2008 49th Annual IEEE Symposium on Foundations of Computer Science

Algorithmic Barriers from Phase Transitions

Dimitris Achlioptas* Amin Coja-Oghlan†

Abstract

For many random Constraint Satisfaction Problems, by now there exist asymptotically tight estimates of the largest constraint density for which solutions exist. At the same time, for many of these problems, all known polynomial-time algorithms stop finding solutions at much smaller densities. For example, it is well-known that it is easy to color a random graph using twice as many colors as its chromatic number. Indeed, some of the simplest possible coloring algorithms achieve this goal. Given the simplicity of those algorithms, one would expect room for improvement. Yet, to date, no algorithm is known that uses $(2-\epsilon)\chi$ colors, in spite of efforts by numerous researchers over the years.

In view of the remarkable resilience of this factor of 2 against every algorithm hurled at it, we find it natural to inquire into its origin. We do so by analyzing the evolution of the set of k-colorings of a random graph, viewed as a subset of $\{1,\ldots,k\}^n$, as edges are added. We prove that the factor of 2 corresponds in a precise mathematical sense to a phase transition in the geometry of this set. Roughly speaking, we prove that the set of k-colorings looks like a giant ball for $k \geq 2\chi$, but like an error-correcting code for $k \leq (2-\epsilon)\chi$. We also prove that an analogous phase transition occurs both in random k-SAT and in random hypergraph 2-coloring. And that for each of these three problems, the location of the transition corresponds to the point where all known polynomial-time algorithms fail. To prove our results we develop a general technique that allows us to establish rigorously much of the celebrated 1-step Replica-Symmetry-Breaking hypothesis of statistical physics for random CSPs.

*UC Santa Cruz, Santa Cruz, CA 95064, USA and CTI, Greece. Supported by NSF CAREER award CCF-0546900, an Alfred P. Sloan Fellowship, and IDEAS grant 210743 from the European Research Council. optas@cs.ucsc.edu.

†School of Informatics, University of Edinburgh, UK. Supported by DFG CO 646. acoghlan@inf.ed.ac.uk.

1 Introduction

For many random Constraint Satisfaction Problems (CSP), such as random graph coloring, random k-SAT, random Max k-SAT, and hypergraph 2-coloring, by now there exist asymptotically tight estimates for the largest constraint density for which typical instances have solutions (see [5]). At the same time, all known efficient algorithms for each of these problems fare very poorly, i.e., they stop finding solutions at constraint densities *much* lower than those for which we can prove that solutions exist. Adding insult to injury, for each problem the best known algorithm asymptotically fares no better than some extremely naive algorithm.

For example, it has been known for nearly twenty years [10] that the following very simple algorithm will find a satisfying assignment of a random k-CNF formula with $m = rn$ clauses for $r = O(2^k/k)$: if there is a unit clause satisfy it; otherwise assign a random value to a random unassigned variable. While it is known that random k-CNF remain satisfiable for $r = \Theta(2^k)$, no polynomial-time algorithm is known to find satisfying assignments for $r = (2^k/k)\cdot\omega(k)$ for some function $\omega(k) \to \infty$.

Similarly, for all $k \geq 3$, the following algorithm [18, 2] will k-color a random graph with average degree $d \leq k\ln k$: select a random vertex with fewest available colors left and assign it a random available color. While it is known that random graphs remain k-colorable for $d \sim 2\,k\ln k$, no polynomial-time algorithm is known to k-color a random graph of average degree $(1+\epsilon)k\ln k$ for some fixed $\epsilon > 0$ and arbitrarily large k. Equivalently, it is trivial to color a random graph using twice as many colors as its chromatic number, but no polynomial-time algorithm is known that gets by with $(2-\epsilon)\chi$ colors, for some fixed $\epsilon > 0$.

