Why almost all k-colorable graphs are easy

Amin Coja-Oghlan[1], Michael Krivelevich[2] and Dan Vilenchik[3]

[1] Institute for Informatics, Humboldt-University, Berlin, Germany. coja@informatik.hu-berlin.de.
[2] School Of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv, Israel. krivelev@post.tau.ac.il.
[3] School of Computer Science, Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv, Israel. vilenchi@post.tau.ac.il.

Abstract. Coloring a k-colorable graph using k colors ($k \geq 3$) is a notoriously hard problem. Considering average case analysis allows for better results. In this work we consider the uniform distribution over k-colorable graphs with n vertices and exactly cn edges, c greater than some sufficiently large constant. We rigorously show that all proper k-colorings of most such graphs are clustered in one cluster, and agree on all but a small, though constant, number of vertices. We also describe a polynomial time algorithm that finds a proper k-coloring for $(1-o(1))$-fraction of such random k-colorable graphs, thus asserting that most of them are "easy". This should be contrasted with the setting of very sparse random graphs (which are k-colorable *whp*), where experimental results show some regime of edge density to be difficult for many coloring heuristics. One explanation for this phenomena, backed up by partially non-rigorous analytical tools from statistical physics, is the complicated clustering of the solution space at that regime, unlike the more "regular" structure that denser graphs possess. Thus in some sense, our result rigorously supports this explanation.

1 Introduction

A *k-coloring* f of a graph $G = (V, E)$ is a mapping from its set of vertices V to $\{1, 2, ..., k\}$. f is a *proper coloring* of G if for every edge $(u, v) \in E$, $f(u) \neq f(v)$. The minimal k s.t. G admits a proper k-coloring is called the chromatic number, commonly denoted by $\chi(G)$. In this work we think of $k > 2$ as some fixed integer, say $k = 3$ or $k = 100$.

1.1 Phase Transitions, Clusters, and Graph Coloring Heuristics

The problem of properly k-coloring a k-colorable graph is one of the most famous NP-hard problems. The plethora of worst-case NP-hardness results for problems in graph theory motivates the study of heuristics that give "useful" answers for "typical" subset of the problem instances, where "useful" and "typical" are usually not well defined. One way of evaluating and comparing heuristics is by running them on a collection of input graphs ("benchmarks"), and checking which heuristic usually gives better results. Though empirical results are sometimes informative, we seek more rigorous measures of evaluating heuristics. Although satisfactory approximation algorithms are known for several NP-hard problems, the coloring problem is not amongst them. In fact, Feige and Kilian [13] prove that no polynomial time algorithm approximates $\chi(G)$ within a factor of $n^{1-\varepsilon}$ for all input graphs G on n vertices, unless ZPP=NP.

When very little can be done in the "worst case", comparing heuristics' behavior on "typical", or "average", instances comes to mind. One possibility of rigourously modeling such "average" instances is to use random models. In the context of graph coloring, the $\mathcal{G}_{n,p}$ and $\mathcal{G}_{n,m}$ models, pioneered by Erdős and Rényi, might appear to be the most natural candidates. A random graph G in $\mathcal{G}_{n,p}$ consists of n vertices, and each of the $\binom{n}{2}$ possible edges is included w.p. $p = p(n)$ independently of the others. In $\mathcal{G}_{n,m}$, $m = m(n)$ edges are picked uniformly at random. Bollobás [7] and Łuczak [21] calculated the probable value of $\chi(\mathcal{G}_{n,p})$ to be *whp* [4] approximately $n \ln(1-p)/(2\ln(np))$ for $p \in [C_0/n, 0.99]$. Thus, the chromatic number of $\mathcal{G}_{n,p}$ is typically rather high (roughly comparable with the average degree np of the random graph) – higher than k, when thinking of k as some fixed integer, say $k = 3$, and allowing the average degree np to be arbitrarily large.

Remarkable phenomena occurring in the random graph $\mathcal{G}_{n,m}$ are **phase transitions**. With respect to the property of being k-colorable, such a phase transition takes place too. More precisely, there exists a threshold $d_k = d_k(n)$ such that graphs with average degree $2m/n > (1+\varepsilon)d_k$ do not admit any proper k-coloring *whp*, while graphs with a lower average degree $2m/n < (1-\varepsilon)d_k$ will have one *whp* [1]. In fact, experimental results show that random graphs with average degree just below the k-colorability threshold (which are thus k-colorable *whp*) are "hard" for many coloring heuristics. One possible explanation for this empirical observation, backed up by partially non-rigorous analytical tools from statistical physics [22], is the surmise that k-colorable graphs with average degree just below the threshold show a **clustering phenomenon** of the solution space. That is, typically random graphs with density close to the threshold d_k have an exponential number of **clusters** of k-colorings. While any two k-colorings in distinct clusters disagree on at least εn vertices, any two k-colorings within one cluster coincide on $(1-\varepsilon)n$ vertices. Furthermore, each cluster has a linear number of "frozen" vertices whose colors coincide in **all** colorings within that cluster.

