eta \cdot \left(\tfrac{p}{q}\right)^{-4h_0} \leq \tfrac{1}{4}\left(\tfrac{p}{q}\right)^{h_0}.\end{aligned}$$

Since

$$\mathcal{I}[B_1]' \subseteq \{F : F|_{(D_1 \setminus C_1)^c} \in \mathcal{F}''\}$$

we have our claim.

Now since $\mu_{p'}^{C_1}(F_1^\flat) \geq (p/q)^{h_0}$ we can use $\mathbb{P}[A|B] \leq \mathbb{P}[A]/\mathbb{P}[B]$ to get

$$\begin{aligned}&\mathbb{P}_{F \in \mu_{p'}^{R_{B_1}}}\left[F \in \mathcal{I}[B_1]' | F|_{C_1} = F_1^\flat\right] \\ &\leq \mathbb{P}_{F \in \mu_{p'}^{R_{B_1}}}\left[F \in \mathcal{I}[B_1]'\right] \cdot \frac{1}{(p'/q)^{h_0}} < \frac{1}{4}\end{aligned}$$

we can write

$$\begin{aligned}\frac{3}{4} < \quad & \mathbb{P}_{F \in \mu_{p'}^{R_{B_1}}}\left[F \in \mathcal{I}[B_1] | F|_{C_1} = F_1^\flat\right] \\ = \quad & \mathbb{P}_{F \in \mu_{p'}^{R_{B_1}}}\left[F \in \mathcal{I}[B_1] \setminus \mathcal{I}[B_1]' | F|_{C_1} = F_1^\flat\right] \\ + \quad & \mathbb{P}_{F \in \mu_{p'}^{R_{B_1}}}\left[F \in \mathcal{I}[B_1]' | F|_{C_1} = F_1^\flat\right].\end{aligned}$$

The second term is $< \frac{1}{4}$, and the first term can be written as $\mathbb{P}[F \in \mathcal{I}[B_1] \mid F|_{D_1} = F_1^*]$, so we have

$$\mu_{p'}^{R_1}(\mathcal{I}_1) = \mathbb{P}[F \in \mathcal{I}[B_1] \mid F|_{D_1} = F_1^*] > \frac{1}{2}$$

■

Finally we extend the colorings F_1^* and F_2^* to colorings of R_{B_1} and R_{B_2}, respectively, that are both in the independent set, but non-agreeing, and hence completing the proof. To extend the colorings we need colorings $H_1 \in \mathcal{I}_1$ and $H_2 \in \mathcal{I}_2$ so that for all $a_1 \in R_1, a_2 \in R_2$ such that $H_1(a_1) = H_2(a_2) \in [q]$ we have $(a_1, a_2) \in E_{\mathcal{B}}$.

Proposition 18. *Given $\mathcal{I}_1$ and $\mathcal{I}_2$ as above, there exist $H_1 \in \mathcal{I}_1, H_2 \in \mathcal{I}_2$ such that for all $a_1 \in R_1$, $a_2 \in R_2$, if $H_1(a_1) = H_2(a_2) \in [q]$, then $(a_1, a_2) \in E_{\mathcal{B}}$.*

Proof: The proof is similar to the proof that two cross-agreeing families cannot be too large. By the monotonicity of $\mathcal{I}_1$ and $\mathcal{I}_2$ we can assume $p' = 1$. In what follows, all additions with elements in $[q]$ are defined modulo q. Define a mapping $\pi : \{1, \dots, q\}^{R_1} \to \{1, \dots, q\}^{R_2}$ as follows: let $F \in \{1, \dots, q\}^{R_1}$ and fix an assignment $\hat{a}$ to $\hat{B}$. Let the extensions of $\hat{a}$ be assignments a_1^T, a_1^F and a_2^T, a_2^F respectively. We describe how $\pi(F)$ colors a_2^T, a_2^F.

If $F(a_1^T) = F(a_1^F) = l \in [q]$, then we assign $\pi(F)(a_2^T) = \pi(F)(a_2^F) = l + 1$. If $F(a_1^T) = l_1$ and $F(a_1^F) = l_2 \neq l_1$, then we assign

$$\pi(F)(a_2^T) = l_1 \text{ and } \pi(F)(a_2^F) = L(l_1, l_2)$$

where $L(l_1, l_2) = l_2 + 2$ if $l_2 + 1 = l_1$, and $L(l_1, l_2) = l_2 + 1$ otherwise.

