Testing k-colorability

Noga Alon * Michael Krivelevich †

Abstract

Let G be a graph on n vertices and suppose that at least ϵn^2 edges have to be deleted from it to make it k-colorable. It is shown that in this case most induced subgraphs of G on $ck \ln k/\epsilon^2$ vertices are not k-colorable, where $c > 0$ is an absolute constant. If G is as above for $k = 2$, then most induced subgraphs on $\frac{(\ln(1/\epsilon))^b}{\epsilon}$ are non-bipartite, for some absolute positive constant b, and this is tight up to the poly-logarithmic factor. Both results are motivated by the study of testing algorithms for k-colorability, first considered by Goldreich, Goldwasser and Ron in [3], and improve the results in that paper.

1 Introduction

Suppose that for a fixed integer k and a small $\epsilon > 0$, a graph $G = (V, E)$ on n vertices is such that at least ϵn^2 edges should be deleted to make G k-colorable. Clearly G contains many non-k-colorable subgraphs. Some of them are probably quite small in order. What is then the smallest non-k-colorable subgraph of G? How many small non-k-colorable subgraphs are there?

In order to address the above questions quantitatively, we introduce a suitable notation. First, we call a graph G on n vertices *ϵ-robustly non-k-colorable* or alternatively *ϵ-far from being k-colorable*, if after deleting any subset of less than ϵn^2 edges of G the remaining graph is still not k-colorable. Of course, it follows that G itself is not k-colorable. Define

$$f_k(G) = \min\{|V_0| : V_0 \subseteq V(G), G[V_0] \text{ is non-}k\text{-colorable}\} \ ,$$

where $G[V_0]$ denotes the subgraph of G induced by V_0. If $\chi(G) \le k$, we set $f_k(G) = \infty$. For an integer n and $0 < \epsilon < 1/(2k)$ let

$$f_k(n, \epsilon) = \max\{f_k(G) : G \text{ is an } \epsilon\text{-robustly non-}k\text{-colorable graph on } n \text{ vertices}\} \ .$$

*Department of Mathematics, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Email: noga@math.tau.ac.il. Research supported in part by a USA-Israeli BSF grant, by the Israel Science Foundation and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

†Department of Mathematics, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: krivelev@math.tau.ac.il. Research was partially performed while this author was with DIMACS Center, Rutgers University, Piscataway, NJ 08854, USA. Research supported by a DIMACS Postdoctoral Fellowship, by a USA-Israeli BSF Grant and by a Bergmann Memorial Grant.

AMS Subject classification: 05C15, 05C35, 05C85.

(Note that the assumption $\epsilon < 1/(2k)$ can be made without loss of generality as every graph on n vertices is at most $n^2/(2k)$ edges far from being k-colorable). Similarly, let

$$\begin{aligned} g_k(G) \ &= \min\{t : \text{if } R \subseteq V(G) \text{ is chosen uniformly at random from all subsets of } V \text{ of size } t, \\ &\qquad \text{then } Pr[\chi(G[R]) > k] \geq 1/2\} \ . \end{aligned}$$

Again, $g_k(G) = \infty$ if $\chi(G) \leq k$. Let

$$g_k(n, \epsilon) = \max\{g_k(G) : G \text{ is an } \epsilon\text{-robustly non-}k\text{-colorable graph on } n \text{ vertices}\} \ .$$

Obviously, $f_k(G) \leq g_k(G)$ for any graph G, thus implying $f_k(n, \epsilon) \leq g_k(n, \epsilon)$.

A few comments on the above definitions are in order. The function $f_k(n, \epsilon)$ represents a very natural extremal graph theory problem, seeking to link the size of a smallest non-k-colorable subgraph of a non-k-colorable graph with its distance from the set of k-colorable graphs. For example, for $k = 2$ one can say that if the odd girth (i.e. the minimal length of an odd cycle) of a graph G on n vertices is more than $f_2(n, \epsilon)$ for some $\epsilon > 0$, then G can be made bipartite by deleting less than ϵn^2 edges. The function $g_k(n, \epsilon)$ says that if G is ϵ-robustly non-k-colorable, then it contains not only one, but very many non-k-colorable subgraphs on $g_k(n, \epsilon)$ vertices. The somewhat artificially looking definition of $g_k(n, \epsilon)$ has actually a very natural algorithmic background in terms of graph property testing, as considered by Goldreich, Goldwasser and Ron in [3]. Applied to the particular problem of testing k-colorability, their approach reads as follows. Suppose our aim is to design an algorithm, which for a given (large enough) integer n and a (small enough) parameter $\epsilon > 0$, distinguishes with high probability between an input graph on n vertices, which is k-colorable, and that in which at least ϵn^2 edges should be deleted to create a k-colorable graph. The algorithm can query whether or not a specific pair of vertices of the input graph is connected by an edge. In general, it is NP-complete to check k-colorability for any $k \geq 3$. However, given the assumption that the input is either k-colorable or very far from being such, one may hope to devise very efficient randomized algorithms. We refer the reader to [3] for a general discussion of graph property testing.

Returning to the definition of the function $g_k(n, \epsilon)$, one can propose the following very simple algorithm for testing k-colorability. Given an input graph $G = (V, E)$, choose uniformly at random $g_k(n, \epsilon)$ vertices of G and denote the chosen set by R. Now, check whether the induced subgraph $G[R]$ is k-colorable. If it is, output "G is k-colorable", otherwise output "G is not-k-colorable". Note that if G is k-colorable, then every subgraph of it is k-colorable as well. Thus, in this case we *always* output a correct answer. On the other hand, if G is ϵ-far from being k-colorable, it follows from the definition of $g_k(n, \epsilon)$ that a sample of size $g_k(n, \epsilon)$ induces a non-k-colorable graph with probability at least $1/2$. Therefore, in this case we output a correct answer with probability at least $1/2$. Having in mind the above discussion, sometimes we will refer to bounding the function $g_k(n, \epsilon)$ as to *testing k-colorability.*

The problem of estimating $f_k(n,\epsilon)$ and $g_k(n,\epsilon)$ will be treated in this paper as an *asymptotic* one. This means that whenever needed, we will assume that the number of vertices n is large enough, and that the robustness parameter ϵ is small enough, but fixed as n is growing.

It is important to observe that the values of the functions defined above are of interest only for graphs with a *quadratic* number of edges. Indeed, if G has n vertices and is ϵ-far from being k-colorable, it contains at least ϵn^2 edges. This observation, together with the asymptotic nature of the problem, prompts the use of graph theoretic methods designed for dense graphs, most notably the well known Regularity Lemma of Szemerédi [6].

Let us now survey the previous research on these problems. Somewhat surprisingly it turned out that the above defined function $g_k(n,\epsilon)$ can be bounded from above by a function of ϵ only. This has been proven by Bollobás, Erdős, Simonovits and Szemerédi [2] for the case $k = 2$ and by Rödl and Duke [5] for every $k \geq 3$. Both papers rely on the Regularity Lemma. As is the case with most applications of the Regularity Lemma, the resulting bounds are extremely fast growing functions of $1/\epsilon$ (towers of height polynomial in $1/\epsilon$), thus making the results hardly applicable from the practical point of view. Note that both papers [2] and [5] formulate their results in a somewhat different language and do not define the function $g_k(n,\epsilon)$ explicitly.

For $k = 2$, Komlós showed in [4] that $f_2(n,\epsilon) = O(1/\epsilon^{1/2})$. This result is easily seen to be tight by considering a blow-up of an odd cycle of length about $1/\epsilon^{1/2}$. (A graph G on n vertices is a *blow-up* of a graph H on m vertices with vertex set $V(H) = \{v_1,\ldots,v_m\}$ if the vertex set of G can be partitioned into m disjoint sets $V_1,\ldots,V_m$, each of size $|V_i| = n/m$, so that V_i and V_j are joined completely if $(v_i,v_j) \in E(H)$ and are not joined by any edge otherwise.)

Motivated by testing k-colorability, Goldreich, Goldwasser and Ron [3] came up with a completely different approach for bounding $g_k(n,\epsilon)$. Using direct combinatorial arguments (and thus avoiding the Regularity Lemma), they were able to prove that $g_2(n,\epsilon) = O(\log(1/\epsilon)/\epsilon^2)$ – a tremendous progress compared to the bound of [2]. Similarly, they proved that for every fixed $k \geq 3$ one has $g_k(n,\epsilon) = O(k^2 \log k/\epsilon^3)$. The difference between the cases $k = 2$ and $k \geq 3$ can be intuitively explained by the fact that for $k = 2$ the family of minimal non-2-colorable graphs coincides with the family of odd cycles and is thus very simple to describe. For every $k \geq 3$ the family of minimal non-k-colorable graphs (usually called $(k+1)$-color-critical graphs) is very complicated. Goldreich et al. did not discuss the function $f_k(n,\epsilon)$ and did not provide any separate bounds for it.