Now, the algorithmic difficulty with such a clustered solution space seems to be that the algorithm does not "steer" into one cluster but tries to find a "compromise" between the colorings in distinct clusters, which actually is impossible. By contrast, the recent Survey Propagation algorithm can apparently cope with the existence of a huge number of clusters [9], though no rigorous analysis of the algorithm is known.

[4] Writing *whp* ("with high probability") we mean with probability tending to 1 as n goes to infinity.

In this work we consider the regime of denser graphs, i.e., the average degree will be by a constant factor higher than the k-colorability threshold. In this regime, almost all graphs are not k-colorable, and therefore we shall condition on the event that the random graph is k-colorable. Thus, we consider the most natural distribution on k-colorable graphs with given numbers n of vertices and m of edges, namely, the uniform distribution $\mathcal{G}^{\text{uniform}}_{n,m,k}$. For $m/n \geq C_0$, C_0 a sufficiently large constant, we are able to **rigorously** prove that the space of all legal k-colorings of a typical graph in $\mathcal{G}^{\text{uniform}}_{n,m,k}$ has the following structure.

- There is an exponential number of legal k-colorings, which are arranged in a **single cluster**.
- We describe a coloring algorithm, and using the same tools that provide the latter observation, we prove that it k-colors *whp* $\mathcal{G}^{\text{uniform}}_{n,m,k}$ with $m \geq C_0 n$ edges using polynomial time.

Thus, our result shows that when a k-colorable graph has a single cluster of k-colorings, though its volume might be exponential, then typically, the problem is easy. This in some sense complements the results in [22] in a rigorous way (where it is conjectured that when the clustering is complicated, more sophisticated algorithms are needed). Besides, standard probabilistic calculations show that when $m \geq Cn \log n$, C a sufficiently large constant, a random k-colorable graph will have *whp* only one proper k-coloring; indeed, it is known that such graphs are even easier to color than in the case $m = O(n)$, which is the focus of this paper. A further appealing implication of our result is the fact that almost all k-colorable graphs, sparse or dense, can be efficiently colored. This extends a previous result from [24] concerning dense graphs (i.e., $m = \Theta(n^2)$).

1.2 Results and Techniques

A subset of vertices $U \subseteq V$ is said to be **frozen** in G if in every proper k-coloring of G, all vertices in U receive the same color. A vertex is said to be **frozen** if it belongs to a frozen subset of vertices. Here and throughout we consider two k-colorings to be the same if one is a permutation of the color classes of the other.

Theorem 1. *(clustering phenomena) Let G be random graph from $\mathcal{G}^{\text{uniform}}_{n,m,k}$, $m \geq C_0(k)n$, $C_0(k)$ a sufficiently large constant that depends on k. Then whp G enjoys the following properties:*

1. *All but $e^{-\Theta(m/n)}n$ of the vertices are frozen.*
2. *The graph induced by the non-frozen vertices decomposes into connected components of at most logarithmic size.*
3. *Letting $\beta(G)$ be the number of proper k-colorings of G, we have $\frac{1}{n}\log \beta(G) = e^{-\Theta(m/n)}$.*

Theorem 2. *(algorithm) There exists a polynomial time algorithm that whp properly k-colors a random graph from $\mathcal{G}^{\text{uniform}}_{n,m,k}$, $m \geq C_1(k)n$, $C_1(k)$ a sufficiently large constant that depends on k.*

It is not hard to see that Property 1 in Theorem 1 implies in particular that any two proper k-colorings of G differ on at most $e^{-\Theta(m/n)}n$ vertices.

In Theorem 1, our analysis gives for $C_0 = \Theta(k^4)$, and in Theorem 2, $C_1 = \Theta(k^6)$, but no serious attempt is made to optimize the power of k.

The Erdős-Rényi graph $\mathcal{G}_{n,m}$ and its well known variant $\mathcal{G}_{n,p}$ are both very well understood and have received much attention during the past years. However the distribution $\mathcal{G}^{\text{uniform}}_{n,m,k}$ differs from $\mathcal{G}_{n,m}$ significantly, as the event of a random graph in $\mathcal{G}_{n,m}$ being k colorable, when k is fixed, and $2m/n$ is some constant above the k-colorability threshold, is very unlikely. In effect, many techniques that have become standard in the study of $\mathcal{G}_{n,m}$ just do not carry over to $\mathcal{G}^{\text{uniform}}_{n,m,k}$ – at least not directly. In particular, the contriving event of being k-colorable causes the edges in $\mathcal{G}^{\text{uniform}}_{n,m,k}$ to be dependent. The inherent difficulty of $\mathcal{G}^{\text{uniform}}_{n,m,k}$ has led many researchers to consider the more approachable, but considerably less natural, **planted distribution** introduced by Kučera [20] and denoted throughout by $\mathcal{G}^{\text{plant}}_{n,m,k}$. In this context we can selectively mention [4, 6, 8, 11, 19]. In the planted distribution, one first fixes some k-coloring, and then picks uniformly at random m edges that respect this coloring. Due to the "constructive" definition of $\mathcal{G}^{\text{plant}}_{n,m,k}$, the techniques developed in the study of $\mathcal{G}_{n,m}$ can be applied to $\mathcal{G}^{\text{plant}}_{n,m,k}$ immediately, whence the model is rather well understood [4].

