Electronic Colloquium on Computational Complexity, Report No. 74 (2015)

ETH Hardness for Densest-k-Subgraph with Perfect Completeness

Mark Braverman * Young Kun Ko † Aviad Rubinstein ‡ Omri Weinstein §

Abstract

We show that, assuming the (deterministic) Exponential Time Hypothesis, distinguishing between a graph with an induced k-clique and a graph in which all k-subgraphs have density at most $1-\varepsilon$, requires $n^{\tilde{\Omega}(\log n)}$ time. Our result essentially matches the quasi-polynomial algorithms of Feige and Seltser [FS97] and Barman [Bar15b] for this problem, and is the first one to rule out an additive PTAS for Densest k-Subgraph. We further strengthen this result by showing that our lower bound continues to hold when, in the soundness case, even subgraphs smaller by a near-polynomial factor ($k' = k \cdot 2^{-\tilde{\Omega}(\log n)}$) are assumed to be at most $(1-\varepsilon)$-dense.

Our reduction is inspired by recent applications of the "birthday repetition" technique [AIM14, BKW15]. Our analysis relies on information theoretical machinery and is similar in spirit to analyzing a parallel repetition of two-prover games in which the provers may choose to answer some challenges multiple times, while completely ignoring other challenges.

*Department of Computer Science, Princeton University, email: mbraverm@cs.princeton.edu. Research supported in part by an an NSF CAREER award (CCF-1149888), a Turing Centenary Fellowship, a Packard Fellowship in Science and Engineering, and the Simons Collaboration on Algorithms and Geometry.

†Department of Computer Science, Princeton University, email: yko@cs.princeton.edu

‡Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, email: aviad@eecs.berkeley.edu

§Department of Computer Science, Princeton University, email: oweinste@cs.princeton.edu

ISSN 1433-8092

1 Introduction

k-Clique is one of the most fundamental problems in computer science: given a graph, decide whether it has a fully connected induced subgraph on k vertices. Since it was proven **NP**-complete by Karp [Kar72], extensive research has investigated the complexity of relaxed versions of this problem.

This work focuses on two natural relaxations of k-Clique which have received significant attention from both algorithmic and complexity communities: The first one is to relax "k", i.e. looking for a smaller subgraph:

Problem 1.1 (Approximate Max Clique, Informal)**.** *Given an n-vertex graph G, decide whether G contains a clique of size k, or all induced cliques of G are of size at most δk for some $1 > \delta(n) > 0$.*

The second natural relaxation is to relax the "Clique" requirement, replacing it with the more modest goal of finding a subgraph that is almost a clique:

Problem 1.2 (Densest k-Subgraph with perfect completeness, Informal)**.** *Given an n-vertex graph G containing a clique of size k, find an induced subgraphs of G of size k with (edge) density at least $(1-\varepsilon)$, for some $1 > \varepsilon > 0$. (More modestly, given an n-vertex graph G, decide whether G contains a clique of size k, or all induced k-subgraphs of G have density at most $(1-\varepsilon)$).*

Today, after a long line of research [FGL+96, AS98, ALM+98, Hås99, Kho01, Zuc07] we have a solid understanding of the inapproximability of Problem 1.1. In particular, we know that it is **NP**-hard to distinguish between a graph that has a clique of size k, and a graph whose largest induced clique is of size at most $k' = \delta k$ for $\delta = 1/n^{1-\varepsilon}$ [Zuc07]. The computational complexity of the second relaxation (Problem 1.2) remained largely open. There are a couple of quasi-polynomial algorithms that guarantee finding a $(1-\varepsilon)$-dense k subgraph in every graph containing a k-clique [FS97, Bar15b][1], suggesting that this problem is not **NP**-hard. Yet we know neither polynomial-time algorithms, nor general impossibility results for this problem.

In this work we provide a strong evidence that the aforementioned quasi-polynomial time algorithms for Problem 1.2 [FS97, Bar15b] are essentially tight, assuming the (deterministic) *Exponential Time Hypothesis* (ETH), which postulates that any deterministic algorithm for 3SAT requires $2^{\Omega(n)}$ time [IP01]. In fact, we show that under ETH, both parameters of the above relaxations are simultaneously hard to approximate:

Theorem 1.3 (Main Result)**.** *There exists a universal constant $\varepsilon > 0$ such that, assuming the (deterministic) Exponential Time Hypothesis, distinguishing between the following requires time $n^{\tilde{\Omega}(\log n)}$, where n is the number of vertices of G.*

Completeness *G has an induced k-clique; and*

Soundness *Every induced subgraph of G size $k' = k \cdot 2^{-\Omega(\frac{\log n}{\log\log n})}$ has density at most $1-\varepsilon$,*

Our result has implications for two major open problems whose computational complexity remained elusive for more than two decades: The (general) Densest k-Subgraph problem, and the Planted Clique problem.

The Densest k-Subgraph problem, $\mathsf{DkS}(\eta, \varepsilon)$, is the same as (the decision version of) Problem 1.2, except that in the "completeness" case, G has a k-subgraph with density η, and in the

[1] Barman [Bar15b] approximates the Densest k-Bi-Subgraph problem. Densest k-Subgraph can be handled via a simple modification [Bar15a].

"soundness" case, every k-subgraph is of density at most ε, where $\eta \gg \varepsilon$. Since Problem 1.2 is a special case of this problem, our main theorem can also be viewed as a new inapproximability result for $\mathsf{DkS}(1, 1-\varepsilon)$. We remark that the aforementioned quasi-polynomial algorithms for the "perfect completeness" regime completely break in the sparse regime, and indeed it is believed that $\mathsf{DkS}\left(n^{-\alpha}, n^{-\beta}\right)$ (for $k = n^{\varepsilon}$) in fact requires much more than quasi-polynomial time [BCV+12]. The best to-date algorithm for Densest k-Subgraph due to Bhaskara et. al, is guaranteed to find a k-subgraph whose density is within an $\sim n^{1/4}$-multiplicative factor of the densest subgraph of size k [BCV+12], and thus $\mathsf{DkS}(\eta, \varepsilon)$ can be solved efficiently whenever $\eta \gg n^{1/4} \cdot \varepsilon$ (this improved upon a previous $n^{1/3}$-approximation of Feige et. al [FKP01]). Making further progress on either the lower or upper bound frontier of the problem is a major open problem.

Several inapproximability results for Densest k-Subgraph were known against specific classes of algorithms [BCV+12] or under assumptions that are incomparable or stronger (thus giving weaker hardness results) than ETH: $\mathbf{NP} \nsubseteq \bigcap_{\varepsilon>0} \mathbf{BPTIME}\left[2^{n^{\varepsilon}}\right]$ [Kho06], Unique Games with expansion [RS10], and hardness of random k-CNF [Fei02, AAM+11]. The most closely related result is by Khot [Kho06], who shows that the Densest k-Subgraph problem has no PTAS unless SAT is in *randomized* subexponential time. The result of [Kho06], as well as other aforementioned works, focus on the sub-constant density regime, i.e. they show hardness for distinguishing between a graph where every k-subgraph is sparse, and one where every k-subgraph is extremely sparse. In contrast, our result has perfect completeness and provides the first *additive* inapproximability for Densest k-Subgraph — the best one can hope for as per the upper bound of [Bar15b].

The Planted Clique problem is a special case of our problem, where the inputs come from a specific distribution ($G(n,p)$ versus $G(n,p)+$ "a planted clique of size k", where p is some constant, typically 1/2). The *Planted Clique Conjecture* ([AAK+07, AKS98, Jer92, Kuc95, FK00, DGGP10]) asserts that distinguishing between the aforementioned cases for $p = 1/2, k = o(\sqrt{n})$ cannot be done in polynomial time, and has served as the underlying hardness assumption in a variety of recent applications including machine-learning and cryptography (e.g. [AAK+07, BBB+13, BR13]) that inherently use the average-case nature of the problem, as well as in reductions to worst-case problems (e.g. [HK11, AAM+11, CLLR15, BPR+15b]).

The main drawback of average-case hardness assumptions is that many average-case instances (even those of worst-case-hard problems) are in fact tractable. In recent years, the centrality of the planted clique conjecture inspired several works that obtain lower bounds in restricted models of computation [FGR+13, MPW15, DM15]. Nevertheless, a general lower bound for the average-case planted clique problem appears out of reach for existing lower bound techniques. Therefore, an important potential application of our result is replacing average-case assumptions such as the planted-clique conjecture, in applications that do not inherently rely on the distributional nature of the inputs (e.g., when the ultimate goal is to prove a worst-case hardness result). In such applications, there is a good chance that planted clique hardness assumptions can be replaced with a more "conventional" hardness assumption, such as the ETH, even when the problem has a quasi-polynomial algorithm. Recently, such a replacement of the planted clique conjecture with ETH was obtained for the problem of finding an approximate Nash equilibrium with approximately optimal social welfare [BKW15].

We also remark that, while showing hardness for Planted Clique from worst-case assumptions seems beyond the reach of current techniques, our result can also be seen as circumstantial evidence that this problem may indeed be hard. In particular, any polynomial time algorithm (if exists) would have to inherently use the (rich and well-understood) structure of $G(n,p)$.

Techniques

Our simple construction is inspired by the "birthday repetition" technique which appeared recently in [AIM14, BKW15, BPR15a]: given a 2CSP (e.g. 3COL), we have a vertex for each $\tilde{\Omega}(\sqrt{n})$-tuple of variables and assignments (respectively, 3COL vertices and colorings). We connect two vertices by an edge whenever their assignments are consistent and satisfy all 2CSP constraints induced on these tuples. In the completeness case, a clique consists of choosing all the vertices that correspond to a fixed satisfying assignment. In the soundness case (where the value of the 2CSP is low), the "birthday paradox" guarantees that most pairs of vertices vertices (i.e. two $\tilde{\Omega}(\sqrt{n})$-tuples of variables) will have a significant intersection (nonempty CSP constraints), thus resulting in lower densities whenever the 2CSP does not have a satisfying assignment. In the language of two-prover games, the intuition here is that the verifier has a "constant chance in catching the players in a lie if thy are trying to cheat" in the game while not satisfying the CSP.

While our construction is simple, analyzing it is intricate. The main challenge is to rule out a "cheating" dense subgraph that consists of different assignments to the same variables (inconsistent colorings of the same vertices in 3COL). Intuitively, this is similar in spirit to *proving a parallel repetition theorem where the provers can answer some questions multiple times, and completely ignore other questions.* Continuing with the parallel repetition metaphor, notice that the challenge is doubled: in addition to a cheating prover correlating her answers (the standard obstacle to parallel repetition), each prover can now also correlate which questions she chooses to answer. Our argument follows by showing that a sufficiently large subgraph must accumulate many non-edges (violations of either 2CSP or consistency constraints). To this end we introduce an information theoretic argument that carefully counts the entropy of choosing a random vertex in the dense subgraph.

1.1 Open problems

There are several interesting open problems related to our work. We henceforth list four of them that are of particular interest and potential applications.