The purpose of the present paper is to provide improved bounds for both functions f_k and g_k. We prove the following results.

Theorem 1

1. *For all* $\epsilon \leq 1/9$, $g_2(n,\epsilon) \geq \frac{1}{6\epsilon}$.

2. For every fixed $k \geq 3$ and every small enough $\epsilon > 0$, for infinitely many n one has

$$g_k(n,\epsilon) \geq f_k(n,\epsilon) \geq \frac{1}{110}\left(\frac{1}{330\ln k}\right)^{\frac{2}{k-2}}\frac{1}{\epsilon}\ .$$

.

Theorem 2

$$g_2(n,\epsilon) \leq \frac{34\ln^4\left(\frac{1}{\epsilon}\right)\ln\ln\left(\frac{1}{\epsilon}\right)}{\epsilon}\ .$$

Theorem 3 *For every fixed $k \geq 3$,*

$$g_k(n,\epsilon) \leq \frac{36k\ln k}{\epsilon^2}\ .$$

These results improve upon the above mentioned bounds of Goldreich et al. Still, for every $k \geq 3$, the gap between the lower and the upper bounds, given by Theorems 1 and 3 respectively, remains quite substantial.

The rest of the paper is organized as follows. In Section 2 we discuss lower bounds for the functions f_k, g_k and prove Theorem 1. In Section 3 we prove Theorem 2. Section 4 is devoted to proving Theorem 3. The final Section 5 contains some concluding remarks and open problems.

During the course of the proof we make no serious attempts to optimize the constants involved. Also, we omit routinely floor and ceiling signs to simplify the presentation. Given a graph $G = (V, E)$ and a vertex $v \in V$, we denote by $N(v)$ the set of all neighbors of v in G. The degree of v in G is denoted by $d(v)$. For a vertex $v \in V$ and a subset $U \subset V$, we denote by $d(v, U)$ the number of neighbors of v in U. The number of edges of G spanned by U, i.e. having both endpoints in U, is denoted by $e(U)$. A vertex $v \in V(G)$ is *dominated* by a subset $S \subseteq V(G)$ if v has a neighbor inside S in G. Recall that, whenever needed, the number of vertices n is assumed to be large enough, while $\epsilon > 0$ is small enough.

2 Lower bounds

In this section we prove lower bounds for the functions f_k, g_k. For many graph testing problems, the lower bound of order $1/\epsilon$ is very natural and can be proven quite easily. The property of k-colorability is not an exception, and the bound $g_k(n,\epsilon) \geq c(k)/\epsilon$ can be obtained by considering a complete $(k+1)$-partite graph with one part of size $\Theta(\epsilon n)$ and the other k of equal size. This is how we prove Theorem 1, Part 1. For every $k \geq 3$ we prove a stronger statement. Namely, we show the existence of an ϵ-robustly non-k-colorable graph on n vertices, in which, for a fixed constant $c = c(k)$, not only does a typical subset of size c/ϵ induce a k-colorable graph, but *every* subgraph of this order is k-colorable. This is done by considering a random graph with a linear number of edges and then blowing it up to get an ϵ-robustly non-k-colorable graph which is locally k-colorable. This supplies a lower bound for the function $f_k(n,\epsilon)$. It is worth noting here that the case $k=2$ is different, as it follows from the result of Komlós [4] that $f_2(n,\epsilon)=\Theta(1/\epsilon^{1/2})$.

Proof of Theorem 1, Part 1. Given n, ϵ, let G be a complete tripartite graph with parts V_0, V_1, V_2 of sizes $|V_0|=3\epsilon n$, $|V_1|=|V_2|=\frac{1-3\epsilon}{2}n$. Notice that each edge of G participates in at most $(1-3\epsilon)n/2$ triangles. As the total number of triangles in G is $3\epsilon(1-3\epsilon)^2n^3/4$, at least $3\epsilon(1-3\epsilon)n^2/2 \geq \epsilon n^2$ edges should be deleted to destroy all the triangles of G. Therefore, G is ϵ-robustly non-2-colorable. In order to estimate $g_2(G)$, note that if $R \subset V(G)$ is such that $R \cap V_0 = \emptyset$, then the subgraph $G[R]$ is bipartite. Thus in order to have $\chi(G[R])=3$, the set R has to hit V_0. If R is chosen uniformly at random from all subset of $V(G)$ of size r, then

$$Pr[R \cap V_0 \neq \emptyset] \leq \frac{|V_0|\binom{n-1}{r-1}}{\binom{n}{r}} = \frac{|V_0|r}{n} = 3\epsilon r\ .$$

Thus, requiring $Pr[\chi(G[R])=3] \geq 1/2$ implies $3\epsilon r \geq 1/2$, which in turn gives $r \geq 1/(6\epsilon)$. As G is ϵ-robustly non-bipartite, the statement follows. $\square$

Proof of Theorem 1, Part 2. Let us define

$$\begin{aligned} c_1 &= c_1(k) = \left(\frac{1}{3}\right)^{\frac{2}{k-2}} \left(\frac{1}{40e\ln k}\right)^{\frac{k}{k-2}}\ , \\ c_2 &= c_2(k) = 2\ln k\ , \\ c_3 &= c_3(k) = 40k\ln k\ . \end{aligned}$$

The key ingredient of the proof is the following lemma.

Lemma 2.1 *For every fixed $k \geq 3$ and a sufficiently large integer m, there exists a graph $H = H_{k,m}$ on m vertices, having the following properties:*

1. *Every subset of c_1m vertices of H spans a k-colorable graph;*

2. *Every subset $U \subset V(H)$ of size $|U| > m/(2k)$ spans at least $c_2|U|$ edges;*

3. *At least $c_2m/2$ edges need to be deleted from H to create a k-colorable graph.*

Proof. Set $p=p(m)=c_3/m$ and consider the random graph $G(m,p)$. This is a random graph with vertex set $\{1,\ldots,m\}$ in which every pair $1 \leq i < j \leq m$ is an edge independently and with probability p. We will prove that almost surely $G(m,p)$ has the desired properties. In this proof the term "almost surely" (or a.s. for short) means that the probability that all desired properties hold tends to 1 as $m \to \infty$.

In order to prove that the first assertion of the lemma holds a.s. for the random graph $G(m,p)$, note that a non-k-colorable graph contains a subgraph in which all degrees are at least k. Thus, if the first assertion fails, the random graph contains a subset U of size $|U| \le c_1 m$, spanning at least $(k/2)|U|$ edges. The probability of this event can be bounded from above by the following expression:

$$\sum_{i=k+1}^{c_1 m} \binom{m}{i} \binom{\binom{i}{2}}{\frac{k}{2}i} p^{\frac{k}{2}i} < \sum_{i=k+1}^{c_1 m} \left(\frac{em}{i}\right)^i \left(\frac{ei}{k}\right)^{\frac{ki}{2}} p^{\frac{ki}{2}} = \sum_{i=k+1}^{c_1 m} \left[\frac{em}{i}\left(\frac{eip}{k}\right)^{\frac{k}{2}}\right]^i .$$

Denote the i-th summand of the last sum by a_i. Then, if $m^{1/2} \le i \le c_1 m$ we have:

$$\begin{aligned} a_i &\le \left[\frac{em}{c_1 m}\left(\frac{ec_1c_3 m}{km}\right)^{\frac{k}{2}}\right]^i = \left[\frac{e}{c_1}\left(\frac{ec_1c_3}{k}\right)^{\frac{k}{2}}\right]^i = \left[e\left(\frac{ec_3}{k}\right)^{\frac{k}{2}} c_1^{\frac{k}{2}-1}\right]^i \\ &= \left[e(40e\ln k)^{\frac{k}{2}}\frac{1}{3}\left(\frac{1}{40e\ln k}\right)^{\frac{k}{2}}\right]^i = \left(\frac{e}{3}\right)^i = o(m^{-1}) . \end{aligned}$$

If $4 \le i < m^{1/2}$, we get

$$a_i < \left[em^{1/2}\left(\frac{ec_3}{km^{1/2}}\right)^{\frac{k}{2}}\right]^4 = \left(\frac{e^{\frac{k}{2}+1}(40\ln k)^{\frac{k}{2}}}{m^{\frac{k}{4}-\frac{1}{2}}}\right)^4 = o(m^{-1/2}) .$$

Thus, $\sum_{i=k+1}^{c_1 m} a_i = o(1)$, showing that the first part of the lemma holds with high probability in $G(m,p)$.