Of course the $\mathcal{G}^{\text{plant}}_{n,m,k}$ model is somewhat artificial and therefore provides a less natural model of random instances than $\mathcal{G}^{\text{uniform}}_{n,m,k}$. Nevertheless, devising new ideas for analyzing $\mathcal{G}^{\text{uniform}}_{n,m,k}$, in this paper we show that $\mathcal{G}^{\text{uniform}}_{n,m,k}$ and $\mathcal{G}^{\text{plant}}_{n,m,k}$ actually share many structural graph properties such as the existence of a single cluster of solutions. As a consequence, we can prove that a certain algorithm, designed with $\mathcal{G}^{\text{plant}}_{n,m,k}$ in mind, works for $\mathcal{G}^{\text{uniform}}_{n,m,k}$ as well. In other words, presenting new methods for analyzing heuristics on random graphs, we can show that algorithmic techniques invented for the somewhat artificial $\mathcal{G}^{\text{plant}}_{n,m,k}$ model extend to the canonical $\mathcal{G}^{\text{uniform}}_{n,m,k}$ model.

To obtain these results, we use two main techniques. As we mentioned, $\mathcal{G}^{\text{plant}}_{n,m,k}$ (and the analogous $\mathcal{G}^{\text{plant}}_{n,p,k}$ in which every edge respecting the planted k-coloring is included with probability p) is already very well understood, and the probability of some graph properties that we discuss can be easily estimated for $\mathcal{G}^{\text{plant}}_{n,m,k}$ using standard probabilistic calculations. It then remains to find a reasonable "exchange rate" between $\mathcal{G}^{\text{plant}}_{n,m,k}$ and $\mathcal{G}^{\text{uniform}}_{n,m,k}$. We use this approach to estimate the probability of "complicated" graph properties, which hold with extremely high probability in $\mathcal{G}^{\text{plant}}_{n,m,k}$. The other method is to directly analyze $\mathcal{G}^{\text{uniform}}_{n,p,k}$, crucially overcoming the edge-dependency issues. This method tends to be more complicated than the first one, and involves intricate counting arguments.

1.3 Related Work

As mentioned above, the k-colorability problem exhibits a sharp threshold phenomenon, in the sense that there exists a function $d_k(n)$ s.t. a random graph from $\mathcal{G}_{n,m}$ is *whp* k-colorable if $2m/n < (1-\varepsilon)d_k(n)$ and is *whp* not k-colorable if $2m/n > (1+\varepsilon)d_k(n)$ (cf. [1]). For example, it is known that $d_3(n) \geq 4.03n$ [3] and $d_3(n) \leq 5.044n$ [2]. Therefore, a typical graph in $\mathcal{G}_{n,m}$ with $m = cn$ will not be k-colorable (when thinking of k as a fixed integer, say $k = 3$, and allowing the average degree c to be an arbitrary constant, say $c = 100$, or even a growing function of n). Therefore, when considering relatively dense random graphs, one should take care when defining the underlying distribution, e.g. consider $\mathcal{G}_{n,m,k}^{\text{plant}}$ or $\mathcal{G}_{n,m,k}^{\text{uniform}}$.

Almost all polynomial-time graph-coloring heuristics suggested so far for finding a proper k-coloring of the input graph (or return a failure), were analyzed when the input is sampled according to $\mathcal{G}_{n,p,k}^{\text{plant}}$, or various semi-random variants thereof (and similarly for other graph problems such as clique, independent set, and random satisfiability problems). Alon and Kahale [4] suggest a polynomial time algorithm, based on spectral techniques, that *whp* properly k-colors a random graph from $\mathcal{G}_{n,p,k}^{\text{plant}}$, $np \geq C_0k^2$, C_0 a sufficiently large constant. Combining techniques from [4] and [11], Böttcher [8] suggests an expected polynomial time algorithm for $\mathcal{G}_{n,p,k}^{\text{plant}}$ based on SDP (semi-definite programming) for the same p values. Much work was done also on semi-random variants of $\mathcal{G}_{n,p,k}^{\text{plant}}$, e.g. [6, 11, 14, 19].

On the other hand, very little work has been done on non-planted k-colorable graph distributions, such as $\mathcal{G}_{n,m,k}^{\text{uniform}}$. In this context one can mention the work of Prömel and Steger [23] who analyze $\mathcal{G}_{n,m,k}^{\text{uniform}}$ but with a parametrization which causes $\mathcal{G}_{n,m,k}^{\text{uniform}}$ to collapse to $\mathcal{G}_{n,m,k}^{\text{plant}}$, thus not shedding light on the setting of interest in this work. Similarly, Dyer and Frieze [12] deal with very dense graphs (of average degree $\Omega(n)$).