Strengthening the inapproximability factor Our result states that it is hard to distinguish between a graph containing a k-clique and a graph that does not contain a very dense $(1-\delta)$ k-subgraph. The latter $(1-\delta)$ seems to be a limitation of our technique. None of the algorithms we know (including the two quasi-polynomial time algorithms mentioned above) can distinguish in polynomial time between a graph containing a k-clique and a graph that does not contain even a slightly dense (δ) k-subgraph; for any constant $\delta > 0$, and in fact even for some sub-constant values of δ. Furthermore, there is evidence [AAM+11] that this problem may indeed be hard. This naturally leads to the following problem.

Problem 1.4 (Hardness Amplification)**.** *Show that for every given constant $\delta > 0$, distinguishing between the following two cases is ETH-hard:*

- *There exists $S \subset V$ of size k such that* $\mathrm{den}(S) = 1$.
- *All $S \subset V$ of size k have* $\mathrm{den}(S) \leq \delta$.

We remark that a similar amplification, from "clique versus dense" ($\mathrm{den}(S) = 1$ vs. $\mathrm{den}(S) = 1-\delta$) to "clique versus sparse" ($\mathrm{den}(S) = 1$ vs. $\mathrm{den}(S) = \delta$), was shown by Alon et al. when the "clique vs. dense" instance is drawn at random according to the planted clique model [AAM+11]. (Unfortunately, their techniques do not seem to apply to our hard instance.)

An easier variant of Problem 1.4 is to show hardness for a large gap in the imperfect completeness regime.

Problem 1.5 (Hardness Amplification - imperfect completeness)**.** *Show that there exist parameters* $0 < \varepsilon \ll \eta < 1$ *for which distinguishing between the following two cases is ETH-hard:*

- *There exists* $S \subset V$ *of size* k *such that* $\mathrm{den}(S) \geq \eta$.
- *All* $S \subset V$ *of size* k *have* $\mathrm{den}(S) \leq \varepsilon$.

We note that such gaps can be obtained from average-case hardness for a random k-CNF [AAM+11] and from Unique Games with expansion [RS10].

Beyond quasi-polynomial hardness Another interesting challenge is to trade the perfect completeness in our main result for stronger notions of hardness. Indeed, there are substantial evidences which suggest that the "sparse vs. very-sparse" regime ($\mathsf{DkS}(\eta, \varepsilon)$) is much harder to solve. The gap instance in [BCV+12] where all known linear and semidefinite programming techniques fail is a very sparse instance and has integrality gap of $\Omega(n^{2/53-\varepsilon})$. In particular, every vertex has degree $n^{1/2+o(1)}$, compared to almost linear average degree in our instance. Since no other algorithms succeed in this regime (even in quasi-polynomial time), it is natural to look for stronger lower bounds on the running time.

Problem 1.6 (Trading-off perfect completeness for stronger lower bounds)**.** *Show that there exist parameters* $0 < \varepsilon < \eta \ll 1$ *for which distinguishing between the following two cases is* **NP**-*hard:*

- *There exists* $S \subset V$ *of size* k *such that* $\mathrm{den}(S) \geq \eta$.
- *All* $S \subset V$ *of size* k *have* $\mathrm{den}(S) \leq \varepsilon$.

Finding Stable Communities The problem of finding *Stable Communities* is tightly related to Densest k-Subgraph, and has received recent attention in the context of social networks and learning theory [AGSS12, AGM13, BL13].

Definition 1.7 (Stable Communities [BBB+13])**.** *Let* α, β *with* $\beta < \alpha \leq 1$ *be two positive parameters. Given an undirected graph,* $G = (V, E)$, $S \subset V$ *is an* (α, β)*-cluster if* S *is :*

1. *Internally Dense:* $\forall i \in S$, $|\mathcal{N}(i) \cap S| \geq \alpha |S|$.
2. *Externally Sparse:* $\forall i \notin S$, $|\mathcal{N}(i) \cap S| \leq \beta |S|$.

Currently, only planted clique based hardness is known.

Theorem 1.8 ([BBB+13])**.** *For sufficiently small (constant)* γ*, finding a* $(1, 1-\gamma)$ *cluster is at least as hard as* Planted Clique.

As insinuated in the introduction, we believe it is plausible and interesting to see whether the hardness assumption of the theorem above can be replaced with ETH.

Problem 1.9 (Hardness of Stable Communities)**.** *Show that for some* α, β *with* $\beta < \alpha \leq 1$*, finding an* (α, β)*-cluster* S *is ETH-hard.*

2 Preliminaries

Throughout the paper we use $\mathrm{den}(S) \in [0,1]$ to denote the density of subgraph S,

$$\mathrm{den}(S) := \frac{\left|(S \times S) \cap E\right|}{|S \times S|}.$$

2.1 Information theory

In this section, we introduce information-theoretic quantities used in this paper. For a more thorough introduction, the reader should refer to [CT12]. Unless stated otherwise, all log's in this paper are base-2.

Definition 2.1. *Let μ be a probability distribution on sample space Ω. The* Shannon entropy *(or just* entropy*) of μ, denoted by $H(\mu)$, is defined as $H(\mu) := \sum_{\omega \in \Omega} \mu(\omega) \log \frac{1}{\mu(\omega)}$.*

Definition 2.2 (Binary Entropy Function)**.** *For $p \in [0,1]$, the binary entropy function is defined as follows (with a slight abuse of notation) $H(p) := -p \log p - (1-p)\log(1-p)$.*

Fact 2.3 (Concavity of Binary Entropy)**.** *Let μ be a distribution on $[0,1]$, and let $p \sim \mu$. Then $H(\mathbb{E}_\mu[p]) \geq \mathbb{E}_\mu[H(p)]$.*

For a random variable A we shall write $H(A)$ to denote the entropy of the induced distribution on the support of A. We use the same abuse of notation for other information-theoretic quantities appearing later in this section.

Definition 2.4. *The* Conditional entropy *of a random variable A conditioned on B is defined as*

$$H(A|B) = \mathbb{E}_b(H(A|B=b)).$$

Fact 2.5 (Chain Rule)**.** $H(AB) = H(A) + H(B|A)$.

Fact 2.6 (Conditioning Decreases Entropy)**.** $H(A|B) \geq H(A|BC)$.

Another measure we will use (briefly) in our proof is that of *Mutual Information*, which informally captures the correlation between two random variables.

Definition 2.7 (Conditional Mutual Information)**.** *The* mutual information *between two random variable A and B, denoted by $I(A;B)$ is defined as*

$$I(A;B) := H(A) - H(A|B) = H(B) - H(B|A).$$

The conditional mutual information *between A and B given C, denoted by $I(A;B|C)$, is defined as*

$$I(A;B|C) := H(A|C) - H(A|BC) = H(B|C) - H(B|AC).$$

The following is a well-known fact on mutual information.

Fact 2.8 (Data processing inequality)**.** *Suppose we have the following Markov Chain:*

$$X \to Y \to Z$$

where $X \perp Z|Y$. Then $I(X;Y) \geq I(X;Z)$ or equivalently, $H(X|Y) \leq H(X|Z)$.

Mutual Information is related to the following distance measure.

Definition 2.9 (Kullback-Leiber Divergence)**.** *Given two probability distributions μ_1 and μ_2 on the same sample space Ω such that $(\forall \omega \in \Omega)(\mu_2(\omega) = 0 \Rightarrow \mu_1(\omega) = 0)$, the* Kullback-Leibler Divergence *between is defined as (also known as relative entropy)*

$$\mathsf{D}_{\mathsf{KL}}\left(\mu_1 \middle\| \mu_2\right) = \sum_{\omega \in \Omega} \mu_1(\omega) \log \frac{\mu_1(\omega)}{\mu_2(\omega)}.$$

The connection between the mutual information and the Kullback-Leibler divergence is provided by the following fact.

Fact 2.10. *For random variables A, B, and C we have*

$$I(A; B|C) = \mathbb{E}_{b,c}\left[\mathsf{D}_{\mathsf{KL}}\left(A_{bc} \middle\| A_c\right)\right].$$

2.2 2CSP and the PCP Theorem

In the **2CSP** problem, we are given a graph $G = (V, E)$ on $|V| = n$ vertices, where each of the edges $(u, v) \in E$ is associated with some constraint function $\psi_{u,v} : A \times A \to \{0, 1\}$ which specifies a set of legal "colorings" of u and v, from some finite alphabet A (2 in the term "**2CSP** " stands for the "arity" of each constraint, which always involves two variables). Let us denote by ψ the entire **2CSP** instance, and define by $\mathrm{OPT}(\psi)$ the maximum fraction of satisfied constraints in the associated graph G, over all possible assignments (colorings) of V.

The starting point of our reduction is the following version of the PCP theorem, which asserts that it is **NP**-hard to distinguish a **2CSP** instance whose value is 1, and one whose value is $1 - \eta$, where η is some small constant:

Theorem 2.11 (PCP Theorem [Din07])**.** *Given a* **3SAT** *instance φ of size n, there is a polynomial time reduction that produces a* **2CSP** *instance ψ, with size $|\psi| = n \cdot \mathrm{polylog}\, n$ variables and constraints, and constant alphabet size such that*

- *(Completeness) If $\mathrm{OPT}(\varphi) = 1$ then $\mathrm{OPT}(\psi) = 1$.*
- *(Soundness) If $\mathrm{OPT}(\varphi) < 1$ then $\mathrm{OPT}(\psi) < 1 - \eta$, for some constant $\eta = \Omega(1)$*
- *(Balance) Every vertex in ψ has degree d for some constant d.*

In the appendix, we describe in detail how to derive this formulation of the PCP Theorem from that of e.g. [AIM14].

Notice that since the size of the reduction is near linear, ETH implies that solving the above problem requires near exponential time.

Corollary 2.12. *Let ψ be as in Theorem 2.11. Then assuming ETH, distinguishing between $\mathrm{OPT}(\psi) = 1$ and $\mathrm{OPT}(\psi) < 1 - \eta$ requires time $2^{\tilde{\Omega}(|\psi|)}$.*

3 Main Proof

3.1 Construction

Let ψ be the **2CSP** instance produced by the reduction in Theorem 2.11, i.e. a constraint graph over n variables with alphabet A of constant size. We construct the following graph $G_\psi = (V, E)$:

- Let $\rho := \sqrt{n}\log\log n$ and $k := \binom{n}{\rho}$.
- Vertices of G_ψ correspond to all possible assignments (colorings) to all ρ-tuples of variables in ψ, i.e $V = [n]^\rho \times |A|^\rho$. Each vertex is of the form $v = (y_{x_1}, y_{x_2}, \ldots, y_{x_\rho})$ where $\{x_1, \ldots, x_\rho\}$ are the chosen variables of v, and y_{x_i} is the corresponding assignment to variable x_i.
- If $v \in V$ *violates any* **2CSP** *constraints*, i.e. if there is a constraint on (x_i, x_j) in ψ which is not satisfied by $\left(y_{x_i}, y_{x_j}\right)$, then v is an isolated vertex in G_ψ.
- Let $u = (y_{x_1}, y_{x_2}, \ldots, y_{x_\rho})$ and $v = (y'_{x'_1}, y'_{x'_2}, \ldots, y'_{x'_\rho})$. $(u, v) \in E$ iff:
  - (u, v) does not *violate any consistency constraints*: for every shared variable x_i, the corresponding assignments agree, $y_{x_i} = y'_{x_i}$;
  - and (u, v) also does not *violate any* **2CSP** *constraints*: for every **2CSP** constraint on $\left(x_i, x'_j\right)$ (if exists), the assignment $\left(y_{x_i}, y'_{x'_j}\right)$ satisfy the constraint.