For the second part of the lemma, note that for a fixed subset $U \subseteq V(G(m,p))$ of size $|U| = i$, the number of edges spanned by U in $G(m,p)$ is a binomial random variable with parameters $\binom{i}{2}$ and p. Using the well known Chernoff bounds on the tails of binomial distribution (cf., e.g., [1], Appendix A), we get $Pr[|E(G[U])| < \binom{i}{2}p - a] < \exp\{-a^2/(2\binom{i}{2}p)\}$. Therefore, the probability of existence of a subset U, violating the assertion of the second part of the lemma, can be bounded from above by

$$\begin{aligned} &\sum_{i>m/2k} \binom{m}{i} \exp\left\{-\frac{(\binom{i}{2}p - c_2 i)^2}{2\binom{i}{2}p}\right\} < \sum_{i>m/2k} \left(\frac{em}{i}\right)^i \exp\left\{-\frac{\left(\frac{i-1}{2}p - c_2\right)^2 i}{(i-1)p}\right\} \\ &< \sum_{i>m/2k} (2ek)^i \exp\left\{-\frac{m\left(\frac{c_3}{2}\frac{i-1}{m} - c_2\right)^2}{c_3}\right\} . \end{aligned}$$

Denote the i-th summand in the sum above by b_i. Notice that $c_2 = c_3/(20k) \le (1/5)(i-1)c_3/(2m)$ for $i > m/(2k)$. Hence

$$b_i < (2ek)^i e^{-\frac{m}{c_3}\left(\frac{2c_3(i-1)}{5m}\right)^2} < e^{\ln(2ek)i - \frac{4c_3(i-1)^2}{25m}} < e^{3i\ln k - 3.2(i-1)\ln k} = o(m^{-1}) .$$

Finally, we prove the third part of the lemma. Let $V(H) = C_1 \cup \ldots \cup C_k$ be a k-partition of the vertex set of H. Then by Part 2 of the lemma,

$$\begin{aligned} \sum_{j:|C_j|>\frac{m}{2k}} |\{(u,v) \in E(H) : u,v \in C_j\}| &\ge \sum_{j:|C_j|>\frac{m}{2k}} c_2 \cdot |C_j| \\ &= c_2 m - \sum_{j:|C_j|\le\frac{m}{2k}} c_2 \cdot |C_j| \ge \frac{c_2 m}{2} . \end{aligned}$$

We have thus proven that the random graph $G(m,p)$, with p as defined above, has almost surely the desired properties. $\square$

In order to prove Theorem 1, Part 2, we take the output of Lemma 2.1 and blow it up to show the existence of a graph with the desired properties. Set

$$m = \left\lfloor \frac{c_2}{2\epsilon} \right\rfloor = \left\lfloor \frac{\ln k}{\epsilon} \right\rfloor .$$

Assume that $\epsilon > 0$ is such that the conclusion of Lemma 2.1 holds for $m = m(\epsilon)$ as defined above. Let $H = H_{k,m}$ be the graph from Lemma 2.1. Label the vertices of H by the integers $1, \ldots, m$. For an integer n divisible by m, define a graph $G = (V, E)$ on n vertices as follows. The vertex set $V(G)$ is a union of m disjoint subsets $V_1, \ldots, V_m$, each of size n/m. For each pair $1 \leq i \neq j \leq m$, vertices $u \in V_i$, $u \in V_j$ are connected by an edge in G if and only if $(i,j) \in E(H)$.

Let us now state some properties of the obtained graph G. First, G is easily seen to be homomorphic to H. (We say that G_1 is *homomorphic* to G_2 if there exists a mapping $\phi : V(G_1) \rightarrow V(G_2)$ so that for every edge $(u,v) \in E(G_1)$, $(\phi(u), \phi(v)) \in E(G_2)$). Therefore every subgraph of G is homomorphic to a subgraph of H. As a homomorphism does not decrease the chromatic number, we derive from Lemma 2.1 that every subgraph of G on at most $c_1 m$ vertices is k-colorable.

Next, we need to estimate the distance from G to the set of k-colorable graphs on n vertices. Let $V = C_1 \cup \ldots \cup C_k$ be a k-partition of $V(G)$ with a minimal number of monochromatic edges. Denote the latter by s. Consider a random k-partition of $V(H)$ induced by assigning a color j, $1 \leq j \leq k$, to a vertex i, $1 \leq i \leq m$, with probability $|C_j \cap V_i|/|V_i|$. The expected number of monochromatic edges of H under such a partition is

$$\begin{aligned} \sum_{(i_1,i_2) \in E(H)} \sum_{j=1}^{k} \frac{|C_j \cap V_{i_1}|}{|V_{i_1}|} \frac{|C_j \cap V_{i_2}|}{|V_{i_2}|} &= \frac{1}{(n/m)^2} \sum_{j=1}^{k} \sum_{(i_1,i_2) \in E(H)} |C_j \cap V_{i_1}||C_j \cap V_{i_2}| \\ &= \frac{m^2}{n^2} \sum_{j=1}^{k} |\{(u,v) \in E(G) : u, v \in C_j\}| = \frac{m^2 s}{n^2} . \end{aligned}$$

As by our assumption on H we have that the distance from H to the family of k-colorable graphs on m vertices is at least $c_2 m/2$, we get $s \geq c_2 n^2/(2m)$.

Recalling now the definitions of m and of the constants c_1, c_2, we conclude that G has the following properties:

1. G is ϵ-robustly non-k-colorable;

2. every subgraph of G on at most $c_1 m = \frac{c_1 c_2}{2\epsilon} = \frac{1}{40e}(\frac{1}{120e \ln k})^{\frac{2}{k-2}} \frac{1}{\epsilon}$ vertices is k-colorable.

This implies Theorem 1, Part 2. $\square$

3 Testing bipartiteness

In this section we prove Theorem 2. Our proof exploits the basic elegant idea of Goldreich, Goldwasser and Ron [3]. It is however far more involved technically.

Let us first describe briefly the main idea of the argument of Goldreich et al. for testing bipartiteness. Let $G = (V, E)$ be an ϵ-robustly non-bipartite graph on n vertices. We need to show that a random sample R of size $|R| = \tilde{O}(1/\epsilon^2)$ contains a non-two-colorable subgraph (i.e. an odd cycle) with probability at least $1/2$. The set R will be generated in two stages: $R = S \cup T$, where S is a random subset of size $|S| = \tilde{O}(1/\epsilon)$ and T is a random subset of size $|T| = \tilde{O}(1/\epsilon^2)$. First, it is easy to see that with probability at least $3/4$ such S as above will dominate most of the vertices of G of degree at least $\epsilon n/2$. We assume that S indeed has this property. For a partition $S = S^1 \cup S^2$, denote by U^1 the set of vertices of G of degree at least $\epsilon n/2$, dominated by S^1, let also U^2 be the remaining vertices of degree at least $\epsilon n/2$, dominated by S. We call any edge $e \in E(G)$ spanned by U^1 or by U^2 a *witness* for the partition $S = S^1 \cup S^2$. If a random set T contains a witness for every partition $S = S^1 \cup S^2$, then the union $S \cup T$ is easily seen to span a non-bipartite subgraph.

Recall that G is ϵ-robustly non-bipartite. Therefore, for every partition $S = S^1 \cup S^2$, dominating most of the vertices of degree at least $\epsilon n/2$, at least one of the sets U^1, U^2 should span at least $\epsilon n^2/4$ edges, each of them being a witness for $S^1 \cup S^2$. If we choose the vertices of T of size $|T| = \tilde{O}(1/\epsilon^2)$ pair after pair, then the probability that T does not contain a witness for a fixed partition $S^1 \cup S^2$ is at most $2^{-\tilde{\Omega}(1/\epsilon)} << 2^{-|S|}$. As S has $2^{|S|}$ partitions, by the union bound we obtain that the probability that T does not contain a witness for one of the partitions is much less than $2^{|S|} \cdot 2^{-|S|} = 1$. This implies that the probability that $G[S \cup T]$ is non-bipartite is at least $1/2$.