1.4 Paper's Structure

The rest of the paper is structured as follows. In Section 2 we present the algorithm Color that is used to prove Theorem 2. In Section 3 we discuss some properties that a typical graph in $\mathcal{G}_{n,m,k}^{\text{uniform}}$ possesses. Using these properties we then prove Theorem 1 in Section 4, and prove that the algorithm Color indeed meets the requirements of Theorem 2. Due to lack of space, most propositions are given without a proof, which can be found in complete in the journal version of this paper.

2 The Coloring Algorithm

In Section 3 we prove that a typical graph in $\mathcal{G}_{n,m,k}^{\text{uniform}}$ and in $\mathcal{G}_{n,m,k}^{\text{plant}}$ share many structural properties such as the existence of a single cluster of solutions. In effect, it will turn out that coloring heuristics that prove efficient for $\mathcal{G}_{n,m,k}^{\text{plant}}$ (e.g. [4, 11]) are useful in the uniform setting as well. Therefore, our coloring algorithm builds on ideas from [4] and [11].

When describing the algorithm we have a sparse graph in mind, namely $m/n = c$, c a constant satisfying $c \geq C_0 k^6$ (in the denser setting where $m/n = \omega(1)$, matters actually get much simpler). For simplicity of exposition (to avoid the cumbersome floor and ceiling brackets) we assume that k divides n. The algorithm proceeds in several phases. First, using the Semi-Definite Programming ("SDP")-based subroutine SDPColor, a k-coloring of the vertices is obtained. This coloring may not be proper, but *whp* differs from a proper k-coloring on the colors of at most, say, $n/(200k)$ vertices. Next, this coloring is refined using an iterative recoloring procedure, after which the obtained coloring differs on the colors of at most $e^{-\Theta(m/n)}n$ vertices from some proper k-coloring. The next step is to obtain a partial but *correct* k-coloring of the graph (correct in the sense that the coloring can be completed to a proper k-coloring of the entire graph). This is done using a careful uncoloring procedure, in which the color of "suspicious" vertices is removed. Finally, the graph induced by the uncolored vertices is sparse enough so that *whp* the largest connected component in it is of at most logarithmic size. Therefore, one can simply use exhaustive search, separately in every connected component, to extract the k-coloring of the remaining vertices. Steps 2–5 are similar to the work in [4] on $\mathcal{G}^{\text{plant}}_{n,p,k}$, Step 1 is inspired by [8, 11].

Color(G, k):
step 1: first approximation.
1. SDPColor(G, k).

step 2: recoloring procedure.
2. for $i = 1$ to $\log n$ do:
   2.a for all $v \in V$ simultaneously color v with the least popular color in $N_G(v)$.

step 3: uncoloring procedure.
3. while $\exists v \in V$ with <3 neighbors colored in some other color do:
   3.a uncolor v.

step 4: Exhaustive Search.
4. let $U \subseteq V$ be the set of uncolored vertices.
5. consider the graph $G[U]$.
   5.a if $\exists$ a connected component of size at least $\log n$ - fail.
   5.b otherwise, exhaustively color $G[U]$ according to $V \setminus U$.

We proceed by discussing the subroutine SDPColor in detail. The procedure is based on a SDP relaxation of the max k-cut problem ("partition the vertices in a given graph into k classes so as to maximize the number of edges that join vertices in different classes") suggested by Frieze and Jerrum [17]. For a graph $G = (V, E)$, $SDP_k(G)$ is defined as follows (here $\langle x, y\rangle$ stands for the scalar product of two vectors $x, y \in \mathbb{R}^{|V|}$):

$$SDP_k(G) = \max \sum_{(u,v)\in E} \frac{k-1}{k}(1 - \langle x_u, x_v\rangle) \quad \text{s.t. } \forall\, u, v \in V,\ \langle x_u, x_v\rangle \geq -\frac{1}{k-1},$$

where the max is taken over all families $(x_v)_{v\in V}$ of unit vectors in $\mathbb{R}^{|V|}$ (the vector x_v corresponds to the vertex v). Since SDP_k is a semi-definite program, its optimal value can be computed up to an arbitrary high precision $\varepsilon > 0$, in time polynomial in $|V|, k, \log\frac{1}{\varepsilon}$ (e.g. using the Ellipsoid algorithm [18]).

To get some intuition for the usefulness of SDP_k in the context of the coloring problem, consider the same objective function as $SDP_k(G)$ only restrict the x_v's to be one of $\{a_1, a_2, ..., a_k\}$, where a_i is the vector connecting the centroid of a simplex in $\mathbb{R}^{k-1}$ to its i'th vertex (scaled to be of length 1). It is not hard to see that for $i \neq j$, $\langle a_i, a_j\rangle = -\frac{1}{k-1}$, and that $\frac{k-1}{k}(1 - \langle a_i, a_j\rangle)$ is 1 if $i \neq j$ and 0 otherwise. Furthermore, if the graph is k-colorable, then $SDP_k = |E(G)|$, and therefore the assignment of the a_i's must imply the k color classes of some proper k-coloring (all vertices receiving the same a_i are placed in the same color class).

