A Linear Lower Bound on the Query Complexity of Property Testing Algorithms for 3-Coloring in Bounded-degree Graphs

Andrej Bogdanov[*] Kenji Obata[†] Luca Trevisan[‡]

May 4, 2002

Abstract

We consider the problem of testing 3-colorability in the bounded-degree model. A 3-colorability tester is an algorithm A that is given oracle access to the adjacency list representation of a graph G of maximum degree d with n vertices; A is required to, say, accept with probability at least 2/3 if G is 3-colorable, and to accept with probability at most 1/3 if G is ϵ-far from 3-colorable (meaning that at least an ϵ fraction of edges must be removed from G to make it 3-colorable); there is no requirement on A in the remaining cases. If A accepts 3-colorable graphs with probability one, then it is said to have one-sided error.

For sufficiently small ϵ, the testing problem is NP-complete, so it is unlikely that polynomial-time, or even sub-exponential time testers exist. In this paper we are interested in *unconditional* lower bounds on *query complexity.* The strongest known lower bound is due to Goldreich and Ron, who show that, for small enough ϵ, every tester must have query complexity $\Omega(\sqrt{n})$.

In this paper we show unconditionally that, for small enough ϵ, every tester for 3-colorability must have query complexity $\Omega(n)$. This is the first linear lower bound for testing a natural graph property in the bounded-degree model.

For one-sided error testers, we also show an $\Omega(n)$ lower bound for testers that distinguish 3-colorable graphs from graphs that are $(1/3 - \alpha)$-far from 3-colorable, for arbitrarily small α. In contrast, a polynomial time algorithm by Frieze and Jerrum distinguishes 3-colorable graphs from graphs that are 1/5-far from 3-colorable.

As a by-product of our techniques, we obtain tight unconditional lower bounds on the approximation ratios achievable by sub-linear time algorithms for Max E3SAT and Max E3LIN-2.

[*] adib@cs.berkeley.edu. Computer Science Division, University of California, Berkeley.

[†] kenjioba@cs.berkeley.edu. Computer Science Division, University of California, Berkeley. Work supported by an NSF graduate fellowship.

[‡] luca@cs.berkeley.edu. Computer Science Division, University of California, Berkeley. Work supported by NSF grant CCR 9984703 and a Sloan Research Fellowship.

1 Introduction

A property testing algorithm A for a graph property $\mathcal{P}$ is an algorithm that, given an approximation parameter ϵ and oracle access to the representation of a graph G, accepts with probability 2/3 if G has property $\mathcal{P}$ and rejects with probability 2/3 if G is ϵ-far from every graph having property $\mathcal{P}$. There is no requirement on A if G satisfies neither condition. Graphs G and H are ϵ-close if a representation of H can be obtained by modifying an ϵ-fraction of the representation of G.

The complexity of graph property testing problems is highly dependent on the representation. In the *adjacency matrix* representation, introduced in the original paper on graph property testing [?], two graphs are ϵ-close if they differ in at most about $\epsilon n^2/2$ edges. This model is interesting for studying properties of dense graphs. To study sparse graph properties, Goldreich and Ron [?] considered the model where a bounded-degree graph is represented by its *adjacency list*. In this model, the vertex degrees are bounded by a constant d independent on the number of vertices n. Two graphs are ϵ-close if they differ by at most $\epsilon dn/2$ edges.

The difference in complexity between the two models can be striking. For example, for $\epsilon = 1/100$, bipartiteness can be tested in constant time in the adjacency matrix representation [?] but it requires $\Omega(\sqrt{n})$ queries in the adjacency list representation [?], even for $d = 3$.

Indeed, a wide variety of graph properties are known to be testable in time constant in the number of vertices (dependent only on ϵ) in the adjacency matrix representation,[1] while much fewer algorithms running even in sub-linear time (leave alone constant time) are known for the adjacency list representation. This is particularly unfortunate considering that bounded degree graphs are more likely to occur in settings where sub-linear time testing algorithm are useful. Even fewer lower bounds are known for the adjacency list model. Apart from the $\Omega(\sqrt{n})$ lower bound on the query complexity of bipartiteness (which extends trivially to 3-colorability and other problems), there is an $\Omega(n^{1/3})$ lower bound for testing acyclicity in directed graphs [?]. We are not aware of any other nontrivial query complexity bound in this model.[2]

In this paper, we prove a tight $\Omega(n)$ lower bound on the query complexity of testing 3-colorability in bounded-degree graphs.

The problem of 3-colorability is interesting not only as a natural extension of bipartiteness (whose query complexity was resolved in [?, ?], but also as a canonical problem from which we obtain lower bounds for other problems using appropriate reductions.

Related results

Goldreich, Goldwasser and Ron [?] present a tester for 3-colorability in the adjacency matrix representation that makes $\tilde{O}(1/\epsilon^4)$ queries and runs in $2^{\tilde{O}(1/\epsilon^2)}$ time. (Alon and Krivelevich

[1] For example, all graph properties recognized by finite-state automata [?] and all properties expressed by a certain fragment of first-order logic [?] can be tested in time dependent only on ϵ.

[2] Bender and Ron [?] also prove an $\Omega(\sqrt{n})$ lower bound for the problem of testing strong connectivity in directed graphs, assuming that the adjacency list representation only contains outgoing edges; on the other hand, a constant-time algorithm exists for the representation where both outgoing and incoming edges are contained in each adjacency list.

[?] have improved the number of queries to $\tilde{O}(1/\epsilon^2)$ and the running time to $2^{\tilde{O}(1/\epsilon)}$.)

In the bounded-degree model, a testing algorithm must tell apart 3-colorable bounded-degree graphs from bounded-degree graphs where every 3-coloring violates an ϵ-fraction of the edges. For sufficiently small ϵ, this problem is NP-hard for general graphs [?]. Using the reduction from Section 5 (that we introduce for a different purpose), this problem can be shown to be NP-hard when restricted to bounded-degree graphs. This provides strong evidence against the existence of polynomial-time algorithms (and consequently, sub-linear time algorithms) for this problem. Using our reduction, together with the Polishuck-Spielman version of the PCP theorem [?], it can be shown that the testing problem has query complexity $\Omega(n^{1-\epsilon^c})$ for some constant c, assuming 3SAT on n variables has circuit complexity $2^{n^{1-o(1)}}$. This is an extremely strong assumption (although a refutation of it would constitute a major breakthrough).