How tight is the above analysis? At the first stage, $\tilde{\Omega}(1/\epsilon)$ random vertices are needed indeed to dominate most of the vertices of G of degree at least $\epsilon n/2$. As for the second stage, an example of a complete bipartite subgraph $K_{\frac{\epsilon n}{2}, \frac{n}{2}}$ (for the induced subgraph on U^1, say) shows that $\tilde{\Omega}(1/\epsilon^2)$ random vertices are necessary to catch one of its edges with probability $1 - 2^{-\tilde{\Omega}(1/\epsilon)}$. Note however that the subgraph $K_{\frac{\epsilon n}{2}, \frac{n}{2}}$ has $\epsilon n/2$ vertices of degree $n/2$. As this degree is much larger than $\epsilon n/2$, we need to sample only $O(1)$ vertices to dominate most of those high degree vertices. Thus in this case the set S of the first stage does not need to be that large. This in turn reduces the number of partitions of S and makes the requirement for the success probability for a fixed partition of S much less severe.

Our idea will be to represent the first random subset S of size $|S| = \tilde{O}(1/\epsilon)$ as a union of several subsets $S = S_1 \cup \ldots \cup S_t$ with $t = \tilde{O}(\ln(1/\epsilon))$, where each S_i dominates most of the vertices of G of degree about n/e^i (we denote this set by U_i). Each partition $S = S^1 \cup S^2$ induces then partitions of the subsets: $S_i = S_i^1 \cup S_i^2$ and corresponding partitions $U_i = U_i^1 \cup U_i^2$ of the dominated subsets U_i of $V(G)$. Then if G is an ϵ-robustly non-bipartite graph on n vertices, for each partition $S = S^1 \cup S^2$, one of the sets U_i^l, $l = 1, 2$, will span $\tilde{\Omega}(\epsilon n^2)$ edges. Catching any of them will provide a desired witness for this partition of S. As all degrees in U_i^l are at most n/e^i, this will allow us to apply the so called generalized Janson Inequality to show that if $|T| = \tilde{O}(1/\epsilon)$, then T catches one of the edges inside U_i^l with probability at least $1 - O(2^{|S_i|})$. Then applying the union bound will prove the desired result.

The actual proof will deviate somewhat from the above outline as we will need to overcome some further complications.

In the course of the proof we will need the following lemma.

Lemma 3.1 *Let $G = (V, E)$ be a graph on n vertices and let $0 < \delta_2 < \delta_1 < 1/2$ be constants. Suppose A, B are disjoint subsets of V. Then with probability at least $1/2$ a random subset $R \subset V$ of size $|R| = (6/\delta_2)\ln^2(1/\delta_1)$ contains a subset $T \subset A$ having the following properties:*

1. $|T| \leq \frac{1}{\delta_1}$;
2. *Denote*
$$B^* = \{v \in B : N(v) \cap T = \emptyset\} \ . \tag{1}$$
Then
$$\sum_{v \in A:\, d(v,B^*) > \delta_1 n} d(v, B^*) \leq \delta_2 n^2 \ . \tag{2}$$

The lemma asserts the existence of a set T such that if we remove T from A and the neighbors of T from B, most vertex degrees from A to B will be bounded from above by $\delta_1 n$.

Proof of Lemma 3.1. We will generate a random subset R in several steps, each time choosing a random subset R_i of V, where the cardinality of R_i may vary from step to step. At each step we will update the value of T until we will reach T with the desired properties. Then R will be a union of all chosen random subsets R_i.

Denote $s = \ln(1/\delta_1)$. Initially we set $T = \emptyset$, $i = 1$. Define B^* by (1). As long as condition (2) is not satisfied we do the following. For $1 \leq j \leq s$ define $A_j = \{v \in A : e^{j-1}\delta_1 n < d(v, B^*) \leq e^j \delta_1 n\}$. If for all $1 \leq j \leq s$ one has $|A_j| < \frac{\delta_2 n}{e^j \delta_1 s}$, then

$$\begin{aligned}
\sum_{v \in A: d(v,B^*) > \delta_1 n} d(v, B^*) &= \sum_{j=1}^{s} \sum_{v \in A_j} d(v, B^*) \\
&\leq \sum_{j=1}^{s} \frac{\delta_2 n}{e^j \delta_1 s} \cdot e^j \delta_1 n \\
&= \delta_2 n^2
\end{aligned}$$

– a contradiction. Therefore, there exists an index $1 \leq j_0 = j_0(i) \leq s$ for which $|A_{j_0}| \geq \frac{\delta_2 n}{e^{j_0} \delta_1 s}$. Choose a random subset $R_i \subset V$ of size $|R_i| = (e^{j_0}\delta_1 s/\delta_2)\ln(2/\delta_1)$. The probability that R_i does not intersect A_{j_0} is

$$Pr[R_i \cap A_{j_0} = \emptyset] = \frac{\binom{n-|A_{j_0}|}{|R_i|}}{\binom{n}{|R_i|}} \le \left(1 - \frac{|A_{j_0}|}{n}\right)^{|R_i|} \le e^{-\frac{|A_{j_0}||R_i|}{n}} \le \frac{\delta_1}{2} \ .$$

We call step i successful if $R_i \cap A_{j_0} \neq \emptyset$. In this case we choose an arbitrary vertex $v_i \in R_i \cap A_{j_0}$ and denote $d_i = d(v_i, B^*)$. Note that $d_i > e^{j_0-1}\delta_1 n$, implying $|R_i|/d_i \le (es/(\delta_2 n))\ln(2/\delta_1)$. We then add v_i to T, update B^* according to (1), and repeat the above described procedure.

Note that after a successful step has been performed, the size of B^* is decreased by at least $\delta_1 n$. Hence at most $1/\delta_1$ successful steps were executed. Consider the event where all steps were successful until the end of the above described iterative procedure. The probability of this event is at least $1 - (1/\delta_1)(\delta_1/2) = 1/2$. As the size of T is equal to the number of successful steps, we get $|T| \le 1/\delta_1$.

Define now $R = \bigcup_{i=1}^{|T|} R_i$. As $\sum_{i=1}^{|T|} d_i \le |B| \le n$, the size of R can be bounded by

$$|R| = \sum_{i=1}^{|T|} |R_i| \le \sum_{i=1}^{|T|} \frac{es}{\delta_2 n} \ln\left(\frac{2}{\delta_1}\right) d_i \le \frac{es}{\delta_2} \ln\left(\frac{2}{\delta_1}\right) \le \frac{e}{\delta_2} \ln\left(\frac{1}{\delta_1}\right) \ln\left(\frac{2}{\delta_1}\right) < \frac{6}{\delta_2} \ln^2\left(\frac{1}{\delta_1}\right) \ . \quad \Box$$

Now we briefly outline the proof of Theorem 2. A random set R of size $|R| = 34\ln^4(1/\epsilon)\ln\ln(1/\epsilon)/\epsilon$ will be generated in three stages, with each stage producing its own set of random vertices R^j, $j = 1, 2, 3$. At the first stage we construct inside R^1 a family of sets $\{S_i\}$, where each S_i has size about $e^i \ln(1/\epsilon)$ and dominates most of the vertices of G with degrees about n/e^i. We denote by U_i the set of vertices of G of degree about n/e^i, dominated by S_i. Note that U_i is not a subset of R^1, in fact, with high probability most of U_i will be outside R^1. At the second stage we use R^2 to adjust the families $\{S_i\}, \{U_i\}$ in such a way that each S_i still dominates U_i, and for each U_i almost all vertices of $\bigcup_{j=1}^{i-1} U_j$ have their degrees into U_i bounded from above by n/e^i. This is a crucial stage which enables us to complete the union of S_i to a non-bipartite subgraph by choosing a random subset R^3 at the third stage.

Let us now introduce some notation. From now till the end of the section we assume that $G = (V, E)$ is an ϵ-robustly non-2-colorable graph on n vertices. Let

$$t = \ln\left(\frac{1}{\epsilon}\right) \ .$$

Let also, for each $1 \le i \le t + 2$,

$$I_i = \left(\frac{n}{e^i}, \frac{n}{e^{i-1}}\right] \ .$$

Stage 1: defining S_i's, U_i's.

Proposition 3.1 *With probability at least* $5/6$ *a random subset* R^1 *of* V *of size* $|R^1| = 55t/\epsilon$ *contains* $t+2$ *disjoint subsets* $S_1, \ldots, S_{t+2}$ *of cardinalities* $|S_i| = e^{i+1}t$, $i = 1, \ldots, t+2$, *so that for each* $1 \le i \le t+2$ *the number of vertices of* G *with degrees in* I_i, *not dominated by* S_i, *does not exceed* $\frac{\epsilon n}{4(t+2)}$.

