Electronic Colloquium on Computational Complexity, Report No. 195 (2016)

Almost-Polynomial Ratio ETH-Hardness of Approximating Densest k-Subgraph

Pasin Manurangsi[*]
UC Berkeley

November 18, 2016

Abstract

In the Densest k-Subgraph problem, given an undirected graph G and an integer k, the goal is to find a subgraph of G on k vertices that contains maximum number of edges. Even though the state-of-the-art algorithm for the problem achieves only $O(n^{1/4+\varepsilon})$ approximation ratio [BCC+10], previous attempts at proving hardness of approximation, including those under average case assumptions, fail to achieve a polynomial ratio; the best ratios ruled out under any worst case assumption and any average case assumption are only any constant [RS10] and $2^{\Omega(\log^{2/3} n)}$ [AAM+11] respectively.

In this work, we show, assuming the exponential time hypothesis (ETH), that there is no polynomial-time algorithm that approximates Densest k-Subgraph to within $n^{1/(\log\log n)^c}$ factor of the optimum, where $c > 0$ is a universal constant independent of n. In addition, our result has *perfect completeness*, meaning that we prove that it is ETH-hard to even distinguish between the case in which G contains a k-clique and the case in which every induced k-subgraph of G has density at most $1/n^{-1/(\log\log n)^c}$ in polynomial time.

Moreover, if we make a stronger assumption that there is some constant $\varepsilon > 0$ such that no subexponential-time algorithm can distinguish between a satisfiable 3SAT formula and one which is only $(1-\varepsilon)$-satisfiable (also known as Gap-ETH), then the ratio above can be improved to $n^{f(n)}$ for any function f whose limit is zero as n goes to infinity (i.e. $f \in o(1)$).

[*] Email: pasin@berkeley.edu. This material is based upon work supported by the National Science Foundation under Grants No. CCF 1540685 and CCF 1655215

ISSN 1433-8092

1 Introduction

In the Densest k-Subgraph problem, we are given an undirected graph G on n vertices and a positive integer $k \leqslant n$. The goal is to find a set S of k vertices such that the induced subgraph on S has maximum number of edges. Since the size of S is fixed, the problem can be equivalently stated as finding a k-subgraph (i.e. subgraph on k vertices) with maximum density where density[1] of the subgraph induced on S is $\frac{|E(S)|}{\binom{|S|}{2}}$ and $E(S)$ denotes the set of all edges among the vertices in S.

Densest k-Subgraph (DkS), a natural generalization of k-Clique [Kar72], was first formulated and studied by Kortsarz and Peleg [KP93] in the early 90s. Since then, it has been the subject of intense study in the context of approximation algorithm and hardness of approximation [FS97, SW98, FL01, FKP01, AHI02, Fei02, Kho06, GL09, RS10, BCC+10, AAM+11, BCV+12, Bar15, BKRW17]. Despite this, its approximability still remains wide open and is considered by some to be an important open question in approximation algorithms [BCC+10, BCV+12, CL15, BKRW17].

On the approximation algorithm front, Kortsarz and Peleg [KP93], in the same work that introduced the problem, gave a polynomial-time $\tilde{O}(n^{0.3885})$-approximation algorithm for DkS. Feige, Kortsarz and Peleg [FKP01] later provided an $O(n^{1/3-\delta})$-approximation for the problem for some constant $\delta \approx 1/60$. This approximation ratio was the best known for almost a decade[2] until Bhaskara, Charikar, Chlamtac, Feige and Vijayaraghavan [BCC+10] invented a log-density based approach which yielded an $O(n^{1/4+\varepsilon})$-approximation for any constant $\varepsilon > 0$. This remains the state-of-the-art approximation algorithm for DkS.

While the above algorithms demonstrate the main progresses of approximations of DkS in general case over the years, many special cases have also been studied. Most relevant to our work is the case where the optimal k-subgraph has high density, in which better approximations are known [FS97, ST08, MM15, Bar15]. The first and most representative algorithm of this kind is that of Feige and Seltser [FS97], which provides the following guarantee: when the input graph contains a k-clique, the algorithm can find an $(1-\varepsilon)$-dense k-subgraph in $n^{O(\log n/\varepsilon)}$ time. We will refer to this problem of finding densest k-subgraph when the input graph is promised to have a k-clique Densest k-Subgraph *with perfect completeness*. In other words, the Feige-Seltser algorithm is a quasi-polynomial time approximation scheme for DkS with perfect completeness.

Although many algorithms have been devised for DkS, relatively little is known regarding its hardness of approximation. While it is commonly believed that the problem is hard to approximate to within n^c ratio for some constant $c > 0$ [AAM+11, BCV+12], not even a constant factor NP-hardness of approximation is known. To circumvent this, Feige [Fei02] came up with a hypothesis that a random 3SAT formula is hard to refute in polynomial time and proved that, assuming this hypothesis, Densest k-Subgraph is hard to approximate within some constant factor.

Alon, Arora, Manokaran, Moshkovitz and Weinstein [AAM+11] later used a similar conjecture regarding random k-AND to rule out polynomial-time algorithms for DkS with any constant approximation ratio. Moreover, they proved hardnesses of approximation of DkS under the following *Planted Clique Hypothesis* [Jer92, Kuč95]: there is no polynomial-time algorithm that can distinguish between a typical Erdős–Rényi random graph $\mathcal{G}(n, 1/2)$ and one in which a clique of size polynomial in n (e.g. $n^{1/3}$) is planted. Assuming this hypothesis, Alon et al. proved that no polynomial-time algorithm approximates DkS to within any constant factor. They also showed that, when the hypothesis is strengthened to rule out not only polynomial-time but also super-polynomial time algorithms for the Planted Clique problem, their inapproximability guarantee for DkS can be improved. In particular, if one assumes that no $n^{O(\sqrt{\log n})}$-time algorithm solves the Planted Clique problem, then $2^{\Omega(\log^{2/3} n)}$-approximation for DkS cannot be achieved in polynomial time.

There are also several inapproximability results of DkS based on worst-case assumptions. Khot [Kho06] showed, assuming NP $\nsubseteq$ BPTIME$(2^{n^{\varepsilon}})$ for some constant $\varepsilon > 0$, that no (possibly randomized) polynomial-time algorithm can approximate DkS to within $(1+\delta)$ factor where $\delta > 0$ is a constant depending only on ε;

[1] It is worth noting that sometimes density is defined as $|E(S)|/|S|$. For DkS, both definitions of density result in the same objective since $|S| = k$ is fixed. However, our notion is more convenient to deal with as it always lies in $[0, 1]$.

[2] Around the same time as Bhaskara et al.'s work [BCC+10], Goldstein and Langberg [GL09] presented an algorithm with approximation ratio $O(n^{0.3159})$, which is slightly better than [FKP01] but is worse than [BCC+10].

the proof is based on a construction of a "quasi-random" PCP, which is then used in placed of a random 3SAT in a reduction similar to that from [Fei02].

While no inapproximability of DkS is known under the Unique Games Conjecture, Raghavendra and Steurer [RS10] showed that a strengthened version of it, in which the constraint graph is required to satisfy a "small-set expansion" property, implies that DkS is hard to approximate to within any constant ratio.

Recently, Braverman et al. [BKRW17], showed, under the exponential time hypothesis (ETH), which will be stated shortly, that, for some constant $\varepsilon > 0$, no $n^{\tilde{O}(\log n)}$-time algorithm can approximate Densest k-Subgraph with perfect completeness to within $(1+\varepsilon)$ factor. It is worth noting here that their result matches almost exactly with the previously mentioned Feige-Seltser algorithm [FS97].

Since none of these inapproximability results achieve a polynomial ratio, there have been efforts to prove better lower bounds for more restricted classes of algorithms. For example, Bhaskara et al. [BCV$^+$12] provided polynomial ratio lower bounds against strong SDP relaxations of DkS. Specifically, for the Sum-of-Square hierarchy, they showed integrality gaps of $n^{2/53-\varepsilon}$ and n^{ε} against $n^{\Omega(\varepsilon)}$ and $n^{1-O(\varepsilon)}$ levels of the hierarchy respectively. (See also [Man15, CMMV17] in which 2/53 in the exponent was improved to 1/14.) Unfortunately, it is unlikely that these lower bounds can be translated to inapproximability results and the question of whether any polynomial-time algorithm can achieve subpolynomial approximation ratio for DkS remains an intriguing open question.