Notice that the size of our reduction (number of vertices of G_ψ) is $N = \binom{n}{\rho} \cdot |A|^\rho = 2^{\tilde{O}(\sqrt{n})}$.

Completeness If $\mathrm{OPT}(\psi) = 1$, then G_ψ has a k-clique: Fix a satisfying assignment for ψ, and let S be the set of all vertices that are consistent with this assignment. Notice that $|S| = \binom{n}{\rho} = k$. Furthermore its vertices do not violate any consistency constraints (since they agree with a single assignment), or **2CSP** constraints (since we started from a satisfying assignment).

4 Soundness

Suppose that $\mathrm{OPT}(\psi) < 1 - \eta$, and let $\varepsilon_0 > 0$ be some constant to be determined later. We shall show that for any subset S of size $k' = k \cdot |V|^{-\varepsilon_0/\log\log|V|}$, $\mathrm{den}(S) < 1 - \delta$, where δ is some constant depending on η. The remainder of this section is devoted to proving the following theorem:

Theorem 4.1 (Soundness)**.** *If* $\mathrm{OPT}(\psi) < 1 - \eta$*, then* $\forall S \subset V$ *of size* $k' = k \cdot |V|^{-\varepsilon_0/\log\log|V|}$, $\mathrm{den}(S) < 1 - \delta$ *for some constant* δ*.*

4.1 Setting up the entropy argument

Fix some subset S of size k', and let $v \in_R S$ be a uniformly chosen vertex in S (recall that v is a vector of ρ coordinates, corresponding to labels for a subset of ρ chosen variables). Let X_i denote the indicator variable associated with v such that $X_i = 1$ if the i'th variable appears in v and 0 otherwise. We let Y_i represent the coloring assignment (label) for the i'th variable whenever $X_i = 1$, which is of the form $l \in A$. Throughout the proof, let

$$W_{i-1} = X_{<i}, Y_{<i}$$

denote the i'th prefix corresponding to v. We can write :

$$\begin{aligned} H(Y_i|W_{i-1}, X_i) &= \Pr[X_i = 0] \cdot H(Y_i|W_{i-1}, X_i = 0) + \Pr[X_i = 1] \cdot H(Y_i|W_{i-1}, X_i = 1) \\ &= \Pr[X_i = 1] \cdot H(Y_i|W_{i-1}, X_i = 1) \end{aligned}$$

since $H(Y_i|W_{i-1}, X_i = 0) = 0$. Notice that since (XY) and v determine each other, and v was uniform on a set of size $|S| = k'$, we have

Observation 4.2. $H(XY) = \log k'$.

Thus, in total, the choice of challenge and the choice of assignments should contribute $\log k'$ to the entropy of v. If much of the entropy comes from the assignment distribution (conditioned on the fixed challenge variables), we will show that S must have many consistency violations, implying that S is sparse. If, on the other hand, almost all the entropy comes from the challenge distribution, we will show that this implies many CSP constraint violations (implied by the soundness assumption). From now on, we denote

$$\alpha_i := H(X_i|X_{<i}, Y_{<i}) \quad \text{and} \quad \beta_i := H(Y_i|X_{\leq i}, Y_{<i}).$$

When conditioning on the i'th prefix, we shall write $\alpha_i(w_{i-1}) := H(X_i|X_{<i}, Y_{<i} = w_{i-1})$, and similarly for $\beta_i(\cdot)$. Also for brevity, we denote

$$q_i := \Pr[X_i = 1] \quad \text{and} \quad q_i(w_{i-1}) := \Pr[X_i = 1|w_{i-1}].$$

Prefix graphs

The consistency constraints induce, for each i, a graph over the prefixes: the vertices are the prefixes, and two prefixes are connected by an edge if their labels are consistent. (We can ignore the **2CSP** constraints for now — the prefix graph will be used only in the analysis of the consistency constraints.) Formally,

Definition 4.3 (Prefix graph)**.** *For $i \in [n+1]$ let the i-th* prefix graph*, G_i be defined over the prefixes of length $i-1$ as follows. We say that w_{i-1} is a neighbor of σ_{i-1} if they do not violate any consistency constraints. Namely, for all $j < i$, if $X_j = 1$ for both w_{i-1} and σ_{i-1}, then w_i and σ_i assign the same label Y_j.*

In particular, we will heavily use the following notation: let $\mathcal{N}(w_{i-1})$ be the prefix neighborhood *of w_{i-1}; i.e. it is the set of all prefixes (of length $i-1$) that are consistent with w_{i-1}. For technical issues of normalization, we let $w_{i-1} \in \mathcal{N}(w_{i-1})$, i.e. all the prefixes have self-loops.*

Notice that G_{n+1} is defined over the vertices of S (the original subgraph). The set of edges on S is contained in the set of edges of G_{n+1}, since in the latter we only remove pairs that violated consistency constraints (recall that we ignore the **2CSP** constraints).

Unless stated otherwise, we always think of prefixes as weighted by their probabilities. Naturally, we also define the weighted degree and weighted edge density of the prefix graph.

Definition 4.4 (Prefix degree and density)**.** *The* prefix degree *of w_{i-1} is given by:*

$$\deg(w_{i-1}) = \sum_{\sigma_{i-1} \in \mathcal{N}(w_{i-1})} \Pr[\sigma_{i-1}].$$

Similarly, we define the prefix density *of G_i as:*

$$\mathrm{den}(G_i) = \sum_{w_{i-1}} \sum_{\sigma_{i-1} \in \mathcal{N}(w_{i-1})} \Pr[w_{i-1}] \cdot \Pr[\sigma_{i-1}].$$

When it is clear from the context, we henceforth drop the *prefix* qualification, and simply refer to the *neighborhood* or *degree*, etc., of w_{i-1}.

Notice that in G_{n+1}, the probabilities are uniformly distributed. In particular, $\mathrm{den}(G_{n+1}) \geq \mathrm{den}(S)$, since, as we mentioned earlier, the set of edges in S is contained in that of G_{n+1}. Finally, observe also that because we accumulate violations, the density of the prefix graphs is monotonically non-increasing with i.

Observation 4.5.

$$\mathrm{den}(G_1) \geq \cdots \geq \mathrm{den}(G_{n+1}) \geq \mathrm{den}(S).$$

Useful approximations

We use the following bounds on α_i and β_i many times throughout the proof:

Fact 4.6.

$$\alpha_i = \mathbb{E}\left[H(q_i(w_{i-1}))\right] \leq H(\mathbb{E}\left[q_i(w_{i-1})\right]) = H(q_i)$$

Fact 4.7.

$$\beta_i = \mathbb{E}\left[\beta_i(w_{i-1}))\right] \leq \mathbb{E}\left[q_i(w_{i-1}) \cdot \log|A|\right] = q_i \log|A|$$

Proof. The bound on α_i follows from concavity of entropy (Fact 2.3). For the second bound, observe that β_i is maximized by spreading q_i mass uniformly over alphabet A. □

We also recall some elementary approximations to logarithms and entropies that will be useful in the analysis. The proofs are deferred to the appendix.

Fact 4.8. *For $k = \binom{n}{\rho}$ then,*

$$\log k = nH\left(\frac{\rho}{n}\right) \pm O\left(\log n\right) = \left(\frac{1}{2} - o\left(1\right)\right)\rho \log n$$

More useful to us will be the following bounds on $\log k'$:

Fact 4.9. *Let $\varepsilon_1 \geq 5\varepsilon_0$, and k, k', V, n, ρ as specified in the construction. Then,*

$$\log k' \geq \max\left\{\log k, nH\left(\frac{\rho}{n}\right)\right\} - \underbrace{\varepsilon_1 \log k / \log n}_{\approx \frac{\varepsilon_1}{2} \cdot \rho}.$$

In particular, this means that most indices i should contribute roughly $H\left(\frac{\rho}{n}\right)$ entropy to the choice of v.

We will also need the following bound which relates the entropies of a very biased coin and a slightly less biased one:

Fact 4.10. *Let $1/n \ll |\upsilon| \ll 1$, then*

$$H\left(\frac{1+\upsilon}{n}\right) = H\left(\frac{1}{n}\right) - \frac{\upsilon}{n}\log\frac{1}{n} - (\log \mathrm{e})\frac{\upsilon^2}{2n} + O\left(n^{-2}\right) + O\left(\frac{\upsilon^3}{n}\right)$$

4.2 Consistency violations

In this section, we show that if the entropy contribution of the assignments ($\sum_i H(Y_i|X_{\leq i}, Y_{<i})$) is large, there are many consistency violations between vertices, which lead to constant density loss. First, we show that if $H(Y_i|X_{\leq i}, Y_{<i}) > 5\varepsilon_1 \log k / \log n$, then at least a constant fraction of such entropy is concentrated on "good" variables.

Definition 4.11 (Good Variables)**.** *We say that an index i is* good *if*

- $\alpha_i \geq H(q_i) - 2q_i \log|A|$
- $\beta_i \geq \frac{1}{2}\varepsilon_1 q_i$

where ε_1 is a constant to be determined later in the proof.

Claim 4.12. *For any constant ε_1, if $\sum_i \beta_i > 5\varepsilon_1 \log k/\log n$,*

$$\sum_{\text{good } i\text{'s}} q_i^2 \geq \left(\frac{1}{5}\varepsilon_1\rho\right)^2/(n\log^2|A|) = \Omega(\rho^2/n).$$

Proof. We want to show that many of the indices i have both a large α_i and a large β_i simultaneously. We can write

$$\sum_{i\in[n]}(\alpha_i+\beta_i) = \sum_{i\colon \alpha_i+\beta_i<H(q_i)-q_i\log|A|}(\alpha_i+\beta_i) + \sum_{i\colon \alpha_i+\beta_i\geq H(q_i)-q_i\log|A|}(\alpha_i+\beta_i)$$

Using Facts 4.6 and 4.7, we have

$$\sum_{i\in[n]}(\alpha_i+\beta_i) \leq \sum_{i\colon \alpha_i+\beta_i<H(q_i)-q_i\log|A|}(H(q_i)-\beta_i) + \sum_{i\colon \alpha_i+\beta_i\geq H(q_i)-q_i\log|A|}(H(q_i)+\beta_i). \tag{1}$$

Because the subgraph is of size k', from the expansion of $\log k'$ (Fact 4.9),

$$\sum_{i\in[n]}(\alpha_i+\beta_i) \geq nH\left(\frac{\rho}{n}\right) - \varepsilon_1\log k/\log n \geq \sum H(q_i) - \varepsilon_1\log k/\log n,$$

where the second inequality follows from the concavity of entropy. Plugging into (1), we have

$$\begin{aligned}\sum_{i\colon \alpha_i+\beta_i\geq H(q_i)-q_i\log|A|}\beta_i &\geq \sum_{i\colon \alpha_i+\beta_i<H(q_i)-q_i\log|A|}\beta_i - \varepsilon_1\log k/\log n\\ &= \left(\sum_i \beta_i - \sum_{i\colon \alpha_i+\beta_i\geq H(q_i)-q_i\log|A|}\beta_i\right) - \varepsilon_1\log k/\log n\end{aligned}$$

Rearranging, we get

$$\sum_{i\colon \alpha_i+\beta_i\geq H(q_i)-q_i\log|A|}\beta_i \geq \frac{1}{2}\sum_i\beta_i - \varepsilon_1\log k/\log n \tag{2}$$

For all the i's in the LHS summation, $\alpha_i \geq H(q_i) - 2q_i\log|A|$ by Fact 4.7. From now on, we will consider only i's that satisfy this condition. Now, using the premise on $\sum_i\beta_i$ and (2) we have:

$$\sum_{i\colon \alpha_i\geq H(q_i)-2q_i\log|A|}\beta_i \geq (5/2-1)\varepsilon_1\log k/\log n \geq 0.7\varepsilon_1\rho,$$

where the second inequality follows from our approximation for $\log k$ (Fact 4.8).

