The Condensation Phase Transition in Random Graph Coloring*

Victor Bapst[1], Amin Coja-Oghlan[1], Samuel Hetterich[1], Felicia Raßmann[1], and Dan Vilenchik[2]

1 **Mathematics Institute, Goethe University**
10 Robert Mayer St, Frankfurt 60325, Germany
{bapst, acoghlan, hetterich, rassmann}@math.uni-frankfurt.de

2 **Faculty of Mathematics & Computer Science, The Weizmann Institute**
Rehovot, Israel
dan.vilenchik@weizmann.ac.il

Abstract

Based on a non-rigorous formalism called the "cavity method", physicists have made intriguing predictions on phase transitions in discrete structures. One of the most remarkable ones is that in problems such as random k-SAT or random graph k-coloring, very shortly before the threshold for the existence of solutions there occurs another phase transition called *condensation* [Krzakala et al., PNAS 2007]. The existence of this phase transition seems to be intimately related to the difficulty of proving precise results on, e. g., the k-colorability threshold as well as to the performance of message passing algorithms. In random graph k-coloring, there is a precise conjecture as to the location of the condensation phase transition in terms of a distributional fixed point problem. In this paper we prove this conjecture, provided that k exceeds a certain constant k_0.

1998 ACM Subject Classification G.2.1 Combinatorics, G.2.2 Graph Theory

Keywords and phrases random graphs, graph coloring, phase transitions, message-passing algorithm

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2014.449

1 Introduction

Let $G(n,p)$ denote the random graph on the vertex set $V = \{1,\dots,n\}$ obtained by connecting any two vertices with probability $p \in [0,1]$ independently. Throughout the paper, we are concerned with the setting that $p = d/n$ for a number $d > 0$ that remains fixed as $n \to \infty$. We say that $G(n,d/n)$ has a property with high probability *('w.h.p.') if its probability converges to 1 as $n \to \infty$.*

The study of random constraint satisfaction problems started with experimental work in the 1990s, which led to two hypotheses [5, 23]. First, that in problems such as random k-SAT or random graph coloring there is a *satisfiability threshold*, i. e., a critical "constraint density" below which the instance admits a solution and above which it does not w.h.p. Second, that this threshold is associated with the algorithmic "difficulty" of actually computing a solution, where "difficulty" has been quantified in various ways, albeit not in the formal sense

* The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 278857–PTCC.

17th Int'l Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'14) /
18th Int'l Workshop on Randomization and Computation (RANDOM'14).
Editors: Klaus Jansen, José Rolim, Nikhil Devanur, and Cristopher Moore; pp. 449–464

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of computational complexity. These findings have led to a belief that random instances of k-SAT or graph k-colorability near the threshold for the existence of solutions are challenging algorithmic benchmarks, at the very least.

These two hypotheses have inspired theoretical work. Short of establishing the existence of an actual satisfiability threshold, Friedgut [13] and Achlioptas and Friedgut [1] proved that in random k-SAT and random graph k-coloring there exists a *sharp threshold sequence*. For instance, in the graph k-coloring problem, this is a sequence $d_{k-\mathrm{col}}(n)$ that marks the point where the probability of being k-colorable drops from 1 to 0.[1] The dependence on n allows for the possibility that this point might vary with the number of vertices, although this is broadly conjectured not to be the case. In fact, proving that $(d_{k-\mathrm{col}}(n))_{n\geq 1}$ converges to a single number $d_{k-\mathrm{col}}$ is a well-known open problem. So is determining the location of $d_{k-\mathrm{col}}(n)$ (or its limit), as [1] is a pure existence result.

In addition, inspired by predictions from statistical physics, the geometry of the set of solutions of random k-SAT or k-colorability instances has been investigated [2, 27]. The result is that at a certain point well before the satisfiability threshold the set of solutions shatters into a multitude of well-separated "clusters". Inside a typical cluster, all solutions agree on most of the variables/vertices, the so-called "frozen" ones. The average degree d at which these "frozen clusters" arise (roughly) matches the point up to which efficient algorithms provably find solutions.[2] Hence, on the one hand it is tempting to think that there is a connection between clustering and the computational "difficulty" of finding a solution [2, 27, 30]. On the other hand, physicists have suggested new *message passing algorithms* specifically to cope with a clustered geometry [4, 26]. A satisfactory analysis of these algorithms remains elusive.

Remarkably, the physics predictions are not merely circumstantial or experimental findings. They derive from a non-rigorous but systematic formalism called the *cavity method* [25]. This technique yields, among other things, a prediction as to the precise location of the k-SAT or k-colorability threshold. But perhaps even more remarkably, according to the cavity method shortly before the threshold for the existence of solutions there occurs another phase transition called *condensation* [20]. This phase transition marks a further change in the geometry of the solution space. While prior to the condensation phase transition each cluster contains only an exponentially small fraction of all solutions, thereafter a sub-exponential number of clusters contain a constant fraction of the entire set of solutions. As we will see in Section 3 below, condensation seems to hold the key to a variety of problems, including that of finding the k-colorability threshold and of analyzing message passing algorithms rigorously. More generally, the physicists' cavity method is extremely versatile. It has been used to put forward tantalizing conjectures in a variety of areas, including coding theory, probabilistic combinatorics, unsurprisingly, mathematical physics (see [25] for an overview) or, more recently, compressed sensing [19]. Hence the importance of providing a rigorous foundation for this technique.

2 Results

In this paper we prove that, indeed, a condensation phase transition occurs in random graph coloring, and that it occurs at the *precise* location predicted by the cavity method. This is

[1] Formally, for any $k \geq 3$ there is a sequence $(d_{k-\mathrm{col}}(n))_n$ such that for any fixed $\varepsilon > 0$, $G(n,p)$ is k-colorable w.h.p. if $p < (1-\varepsilon)d_{k-\mathrm{col}}(n)/n$, while $G(n,p)$ fails to be k-colorable w.h.p. if $p > (1+\varepsilon)d_{k-\mathrm{col}}(n)/n$.

[2] Actually the appearance of clusters does not quite match the appearance of frozen variables/vertices. For a more detailed explanation on the connection between clusters, frozen variables and computational hardness see [18, 21].

the first rigorous result to determine the exact location of the condensation transition in a model of this kind. Additionally, the proof yields a direct combinatorial explanation of how this phase transition comes about.

2.1 Catching a Sharp Threshold

To state the result, let us denote by $Z_k(G)$ the number of k-colorings of a graph G. We would like to study the "typical value" of $Z_k(G(n, d/n))$ in the limit as $n \to \infty$. As it turns out, the correct scaling of this quantity (to obtain a finite limit) is[3]

$$\Phi_k(d) \equiv \lim_{n\to\infty} \mathbb{E}[Z_k(G(n, d/n))^{1/n}].$$

In physics terminology, a "phase transition" is a point d_0 where the function $d \mapsto \Phi_k(d)$ is non-analytic. However, the limit $\Phi_k(d)$ is not currently known to exists for all d, k.[4] Hence, we need to tread carefully. For a given $k \geq 3$ we call $d_0 \in (0, \infty)$ *smooth* if there exists $\varepsilon > 0$ such that

- for any $d \in (d_0 - \varepsilon, d_0 + \varepsilon)$ the limit $\Phi_k(d)$ exists, and
- the map $d \in (d_0 - \varepsilon, d_0 + \varepsilon) \mapsto \Phi_k(d)$ has an expansion as an absolutely convergent power series around d_0.

If d_0 fails to be smooth, we say that a *phase transition* occurs at d_0.

For a smooth d_0 the sequence of random variables $(Z_k(G(n, d_0/n))^{1/n})_n$ converges to $\Phi_k(d_0)$ in probability. This follows from a concentration result for the number of k-colorings from [2]. Hence, $\Phi_k(d)$ really captures the "typical" value of $Z_k(G(n, d/n)$ (up to a factor of $\exp(o(n))$).