1.1 Our Results

In this work, we rule out, under the exponential time hypothesis (i.e. there is no subexponential time algorithm that solves 3SAT; see Hypothesis 4), polynomial-time approximation algorithms for Densest k-Subgraph (even with perfect completeness) with slightly subpolynomial ratio:

Theorem 1 *There is a constant $c > 0$ such that, assuming ETH, no polynomial-time algorithm can, given a graph G on n vertices and a positive integer $k \leqslant n$, distinguish between the following two cases:*

- *There exist k vertices of G that induce a k-clique.*
- *Every k-subgraph of G has density at most $n^{-1/(\log\log n)^c}$.*

If we assume a stronger assumption that it takes exponential time to even distinguish between a satisfiable 3SAT formula and one which is only $(1-\varepsilon)$-satisfiable for some constant $\varepsilon > 0$ (also known as Gap-ETH; see Hypothesis 5), then the ratio can be improved to $n^{f(n)}$ for any[3] $f \in o(1)$:

Theorem 2 *For every function $f \in o(1)$, assuming Gap-ETH, no polynomial-time algorithm can, given a graph G on n vertices and a positive integer $k \leqslant n$, distinguish between the following two cases:*

- *There exist k vertices of G that induce a k-clique.*
- *Every k-subgraph of G has density at most $n^{-f(n)}$.*

We remark that, for DkS with perfect completeness, the FS algorithm [FS97] mentioned earlier can achieve an n^{ε}-approximation in time $n^{O(1/\varepsilon)}$ for every $\varepsilon > 0$. Hence, the ratios in our theorems cannot be improved to some fixed polynomial and the ratio in Theorem 2 is tight in this sense.

Comparison to Previous Results. In terms of the inapproximability ratio, the ratios ruled out in this work are almost polynomial and provides a vast improvement over previous results. Prior to our result, the best known ratio ruled out under any worst case assumption is only any constant [RS10] and the best ratio ruled out under any average case assumption is only $2^{\Omega(\log^{2/3} n)}$ [AAM$^+$11]. In addition, our result also has perfect completeness, which was only achieved in [BKRW17] under ETH and in [AAM$^+$11] under the Planted Clique Hypothesis but not in [Kho06, Fei02, RS10].

Regarding the assumptions our result is based upon, the average case assumptions used in [Fei02, AAM$^+$11] are incomparable to ours. The assumption NP $\not\subseteq$ BPTIME$(2^{n^{\varepsilon}})$ used in [Kho06] is also incomparable to ours since, while not stated explicitly, ETH (Gap-ETH) by default focuses only on deterministic algorithms and our reduction is also deterministic. The Unique Games with Small Set Expansion Conjecture used in [RS10] is again incomparable to ours as it is a statement whether a specific problem is NP-hard. Finally, although

[3] Recall that $f \in o(1)$ if and only if $\lim_{n\to\infty} f(n) = 0$.

Braverman et al.'s result [BKRW17] also relies on ETH, its relation to our result is more subtle. Specifically, their reduction time is only $2^{\tilde{\Theta}(\sqrt{m})}$ where m is the number of clauses, meaning that the assumption they need to rule out a constant ratio polynomial-time approximation for DkS is in fact 3SAT $\notin$ DTIME$(2^{\tilde{\Theta}(\sqrt{m})})$. However, as we will see later (Theorem 8), even to achieve a constant gap, our reduction time is at least $2^{\tilde{\Theta}(m^{3/4})}$. Hence, if 3SAT somehow ends up in DTIME$(2^{\tilde{\Theta}(m^{3/4})})$ but outside of DTIME$(2^{\tilde{\Theta}(\sqrt{m})})$, then their result will still hold whereas ours will not even imply constant ratio inapproximability for DkS.

Implications of Our Results. One of the reasons that Densest k-Subgraph has received significant attention in approximation algorithm community is due to its connections to many other problems. Most relevant to our work are the problems which generalize DkS and those with known reductions from DkS that preserve approximation ratio within polynomial[4]. These problems include Densest At-Most-k-Subgraph [AC09], Smallest m-Edge Subgraph [CDK12], Steiner k-Forest [HJ06] and Quadratic Knapsack [Pis07]. For brevity, we do not give definitions of the problems here and we refer the readers to cited sources for their definitions and, in the cases where the connections to DkS are not apparent, reductions from DkS to the respective problems. We also note that this list is by no means exhaustive and there are indeed numerous other problems with similar known connections to DkS (see e.g. [HJL$^+$06, KS07, KMNT11, CHK11, HIM11, LNV14, CLLR15, CL15, CZ15, SFL15, TV15, CDK$^+$16, CMVZ15, Lee16]). Our results also imply hardness of approximation with similar ratio to DkS for such problems:

Corollary 3 *For some constant $c > 0$, assuming ETH, there is no polynomial-time $n^{1/(\log\log n)^c}$-approximation algorithm for* Densest At-Most-k-Subgraph*,* Smallest m-Edge Subgraph*,* Steiner k-Forest*,* Quadratic Knapsack*. Moreover, for any function $f \in o(1)$, assuming Gap-ETH, there is no polynomial-time $n^{f(n)}$-approximation algorithm for any of these problems.*

2 Preliminaries and Notations

Before we proceed to describe the reduction and prove our main theorems, we first define additional notations and provide some preliminaries. Throughout the paper, $\exp(x)$ and $\log(x)$ denote e^x and $\log_2(x)$ respectively. We use $\operatorname{polylog} n$ as a shorthand for $O(\log^c n)$ for some absolute constant c. For any set S, $\mathscr{P}(S)$ denotes the power set of S, i.e., $\mathscr{P}(S) = \{T \mid T \subseteq S\}$. Moreover, for any non-negative integer $t \leqslant |S|$, we use $\binom{S}{t}$ to denote the collection of all subsets of S of size t, i.e., $\binom{S}{t} = \{T \in \mathscr{P}(S) \mid |T| = t\}$.

For an undirected graph $G = (V, E)$ and a positive integer $t \leqslant |V|$, $(L, R) \in \binom{V}{t} \times \binom{V}{t}$ is said to be a *copy* or an *occurrence* of a t-biclique (or $K_{t,t}$) in G if every $(u, v) \in L \times R$ is an edge. The number of copies of $K_{t,t}$ in G is the number of all such (L, R)'s. G is said to be $K_{t,t}$*-free* if G contains no copy of $K_{t,t}$.

2.1 Exponential Time Hypotheses

One of our results is based on the exponential time hypothesis, a conjecture proposed by Impagliazzo and Paturi [IP01] which asserts that 3SAT cannot be solved in subexponential time:

Hypothesis 4 (Exponential Time Hypothesis (ETH) [IP01]) *No algorithm that can decide whether any 3SAT formula with m clauses[5] is satisfiable runs in $2^{o(m)}$ time.*

Another hypothesis used in this work is Gap-ETH, a strengthened version of the ETH, which essentially states that even approximating 3SAT to some constant ratio takes exponential time:

Hypothesis 5 (Gap Exponential Time Hypothesis (Gap-ETH) [Din16, MR16]) *There exists a constant $\varepsilon > 0$ such that no algorithm that can, given a 3SAT formula ϕ with m clauses[6], distinguish between the case where ϕ is satisfiable and the case where $\operatorname{val}(\phi) \leqslant 1 - \varepsilon$ runs in $2^{o(m)}$ time. Here $\operatorname{val}(\phi)$ denote the maximum fraction of clauses of ϕ satisfied by any assignment.*

[4] These are problems whose $O(\rho)$-approximation gives an $O(\rho^c)$-approximation for DkS for some constant c.

[5] In its original form, the running time lower bound is exponential in the number of variables not the number of clauses; however, thanks to the sparsification lemma of Impagliazzo et al. [IPZ01], both versions are equivalent.

[6] As noted by Dinur [Din16], a subsampling argument can be used to make the number of clauses linear in the number of variables, meaning that the conjecture remains the same even when m denotes the number of variables.

2.2 Nearly-Linear Size PCPs and Subexponential Time Reductions

The celebrated PCP Theorem [AS98, ALM$^+$98], which lies at the heart of virtually all known NP-hardness of approximation results, can be viewed as a polynomial-time reduction from 3SAT to a gap version of 3SAT, as stated below. While this perspective is a rather narrow viewpoint of the PCP Theorem that leaves out the fascinating relations between parameters of PCPs, it will be the most convenient for our purpose.

Theorem 6 (PCP Theorem [AS98, ALM$^+$98]) *For some constant $\varepsilon > 0$, there exists a polynomial-time reduction that takes a 3SAT formula φ and produces a 3SAT formula ϕ such that*

- (Completeness) *if φ is satisfiable, then ϕ is satisfiable, and,*
- (Soundness) *if φ is unsatisfiable, then* $\mathrm{val}(\phi) \leqslant 1 - \varepsilon$.