Proof. For each $1 \le i \le t+2$ we choose a subset $S_i \subset V$ of size $|S_i| = e^{i+1}t$ uniformly at random and then take R^1 to be the union of the sets S_i. Note that with probability $1 - o(1)$ the sets S_i are pairwise disjoint. Also,

$$\sum_{i=1}^{t+2} |S_i| = \sum_{i=1}^{t+2} e^{i+1}t \le te^{t+4} = e^4 \ln\left(\frac{1}{\epsilon}\right) e^{\ln\left(\frac{1}{\epsilon}\right)} < \frac{55 \ln\left(\frac{1}{\epsilon}\right)}{\epsilon} .$$

Let X_i be a random variable, counting the number of vertices of G with degrees in I_i, not dominated by S_i. If $v \in V$ has its degree in I_i, then the probability that v is not dominated by S_i can be estimated from above by:

$$\frac{\binom{n-d(v)}{|S_i|}}{\binom{n}{|S_i|}} < \left(1 - \frac{d(v)}{n}\right)^{|S_i|} < e^{-\frac{|S_i| n}{e^i n}} = e^{-\frac{e^{i+1}t}{e^i}} = e^{-et} = \epsilon^e.$$

By linearity of expectation we get

$$E[X_i] < n\epsilon^e < \frac{\epsilon n}{80(t+2)^2} .$$

By the Markov inequality $Pr[X_i > \frac{\epsilon n}{4(t+2)}] < 1/(20(t+2))$. Therefore the probability that the family $\{S_i\}_{i=1}^{t+2}$ does not satisfy the claim of the lemma is less than $(t+2)\frac{1}{20(t+2)} + o(1) < 1/6$. $\Box$

Suppose now that the first stage is successful and the family $\{S_i\}_{i=1}^{t+2}$ has the property described in the above proposition. For $1 \le i \le t+2$ we define

$$U_i = \{v \in V : d(v) \in I_i, N(v) \cap S_i \neq \emptyset\} .$$

It follows from Proposition 3.1 that

$$\begin{aligned}
\sum_{v \notin \bigcup_{i=1}^{t+2} U_i} d(v) &\le \sum_{i=1}^{t+2} \sum_{v \in V : d(v) \in I_i, N(v) \cap S_i = \emptyset} d(v) + \sum_{v \in V : d(v) \le n/e^{t+2}} d(v) \\
&\le \sum_{i=1}^{t+2} \frac{\epsilon n}{4(t+2)} \cdot \frac{n}{e^{i-1}} + n \cdot \frac{\epsilon n}{e^2} \\
&< \frac{\epsilon n^2}{2(t+2)} + \frac{\epsilon n^2}{e^2} \\
&< \frac{\epsilon n^2}{2} .
\end{aligned}$$

Stage 2: adjusting S_i's, U_i's.

The purpose of this stage is to achieve the situation, in which for all $2 \le i \le t+2$ most of the degrees of vertices from $\bigcup_{j=1}^{i-1} U_j$ to U_i are bounded from above by n/e^{i-1}. We also want S_i to dominate U_i and the size of S_i to remain basically unchanged. Our main technical tool is Lemma 3.1.

For $i = t+2$ down to 2 we repeat the following procedure. Denote $A = U_1 \cup \ldots U_{i-1}$, $B = U_i$, $\delta_1 = 1/e^{i-1}$, $\delta_2 = \epsilon/(8(t+2))$. Applying Lemma 3.1 $2\ln t$ times we get that with probability at least $1 - 1/(6(t+1))$ a random subset $R_i^2 \subset V$ of size $|R_i^2| = 12 \ln t \ln^2(1/\delta_1)/\delta_2 = 96(t+2)\ln t(i-1)^2/\epsilon$ contains a subset T_i of size $|T_i| = e^{i-1}$ having property (2) with A, B, δ_1 and δ_2 as defined above.

Now we update

$$\begin{aligned} S_{i-1} &:= S_{i-1} \cup T_i \ , \\ U_{i-1} &:= U_{i-1} \cup \{v \in U_i : N(v) \cap T_i \neq \emptyset\} \ , \\ U_i &:= U_i \setminus U_{i-1} \ . \end{aligned}$$

Proposition 3.2 *After having executed the above loop, with probability at least 5/6, the families $\{S_i\}_{i=1}^{t+2}$, $\{U_i\}_{i=1}^{t+2}$ have the following properties:*

1. *for every $2 \le i \le t+2$*

$$\sum_{v \in \bigcup_{j=1}^{i-1} U_j,\, d(v,U_i) > \frac{n}{e^{i-1}}} d(v, U_i) \le \frac{\epsilon n^2}{8(t+2)} \ ; \tag{3}$$

2. *for every $1 \le i \le t+2$*

$$|S_i| \le te^{i+2} \ ; \tag{4}$$

3. *for every $1 \le i \le t+2$ and for every vertex $v \in U_i$,*

$$d(v) \le \frac{n}{e^{i-1}}; \tag{5}$$

4. *Still*

$$\sum_{v \notin \bigcup_{i=1}^{t+2} U_i} d(v) \le \frac{\epsilon n^2}{2} \ . \tag{6}$$

Proof. Note that before starting Stage 2, all vertices in U_i have their degrees bounded from above by n/e^{i-1}. Therefore, moving some of them to U_{i-1} cannot create vertices $v \in \bigcup_{j=1}^{i-1} U_j$ for which $d(v, U_i) > n/e^{i-1}$. Also, as we proceed downwards from $i = t+2$ to $i = 2$, once we have moved vertices from U_i to U_{i-1}, the set U_i remains unchanged. Therefore, (3) follows from Lemma 3.1. Similarly, (4) follows from the estimate $|S_i| \le te^{i+1}$ before the execution of Stage 2 and the fact $|T_{i+1}| = e^i$. Note that the new U_i is a subset of the union of the old U_j, $j = i, \ldots, t+2$. As before

Stage 2 we have $d(v) \le n/e^{i-1}$ for all $v \in \bigcup_{j=i}^{t+2} U_j$, (5) follows. Finally, as the union $\bigcup_{i=1}^{t+2} U_i$ remains the same after Stage 2, (6) follows from the corresponding property of the old sets U_i. $\square$

Let $R^2 = \bigcup_{i=2}^{t+2} R_i^2$ be the random vertices consumed at Stage 2. We have

$$|R^2| = \sum_{i=2}^{t+2} |R_i^2| = \sum_{i=2}^{t+2} \frac{96(t+2)\ln t(i-1)^2}{\epsilon} < \frac{33t^4 \ln t}{\epsilon} .$$

Stage 3: Completing $\bigcup_{i=1}^{t+2} S_i$ to a non-bipartite subgraph.

Assume now that the graph G on n vertices is ϵ-far from being bipartite. Our aim is to show that with probability at least $11/12$ the union of $\bigcup_{i=1}^{t+2} S_i$, with S_i as defined in the end of Stage 2, and a random subset $R^3 \subset V$ of an appropriately chosen size forms a non-bipartite subgraph of G. This will follow easily from the proposition below.

Proposition 3.3 *Let $G = (V, E)$ be an ϵ-robustly non-bipartite graph on n vertices. Let the subsets $\{S_i\}_{i=1}^{t+2}$, $\{U_i\}_{i=1}^{t+2}$ satisfy (3)–(6). Denote $S = \bigcup_{i=1}^{t+2} S_i$. Then with probability at least $5/6$ a random subset R^3 of size $|R^3| = 2700t^2/\epsilon$ has the following property. For every partition $S = S^1 \cup S^2$ of S there exists an edge $e = (u, v) \in E(G)$ with $u, v \in R^3$ and both u, v having neighbors in the same S^l for some $l \in \{1, 2\}$.*

Proof. For a fixed partition $S = S^1 \cup S^2$ we denote, for $1 \le i \le t+2$, $l = 1, 2$, $S_i^l = S^l \cap S_i$. We also set $U_i^1 = \{v \in U_i : N(v) \cap S_i^2 \neq \emptyset\}$, $U_i^2 = U_i \setminus U_i^1$. Let G_i^l be the following graph. The vertex set of G_i^l is $\bigcup_{j=1}^{i} U_j^l$; an edge $e = (u, v) \in E(G)$ is an edge of G_i^l if and only if $u, v \in U_i^l$ or $u \in \cup_{j=1}^{i-1} U_j^l$, $v \in U_i^l$ and $d(u, U_i^l) \le n/e^{i-1}$. Note that by (5) all degrees in G_i^l are at most n/e^{i-1}.