Random k-SAT and random graph coloring are not alone. In fact, for nearly every random CSP of interest, the known results establish an analogous state of the art:

1. There is a trivial upper bound on the largest constraint density for which solutions exist.
2. There is a non-constructive proof, usually via the second moment method, that the bound from (1) is essentially tight, i.e., that solutions do exist for densities nearly as high as the trivial upper bound.

0272-5428/08 $25.00 © 2008 IEEE
DOI 10.1109/FOCS.2008.11

3. A simple algorithm finds solutions up to a constraint density much below the one from (2).

4. No polynomial-time algorithm is known to succeed for a density asymptotically greater than that in (3).

In this paper we prove that this is not a coincidence. Namely, for random graph coloring, random k-SAT, and random hypergraph 2-coloring, we prove that the point where all known algorithms stop is the point where the geometry of the space of solutions undergoes a dramatic change. This is known as a "dynamical" phase transition in statistical physics and our results establish rigorously for random CSPs a large part of the "1-step Replica Symmetry Breaking" hypothesis [20]. Roughly speaking, this hypothesis asserts that while the set of solutions for low densities looks like a giant ball, at some critical point this ball shatters into exponentially many pieces that are far apart from one another and separated by huge "energy barriers", like an error-correcting code. Algorithms (even extremely simple ones) have no problem finding solutions in the "ball" regime, but no algorithm is known to find solutions in the "error-correcting code" regime.

We believe that the presence of dynamical phase transitions in random CSPs is a very general phenomenon whose qualitative characteristics should be problem-independent, i.e., *universal*. The fact that we can establish the exact same qualitative picture for a problem with binary constraints over k-ary variables (random graph k-coloring) and a problem with k-ary constraints over binary variables (hypergraph 2-colorability) lends support to this notion.

Perhaps the following is an intuitive model of how a dynamical phase transition comes about. In random graph coloring, rather than thinking of the number of available colors as fixed and the constraint density (number of edges) as increasing, imagine that we keep the constraint density fixed, but we keep decreasing the number of available colors. If we start with q available colors where $q \gg \chi$, it is reasonable to imagine that the set of valid q-colorings, viewed as a subset of $\{1, 2, \ldots, q\}^n$, has a nice "round" shape, the rounder the greater q is relative to χ. By the same token, when we restrict our attention to the set of those q-colorings that only use colors $\{1, 2, \ldots, q-1\}$, we are taking a "slice' of the set of q-colorings. With each slicing the connectivity of the set at hands deteriorates, until at some point the set shatters. As an analogy, slicing the 2-dimensional unit sphere through the origin yields a circle, but slicing the circle, yields a pair of points.

Having said the above, we wish to emphasize that determining the location of the dynamical phase transition of a given CSP requires non-trivial, problem-specific ideas and computations. In this paper we do this for the three problems mentioned above, allowing us to demonstrate that the transition coincides with the demise of algorithms.

We conclude the introduction with a few words about the technical foundation for our work. To prove the existence (and determine the location) of a dynamical phase transition one needs to access the uniform measure over the solutions of random CSP instances. A geometric way of thinking about this is as follows. Given a CSP instance, say a random k-CNF formula with m clauses over n variables, consider the function H on $\{0,1\}^n$ that assigns to each truth assignment the number of clauses it violates. In this manner, H defines a "landscape" in which satisfying assignments correspond to (valleys at) sea-level. Understanding statistical properties of the uniform measure over solutions entails understanding "the view" one enjoys from the bottom of a random such valley, a probabilistically formidable task.

As we discuss in Section 4, we establish the following: if the number of solutions of a random CSP is sufficiently concentrated around its exponentially large expectation, then the view from a random sea-level valley is "the same" as the view from an "artificial" sea-level valley. That is, in terms of our random k-CNF formula example, from the valley created by first selecting a random $\sigma \in \{0,1\}^n$ and then forming a random k-CNF formula with m clauses chosen uniformly among the clauses satisfied by σ, i.e., the view from the *planted* satisfying assignment. This is a *much* easier view to understand and we believe that the "transfer" theorems we establish in this paper will significantly aid in the analysis of random CSPs in general.