Following the first proofs of the PCP Theorem, considerable efforts have been made to improve the trade-offs between the parameters in the PCP Theorem. One such direction is to try to reduce the size of the PCP, which, in the context of the above formulation, translates to reducing the size of ϕ relative to φ. On this front, it is known that the size of ϕ can be made as small as nearly-linear in the size of φ [Din07, MR08] (see also [BSS08] which [Din07] builds on). For our purpose, we will use Dinur's PCP Theorem [Din07], which has a blow-up of only polylogarithmic in the size of ϕ:

Theorem 7 (Dinur's PCP Theorem [Din07]) *For some constant $\varepsilon, d > 0$, there exists a polynomial-time reduction that takes a 3SAT formula φ with m clauses and produces another 3SAT formula ϕ with $m' = O(m \operatorname{polylog} m)$ clauses such that*

- (Completeness) *if φ is satisfiable, then ϕ is satisfiable, and,*
- (Soundness) *if φ is unsatisfiable, then* $\mathrm{val}(\phi) \leqslant 1 - \varepsilon$, *and,*
- (Bounded Degree) *each variable of ϕ appears in at most d clauses.*

Note that Dinur's PCP Theorem, combined with ETH, implies a lower bound of $2^{\Omega(m/\operatorname{polylog} m)}$ on the running time of algorithms that solve the gap version of 3SAT, which is only a factor of $O(\operatorname{polylog} m)$ in the exponent off from Gap-ETH. Putting it differently, Gap-ETH is closely related to the question of whether a linear size PCP, one where the size blow-up is only constant instead of polylogarithmic, exists; its existence would mean that Gap-ETH is implied by ETH.

Under the exponential time hypothesis, nearly-linear size PCPs allow us to start with an instance ϕ of the gap version of 3SAT and reduce, in subexponential time, to another problem. As long as the time spent in the reduction is $2^{o(m/\operatorname{polylog} m)}$, we arrive at a lower bound for the problem. Arguably, Aaronson, Impagliazzo and Moshkovitz [AIM14] popularized this method, under the name *birthday repetition*, by using such a reduction of size $2^{\tilde{\Omega}(\sqrt{m})}$ to prove ETH-hardness for free games and dense CSPs. Without going into any detail now, let us mention that the name birthday repetition comes from the use of the birthday paradox in their proof and, since its publication, their work has inspired many inapproximability results [BKW15, Rub15, BPR16, MR16, Rub16b, DFS16, BKRW17]. Our result too is inspired by [AIM14] and, as we will see soon, part of our proof also contains a birthday-type paradox.

3 The Reduction and Proofs of The Main Theorems

The reduction from the gap version of 3SAT to DkS is simple and intuitive. Given a 3SAT formula ϕ on n variables $x_1, \dots, x_n$ and an integer $1 \leqslant \ell \leqslant n$, we construct a graph[7] $G_{\phi,\ell} = (V_{\phi,\ell}, E_{\phi,\ell})$ as follows:

- Its vertex set $V_{\phi,\ell}$ contains all partial assignments to ℓ variables, i.e., each vertex is $\{(x_{i_1}, b_{i_1}), \dots, (x_{i_\ell}, b_{i_\ell})\}$ where $x_{i_1}, \dots, x_{i_\ell}$ are ℓ distinct variables and $b_{i_1}, \dots, b_{i_\ell} \in \{0, 1\}$ are the bits assigned to them.
- Two vertices $\{(x_{i_1}, b_{i_1}), \dots, (x_{i_\ell}, b_{i_\ell})\}$ and $\{(x_{i'_1}, b_{i'_1}), \dots, (x_{i'_\ell}, b_{i'_\ell})\}$ are connected by an edge if and only if (1) the two partial assignments are consistent (i.e. no variable x is assigned 0 in one vertex and 1 in another) and (2) every clause in ϕ all of whose variables are from $x_{i_1}, \dots, x_{i_\ell}, x_{i'_1}, \dots, x_{i'_\ell}$ is satisfied by the partial assignment induced by the two vertices.

[7] For interested readers, we note that our graph is not the same as the FGLSS graph [FGL$^+$91] of the PCP in which the verifier reads ℓ random variables and accepts if no clause is violated; while this graph has the same vertex set as ours, the edges are different since we check that no clause between the two vertices is violated, which is not checked in the FGLSS graph. It is possible to modify our proof to make it work for this FGLSS graph. However, the soundness there would become $2^{-\delta\ell^6/n^5}$, which is worse than $2^{-\delta\ell^4/n^3}$ we have in Theorem 8 for our graph.

Clearly, if ϕ is satisfiable, then the $\binom{n}{\ell}$ vertices corresponding to a satisfying assignment induce a clique. Our main technical contribution is proving that, when $\mathrm{val}(\phi) \leqslant 1 - \varepsilon$, every $\binom{n}{\ell}$-subgraph is sparse:

Theorem 8 *For any $d, \varepsilon > 0$, there is a constant $\delta > 0$ such that, for any sufficiently large n, if ϕ is a 3SAT formula on n variables with $\mathrm{val}(\phi) \leqslant 1 - \varepsilon$ and each variable appears in at most d clauses, then for any integer $n^{3/4} \log n \leqslant \ell \leqslant n/2$, any $\binom{n}{\ell}$-subgraph of $G_{\phi,\ell}$ has density at most $2^{-\delta \ell^4 / n^3}$.*

Before we prove Theorem 8, we remark that there is nothing special about 3SAT; we can start with any boolean CSP and end up with a similar result, albeit the soundness deteriorates as the arity of the CSP grows. However, it is crucial that the variables are boolean; in fact, Braverman et al. [BKRW17] considered a graph similar to ours for 2CSPs but they were unable to achieve subconstant soundness since their variables were not boolean[8]. In particular, there is a non-boolean 2CSP with low value which results in the graph having a biclique of size larger than $\binom{n}{\ell}$ (see Appendix A), i.e., one cannot get an inapproximability ratio more than two starting from a non-boolean CSP.

Once we have Theorem 8, the inapproximability results of DkS (Theorem 1 and 2) can be easily proved by applying the theorem with appropriate choices of ℓ. We defer these proofs to Subsection 3.3. For now, let us turn our attention to the proof of Theorem 8. To prove the theorem, we resort to a well-known result from extremal graph theory called the Kővári-Sós-Turán Theorem:

Theorem 9 (Kővári-Sós-Turán (KST) Theorem [KST54]) *For every positive integer N and $t \leqslant N$, every $K_{t,t}$-free graph on N vertices has at most $O(N^{2-1/t})$ edges (i.e. $O(N^{-1/t})$-dense).*

Thanks to the KST Theorem, to certify that a graph is sparse, it is enough to show that the graph is $K_{t,t}$-free for some small t; observe also that, since the density we get from the theorem is $O(N^{-1/t})$, we need $t \ll \log N$ to get a subconstant bound on the density. Since we want to show that any subgraph of $G_{\phi,\ell}$ of size $\binom{n}{\ell}$ is sparse, it suffices to show that $G_{\phi,\ell}$ is $K_{t,t}$-free for some $t \ll \log \binom{n}{\ell} = \Theta(\ell \log(n/\ell))$. Alas, this is in fact not always true. Fortunately, we can relax the condition of the KST Theorem so that, instead of requiring the graph to be $K_{t,t}$-free, it only requires that the number of occurrences of $K_{t,t}$ in the graph is significantly smaller than that in a complete graph. Formally, this robust version of the theorem is stated below.

Lemma 10 (Robust Kővári-Sós-Turán Theorem) *Let G be any undirected graph on N vertices, t be any positive integer and α be any positive real number. If G contains at most αN^{2t} copies of $K_{t,t}$, then G has at most $2\alpha^{1/t^2} N^2 + tN^{2-1/t}$ edges, i.e., it is $O(\alpha^{1/t^2} + tN^{-1/t})$-dense.*

The parameter α in Lemma 10 can be thought of as the sparsity of $K_{t,t}$ occurrences in G. When G is a complete graph, it contains $2\binom{N}{2t}$ $K_{t,t}$'s, meaning that α is $\Omega_t(1)$. As α decreases, the upper bound on the density of the graph also decreases. This culminates in the bound of $O(tN^{-1/t})$ when $\alpha = 1/N^t$, which is, up to an $O(t)$ factor, the same as that in the original KST Theorem.

We are not aware of any previous result similar to Lemma 10 and we provide its proof, which consists of counting arguments reminiscent of the proof of the original KST Theorem, in Subsection 3.1.

Equipped with Lemma 10, our proof strategy is to bound the number of occurrences of $K_{t,t}$ in $G_{\phi,\ell}$ where t is to be chosen later. To argue this, we will need some additional notations:

- First, let $A_\phi := \{(x_1, 0), (x_1, 1), \dots, (x_n, 0), (x_n, 1)\}$ be the set of all single-variable partial assignments. Observe here that $V_{\phi,\ell} \subseteq \binom{A_\phi}{\ell}$, i.e., each $u \in V_{\phi,\ell}$ is a subset of A_ϕ of size ℓ.
- Let $\mathcal{A} : \mathscr{P}(V_{\phi,\ell}) \to \mathscr{P}(A_\phi)$ be a "flattening" function that, on input $T \subseteq V_{\phi,\ell}$, outputs the set of all single-variable partial assignments that appear in at least one vertex in T. In other words, when each vertex u is viewed as a subset of A_ϕ, we can write $\mathcal{A}(T)$ simply as $\bigcup_{u \in T} u$.
- Finally, let $\mathcal{K}_{t,t} := \{(L, R) \in \binom{V_{\phi,\ell}}{t} \times \binom{V_{\phi,\ell}}{t} \mid \forall u \in L, \forall v \in R, (u, v) \in E_{\phi,\ell}\}$ denote the set of all copies of $K_{t,t}$ in $G_{\phi,\ell}$ and, for each $A_L, A_R \subseteq A_\phi$, let $\mathcal{K}_{t,t}(A_L, A_R) := \{(L, R) \in \mathcal{K}_{t,t} \mid \mathcal{A}(L) = A_L, \mathcal{A}(R) = A_R\}$ denote the set of all $(L, R) \in \mathcal{K}_{t,t}$ with $\mathcal{A}(L) = A_L$ and $\mathcal{A}(R) = A_R$.