Goldreich and Ron [?] prove an unconditional $\Omega(\sqrt{n})$ lower bound on query complexity for sufficiently small ϵ. On the positive side, Frieze and Jerrum [?] give a polynomial time algorithm that distinguishes between 3-colorable graphs and graphs that are 1/5-far from 3-colorable.

Lower bound for one-sided error testers

Our goal is to prove that no property tester with one-sided error, given a degree-d graph with n vertices, can look at fewer than δn entries of the adjacency list representation of the graph, yet reject with constant probability graphs that are ϵ-far from 3-colorable. A simple observation is that a one-sided error tester must accept whenever its "view" of the graph is 3-colorable. In other words, it is sufficient to construct a graph G that is ϵ-far from 3-colorable, yet every one of its induced subgraphs on δn edges is 3-colorable.

In Section 3 we give a probabilistic construction of such graphs, based on a technique due to Erdős [?]. For every $\alpha > 0$, there are constants $d = O(1/\alpha^2)$ and $\delta > 0$ such that some d-regular graph on n vertices is $(1/3 - \alpha)$-far from 3-colorable, yet every subgraph induced by $\leq \delta n$ edges is 3-colorable, for arbitrarily small α. The consequence is the following result.

Theorem 1 *For every $\alpha > 0$ there are constants d and $\delta > 0$ such that if A is a one-sided error tester for degree-d graphs that distinguishes 3-colorable graphs from graphs that are $(1/3 - \alpha)$-far from being 3-colorable, then the query complexity of A is at least δn, where n is the number of vertices.*

Notice that no graph is more than 1/3-far from being 3-colorable, so our result applies to the full spectrum of gaps for which the testing problem is well defined.

Furthermore, for small enough α, the testing problem is solvable deterministically in polynomial time with the Frieze-Jerrum algorithm [?]. This gives a separation of the testing ability of polynomial time versus (one-sided error) sub-linear time algorithms for a natural problem.

We consider the problem of constructing graphs that are simultaneously far from being 3-colorable, and free of small non-3-colorable subgraphs as an independently interesting combinatorial question. In section 5 we give an *explicit* construction of d-regular graphs that are ϵ-far from 3-colorable, yet any subgraph induced by a δ-fraction of edges is 3-colorable, where d, $\epsilon > 0$, $\delta > 0$ are absolute constants. To this end, we first construct an instance of kCSP (a set of constraints over binary variables, with k variables per constraint) that is ϵ'-far from being satisfiable, yet every δ' fraction of constraints is satisfiable (with k, δ', ϵ' constants, and each variable occurring in exactly two constraints). We then apply a reduction from kCSP to 3SAT and from 3SAT to 3-coloring, and argue that the reduction preserves distance from satisfiability (respectively, colorability) and the satisfiability (respectively, 3-colorability) of small enough subsets of the instance. The reduction from kCSP to 3SAT is the standard approximation-preserving reduction between the two problems [**?**], while the reduction from 3SAT to 3-coloring is a new one (the new reduction is needed to produce a constant-degree graph).

Lower bound for two-sided error testers

To prove a lower bound for two-sided error testers, by Yao's principle, it is enough to produce two distributions $\mathcal{G}_{3col}$ and $\mathcal{G}_{far}$ over bounded-degree graphs, such that graphs in $\mathcal{G}_{3col}$ are always 3-colorable, graphs in $\mathcal{G}_{far}$ are typically far from being 3-colorable, and the two distributions are indistinguishable for testers of sub-linear query complexity.

Towards this goal, we first create two distributions of instances of E3LIN-2, $\mathcal{D}_{sat}$ and $\mathcal{D}_{far}$, such that instances in $\mathcal{D}_{sat}$ are always satisfiable and instances in $\mathcal{D}_{far}$ are typically far from satisfiable.[3] yet the two distributions look the same to sub-linear time algorithms with oracle access to their input. We then reduce E3LIN-2 to 3SAT and then 3SAT to 3-coloring and argue that the transformation preserves satisfiability/3-colorability, as well as farness from satisfiability/3-colorability. Moreover, an oracle for a reduced instance can be implemented in constant time given the original instance.

In order to define $\mathcal{D}_{sat}$ and $\mathcal{D}_{far}$, we first show that for every c there is a δ such that there is a 3LIN-2 instance I with n variables and cn equations such that any subset of δn equations are linearly independent. We do so using a probabilistic argument. Then we define $\mathcal{D}_{sat}$ to be the distribution of instances obtained by first picking an assignment to the variables, and then setting the right-hand side of I to be consistent with the assignment. In $\mathcal{D}_{far}$ we set the right-hand side of I uniformly at random. For algorithms that look at less than a δ fraction of equations, the two distributions are identical, however instances in $\mathcal{D}_{sat}$ are always satisfiable and instances in $\mathcal{D}_{far}$ are about $(1/2 - O(1/\sqrt{c}))$-far from satisfiable, except with negligibly small probability. In summary, we have a proof of the following theorem.

Theorem 2 *Constants δ, ϵ, d exist such that if A is a two-sided error tester for degree-d graphs that distinguishes 3-colorable graphs from graphs that are ϵ-far from being 3-colorable, then the query complexity of A is at least δn, where n is the number of vertices.*

Other applications

Given a graph optimization problem, one can derive a property testing problem by first turning the optimization problem into a decision problem. For example, in the property testing version of Max CUT, one is given a fraction ρ and a parameter ϵ and wants to

[3] E3LIN-2 is the problem of deciding the satisfiability of a system of linear equations modulo 2, with three variables per equation.

distinguish graphs whose optimal cut cuts at least a ρ fraction of edges from graphs that are ϵ-far from having the above property.

A more natural (and often equivalent) way of studying sublinear time algorithms for graph optimization problems is to consider algorithms that produce in output an approximation of the cost of an optimal solution. For example, Goldreich, Goldwasser and Ron [**?**] give an algorithm running in $2^{\text{poly}(1/\epsilon)}$ time that returns an estimate of the cost of the max cut of a given graph within an *additive* error ϵn^2, which is a good approximation for dense graphs. Similar results are known for other problems in dense graphs [**?**].