Thus, grouping vertices into color classes according to the distances between the vectors assigned to them by an optimal solution to $SDP_k(G)$, seems like a good heuristic to get a fair approximation of some proper k-coloring. This is done by the following procedure.

SDPColor(G, k):

1. `solve` $SDP_k(G)$, `and let` $(x_v)_{v\in V(G)}$ `be the optimal solution.`
2. `for all choices of` k `distinct vectors` $x_1^*, x_2^*, ..., x_k^* \in (x_v)_{v\in V(G)}$ `do:`
   2.a `for every` $i \in [1..k]$ `compute` $S_{x_i^*} = \{w \in V : \langle x_i^*, x_w\rangle \geq 0.99\}$.
   2.b `if for every` i, $|S_{x_i^*}| \geq n/k - n/(400k^2)$ `then:`
      2.b.1 `for every` i, `color` $S_{x_i^*}$ `in color` i `(break ties arbitrarily).`
      2.b.2 `color uncolored vertices in color 1.`
      2.b.3 `return the resulting coloring.`
3. `return failure`

3 Properties of a random instance from $\mathcal{G}_{n,m,k}^{\text{uniform}}$

In this section we analyze the structure of a typical graph in $\mathcal{G}_{n,m,k}^{\text{uniform}}$.

3.1 Balancedly k-colorable graphs

We say that a graph G is ε-**balanced** if it admits a proper k-coloring in which every color class is of size $(1\pm\varepsilon)\frac{n}{k}$. We say that a graph is **balancedly** k-colorable if it is 0-balanced.

In the common definition of $\mathcal{G}_{n,m,k}^{\text{plant}}$, all color classes of the planted k-coloring are of the same cardinality, namely n/k. Therefore, all graphs in $\mathcal{G}_{n,m,k}^{\text{plant}}$ have at least one balanced k-coloring (the planted one). Similarly, for the uniform case:

Proposition 1. *Let $m \geq (10k)^4$, then whp a random graph in $\mathcal{G}_{n,m,k}^{\text{uniform}}$ is 0.01-balanced.*

Therefore in order to prove Theorems 1 and 2 , we may just as well confine our discussion to 0.01-balanced k-colorable graphs. To simplify the presentation we will analyze the case $\varepsilon = 0$, namely $\mathcal{G}_{n,m,k}^{\text{uniform}}$ restricted to balancedly k-colorable graphs. Nevertheless, the result easily extends to any $\varepsilon \leq 0.01$ – details omitted. Somewhat abusing notation, from now on we use $\mathcal{G}^{\text{uniform}}_{n,m,k}$ to denote $\mathcal{G}^{\text{uniform}}_{n,m,k}$ restricted to *balancedly* k-colorable graphs. Propositions of similar flavor to Proposition 1 were proven in similar contexts, e.g. [23], and involve rather simple counting arguments.

3.2 Setting the exchange rate

Let $\mathcal{A}$ be some graph property (it would be convenient for the reader to think of $\mathcal{A}$ as a "bad" property). We start by determining the exchange rate for $Pr[\mathcal{A}]$ between the different distributions.

Notation. For a graph property $\mathcal{A}$ we use the following notation to denote the probability of $\mathcal{A}$ under the various distributions: $Pr^{\text{uniform,m}}[\mathcal{A}]$ denotes the probability of property $\mathcal{A}$ occurring under $\mathcal{G}^{\text{uniform}}_{n,m,k}$, $Pr^{\text{planted,m}}[\mathcal{A}]$ for $\mathcal{G}^{\text{plant}}_{n,m,k}$, and $Pr^{\text{planted,n,p}}[\mathcal{A}]$ for $\mathcal{G}^{\text{plant}}_{n,p,k}$. We shall be mostly interested in the case $m = \binom{k}{2}\left(\frac{n}{k}\right)^2 p$, namely m is the expected number of edges in $\mathcal{G}^{\text{plant}}_{n,p,k}$. The following lemma, which is proved using rather standard probabilistic calculations, establishes the exchange rate for $\mathcal{G}^{\text{plant}}_{n,p,k} \to \mathcal{G}^{\text{plant}}_{n,m,k}$.

Lemma 1. *($\mathcal{G}^{\text{plant}}_{n,p,k} \to \mathcal{G}^{\text{plant}}_{n,m,k}$) Let $\mathcal{A}$ be some graph property, then if $m = \binom{k}{2}\left(\frac{n}{k}\right)^2 p$ it holds that*

$$Pr^{\text{planted,m}}[\mathcal{A}] \leq O(\sqrt{m}) \cdot Pr^{\text{planted,n,p}}[\mathcal{A}]$$

Next, we establish the exchange rate $\mathcal{G}^{\text{plant}}_{n,m,k} \to \mathcal{G}^{\text{uniform}}_{n,m,k}$, which is rather involved technically and whose proof embeds interesting results of their own – for example, bounding the expected number of proper k-colorings of a graph in $\mathcal{G}^{\text{uniform}}_{n,m,k}$.