As the graph G is at least ϵn^2 edges far from any bipartite graph, we get, recalling (6), that either U^1 or U^2 span at least $\epsilon n^2/4$ edges. Therefore, for some $1 \le i \le t+2$, $l \in \{1, 2\}$ we have:

$$e(\bigcup_{j=1}^{i-1} U_j^l, U_i^l) + e(U_i^l) \ge \frac{\epsilon n^2}{4(t+2)} .$$

A partition (S^1, S^2) of S is called (i, l)-*bad*, if $|E(G_i^l)| \ge \epsilon n^2/(8(t+2))$ and $|E(G_j^{l'})| < \epsilon n^2/(8(t+2))$ for all $j < i$, $l' \in \{1, 2\}$. From the definition of G_i^l we get, using (3), that any partition (S^1, S^2) is (j, l)-bad for some $1 \le j \le t+2$, $l \in \{1, 2\}$.

Two (j, l)-bad partitions (S^1, S^2), $((S^1)', (S^2)')$ are called *equivalent* if $S_i^1 = (S_i^1)'$ for $1 \le i \le j$ (and thus $U_i^1 = (U_i^1)'$). By (4) for a fixed $1 \le j \le t+2$, the total number of equivalence classes of (j, l)-bad partitions, where $l \in \{1, 2\}$, is at most

$$2 \cdot 2^{\sum_{i=1}^{j} |S_i|} \le 2^{1 + \sum_{i=1}^{j} e^{i+2} t} \le e^{e^{j+3} t} .$$

Note crucially that two (j, l)-bad partitions in the same equivalence class have the same graph G_j^l. It follows easily from this observation that it is enough to prove that with probability at least

5/6 the random subset R^3 spans an edge of G_j^l, for every $1 \le j \le t+2$, every $l \in \{1,2\}$ and every equivalence class of (j,l)-bad partitions.

In this proof it is convenient to generate R^3 by choosing each vertex $v \in V$ independently with probability $p = 2700t^2/\epsilon n$. This will allow us to use the so called Generalized Janson Inequality (see, e.g. [1], Ch. 8) to estimate the probability that R^3 misses all edges of G_j^l for some fixed equivalence class of (j,l)-bad partitions.

Consider some fixed equivalence class of (j,l)-bad partitions and its graph G_j^l. Note that $|E(G_j^l| \ge \epsilon n^2/(8(t+2))$ and also that the maximal degree of G_j^l is bounded from above by n/e^{j-1}. Denote by Y the random variable counting the number of edges of G_j^l, spanned by R^3. Then $E[Y] = |E(G_j^l)|p^2$. Our aim is to estimate from above the probability that R^3 spans no edges of G_j^l, i.e. $Pr[Y=0]$. A naive analysis performed by choosing the vertices of R^3 pair after pair and requiring that each pair does not coincide with an edge of G_j^l gives only $Pr[Y=0] \le (1 - |E(G_j^l)|/\binom{n}{2})^{|R^3|/2}$. We will get a better estimate, using the assumption on the maximal degree in G_j^l. Let

$$\Delta = 2 \sum_{\substack{e \ne e' \in E(G_j^l) \\ e \cap e' \ne \emptyset}} Pr[e, e' \subset R^3] \,.$$

Then

$$\begin{aligned}
\Delta &= \sum_{e \in E(G_j^l)} \sum_{\substack{e \ne e' \in E(G_j^l) \\ e' \cap e \ne \emptyset}} Pr[e, e' \subset R^3] \\
&= \sum_{e=(u,v) \in E(G_j^l)} ((d_{G_j^l}(u) - 1) + (d_{G_j^l}(v) - 1))p^3 \\
&< \sum_{e \in E(G_j^l)} \frac{2np^3}{e^{j-1}} = \frac{2|E(G_j^l)|np^3}{e^{j-1}} \,.
\end{aligned}$$

By the Generalized Janson Inequality,

$$Pr[Y=0] \le e^{-\frac{(E[Y])^2}{3\Delta}} = e^{-\frac{|E(G_j^l)|e^{j-1}p}{6n}} \le e^{-\frac{\epsilon n e^{j-1} p}{48(t+2)}} < \frac{e^{-e^{j+3}t}}{6(t+2)} \,.$$

Recalling the estimate on the number of equivalence classes of (j,l)-bad partitions, we conclude that the probability that R^3 does not contain an edge of G_j^l for some equivalence class is at most

$$\sum_{j=1}^{t+2} e^{e^{j+3}t} \frac{e^{-e^{j+3}t}}{6(t+2)} = 1/6 \,. \qquad \Box$$

Assume now that Stages 1 and 2 were successful and the set R^3 has the property stated in Proposition 3.3. Then it is easy to see that the spanned subgraph $G[S \cup R^3]$ is not bipartite.

Indeed, let $c : S \cup R^3 \to \{1,2\}$ be a 2-coloring of $S \cup R$. Define a partition $S = S^1 \cup S^2$ of S by $S^1 = \{v \in S : c(v) = 1\}$, $S^2 = \{v \in S : c(v) = 2\}$. Then R^3 contains an edge $e = (u,v) \in E(G)$ with both endpoints u, v connected to one color class, say, S^1. If c colors u or v in color 1, we get a monochromatic edge connecting u or v, respectively, with S^1. Otherwise, $c(u) = c(v) = 2$, but then e is monochromatic under c. By the above analysis with probability at least $2/3$ the random sets R^1 and R^2 define a subset $S \subset R^1 \cup R^2$ with the properties stated in Proposition 3.2. Therefore, with probability at least $1/2$ the union $S \cup R^3$ spans a non-bipartite subgraph of G.

It remains only to estimate the size of the random set $R = R^1 \cup R^2 \cup R^3$. We have

$$|R| = |R^1| + |R^2| + |R^3| = \frac{55t}{\epsilon} + \frac{33t^4 \ln t}{\epsilon} + \frac{2700t^2}{\epsilon} < \frac{34 \ln^4\left(\frac{1}{\epsilon}\right) \ln\ln\left(\frac{1}{\epsilon}\right)}{\epsilon}\ .$$

The proof of Theorem 2 is complete. $\square$

4 Testing k-colorability

In this section we prove Theorem 3. It will be convenient to generate a random subset $R \subset V(G)$ of size $|R| = s = 36k \ln k/\epsilon^2$ in s rounds, each time choosing uniformly at random a single vertex $r_j \in V(G)$. This in principle may result in choosing one vertex several times and thus getting a set of cardinality less than s. However, the probability of this event is $o(1)$, and therefore the approach for generating R we take here is asymptotically equivalent to choosing a subset of V of size s uniformly at random.

Our basic approach is similar to the one of Goldreich et al. [3]. At the end of the section we explain the main differences and the reason our argument saves a factor of $\Theta(1/\epsilon)$ in the number of vertices sampled.

Let G be an ϵ-robustly non-k-colorable graph on n vertices. Suppose we are given a subset $S \subset V(G)$ (of the sample set R), and its k-partition $\phi : S \to [k]$, our aim is to find with high probability inside the next several random vertices a succinct witness to the fact that ϕ can not be extended to a proper coloring of the sample. If a k-coloring $c : V(G) \to [k]$ of G is to coincide with ϕ on S, then for every vertex $v \in V \setminus S$, the colors of neighbors of v in S under ϕ are forbidden for v in c. The rest of the colors are still feasible for v. It could be that v has no feasible colors left at all. Such a vertex will be called colorless with respect to S and ϕ. If the number of colorless vertices is large, then there is a decent chance that between the next few random vertices of R there will be one such colorless vertex v^*. Obviously, adding v^* to S provides the desired witness for non-extendibility of ϕ.

If the set of colorless vertices is small, then one can show that, as G is ϵ-far from being k-colorable, there is a relatively large subset W of vertices (which will be called restricting) such that adding any vertex $v \in W$ to S and coloring it by any feasible color with regard to ϕ excludes this color from the lists of feasible colors of at least $\Omega(\epsilon)n$ neighbors of v. If such v is caught in the next few vertices of the random sample R, then adding v to S and coloring it by any of its feasible colors reduces substantially the total length of the lists of feasible colors for the vertices of V, thus helping to approach the first situation, i.e. the case when there are many colorless vertices. As the reader can guess, the above described process can be represented by a tree in which every node corresponds to a colorless or restricting vertex v and each edge corresponds to a feasible color for v. As the degree of such a node can be as large as k, the size of the tree grows quickly as we proceed with choosing vertices from R, and can reach size exponential in $1/\epsilon$. We therefore will need the probability of success (i.e. the probability of catching a colorless/restricting vertex) along several consecutive steps to be exponentially close to 1.