We want to further restrict our attention to i's for which β_i is at least $\frac{1}{2}\varepsilon_1 q_i$ (aka good i's). Note that the above inequality can be decomposed to

$$\sum_{\text{good } i\text{'s}}\beta_i + \sum_{\substack{i\colon \alpha_i\geq H(q_i)-2q_i\log|A|\\ \beta_i<\frac{1}{2}\varepsilon_1 q_i}}\beta_i \geq 0.7\varepsilon_1\rho$$

Now via a simple sum bound,

$$\sum_{\substack{i\colon \alpha_i\geq H(q_i)-2q_i\log|A|\\ \beta_i<\frac{1}{2}\varepsilon_1 q_i}}\beta_i \leq \frac{1}{2}\varepsilon_1\sum_i q_i = \frac{1}{2}\varepsilon_1\rho$$

Rearranging, we get,

$$\sum_{\text{good } i\text{'s}} \beta_i \geq \frac{1}{5}\varepsilon_1 \rho$$

By Cauchy-Schwartz we have:

$$\sum_{\text{good } i\text{'s}} \beta_i^2 \geq \left(\frac{1}{5}\varepsilon_1 \rho\right)^2 / n$$

Finally, since $\beta_i \leq q_i \log |A|$,

$$\sum_{\text{good } i\text{'s}} q_i^2 \geq \left(\frac{1}{5}\varepsilon_1 \rho\right)^2 / (n \log^2 |A|).$$

□

In the same spirit, we now define a notion of a "good" prefix. Intuitively, conditioning on a good prefix leaves a significant amount of entropy on the i'th index. We also require that a good prefix has a high prefix degree; that is, it has many neighbors it could potentially lose when revealing the i-th label.

Definition 4.13 (Good Prefixes)**.** *We say w_{i-1} is a* good *prefix if*

- *i is good;*
- $\sum_{\sigma_{i-1} \in \mathcal{N}(w_{i-1})} q_i(\sigma_{i-1}) \Pr[\sigma_{i-1}] \geq (1 - \varepsilon_2) q_i$*;*
- $\beta_i(w_{i-1}) \geq \varepsilon_3 q_i(w_{i-1})$

where $\varepsilon_3 = (\varepsilon_4 + \kappa) \log |A|$, with ε_4 an an arbitrarily small constant that denotes the fraction of assignments that disagree with the majority of the assignments, $\kappa = \Theta(1/\log |A|)$ factor, and ε_2 a constant that satisfies $\delta = \left(\frac{\varepsilon_2}{|A|^{2/\varepsilon_2}}\right)^4$, with $\mathrm{den}(S) = 1 - \delta$.

In the following claim, we show that these prefixes contribute some constant fraction of entropy, assuming that our subset is dense.

Claim 4.14. *If $\mathrm{den}(S) > 1 - \delta$, where $\delta = \left(\frac{\varepsilon_2}{|A|^{2/\varepsilon_2}}\right)^4$ and $\varepsilon_1 \geq 4\varepsilon_2 \log |A| + 8\varepsilon_3$, then for every good index i, it holds that*

$$\sum_{\text{good } w_{i-1}\text{'s}} \Pr[w_{i-1}] \beta_i (w_{i-1}) \geq \beta_i / 4$$

Proof. We begin by proving that most prefixes satisfy the degree condition of Definition 4.13. Let w_{i-1} be *popular* if i is a good variable and its degree in the prefix graph G_i is at least $deg(w_{i-1}) := \sum_{\sigma_{i-1} \in \mathcal{N}(w_{i-1})} \Pr[\sigma_{i-1}] \geq 1 - \sqrt{\delta}$. Recall that $\mathrm{den}(G_i) \geq \mathrm{den}(S) \geq (1 - \delta)$ (by Observation 4.5). Thus by Markov inequality, at most $\sqrt{\delta}$-fraction of the prefixes are unpopular.

Let $Z(\cdot)$ be the indicator variable for W_{i-1} being popular. For the sake of contradiction, suppose that more than ε_2-fraction of the q_i-mass is concentrated on unpopular prefixes, that is:

$$\sum_{\text{unpopular } w_{i-1}\text{'s}} \Pr[w_{i-1}] q_i (w_{i-1}) = \Pr[Z(W_{i-1}) = 0] \cdot \Pr[X_i = 1 \mid Z(W_{i-1}) = 0] > \varepsilon_2 q_i. \quad (3)$$

We would like to argue that this condition implies that the distribution on the X_i's is highly biased by the conditioning on the (popularity of the) prefix; this in turn implies that α_i, the expected conditional entropy of X_i, must be low, contradicting the assumption that i is good. Indeed, by data-processing inequality (Fact 2.8),

$$\begin{aligned} \alpha_i &= H\left(X_i \mid W_{i-1}\right) \\ &\leq H\left(X_i \mid Z(W_{i-1})\right) \\ &= H\left(X_i\right) - I\left(X_i; Z(W_{i-1})\right) \end{aligned} \tag{4}$$

Since we can write mutual information as expected KL-divergence (Fact 2.10), and KL-divergence is non-negative, we get

$$\begin{aligned} I(X_i; Z(W_{i-1})) &= \mathbb{E}_{x_i}\left[\mathsf{D}_{\mathsf{KL}}\left(Z(W_{i-1})|x_i \middle\| Z(W_{i-1})\right)\right] \\ &\geq q_i \cdot \mathsf{D}_{\mathsf{KL}}\left(\Pr[Z(W_{i-1}) = 1 \mid x_i = 1] \middle\| Z(W_{i-1}) = 1\right) \\ &\geq q_i \cdot \mathsf{D}_{\mathsf{KL}}\left(1 - \varepsilon_2 \middle\| 1 - \sqrt{\delta}\right) = q_i \mathsf{D}_{\mathsf{KL}}\left(\varepsilon_2 \middle\| \sqrt{\delta}\right), \end{aligned}$$

where the second inequality follows from the fact that for all good i's, our degree assumption implies $\Pr[Z(W_{i-1})] \geq (1 - \sqrt{\delta})$, and our assumption in (3) implies, via Bayes rule, that $\Pr[Z(W_{i-1} = 0 \mid x_i = 1] \geq \varepsilon_2$, and therefore $\Pr[W_{i-1} = 1 \mid x_i = 1] \leq 1 - \varepsilon_2$. Note that by our setting of parameters $1 - \sqrt{\delta} > 1 - \varepsilon_2$.

Plugging into (4) we have:

$$\alpha_i \leq H\left(q_i\right) - q_i \mathsf{D}_{\mathsf{KL}}\left(\varepsilon_2 \middle\| \sqrt{\delta}\right). \tag{5}$$

On the other hand, recall that since i is good, $\alpha_i \geq H\left(q_i\right) - 2q_i \log|A|$. Recall also that $\delta = \left(\frac{\varepsilon_2}{|A|^{2/\varepsilon_2}}\right)^4$, and therefore $\mathsf{D}_{\mathsf{KL}}\left(\varepsilon_2 \middle\| \sqrt{\delta}\right) \geq 2\log|A|$. Thus, we get a contradiction to (3). From now on we assume

$$\sum_{\text{unpopular } w_{i-1}\text{'s}} \Pr[w_{i-1}] q_i\left(w_{i-1}\right) \leq \varepsilon_2 q_i. \tag{6}$$

This implies that even if the assignment is uniform over the alphabet, the contribution to $\sum \beta_i$ from unpopular prefixes is small:

$$\begin{aligned} \sum_{\text{unpopular } w_{i-1}\text{'s}} \Pr[w_{i-1}] \beta_i\left(w_{i-1}\right) &\leq \sum_{\text{unpopular } w_{i-1}\text{'s}} \Pr[w_{i-1}] q_i\left(w_{i-1}\right) \log|A| \\ &\leq \varepsilon_2 q_i \log|A| \leq \frac{1}{4}\varepsilon_1 q_i \leq \frac{1}{2}\beta_i \end{aligned}$$

where first inequality follows from Fact 4.7, second from (6), third from our setting of $\varepsilon_1 \geq 4\varepsilon_2 \log|A|$, and fourth from $\beta_i \geq \frac{1}{2}\varepsilon_1 q_i$ since i is good. Therefore,

$$\sum_{\text{popular } w_{i-1}\text{'s}} \Pr[w_{i-1}] \beta_i\left(w_{i-1}\right) = \beta_i - \sum_{\text{unpopular } w_{i-1}\text{'s}} \Pr[w_{i-1}] \beta_i\left(w_{i-1}\right) \geq \beta_i/2$$

Using a similar argument, we show that for any popular w_{i-1}, most of the q_i mass is concentrated on its neighbors. Consider any popular w_{i-1}, and let $\mathcal{N}^C\left(w_{i-1}\right)$ denote the complement of $\mathcal{N}\left(w_{i-1}\right)$. Then we can rewrite α_i as:

$$\alpha_i = \sum_{\sigma_{i-1} \in \mathcal{N}(w_{i-1})} \Pr[\sigma_{i-1}] \alpha_i\left(\sigma_{i-1}\right) + \sum_{\sigma_{i-1} \in \mathcal{N}^C(w_{i-1})} \Pr[\sigma_{i-1}] \alpha_i\left(\sigma_{i-1}\right)$$

Notice that since w_{i-1} is popular, $\mathcal{N}^C(w_{i-1})$ has measure at most $\sqrt{\delta}$. Thus, if an ε_2-fraction of the q_i mass is concentrated on $\mathcal{N}^C(w_{i-1})$, we once again (like in (5)) have

$$\alpha_i \leq H(q_i) - q_i \mathsf{D}_{\mathsf{KL}}\left(\varepsilon_2 \middle\| \sqrt{\delta}\right),$$

which would again yield a contradiction to i being a good variable. Therefore every popular prefix also satisfies the q_i-weighted condition on the degree:

$$\sum_{\sigma_{i-1} \in \mathcal{N}(w_{i-1})} \Pr[\sigma_{i-1}] q_i(\sigma_{i-1}) \geq (1-\varepsilon_2) q_i \tag{7}$$

Recall that a prefix w_{i-1} is good if it also satisfies $\beta_i(w_{i-1}) \geq \varepsilon_3 \cdot q_i(w_{i-1})$. Fortunately, prefixes that violate this condition (i.e. those with small $\beta_i(w_{i-1})$), cannot account for much of the weight on β_i:

$$\sum_{\beta_i(w_{i-1}) < \varepsilon_3 q_i(w_{i-1})} \Pr[w_{i-1}] \beta_i(w_{i-1}) \leq \varepsilon_3 q_i.$$