2 Statement of Results

To present our results in a uniform manner we need to introduce some common notions. Let V be a set of n variables, all with the same domain D, and let C be an arbitrary set of constraints over the variables in V. A CSP instance is a subset of C. We let $\mathrm{dist}(\sigma, \tau)$ denote the Hamming distance between $\sigma, \tau \in D^n$ and we turn D^n into a graph by saying that σ, τ are adjacent if $\mathrm{dist}(\sigma, \tau) = 1$. For a given instance I, we let $H = H_I : D^n \to \mathbf{N}$ be the function counting the number of constraints of I violated by each $\sigma \in D^n$.

Definition 1 *We say that $\sigma \in D^n$ is a solution of an instance I, if $H(\sigma) = 0$. We will denote by $\mathcal{S}(I)$ the set of all solutions of an instance I. The **clusters** of an instance I are the connected components of $\mathcal{S}(I)$. A **region** is a non-empty union of clusters. The **height** of a path $\sigma_0, \sigma_1, \ldots, \sigma_t \in D^n$ is $\max_i H(\sigma_i)$.*

Remark 1 *The term cluster comes from physics. Requiring $\mathrm{dist}(\sigma, \tau) = 1$ to say that σ, τ are adjacent is somewhat arbitrary (but conceptually simplest) and a number of our results hold if one replaces 1 with $o(n)$.*

We will be interested in distributions of CSP instances as the number of variables n grows. The set $C = C_n$ will

typically consist of all possible constraints of a certain type, e.g., the set of all $\binom{n}{k}$ possible hyperedges in the problem of 2-coloring random k-uniform hypegraphs. We let $I_{n,m}$ denote the set of all CSP instances with precisely m distinct constraints from C_n and we let $\mathcal{I}_{n,m}$ denote the uniform distribution on the set of all instances $I_{n,m}$. We will say that a sequence of events $\mathcal{E}_n$ holds *with high probability* (w.h.p.) if $\lim_{n\to\infty} \Pr[\mathcal{E}_n] = 1$ and *with uniformly positive probability* (w.u.p.p.) if $\liminf_{n\to\infty} \Pr[\mathcal{E}_n] > 0$. As per standard practice in the study of random structures, we will take the liberty of writing $\mathcal{I}_{n,m}$ to also denote the underlying random variable and, thus, write things like "The probability that $\mathcal{S}(\mathcal{I}_{n,m})$..."

2.1 Shattering

Definition 2 *We say that the set of solutions of $\mathcal{I}_{n,m}$ **shatters** if there exist constants $\beta, \gamma, \zeta, \theta > 0$ such that w.h.p. $\mathcal{S}(\mathcal{I}_{n,m})$ can be partitioned into regions so that:*

1. *The number of regions is at least $e^{\beta n}$.*
2. *Each region contains at most an $e^{-\gamma n}$ fraction of all solutions.*
3. *The Hamming distance between any two regions is at least ζn.*
4. *Every path between vertices in distinct regions has height at least θn.*

Our first main result asserts that the space of solutions for random graph coloring, random k-SAT, and random hypergraph 2-colorability shatters and that this shattering occurs just above the largest density for which any polynomial-time algorithm is known to find solutions for the corresponding problem. Moreover, we prove that the space remains shattered until, essentially, the CSP's satisfiability threshold. More precisely:

– A random graph with average degree d, i.e., $m = dn/2$, is w.h.p. k-colorable for $d \le (2-\gamma_k)k\ln k$, where $\gamma_k \to 0$. The best rigorously analyzed poly-time k-coloring algorithm w.h.p. fails for $d \ge (1+\delta_k)k\ln k$, where $\delta_k \to 0$.