Lemma 2. *($\mathcal{G}^{\text{plant}}_{n,m,k} \to \mathcal{G}^{\text{uniform}}_{n,m,k}$) Let $\mathcal{A}$ be some graph property, then*

$$Pr^{\text{uniform,m}}[\mathcal{A}] \leq e^{ke^{-m/(6nk^3)}n} \cdot Pr^{\text{planted,m}}[\mathcal{A}]$$

3.3 Coloring using SDP

In this section we analyze the behavior of $SDP_k(G)$, where G is sampled according to $\mathcal{G}^{\text{uniform}}_{n,m,k}$. We start by analyzing SDP_k on $\mathcal{G}^{\text{plant}}_{n,p,k}$, then use the discussion in Section 3.2 to obtain basically the same behavior for $\mathcal{G}^{\text{uniform}}_{n,m,k}$. The following lemma appears in [8].

Lemma 3. *Let G be a random graph sampled according to $\mathcal{G}^{\text{plant}}_{n,p,k}$ with φ its planted k-coloring, $np \geq C_0k^6$, C_0 a sufficiently large constant. Then with probability $(1 - e^{-n})$ SDPColor(G,k) obtains a k-coloring which differs from φ on the colors of at most $n/(200k)$ vertices.*

Proposition 2. *Let G be a random graph in $\mathcal{G}^{\text{uniform}}_{n,m,k}$, $m \geq C_0 k^6 n$, C_0 a sufficiently large constant. Then whp there exists a proper balanced k-coloring φ of G s.t. SDPColor(G,k) obtains a k-coloring which differs from φ on the colors of at most $n/(200k)$ vertices.*

Proof. Let $\mathcal{A}$ be "there exists no balanced k-coloring s.t. SDPColor(G,k) obtains a k-coloring which differs from it on the colors of at most $n/(200k)$ vertices", and set $m = \binom{k}{2}\left(\frac{n}{k}\right)^2 p$. Using the "exchange rate" technique:

$$Pr^{uniform,m}[\mathcal{A}] \underbrace{\leq}_{\text{Lemma 2}} e^{ke^{-m/(6nk^3)}n} \cdot Pr^{\text{planted,m}}[\mathcal{A}] \underbrace{\leq}_{\text{Lemma 1}}$$
$$\sqrt{m} \cdot e^{ke^{-m/(6nk^3)}n} \cdot Pr^{planted,n,p}[\mathcal{A}] \underbrace{\leq}_{\text{Lemma 3}} O(\sqrt{m}) \cdot e^{ke^{-m/(6nk^3)}n} \cdot e^{-n} = o(1).$$

The last inequality is due to $m \geq C_0 k^6 n/2$.

3.4 Dense Subgraphs

A random graph in $\mathcal{G}^{\text{plant}}_{n,m,k}$ (also in $\mathcal{G}_{n,m}$) *whp* will not contain a small yet unexpectedly dense subgraph. This property holds only with probability $1-1/poly(n)$, and therefore the "exchange rate" technique, implemented in Section 3.3 for example, is of no use in this case. Overcoming the edge-dependency issue, using an intricate counting argument, we directly analyze $\mathcal{G}^{\text{uniform}}_{n,p,k}$ to prove:

Proposition 3. *Let G be a random graph in $\mathcal{G}^{\text{uniform}}_{n,m,k}$, $m \geq C_0 k^2 n$, C_0 a sufficiently large constant. Then whp there exists* no *subgraph of G containing at most $n/(100k)$ vertices whose average degree is at least $m/(6nk)$.*

3.5 The Core Vertices

We describe a subset of the vertices, referred to as the *core vertices*, which plays a crucial role in the analysis of the algorithm and in the understanding of $\mathcal{G}^{\text{uniform}}_{n,m,k}$. Recall that a set of vertices is said to be frozen in G if in every proper k-coloring of G, all vertices of that set receive the same color. A vertex v is said to be frozen if it belongs to a frozen set. The notion of core captures this phenomenon. In addition, a core typically contains all but a small (though constant) fraction of the vertices. This implies that a large fraction of the vertices is frozen, a fact which must leave imprints on the structure of the graph. These imprints allow efficient heuristics to recover the k-coloring of the core. A second implication of this, is an upper bound on the number of possible k-colorings, and on the distance between every such two (namely, a catheterization of the cluster structure of the solution space).

There are several ways to define a core, we choose a constructive way. $\mathcal{H} = \mathcal{H}(\varphi, t)$ is defined using the following iterative procedure. Set $\mathcal{H} = V$, and remove all vertices v s.t. v has less than $(1-1/200)t$ neighbors in some color class other than $\varphi(v)$. Then, iteratively, while there exists a vertex v in $\mathcal{H}$, s.t. v has more than $t/200$ neighbors outside $\mathcal{H}$, remove v.