Since i is good and $\varepsilon_1 \geq 8\varepsilon_3$, this implies:

$$\sum_{\text{good } w_{i-1}\text{'s}} \Pr[w_{i-1}] \beta_i(w_{i-1}) \geq \beta_i/2 - \varepsilon_3 q_i \geq \beta_i/4$$

since

$$\varepsilon_3 q_i \leq \frac{1}{8}\varepsilon_1 q_i \leq \frac{1}{4}\beta_i$$

where last inequality follows from i being good. □

Corollary 4.15. *For every good index i,*

$$\sum_{\textit{good } w_{i-1}\textit{'s}} \Pr[w_{i-1}] q_i(w_{i-1}) \geq \frac{\varepsilon_1}{8 \log |A|} q_i.$$

Proof.

$$\begin{aligned}
\sum_{\text{good } w_{i-1}\text{'s}} \Pr[w_{i-1}] q_i(w_{i-1}) &\geq \sum_{\text{good } w_{i-1}\text{'s}} \Pr[w_{i-1}] \beta_i / \log |A| && \text{(Fact 4.7)} \\
&\geq \beta_i / (4 \log |A|) && \text{(Claim 4.14)} \\
&\geq \frac{\varepsilon_1}{8 \log |A|} q_i && \text{(Definition of good } i\text{)}
\end{aligned}$$

□

With Claim 4.12 and Corollary 4.15, we are ready to prove the main lemma of this section:

Lemma 4.16 (Labeling Entropy Bound)**.** *If* $\sum_i H(Y_i | X_{\leq i}, Y_{<i}) > \frac{5\varepsilon_1 \log k}{\log n}$*, then* $\mathrm{den}(S) < 1 - \delta$*.*

Proof. Assume for a contradiction that $\mathrm{den}(S) \geq 1-\delta$. For prefix w_{i-1}, let $\mathcal{D}_{w_{i-1}}$ denote the induced distribution on labels to the i-th variable, conditioned on w_{i-1} and $x_i = 1$. (If $q_i(w_{i-1}) = 0$, take an arbitrary distribution.) After revealing each variable i, the loss in prefix density is given by the

probability of "fresh violations": the sum over all prefix edges (w_{i-1}, σ_{i-1}) of the probability that they assign different labels to the i-th variable:

$$\mathrm{den}(G_i) - \mathrm{den}(G_{i+1}) = \sum_{w_{i-1}} \sum_{\sigma_{i-1} \in \mathcal{N}(w_{i-1})} \Big(\Pr[w_{i-1}] \Pr[\sigma_{i-1}] q_i(w_{i-1}) q_i(\sigma_{i-1}) \Big) \Pr_{\substack{Y_i \sim \mathcal{D}_{w_{i-1}} \\ Y_i' \sim \mathcal{D}_{\sigma_{i-1}}}} [Y_i \neq Y_i'] \quad (8)$$

We now lower-bound $\Pr_{\mathcal{D}_{w_{i-1}} \times \mathcal{D}_{\sigma_{i-1}}}[Y_i \neq Y_i']$ for good w_{i-1} (notice that we assume nothing about σ_{i-1}). A simple calculation shows that for $\kappa < 1/2$, if

$$\beta_i(w_{i-1}) \geq (\kappa \log |A| - \kappa \log \kappa - (1-\kappa) \log(1-\kappa))\, q_i(w_{i-1}),$$

then the probability mass (under $\mathcal{D}(w_{i-1})$) on the most common label is at most $1-\kappa$. Observe that this probability is an upper bound on $\Pr_{\mathcal{D}_{w_{i-1}} \times \mathcal{D}_{\sigma_{i-1}}}[Y_i = Y_i']$. For good w_{i-1}, we indeed have

$$\beta_i(w_{i-1}) \geq \varepsilon_3 q_i(w_{i-1}) \geq (\varepsilon_4 \log |A| - \varepsilon_4 \log \varepsilon_4 - (1-\varepsilon_4) \log(1-\varepsilon_4))\, q_i(w_{i-1}),$$

where the second inequality follows from choice of ε_4. Therefore $\Pr_{\mathcal{D}_{w_{i-1}} \times \mathcal{D}_{\sigma_{i-1}}}[Y_i \neq Y_i'] \geq \varepsilon_4$.

We now have, for every good index i,

$$\begin{aligned}
\mathrm{den}(G_i) - \mathrm{den}(G_{i+1}) &\geq \sum_{\text{good } w_{i-1}\text{'s}} \sum_{\sigma_{i-1} \in \mathcal{N}(w_{i-1})} \Big(\Pr[w_{i-1}] \Pr[\sigma_{i-1}] q_i(w_{i-1}) q_i(\sigma_{i-1}) \Big) \varepsilon_4 && \text{(Eq. (8))} \\
&\geq \varepsilon_4 q_i (1-\varepsilon_2) \sum_{\text{good } w_{i-1}\text{'s}} \Pr[w_{i-1}] q_i(w_{i-1}) && \text{(Definition of good prefix)} \\
&\geq \frac{\varepsilon_1 \varepsilon_4}{10 \log |A|} q_i^2 && \text{(Corollary 4.15 + } \varepsilon_2 < 0.2\text{)}
\end{aligned}$$

Finally, summing over all good i's, we get a negative density for S, which is, of course, a contradiction.

$$\begin{aligned}
1 - \mathrm{den}(S) &\geq \mathrm{den}(G_1) - \mathrm{den}(G_{n+1}) && \text{(Observation 4.5)} \\
&= \sum_i \mathrm{den}(G_i) - \mathrm{den}(G_{i+1}) && \text{(telescoping sum)} \\
&\geq \sum_{\text{good } i\text{'s}} \mathrm{den}(G_i) - \mathrm{den}(G_{i+1}) \\
&\geq \sum_{\text{good } i\text{'s}} \left(\frac{\varepsilon_1 \varepsilon_4}{10 \log |A|} \right) q_i^2 \\
&\geq \left(\frac{\varepsilon_1^3 \varepsilon_4}{250 \log^3 |A|} \right) \rho^2/n = \Omega(\rho^2/n). && \text{(Claim 4.12)}
\end{aligned}$$

□

4.3 2CSP violation

Intuitively, if $\sum_i H(X_i | X_{<i}, Y_{<i})$ is large, then the subgraph approximately corresponds to assignments to all subsets in $\binom{[n]}{\rho}$. More specifically, in this section we show that most of the constraints appear approximately as frequently as we expect. Since in any assignment, a constant fraction

of them must be violated, this implies (eventually) that a constant fraction of the edges have a violated constraint.

First, we show that most of the variables appear approximately as frequently as we expect by showing that most of them are "typical."

Definition 4.17 (Typical variables)**.** *Prefix w_{i-1} is* typical *if*

$$(1-\varepsilon_5)\cdot\rho/n < \Pr[X_i = 1|w_{i-1}] < (1+\varepsilon_5)\cdot\rho/n,$$

where ε_5 is some constant such that $\left(\frac{\log e}{8}\right)\varepsilon_5^4 > 14\varepsilon_1$.
Similarly, we say that variable x_i is typical *if*

$$\sum_{\text{typical } w_{i-1}\text{'s}} \Pr[w_{i-1}] \geq 1-\varepsilon_5$$

Claim 4.18. *If $\sum_i H(X_i|X_{<i},Y_{<i}) \geq \left(1-\frac{6\varepsilon_1}{\log n}\right)\log k = \log k - \Theta(\rho)$, then all but at most $\varepsilon_5 n$ variables are typical.*

Proof. Assume by contradiction that there are $\varepsilon_5 n$ atypical variables. That is $\varepsilon_5 n/2$ variables x_i appear with probability at least $(1+\varepsilon_5)\cdot\rho/n$ (or at most $(1-\varepsilon_5)\cdot\rho/n$) for an $(\varepsilon_5/2)$-fraction of the prefixes w_{i-1}. Now, subject only to this constraint and maintaining the correct expected number of variables in each vertex, the entropy is maximized by spreading the $(\varepsilon_5^3/4)$-loss in frequency evenly across all other prefixes and variables. That is on the atypical prefixes, labels are assigned with probability $(1+\varepsilon_5)\,\rho/n$, and with probability $\left(1-\frac{\varepsilon_5^3/4}{1-\varepsilon_5^2/4}\right)\rho/n$ on the rest. Thus,

$$\sum_i H(X_i|X_{<i},Y_{<i}) < \frac{\varepsilon_5^2}{4}n\cdot H\left((1+\varepsilon_5)\,\rho/n\right) + \left(1-\frac{\varepsilon_5^2}{4}\right)nH\left(\left(1-\frac{\varepsilon_5^3/4}{1-\varepsilon_5^2/4}\right)\rho/n\right)$$

Recall from Fact 4.10 the expansion of the entropy function:

$$H\left(\frac{1+\upsilon}{n}\right) = H\left(\frac{1}{n}\right) - \frac{\upsilon}{n}\log\frac{1}{n} - \left(\frac{\log e}{2}\right)\frac{\upsilon^2}{n} + O\left(n^{-2}\right) + O\left(\frac{\upsilon^3}{n}\right)$$

Therefore,

$$\begin{aligned}\sum_i H(X_i|X_{<i},Y_{<i}) \;&<\; \frac{\varepsilon_5^2}{4}n\left[H\left(\frac{\rho}{n}\right) - \varepsilon_5\frac{\rho}{n}\log\frac{\rho}{n} - \left(\frac{\log e}{2}\right)\frac{\rho}{n}\cdot\varepsilon_5^2 + O\left(\left(\frac{\rho}{n}\right)^2\right) + O\left(\frac{\rho}{n}\varepsilon_5^3\right)\right]\\ &\quad + \left(1-\frac{\varepsilon_5^2}{4}\right)n\left[H\left(\frac{\rho}{n}\right) + \left(\frac{\varepsilon_5^3/4}{1-\varepsilon_5^2/4}\right)\frac{\rho}{n}\log\frac{\rho}{n} + O\left(\left(\frac{\rho}{n}\right)^2\right) + O\left(\frac{\rho}{n}\varepsilon_5^6\right)\right]\\ &= \; n\left[H\left(\frac{\rho}{n}\right) - \left(\frac{\log e}{8}\right)\frac{\rho}{n}\cdot\varepsilon_5^4 + O\left(\left(\frac{\rho}{n}\right)^2\right) + O\left(\frac{\rho}{n}\varepsilon_5^5\right)\right]\end{aligned}$$

Recall that $-2\log\frac{\rho}{n} < \log n$. Thus for $\left(\frac{\log e}{8}\right)\varepsilon_5^4 > 14\varepsilon_1$, we have that

$$\left(\frac{\log e}{8}\right)\frac{\rho}{n}\cdot\varepsilon_5^4 - O\left(\left(\frac{\rho}{n}\right)^2\right) - O\left(\frac{\rho}{n}\varepsilon_5^5\right) > \frac{\rho}{n}\cdot 12\varepsilon_1 > -\frac{\rho}{n}\log\frac{\rho}{n}\cdot 24\varepsilon_1/\log n > (12\varepsilon_1/\log n)\,H\left(\frac{\rho}{n}\right)$$

and therefore,

$$\sum_i H(X_i|X_{<i},Y_{<i}) < (1-12\varepsilon_1/\log n)\,nH\left(\frac{\rho}{n}\right) < (1-6\varepsilon_1/\log n)\log k,$$

where the second inequality follows from Fact 4.8. Thus we have reached a contradiction. Notice that the $\left(\frac{\log e}{8}\right)\frac{\rho}{n}\cdot\varepsilon_5^4$ term of missing entropy is symmetric (but not the negligible higher order terms); i.e. the same derivation can be used to show a contradiction when many variables appear with probability less than $(1-\varepsilon_5)\,\rho/n$. □

Definition 4.19. *Let $\mathcal{I}(u,v)$ be defined as the number of (i,j) pairs such that*

- *In the original* **2CSP** *instance ψ, there exists an edge (constraint) between typical variables x_i and x_j.*
- *$X_i = 1$ for u and $X_j = 1$ for v.*
- *u_{i-1} and v_{j-1} are typical prefixes, where u_{i-1} denotes the prefix represented by u for $X_{<i}, Y_{<i}$, similarly for v_{j-1}.*

Intuitively, $\mathcal{I}(u,v)$ is the number of "tests" of **2CSP**-constraints between vertices u,v, when restricting to typical prefixes and variables. We now use the properties of typical prefixes and constraints to show that $\mathcal{I}(u,v)$ behaves "nicely".