Theorem 1 *There exists a sequence $\epsilon_k \to 0$, such that the space of k-colorings of a random graph with average degree d shatters for all*

$$(1+\epsilon_k)k\ln k \le d \le (2-\epsilon_k)k\ln k \ . \qquad (1)$$

– A random k-CNF formula with n variables and rn clauses is w.h.p. satisfiable for $r \le 2^k \ln 2 - k$. The best rigorously analyzed poly-time satisfiability algorithm w.h.p. fails for $r > 2^{k+1}/k$. In [23], non-rigorous, but mathematically sophisticated evidence is given that a different algorithm succeeds for $r = \Theta((2^k/k)\ln k)$ but not higher.

Theorem 2 *There exists a sequence $\epsilon_k \to 0$ such that the space of satisfying assignments of a random k-CNF formula with rn clauses shatters for all*

$$(1+\epsilon_k)\,\frac{2^k}{k}\,\ln k \le r \le (1-\epsilon_k)2^k \ln 2 \ . \qquad (2)$$

– A random k-uniform hypergraph with n variables and rn edges is w.h.p. 2-colorable for $r \le 2^{k-1}\ln 2 - \frac{3}{2}$. The best rigorously analyzed poly-time 2-coloring algorithm w.h.p. fails for $r > 2^k/k$. In [23], non-rigorous, but mathematically sophisticated evidence is given that a different algorithm succeeds for $r = \Theta((2^k/k)\ln k)$, but not higher.

Theorem 3 *There exists a sequence $\epsilon_k \to 0$ such that the space of 2-colorings of a random k-uniform hypergraph with rn edges shatters for all*

$$(1+\epsilon_k)\,\frac{2^{k-1}}{k}\,\ln k \le r \le (1-\epsilon_k)2^{k-1} \ln 2 \ . \qquad (3)$$

Remark 2 *As the notation in Theorems 1,2,3 is asymptotic in k, the stated intervals may be empty for small values of k. In this extended abstract we have not optimized the proofs to deliver the smallest values of k for which the intervals are non-empty. Quick calculations suggest $k \ge 6$ for hypergraph 2-colorability, $k \ge 8$ for k-SAT, and $k \ge 20$ for k-coloring.*

2.2 Rigidity

The regions mentioned in Theorems 1, 2 and 3 can be thought of as forming an error-correcting code in the solution-space of each problem. To make this precise we need to introduce the following definition and formalize the notion of "a random solution of a random instance".

Definition 3 *Given an instance I, a solution $\sigma \in \mathcal{S}(I)$ and a variable $v \in V$, we say that v in (I,σ):*

- *Is $f(n)$-**rigid**, if every $\tau \in \mathcal{S}(I)$ such that $\tau(v) \neq \sigma(v)$ has dist$(\sigma,\tau) \ge f(n)$.*
- *Is $f(n)$-**loose**, if for every $j \in D$, there exists $\tau \in \mathcal{S}(I)$ such that $\tau(v) = j$ and dist$(\sigma,\tau) \le f(n)$.*

We will prove that in a typical solution, while before the phase transition every variable is loose, after the phase transition nearly every variable is rigid. To formalize the notion of a random/typical solution, recall that $I_{n,m}$ denotes the set of all instances with m constraints over n variables and let $\Lambda = \Lambda_{n,m}$ denote the set of all instance–solution pairs, i.e., $\Lambda_{n,m} = \{(I,\sigma) : I \in I_{n,m},\ \sigma \in \mathcal{S}(I)\}$. We let $\mathcal{U} = \mathcal{U}_{n,m}$ be the probability distribution induced on $\Lambda_{n,m}$ by the following experiment:

Choose an instance $I \in I_{n,m}$ uniformly at random.
If $\mathcal{S}(I) \neq \emptyset$, select $\sigma \in \mathcal{S}(I)$ uniformly at random.