[8] Any satisfiable boolean 2CSP is solvable in polynomial time so one cannot start with a boolean 2CSP either.

The number of copies of $K_{t,t}$ in $G_{\phi,\ell}$ can now be written as

$$|\mathcal{K}_{t,t}| = \sum_{A_L, A_R \subseteq A_\phi} |\mathcal{K}_{t,t}(A_L, A_R)|. \tag{1}$$

To bound $|\mathcal{K}_{t,t}|$, we will prove the following bound on $|\mathcal{K}_{t,t}(A_L, A_R)|$.

Lemma 11 *Let ϕ, n, ℓ, d and ε be as in Theorem 8. There exists a constant $\lambda > 0$ depending only on d and ε such that, for any $t > 0$ and any $A_L, A_R \subseteq A_\phi$, $|\mathcal{K}_{t,t}(A_L, A_R)| \leqslant \left(2^{-\lambda\ell^2/n}\binom{n}{\ell}\right)^{2t}$.*

Before we prove the above lemma, let us see how Lemma 10 and Lemma 11 imply Theorem 8.

Proof of Theorem 8. We assume without loss of generality that $\lambda \leqslant 1$. Pick $\delta = \lambda^2/12$ and $t = (4/\lambda)(n^2/\ell^2)$. From Lemma 11 and (1), we can bound the number of copies of $K_{t,t}$ in $G_{\phi,\ell}$ by

$$|\mathcal{K}_{t,t}| \leqslant \sum_{A_L, A_R \subseteq A_\phi} \left(2^{-\lambda\ell^2/n}\binom{n}{\ell}\right)^{2t} = 2^{4n} \cdot \left(2^{-\lambda\ell^2/n}\binom{n}{\ell}\right)^{2t} \leqslant (2^{-\lambda\ell^2/n})^t \cdot \binom{n}{\ell}^{2t}.$$

where the last inequality comes from our choice of t; note that t is chosen so that the 2^{4n} factor is consumed by $2^{-\lambda\ell^2/n}$ from Lemma 11. Next, consider any $\binom{n}{\ell}$-subgraph of $G_{\phi,\ell}$. By the above bound, it contains at most $(2^{-\lambda\ell^2/n})^t \cdot \binom{n}{\ell}^{2t}$ copies of $K_{t,t}$. As a result, from Lemma 10, its density is

$$O\left((2^{-\lambda\ell^2/n})^{1/t} + t\binom{n}{\ell}^{-1/t}\right) \leqslant O\left((2^{-\lambda\ell^2/n})^{1/t} + t2^{-\ell/t}\right) = O\left(2^{-3\delta\ell^4/n^3} + 2^{-(\ell/t - \log t)}\right).$$

Lastly, note that $\ell/t - \log t = (\lambda/4)(\ell^3/n^2) - \log(4/\lambda) - \log(n^2/\ell^2) \geqslant 3\delta\ell^4/n^3 - \log(4/\lambda) - 2\log n$. Since $\ell \geqslant n^{3/4}\log n$, we have $\ell^4/n^3 \geqslant \log^4 n$, which means that, when n is sufficiently large, $\ell/t - \log t \geqslant 2\delta\ell^4/n^3$. As a result, when n is sufficiently large, the density of the subgraph is at most $2^{-\delta\ell^4/n^3}$ as desired. □

We now move on to the proof of Lemma 11.

Proof of Lemma 11. First, notice that if (x, b) appears in A_L and $(x, \neg b)$ appears in A_R for some variable x and bit b, then $\mathcal{K}_{t,t}(A_L, A_R) = \emptyset$; this is because, for any L with $\mathcal{A}(L) = A_L$ and R with $\mathcal{A}(R) = A_R$, there exist $u \in L$ and $v \in R$ that contain (x, b) and $(x, \neg b)$ respectively, meaning that there is no edge between u and v and, thus, $(L, R) \notin \mathcal{K}_{t,t}(A_L, A_R)$. Hence, from now on, we can assume that, if (x, b) appears in one of A_L, A_R, then the other does not contain $(x, \neg b)$. Observe that this implies that, for each variable x, its assignments can appear in A_L and A_R at most two times[9] in total. This in turn implies that $|A_L| + |A_R| \leqslant 2n$.

Let us now argue that $|\mathcal{K}_{t,t}(A_L, A_R)| \leqslant \binom{n}{\ell}^{2t}$; while this is not the bound we are looking for yet, it will serve as a basis for our argument later. For every $(L, R) \in \mathcal{K}_{t,t}(A_L, A_R)$, since $\mathcal{A}(L) = A_L$ and $\mathcal{A}(R) = A_R$, we have $L \subseteq \binom{A_L}{\ell}$ and $R \subseteq \binom{A_R}{\ell}$, meaning that $\mathcal{K}_{t,t}(A_L, A_R) \subseteq \binom{\binom{A_L}{\ell}}{t} \times \binom{\binom{A_R}{\ell}}{t}$. Hence,

$$|\mathcal{K}_{t,t}(A_L, A_R)| \leqslant \binom{|A_L|}{\ell}^t \binom{|A_R|}{\ell}^t. \tag{2}$$

Moreover, $\binom{|A_L|}{\ell}\binom{|A_R|}{\ell}$ can be further bounded as

$$\binom{|A_L|}{\ell}\binom{|A_R|}{\ell} = \frac{1}{(\ell!)^2}\prod_{i=0}^{\ell-1}(|A_L| - i)(|A_R| - i) \leqslant \frac{1}{(\ell!)^2}\prod_{i=0}^{\ell-1}\left(\frac{|A_L| + |A_R|}{2} - i\right)^2 \leqslant \binom{n}{\ell}^2 \tag{3}$$

[9] This is where we use the fact that the variables are boolean. For non-boolean CSPs, each variable x can appear more than two times in one of A_L or A_R alone, which can indeed be problematic (see Appendix A).

where the inequalities come from the AM-GM Inequality and from $|A_L| + |A_R| \leqslant 2n$ respectively. Combining (2) and (3) indeed yields $|\mathcal{K}_{t,t}(A_L, A_R)| \leqslant \binom{n}{\ell}^{2t}$.

Inequality (2) is very crude; we include all elements of $\binom{A_L}{\ell}$ and $\binom{A_R}{\ell}$ as candidates for vertices in L and R respectively. However, as we will see soon, only tiny fraction of elements of $\binom{A_L}{\ell}, \binom{A_R}{\ell}$ can actually appear in L, R when $(L, R) \in \mathcal{K}_{t,t}(A_L, A_R)$. To argue this, let us categorize the variables into three groups:

- x is said to be *terrible* iff its assignments appear at most once in total in A_L and A_R.
- x is said to be *good* iff, for some $b \in \{0, 1\}$, $(x, b) \in A_L, A_R$ and $(x, \neg b) \notin A_L, A_R$.
- x is said to be *bad* iff either $(x, 0), (x, 1) \in A_L$ or $(x, 0), (x, 1) \in A_R$.

The next and last step of the proof is where birthday-type paradoxes come in. Before we continue, let us briefly demonstrate the ideas behind this step by considering the following extreme cases:

- If all variables are terrible, then $|A_L| + |A_R| \leqslant n$ and (3) can be immediately tightened.
- If all variables are bad, consider a random element u of $\binom{A_L}{\ell}$. Since u is a set of random ℓ distinct elements of A_L, there will, in expectation, be $\Omega(\ell^2/n)$ variables x's with $(x, 0), (x, 1) \in u$. However, the presence of such x's means that u is not a valid vertex. Moreover, it is not hard to turn this into the following probabilistic statement: with probability at most $2^{-\Omega(\ell^2/n)}$, u contains at most one of $(x, 0), (x, 1)$ for every variable x. In other words, only $2^{-\Omega(\ell^2/n)}$ fraction of elements of $\binom{A_L}{\ell}$ are valid vertices, which yields the desired bound on $|\mathcal{K}_{t,t}(A_L, A_R)|$.
- If all variables are good, then $A_L = A_R$ is simply an assignment to all the variables. Since $\mathrm{val}(\phi) \leqslant 1 - \varepsilon$, at least εm clauses are unsatisfied by this assignment. As we will argue below, every element of $\binom{A_L}{\ell}$ that contains two variables from some unsatisfied clause cannot be in L for any $(L, R) \in \mathcal{K}_{t,t}(A_L, A_R)$. This means that there are $\Theta_\varepsilon(m) \geqslant \Omega_\varepsilon(n)$ prohibited pairs of variables that cannot appear together. Again, similar to the previous case, it is not hard to argue that only $2^{-\Omega_{\varepsilon,d}(\ell^2/n)}$ fraction of elements of $\binom{A_L}{\ell}$ can be candidates for vertices of L.