Claim 4.20. *$\mathbb{E}_{u,v}\left[\mathcal{I}\left(u,v\right)\right] \geq (1-\varepsilon_7)\,\rho^2/n$ and $\mathbb{E}_{u,v}\left[\mathcal{I}^2\left(u,v\right)\right] \leq (1+2\varepsilon_7)\,d^4\left(\mathbb{E}_{u,v}\left[\mathcal{I}\left(u,v\right)\right]\right)^2$, where ε_7 is some constant $\varepsilon_7 \geq 6\varepsilon_5 + \Theta(\varepsilon_5^2)$.*

Proof. For any $i,j \in [n]$, we say that $i \in \mathcal{N}^{2CSP}(j)$ if there is a constraint on (x_i, x_j). For the proof of this claim, we also abuse notation and denote $i \in v$ when i is typical, v_{i-1} is a typical prefix, and $X_i = 1$ for v. We also say that $i \in \mathcal{N}(u)$ if i is a typical variable, $i \in \mathcal{N}^{2CSP}(j)$, and $j \in u$ (for some $j \in [n]$). (Do not confuse this notation with prefix neighborhood in the prefix graph.) We can now lower bound the expectation of $\mathcal{I}\left(u,v\right)$ as:

$$\mathbb{E}_{u,v}\left[\mathcal{I}\left(u,v\right)\right] \geq \mathbb{E}_u\left[\sum_{i\in\mathcal{N}(u)} \Pr_v\left[i \in v\right]\right]$$

Notice that this bound may not be tight since any $i \in v$ can potentially have d neighbors in u. Thus our upper bound is:

$$\mathbb{E}_{u,v}\left[\mathcal{I}\left(u,v\right)\right] \leq d\cdot\mathbb{E}_u\left[\sum_{i\in\mathcal{N}(u)} \Pr_v\left[i \in v\right]\right]$$

By definition of typical variables, for each typical i, $i \in v$ with probability at least $(1-\varepsilon_5)^2\,\rho/n$; thus,

$$\mathbb{E}_{u,v}\left[\mathcal{I}\left(u,v\right)\right] \geq \mathbb{E}_u\left[\sum_{i\in\mathcal{N}(u)} (1-\varepsilon_5)^2\,\rho/n\right] = (1-\varepsilon_5)^2\,\rho/n\cdot\mathbb{E}_u\left[\left|\mathcal{N}\left(u\right)\right|\right] \tag{9}$$

All but $\varepsilon_5 n$ variables are typical, so all but $2\varepsilon_5 n$ variables are typical and have at least one typical neighbor. We restrict our attention to the set of such variables and fix one typical neighbor for each; this neighbor appears in u with probability at least $(1-\varepsilon_5)^2\,\rho/n$. Therefore,

$$\mathbb{E}_u\left[|\mathcal{N}(u)|\right] \geq (1-2\varepsilon_5)n\cdot\left((1-\varepsilon_5)^2\rho/n\right) \geq (1-4\varepsilon_5)\rho \tag{10}$$

Combining (9) and (10), we get the desired bound:

$$\mathbb{E}_{u,v}\left[\mathcal{I}\left(u,v\right)\right] \geq \left((1-\varepsilon_5)^2\,\rho/n\right)(1-4\varepsilon_5)\rho \geq (1-\varepsilon_7)\,\rho^2/n. \tag{11}$$

Similarly, for the variance we have

$$\begin{aligned}
\mathbb{E}_{u,v}\left[\mathcal{I}^2(u,v)\right] &\leq d^2 \cdot \mathbb{E}_{u,v}\left(\sum_{i\in v\cap\mathcal{N}(u)} 1\right)^2 \\
&= d^2 \cdot \mathbb{E}_{u,v}\left[\sum_{i\neq j\in v\cap\mathcal{N}(u)} 1 + \sum_{i\in v\cap\mathcal{N}(u)} 1\right] \\
&\leq d^2 \cdot \mathbb{E}_{u}\left[2\sum_{i<j\in\mathcal{N}(u)} \Pr_v[i\in v]\Pr_v[j\in v \mid i\in v]\right] + d^2\cdot\mathbb{E}_{u,v}\left[\mathcal{I}(u,v)\right].
\end{aligned}$$

Since for every prefix, each variable receives a typical assignment with probability at most $(1+\varepsilon_5)\cdot\rho/n$, we have that

$$\begin{aligned}
\mathbb{E}_{u,v}\left[\mathcal{I}^2(u,v)\right] &\leq 2d^2\cdot\mathbb{E}_u\left[\sum_{i<j\in\mathcal{N}(u)}\left((1+\varepsilon_5)\cdot\rho/n\right)^2\right] + d^2\cdot\mathbb{E}_{u,v}\left[\mathcal{I}(u,v)\right] \\
&\leq \left((1+\varepsilon_5)\cdot\rho/n\right)^2\cdot 2d^2\cdot\mathbb{E}_u\binom{|\mathcal{N}(u)|}{2} + d^2\cdot\mathbb{E}_{u,v}\left[\mathcal{I}(u,v)\right] \qquad (12)
\end{aligned}$$

We would like to bound $\mathbb{E}_u\binom{\mathcal{N}(u)}{2}$.

$$\begin{aligned}
\mathbb{E}_u\binom{\mathcal{N}(u)}{2} &= \sum_{i<j}\sum_{k\in\mathcal{N}^{2CSP}(i)}\sum_{l\in\mathcal{N}^{2CSP}(j)}\Pr_u[k\in u]\Pr_u[l\in u\mid k\in u] \\
&= \sum_{i<j}\sum_{\substack{k\in\mathcal{N}^{2CSP}(i)\\ l\in\mathcal{N}^{2CSP}(j)\\ \text{and } k<l}}\Pr_u[k\in u]\Pr_u[l\in u\mid k\in u] \qquad (13) \\
&\quad + \sum_{i<j}\sum_{\substack{k\in\mathcal{N}^{2CSP}(i)\\ l\in\mathcal{N}^{2CSP}(j)\\ \text{and } k>l}}\Pr_u[l\in u]\Pr_u[k\in u\mid l\in u] \qquad (14) \\
&\quad + \sum_{i<j}\sum_{k\in\mathcal{N}^{2CSP}(i)\cap\mathcal{N}^{2CSP}(j)}\Pr_u[k\in u] \qquad (15)
\end{aligned}$$

For the first two summands, we can use the condition on the prefixes to conclude that

$$(13)+(14)\leq\binom{n}{2}d^2\left((1+\varepsilon_5)\cdot\rho/n\right)^2$$

Whereas to bound the third summand we first change the order of summation:

$$\begin{aligned}
(15) &= \sum_k \Pr_u[k\in u]\cdot\left|\left\{(i,j): i\neq j \text{ and } k\in\mathcal{N}^{2CSP}(i)\cap\mathcal{N}^{2CSP}(j)\right\}\right| \\
&\leq \left((1+\varepsilon_5)\cdot\rho\right)\binom{d}{2} = O(\rho)
\end{aligned}$$

Summing the last two inequalities, we have

$$2\cdot\mathbb{E}_u\binom{|\mathcal{N}(u)|}{2}\leq d^2\left((1+\varepsilon_5)\cdot\rho\right)^2 + O(\rho)\leq(1+\varepsilon_5)^3 d^2\rho^2$$

Plugging back into (12):

$$\mathbb{E}_{u,v}\left[\mathcal{I}^2(u,v)\right] \quad \leq \quad (1+\varepsilon_5)^5 d^4\rho^4/n^2 + d^2 \cdot \mathbb{E}_{u,v}\left[\mathcal{I}(u,v)\right]$$

Using (11) and the fact that $\rho = \sqrt{n}\log\log n \gg \sqrt{n}$, this gives

$$\begin{aligned}\mathbb{E}_{u,v}\left[\mathcal{I}^2(u,v)\right] &\leq \frac{d^4(1+\varepsilon_5)^5}{1-\varepsilon_7}\left(\mathbb{E}_{u,v}\left[\mathcal{I}(u,v)\right]\right)^2 + d^2 \cdot \mathbb{E}_{u,v}\left[\mathcal{I}(u,v)\right] \\ &\leq (1+2\varepsilon_7)\, d^4 \left(\mathbb{E}_{u,v}\left[\mathcal{I}(u,v)\right]\right)^2\end{aligned}$$

□

It will also be convenient to count the number of tests between a pair of variables.

Definition 4.21. *For any pair of typical $(i,j) \in \psi$, let $\mathcal{I}^\top(i,j)$ be defined as the number of $(u,v) \in (S \times S)$ pairs such that*

- *$X_i = 1$ for u and $X_j = 1$ for v.*
- *u_{i-1} and v_{j-1} are typical prefixes, where u_{i-1} denotes the prefix represented by u for $X_{<i}, Y_{<i}$, similarly for v_{j-1}.*

We now have two ways to count the total number of tests between typical prefixes to typical variables:

Observation 4.22. $\sum_{(u,v)\in(S\times S)} \mathcal{I}(u,v) = \sum_{(i,j)\in\psi} \mathcal{I}^\top(i,j)$.

Furthermore, since i and j are typical, the number of tests between also behaves "nicely":

Observation 4.23. *For every typical $(i,j) \in \psi$, we have $\mathcal{I}^\top(i,j) \in |S|^2\rho^2/n^2\left[(1-\varepsilon_5)^4, (1+\varepsilon_5)^2\right]$.*

Proof.

$$\begin{aligned}\mathcal{I}^\top(i,j) &= \sum_{\text{typical } u_{i-1}\text{'s}} |S| \cdot \Pr[u_{i-1}] \Pr\left[X_i = 1 \mid u_{i-1}\right] \sum_{\text{typical } v_{j-1}\text{'s}} |S| \cdot \Pr\left[v_{j-1}\right] \Pr\left[X_j = 1 \mid v_{j-1}\right] \\ &\in |S|^2\rho^2/n^2\left[(1-\varepsilon_5)^4, (1+\varepsilon_5)^2\right]\end{aligned}$$

□

Armed with these Claims 4.18 and 4.20 and Observations 4.22 and 4.23, we are now ready to prove the main lemma of this section. Recall that the soundness of the 2CSP we started with is $1-\eta$ for a small constant η.