Chazelle, Rubinfeld and Trevisan [**?**] show how to approximate within a multiplicative error $1 + \epsilon$ the cost of the minimum spanning tree in a given bounded-degree graph; the algorithm runs in time $\tilde{O}(dw\epsilon^{-2})$ where d is the maximum degree and the edge weights are integers in the range $\{1, \ldots, w\}$.

What about problems that can be approximated to within some constant in polynomial time but that do not have a PTAS, such as Max SAT and Max CUT? Can one achieve reasonably good approximation factors in sublinear time? Can unconditional inapproximability results be proved?

In Section 7 we show *unconditional* inapproximability results for sublinear time approximation algorithms that match the inapproximability results proved by Håstad [**?**] for polynomial time algorithms assuming $P \neq NP$.

Specifically, we prove that no sub-linear time approximation algorithm can approximate Max E3SAT better than 7/8, Max E3LIN-2 better than 1/2, Vertex Cover better than 7/6, Max CUT better than 16/17, or Max 2SAT better than 21/22.

2 Preliminaries and Definitions

Let $\mathcal{X}$ be a collection of combinatorial objects with distance function $d : \mathcal{X} \to [0, 1]$, such that $\text{diam}_d(\mathcal{X}) = 1$. An instance $X \in \mathcal{X}$ is *ϵ-far* from property $\mathcal{P} \subseteq \mathcal{X}$ if for any $P \in \mathcal{P}$, $d(X, P) > \epsilon$. An *ϵ-tester* for property $\mathcal{P}$ is a randomized algorithm that, given oracle access to an object $X \in \mathcal{X}$:

- If $X \in \mathcal{P}$, accepts X with probability at least 2/3,
- If X is ϵ-far from $\mathcal{P}$, rejects X with probability at least 2/3.

A tester is *one-sided* if the accepting probability above is 1. We are interested in testers for the following problems: 3-colorability in bounded degree graphs, $(3, c)$SAT (3CNF satisfiability where each literal occurs in at most c clauses), and E$(3, c)$LIN-2 (satisfiability of E3LIN-2 systems where each variable occurs in at most c equations).

We represent n-vertex graphs with degree bound d by an adjacency list $f_G : [n] \times [d] \to [n] \cup \{\varnothing\}$, where $f_G[v, i] = w$ if vertex w the i-th neighbor of vertex v, or $\varnothing$ if v has fewer than i neighbors. A graph G is ϵ-far from 3-colorable if no graph that is obtained by deleting $\epsilon dn/2$ edges of G is 3-colorable.

Similarly, we represent $(3, c)$CNF formulas (resp. E$(3, c)$LIN-2 systems) φ as a membership list M_φ, which provides for each literal (resp. variable) v and index $0 \leq i < c$ the i-th clause (resp. equation) in which v appears, or $\varnothing$ if v appears in fewer than i clauses (resp.

equations). A formula (resp. system) φ is ϵ-far from satisfiable if no subformula (resp. subsystem) of φ obtained by removing $\leq \epsilon cn/3$ clauses (resp. equations) is satisfiable.

3 Probabilistic constructions

In this section we provide probabilistic constructions of combinatorial objects (graphs and 3-hypergraphs) that will be used to obtain problem instances for 3-colorability and E3LIN-2 that are difficult to test.

Graphs and Hypergraphs with no Small Dense Subgraph

It will be somewhat more convenient to work with multigraphs instead of graphs. We consider a distribution $\mathcal{G}$ on n-vertex multigraphs G (where n is even) obtained as follows: Let $C_1, \ldots, C_d$ be independent random perfect matchings on the vertices of G. The edge set of G is the multiset union of the C_i, so that the multiplicity of an edge equals the number of matchings C_i in which it appears. If $(u, v) \in C_i$, we say that v is the i-th neighbor of u in G.

We denote by $G|_S$ the restriction of multigraph G on vertex set $S \subseteq V(G)$. Let X_S be the number of edges in $G|_S$. Then $\mathrm{E}[X_S] = d\binom{|S|}{2}\frac{1}{n-1}$. Fix a partition $\{S_1, S_2, S_3\}$ of $V(G)$. We are interested in bounding the probability that this partition is 1/3-close to a valid coloring of G. Let $X = X_{S_1} + X_{S_2} + X_{S_3}$.

Lemma 3 *For every partition $\{S_1, S_2, S_3\}$ of $V(G)$ and every constant $\alpha > 0$,*

$$\Pr[X < (1/6 - \alpha)dn] \leq \exp(-(\alpha - o(1))^2 dn).$$

Proof Consider the random process $I_1, \ldots, I_{dn/2}$ on G, which reveals the edges of G one by one. For a fixed partition $\{S_1, S_2, S_3\}$, the random variable X determines a Doob martingale with respect to this process. A simple computation shows that for $1 < j \leq dn/2$,

$$|\mathrm{E}[X|I_1, \ldots, I_j] - \mathrm{E}[X|I_1, \ldots, I_{j-1}]| \leq 1.$$

By convexity, $\mathrm{E}[X] \geq \frac{dn}{6}\frac{n-3}{n-1}$ (this value is attained when $|S_1| = |S_2| = |S_3| = n/3$). Azuma's inequality yields

$$\Pr\left[X < \left(\frac{1}{6}\frac{n-3}{n-1} - \alpha'\right) dn\right] \leq \exp(-\alpha'^2 dn).$$

The conclusion follows, with $\alpha = \alpha' + \frac{1}{3(n-1)}$. ■

Denote by $\bar{G}$ the graph obtained by identifying every multiedge of G with an ordinary edge.

Lemma 4 *For any constant $\alpha > 0$ there exists a constant d such that with probability $1 - o(1)$ any 3-coloring of the vertices of $\bar{G}$ has at least $(1/6 - \alpha)dn$ violating edges.*

Proof First we show that the conclusion holds for G. The number of tri-partitions of $V(G)$ is 3^n. By combining a union bound with the bound from Lemma 3, it follows that any such partition has $(1/6 - \alpha)dn$ violating edges if $d > \ln 3/\alpha^2$.