To turn this intuition into an actual bound on the size of $\mathcal{K}_{t,t}(A_L, A_R)$, we need the following inequality. Its proof is straightforward and is deferred to Subsection 3.2.

Proposition 12 *Let U be any set and $P \subseteq \binom{U}{2}$ be any set of pairs of elements of U such that each element of U appears in at most q pairs. For any positive integer $2 \leqslant r \leqslant |U|$, the probability that a random element of $\binom{U}{r}$ does not contain both elements of any pair in P is at most* $\exp\left(-\frac{|P|r^2}{4q|U|^2}\right)$.

We are now ready to formalize the above intuition and finish the proof of Lemma 11. For the sake of convenience, denote the sets of good, bad and terrible variables by X_g, X_b and X_t respectively. Moreover, let $\beta := \varepsilon/(100d)$ and pick $\lambda = \min\{-\log(1 - \beta/2), \beta/64, \varepsilon/(384d)\}$. To refine the bound on the size of $\mathcal{K}_{t,t}(A_L, A_R)$, consider the following three cases:

1. $|X_t| \geqslant \beta n$. Since each $x \in X_t$ contributes at most one to $|A_L| + |A_R|$, we have $|A_L| + |A_R| \leqslant (1 - \beta/2)(2n)$. Hence, we can improve (3) to $\binom{|A_L|}{\ell}\binom{|A_R|}{\ell} \leqslant \binom{(1-\beta/2)n}{\ell}^2$. Thus, we have

$$|\mathcal{K}_{t,t}(A_L, A_R)| \overset{(2)}{\leqslant} \binom{|A_L|}{\ell}^t \binom{|A_R|}{\ell}^t \leqslant \binom{(1-\beta/2)n}{\ell}^{2t} \leqslant \left((1-\beta/2)^\ell \binom{n}{\ell}\right)^{2t} \leqslant \left(2^{-\lambda\ell^2/n}\binom{n}{\ell}\right)^{2t}$$

   where the last inequality comes from $\lambda \leqslant -\log(1 - \beta/2)$ and $\ell > \ell^2/n$.

2. $|X_b| \geqslant \beta n$. Since each $x \in X_b$ appears either in A_L or A_R, one of A_L and A_R must contain assignments to at least $(\beta/2)n$ variables in X_b. Assume without loss of generality that A_L satisfies this property. Let X_b^L be the set of all $x \in X_b$ whose assignments appear in A_L. Note that $|X_b^L| \geqslant (\beta/2)n$.
   Observe that an element $u \in \binom{A_L}{\ell}$ is not a valid vertex if it contains both $(x, 0)$ and $(x, 1)$ for some $x \in X_b^L$. We invoke Proposition 12 with $U = A_L$, $P = \{\{(x, 0), (x, 1)\} \mid x \in X_b^L\}, q = 1$ and $r = \ell$, which implies that a random element of $\binom{A_L}{\ell}$ does not contain any prohibited pairs in P with probability at most $\exp\left(-\frac{|X_b^L|\ell^2}{4|A_L|^2}\right) \leqslant \exp\left(-\frac{(\beta/2)n\ell^2}{4(2n)^2}\right)$, which is at most $2^{-2\lambda\ell^2/n}$ because $\lambda \leqslant \beta/64$. In other

words, at most $2^{-2\lambda\ell^2/n}$ fraction of elements of $\binom{A_L}{\ell}$ are valid vertices. This gives us the bound

$$|\mathcal{K}_{t,t}(A_L, A_R)| \leqslant \left(2^{-2\lambda\ell^2/n} \cdot \binom{|A_L|}{\ell}\right)^t \cdot \binom{|A_R|}{\ell}^t \overset{(3)}{\leqslant} \left(2^{-\lambda\ell^2/n}\binom{n}{\ell}\right)^{2t}$$

3. $|X_t| < \beta n$ and $|X_b| < \beta n$. In this case, $|X_g| > 1 - 2\beta n$. Let S denote the set of clauses whose variables all lie in X_g. Since each variable appears in at most d clauses, $|S| > m - (2\beta n)d \geqslant (1 - \varepsilon/2)m$ where the second inequality comes from our choice of β and from $m \geqslant n/3$.
   Consider the partial assignment $f : X_g \to \{0, 1\}$ induced by A_L and A_R, i.e., $f(x) = b$ iff $(x, b) \in A_L, A_R$. Since $\text{val}(\phi) \leqslant 1 - \varepsilon$, the number of clauses in S satisfied by f is at most $(1 - \varepsilon)m$. Hence, at least $\varepsilon m/2$ clauses in S are unsatisfied by f. Denote the set of such unsatisfied clauses by S_{UNSAT} .
   Fix a clause $C \in S_{\text{UNSAT}}$ and let x, y be two different variables in C. We claim that x, y cannot appear together in any vertex of L for any $(L, R) \in \mathcal{K}_{t,t}(A_L, A_R)$. Suppose for the sake of contradiction that $(x, f(x))$ and $(y, f(y))$ both appear in $u \in L$ for some $(L, R) \in \mathcal{K}_{t,t}(A_L, A_R)$. Let $z \in X_g$ be another variable[10] in C. Since $(z, f(z)) \in A_R$, some vertex $v \in R$ contains $(z, f(z))$. Thus, there is no edge between u and v in $G_{\phi,\ell}$, which contradicts with $(L, R) \in \mathcal{K}_{t,t}$.
   In other words, such pairs $(x, f(x)), (y, f(y))$ are prohibited pairs that cannot appear together in any vertex of L for any $(L, R) \in \mathcal{K}_{t,t}(A_L, A_R)$. Since $|S_{\text{UNSAT}}| \geqslant \varepsilon m/2 \geqslant \varepsilon n/6$, the number of such prohibited pairs is also at least $\varepsilon n/6$. Moreover, since each variable appears in at most d clauses, each $(x, f(x))$ appears in at most $2d$ prohibited pairs.
   We can now appeal to Proposition 12 with $U = A_L$, $q = 2d$, $r = \ell$ and P be the prohibited pairs described above. This implies that with probability at most $\exp\left(-\frac{|P|\ell^2}{8d|A_L|^2}\right) \leqslant \exp\left(-\frac{\varepsilon\ell^2}{192dn}\right)$, a random element of $\binom{A_L}{\ell}$ contains no prohibited pair from P. In other words, at most $\exp\left(-\frac{\varepsilon\ell^2}{192dn}\right)$ fraction of elements of $\binom{A_L}{\ell}$ can be candidates for each element of L for $(L, R) \in \mathcal{K}_{t,t}(A_L, A_R)$. This gives the following sharpened bound:

   $$|\mathcal{K}_{t,t}(A_L, A_R)| \leqslant \left(\exp\left(-\frac{\varepsilon\ell^2}{192dn}\right) \cdot \binom{|A_L|}{\ell}\right)^t \cdot \binom{|A_R|}{\ell}^t \overset{(3)}{\leqslant} \left(2^{-\varepsilon\ell^2/(384dn)} \cdot \binom{n}{\ell}\right)^{2t}.$$

   Since we picked $\lambda \leqslant \varepsilon/(384d)$, $|\mathcal{K}_{t,t}(A_L, A_R)|$ is once again bounded above by $\left(2^{-\lambda\ell^2/n}\binom{n}{\ell}\right)^{2t}$.

In all three cases, we have $|\mathcal{K}_{t,t}(A_L, A_R)| \leqslant \left(2^{-\lambda\ell^2/n}\binom{n}{\ell}\right)^{2t}$, completing the proof of Lemma 11. □

3.1 Proof of the Robust Kővári-Sós-Turán Theorem

In this subsection, we prove Lemma 10. While the connection between this robust version and the original theorem may not be clear in the proof below, one way to see the connection is as follows. Assume that $1 \gg \alpha \gg 1/N^{2t}$ and that we have a graph G that contains at most αN^{2t} copies of $K_{t,t}$. As a thought experiment, let S be a set of vertices of G where each vertex is included independently with probability $p := 1/(10\alpha^{\frac{1}{2t}}N)$. With high probability, $|S| = \Theta(pN) = \Theta(\alpha^{-\frac{1}{2t}})$. Let G_S be the subgraph of G induced on S. Since each copy of $K_{t,t}$ in G remains in G_S with probability p^{2t}, the expected number of copies of $K_{t,t}$ in G_S is at most $p^{2t}(\alpha N^{2t}) < 1/10$. By Markov's inequality, with probability at least 0.9, G_S is $K_{t,t}$-free. When G_S is $K_{t,t}$-free and $|S| = \Theta(\alpha^{-\frac{1}{2t}})$, the original KST Theorem implies that G_S is $O(|S|^{-1/t}) = O((\alpha^{-\frac{1}{2t}})^{(-1/t)}) = O(\alpha^{\frac{1}{2t^2}})$-dense. However, vertex sampling intuitively should not decrease the density of the graph. Hence, we expect the graph G to also be $O(\alpha^{\frac{1}{2t^2}})$-dense, which is roughly the same bound we get in our robust KST Theorem.