Lemma 4.24. *If $\sum_i H(X_i|X_{<i}, Y_{<i}) \geq \left(1 - \frac{6\varepsilon_1}{\log n}\right)\log k$, then $\delta(S) < 1-\delta$, where $\delta < \frac{\varepsilon_6^2}{d^4(1+2\varepsilon_7)}$ and $\varepsilon_6 = (\eta/2 - \varepsilon_5)\left(1/|A|^2\right)\frac{(1-\varepsilon_5)^4}{(1+\varepsilon_5)^2}$.*

Proof. Let the *mode assignment* be the assignment $\mathcal{A}\colon [n] \to \Sigma$ which assigns to each variable x_i its most common typical assignment (i.e. assignment after a typcial prefix), breaking ties arbitrarily. In particular, at least $1/|A|$ of the typical assignments for x_i are equal to $\mathcal{A}(i)$. Of course, this assignment cannot satisfy more than a $(1-\eta)$-fraction of the constraints in the original **2CSP**;

after removing the $\varepsilon_5 n$ atypical variables, $(\eta/2 - \varepsilon_5)\, dn$ constraints out of the $dn/2$ constraints must still be unsatisfied.

Recall that the number of tests for each constraint over typcial variables, $\mathcal{I}^\top(i,j)$, is approximately the same for every pair of (i,j) — up to a $\frac{(1-\varepsilon_5)^4}{(1+\varepsilon_5)^2}$-multiplicative factor (Observation 4.23). Therefore, the total fraction of tests over unsatisfied constraints, out of all tests, is approximately proportional to the fraction of unsatisfied constraints:

$$\begin{aligned}
\sum_{\text{typcial, unsatisfied } (i,j)\text{'s}} \mathcal{I}^\top(i,j) &\geq \frac{(1-\varepsilon_5)^4}{(1+\varepsilon_5)^2} \cdot \frac{\left|\left\{\text{typical, unsatisfied } (i,j)\text{'s}\right\}\right|}{\left|\left\{\text{typical } (i,j) \in \psi\right\}\right|} \cdot \sum_{(i,j)\in\psi} \mathcal{I}^\top(i,j) \\
&\geq \frac{(1-\varepsilon_5)^4}{(1+\varepsilon_5)^2} \cdot \frac{(\eta/2 - \varepsilon_5)\, dn}{dn/2} \cdot \sum_{(i,j)\in\psi} \mathcal{I}^\top(i,j) \\
&= \frac{(1-\varepsilon_5)^4}{(1+\varepsilon_5)^2} \cdot (\eta - 2\varepsilon_5) \cdot \sum_{(u,v)\in(S\times S)} \mathcal{I}(u,v) \qquad \text{(Observation 4.22)}
\end{aligned}$$

For each such pair (i,j), on at least a $1/|A|^2$-fraction of the tests both variables receive the mode assignment, so the constraint is violated[2]. Thus the total number of violations is at least $\varepsilon_6 \sum_{(u,v)\in(S\times S)} \mathcal{I}(u,v)$ (where $\varepsilon_6 = (\eta/2 - \varepsilon_5)\left(1/|A|^2\right)\frac{(1-\varepsilon_5)^4}{(1+\varepsilon_5)^2}$).

Finally, we show that so many violations cannot concentrate on less than a δ-fraction of the pairs $u, v \in S$; otherwise:

$$\begin{aligned}
\sum_{(u,v)\in(S\times S)\setminus E} \mathcal{I}^2(u,v) &\geq \frac{1}{\delta\,|S|^2}\left(\sum_{(u,v)\in(S\times S)\setminus E} \mathcal{I}(u,v)\right)^2 \qquad \text{(Cauchy-Schwartz)} \\
&\geq \frac{1}{\delta\,|S|^2}\left(\varepsilon_6 \sum_{(u,v)\in(S\times S)} \mathcal{I}(u,v)\right)^2 \\
&= \frac{|S|^2\varepsilon_6^2}{\delta}\left(\mathbb{E}_{u,v}\left[\mathcal{I}(u,v)\right]\right)^2;
\end{aligned}$$

yet by Claim 4.20,

$$\sum_{(u,v)\in(S\times S)\setminus E} \mathcal{I}^2(u,v) \leq \sum_{(u,v)\in S\times S} \mathcal{I}^2(u,v) \leq (1+2\varepsilon_7)\, d^4|S|^2\left(\mathbb{E}_{u,v}\left[\mathcal{I}(u,v)\right]\right)^2.$$

Thus we have a contradiction since $d^4(1+2\varepsilon_7) < \varepsilon_6^2/\delta$ by our setting of δ. Therefore we have **2CSP**-violations in more than a δ-fraction of the pairs $u, v \in S$. □

With Lemma 4.16 and Lemma 4.24, we can now complete the proof of Theorem 4.1.

Theorem 4.1 (Soundness)**.** *If* $\mathrm{OPT}(\psi) < 1-\eta$*, then* $\forall S \subset V$ *of size* $k' = k \cdot |V|^{-\varepsilon_0/\log\log|V|}$*,* $\mathrm{den}(S) < 1-\delta$ *for some constant* δ*.*

Proof. Recall that $\sum_i \alpha_i + \beta_i = \log k' \geq (1 - \frac{\varepsilon_1}{\log n})\log k$ by Fact 4.9. If $\sum_i \beta_i > (\frac{5\varepsilon_1}{\log n})\log k$, then by Lemma 4.16, $\delta(S) < 1-\delta$. Otherwise, if $\sum_i \alpha_i > (1 - \frac{6\varepsilon_1}{\log n})\log k$, by Lemma 4.24, $\delta(S) < 1-\delta$. □

[2]We remark that a more careful analysis of the expected number of violations would allow one to save an $|A|^2$-factor in the value of ε_6. Since it does not qualitatively affect the result, we opt for the simpler analysis.

References

[AAK+07] Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *STOC*, pages 496–505. ACM, 2007.

[AAM+11] Noga Alon, Sanjeev Arora, Rajsekar Manokaran, Dana Moshkovitz, and Omri Weinstein. Inapproximability of densest κ-subgraph from average case hardness. *Unpublished manuscript*, 2011.

[AGM13] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. *arXiv preprint arXiv:1308.6273*, 2013.

[AGSS12] Sanjeev Arora, Rong Ge, Sushant Sachdeva, and Grant Schoenebeck. Finding overlapping communities in social networks: toward a rigorous approach. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 37–54. ACM, 2012.

[AIM14] Scott Aaronson, Russell Impagliazzo, and Dana Moshkovitz. Am with multiple merlins. In *Computational Complexity (CCC), 2014 IEEE 29th Conference on*, pages 44–55. IEEE, 2014.

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. In *SODA*, pages 594–598, 1998.

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *J. ACM*, 45(3):501–555, 1998.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of NP. *J. ACM*, 45(1):70–122, 1998.

[Bar15a] Siddharth Barman. personal communication, 2015.

[Bar15b] Siddharth Barman. Approximating carathéodory's theorem and nash equilibria. In *STOC*, 2015.

[BBB+13] Maria-Florina Balcan, Christian Borgs, Mark Braverman, Jennifer Chayes, and Shang-Hua Teng. Finding endogenously formed communities. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 767–783. SIAM, 2013.

[BCV+12] Aditya Bhaskara, Moses Charikar, Aravindan Vijayaraghavan, Venkatesan Guruswami, and Yuan Zhou. Polynomial integrality gaps for strong sdp relaxations of densest k-subgraph. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 388–405. SIAM, 2012.

[BKW15] Mark Braverman, Young Kun Ko, and Omri Weinstein. Approximating the best nash equilibrium in $n^{o(\log n)}$-time breaks the exponential time hypothesis. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2015.

[BL13] Maria Florina Balcan and Yingyu Liang. Modeling and detecting community hierarchies. In *Similarity-Based Pattern Recognition*, pages 160–175. Springer, 2013.

[BPR15a] Yakov Babichenko, Christos Papadimitriou, and Aviad Rubinstein. Can Almost Everybody be Almost Happy? PCP for PPAD and the Inapproximability of Nash. In submission, 2015.

[BPR+15b] Ashwinkumar Badanidiyuru, Christos Papadimitriou, Aviad Rubinstein, Lior Seeman, and Yaron Singer. Submodular adaptive seeding. In submission, 2015.

[BR13] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[CLLR15] Wei Chen, Fu Li, Tian Lin, and Aviad Rubinstein. Combining traditional marketing and viral marketing with amphibious influence maximization. In submission, 2015.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[DGGP10] Yael Dekel, Ori Gurel-Gurevich, and Yuval Peres. Finding hidden cliques in linear time with high probability. *CoRR*, abs/1010.2997, 2010.

[Din07] Irit Dinur. The pcp theorem by gap amplification. *Journal of the ACM (JACM)*, 54(3):12, 2007.

[DM15] Yash Deshpande and Andrea Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. *CoRR*, abs/1502.06590, 2015.

[Fei02] Uriel Feige. Relations between average case complexity and approximation complexity. In *STOC*, pages 534–543. ACM Press, 2002.

[FGL+96] Uriel Feige, Shafi Goldwasser, Laszlo Lovász, Shmuel Safra, and Mario Szegedy. Interactive proofs and the hardness of approximating cliques. *Journal of the ACM (JACM)*, 43(2):268–292, 1996.

[FGR+13] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 655–664, 2013.

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

[FKP01] Uriel Feige, Guy Kortsarz, and David Peleg. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.

[FS97] Uriel Feige and Michael Seltser. *On the densest k-subgraph problem*. Citeseer, 1997.

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

[HK11] Elad Hazan and Robert Krauthgamer. How hard is it to approximate the best nash equilibrium? *SIAM J. Comput.*, 40(1):79–91, 2011.

[IP01] Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-sat. *J. Comput. Syst. Sci.*, 62(2):367–375, 2001.

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

[Kar72] Richard M. Karp. Reducibility among combinatorial problems. In *Proceedings of a symposium on the Complexity of Computer Computations, held March 20-22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York.*, pages 85–103, 1972.

[Kho01] Subhash Khot. Improved inaproximability results for maxclique, chromatic number and approximate graph coloring. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*, pages 600–609, 2001.

[Kho06] Subhash Khot. Ruling out ptas for graph min-bisection, dense k-subgraph, and bipartite clique. *SIAM Journal on Computing*, 36(4):1025–1071, 2006.

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

[MPW15] Raghu Meka, Aaron Potechin, and Avi Wigderson. Sum-of-squares lower bounds for planted clique. In *STOC*, 2015.

[RS10] Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 755–764. ACM, 2010.

[Zuc07] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3(1):103–128, 2007.