For any pair of vertices (u, v), let $M_{u,v}$ indicate the event that (u, v) is an edge of G with multiplicity two or more. Then $\Pr[M_{u,v} = 1] = O(d/n^2)$. By Markov's inequality, the probability that there are $d \log n$ or more pairs (u, v) with $M_{u,v} = 1$ is $o(1)$. Since no edge of G has multiplicity more than d, it follows that $|E(G)| - |E(\bar{G})| \leq d^2 \log n = o(n)$. Therefore the conclusion of the lemma carries over to $\bar{G}$. ■

Lemma 5 *For every $K > 1$ there exists a $\delta > 0$ such that with probability $1 - o(1)$ all graphs $\bar{G}|_S$ with $|S| \leq \delta n$ have at most $K|S|$ edges.*

Proof Suppose some set S of cardinality s contains Ks edges $(u_1, v_1), \ldots, (u_{Ks}, v_{Ks})$. Denote by $X_{i,k}, Y_{i,k}$ the vertices matched to u_i and v_i, respectively, in the matching C_k. Then

$$\Pr[\exists k : X_{i,k} = v_i \wedge Y_{i,k} = u_i | X_{p,q}, Y_{p,q} : 1 \leq p \leq i-1, 1 \leq q \leq d] \leq d/(n-2s),$$

since for any fixed q, the variables $X_{p,q}$ and $Y_{p,q}$ determine the neighbors of at most $2s$ vertices in matching C_k. It follows that

$$\Pr[\forall i, 1 \leq i \leq d : \exists k : X_{i,k} = b_i \wedge Y_{i,k} = a_i] \leq \left(\frac{d}{n-2s}\right)^{Ks} < \left(\frac{d}{(1-2\delta)n}\right)^{Ks}.$$

For fixed s, the set S can be chosen in $\binom{n}{s}$ ways, while the set $\{(u_1, v_1), \ldots, (u_{Ks}, v_{Ks})\}$ can be chosen in $\binom{\binom{s}{2}}{Ks}$ ways. Therefore for some constant s_0,

$$\begin{aligned}\Pr[\exists S, s_0 \leq |S| < \delta n : |E(G|_S)| \geq K|S|] &\leq \sum_{s=s_0}^{\delta n} \binom{n}{s} \binom{\binom{s}{2}}{Ks} \left(\frac{d}{(1-2\delta)n}\right)^{Ks} \\ &\leq \sum_{s=s_0}^{\delta n} \left(\frac{ne}{s}\right)^s \left(\frac{s^2 e/2}{Ks}\right)^{Ks} \left(\frac{d}{(1-2\delta)n}\right)^{Ks} \\ &= \left[\frac{e^2 d}{2} \left(\frac{ed}{2K(1-2\delta)}\right)^K \left(\frac{s}{n}\right)^{K-1}\right]^s = o(1).\end{aligned}$$

It is easy to see that the contribution of sets S of size less than s_0 is also $o(1)$. ■

We define an analogous distribution $\mathcal{H}$ on 3-hypergraphs (hypergraphs with multiple hyperedges where each hyperedge has cardinality 3) with n vertices, where n is a multiple of 3. To obtain a graph $H \sim \mathcal{H}$, we choose d independent uniformly random partitions of the vertex set $V(H)$ into 3-hyperedges (i.e., 3-element subsets). With probability $1 - o(1)$, all hyperedges of H have multiplicity one. An argument similar to the proof of Lemma 5 shows the following property:

Lemma 6 *For every $K > 1/2$ there exists a $\delta > 0$ such that with probability $1 - o(1)$ all 3-hypergraphs $H|_S$ with $|S| \leq \delta n$ have at most $K|S|$ edges.*

Hard Instances

We show the existence of graphs that are almost 1/3-far from 3-colorable, yet for some $\delta > 0$ all their subgraphs of size δn are 3-colorable. Choose a multigraph G according to the distribution $\mathcal{G}$ of section 3, and let $\bar{G}$ denote the graph obtained from G by ignoring multiplicities. We show that the graph $\bar{G}$ has the desired property.

Theorem 7 *For every $\alpha > 0$ there exists a $\delta > 0$ such that with probability $1 - o(1)$, the graph $\bar{G}$ is $(1/3-\alpha)$-far from 3-colorable, yet all subgraphs $G|_S$ with $|S| < \delta n$ are 3-colorable.*

Proof By Lemma 4 (with parameter $\alpha/2$), every tri-partition of $V(\bar{G})$ has at least $(1/3-\alpha)dn/2$ violating edges, so $\bar{G}$ is 1/3-far from 3-colorable.

Suppose that there exists a set S of size $s < \delta n$ such that $\bar{G}|_S$ is not 3-colorable. We may assume that S is a minimal set with this property. Suppose that $\bar{G}|_S$ contains a vertex v of degree two or less (with respect to $\bar{G}|_S$). By the minimality of S, there is a 3-coloring of the graph $\bar{G}|_{S-\{v\}}$. However, this coloring extends to a 3-coloring of $\bar{G}|_S$, by picking a color for v that does not match any of its neighbors. It follows that any vertex in $\bar{G}|_S$ must have degree at least 3. Therefore, $\bar{G}|_S$ must contain at least $3s/2$ edges. By Lemma 5 with $K = 3/2$, this is not possible. ■

Using the 3-hypergraph construction, we prove the existence of certain matrices that will be used as the left hand side of E3LIN-2 instances.