Proposition 4. *Let G be a random graph in $\mathcal{G}^{\text{uniform}}_{n,m,k}$, $m \geq C_0 k^4 n$, C_0 a sufficiently large constant, and set $t = (1-1/k)2m/n$. Then whp there exists a proper k-coloring φ s.t. $\mathcal{H} = \mathcal{H}(\varphi, t)$ enjoys the following properties (by $V_1, V_2 \ldots V_k$ we denote φ's k color classes):*

1. *$|\mathcal{H}| \geq (1 - e^{-m/(20nk^3)})n$.*
2. *Every $v \in \mathcal{H} \cap V_i$ has the property that $e(v, \mathcal{H} \cap V_j) \geq 99t/100$ for all $j \in \{1, \ldots, k\} \setminus \{i\}$.*
3. *For all $v \in \mathcal{H}$ we have $e(v, V \setminus \mathcal{H}) \leq t/200$.*
4. *The graph induced by the vertices of $\mathcal{H}$ is uniquely k-colorable.*

Observe that t is chosen to be the expected degree of a vertex $v \in V_i$ in color class V_j, $j \neq i$. Properties 2 and 3 follow immediately from the construction of $\mathcal{H}$. To obtain property 1 we first establish the following fact, which appears in [8] (with a complete proof).

Lemma 4. *Let G be a graph sampled according to $\mathcal{G}^{\text{plant}}_{n,p,k}$, $np \geq C_0 k^4 n$, C_0 a sufficiently large constant, and let φ be its planted k-coloring. Then $Pr[|\mathcal{H}(\varphi, np/k)| \leq (1 - e^{-np/(40k^3)})n] \leq e^{-e^{-np/(40k^3)}n}$.*

Using the "exchange rate" technique we obtain:

Proposition 5. *Let G be a random graph in $\mathcal{G}^{\text{uniform}}_{n,m,k}$, $m \geq C_0 k^4 n$, C_0 a sufficiently large constant, and set $t = (1 - 1/k)2m/n$. Then whp there exists a proper k-coloring φ of G s.t. $|\mathcal{H}(\varphi, t)| \geq (1 - e^{-m/(20nk^3)})n$.*

Lastly, we establish the frozenness property (property 4).

Proposition 6. *Let G be a graph for which Proposition 3 holds. Then every core satisfying Properties 1 and 2 in Proposition 4 is uniquely k-colorable.*

The next proposition ties between the core vertices and the approximation ratio of SDPColor, and is crucial to the analysis of the algorithm. The proposition follows by noticing that φ in Lemmas 3 and 4 is the same – the planted coloring, and then using the "exchange rate" technique on the combined property.

Proposition 7. *Let G be a random graph in $\mathcal{G}^{\text{uniform}}_{n,m,k}$, $m \geq C_0 k^6 n$, C_0 a sufficiently large constant. Then whp there exists a proper k-coloring φ of G s.t. the coloring returned by SDPColor(G,k) differs from φ on the colors of at most $n/(200k)$ vertices, and there exists a core $\mathcal{H} = \mathcal{H}(\varphi, t)$ s.t. Proposition 4 holds for $\mathcal{H}$, where $t = (1 - 1/k)2m/n$ as in Proposition 4.*

The next proposition characterizes the structure of the graph induced by the non-core vertices.

Proposition 8. *Let G be a random graph in $\mathcal{G}^{\text{uniform}}_{n,m,k}$, $m \geq C_0 k^2 n$, C_0 a sufficiently large constant. Let $G[V \setminus \mathcal{H}]$ be the graph induced by the non-core vertices. Then whp the largest connected component in $G[V \setminus \mathcal{H}]$ is of size $O(\log n)$.*

This fact is proven in [4] for $\mathcal{G}^{\text{plant}}_{n,m,k}$, however it holds w.p. $1 - 1/poly(n)$. Therefore the "exchange rate" technique is of no use. Thus, in the uniform case the analysis is much more involved due to dependency issues (an intricate counting argument). Full details are in the journal version.

4 Proofs of Theorems 1 and 2

Theorem 1 is an immediate corollary of Proposition 4, as it implies that all but $e^{-m/(20nk^3)}n$ of the vertices are uniquely colorable, and in particular are frozen. There are at most $k^{e^{-m/(20nk^3)}n} = \exp\{ne^{-\Theta(m/n)}\}$ possible ways to set the colors of the non-frozen vertices. Proposition 8 characterizes the graph induced by the non-core (which contain the non-frozen) vertices.

To prove Theorem 2, we prove that the algorithm Color meets the requirements of Theorem 2. In particular, we prove that if G is *typical* (in the sense that the properties discussed in Sections 3.3, 3.4, and 3.5 hold for it), then Color k-colors it properly in polynomial time. Since G is typical *whp* (the discussion in Section 3), Theorem 2 follows. The proofs of the following propositions can be found in [4], and are based on the discussion in Section 3 (while a similar discussion exists in [4] for the planted setting). In the following propositions we assume G is typical.