A PCP theorem

Theorem 2.11 (PCP Theorem [Din07])**.** *Given a* **3SAT** *instance φ of size n, there is a polynomial time reduction that produces a* **2CSP** *instance ψ, with size $|\psi| = n \cdot \operatorname{polylog} n$ variables and constraints, and constant alphabet size such that*

- *(Completeness) If* $\mathrm{OPT}(\varphi) = 1$ *then* $\mathrm{OPT}(\psi) = 1$.
- *(Soundness) If* $\mathrm{OPT}(\varphi) < 1$ *then* $\mathrm{OPT}(\psi) < 1 - \eta$*, for some constant* $\eta = \Omega(1)$
- *(Balance) Every vertex in ψ has degree d for some constant d.*

Proof. We start with the following version of PCP of near linear size.

Theorem A.1 ([Din07], version as in [AIM14])**.** *Given a* **3SAT** *instance φ of size n, there is a polynomial time reduction that produces a* **3SAT** *instance ξ, with size $|\xi| = n \cdot \operatorname{polylog} n$ variables and constraints such that*

- *(Completeness) If* $\mathrm{OPT}(\varphi) = 1$ *then* $\mathrm{OPT}(\xi) = 1$.
- *(Soundness) If* $\mathrm{OPT}(\varphi) < 1$ *then* $\mathrm{OPT}(\xi) < 1 - \varepsilon$*, for some constant* $0 < \varepsilon < 1/8$
- *(Balance) Every clause of ψ involves exactly 3 variables, and every variable of ψ appears in exactly d clauses, for some constant d.*

We use the following definition to reduce ξ given by Theorem A.1 to a **2CSP** instance ψ.

Definition A.2 ([AIM14], Clause/Variable game)**.** *Given a* **3SAT** *instance ξ with n variables $x_1, \ldots x_n$ and m clauses $C_1, \ldots, C_m$, the clause/variable game G_ξ is defined as follows: Referee chooses an index $i \in m$ uniformly at random, then chooses $j \in [n]$ uniformly at random conditioned on x_j or $\overline{x_j}$ appearing in C_i as a literal. He sends i to Alice and j to Bob. Referee accepts if and only if*

- *Alice sends back a satisfying assignment to the variables in C_i.*
- *Bob sends back a value for x_j that agrees with the value sent by Alice.*

In particular, we can think of following explicit reduction.

1. $X = [m]$ represents clauses; $Y = [n]$ represents variables; $A = \{0,1\}^3$ represents assignment to all 3 variables in a clause; $B = \{0,1\}$ represents assignment to a singleton variable.
2. $(i,j) \in E$ if x_j or $\overline{x_j}$ appears in ith clause (C_i).
3. $V_{(i,j)}$ checks for the following :
   - Assignment on $i \in [m]$ indeed satisfies the clause C_i and,
   - Assignment on $i \in [m]$ agrees with the assignment on $j \in [n]$.

The size blowup is indeed only constant, since we have linear number of vertices, and constant alphabet size. Also any vertex in X has degree 3, and any vertex in Y has degree d since we we started with Dinur's PCP. Completeness follows by assigning satisfying assignment for **3SAT** to this **2CSP**. Soundness follows from the following claim:

Claim A.3 ([AIM14])**.** $\mathrm{OPT}(\xi) \leq 1 - \varepsilon$*, then* $\mathrm{OPT}(\psi) \leq 1 - \varepsilon/3$

Proof. Consider fixing an assignment x on Y's. By our assumption on ξ, this violates the clause C_i with probability at least ε over i. And if x violates C_i, regardless of assignments on X, at least one out of 3 edges of $i \in X$ is not satisfied. Therefore, at least $\varepsilon/3$-fraction of the edges are violated, thus $\mathrm{OPT}(\psi) \leq 1 - \varepsilon/3$. □

Now we add trivial constraints (i.e. always satisfying edges) between vertices in X to make the overall graph of ψ d-regular. (we lose bipartite property, which is not necessary in our reduction) Take a regular graph on X with degree $d-3$. Add the edges with constraints on them as trivial constraints to our **2CSP** instance ψ generated via the reduction. Now the graph is indeed d-regular, completeness is preserved since we only added trivial constraints. For soundness, we know that there are now total $3|X| + \frac{d-3}{2}|X|$ edges. Among them $\frac{d-3}{2}|X|$ are always satisfied. Out of $3|X|$ edges, at most $1-\varepsilon/3$ fraction of them are satisfied, i.e. $(3-\varepsilon)|X|$ edges. So the fraction of satisfied edges is at most :

$$\mathrm{OPT}(\psi) \leq \frac{(3-\varepsilon)|X| + \frac{d-3}{2}|X|}{3|X| + \frac{d-3}{2}|X|} = \frac{d+3-2\varepsilon}{d+3} \leq 1 - \frac{\varepsilon}{d} = 1 - \eta$$

□

B Useful approximations

We recall some elementary approximations to logarithms and entropies that will be useful in the analysis.

Fact B.1. (Fact 4.8) *If* $k = \binom{n}{\rho}$ *then,*

$$\log k = nH\left(\frac{\rho}{n}\right) \pm O\left(\log n\right) = \left(\frac{1}{2} - o\left(1\right)\right)\rho \log n$$

Proof. By Stirling's approximation, we have

$$\log n! \quad = \quad n \log n - \left(\log \mathrm{e}\right) n + O\left(\log n\right)$$

Therefore the total entropy is given by

$$\begin{aligned}
\log k &= \log \binom{n}{\rho} \\
&= \log n! - \log \rho! - \log\left(n-\rho\right)! \\
&= n \log n - \rho \log \rho - \left(n-\rho\right) \log\left(n-\rho\right) \pm O\left(\log n\right) \\
&= nH\left(\frac{\rho}{n}\right) \pm O\left(\log n\right),
\end{aligned}$$

For small ε, we have

$$\log\left(1+\varepsilon\right) = \left(\log \mathrm{e}\right)\left(\varepsilon - \frac{\varepsilon^2}{2} + O\left(\varepsilon^3\right)\right);$$

and in particular,

$$\log \frac{n-\rho}{n} = O\left(-\frac{\rho}{n}\right)$$

Therefore,

$$\begin{aligned}
\log k &= \rho \cdot \log \frac{n}{\rho} + \left(n-\rho\right) \cdot \log \frac{n}{n-\rho} + O\left(\log n\right) \\
&= \rho \cdot \left(\frac{1}{2} - o\left(1\right)\right) \log n + \left(n-\rho\right) \cdot O\left(\frac{\rho}{n}\right) + O\left(\log n\right) \\
&= \left(\frac{1}{2} - o\left(1\right)\right) \rho \log n
\end{aligned}$$

□

More useful to us will be the following bounds on $\log k'$:

Fact B.2. (Fact 4.9) *Let* $\varepsilon_1 \geq 5\varepsilon_0$*, and* k, k', V, n, ρ *as specified in the construction. Then,*

$$\log k' \geq \max\left\{\log k, nH\left(\frac{\rho}{n}\right)\right\} - \varepsilon_1 \log k / \log n.$$

In particular, this means that most indices i *should contribute roughly* $H\left(\frac{\rho}{n}\right)$ *entropy to the choice of* v*.*

Proof. Observing that since $k = \binom{n}{\rho}$, we have

$$\log |V| = \log \binom{n}{\rho} + \rho \log |A| = (1 + o(1)) \log k. \tag{16}$$

We also have that

$$\log \log |V| = \log(1 + o(1)) + \log \log k > \log \rho > \frac{1}{2} \log n; \tag{17}$$

where the first inequality follows from Fact 4.8, and the second from the definition of ρ.

Finally, we have

$$\begin{aligned}
\log k' &= \log k - \varepsilon_0 \log |V| / \log \log |V| \\
&\geq \log k - \varepsilon_0 (1 + o(1)) \log k / \frac{1}{2} \log n && \text{(Using (16) and (17))} \\
&\geq \log k - \frac{1}{2} \varepsilon_1 \log k / \log n && \text{(Using } \varepsilon_1 \geq 5\varepsilon_0\text{)}
\end{aligned}$$

Using Fact 4.8 completes the proof. □

We will also need the following bound which relates the entropies of a very biased coin and a slightly less biased one:

Fact B.3. (Fact 4.10)

$$H\left(\frac{1+\upsilon}{n}\right) = H\left(\frac{1}{n}\right) - \frac{\upsilon}{n} \log \frac{1}{n} - (\log \mathrm{e}) \frac{\upsilon^2}{2n} + O\left(n^{-2}\right) + O\left(\frac{\upsilon^3}{n}\right)$$

Proof. By definition,

$$H\left(\frac{1+\upsilon}{n}\right) = -\left(\frac{1+\upsilon}{n}\right) \log \left(\frac{1+\upsilon}{n}\right) - \left(1 - \frac{1+\upsilon}{n}\right) \log \left(1 - \frac{1+\upsilon}{n}\right)$$

In order to relate this quantity to $H\left(\frac{1}{n}\right)$, we rewrite as:

$$\begin{aligned}
H\left(\frac{1+\upsilon}{n}\right) &= -\frac{1}{n} \log \frac{1}{n} - \frac{\upsilon}{n} \log \frac{1}{n} - \left(\frac{1+\upsilon}{n}\right) \cdot \underbrace{\log(1+\upsilon)}_{(\log \mathrm{e})(\upsilon - \upsilon^2/2 + O(\upsilon^3))} \\
&\quad - \left(1 - \frac{1}{n}\right) \log \left(1 - \frac{1}{n}\right) + \upsilon \underbrace{\frac{1}{n} \log \left(1 - \frac{1}{n}\right)}_{O(n^{-2})} - \left(1 - \left(\frac{1+\upsilon}{n}\right)\right) \cdot \underbrace{\log \left(\frac{1 - \left(\frac{1+\upsilon}{n}\right)}{1 - \frac{1}{n}}\right)}_{(\log \mathrm{e})(-(\upsilon/n) - O(\upsilon/n^2))} \\
&= H\left(\frac{1}{n}\right) - \frac{\upsilon}{n} \log \frac{1}{n} - (\log \mathrm{e}) \frac{\upsilon^2}{2n} + O\left(n^{-2}\right) + O\left(\frac{\upsilon^3}{n}\right)
\end{aligned}$$

□

C Small constants in the proof of Theorem 4.1

To help verify the correctness of the proof, we concentrate all the definitions of the small ε's used in the following list:

- $\varepsilon_0 \leq \varepsilon_1/5$
- $\varepsilon_1 \geq 4\varepsilon_2 \log |A| + 8\varepsilon_3$
- ε_2: $\varepsilon_2 < 0.2$, $\delta = \left(\frac{\varepsilon_2}{|A|^{2/\varepsilon_2}}\right)^4$
- $\varepsilon_3 \geq \varepsilon_4 \log |A| - \varepsilon_4 \log \varepsilon_4 - (1-\varepsilon_4)\log(1-\varepsilon_4)$
- $\varepsilon_4 = \omega(n/\rho^2)$
- ε_5: $\left(\frac{\log \mathrm{e}}{8}\right)\varepsilon_5^4 > 14\varepsilon_1$
- ε_6: $\varepsilon_6 = (\eta/2 - \varepsilon_5)\left(1/|A|^2\right)\frac{(1-\varepsilon_5)^4}{(1+\varepsilon_5)^2}$ and $d^4(1+2\varepsilon_7) < \varepsilon_6^2/\delta$
- ε_7: $\varepsilon_7 \geq 6\varepsilon_5 + \Theta(\varepsilon_5^2)$

ECCC ISSN 1433-8092

http://eccc.hpi-web.de