We will refer to instance-solution pairs generated according to $\mathcal{U}_{n,m}$ as **uniform** instance-solution pairs. We note that although the definition of uniform pairs allows for $\mathcal{S}(I)$ to be typically empty, i.e., to be in the typically unsatisfiable regime, we will employ the definition for constraint densities such that w.h.p. $\mathcal{S}(I)$ contains exponentially many solutions. Hence, our liberty in using the term a "typical" solution. At the same time, we emphasize that $\mathcal{U}_{n,m}$ is in general *not* the uniform distribution over $\Lambda_{n,m}$.

Theorem 4 *Let (I, σ) be a uniform instance-solution pair where:*

- *I is a graph with $dn/2$ edges, where d is as in* (1)*, and σ is a k-coloring of I, or,*
- *I is a k-CNF formula with rn clauses, where r is as in* (2)*, and σ is a satisfying assignment of I, or,*
- *I is a k-uniform hypergraph with rn edges, where r is as in* (3)*, and σ is a 2-coloring of I.*

W.h.p. the number of $\Omega(n)$-rigid variables in (I, σ) is at least $\gamma_k n$, for some sequence $\gamma_k \to 1$.

Remark 3 *Theorem 4 is tight since random instances as above w.h.p. have $\Omega(n)$ variables that are not bound by any constraint and hence can never be rigid.*

The picture drawn by Theorem 4, whereby nearly all variables are rigid in typical solutions above the dynamical phase transition, is in sharp contrast with our results for densities below the transition for graph coloring and hypergraph 2-colorability. While we believe that an analogous picture holds for k-SAT, see Conjecture 1, for technical reasons we cannot establish this presently. (We discuss the general additional difficulties imposed by random k-SAT in Section 4.)

Theorem 5 *Let (I, σ) be a uniform instance-solution pair where:*

- *I is a graph with $dn/2$ edges, where $d \leq (1-\epsilon_k)k \ln k$, and σ is a k-coloring of I, or,*
- *I is a k-uniform hypergraph with rn edges, where $r \leq (1-\epsilon_k)(2^{k-1}/k) \ln k$, and σ is a 2-coloring of I.*

There exists a sequence $\epsilon_k \to 0$ such that w.h.p. every variable in (I, σ) is $o(n)$-loose.

We note that in fact, for all d and r as in Theorem 5, w.u.p.p. (I, σ) is such that setting the color of any vertex to any color only requires changing the color of $O(\log n)$ other vertices.

Conjecture 1 *Let (I, σ) be a uniform instance-solution pair where I is a k-CNF formula with rn clauses, where $r \leq (1-\epsilon_k)(2^k/k) \ln k$, and σ is a satisfying assignment of I. There exists a sequence $\epsilon_k \to 0$ such that w.h.p. every variable in (I, σ) is $o(n)$-loose.*

3 Background and Related Work

3.1 Algorithms

Attempts for a "quick improvement" upon either of the naive algorithms mentioned in the introduction for satisfiability/graph coloring stumble upon the following general fact. Given a CSP instance, consider the bipartite graph in which every variable is adjacent to precisely those constraints in which it appears, known as the factor graph of the instance. For random formulas/graphs, factor graphs are locally tree-like, i.e., for any arbitrarily large constant D, the depth-D neighborhood of a random vertex is a tree w.h.p. In other words, locally, random CSPs are trivial, e.g., random graphs of any finite average degree are locally 2-colorable. Moreover, as the constraint density is increased, the factor graphs of random CSPs get closer and closer to being biregular, so that degree information is not useful either. Combined, these two facts render all known algorithms impotent, i.e., as the density is increased, their asymptotic performance matches that of trivial algorithms.