Theorem 8 *For every $c > 0$ there exists a $\delta > 0$ such that for every n there exists a matrix $A \in \{0,1\}^{n \times cn}$ with n columns and cn rows, such that each row has exactly three non-zero entries, each column has exactly $3c$ non-zero entries, and every collection of δn rows is linearly independent.*

Proof By Lemma 6, there exists a $3c$-regular 3-hypergraph H on n vertices such that any $H|_S$ with $|S| \leq 3\delta n$ has strictly fewer than $2|S|/3$ edges. Let A be the incidence matrix of H: The columns of A correspond to vertices of H, the rows of A correspond to hyperedges of H, and $A_{ve} = 1$ if and only if $v \in e$. Suppose that there is a set R of δn rows of A (or hyperedges of H) that are linearly dependent. We may assume that R is a minimal set with this property. Let $S \subseteq V(H)$ denote the set of vertices incident to hyperedges in R, so that $|S| \leq 3\delta n$. By minimality of R, every element of S must appear in at least two rows of R. Therefore, R contains at least $2|S|/3$ hyperedges. Contradiction. ■

4 Reductions

In this section, we define a notion of reducibility between constraint satisfaction problems which preserves up to modification of constants the property that a family of problems has a sub-linear testing algorithm, and exhibit such a reduction from $(3,k)$-SAT to 3-colorability in bounded degree graphs.

For our purposes, the following notion of reduction will be appropriate:

Definition 9 (Gap-preserving local reduction) *Let A, B be decision problems. We say that a mapping $\varphi()$ is a gap-preserving local reduction from A to B if there exist universal constants $c_1, c_2 > 0$ such that the following properties hold:*

- *If x is a YES-instance of A, then $\varphi(x)$ is a YES-instance of B.*
- *If x is ϵ-far from being a YES-instance of A then $\varphi(x)$ is ϵ/c_1-far from being a YES-instance of B.*
- *The answer to an oracle query to $\varphi(x)$ can be computed by making c_2 oracle queries to x.*

Since we will be dealing frequently with partially satisfiable constraint satisfaction problems, we introduce the following notation:

Definition 10 (($\delta, 1-\epsilon$)-satisfiability) *A constraint satisfaction problem on m clauses is* ($\delta, 1-\epsilon$)-satisfiable *if any subset of at most δm constraints is satisfiable, but no assignment satisfies more than $(1-\epsilon)m$ constraints.*

We note three easy lemmas, which will allow us to move between various CSP formulations:

Lemma 11 *Let H be an arbitrary fixed set of boolean predicates on a finite number of variables. There exists a gap-preserving local reduction from CSPs defined on H which carries an instance f with n variables and m clauses into a 3-CNF formula with $O(n+m)$ variables and $O(m)$ clauses.*

Proof It is a basic fact that an arbitrary boolean predicate on a finite number of variables can be expressed as a 3-CNF formula, possibly with introduction of a constant number of auxiliary variables. It is easy to check that applying this transformation to each clause of f gives a reduction which has the claimed properties. ■

Lemma 12 *Gap-preserving local reductions are closed under composition.*

Proof Clearly, if φ, φ' are gap-preserving local reductions with distortion constants c_1, c_2 and c_1', c_2' respectively, then $\varphi \circ \varphi'$ is a gap-preserving local reduction with distortion constants $c_1 c_1', c_2 c_2'$. ■

Lemma 13 *If $\varphi : A \to B$ is a gap-preserving local reduction with distortion constants c_1, c_2 and f is a $(\delta, 1-\epsilon)$-satisfiable CSP, then $\varphi(f)$ is a $(\frac{\delta}{c_2}, 1-\frac{\epsilon}{c_1})$-satisfiable CSP.*

Proof Let f_A be a $(\delta, 1-\epsilon)$-satisfiable instance of A, and $f_B = \varphi(f_A)$. That the problem f_B is $\frac{\epsilon}{c_1}$-far from satisfiable is immediate from the definition of a gap-preserving local reduction. Now, let m be the number of clauses in problem f_B and consider any subset $C_1', \ldots, C_{k'}'$ of $\frac{\delta}{c_2}m$ of these clauses. By the locality property, these clauses are a function of some set of clauses $C_1, \ldots, C_k$ of f_A with $k \leq c_2 \frac{\delta}{c_2} m = \delta m$. Since f_A is $(\delta, 1-\epsilon)$-satisfiable, the clauses $C_1, \ldots, C_k$ are satisfiable, and we can extend these clauses to a new, satisfiable instance f_A' of A by setting every clause other than $C_1, \ldots, C_k$ to a satisfiable clause on fresh variables. φ must send f_A' into a satisfiable instance, and this instance contains clauses $C_1', \ldots, C_{k'}'$. In particular, the clauses $C_1', \ldots, C_{k'}'$ must be satisfiable. ■

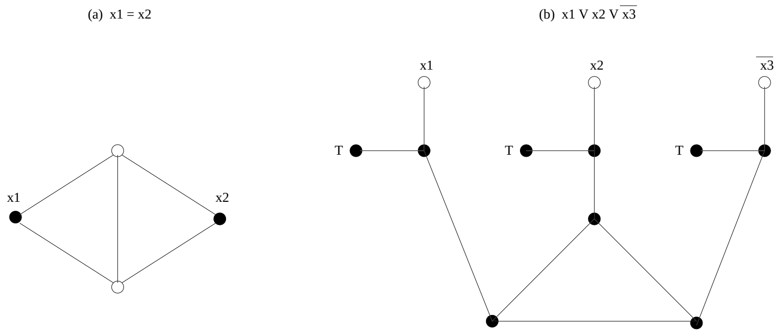

Figure 1: Gadgets for Theorem 14

We now exhibit a gap-preserving local reduction $\varphi()$ from $(3, k)$-SAT to 3-coloring in bounded degree graphs. We comment that a reduction with essentially the same properties was given by Petrank in [**?**]. However, Petrank's construction does not yield a bounded degree graph, which is essential in our context. Also, our construction is somewhat simpler to describe and analyze.

Construction: Let f be the $(3, k)$-CNF formula on n variables and m clauses to be mapped. First, we introduce a large set of nodes which are independent of the clauses of f which we label D_i, T_i, and F_i for $i = 1, \ldots, 2kn$. The nodes D_i will all assume the color corresponding to the "dummy" color (this color is used as in the standard 3-coloring reduction), T_i to the "true" color, and F_i to the "false" color. To assure that nodes in a given *color class* are the same color, we introduce equality gadgets (Figure XYZ.a) between nodes D_i and D_j for all $(i, j) \in E_{2kn}$ where $G_{2kn}(V_{2kn}, E_{2kn})$ is a $(2kn, d)$-expander as in Lemma 2 (similarly for the classes T and F). To assure that nodes in distinct color classes have distinct colors, for $i = 1, \ldots, 2kn$ we introduce triangles $\{(D_i, T_i), (D_i, F_i), (T_i, F_i)\}$.