Proposition 9. *After the recoloring step ends, the core vertices $\mathcal{H}$ are colored according to the proper k-coloring promised in Proposition 7 – let φ denote this coloring.*

Proposition 10. *Assuming Proposition 9 holds, $\mathcal{H}$ survives the uncoloring step, and every vertex that survives the uncoloring step is colored according to φ.*

Proposition 11. *Assuming Proposition 10 holds, the exhaustive search completes in polynomial time with a legal k-coloring of the graph.*

5 Discussion

In this paper we explore the uniform distribution over k-colorable graphs with cn edges, c greater than some constant. We obtain a rather comprehensive understanding of the structure of the space of proper k-colorings of a typical graph in it, and describe a polynomial time algorithm that properly k-colors most such graphs.

The techniques of this paper apply to a number of further NP-hard problems, including random instances of k-SAT. More precisely, we can show that a uniformly distributed **satisfiable** k-SAT formula with sufficiently large, yet constant, clause-variable ratio (above the satisfiability threshold) typically exhibits a single cluster of exponentially many satisfying assignments. Our result implies that the algorithmic techniques developed for the planted k-SAT distribution [16, 15] extend to the significantly more natural uniform distribution, thus improving Chen's [10] *exponential* time algorithm for the same problem. In addition, our result answers questions posed in [5]. Full details will appear in a separate paper.

Acknowledgements: we thank Uriel Feige for many useful discussions. Part of this work was done while the third author was visiting Humboldt University.

References

1. D. Achlioptas and E. Friedgut. A sharp threshold for k-colorability. *Random Struct. Algorithms*, 14(1):63–70, 1999.
2. D. Achlioptas and M. Molloy. Almost all graphs with 2.522n edges are not 3-colorable. *Elec. Jour. Of Comb.*, 6(1), R29, 1999.
3. D. Achlioptas and C. Moore. Almost all graphs with average degree 4 are 3-colorable. In *STOC '02*, pages 199–208, 2002.
4. N. Alon and N. Kahale. A spectral technique for coloring random 3-colorable graphs. *SIAM J. on Comput.*, 26(6):1733–1748, 1997.
5. E. Ben-Sasson, Y. Bilu, and D. Gutfreund. Finding a randomly planted assignment in a random $3CNF$. *manuscript*, 2002.
6. A. Blum and J. Spencer. Coloring random and semi-random k-colorable graphs. *J. of Algorithms*, 19(2):204–234, 1995.
7. B. Bollobás. The chromatic number of random graphs. *Combin.*, 8(1):49–55, 1988.
8. J. Böttcher. Coloring sparse random k-colorable graphs in polynomial expected time. In *Proc. 30th MFCS*, pages 156–167, 2005.
9. A. Braunstein, M. Mézard, M. Weigt, and R. Zecchina. Constraint satisfaction by survey propagation. *Computational Complexity and Statistical Physics*, 2005.
10. H. Chen. An algorithm for sat above the threshold. In *6th International Conference on Theory and Applications of Satisfiability Testing*, pages 14–24, 2003.
11. A. Coja-Oghlan. Coloring semirandom graphs optimally. In *Proc. 31st ICALP*, pages 383–395, 2004.
12. M. E. Dyer and A. M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *J. Algorithms*, 10(4):451–489, 1989.
13. U. Feige and J. Kilian. Zero knowledge and the chromatic number. *J. Comput. and Syst. Sci.*, 57(2):187–199, 1998.
14. U. Feige and J. Kilian. Heuristics for semirandom graph problems. *J. Comput. and Syst. Sci.*, 63(4):639–671, 2001.
15. U. Feige, E. Mossel, and D. Vilenchik. Complete convergence of message passing algorithms for some satisfiability problems. In *RANDOM*, pages 339–350, 2006.
16. A. Flaxman. A spectral technique for random satisfiable 3CNF formulas. In *Proc. 14th ACM-SIAM Symp. on Discrete Algorithms*, pages 357–363, 2003.
17. A. Frieze and M. Jerrum. Improved approximation algorithms for MAX k-CUT and MAX BISECTION. *Algorithmica*, 18(1):67–81, 1997.
18. M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, second edition, 1993.
19. M. Krivelevich and D. Vilenchik. Semirandom models as benchmarks for coloring algorithms. In *ANALCO*, pages 211–221, 2006.
20. L. Kučera. Expected behavior of graph coloring algorithms. In *Proc. Fundamentals of Computation Theory*, volume 56 of *Lecture Notes in Comput. Sci.*, pages 447–451. Springer, Berlin, 1977.
21. T. Łuczak. The chromatic number of random graphs. *Combin.*, 11(1):45–54, 1991.
22. R. Mulet, A. Pagnani, M. Weigt, and R. Zecchina. Coloring random graphs. *Phys. Rev. Lett.*, 89(26):268701, 2002.
23. H. Prömel and A. Steger. Random l-colorable graphs. *Random Structures and Algorithms*, 6:21–37, 1995.
24. J. S. Turner. Almost all k-colorable graphs are easy to color. *J. Algorithms*, 9(1):63–82, 1988.