In [22], Mézard, Parisi, and Zecchina proposed a new satisfiability algorithm called Survey Propagation (SP) which performs extremely well experimentally on instances of random 3-SAT. This was very surprising at the time and allowed for optimism that, perhaps, random k-SAT instances might not be so hard. Later, SP was extended to other problems, e.g., k-coloring [9] and Max k-SAT [8]. An experimental evaluation of SP for values of k even as small as 5 or 6 is already somewhat problematic, but to the extent it is reliable it strongly suggests that SP does not find solutions for densities as high as those for which solutions are known to exist. The problem seems to lie with the "decimation" aspect of the algorithm, i.e., the process of repeatedly selecting (and setting permanently) the most "biased" variables in the formula (we comment on this further below). Perhaps more importantly, it can be shown that for densities at least as high as $2^k \ln 2 - k$, if SP can succeed at its main task (approximating the marginal probability distribution of the variables with respect to the uniform measure over an approximation of the cluster projections), then so can a much simpler algorithm, namely Belief Propagation (BP), i.e., dynamic programming on trees (for approximating the marginal probability distribution of the variables with respect to the uniform measure over satisfying assignments).

The trouble is that to use either BP or SP to find satisfying assignments one sets variables iteratively. So, even if it is possible to compute approximately correct marginals at the beginning of the execution (for the entire formula), this can stop being the case after some variables are set. Concretely, in [23], Montanari et al. showed that (even within the relatively generous assumptions of statistical physics computations) the following Gibbs-sampling algorithm fails above density $e(2^k/k)$, i.e., step 2 below fails to converge after only a small fraction of all variables have been assigned a value:

1. Select a variable v at random.
2. Compute the marginal distribution of v using Belief Propagation.
3. Set v to $\{0, 1\}$ according to the computed marginal distribution; simplify the formula; go to step 1.

3.2 Relating the Uniform and the Planted Model

The idea of deterministically embedding a property inside a random structure is very old and, in general, the process of doing this is referred to as "planting" the property. In our case, we plant a solution σ in a random CSP by only including constraints compatible with σ. Juels and Peinado [19] were perhaps the first to explore the relationship between the planted and the uniform model and they did so for the clique problem in dense random graphs $G_{n,1/2}$, i.e., where each edge appears independently with probability 1/2. They showed that the distribution resulting from first choosing $G = G_{n,1/2}$ and then planting a clique of size $(1 + \varepsilon) \log_2 n$ is very close to $G_{n,1/2}$ and suggested this as a scheme to obtain a one-way-function. Since the planted clique has size only $(1 + \varepsilon) \log_2 n$, the basic argument in [19] is closely related to subgraph counting. In contrast, the objects under consideration in our work (k-colorings, satisfying assignments, etc.) have an immediate impact on the *global* structure of the combinatorial object being considered, rather than just being local features, such as a clique on $O(\log n)$ vertices.

Coja-Oghlan, Krivelevich, and Vilenchik [12, 13] proved that for constraint densities well above the threshold for the existence of solutions, the planted model for k-coloring and k-SAT is equivalent to the uniform distribution *conditional* on the (exponentially unlikely) existence of at least one solution. In this conditional distribution as well as in the high-density planted model, the geometry of the solution space is very simple, as there is precisely one cluster of solutions, in stark contrast with the regime we analyze.

3.3 Solution-space Geometry

In [7, 21] the first steps were made towards understanding the solution-space geometry of random k-CNF formulas by proving the existence of shattering and the presence of rigid variables for $r = \Theta(2^k)$. This was a far cry from the true $r \sim (2^k/k) \ln k$ threshold for the onset of both phenomena, as we establish here. Besides the quantitative aspect, there is also a fundamentally important difference in the methods employed in [7, 21] vs. those employed here. In those works, properties such as the existence of frozen variables were shown to hold for *all* satisfying assignments and were correspondingly established by taking a union bound over all satisfying assignments. It is not hard to show that the derived results are best possible using those methods and, in fact, there is good reason to believe that the results are genuinely tight, i.e., that for densities $o(2^k)$ the derived properties simply do not hold for *all* satisfying assignments. Here, we instead establish a systematic connection between the planted model and random solutions of random instances. This argument allows us to analyze "typical" solutions while allowing for the possibility that a (relatively small, though exponential) number of "atypical" solutions exist. Therefore, we are for the first time in a position to analyze the extremely complex energy landscape of below-threshold instances of random CSPs, and to estimate quantities that appeared completely out of reach prior to this work.