For each variable x_i in f, we introduce $2k$ literal nodes $x_i^1, \ldots, x_i^k, \overline{x_i^1}, \ldots, \overline{x_i^k}$. Literal nodes for a particular variable and sign should be colored identically, so we introduce equality constraints between x_i^j and $x_i^{j'}$ for all $1 \leq i, j \leq k$ with $i \neq j$ (similarly for $\overline{x_i^j}$ and $\overline{x_i^{j'}}$). We fix some one-to-one correspondence between the literal nodes and the color class nodes for each color class (we can do so since we have $2kn$ nodes in each color class). Since literal nodes should be colored only with "true" or "false", every literal node is connected to its corresponding node D_i. Since only one of $x_i, \overline{x_i}$ can be true, we introduce edges $(x_i^j, \overline{x_i^j})$ for all i, j. Finally, for each clause in f, we introduce a clause gadget (Figure XYZ.b) on the literals appearing in the clause. We can do so in such a way that each literal node is used in at most one clause gadget since we have k literal nodes for each literal, and each variable appears in at most k clauses. Similarly, we can have each T node used in at most one clause gadget, since the gadgets consume at most $kn < 2kn$ T nodes. The clause gadget allows any coloring of the literal nodes with "true" or "false" other than the coloring which corresponds to an assignment where all literals are false (and the clause goes unsatisfied).

Theorem 14 *The mapping φ is a gap-preserving local reduction from $(3, k)$-SAT to 3-coloring in bounded degree graphs. In particular, if f is a $(\delta, 1 - \epsilon)$-satisfiable $(3, k)$-CNF* *formula, then the graph $\varphi(f)$ has degree bounded by some universal constant b and the 3-coloring CSP of $\varphi(f)$ is $\left(\frac{\delta}{bc}, 1-\frac{\epsilon}{8}\right)$-satisfiable.*

Proof It is clear by observation that the mapping φ always produces graphs bounded by some constant degree b, and that there exists a constant c such that φ converts a $(3,k)$-CNF formula on n variables to a graph on at most cn nodes. Furthermore, one can answer a query for an edge of $\varphi(f)$ making at most one query into f, namely, for the clause in which the queried edge is a part (if any). Write n' for the number of nodes in $\varphi(f)$, and $m' < bn' \leq bcn$ for the number of edges.

Suppose that the original $(3,k)$-CNF formula is $(\delta, 1-\epsilon)$-satisfiable. Clearly any subgraph of $\varphi(f)$ induced by δn edges is 3-colorable – such a subgraph contains nodes "involved" with at most δn clause gadgets, where a node is involved with a clause gadget if it is contained in the clause gadget, or is a color class node corresponding to a literal node contained in the clause gadget. By definition, there exists a boolean assignment satisfying these δn clauses of f. The coloring which sets all color classes to their intended colors and colors the literal nodes "true" or "false" as in this assignment satisfies these $\delta n > \frac{\delta}{bc} m'$ 3-coloring constraints.

Note that if we delete γt edges from the expander graph G_t with $\gamma \leq \frac{1}{2}$, then there must remain a connected component of size at least $(1-\gamma)t$, for disconnecting a set S of nodes with $|S| \leq \frac{1}{2}t$ requires at least $|\Gamma(S)|$ edge deletions which, by the expansion property, is at least $|S|$. Applying this to the equality gadgets between color class nodes, we see that deletion of $\gamma(2kn)$ edges leaves each color class with at least $(1-\gamma)(2kn)$ color class nodes in a connected component with equality constraints intact. Therefore, it leaves at least $(1-3\gamma)(2kn)$ triples $\{D_i, T_i, F_i\}_{i\in S}$ such that the D_i must be colored the same as D_j for $i, j \in S$ (similarly for T_i and F_i). The disconnected triples $\overline{S}$ are involved in at most $2 \cdot 3\gamma(2kn)$ clause gadgets. Furthermore, deleting $\gamma(2kn)$ edges modifies constraints about nodes involved with at most $2 \cdot \gamma(2kn)$ clauses of f. Summing up, deletion of $\gamma(2kn)$ edges leaves the 3-coloring construction for at least $m - (2 \cdot 3\gamma(2kn) + 2 \cdot \gamma(2kn)) = m - 16\gamma kn$ clauses of f intact. If f is $(\delta, 1-\epsilon)$-satisfiable, then no coloring of the remaining graph can be valid if

$$m - 16\gamma kn > (1-\epsilon)m$$

or, equivalently, $\gamma < \frac{\epsilon}{16k}$. Changing notation so that

$$\gamma' m' = \gamma(2kn)$$

(i.e. we have deleted a fraction γ' of the edges of $\varphi(f)$ in the above discussion) and noting that $m' > n$, we get that

$$\frac{\epsilon}{16k} > \gamma = \frac{\gamma' m'}{2kn} > \frac{\gamma'}{2k}$$

or $\gamma' < \frac{\epsilon}{8}$.

Combining the conclusions of the previous two paragraphs, we see that the graph 3-coloring problem $\varphi(f)$ is $\left(\frac{\delta}{bc}, 1-\frac{\epsilon}{8}\right)$-satisfiable. ■

5 Explicit Constructions

In this section, we give an explicit construction of an infinite family of $(\delta, 1-\epsilon)$-satisfiable CSPs on n variables and $m = O(n)$ clauses over a fixed boolean predicate. By applying the gap-preserving local reductions presented in Section 4, we achieve an explicit construction of an infinite family of $(3,k)$-CNF formulas on n variables and $O(n)$ clauses with analogous properties, and of bounded degree graphs G on n vertices and m edges such that every subgraph induced by δm edges is 3-colorable, but any 3-coloring of G has at least ϵm monochromatic edges. (In the proof of Theorem 7 we used the probabilistic method to prove only the *existence* of such graphs.)