The above notion of "phase transition" is in line with the intuition held in combinatorics. For instance, the classical result of Erdős and Rényi [11] implies that the function that maps d to the limit as $n \to \infty$ of the expected fraction of vertices that belong to the largest component of $G(n, d/n)$ is non-analytic at $d = 1$. Similarly, if there actually is a sharp threshold $d_{k-\mathrm{col}}$ for k-colorability, then $d_{k-\mathrm{col}}$ is easily seen to be a phase transition in the above sense.[5]

As a next step, we state (an equivalent but slightly streamlined version of) the physics prediction from [22] as to the location of the condensation phase transition. As most predictions based on the "cavity method", this one comes in terms of a distributional fixed point problem. To be specific, let Ω be the set of probability measures on the set $[k] = \{1, \ldots, k\}$. We identify Ω with the k-simplex, i. e., the set of maps $\mu : [k] \to [0, 1]$ such that $\sum_{h=1}^{k} \mu(h) = 1$, equipped with the topology and Borel algebra induced by $\mathbb{R}^k$. Moreover, we define a map $\mathcal{B} : \bigcup_{\gamma=1}^{\infty} \Omega^\gamma \to \Omega$, $(\mu_1, \ldots, \mu_\gamma) \mapsto \mathcal{B}[\mu_1, \ldots, \mu_\gamma]$ by letting

$$\mathcal{B}[\mu_1, \ldots, \mu_\gamma](i) = \begin{cases} 1/k & \text{if } \sum_{h\in[k]} \prod_{j=1}^{\gamma} 1 - \mu_j(h) = 0, \\ \frac{\prod_{j=1}^{\gamma} 1-\mu_j(i)}{\sum_{h\in[k]} \prod_{j=1}^{\gamma} 1-\mu_j(h)} & \text{otherwise,} \end{cases} \quad \text{for any } i \in [k]\,. \tag{2.1}$$

[3] In the physics literature, one typically considers $n^{-1} \ln Z$ instead of $Z^{1/n}$, where Z is the so-called "partition function". We work with the nth root because our "partition function" Z_k may be equal to 0.

[4] It seems natural to conjecture that the limit $\Phi_k(d)$ exists for all d, k, but proving this might be difficult.In fact, the existence of the limit for all d, k would imply that $d_{k-\mathrm{col}}(n)$ converges.

[5] For $d < d_{k-\mathrm{col}}$, $G(n, d/n)$ has a k-coloring w.h.p., and thus the number of k-colorings is, in fact, exponentially large in n as there are $\Omega(n)$ isolated vertices w.h.p. Hence, if $\Phi_k(d)$ exists for $d < d_{k-\mathrm{col}}$, then $\Phi_k(d) > 0$. By contrast, for $d > d_{k-\mathrm{col}}$ the random graph $G(n, d/n)$ fails to be k-colorable w.h.p., and therefore $\Phi_k(d) = 0$. Thus, $\Phi_k(d)$ cannot be analytic at $d_{k-\mathrm{col}}$.

$$\begin{aligned}
\phi_{d,k}(\pi) &= \phi^e_{d,k}(\pi) + \frac{1}{k}\sum_{i\in[k]}\sum_{\gamma_1,\ldots,\gamma_k=0}^{\infty}\phi^v_{d,k}(\pi;i;\gamma_1,\ldots,\gamma_k)\prod_{h\in[k]}\left(\frac{d}{k-1}\right)^{\gamma_h}\frac{\exp(-d/(k-1))}{\gamma_h!}, \quad \text{where}\\
\phi^e_{d,k}(\pi) &= -\frac{d}{2k(k-1)}\sum_{h_1=1}^{k}\sum_{h_2\in[k]\setminus\{h_1\}}\int_{\Omega^2}\ln\left[1-\sum_{h\in[k]}\mu_1(h)\mu_2(h)\right]\bigotimes_{i=1}^{2}\mathrm{d}\pi_{h_i}(\mu_i), \qquad (2.4)\\
\phi^v_{d,k}(\pi;i;\gamma_1,\ldots,\gamma_k) &= \begin{cases}\ln k & \text{if } \sum_{i=1}^k\gamma_i=0,\\ \displaystyle\int_{\Omega^{\gamma_1+\cdots+\gamma_k}}\ln\left[\sum_{h=1}^{k}\prod_{h'\in[k]\setminus\{i\}}\prod_{j=1}^{\gamma_{h'}}1-\mu^{(j)}_{h'}(h)\right]\bigotimes_{h'\in[k]}\bigotimes_{j=1}^{\gamma_{h'}}\mathrm{d}\pi_{h'}(\mu^{(j)}_{h'}) & \text{if } \sum_{i=1}^k\gamma_i>0.\end{cases} \qquad (2.5)
\end{aligned}$$

Figure 1 The function $\phi_{d,k}$.

Further, let $\mathcal{P}$ be the set of all probability measures on Ω. For each $\mu\in\Omega$ let $\delta_\mu\in\mathcal{P}$ denote the Dirac measure that puts mass one on the single point μ. In particular, $\delta_{k^{-1}\mathbf{1}}\in\mathcal{P}$ signifies the measure that puts mass one on the uniform distribution $k^{-1}\mathbf{1}=(1/k,\ldots,1/k)$. For $\pi\in\mathcal{P}$ and $\gamma\geq 0$ let

$$Z_\gamma(\pi)=\sum_{h=1}^{k}\left(1-\int_\Omega\mu(h)\mathrm{d}\pi(\mu)\right)^\gamma. \qquad (2.2)$$

Further, define a map $\mathcal{F}_{d,k}:\mathcal{P}\to\mathcal{P}$, $\pi\mapsto\mathcal{F}_{d,k}[\pi]$ by letting

$$\begin{aligned}
\mathcal{F}_{d,k}[\pi] &= \exp(-d)\cdot\delta_{k^{-1}\mathbf{1}}\\
&\quad+\sum_{\gamma=1}^{\infty}\frac{\gamma^d\exp(-d)}{\gamma!\cdot Z_\gamma(\pi)}\int_{\Omega^\gamma}\left[\sum_{h=1}^{k}\prod_{j=1}^{\gamma}1-\mu_j(h)\right]\cdot\delta_{\mathcal{B}[\mu_1,\ldots,\mu_\gamma]}\bigotimes_{j=1}^{\gamma}\mathrm{d}\pi(\mu_j). \qquad (2.3)
\end{aligned}$$

Thus, in (2.3) we integrate a function with values in $\mathcal{P}$, viewed as a subset of the Banach space [6] of signed measures on Ω. The normalising term $Z_\gamma(\pi)$ ensures that $\mathcal{F}_{d,k}[\pi]$ really is a probability measure on Ω.

The main theorem is in terms of a fixed point of the map $\mathcal{F}_{d,k}$, i. e., a point $\pi^*\in\mathcal{P}$ such that $\mathcal{F}_{d,k}[\pi^*]=\pi^*$. In general, the map $\mathcal{F}_{d,k}$ has several fixed points. Hence, we need to single out the correct one. For $h\in[k]$ let $\delta_h\in\Omega$ denote the vector whose hth coordinate is one and whose other coordinates are 0 (i. e., the Dirac measure on h). We call a measure $\pi\in\mathcal{P}$ *frozen* if $\pi(\{\delta_1,\ldots,\delta_k\})\geq 2/3$; in words, the total probability mass concentrated on the k vertices of the simplex Ω is at least $2/3$.

As a final ingredient, we need a function $\phi_{d,k}:\mathcal{P}\to\mathbb{R}$. To streamline the notation, for $\pi\in\mathcal{P}$ and $h\in[k]$ we write π_h for the measure $\mathrm{d}\pi_h(\mu)=k\mu(h)\mathrm{d}\pi(\mu)$. With this notation, $\phi_{d,k}$ is defined in Figure 1. The integrals in (2.4) and (2.5) are well-defined because the set where the argument of the logarithm vanishes has measure zero.

▶ **Theorem 1.** *There exists a constant $k_0\geq 3$ such that for any $k\geq k_0$ the following holds. If $d\geq(2k-1)\ln k-2$, then $\mathcal{F}_{d,k}$ has precisely one frozen fixed point $\pi^*_{d,k}$. Further, the*

[6] To be completely explicit, the probability mass that a measurable set $A\subset\Omega$ carries under $\mathcal{F}_{d,k}[\pi]$ is

$$\mathcal{F}_{d,k}[\pi](A)=\exp(-d)\cdot\mathbf{1}_{\frac{1}{k}\mathbf{1}\in A}+\sum_{\gamma\geq 1}\frac{\gamma^d\exp(-d)}{\gamma!\cdot Z_\gamma(\pi)}\int\left[\sum_{h=1}^{k}\prod_{j=1}^{\gamma}1-\mu_j(h)\right]\cdot\mathbf{1}_{\mathcal{B}[\mu_1,\ldots,\mu_\gamma]\in A}\bigotimes_{j=1}^{\gamma}\mathrm{d}\pi(\mu_j),$$

where $\mathbf{1}_{\nu\in A}=1$ if $\nu\in A$ and $\mathbf{1}_{\nu\in A}=0$ otherwise.

function

$$\Sigma_k : d \mapsto \ln k + \frac{d}{2}\ln(1-1/k) - \phi_{d,k}(\pi^*_{d,k}) \tag{2.6}$$

has a unique zero $d_{k,\mathrm{cond}}$ *in the interval* $[(2k-1)\ln k - 2, (2k-1)\ln k - 1]$. *For this number* $d_{k,\mathrm{cond}}$ *the following three statments hold.*

(i) *Any* $0 < d < d_{k,\mathrm{cond}}$ *is smooth and* $\Phi_k(d) = k \cdot (1-1/k)^{d/2}$.