The proof given below is not based on the above vertex sampling argument but is instead based on counting arguments similar to those that appear in the proof of the original KST Theorem.

Proof of Lemma 10. Let $\mathcal{K}_{t,t} := \{(L, R) \in \binom{V}{t} \times \binom{V}{t} \mid \forall u \in L,\ \forall v \in R, (u, v) \in E\}$. Suppose for the sake of contradiction that $|\mathcal{K}_{t,t}| \leqslant \alpha N^{2t}$ but $|E| > 2\alpha^{1/t^2}N^2 + tN^{2-1/t}$.

[10] If C contains two variables, let $z = x$. Note that we can assume w.l.o.g. that C contains at least two variables.

For $L \in \binom{V}{t}$, let $\Gamma(L) := \{v \in V \mid \forall u \in L, (u, v) \in E\}$ denote the set of vertices that are neighbors to all vertices in L. Observe that each $v \in V$ appears in $\Gamma(L)$ for $\binom{\deg(v)}{t}$ different L's. Hence,

$$\begin{aligned}
\sum_{L \in \binom{V}{t}} |\Gamma(L)| = \sum_{v \in V} \binom{\deg(v)}{t} &\geqslant \frac{1}{t!} \sum_{v \in V} (\max(\deg(v) - t, 0))^t \\
\text{(Power Mean Inequality)} &\geqslant \frac{N}{t!} \left(\frac{1}{N} \sum_{v \in V} \max(\deg(v) - t, 0) \right)^t \\
&\geqslant \frac{N}{t!} \left(\frac{1}{N} (2|E| - Nt) \right)^t \\
\text{(Since } |E| > 2\alpha^{1/t^2} N^2 + tN^{2-1/t}\text{)} &> \frac{N}{t!} \left(4\alpha^{1/t^2} N + tN^{1-1/t} \right)^t \\
&\geqslant (4^t \alpha^{1/t} N^{t+1})/t! + N^t \\
\text{(Since } 4^t \geqslant t^2 \text{ for every } t \in \mathbb{N}\text{)} &\geqslant (t^2 \alpha^{1/t} N^{t+1})/t! + N^t.
\end{aligned}$$

Next, observe that, for every $L, R \in \binom{V}{t}$, $(L, R) \in \mathcal{K}_{t,t}$ if and only if $R \subseteq \Gamma(L)$. Thus, for each L, there are exactly $\binom{|\Gamma(L)|}{t}$ different R's such that $(L, R) \in \mathcal{K}_{t,t}$. Hence, we can bound $|\mathcal{K}_{t,t}|$ as follows.

$$\begin{aligned}
|\mathcal{K}_{t,t}| = \sum_{L \in \binom{V}{t}} \binom{|\Gamma(L)|}{t} &\geqslant \frac{1}{t!} \sum_{L \in \binom{V}{t}} (\max(|\Gamma(L)| - t, 0))^t \\
\text{(Power Mean Inequality)} &\geqslant \frac{\binom{N}{t}}{t!} \left(\frac{1}{\binom{N}{t}} \sum_{L \in \binom{V}{t}} \max(|\Gamma(L)| - t, 0) \right)^t \\
&\geqslant \frac{1}{t! \binom{N}{t}^{t-1}} \left(\left(\sum_{L \in \binom{V}{t}} |\Gamma(L)| \right) - \binom{N}{t} t \right)^t \\
&\geqslant \frac{(t!)^{t-2}}{N^{t(t-1)}} \left(\left(\sum_{L \in \binom{V}{t}} |\Gamma(L)| \right) - N^t \right)^t .
\end{aligned}$$

Recall that $\sum_{L \in \binom{V}{t}} |\Gamma(L)| > (t^2 \alpha^{1/t} N^{t+1})/t! + N^t$. Plugging this into the above inequality gives

$$|\mathcal{K}_{t,t}| > \frac{(t!)^{t-2}}{N^{t(t-1)}} \left(t^2 \alpha^{1/t} N^{t+1} / t! \right)^t = \frac{1}{(t!)^2 N^{t(t-1)}} (t^2 \alpha^{1/t} N^{t+1})^t \geqslant \alpha N^{2t},$$

which is a contradiction. □

3.2 Proof of Proposition 12

Proof of Proposition 12. We first construct $P' \subseteq P$ such that each element of U appears in at most one pair in P' as follows. Start out by marking every pair in P as active and, as long as there are active pairs left, include one in P' and mark every pair that shares an element of U with this pair as inactive. Since each element of U appears in at most q pairs in P, we mark at most $2q$ pairs as inactive per each inclusion. This implies that $|P'| \geqslant |P|/(2q)$.

Suppose that $P' = \{\{a_1, b_1\}, \dots, \{a_{|P'|}, b_{|P'|}\}\}$ where $a_1, b_1, \dots, a_{|P'|}, b_{|P'|}$ are distinct elements of U. Let u be a random element of $\binom{U}{r}$. For each $i = 1, \dots, |P'|$, we have

$$\Pr[\{a_i, b_i\} \not\subseteq u] = 1 - \frac{\binom{|U|-2}{r-2}}{\binom{|U|}{r}}$$

$$\begin{aligned} &= 1 - \frac{r(r-1)}{|U|(|U|-1)} \\ \text{(Since } r-1 \geqslant r/2 \text{ for all } r \geqslant 2) &\leqslant 1 - \frac{r^2}{2|U|^2} \\ \text{(Since } 1-z \leqslant \exp(-z) \text{ for all } z \in \mathbb{R}) &\leqslant \exp\left(-\frac{r^2}{2|U|^2}\right). \end{aligned}$$

If u does not contain both elements of any pairs in P, it does not contain both elements of any pairs in P'. The probability of the latter can be written as $\Pr\left[\bigwedge_{i=1}^{|P'|}\{a_i, b_i\} \not\subseteq u\right] = \prod_{i=1}^{|P'|}\Pr\left[\{a_i, b_i\} \not\subseteq u \,\middle|\, \bigwedge_{j=1}^{i-1}\{a_j, b_j\} \not\subseteq u\right]$. In addition, since $a_1, b_1, \dots, a_{|P'|}, b_{|P'|}$ are distinct, it is not hard to see that $\Pr\left[\{a_i, b_i\} \not\subseteq u \,\middle|\, \bigwedge_{j=1}^{i-1}\{a_j, b_j\} \not\subseteq u\right] \leqslant \Pr[\{a_i, b_i\} \not\subseteq u]$. Hence, we have

$$\Pr\left[\bigwedge_{i=1}^{|P'|}\{a_i, b_i\} \not\subseteq u\right] \leqslant \prod_{i=1}^{|P'|}\Pr[\{a_i, b_i\} \not\subseteq u] \leqslant \left(\exp\left(-\frac{r^2}{2|U|^2}\right)\right)^{|P'|} \leqslant \exp\left(-\frac{|P|r^2}{4q|U|^2}\right),$$

completing the proof of Proposition 12. □

3.3 Proofs of Inapproximability Results of Densest k-Subgraph

In this subsection, we prove Theorem 1 and 2. The proof of Theorem 1 is simply by combining Dinur's PCP Theorem and Theorem 8 with $\ell = m/\operatorname{polylog} m$, as stated below.

Proof of Theorem 1. For any 3SAT formula φ with m clauses, use Dinur's PCP Theorem (Theorem 7) to produce a 3SAT formula ϕ with $m' = O(m \operatorname{polylog} m)$ clauses such that each variable appears in at most d clauses. Let ζ be a constant such that $m' = O(m \log^\zeta m)$ and let $\ell = m/\log^2 m$. Let us consider the graph $G_{\phi,\ell}$ with $k = \binom{n}{\ell}$ where n is the number of variables of ϕ. Let N be the number of vertices of $G_{\phi,\ell}$. Observe that $N = 2^\ell\binom{n}{\ell} \leqslant n^{2\ell} \leqslant (m')^{O(\ell)} = 2^{O(\ell \log m')} = 2^{o(m)}$.