For a fixed d, we will consider $2d$-ary constraints of the form

$$h : \{0,1\}^d \times \{0,1\}^d \to \{0,1\}$$

where $h(x_1, \ldots, x_d, y_1, \ldots, y_d)$ is satisfied exactly when

$$\sum_{i=1}^{d} x_i = \sum_{i=1}^{d} y_i + 1$$

where we identify the boolean $\{0,1\}$ inputs with the integers 0 and 1 in the obvious way.

Let $G(V, E)$ be an undirected multigraph. We write $\Gamma(v)$ for the neighbor set of vertex $v \in V$, $\Gamma(v, i)$ for the i-th neighbor of v (where we index $\Gamma(v)$ in an arbitary way), and $\Gamma(S)$ for the neighbor set of a vertex-subset $S \subseteq V$.

Definition 15 ((n,d)-Expander) *A multigraph G is an (n,d)-expander if it is d-regular and if, for every subset $S \subset V$ with $|S| \leq \frac{1}{2}|V|$, $|\Gamma(S)| \geq |S|$.*

Explicit constructions of (n, d)-expanders are known [**?**, **?**], and we assume that we are given an infinite family of (n, d)-expanders for some universal constant d.

Define the constraint satisfaction problem f_n on dn variables and n clauses over h as follows: Let $G(V, E)$ be an (n, d)-expander. Begin by converting G into a directed multigraph $G'(V, E')$ by replacing each undirected edge $(i, j) \in E$ with two directed edges $(i, j), (j, i) \in E'$. Each edge $(i, j) \in E'$ is identified with a boolean variable $x_{i,j}$ in f_n. One constraint h is introduced for each $v \in V$, with the predicate variables mapped to the edges incident to v:

$$f_n = \bigwedge_{v \in V} h(x_{v,\Gamma(v,1)}, \ldots, x_{v,\Gamma(v,d)}, x_{\Gamma(v,1),v}, \ldots, x_{\Gamma(v,d),v})$$

Theorem 16 *There exist constants $\delta, \epsilon > 0$ such that the CSP formulas f_n are $(\delta, 1-\epsilon)$-satisfiable.*

Proof We begin by finding ϵ such that no subset of more than $(1-\epsilon)n$ constraints can be satisfied. Suppose there is an assignment satisfying some subset S of constraints with $|S| > (1-\epsilon)n$. Then the following network flow problem is solvable: Contract the vertices corresponding to $\overline{S}$ into a single sink vertex t, create a source vertex s with unit capacity edges from s to every vertex in S, and interpret the remaining edges of G as unit capacity edges (see Figure 1.b). The assignment can then be interpreted as an (s,t)-flow of weight greater than $(1-\epsilon)n$ on this network. However, the cut $(t, G\backslash t)$ has weight at most $d\epsilon n$, so this is impossible if we choose $\epsilon < \frac{1}{d+1}$.

On the other hand, for $\delta \leq \frac{1}{2}$, any subset S of constraints with $|S| = \delta n$ can be satisfied. To see this, we define the following network flow problem: Contract the vertices of G corresponding to the $(1-\delta)n$ constraints in $\overline{S}$ to a sink vertex t, create a source vertex s with unit capacity edges from s to every node in S, and interpret the remaining edges of G as unit capacity edges (see Figure 1.b). We claim that there is a flow of weight at least δn in this system. By the max-flow/min-cut theorem, it is enough to show that there is no (s,t)-cut with weight less than δn (the cut $(s, G\backslash s)$ has weight δn). Let C be an arbitrary (s,t)-cut, and denote by C_s, C_t the vertices of S in the partitions containing s and t respectively. Each node in C_t incurs a cut cost of weight one due to the unit constraint edges we added from s. By the expansion property, $|\Gamma(C_s)| \geq |C_s|$, and each of the edges connecting C_s to $\Gamma(C_s)$ also incurs a cut cost of weight one. Summing up, $|C| \geq |C_s| + |C_t| = \delta n$, so there must exist an flow of weight δn in this system. Furthermore, the *integrality property* of flows implies that we can assume the flow solution is $(0,1)$-valued. Assigning this flow to the edge variables gives a satisfying assignment to the constraints in S. ■

Corollary 17 *Let φ_{3-CNF} be the gap-preserving local reduction of Lemma 11, and φ_{3-Col} that of Theorem 14. The (explictly constructed) set $\{\varphi_{3-Col}(\varphi_{3-CNF}(f_n))\}_n$ is an infinite family of bounded-degree graphs G_n on m_n edges such that, for universal constants $\delta, \epsilon > 0$, every subgraph induced by δm_n edges is 3-colorable, but every 3-coloring of G_n has at least ϵm_n monochromatic edges.*

Proof We need only note that the 3-CNF formulas $\{\varphi_{3-CNF}(f_n)\}_n$ are in fact $(3,k)$-CNF formulas. This is because the variable $x_{i,j}$ corresponding to edge (i,j) appears only in the constraints around vertices i and j. In particular, if l is the number of clauses in a 3-CNF representation of the predicate h, then $x_{i,j}$ can appear in at most $2l$ clauses. The claim then follows from Lemmas 12 and 13.

■

6 Lower Bounds

We now prove Theorems 1 and 2.

Lower Bound for One-Sided Error Algorithms

To prove Theorem 1, we observe that any testing algorithm with one-sided error must accept whenever the subgraph it has queried is 3-colorable. In particular, when presented with the graph from Theorem 7, any algorithm with query complexity at most δn will accept with probability one. However, this graph is $(1/3-\alpha)$-far from being 3-colorable, so the algorithm cannot be a $(1/3-\alpha)$-tester for 3-colorability.

Lower Bounds for Two-Sided Error Algorithms

Our distinguishing instances for two-sided error algorithm are based on the matrix A from Theorem 8. We consider the following two distributions on instances of E3LIN-2 with n variables, cn equations, and each variable appearing in exactly $3c$ equations:

1. Distribution $\mathcal{D}_{far}$ consists of instances $Ax = b$, where $b \in \{0,1\}^{cn}$ is chosen uniformly at random.

2. Distribution $\mathcal{D}_{sat}$ consists of instances $Ax = Az$, where $z \in \{0,1\}^n$ is chosen uniformly at random.