(ii) *There occurs a phase transition at* $d_{k,\mathrm{cond}}$.

(iii) *If* $d > d_{k,\mathrm{cond}}$, *then*

$$\limsup_{n\to\infty} \mathbb{E}[Z_k(G(n,d/n))^{1/n}] < k \cdot (1-1/k)^{d/2}.$$

Thus, if d *is smooth, then* $\Phi_k(d) < k \cdot (1-1/k)^{d/2}$.

The key strength of Theorem 1 and the main achievement of this work is that we identify the *precise* location of the phase transition. In particular, the result $d_{k,\mathrm{cond}}$ is one number rather than a "sharp threshold sequence" that might vary with n. Admittedly, this precise answer is not exactly a simple one. But that seems unsurprising, given the intricate combinatorics of the random graph coloring problem. That said, the proof of Theorem 1 will illuminate matters. For instance, the fixed point $\pi^*_{d,k}$ turns out to have a nice combinatorial interpretation and, perhaps surprisingly, $\pi^*_{d,k}$ emerges to be a *discrete* probability distribution.

The above formulas are derived systematically via the cavity method [25]. For instance, the functional $\phi_{d,k}$ is a special case of a general formula, the so-called "Bethe free entropy". Moreover, the map $\mathcal{B}$ is the distributional version of the "Belief Propagation" operator. In effect, the predictions as to the condensation phase transitions in other problems look very similar to the above. Consequently, it can be expected that the proof technique developed in the present work carries over to many other problems.

While the main point of Theorem 1 is that it gives an exact answer, it is not difficult to obtain a simple asymptotic expansion of $d_{k,\mathrm{cond}}$ in the limit of large k. Namely, $d_{k,\mathrm{cond}} = (2k-1)\ln k - 2\ln 2 + \varepsilon_k$, where $\varepsilon_k \to 0$ as $k \to \infty$. This asymptotic formula was obtained in [8] by means of a *much* simpler argument than the one developed in the present paper. However, this simpler argument does not quite get to the bottom of the combinatorics behind the condensation phase transition.

2.2 The Cluster Size

The proof of Theorem 1 allows us to formalise the physicists' notion that as d tends to $d_{k,\mathrm{cond}}$, the cluster size approaches the total number of k-colorings. Of course, we need to formalise what we mean by "clusters" first. Thus, let G be a graph on n vertices. If σ, τ are k-colorings of G, we define their *overlap* as the $k \times k$-matrix $\rho(\sigma,\tau) = (\rho_{ij}(\sigma,\tau))_{i,j\in[k]}$ with entries

$$\rho_{ij}(\sigma,\tau) = \frac{|\sigma^{-1}(i) \cap \tau^{-1}(j)|}{n},$$

i. e., $\rho_{ij}(\sigma,\tau)$ is the fraction of vertices colored i under σ and j under τ. Now, define the *cluster* of σ in G as

$$\mathcal{C}(G,\sigma) = \{\tau : \tau \text{ is a } k\text{-coloring of } G \text{ and } \rho_{ii}(\sigma,\tau) \geq 0.51/k \text{ for all } i \in [k]\}. \tag{2.7}$$

Suppose that σ, τ are such that $|\sigma^{-1}(i)|, |\tau^{-1}(i)| \sim n/k$ for all $i \in [k]$; most k-colorings of $G(n,d/n)$ have this property w.h.p. [1, 7]. Then $\tau \in \mathcal{C}(G,\sigma)$ means that a little over 50% of the vertices with color i under σ also have color i under τ.

▶ **Corollary 2.** *With the notation and assumptions of Theorem 1, the function Σ_k is continuous, strictly positive and monotonically decreasing on $((2k-1)\ln k - 2, d_{k,\mathrm{cond}})$, and $\lim_{d\to d_{k,\mathrm{cond}}} \Sigma_k(d) = 0$. Further, given that $G(n,d/n)$ is k-colorable, let $\boldsymbol{\tau}$ be a uniformly random k-coloring of this random graph. Then*

$$\lim_{\varepsilon\searrow 0}\lim_{n\to\infty} \mathbb{P}\left[\frac{1}{n}\ln\frac{|\mathcal{C}(G(n,d/n),\boldsymbol{\tau})|}{Z_k(G(n,d/n))} \le \Sigma_k(d)+\varepsilon \,\middle|\, \chi(G(n,d/n))\le k\right] = 1, \quad \textit{and}$$
$$\lim_{\varepsilon\searrow 0}\limsup_{n\to\infty} \mathbb{P}\left[\frac{1}{n}\ln\frac{|\mathcal{C}(G(n,d/n),\boldsymbol{\tau})|}{Z_k(G(n,d/n))} \ge \Sigma_k(d)-\varepsilon \,\middle|\, \chi(G(n,d/n))\le k\right] > 0.$$

We observe that our conditioning on the chromatic number $\chi(G(n,d/n))$ being at most k is necessary to speak of a random k-coloring $\boldsymbol{\tau}$ but otherwise harmless. For the first part of Theorem 1 implies that $G(n,d/n)$ is k-colorable w.h.p. for any $d < d_{k,\mathrm{cond}}$. Indeed, if $d < d_{k,\mathrm{cond}}$, then $\Phi_k(d) = k(1-1/k)^{d/2} > 0$ and thus $Z_k(G(n,d/n))^{1/n} > 0$ w.h.p. because $(Z_k(G(n,d/n))^{1/n})$ converges to $\Phi_k(d)$ in probability.

In words, Corollary 2 states that there is a certain function $\Sigma_k > 0$ such that the total number of k-colorings exceeds the number of k-colorings in the cluster of a randomly chosen k-coloring by at least a factor of $\exp[n(\Sigma_k(d)+o(1))]$ w.h.p. However, as d approaches $d_{k,\mathrm{cond}}$, $\Sigma_k(d)$ tends to 0, and with a non-vanishing probability the gap between the total number of k-colorings and the size of a single cluster is upper-bounded by $\exp[n(\Sigma_k(d)+o(1))]$.

3 Discussion and Related Work

In this section we discuss some relevant related work and also explain the impact of Theorem 1 on some questions that have come up in the literature.

3.1 The k-Colorability Threshold

The problem of determining the chromatic number of random graphs has attracted a great deal of attention since it was first posed by Erdős and Rényi [11] (see [15] for a comprehensive overview). In the case that $p = d/n$ for a fixed real $d > 0$, the problem amounts to calculating the threshold sequence $d_{k-\mathrm{col}}(n)$. The best current bounds are

$$(2k-1)\ln k - 2\ln 2 + \varepsilon_k \le \liminf_{n\to\infty} d_{k-\mathrm{col}}(n) \le \limsup_{n\to\infty} d_{k-\mathrm{col}}(n) \le (2k-1)\ln k - 1 + \delta_k, \quad (3.1)$$

where $\varepsilon_k, \delta_k \to 0$ as $k\to\infty$. The upper bound is by the "first moment" method [7]. The lower bound rests on a "second moment" argument [8], which improves a landmark result of Achlioptas and Naor [3].

While Theorem 1 allows for the possibility that $d_{k,\mathrm{cond}}$ is equal to the k-colorability threshold $d_{k-\mathrm{col}}$ (if it exists), the physics prediction is that these two are different. More specifically, the cavity method yields a prediction as to the precise value of $d_{k-\mathrm{col}}$ in terms of another distributional fixed point problem. An asymptotic expansion in terms of k leads to the conjecture $d_{k-\mathrm{col}} = (2k-1)\ln k - 1 + \eta_k$ with $\eta_k \to 0$ as $k\to\infty$. Thus, the upper bound in (3.1) is conjectured to be asymptotically tight in the limit $k\to\infty$.

The present work builds upon the second moment argument from [8]. Conversely, Theorem 1 yields a small improvement over the lower bound from [8]. Indeed, as we saw above Theorem 1 implies that $\liminf_{n\to\infty} d_{k-\mathrm{col}}(n) \ge d_{k,\mathrm{cond}}$, thereby determining the precise "error term" ε_k in the lower bound (3.1). In fact, $d_{k,\mathrm{cond}}$ is the best-possible lower bound that can be obtained via a certain "natural" type of second moment argument.