If φ is satisfiable, ϕ is also satisfiable and it is obvious that $G_{\phi,\ell}$ contains an induced k-clique. Otherwise, If φ is unsatisfiable, $\operatorname{val}(\phi) \leqslant 1 - \varepsilon$. From Theorem 8, any k-subgraph of $G_{\phi,\ell}$ has density at most $2^{-\Omega(\ell^4/n^3)} \leqslant 2^{-\Omega(m/\log^{3\zeta+8} m)} = N^{-\Omega(1/\log\log^{3\zeta+8} N)}$, which is at most $N^{-1/\log\log^{3\zeta+9} N}$ when m is sufficiently large. Hence, if there is a polynomial-time algorithm that can distinguish between the two cases in Theorem 1 when $c = 3\zeta + 9$, then there also exists an algorithm that solves 3SAT in time $2^{o(m)}$, contradicting with ETH. □

The proof of Theorem 2 is even simpler since, under Gap-ETH, we have the gap version of 3SAT to begin with. Hence, we can directly apply Theorem 8 without going through Dinur's PCP:

Proof of Theorem 2. Let ϕ be any 3SAT formula with m clauses such that each variable appears in at most d clauses[11]. Let $\ell = m\sqrt[5]{f(m)}$ and consider the graph $G_{\phi,\ell}$ with $k = \binom{n}{\ell}$ where n is the number of variables of ϕ. The number of vertices N of $G_{\phi,\ell}$ is $2^\ell\binom{n}{\ell} \leqslant 2^\ell(en/\ell)^\ell \leqslant 2^\ell(3em/\ell)^\ell \leqslant 2^{O(\ell\log(m/\ell))} = 2^{O(m\sqrt[5]{f(m)}\log(1/f(m)))} = 2^{o(m)}$ where the last inequality follows from $f \in o(1)$.

If ϕ is satisfiable, then it is obvious that $G_{\phi,\ell}$ contains an induced k-clique. Otherwise, if $\operatorname{val}(\phi) \leqslant 1 - \varepsilon$, from Theorem 8, any k-subgraph of $G_{\phi,\ell}$ has density at most $2^{-\Omega(\ell^4/n^3)} \leqslant 2^{-\Omega(mf(m)^{4/5})} \leqslant N^{-\Omega(f(m)^{4/5})}$, which is at most[12] $N^{-f(N)}$ when m is sufficiently large. Hence, if there is a polynomial-time algorithm that can distinguish between the two cases in Theorem 2, then there also exists an algorithm that solves the gap version of 3SAT in time $2^{o(m)}$, contradicting with Gap-ETH. □

[11] We can assume w.l.o.g. that each variable appears in at most $d = O(1)$ clauses. As pointed out in [MR16], we can pick $d \gg 1/c$; since the number of variables appearing in d clauses is at most $3m/d$, we can enumerate through all assignments to such variables. This creates $2^{3m/d}$ 3SAT formulae that satisfy the bounded degree condition. From our choice of d, we have $2^{3m/d} \ll 2^{cm}$, meaning that it still takes at least $2^{(c-3/d)m}$ time to solve one of the formulae.

[12] We assume w.l.o.g. that f is decreasing because otherwise we can take $f'(m) = \sup_{m' \geqslant m} f(m')$ instead.

4 Conclusion and Open Questions

In this work, we provide a subexponential time reduction from the gap version of 3SAT to DkS and prove that it establishes an almost-polynomial ratio hardness of approximation of the latter under ETH and Gap-ETH. Even with our results, however, complexity of approximating DkS still remains wide open. Namely, it is still not known whether it is NP-hard to approximate DkS to within some constant factor, and, no polynomial ratio hardness of approximation is yet known.

Although our results appear to almost resolve the second question, it still seems out of reach with our current knowledge of hardness of approximation. In particular, to achieve a polynomial ratio hardness for DkS, it is plausible that one has to prove a long-standing conjecture called *the sliding scale conjecture (SSC)* [BGLR93]. In short (and slightly inaccurate), SSC essentially says that Label Cover, a problem used as starting points of almost all NP-hardness of approximation results, is NP-hard to approximate to some polynomial ratio. We note here that polynomial ratio hardness for Label Cover is not even known under stronger assumptions such as ETH or Gap-ETH; we refer the readers to [Din16] for more detailed discussions on the topic.

Apart from the approximability of DkS, our results also prompt the following natural question: since previous techniques, such as Feige's Random 3SAT Hypothesis [Fei02], Khot's Quasi-Random PCP [Kho06], Unique Games with Small Set Expansion Conjecture [RS10] and the Planted Clique Hypothesis [Jer92, Kuč95], that were successful in showing inapproximability of DkS also gave rise to hardnesses of approximation of many problems that are not known to be APX-hard including Sparsest Cut, Min Bisection, Balanced Separator, Minimum Linear Arrangement and 2-Catalog Segmentation [AMS07, Sak10, RST12], is it possible to modify our construction to prove inapproximability for these problems as well? An evidence suggesting that this may be possible is the case of ε-approximate Nash Equilibrium with ε-optimal welfare, which was first proved to be hard under the Planted Clique Hypothesis by Hazan and Krauthgamer [HK11] before Braverman, Kun-Ko and Weinstein proved that the problem was also hard under ETH [BKW15].

Acknowledgement

I am truly grateful to Aviad Rubinstein, Prasad Raghavendra and Luca Trevisan for invaluable discussions throughout various stages of this work. Without their comments, suggestions and support, this work would not have been possible. I also thank Daniel Reichman and Igor Shinkar for stimulating dicussions on a related problem which inspire part of the proof presented here.

References

[AAM$^+$11] Noga Alon, Sanjeev Arora, Rajsekar Manokaran, Dana Moshkovitz, and Omri Weinstein. Inapproximabilty of densest k-subgraph from average case hardness. Unpublished Manuscript, 2011.

[AC09] Reid Andersen and Kumar Chellapilla. Finding dense subgraphs with size bounds. In *Algorithms and Models for the Web-Graph, 6th International Workshop, WAW 2009, Barcelona, Spain, February 12-13, 2009. Proceedings*, pages 25–37, 2009.

[AHI02] Yuichi Asahiro, Refael Hassin, and Kazuo Iwama. Complexity of finding dense subgraphs. *Discrete Applied Mathematics*, 121(1–3):15 – 26, 2002.

[AIM14] Scott Aaronson, Russell Impagliazzo, and Dana Moshkovitz. AM with multiple merlins. In *IEEE 29th Conference on Computational Complexity, CCC 2014, Vancouver, BC, Canada, June 11-13, 2014*, pages 44–55, 2014.

[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, May 1998.

[AMS07] Christoph Ambuhl, Monaldo Mastrolilli, and Ola Svensson. Inapproximability results for sparsest cut, optimal linear arrangement, and precedence constrained scheduling. In *Foundations of Computer Science, 2007. FOCS '07. 48th Annual IEEE Symposium on*, pages 329–337, Oct 2007.

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

[Bar15] Siddharth Barman. Approximating Nash equilibria and dense bipartite subgraphs via an approximate version of Caratheodory's Theorem. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 361–369, New York, NY, USA, 2015. ACM.

[BCC$^+$10] Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. Detecting high log-densities: an $O(n^{1/4})$ approximation for densest k-subgraph. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 201–210, 2010.

[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, Philadelphia, PA, USA, 2012. Society for Industrial and Applied Mathematics.

[BGLR93] Mihir Bellare, Shafi Goldwasser, Carsten Lund, and Alexander Russell. Efficient probabilistically checkable proofs and applications to approximations. In *Proceedings of the Twenty-fifth Annual ACM Symposium on Theory of Computing*, STOC '93, pages 294–304, New York, NY, USA, 1993. ACM.

[BKRW17] Mark Braverman, Young Kun-Ko, Aviad Rubinstein, and Omri Weinstein. ETH hardness for densest-k-subgraph with perfect completeness. In *Proceedings of the Twenty-eight Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, 2017. To appear.

[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 *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 970–982, 2015.

[BPR16] Yakov Babichenko, Christos H. Papadimitriou, and Aviad Rubinstein. Can almost everybody be almost happy? In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016*, pages 1–9, 2016.

[BSS08] Eli Ben-Sasson and Madhu Sudan. Short PCPs with polylog query complexity. *SIAM J. Comput.*, 38(2):551–607, May 2008.

[CDK12] Eden Chlamtac, Michael Dinitz, and Robert Krauthgamer. Everywhere-sparse spanners via dense subgraphs. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*, pages 758–767, 2012.

[CDK$^+$16] Eden Chlamtác, Michael Dinitz, Christian Konrad, Guy Kortsarz, and George Rabanca. The densest k-subhypergraph problem. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016, September 7-9, 2016, Paris, France*, pages 6:1–6:19, 2016.

[CHK11] Moses Charikar, MohammadTaghi Hajiaghayi, and Howard Karloff. Improved approximation algorithms for label cover problems. volume 61, pages 190–206. Springer US, 2011.

[CL15] Chandra Chekuri and Shi Li. A note on the hardness of approximating the k-WAY HYPERGRAPH CUT problem. Unpublished Manuscript, 2015.

[CLLR15] Wei Chen, Fu Li, Tian Lin, and Aviad Rubinstein. Combining traditional marketing and viral marketing with amphibious influence maximization. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15, pages 779–796, New York, NY, USA, 2015. ACM.

[CMMV17] Eden Chlamtác, Pasin Manurangsi, Dana Moshkovitz, and Aravindan Vijayaraghavan. Approximation algorithms for label cover and the log-density threshold. In *Proceedings of the*

Twenty-eight Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '17, 2017. To appear.

[CMVZ15] Julia Chuzhoy, Yury Makarychev, Aravindan Vijayaraghavan, and Yuan Zhou. Approximation algorithms and hardness of the k-route cut problem. *ACM Trans. Algorithms*, 12(1):2:1–2:40, December 2015.