By construction, every instance in $\mathcal{D}_{sat}$ is satisfiable. On the other hand, instances in $\mathcal{D}_{far}$ are far from satisfiable:

Lemma 18 *For every $\alpha > 0$, there is a c such that, with probability $1 - o(1)$, an instance sampled from $\mathcal{D}_{far}$ is $(1/2 - \alpha)$-far from satisfiable.*

Proof For a fixed assignment x, the vector $Ax - b$ is uniformly distributed in $\{0,1\}^{cn}$. By a Chernoff bound, with probability $1 - \exp(-\alpha^2 cn)$, $Ax - b$ has Hamming weight at least $(1/2 - \alpha)cn$. A union bound over all 2^n possible assignments for x yields the desired result, as long as $c > \ln 2/\alpha^2$. ■

Lemma 19 *For every $\alpha > 0$ there are constants c and $\delta > 0$ such that every algorithm that distinguishes satisfiable instances of E3LIN-2 with n variables and at most c occurrences from instances that are $(1/2 - \alpha)$-far from satisfiable must have query complexity at least δn.*

Proof Consider an instance $Ax = b$ of cn E3LIN-2 equations. Obtain a subinstance $A'x' = b'$ by choosing *any* subset of δn equations. By Theorem 8, the rows of A' are linearly independent. Therefore, for a uniformly random $z' \in \{0,1\}^n$, $A'z'$ is uniformly distributed in $\{0,1\}^{\delta n}$. It follows that the instances $A'x' = b'$ and $A'x' = A'z'$ are generated with the same probability, or $\Pr_{\mathcal{D}_{far}}[A'x' = b'] = \Pr_{\mathcal{D}_{sat}}[A'x' = b']$.

Let D be any algorithm of query complexity less than δn. If D can decide whether a given instance $Ax = b$ is satisfiable with any constant probability, then D has an advantage at distinguishing instances picked from $\mathcal{D}_{sat}$ (that are always satisfiable) from instances picked from $\mathcal{D}_{far}$ (that are $(1/2 - \alpha)$-far from satisfiable with high probability). However, the queries of D only reveal a subinstance $A'x' = b'$ of at most δn equations, and the two distributions are statistically indistinguishable on such a subinstance. ■

The canonical reduction from E3LIN-2 to E3SAT is a gap-preserving local reduction with $c_1 = c_2 = 4$. This observation immediately yields the following lower bound for E3SAT:

Lemma 20 *For every $\alpha > 0$ there are constants c and $\delta > 0$ such that every algorithm that distinguishes satisfiable instances of E3SAT with n variables and at most c occurrences from instances that are $(1/8 - \alpha)$-far from satisfiable must have query complexity at least δn.*

The proof of Theorem 2 now follows from the hardness result of Lemma 20 and from the reduction from 3SAT to 3-coloring described in Section 4.

7 Approximation Algorithms

The following theorem follows directly from Lemmas 19 and 20.

Theorem 21 *For every $\epsilon > 0$, every $(1/2 + \epsilon)$-approximate algorithm for Max E3LIN-2, every $(7/8+\epsilon)$-approximate algorithm for Max E3SAT has query complexity $\Omega(n+m)$, where n is the number of variables and m is the number of equations/clauses. The theorem applies to the special case where every variable occurs in $O(1)$ equations/clauses and $m = O(n)$.*

Indeed, Lemma 19 is the unconditional version for sub-linear time algorithms of the hardness of approximation proved in [?] for Max E3LIN-2. Håstad [?] then uses locally computable reductions to show that the hardness of Max E3LIN-2 implies hardness of approximation results for other problems. Since the reductions used in [?] preserve the existence of sub-linear time algorithms (for proper instance representation), we also have unconditional inapproximability results for other problems, with respect to sublinear time algorithms.

The standard FGLSS reduction from Max E3LIN-2 to Vertex Cover is such that if every variable occurs in $O(1)$ equations in the E3LIN-2 instance, then the graph produced by the reduction has constant degree. Therefore, the following result also follows from Lemma 19 (see [?] for a calculation of the inapproximability factor).

Theorem 22 *For every $\epsilon > 0$, there are constants d, δ such that every $(7/6+\epsilon)$-approximate algorithm for Minimum Vertex Cover in graphs of degree $\leq \delta$ has query complexity at least δn.*

Similarly, we have a linear query complexity lower bound for every $(21/22+\epsilon)$-approximate algorithm for Max 2SAT, even for the restricted case where every variable occurs in $O(1)$ clauses.

Regarding Max CUT, the reduction used in [?] does not create a bounded-degree graph, even if in the original E3LIN-2 instance every variable occurred in a bounded number of equations. However the randomization reduction in [?] can be used to show that every $(16/17+\epsilon)$-approximate algorithm for Max CUT in bounded-degree graphs has linear query complexity.

8 Conclusions

We proved a linear query complexity lower bound for the problem of testing 3-colorability in bounded-degree graphs, and also linear lower bounds for other property testing and approximation problems.

Our results are the first linear lower bounds for property testing in the bounded-degree model for natural graph properties.

It is still open whether it is possible to distinguish 3-colorable graphs from graphs that are, say, 1/4-far from being 3-colorable with $o(n)$ queries in the bounded-degree model; we have shown it is impossible for one-sided error algorithms, but the question is still open for two-sided error algorithms.

We observed that several unconditional inapproximability results follow as corollaries of our main construction. The results for Vertex Cover, Max CUT and Max 2SAT are not tight, and it would be interesting to strengthen our bounds. We mention that one can modify the bipartiteness lower bound argument in [?] to prove that distinguishing bipartite graphs from graph that are $(1/2 - \alpha)$-far from being bipartite requires $\Omega(\sqrt{n})$ queries, which in turn implies that Max CUT cannot be approximated within $(1/2 + \epsilon)$ with $o(\sqrt{n})$ queries, and, by reductions, that Max E2SAT (and, for a stronger reason, Max SAT) cannot be approximated within $(3/4+\epsilon)$ and Vertex Cover cannot be approximated within $(3/2-\epsilon)$ with $o(\sqrt{n})$ queries. It remains an open question to prove such stronger lower bounds for algorithms that make $o(n)$ queries.

References