3.2 "Quiet Planting?"

The notion that for d close to the (hypothetical) k-colorability threshold $d_{k-\mathrm{col}}$ it seems difficult to find a k-coloring of $G(n, d/n)$ algorithmically could be used to construct a candidate one-way function [2] (see also [14]). This function maps a k-coloring σ to a random graph $G(n, p', \sigma)$ by linking any two vertices v, w with $\sigma(v) \neq \sigma(w)$ with some p' independently. The edge probability p' could be chosen such that the average degree of the resulting graph is close to the k-colorability threshold. The resulting distribution on graphs is the so-called *planted model*.

If the planted distribution is close to $G(n, d/n)$, one might think that the function $\sigma \mapsto G(n, p', \sigma)$ is difficult to invert. Indeed, it should be difficult to find *any* k-coloring of $G(n, p', \sigma)$, not to mention the planted coloring σ. As shown in [2], the planted distribution and $G(n, d/n)$ are interchangeable (in a certain precise sense) iff $\Phi_k(d) = k(1-1/k)^{d/2}$. Hence, $d_{k,\mathrm{cond}}$ marks the point where these two distributions start to differ. In particular, Theorem 1 shows that at the k-colorability threshold, the two distributions are *not* interchangeable.

3.3 Message Passing Algorithms

The cavity method has inspired new "message passing" algorithms by the name of Belief/Survey Propagation Guided Decimation [26]. Experiments on random graph k-coloring instances for small values of k show an excellent performance of these algorithms [4, 30, 22]. However, whether these experimental results are reliable and/or extend to larger k remains shrouded in mystery.

For instance, Belief Propagation Guided Decimation can most easily be described in terms of list colorings. Suppose that G is a given input graph. Initially, the list of colors available to each vertex is the full set $[k]$. The algorithm chooses a color for one vertex at a time as follows. First, it performs a certain fixed point iteration to approximate for each vertex the marginal probability of taking some color i in a randomly chosen proper list coloring of G. Then, a vertex v is chosen, say, uniformly at random and a random color i is chosen from the (supposed) approximation to its marginal distribution. The color list of v is reduced to the singleton $\{i\}$, color i gets removed the lists of all the neighbors of v, and we repeat. The algorithm terminates when either for each vertex a color has been chosen ("success") or the list of some vertex becomes empty ("failure"). Ideally, if at each step the algorithm manages to compute precisely the correct marginal distribution, the result would be a uniformly random k-coloring of the input graph. Of course, generating such a random k-coloring is $\#P$-hard in the worst case, and the crux is that the aforementioned fixed point iteration may or may not produce a good approximation to the actual marginal distribution.

Perhaps the most plausible stab at understanding Belief Propagation Guided Decimation is the non-rigorous contribution [28]. Roughly speaking, the result of the Belief Propagation fixed point iteration after t iterations can be expected to yield a good approximation to the actual marginal distribution iff there is no condensation among the remaining list colorings. If so, one should expect that the algorithm actually finds a k-coloring if condensation does not occur at any step $0 \leq t \leq n$. Thus, we look at a two-dimensional "phase diagram" parametrised by the average degree d and the time t/n. We need to identify the line that marks the (suitably defined) condensation phase transition in this diagram. Theorem 1 deals with the case $t = 0$, and it would be most interesting to see if the present techniques extend to $t \in (0, 1)$. Attempts at (rigorously) analysing message passing algorithms along these lines have been made for random k-SAT, but the results have been far from precise [6, 9].

3.4 The Physics Perspective

In physics terminology the random graph coloring problem is an example of a "diluted mean-field model of a disordered system". The term "mean-field" refers to the fact that there is no underlying lattice geometry, while "diluted" indicates that the average degree in the underlying graph is bounded. Moreover, "disordered systems" reflects that the model involves some degree of randomness (i. e., the random graph). Diluted mean-field models are considered a better approximation to "real" disordered systems (such as glasses) than models where the underlying graph is complete, the Sherrington-Kirkpatrick model [25]. From the viewpoint of physics, the question of whether "disordered systems" exhibit a condensation phase transition can be traced back to Kauzmann's experiments in the 1940s [16]. In models where the underlying graph is complete, physicsts predicted an affirmative answer in the 1980s [17], and this has long been confirmed rigorously [29].

With respect to "diluted" models, Coja-Oghlan and Zdeborova [10] showed that a condensation phase transition *exists* in random r-uniform hypergraph 2-coloring. Furthermore, [10] determines the location of the condensation phase transition up to an error ε_r that tends to zero as the uniformity r of the hypergraph becomes large. By contrast, Theorem 1 is the first result that pins down the *exact* condensation phase transition in a diluted mean-field model.

Technically, we build upon some of the techniques that have been developed to study the "geometry" of the set of k-colorings of the random graph, and add to this machinery. Among the techniques that we harness is the "planting trick" from [2] (which, in a sense, we are going to "put into reverse"), the notion of a core [2, 8, 27], techniques for proving the existence of "frozen variables" [27], and a concentration argument from [10]. Additionally, our proof directly incorporates some of the physics calculations from [22, Appendix C]. That said, the cornerstone of the present work is a novel argument that allows us to connect the distributional fixed point problem from [22] rigorously with the geometry of the set of k-colorings.

4 Proof Outline

From now on we assume that $k \geq k_0$ for some large enough constant k_0.

The proof of Theorem 1 is composed of two parallel threads. The first thread is to identify an "obvious" point where a phase transition occurs or, more specifically, a critical degree $d_{k,\mathrm{crit}}$ where statements (i)-(iii) of the theorem are met. The second thread is to identify the frozen fixed point $\pi^*_{d,k}$ of $\mathcal{F}_{d,k}$ and to interpret it combinatorially. Finally, the two threads intertwine to show that $d_{k,\mathrm{crit}} = d_{k,\mathrm{cond}}$, i. e. that the "obvious" phase transition $d_{k,\mathrm{crit}}$ is indeed the unique zero of equation (2.6). The first thread is an extension of ideas developed in [10] for random hypergraph 2-coloring to the (technically far more involved) random graph coloring problem. The second thread and the intertwining of the two require novel arguments.

4.1 The First Thread

Because the nth root sits inside the expectation, the quantity

$$\Phi_k(d) = \lim_{n\to\infty} \mathbb{E}[Z_k(G(n, d/n))^{1/n}]$$

is difficult to calculate for general values of d. However for $d \in [0, 1)$, $\Phi_k(d)$ is easily understood. In fact, the celebrated result of Erdős and Rényi [11] implies that for $d \in [0, 1)$

the random graph $G(n, d/n)$ is basically a forest. Moreover, the number of k-colorings of a forest with n vertices and m edges is well-known to be $k^n(1-1/k)^m$. Since $G(n, d/n)$ has $m \sim dn/2$ edges w.h.p., we obtain

$$Z_k(G(n,d/n))^{1/n} \sim k(1-1/k)^{d/2} \qquad \text{for } d < 1. \tag{4.1}$$

As $Z_k(G)^{1/n} \leq k$ for any graph on n vertices, (4.1) implies that

$$\Phi_k(d) = \lim_{n\to\infty} \mathbb{E}[Z_k(G(n,d/n))^{1/n}] = k(1-1/k)^{d/2} \qquad \text{for } d < 1. \tag{4.2}$$

Clearly, the function $d \mapsto k(1-1/k)^{d/2}$ is analytic on all of $(0,\infty)$. Therefore, the uniqueness of analytic continuations implies that the least $d > 0$ where the limit $\Phi_k(d)$ either fails to exist or differs from $k(1-1/k)^{d/2}$ is going to be a phase transition. Hence, we let

$$d_{k,\mathrm{crit}} = \sup\left\{d \geq 0 : \text{the limit } \Phi_k(d) \text{ exists and } \Phi_k(d) = k(1-1/k)^{d/2}\right\}. \tag{4.3}$$

▶ **Fact 3.** *We have* $d_{k,\mathrm{crit}} \leq (2k-1)\ln k$.

Thus, $d_{k,\mathrm{crit}}$ is a well-defined finite number, and there occurs a phase transition at $d_{k,\mathrm{crit}}$. Moreover, the following proposition yields a lower bound on $d_{k,\mathrm{crit}}$ and implies that $d_{k,\mathrm{crit}}$ satisfies the first condition in Theorem 1.