Now we present a formal description of the above argument. First we need to introduce some notation. We denote the set $\{1,\dots,k\}$ by $[k]$. Suppose $G=(V,E)$ is a graph on n vertices. Given a subset $S\subseteq V$ and its k-partition $\phi: S\rightarrow [k]$, for every $v\in V\setminus S$ let

$$L_\phi(v)=[k]\setminus\{1\le i\le k: \exists u\in S\cap N(v), \phi(u)=i\}\ .$$

If $S=\emptyset$, we set $L_\phi(v)=[k]$ for every $v\in V$. If a k-coloring $c: V\rightarrow [k]$ of G coincides with ϕ on S, then for every $v\in V\setminus S$ the color of v in c belongs to $L(v)$. For this reason, the set $L_\phi(v)$ is called the *list of feasible colors* for v. A vertex $v\in V\setminus S$ is called *colorless* if $L_\phi(v)=\emptyset$. We denote by U the set of all colorless vertices under (S,ϕ).

For every vertex $v\in V\setminus(S\cup U)$ define

$$\delta_\phi(v)=\min_{i\in L_\phi(v)}|\{u\in N(v)\setminus(S\cup U): i\in L(u)\}|\ .$$

Thus coloring v by one of the colors from $L_\phi(v)$ and then adding it to S results in deleting this color and thus shortening the lists of feasible colors of at least $\delta_\phi(v)$ neighbors of v outside S.

Claim 4.1 *For every set $S\subset V$ and every k-partition ϕ of S, the graph G is at most $(n-1)|S\cup U|+\sum_{v\in V\setminus(S\cup U)}\delta_\phi(v)$ edges far from being k-colorable.*

Proof. For every $v\in S$, color v according to $\phi(v)$. For every $v\in U$ we color v in an arbitrary color from $[k]$. For every $v\in V\setminus(S\cup U)$ we color v in color $i\in L_\phi(v)$ for which $\delta_\phi(v)=|\{u\in N(v)\setminus(S\cup U): i\in L_\phi(u)\}|$. Let us estimate the number of monochromatic edges under this coloring. The number of monochromatic edges incident with $S\cup U$ is at most $(n-1)|S\cup U|$. Every vertex $v\in V\setminus(S\cap U)$ has exactly $\delta_\phi(v)$ neighbors $u\in V\setminus(S\cup U)$, whose color list $L_\phi(v)$ contains the color chosen for v. Therefore, v will have at most $\delta_\phi(v)$ neighbors in $V\setminus(S\cup U)$ colored in the same color. Hence the total number of monochromatic edges is at most $(n-1)|S\cup U|+\sum_{v\in V\setminus(S\cup U)}\delta_\phi(v)$, as claimed. $\Box$

Corollary 4.1 *If G is an ϵ-robustly non-k-colorable graph on n vertices, then for any pair (S,ϕ), where $S \subset V(G)$, $\phi : S \to [k]$, one has:*

$$\sum_{v \in V \setminus (S \cup U)} \delta_\phi(v) > \epsilon n^2 - n(|S| + |U|)\,,$$

where U is the set of colorless vertices for the pair (S,ϕ).

Given a pair (S,ϕ), a vertex $v \in V \setminus (S \cup U)$ is called *restricting* if $\delta_\phi(v) \geq \epsilon n/2$. We denote by W the set of all restricting vertices.

Claim 4.2 *If G is an ϵ-robustly non-k-colorable graph on n vertices, then for every pair (S,ϕ), where $S \subset V(G)$ and $\phi : S \to [k]$, one has:*

$$|U \cup W| > \frac{\epsilon n}{2} - |S|\ .$$

Proof. By Corollary 4.1,

$$\epsilon n^2 - n(|S| + |U|) < \sum_{v \in V \setminus (S \cup U)} \delta_\phi(v) \leq |W|(n-1) + \sum_{v \in V \setminus (S \cup U \cup W)} \delta_\phi(v) < |W|n + \frac{n \cdot \epsilon n}{2}\,.$$

This implies $|S| + |U| + |W| \geq \epsilon n/2$. As U and W are disjoint, the result follows. □

Let now G be an ϵ-robustly non-k-colorable graph on n vertices. While choosing random vertices $r_1, \ldots, r_s$ of R we construct an auxiliary k-ary tree T. To distinguish between the vertices of G and those of T we call the latter *nodes*. Each node of T is labeled either by a vertex of G or by the special symbol #, whose meaning will be explained soon. If a node t of T is labeled by #, then t is called a *terminal node*. The edges of T are labeled by integers from $[k]$.

Let t be a node of T. Consider the path from the root of T to t, not including t itself. The labels of the nodes along this path form a subset $S(t)$ of $V(G)$. The labels of the edges along the path define a k-partition $\phi(t)$ of $S(t)$ in the natural way: the label of the edge following a node t' in the path determines the color of its label $v(t')$. The labeling of the nodes and edges of T will have the following property: if t is labeled by v and v has a neighbor in $S(t)$ whose color in $\phi(t)$ is i, then the son of v along the edge labeled by i is labeled by #. This label indicates the fact that in this case color i is infeasible for v, given $(S(t), \phi(t))$.

At each step of the construction of T we will maintain the following: all leafs of T are either unlabeled or are labeled by #. Also, only leafs of T can be labeled by #. We start the construction of T from an unlabeled single node, the root of T.

Suppose that $j-1$ vertices of T have already been chosen, and we are about to choose vertex r_j of R. Consider a leaf t of T. If t is labeled by #, we do nothing for this leaf. (That is the reason such a t is called a terminal node; nothing will ever grow out of this node.) Assume now that t is unlabeled. Define the pair $(S(t), \phi(t))$ as described above. Now, for the pair $(S(t), \phi(t))$ we define the set $U(t)$ of colorless vertices and the set $W(t)$ of restricting vertices as described before. Round j is called *successful* for the node t if the random vertex r_j satisfies: $r_j \in U(t) \cup W(t)$. If round j is indeed successful for t, then we label t by r_j, create k sons of t and label the corresponding edges by $1, \ldots, k$. Now, if color i is infeasible for r_j, given $(S(t), \phi(t))$, we label the son of t along the edge with label i by #, otherwise we leave this son unlabeled. Note that if $r_j \in U(t)$, then none of the colors from $[k]$ is feasible for r_j, and thus all the sons of t will be labeled by #. This completes the description of the process of constructing T.

Now we state some properties of T.

Claim 4.3 *The depth of T is bounded from above by $\frac{2k}{\epsilon}$.*

Proof. Let t^* be a leaf of T. Notice that if the label of a node t of T belongs to $U(t)$, then all sons of t in T are labeled by # and are terminal nodes. Therefore all nodes on the path from the root of T to t^*, but possibly the node immediately preceding t^*, have their labels in the corresponding sets $W(t)$. Since each vertex in $W(t)$ is restricting with respect to $(S(t), \phi(t))$, coloring v in any feasible color decreases the total size of the lists of feasible colors for all vertices of G by at least $\epsilon n/2$. Therefore, each time when on the path from the root of T to t^* we leave a node t, whose label belongs to $W(t)$, the total length of the list of feasible colors shrinks by at least $\epsilon n/2$. As initially all k colors are feasible for all vertices, we start with lists of feasible colors of total length nk. Thus we cannot make more than $nk/(\epsilon n/2)) = 2k/\epsilon$ steps down from the root of T to t^*. This implies that the depth of T is at most $2k/\epsilon$. □

Claim 4.4 *If a leaf t^* of T is labeled by #, then $\phi(t^*)$ is not a proper k-coloring of $S(t^*)$.*

Proof. By the definition of the labeling procedure: let t' be the father of t^* in T. Let v be the label of t', and let i be the label of the edge of T connecting t' and t^*. Since t^* is labeled by '#', i is not a feasible color for v, given $(S(t'), \phi(t'))$. As $\phi(t^*)$ colors v in color i, we get the existence of an edge spanned by $S(t^*)$, incident with v and monochromatic under $\phi(t^*)$. □

Claim 4.5 *If after round j all leafs of the tree T are terminal nodes, then the subgraph $G[\{r_1, \ldots, r_j\}]$ is not k-colorable.*