The two important facts are that π is injective and that F and $\pi(F)$ do not assign the same colors to any pair of consistent block assignments $a_1 \in R_1$ and $a_2 \in R_2$. So if the Proposition were false and such H_1, H_2 did not exist, each $F \in \mathcal{I}_1$ would rule out $\pi(F)$ from being in $\mathcal{I}_2$. Because $p' = 1$ and $|R_1| = |R_2|$, the fraction of the colorings ruled out is $\mu_1^{R_1}(\mathcal{I}_1)$ so $\mu_1^{R_2}(\mathcal{I}_2) \leq 1 - \mu_1^{R_1}(\mathcal{I}_1)$, a contradiction since $\mu_{p'}^{R_1}(\mathcal{I}_1) > \frac{1}{2}$ and $\mu_{p'}^{R_2}(\mathcal{I}_2) > \frac{1}{2}$ by Proposition 17. ■

ACKNOWLEDGMENT

Irit Dinur supported by ISF, BSF, and ERC grants. Subhash Khot supported by NSF CAREER grant CCF-0833228, NSF Expeditions grant CCF-0832795, and BSF grant 2008059. Will Perkins supported in part by NSF grant OISE-0730136.

REFERENCES

[1] M. Bellare, O. Goldreich, and M. Sudan, "Free bits, PCPs and non-approximability," *Electronic Colloquium on Computational Complexity, Technical Report TR95-024*, 1995.

[2] J. Hastad, "Clique is hard to approximate within $n^{1-\epsilon}$," *Acta Mathematica*, vol. 182, pp. 105–142, 1999.

[3] ——, "Some optimal inapproximability results," *Journal of ACM*, vol. 48, pp. 798–859, 2001.

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

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

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

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

[8] S. Khot, "On the power of unique 2-prover 1-round games," in *Proc. 34th ACM Symposium on Theory of Computing*, 2002.

[9] ——, "Inapproximability of NP-complete problems, discrete fourier analysis, and geometry," in *Proc. the International Congress of Mathematicians*, 2010.

[10] ——, "On the unique games conjecture," in *Proc. Annual IEEE Conference on Computational Complexity*.

[11] I. Dinur and S. Safra, "The importance of being biased," in *Proc. 34th Annual ACM Symposium on Theory of Computing*, 2002.

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

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

[14] D. Karger, R. Motwani, and M. Sudan, "Approximate graph coloring by semidefinite programming," in *Proc. 35th IEEE Symposium on Foundations of Computer Science*, 1994, pp. 2–13.

[15] A. Blum and D. Karger, "An $\tilde{O}(n^{3/14})$ coloring algorithm for 3-colorable graphs," *Information Processing Letters*, vol. 61, pp. 49–53, 1997.

[16] S. Arora, E. Chlamtac, and M. Charikar, "New approximation guarantee for chromatic number," in *Proc. ACM Symposium on the Theory of Computing*, 2006, pp. 215–224.

[17] E. Halperin, R. Nathaniel, and U. Zwick, "Coloring k-colorable graphs using smaller palettes," in *Proc. 12th ACM-SIAM Symposium on Discrete Algorithms*, 2001, pp. 319–326.

[18] S. Khanna, N. Linial, and S. Safra, "On the hardness of approximating the chromatic number," in *Proc. 2nd Israel Symposium on Theory and Computing Systems, ISTCS*, 1993, pp. 250–260.

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

[20] S. Khot, "Improved inapproximability results for maxclique, chromatic number and approximate graph coloring," in *Proc. 42nd IEEE Annual Symposium on Foundations of Computer Science*, 2001.

[21] M. Fürer, "Improved hardness results for approximating the chromatic number," in *Proc. 36th IEEE Sumposium on Foundations of Computer Science*, 1995, pp. 414–421.

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

[23] N. Bansal and S. Khot, "Optimal long code test with one free bit," in *Proc. 50th IEEE Symposium on Foundations of Computer Science*, 2009.

[24] L. Russo, "An approximate zero-one law," *Z. Wahrsch. Verw. Gebiete*, vol. 61(1), pp. 129–139, 1982.

[25] E. Friedgut, "Boolean functions with low average sensitivity depend on few coordinates," *Combinatorica*, vol. 18(1), pp. 27–35, 1998.

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

[27] 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.

[28] R. Ahlswede and L. Khachatrian, "The Diametric Theorem in Hamming Spaces–Optimal Anticodes," *Advances in Applied Mathematics*, vol. 20, no. 4, pp. 429–449, 1998.