▶ **Proposition 4.** *For any* $d > 0$ *we have* $\limsup_{n\to\infty} \mathbb{E}[Z_k(G(n,d/n))^{1/n}] \leq k(1-1/k)^{d/2}$. *Moreover, the number* $d_{k,\mathrm{crit}}$ *satisfies*

$$d_{k,\mathrm{crit}} = \sup\left\{d \geq 0 : \liminf_{n\to\infty} \mathbb{E}[Z_k(G(n,d/n))^{1/n}] \geq k(1-1/k)^{d/2}\right\} \geq (2k-1)\ln k - 2. \tag{4.4}$$

Thus, we know that there *exists* a number $d_{k,\mathrm{crit}}$ that satisfies conditions (i)–(ii) in Theorem 1. Of course, to actually calculate this number we need to unearth its combinatorial "meaning". As we saw in Section 2, if $d_{k,\mathrm{crit}}$ really is the condensation phase transition, then the combinatorial interpretation should be as follows. For $d < d_{k,\mathrm{crit}}$, the size of the cluster that a randomly chosen k-coloring $\boldsymbol{\tau}$ belongs to is smaller than $Z_k(G(n,d/n))$ by an exponential factor $\exp(\Omega(n))$ w.h.p. But as d approaches $d_{k,\mathrm{crit}}$, the gap between the cluster size and $Z_k(G(n,d/n))$ diminishes. Hence, $d_{k,\mathrm{crit}}$ should mark the point where the cluster size has the same order of magnitude as $Z_k(G(n,d/n))$.

But how can we possibly get a handle on the size of the cluster that a randomly chosen k-coloring $\boldsymbol{\tau}$ of $G(n,d/n)$ belongs to? No "constructive" argument (or efficient algorithm) is known for obtaining a single k-coloring of $G(n,d/n)$ for d anywhere close to $d_{k,\mathrm{crit}}$, let alone for sampling one uniformly at random. Nevertheless, as observed in [2], in the case that $\Phi_k(d) = k(1-1/k)^{d/2}$, i. e., for $d < d_{k,\mathrm{crit}}$, it is possible to capture the experiment of first choosing the random graph $G(n,d/n)$ and then sampling a k-coloring $\boldsymbol{\tau}$ uniformly at random by means of a different, much more innocent experiment.

In this latter experiment, we *first* choose a map $\boldsymbol{\sigma} : [n] \to [k]$ uniformly at random. Then, we generate a graph $G(n,p',\boldsymbol{\sigma})$ on $[n]$ by connecting any two vertices $v, w \in [n]$ such that $\boldsymbol{\sigma}(v) \neq \boldsymbol{\sigma}(w)$ with probability p' independently. If $p' = dk/(k-1)$ is chosen so that the expected number of edges is the same as in $G(n,d/n)$, then this so-called *planted model* might be a good approximation to the "difficult" experiment of first choosing $G(n,d/n)$ and then picking a random k-coloring. In particular, we might expect that

$$\mathbb{E}[|\mathcal{C}(G(n,p',\boldsymbol{\sigma}),\boldsymbol{\sigma})|^{1/n}] \sim \mathbb{E}[|\mathcal{C}(G(n,d/n),\boldsymbol{\tau})|^{1/n}],$$

i. e., that the suitably scaled cluster size in the planted model is about the same as the cluster size in $G(n, d/n)$. Hence, $d_{k,\mathrm{crit}}$ should mark the point where $\mathbb{E}[|\mathcal{C}(G(n,p',\boldsymbol{\sigma}),\boldsymbol{\sigma})|^{1/n}]$ equals $k(1-1/k)^{d/2}$. The following Proposition verifies that this is indeed so. Let us write $\boldsymbol{G} = G(n,p',\boldsymbol{\sigma})$ for the sake of brevity.

▶ **Proposition 5.** *Assume that* $(2k-1)\ln k - 2 \le d \le (2k-1)\ln k$ *and set*

$$p' = d'/n \quad \text{with } d' = \frac{dk}{k-1}. \tag{4.5}$$

1. *If*

$$\lim_{\varepsilon\searrow 0}\liminf_{n\to\infty}\mathbb{P}\left[|\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})|^{1/n} \le k(1-1/k)^{d/2}-\varepsilon\right] = 1, \tag{4.6}$$

   then $d \le d_{k,\mathrm{crit}}$.
2. *Conversely, if*

$$\lim_{\varepsilon\searrow 0}\liminf_{n\to\infty}\mathbb{P}\left[|\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})|^{1/n} \ge k(1-1/k)^{d/2}+\varepsilon\right] = 1, \tag{4.7}$$

   then $\limsup_{n\to\infty}\mathbb{E}[Z_k(G(n,d/n))^{1/n}] < k(1-1/k)^{d/2}$. *In particular,* $d \ge d_{k,\mathrm{crit}}$.

4.2 The Second Thread

Our next aim is to "solve" the fixed point problem for $\mathcal{F}_{d,k}$ to an extent that gives the fixed point an explicit combinatorial interpretation. This combinatorial interpretation is in terms of a certain random tree process, associated with a concept of "legal colorings". Specifically, we consider a multi-type Galton-Watson branching process. Its set of types is

$$\mathcal{T} = \{(i,\ell) : i \in [k]\,,\ \ell \subset [k]\,,\ i \in \ell\}\,.$$

The intuition is that i is a "distinguished color" and that ℓ is a set of "available colors". The branching process is further parameterized by a vector $\boldsymbol{q} = (q_1,\ldots,q_k) \in [0,1]^k$ such that $q_1+\cdots+q_k \le 1$. Let $d' = dk/(k-1)$ and

$$q_{i,\ell} = \frac{1}{k}\prod_{j\in\ell\setminus\{i\}}\exp(-q_j d')\cdot\prod_{j\in[k]\setminus\ell}1-\exp(-q_j d') \qquad \text{for } (i,\ell)\in\mathcal{T}.$$

Then

$$\sum_{(i,\ell)\in\mathcal{T}} q_{i,\ell} = 1.$$

Further, for each $(i,\ell)\in\mathcal{T}$ such that $|\ell|>1$ we define $\mathcal{T}_{i,\ell}$ as the set of all $(i',\ell')\in\mathcal{T}$ such that $\ell\cap\ell'\neq\emptyset$ and $|\ell'|>1$. In addition, for $(i,\ell)\in\mathcal{T}$ such that $|\ell|=1$ we set $\mathcal{T}_{i,\ell}=\emptyset$.

The branching process $\mathrm{GW}(d,k,\boldsymbol{q})$ starts with a single individual, whose type $(i,\ell)\in\mathcal{T}$ is chosen from the probability distribution $(q_{i,\ell})_{(i,\ell)\in\mathcal{T}}$. In the course of the process, each individual of type $(i,\ell)\in\mathcal{T}$ spawns a Poisson number $\mathrm{Po}(d'q_{i',\ell'})$ of offspring of type (i',ℓ') for each $(i',\ell')\in\mathcal{T}_{i,\ell}$. In particular, only the initial individual may have a type (i,ℓ) with $|\ell|=1$, in which case it does not have any offspring. Let $1\le\mathcal{N}\le\infty$ be the progeny of the process (i. e., the total number of individuals created).

We are going to view $\mathrm{GW}(d,k,\boldsymbol{q})$ as a distribution over trees endowed with some extra information. Let us define a *decorated graph* as a graph $T=(V,E)$ together with a map $\vartheta: V\to\mathcal{T}$ such that for each edge $e=\{v,w\}\in E$ we have $\vartheta(w)\in\mathcal{T}_{\vartheta(v)}$. Moreover, a *rooted*

decorated graph is a decorated graph (T, ϑ) together with a distinguished vertex v_0, the *root*. Further, an *isomorphism* between two rooted decorated graphs T and T' is an isomorphism of the underlying graphs that preserves the root and the types of the vertices.

Given that $\mathcal{N} < \infty$, the branching process $\mathrm{GW}(d, k, \boldsymbol{q})$ canonically induces a probability distribution over isomorphism classes of rooted decorated trees. Indeed, we obtain a tree whose vertices are all the individuals created in the course of the branching process and where there is an edge between each individual and its offspring. The individual from which the process started is the root. Moreover, by construction each individual v comes with a type $\vartheta(v)$. We denote the (random) isomorphism class of this tree by $\boldsymbol{T}_{d,k,\boldsymbol{q}}$. (It is natural to view the branching process as a probability distribution over *isomorphism classes* as the process does not specify the order in which offspring is created.)

To proceed, we define a *legal coloring* of a decorated graph (G, ϑ) as a map $\tau : V(G) \to [k]$ such that τ is a k-coloring of G and such that for any type $(i, \ell) \in \mathcal{T}$ and for any vertex v with $\vartheta(v) = (i, \ell)$ we have $\tau(v) \in \ell$. Combinatorially, if $\vartheta(v) = (i, \ell)$, then we think of ℓ as a list of colors available to v and of i as a "distinguished color". Let $\mathcal{Z}(G, \vartheta)$ denote the number of legal colorings.