Proof. Notice first that the labels of all nodes of T are either # or vertices from $\{r_1, \ldots, r_j\}$. Let $c : \{r_1, \ldots, r_j\} \to [k]$ be a k-partition of $\{r_1, \ldots, r_j\}$. In order to show that c creates some monochromatic edges in the induced subgraph of G on $\{r_1, \ldots, r_j\}$, we start with the root t_0 of T and traverse T guided by c as follows: while at a node t of T, labeled by $v(t) \in \{r_1, \ldots, r_j\}$, we move from t to its son along the edge of T labeled by $c(v(t))$. Once we reach a terminal node t^* of T, we have then $S(t^*) \subseteq \{r_1, \ldots, r_j\}$ and $\phi(t^*)$ coincides with c on $S(t^*)$. As t^* is a terminal node, it follows from Claim 4.4 that c is not a proper k-coloring of $S(t^*)$. □

Claim 4.6 *If G is ϵ-robustly non-k-colorable graph on n vertices, then after $36k\ln k/\epsilon^2$ rounds with probability at least $1/2$ all leaves of T are terminal nodes.*

Proof. As every non-leaf node of T has k sons and by Claim 4.3 T has depth at most $2k/\epsilon$, it can be embedded naturally in the k-ary tree $T_{k,\frac{2k}{\epsilon}}$ of depth $2k/\epsilon$. Moreover, this embedding can be prefixed even before exposing R and T. Note that the number of vertices of $T_{k,\frac{2k}{\epsilon}}$ is $1+k+\ldots+k^{\frac{2k}{\epsilon}} \leq k^{\frac{2k}{\epsilon}+1}$.

Recall that during the construction of the random sample R and the tree T, a successful round for a leaf t of T results in creating k sons of T. Fix some node t of $T_{k,\frac{2k}{\epsilon}}$. If after $36k\ln k/\epsilon^2$ rounds t is a leaf of T, then the total number of successful rounds for the path from the root of T to t is equal to the depth of t. As $S(t) \subseteq R$ and thus $|S(t)| = O(1)$, by Claim 4.2 each round has probability of success at least $\epsilon/3$. Therefore, the probability that t is a non-terminal leaf of T after $36k\ln k/\epsilon^2$ steps can be bounded from above by the probability that the Binomial random variable $B(36k\ln k/\epsilon^2, \epsilon/3)$ is less than $2k/\epsilon$. The latter probability is at most

$$\exp\left\{-\frac{\left(\frac{12k\ln k}{\epsilon}-\frac{2k}{\epsilon}\right)^2}{\frac{24k\ln k}{\epsilon}}\right\} < \exp\left\{-\frac{\left(\frac{9k\ln k}{\epsilon}\right)^2}{\frac{24k\ln k}{\epsilon}}\right\} = e^{-\frac{27k\ln k}{8\epsilon}} < k^{-\frac{3k}{\epsilon}}\ .$$

Thus by the union bound we conclude that the probability that some node of $T_{\epsilon,\frac{2k}{\epsilon}}$ is a leaf of T, non labeled by '#', is at most $|(V(T_{k,\frac{2k}{\epsilon}})|k^{-\frac{3k}{\epsilon}} < \frac{1}{2}$. □

Proof of Theorem 3. Follows immediately from Claims 4.5 and 4.6. □

Note that our proof here is similar to the basic argument of Goldreich et al. in [3]. They also construct (implicitly) the tree T constructed in the course of our proof. Their argument can be briefly described as follows: given a current tree T, Goldreich et al. require that the next subset R_i of a random sample R contains, with high probability, for every leaf $t \in T$, a vertex $v \in U(t) \cup W(t)$. As each random vertex r_j hits $U(t) \cup W(t)$ with probability at least $\epsilon/3$ by Claim 4.2, the probability that for a fixed $t \in T$ the next $\tilde{\Theta}(1/\epsilon^2)$ random vertices will not hit the set $U(t) \cup W(t)$ is at most $(1-1/\epsilon)^{\tilde{\Theta}(1/\epsilon^2)} = 2^{-\tilde{\Theta}(1/\epsilon)}$. The number of leafs of T is at most $2^{O(1/\epsilon)}$. Therefore, by the union bound the set $R_j \subset R$ of $|R_j| = \tilde{O}(1/\epsilon^2)$ random vertices hits the set $U(t) \cup W(t)$ for every leaf $t \in T$ with probability $1 - 2^{O(1/\epsilon)}2^{-\tilde{\Theta}(1/\epsilon)} = 1 - O(1/\epsilon)$. Thus, representing $R = R_1 \cup \ldots \cup R_{\frac{2k}{\epsilon}}$ with $|R_j| = \tilde{O}(1/\epsilon^2)$, they ensure that almost surely each time after having chosen the next piece R_j of random vertices, all non-terminal leaves of T will get k sons each. As by Claim 4.3 the depth of T is bounded by $2k/\epsilon$, after having chosen all $2k/\epsilon$ random pieces $R_1, \ldots, R_{\frac{2k}{\epsilon}}$, almost surely all leaves of T are terminal nodes. In contrast, in our proof we only require that along each path in the tree $T_{k,\frac{2k}{\epsilon}}$ sufficiently many steps will be successful, not insisting on the regularity of appearance of successful steps. This results in saving a factor of $\tilde{\Theta}(1/\epsilon)$.

5 Concluding remarks and open problems

As mentioned in the introduction, the study of the function $g_k(n, \epsilon)$ is motivated by its relevance to the design of efficient testing algorithms for k-colorability. Thus, Theorem 2 shows that bipartiteness can be tested by choosing randomly some $\tilde{O}(1/\epsilon)$ random vertices, and by checking if the induced subgraph on them is 2-colorable. Here, as usual, $\tilde{O}(1/\epsilon)$ denotes $\frac{(\log(1/\epsilon))^{O(1)}}{\epsilon}$. Moreover, by Theorem 1, part 1, any bipartiteness testing algorithm that checks induced subgraphs and is a one-way error algorithm (that is, never errs on bipartite graphs), must check induced subgraphs on at least $\Omega(1/\epsilon)$ vertices.

Similarly, Theorem 3 provides, for every fixed $k \geq 3$, a one-way error algorithm that tests k-colorability by checking random induced subgraphs on $O(1/\epsilon^2)$ vertices. Both algorithms improve the results in [3].

It will be nice to close the gap between our upper and lower bounds for the functions $g_k(n, \epsilon)$ and $f_k(n, \epsilon)$ for $k \geq 3$. It is plausible to conjecture that for every fixed $k \geq 3$, $g_k(n, \epsilon) = \tilde{O}(1/\epsilon)$ and $f_k(n, \epsilon) = \tilde{O}(1/\epsilon)$. This remains open.

Finally we note that Goldreich et al. measure the complexity of their algorithms for graph property testing by the number of pairs of vertices (u, v) of the input graph G queried by the algorithm. The query complexity of our algorithms for testing k-colorability is $\tilde{O}(1/\epsilon^2)$ for $k = 2$ and $\tilde{}(1/\epsilon^4)$ for $k \geq 3$. It is easy to prove a lower bound of $\Omega(1/\epsilon)$ for testing k-colorability. It would be quite interesting to obtain tighter bounds for the query complexity of this problem.

Acknowledgment. We would like to thank the anonymous referees for very careful reading of the first version of the paper and for many helpful comments and suggestions.

References

[1] N. Alon and J. H. Spencer, **The probabilistic method**, Wiley, New York, 1992.

[2] B. Bollobás, P. Erdös, M. Simonovits and E. Szemerédi, Extremal graphs without large forbidden subgraphs, *Annals of Discrete Mathematics* 3 (1978), 29–41.

[3] O. Goldreich, S. Goldwasser and D. Ron, Property testing and its connection to learning and approximation, *Proceedings of the* 37^{th} *Annual IEEE FOCS* (1996), 339–348. Also: Journal of the ACM 45 (1998), 653–750.

[4] J. Komlós, *Covering odd cycles*, Combinatorica 17 (1997), 393–400.

[5] V. Rödl and R. Duke, On graphs with small subgraphs of large chromatic number, *Graphs and Combinatorics* 1 (1985), 91–96.

[6] E. Szemerédi, Regular partitions of graphs, In: *Proc. Colloque Inter. CNRS* (J. C. Bermond, J. C. Fournier, M. Las Vergnas and D. Sotteau, eds.), 1978, 399–401.