[CZ15] Stephen R. Chestnut and Rico Zenklusen. Hardness and approximation for network flow interdiction. *CoRR*, abs/1511.02486, 2015.

[DFS16] Argyrios Deligkas, John Fearnley, and Rahul Savani. Inapproximability results for approximate Nash equilibria. *CoRR*, abs/1608.03574, 2016.

[Din07] Irit Dinur. The PCP theorem by gap amplification. *J. ACM*, 54(3), June 2007.

[Din16] Irit Dinur. Mildly exponential reduction from gap 3SAT to polynomial-gap label-cover. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:128, 2016.

[Fei02] Uriel Feige. Relations between average case complexity and approximation complexity. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 534–543, New York, NY, USA, 2002. ACM.

[FGL$^+$91] Uriel Feige, Shafi Goldwasser, László Lovász, Shmuel Safra, and Mario Szegedy. Approximating clique is almost NP-complete (preliminary version). In *Proceedings of the 32Nd Annual Symposium on Foundations of Computer Science*, SFCS '91, pages 2–12, Washington, DC, USA, 1991. IEEE Computer Society.

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

[FL01] Uriel Feige and Michael Langberg. Approximation algorithms for maximization problems arising in graph partitioning. *J. Algorithms*, 41(2):174–211, November 2001.

[FS97] Uriel Feige and Michael Seltser. On the densest k-subgraph problem. Technical report, Weizmann Institute of Science, Rehovot, Israel, 1997.

[GL09] Doron Goldstein and Michael Langberg. The dense k subgraph problem. *CoRR*, abs/0912.5327, 2009.

[HIM11] Koki Hamada, Kazuo Iwama, and Shuichi Miyazaki. The hospitals/residents problem with quota lower bounds. In *Algorithms - ESA 2011 - 19th Annual European Symposium, Saarbrücken, Germany, September 5-9, 2011. Proceedings*, pages 180–191, 2011.

[HJ06] Mohammad Taghi Hajiaghayi and Kamal Jain. The prize-collecting generalized steiner tree problem via a new approach of primal-dual schema. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, pages 631–640, Philadelphia, PA, USA, 2006. Society for Industrial and Applied Mathematics.

[HJL$^+$06] Mohammad Taghi Hajiaghayi, Kamal Jain, Lap Chi Lau, Ion I. Mandoiu, Alexander Russell, and Vijay V. Vazirani. Minimum multicolored subgraph problem in multiplex PCR primer set selection and population haplotyping. In *Computational Science - ICCS 2006, 6th International Conference, Reading, UK, May 28-31, 2006, Proceedings, Part II*, pages 758–766, 2006.

[HK11] Elad Hazan and Robert Krauthgamer. How hard is it to approximate the best Nash equilibrium? *SIAM Journal on Computing*, 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, March 2001.

[IPZ01] Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which problems have strongly exponential complexity? *J. Comput. Syst. Sci.*, 63(4):512–530, December 2001.

[Jer92] Mark Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 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.

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

[KMNT11] Guy Kortsarz, Vahab S. Mirrokni, Zeev Nutov, and Elena Tsanko. Approximating minimum-power degree and connectivity problems. *Algorithmica*, 60(4):735–742, 2011.

[KP93] Guy Kortsarz and David Peleg. On choosing a dense subgraph (extended abstract). In *34th Annual Symposium on Foundations of Computer Science, Palo Alto, California, USA, 3-5 November 1993*, pages 692–701, 1993.

[KS07] Stavros G. Kolliopoulos and George Steiner. Partially ordered knapsack and applications to scheduling. *Discrete Applied Mathematics*, 155(8):889 – 897, 2007.

[KST54] Tamás Kővári, Vera T. Sós, and Pál Turán. On a problem of K. Zarankiewicz. *Colloquium Mathematicae*, 3(1):50–57, 1954.

[Kuč95] Luděk Kučera. Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, 57(2-3):193–212, February 1995.

[Lee16] Euiwoong Lee. Partitioning a graph into small pieces with applications to path transversal. *CoRR*, abs/1607.05122, 2016.

[LNV14] Zhentao Li, Manikandan Narayanan, and Adrian Vetta. The complexity of the simultaneous cluster problem. *Journal of Graph Algorithms and Applications*, 18(1):1–34, 2014.

[Man15] Pasin Manurangsi. On approximating projection games. Master's thesis, Massachusetts Institute of Technology, January 2015.

[MM15] Pasin Manurangsi and Dana Moshkovitz. Approximating dense Max 2-CSPs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2015, August 24-26, 2015, Princeton, NJ, USA*, pages 396–415, 2015.

[MR08] Dana Moshkovitz and Ran Raz. Two-query PCP with subconstant error. *J. ACM*, 57(5):29:1–29:29, June 2008.

[MR16] Pasin Manurangsi and Prasad Raghavendra. A birthday repetition theorem and complexity of approximating dense CSPs. *CoRR*, abs/1607.02986, 2016.

[Pis07] David Pisinger. The quadratic knapsack problem—a survey. *Discrete Applied Mathematics*, 155(5):623 – 648, 2007.

[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*, STOC '10, pages 755–764, New York, NY, USA, 2010. ACM.

[RST12] Prasad Raghavendra, David Steurer, and Madhur Tulsiani. Reductions between expansion problems. In *Proceedings of the 27th Conference on Computational Complexity, CCC 2012, Porto, Portugal, June 26-29, 2012*, pages 64–73, 2012.

[Rub15] Aviad Rubinstein. ETH-hardness for signaling in symmetric zero-sum games. *CoRR*, abs/1510.04991, 2015.

[Rub16a] Aviad Rubinstein. Personal communication, 2016.

[Rub16b] Aviad Rubinstein. Settling the complexity of computing approximate two-player Nash equilibria. *CoRR*, abs/1606.04550, 2016.

[Sak10] Rishi Saket. Quasi-random PCP and hardness of 2-catalog segmentation. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2010, December 15-18, 2010, Chennai, India*, pages 447–458, 2010.

[SFL15] Piotr Skowron, Piotr Faliszewski, and Jerome Lang. Finding a collective set of items: From proportional multirepresentation to group recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2131–2137. AAAI Press, 2015.

[ST08] Akiko Suzuki and Takeshi Tokuyama. Dense subgraph problems with output-density conditions. *ACM Trans. Algorithms*, 4(4):43:1–43:18, August 2008.

[SW98] Anand Srivastav and Katja Wolf. Finding dense subgraphs with semidefinite programming. In *Proceedings of the International Workshop on Approximation Algorithms for Combinatorial Optimization*, APPROX '98, pages 181–191, London, UK, UK, 1998. Springer-Verlag.

[TV15] Sumedh Tirodkar and Sundar Vishwanathan. On the approximability of the minimum rainbow subgraph problem and other related problems. In *Algorithms and Computation - 26th International Symposium, ISAAC 2015, Nagoya, Japan, December 9-11, 2015, Proceedings*, pages 106–115, 2015.

A A Counterexample to Obtaining a Subconstant Soundness from Non-Boolean CSPs

Below we describe an example of a non-boolean CSP ϕ with low value for which the graph $G_{\phi,\ell}$ contains a large biclique. This example is due to Aviad Rubinstein [Rub16a]. We note that, for a non-boolean CSP, we define the graph $G_{\phi,\ell}$ similar to that of a 3SAT formula. The only difference is that now the vertices contains all $\{(x_{i_1}, \sigma_{i_1}), \dots, (x_{i_\ell}, \sigma_{i_\ell})\}$ for all distinct variables $x_{i_1}, \dots, x_{i_\ell}$ and all $\sigma_{i_1}, \dots, \sigma_{i_\ell} \in \Sigma$ where Σ is the alphabet of the CSP.

Consider any non-boolean CSP instance on variables $x_1, \dots, x_n$ such that there is no constraint between $\{x_1, \dots, x_{n/2}\}$ and $\{x_{n/2+1}, \dots, x_n\}$. Let L and R be the sets of all vertices corresponding to partial assignments to subsets of $\{x_1, \dots, x_{n/2}\}$ and $\{x_{n/2+1}, \dots, x_n\}$ respectively. Clearly, (L, R) forms a biclique. Moreover, $|L| = |R| = |\Sigma|^\ell \binom{n/2}{\ell} \geqslant 3^\ell \binom{n/2}{\ell}$, which is at least $\binom{n}{\ell}$ for all $\ell \leqslant n/4$. Hence, for such ℓ, $G_{\phi,\ell}$ contains a biclique of size $\binom{n}{\ell}$. Finally, we can add arbitrary constraints within $\{x_1, \dots, x_{n/2}\}$ and $\{x_{n/2+1}, \dots, x_n\}$ so that the value of the instance is bounded away from one. Thus, if we start from a non-boolean CSP, the largest gap we can hope to get is only two.

It should be noted that the instance above is rather extreme as it consists of two disconnected components. Hence, it is still possible that, if one starts from a non-boolean CSP with more specific properties (e.g. expanding constraint graph), then one can arrive at a gap of more than two.