Since $\mathcal{Z}(G, \vartheta)$ is isomorphism-invariant, we obtain the integer-valued random variable $\mathcal{Z}(\boldsymbol{T}_{d,k,\boldsymbol{q}})$. We have $\mathcal{Z}(\boldsymbol{T}_{d,k,\boldsymbol{q}}) \geq 1$ with certainty because a legal coloring τ can be constructed by coloring each vertex with its distinguished color (i. e., setting $\tau(v) = i$ if v has type (i, ℓ)). Hence, $\ln \mathcal{Z}(\boldsymbol{T}_{d,k,\boldsymbol{q}})$ is a well-defined non-negative random variable. Additionally, we write $|\boldsymbol{T}_{d,k,\boldsymbol{q}}|$ for the number of vertices in $\boldsymbol{T}_{d,k,\boldsymbol{q}}$.

Finally, consider a rooted, decorated tree (T, ϑ, v_0) and let $\boldsymbol{\tau}$ be a legal coloring of (T, ϑ, v_0) chosen uniformly at random. Then the color $\boldsymbol{\tau}(v_0)$ of the root is a random variable with values in $[k]$. Let $\mu_{T,\vartheta,v_0} \in \Omega$ denote the distribution of this random variable. Clearly, μ_{T,ϑ,v_0} is invariant under isomorphisms. Consequently, the distribution $\mu_{\boldsymbol{T}_{d,k,\boldsymbol{q}}}$ of the color of the root of a tree in the random isomorphism class $\boldsymbol{T}_{d,k,\boldsymbol{q}}$ is a well-defined Ω-valued random variable. Let $\pi_{d,k,\boldsymbol{q}} \in \mathcal{P}$ denote its distribution. Then we can characterise the frozen fixed point of $\mathcal{F}_{d,k}$ as follows.

▶ **Proposition 6.** *Suppose that* $d \geq (2k-1)\ln k - 2$.

1. *The function*

$$q \in [0,1] \mapsto (1 - \exp(-dq/(k-1)))^{k-1} \tag{4.8}$$

has a unique fixed point q^* *in the interval* $[2/3, 1]$. *Moreover, with*

$$\boldsymbol{q}^* = k^{-1}(q^*, \ldots, q^*) \in [0,1]^k \tag{4.9}$$

the branching process $\mathrm{GW}(d, k, \boldsymbol{q}^*)$ *is sub-critical. Thus,* $\mathbb{P}[\mathcal{N} < \infty] = 1$.
2. *The map* $\mathcal{F}_{d,k}$ *has precisely one frozen fixed point, namely* $\pi_{d,k,\boldsymbol{q}^*}$.
3. *We have* $\phi_{d,k}(\pi_{d,k,\boldsymbol{q}^*}) = \mathbb{E}\left[\frac{\ln \mathcal{Z}(\boldsymbol{T}_{d,k,\boldsymbol{q}^*})}{|\boldsymbol{T}_{d,k,\boldsymbol{q}^*}|}\right]$.
4. *The function* Σ_k *from (2.6) is strictly decreasing and continuous on* $[(2k-1)\ln k - 2, (2k-1)\ln k - 1]$ *and has a unique zero* $d_{k,\mathrm{cond}}$ *in this interval.*

4.3 Tying Up the Threads

To prove that $d_{k,\mathrm{cond}} = d_{k,\mathrm{crit}}$, we establish a connection between the random tree $\boldsymbol{T}_{d,k,\boldsymbol{q}^*}$ and the random graph $\boldsymbol{G}$ with planted coloring $\boldsymbol{\sigma}$. We start by giving a recipe for computing the cluster size $|\mathcal{C}(\boldsymbol{G}, \boldsymbol{\sigma})|$, and then let the random tree process "cook" it.

Computing the cluster size hinges on a close understanding of its combinatorial structure. As hypothesised in physics work [25] and established rigorously in [2, 7, 27], typically many vertices v are "frozen" in $\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})$, i. e., $\tau(v)=\tau'(v)$ for any two colorings $\tau,\tau'\in\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})$. More generally, we consider for each vertex v the set

$$\ell(v)=\{\tau(v):\tau\in\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})\}$$

of colors that v may take in colorings τ that belong to the cluster. Together with the "planted" color $\boldsymbol{\sigma}(v)$, we can thus assign each vertex v a type $\vartheta(v)=(\boldsymbol{\sigma}(v),\ell(v))$. This turns $\boldsymbol{G}$ into a decorated graph $(\boldsymbol{G},\vartheta)$.

By construction, each coloring $\tau\in\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})$ is a legal coloring of the decorated graph $\boldsymbol{G}$. Conversely, it turns out that w.h.p. any legal coloring of $(\boldsymbol{G},\vartheta)$ belongs to the cluster $\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})$. Hence, computing the cluster size $|\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})|$ amounts to calculating the number $\mathcal{Z}(\boldsymbol{G},\vartheta)$ of legal colorings of $\boldsymbol{G},\vartheta$.

This calculation is facilitated by the following observation. Let $\widetilde{\boldsymbol{G}}$ be the graph obtained from $\boldsymbol{G}$ by deleting all edges $e=\{v,w\}$ that join two vertices such that $\ell(v)\cap\ell(w)=\emptyset$. Then any legal coloring τ of $\widetilde{\boldsymbol{G}}$ is a legal coloring of $\boldsymbol{G}$, because $\tau(v)\in\ell(v)$ for any vertex v. Hence, $\mathcal{Z}(\boldsymbol{G},\vartheta)=\mathcal{Z}(\widetilde{\boldsymbol{G}},\vartheta)$.

Thus, we just need to compute $\mathcal{Z}(\widetilde{\boldsymbol{G}},\vartheta)$. This task is much easier than computing $\mathcal{Z}(\boldsymbol{G},\vartheta)$ directly because $\widetilde{\boldsymbol{G}}$ turns out to have *significantly* fewer edges than $\boldsymbol{G}$ w.h.p. More precisely, w.h.p. $\widetilde{\boldsymbol{G}}$ (mostly) consists of connected components that are trees of bounded size. In fact, in a certain sense the distribution of the tree components converges to that of the decorated random tree $\boldsymbol{T}_{d,k,\boldsymbol{q}_*}$. In effect, we obtain

▶ **Proposition 7.** *Suppose that $d\geq(2k-1)\ln k-2$ and let p' be as in (4.5). Let $\boldsymbol{q}^*$ be as in (4.9). Then the sequence $\{\frac{1}{n}\ln|\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})|\}_n$ converges to $\mathbb{E}\left[\frac{\ln\mathcal{Z}(\boldsymbol{T}_{d,k,\boldsymbol{q}^*})}{|\boldsymbol{T}_{d,k,\boldsymbol{q}^*}|}\right]$ in probability.*

The proof of Proposition 7 is the centrepiece of this work. It is based on the precise analysis of a further message-passing algorithm called *Warning Propagation* on the random graph $(\boldsymbol{G},\boldsymbol{\sigma})$ chosen from the planted model. The following section contains an outline of this analysis Combining Propositions 5 and 7, we see that $d_{k,\mathrm{crit}}$ is equal to $d_{k,\mathrm{cond}}$ given by Proposition 6. Theorem 1 then follows from Proposition 4.

5 The Cluster Size

5.1 Warning Propagation

A key step towards the proof of Proposition 7 is to determine the set

$$\ell(v)=\{\tau(v):\tau\in\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})\}$$

of colors that a vertex v may take under a k-coloring in $\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})$. In particular, we called a vertex *frozen* if $\ell(v)=\{\boldsymbol{\sigma}(v)\}$. To establish Proposition 7, we will first show that the sets $\ell(v)$ can be determined by means of a message-passing algorithm called *Warning Propagation* ("WP") [25]. WP has been previously analysed on planted k-SAT instances (where it is similar to Unit Clause Propagation) to show that the algorithm actually finds a solution w.h.p. under certain assumptions [12]. Moreover, the work [27] on frozen variables in k-coloring has a WP flavour. But here we use WP to achieve an even more delicate objective: we aim to figure out the *number* of solutions in the cluster of the planted k-coloring.

More precisely, we will see that WP yields color sets $L(v)$ such that $L(v)=\ell(v)$ for all but $o(n)$ vertices w.h.p. Crucially, by tracing WP we will be able to determine for any given

type (i, ℓ) how many vertices of that type there are. Moreover, we will show that the cluster essentially consists of all k-colorings τ of $\boldsymbol{G}$ such that $\tau(v) \in L(v)$ for all v. In addition, the number of such colorings τ can be calculated by considering a reduced graph $\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma})$. This graphs turns out to be mainly a forest, and finally, informally speaking, w.h.p. the statistics of the trees in this forest coincide with the distribution of the random tree $\boldsymbol{T}_{d,k,\boldsymbol{q}^*}$.

Let us begin by describing Warning Propagation on a general graph G endowed with a k-coloring σ. For each edge $e = \{v, w\}$ of G and any color i we define a sequence $(\mu_{v\to w}(i, t|G, \sigma))_{t\geq 1}$ such that $\mu_{v\to w}(i, t|G, \sigma) \in \{0, 1\}$ for all i, v, w, t. The idea is that $\mu_{v\to w}(i, t|G, \sigma) = 1$ indicates that in the tth step of the process vertex v "warns" vertex w that the other neighbors $u \neq w$ of v force v to take color i. We initialize this process by having each vertex v emit a warning about its original $\sigma(v)$ at $t = 0$, i. e.,

$$\mu_{v\to w}(i, 0|G, \sigma) = \mathbf{1}_{i=\sigma(v)} \tag{5.1}$$

for all edges $\{v, w\}$ and all $i \in [k]$. Letting $\partial v = \partial_G(v)$ denote the neighborhood of v in G, for $t \geq 0$ we let

$$\mu_{v\to w}(i, t+1|G, \sigma) = \prod_{j\in[k]\setminus\{i\}} \max\left\{\mu_{u\to v}(j, t|G, \sigma) : u \in \partial v \setminus \{w\}\right\}. \tag{5.2}$$

That is, v warns w about color i in step $t + 1$ iff at step t it received warnings from its other neighbors u (*not* including w) about all colors $j \neq i$. Further, for a vertex v and $t \geq 0$ we let

$$L(v, t|G, \sigma) = \left\{j \in [k] : \max_{u\in\partial v} \mu_{u\to v}(j, t|G, \sigma) = 0\right\} \quad \text{and} \quad L(v|G, \sigma) = \bigcup_{t=0}^{\infty} L(v, t|G, \sigma).$$

Thus, $L(v, t|G, \sigma)$ is the set of colors that vertex v receives no warnings about at step t. To unclutter the notation, we omit the reference to G, σ where it is apparent from the context.

To understand the semantics of this process, observe that by construction the list $L(v, t|G, \sigma)$ only depends on the vertices at distance at most $t + 1$ from v. Further, if we assume that the tth neighborhood $\partial^t v$ in G is a tree, then $L(v, t|G, \sigma)$ is precisely the set of colors that v may take in k-colorings τ of G such that $\tau(w) = \sigma(w)$ for all vertices w at distance greater than t from v. (This can be verified by a straightforward induction on t.) As we will see, this observation together with the fact that the random graph $\boldsymbol{G}$ contains only few short cycles allows us to show that for most vertices v we have $\ell(v) = L(v|\boldsymbol{G}, \boldsymbol{\sigma})$ w.h.p. In effect, the number of k-colorings τ of $\boldsymbol{G}$ with $\tau(v) \in L(v|\boldsymbol{G}, \boldsymbol{\sigma})$ for all v will emerge to be a very good approximation to the cluster size $\mathcal{C}(\boldsymbol{G}, \boldsymbol{\sigma})$.

Counting these k-colorings τ is made possible by the following observation. For a graph G together with a k-coloring σ, let us denote by $G_{\mathrm{WP}}(t|\sigma)$ the graph obtained from G by removing all edges $\{v, w\}$ such that either $|L(v, t)| < 2$, $|L(w, t)| < 2$ or $L(v, t) \cap L(w, t) = \emptyset$. Furthermore, obtain $G_{\mathrm{WP}}(\sigma)$ from G by removing all edges $\{v, w\}$ such that $L(v) \cap L(w) = \emptyset$. We view $G_{\mathrm{WP}}(t|\sigma)$ and $G_{\mathrm{WP}}(\sigma)$ as decorated graphs in which each vertex v is endowed with the color list $L(v, t)$ and $L(v)$ respectively. As before, we let $\mathcal{Z}$ denote the number of legal colorings of a decorated graph. The key statement in this section is

▶ **Proposition 8.** *W.h.p. we have* $\ln \mathcal{Z}(\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma})) = \ln |\mathcal{C}(\boldsymbol{G}, \boldsymbol{\sigma})| + o(n)$.

We begin by proving that $\mathcal{Z}(\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma}))$ is a lower bound on the cluster size w.h.p. First we are going to argue that w.h.p. in $\boldsymbol{G}$ there are many of frozen vertices, and that thus *all* legal colorings τ of $\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma})$ belong to the cluster $\mathcal{C}(\boldsymbol{G}, \boldsymbol{\sigma})$ w.h.p. To exhibit frozen vertices, we

consider an appropriate notion of a "core". More precisely, assume that σ is a k-coloring of a graph G. We denote by $\mathrm{core}(G,\sigma)$ the largest set V' of vertices with the following property.

$$\text{If } v \in V' \text{ and } j \neq \sigma(v), \text{ then } |V' \cap \sigma^{-1}(j) \cap \partial v| \geq 100. \tag{5.3}$$

In words, any vertex in the core has at least 100 neighbors of any color $j \neq \sigma(v)$ that also belong to the core. The core is well-defined; for if V', V'' are two sets with this property, then so is $V' \cup V''$. The following is immediate from the definition.

▶ **Fact 9.** *Assume that $v \in \mathrm{core}(G,\sigma)$. Then $L(v,t) = \{\sigma(v)\}$ for all t.*

The core has become a standard tool in the theory of random structures in general and in random graph coloring in particular (e. g., [2, 8, 27]). Indeed, standard arguments show that $\boldsymbol{G}$ has a very large core w.h.p.

▶ **Proposition 10** ([8])**.** *W.h.p. we have*

$$|\mathrm{core}(\boldsymbol{G},\boldsymbol{\sigma}) \cap \boldsymbol{\sigma}^{-1}(i)| \geq \frac{n}{k}(1-k^{-2/3}) \quad \text{for all } i \in [k]. \tag{5.4}$$

Moreover, if $v \in \mathrm{core}(\boldsymbol{G},\boldsymbol{\sigma})$, then $\boldsymbol{\sigma}(v) = \tau(v)$ for all $\tau \in \mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})$.

▶ **Corollary 11.** *W.h.p. we have $|\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})| \geq \mathcal{Z}(\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma}))$.*

While $\mathcal{Z}(\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma}))$ provides a lower bound on the cluster size, the two numbers do not generally coincide. This is because for a few vertices v, the set $L(v)$ produced by WP may be a proper subset of $\ell(v)$. (Bipartite sub-structures known as "Kempe chains" are for instance responsable for this, cf. [27].) The origin of this problem is that we launched WP from the initialization (5.1), which is the obvious choice but may be too restrictive. Thus, to obtain an upper bound on the cluster size we will start WP from a different initialization. Ideally, this starting point should be such that only vertices that are frozen emit warnings. By Proposition 10, the vertices in the core meet this condition w.h.p. Thus, we are going to compare the above instalment of Warning Propagation with the result of starting WP from an initialization where only the vertices in the core send out warnings.

Thus, given a graph G together with a k-coloring σ we let

$$\begin{aligned}\mu'_{v\to w}(i,0|G,\sigma) &= \mathbf{1}_{i=\sigma(v)} \cdot \mathbf{1}_{v\in\mathrm{core}(G,\sigma)},\\ \mu'_{v\to w}(i,t+1|G,\sigma) &= \prod_{j\in[k]\setminus\{i\}} \max\{\mu'_{u\to v}(j,t|G,\sigma) : u \in \partial v \setminus \{w\}\}\end{aligned}$$

for all edges $\{v,w\}$ of G, all $i \in [k]$ and all $t \geq 0$. Furthermore, let

$$L'(v,t|G,\sigma) = \left\{j \in [k] : \max_{u\in\partial(v)} \mu'_{u\to v}(j,t|G,\sigma) = 0\right\} \quad \text{and} \quad L'(v|G,\sigma) = \bigcap_{t=0}^{\infty} L'(v,t|G,\sigma).$$

As before, we drop G,σ from the notation where possible.

Similarly as before, we can use the lists $L'(v,t)$ to construct a decorated reduced graph denoted by $G'_{\mathrm{WP}}(t|\sigma)$ and $G'_{\mathrm{WP}}(\sigma)$. Proceeding much as above, we obtain

▶ **Lemma 12.** *W.h.p. we have $|\mathcal{C}(\boldsymbol{G},\boldsymbol{\sigma})| \leq \mathcal{Z}(\boldsymbol{G}'_{\mathrm{WP}}(\boldsymbol{\sigma}))$.*

Combining Corollary 11 and Lemma 12, we see that $\mathcal{Z}(\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma})) \leq |\mathcal{C}(\boldsymbol{G},\sigma)| \leq \mathcal{Z}(\boldsymbol{G}'_{\mathrm{WP}}(\boldsymbol{\sigma}))$ w.h.p. To complete the proof of Proposition 8, we are going to argue that $\ln \mathcal{Z}(\boldsymbol{G}'_{\mathrm{WP}}(\boldsymbol{\sigma})) = \ln \mathcal{Z}(\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma})) + o(n)$ w.h.p.

To this end, we need one more general construction. Let G be a graph and let σ be a k-coloring of G. Let $t \geq 0$ be an integer. For each vertex v of G we define a rooted, decorated graph $T(v,t|G,\sigma)$ as follows.

- The graph underlying $T(v,t|G,\sigma)$ is the connected component of v in $G_{\mathrm{WP}}(t|\sigma)$.
- The root of $T(v,t|G,\sigma)$ is v.
- The type of each vertex w of $T(v,t|G,\sigma)$ is $(\sigma(w), L(w,t|G,\sigma))$.

Analogously we obtain a rooted, decorated graph $T(v|G,\sigma)$ from $G_{\mathrm{WP}}(\sigma)$, $T'(v,t|G,\sigma)$ from $G'_{\mathrm{WP}}(t|\sigma)$ and $T'(v|G,\sigma)$ from $G'_{\mathrm{WP}}(\sigma)$. By carefully coupling our two versions of WP, we obtain

▶ **Lemma 13.** *W.h.p.* $\boldsymbol{G},\boldsymbol{\sigma}$ *is such that* $T(v|\boldsymbol{G},\boldsymbol{\sigma}) = T'(v|\boldsymbol{G},\boldsymbol{\sigma})$ *for all but* $o(n)$ *vertices* v.

5.2 Counting Legal Colorings

Proposition 8 reduces the proof of Proposition 7 to the problem of counting the legal colorings of the reduced graph $\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma})$. For a rooted, decorated tree T let H_T be the number of vertices v in $\boldsymbol{G}_{\mathrm{WP}}(\boldsymbol{\sigma})$ such that $T(v|\boldsymbol{G},\boldsymbol{\sigma}) \cong T$. Let us write $\boldsymbol{T} = \boldsymbol{T}_{d,k,\boldsymbol{q}^*}$ for the sake of brevity. Recall that $\boldsymbol{T}$ is an isomorphism class of rooted, decorated trees; thus, it makes sense to write $T \in \boldsymbol{T}$. To complete the proof of Proposition 7 we need to show the following.

▶ **Proposition 14.** *For any* T *the sequence* $(\frac{1}{n}H_T)_{n\geq 1}$ *converges to* $\mathbb{P}[T \in \boldsymbol{T}]$ *in probability.*

This can be shown by proving that the number q^* from Proposition 6 provides a good approximation to the number of vertices v such that $L(v|\boldsymbol{G},\boldsymbol{\sigma}) = \{i\}$ for any i. As a next step, it can be argued that WP "converges quickly". More specifically, for most vertices v the component $T(v|\boldsymbol{G},\boldsymbol{\sigma})$ is already completely determined after just a bounded number t of iterations of WP. This reduces the proof of Proposition 14 to the problem of studying the statistics of the trees $T(v,t|\boldsymbol{G},\boldsymbol{\sigma})$ with $t \geq 0$ (large but) fixed as $n \to \infty$. This problem is *much* simpler than the original one, because we only need to iterate WP for t rounds. Finally, Proposition 7 follows from Propositions 8 and 14.

Acknowledgment. We thank Guilhem Semerjian for helpful discussions and explanations regarding the articles [20, 22] and Nick Wormald for pointing us to [24, Theorem 3.8].

References

1 D. Achlioptas, E. Friedgut: A sharp threshold for k-colorability. Random Struct. Algorithms **14** (1999) 63–70.
2 D. Achlioptas, A. Coja-Oghlan: Algorithmic barriers from phase transitions. Proc. 49th FOCS (2008) 793–802.
3 D. Achlioptas, A. Naor: The two possible values of the chromatic number of a random graph. Annals of Mathematics **162** (2005) 1333–1349.
4 A. Braunstein, R. Mulet, A. Pagnani, M. Weigt, R. Zecchina: Polynomial iterative algorithms for coloring and analyzing random graphs. Phys. Rev. E **68** (2003) 036702.
5 P. Cheeseman, B. Kanefsky, W. Taylor: Where the *really* hard problems are. Proc. IJCAI (1991) 331–337.
6 A. Coja-Oghlan: On belief propagation guided decimation for random k-SAT. Proc. 22nd SODA (2011) 957–966.
7 A. Coja-Oghlan: Upper-bounding the k-colorability threshold by counting covers. Electronic Journal of Combinatorics **20** (2013) P32.
8 A. Coja-Oghlan, Dan Vilenchik: Chasing the k-colorability threshold. Proc. 54th FOCS (2013) 380–389. A full version is available as arXiv:1304.1063.
9 A. Coja-Oghlan, A. Y. Pachon-Pinzon: The decimation process in random k-SAT. SIAM Journal on Discrete Mathematics **26** (2012) 1471–1509.

10 A. Coja-Oghlan, L. Zdeborová: The condensation transition in random hypergraph 2-coloring. Proc. 23rd SODA (2012) 241–250.

11 P. Erdős, A. Rényi: On the evolution of random graphs. Magayar Tud. Akad. Mat. Kutato Int. Kozl. **5** (1960) 17–61.

12 U. Feige, E. Mossel, D. Vilenchik: Complete convergence of message passing algorithms for some satisfiability problems. Theory of Computing **9** (2013) 617–651.

13 E. Friedgut: Sharp thresholds of graph properties, and the k-SAT problem. J. AMS **12** (1999) 1017–1054.

14 O. Goldreich: Candidate one-way functions based on expander graphs. Electronic Colloquium on Computational Complexity (ECCC) **7** (2000).

15 S. Janson, T. Łuczak, A. Ruciński: Random Graphs, Wiley 2000.

16 W. Kauzmann: The nature of the glassy state and the behavior of liquids at low temperatures. Chem. Rev. **43** (1948) 219–256.

17 T. R. Kirkpatrick, D. Thirumalai: p-spin-interaction spin-glass models: Connections with the structural glass problem. Phys. Rev. B **36** (1987) 5388.

18 F. Krzakala, J. Kurchan: A Landscape Analysis of Constraint Satisfaction Problems. Phys. Rev. E **76** (2007) 02112.

19 F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, L. Zdeborová: Statistical physics-based reconstruction in compressed sensing. Phys. Rev. X **2** (2012), 021005.

20 F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, L. Zdeborova: Gibbs states and the set of solutions of random constraint satisfaction problems. Proc. National Academy of Sciences **104** (2007) 10318–10323.

21 F. Krzakala, L. Zdeborová: Generalization of the cavity method for adiabatic evolution of Gibbs states. Phys. Rev. B **81** (2010) 224205.

22 F. Krzakala, L. Zdeborová: Phase transition in the coloring of random graphs. Phys. Rev. E **76** (2007) 031131.

23 H. Levesque, D. Mitchell, B. Selman: Hard and easy distribution of SAT problems. Proc. 10th AAAI (1992) 459–465.

24 C. McDiarmid: Concentration. In Habib et al. (eds): Probabilistic methods for algorithmic discrete mathematics. Springer (1998) 195–248.

25 M. Mézard, A. Montanari: Information, physics and computation. Oxford University Press 2009.

26 M. Mézard, G. Parisi, R. Zecchina: Analytic and algorithmic solution of random satisfiability problems. Science **297** (2002) 812–815.

27 M. Molloy: The freezing threshold for k-colourings of a random graph. Proc. 43rd STOC (2012) 921–930.

28 F. Ricci-Tersenghi, G. Semerjian: On the cavity method for decimated random constraint satisfaction problems and the analysis of belief propagation guided decimation algorithms. J. Stat. Mech. (2009) P09001.

29 M. Talagrand: Spin glasses, a Challenge for Mathematicians. Springer 2003.

30 L. Zdeborová: Statistical Physics of Hard Optimization Problems. Acta Physica Slovaca **59** (2009) 169–303.