4 Our Point of Departure: Symmetry, Randomness and Inversion

As mentioned, the results in this paper are enabled by a set of technical lemmas that allow one to reduce the study of "random solutions of random CSP instances" to the study of "planted CSP solutions". The conceptual origin of these lemmas can be traced to the following humble observation.

Let M be an arbitrary 0-1 matrix with the property that all its rows have the same number of 1s and all its columns have the same the number of 1s. A moment's reflection makes it clear that for such a matrix, both of the following methods select a uniformly random 1 from the entire matrix (and are therefore equivalent):

1. Select a uniformly random column and then a uniformly random 1 in that column.
2. Select a uniformly random row and then a uniformly random 1 in that row.

An example of how we employ this fact for random CSPs is as follows. Let $\mathcal{F}$ be the set of all k-CNF formulas with n variables and m distinct clauses (chosen among all $2^k \binom{n}{k}$ possible k-clauses). Say that $\sigma \in \{0, 1\}^n$ NAE-satisfies a

formula $F \in \mathcal{F}$ if under σ, every clause of F has at least one satisfied and at least one falsified literal. Let M be the $2^n \times |\mathcal{F}|$ matrix where $M_{\sigma,F} = 1$ iff $\sigma \in \{0,1\}^n$ NAE-satisfies F. By the symmetry of $\mathcal{F}$, it is clear that all rows of M have the same number of 1s. Imagine, for a moment, that the same was true for all columns. Then, a uniformly random solution of a uniformly random instance would be distributed *exactly* as a "planted" instance-solution pair: first select $\sigma \in \{0,1\}^n$ uniformly at random; then select m distinct clauses uniformly at random among all $2^{k-1}\binom{n}{k}$ clauses NAE-satisfied by σ.

Our contribution begins with the realization that exact row- and column-balance is not necessary. Rather, it is enough for the 1s in M to be "well-spread". More precisely, it is enough that the marginal distribution induced on the rows of M by selecting a uniformly random 1 from the entire matrix to be "reasonably close to" uniform, and that exactly the same holds for the columns of M. For example, assume we can prove that $\Omega(|\mathcal{F}|)$ columns of M have $\Theta(f(n))$ 1s, where $f(n)$ is the average number of 1s per column in the entire matrix. Indeed, this is precisely the kind of property implied by the success of the second moment method for random NAE-k-SAT [3]. Under this assumption, proving that a property holds w.u.p.p. for a uniformly random solution of a uniformly random instance, reduces to proving that it holds w.h.p. for the planted solution of a planted instance, a dramatically simpler task.

There is a geometric intuition behind our transfer theorems and it is more conveniently described when every constraint is included independently with the same probability $p = m/\left(2^k\binom{n}{k}\right)$. For all $k \geq 3$ and $m = rn$, it was shown in [3] that the resulting NAE-k-SAT instances w.u.p.p. have exponentially many solutions for $r \leq 2^{k-1}\ln 2 - 3/2$. Consider now the following way of generating *planted* NAE k-SAT instances. First, select a formula F by including each clause with probability p, exactly as above. Then, select $\sigma \in \{0,1\}^n$ uniformly at random and remove from F all constraints violated by σ. Call the resulting instance F'. Our results say that as long as $q \equiv r(1-2^{-k+1}) \leq 2^{k-1}\ln 2 - 3/2$, the instance F' is "nearly indistinguishable" from a *uniform* instance created by including each clause with probability q. (We will make this statement precise shortly.)

To see how this happens, recall the function $H : \sigma \to \mathbf{N}$ counting the number of violated constraints under each assignment. Clearly, selecting F specifies such a function H_F, while selecting $\sigma \in \{0,1\}^n$ and removing all constraints violated by σ amounts to modifying H_F so that $H_F(\sigma) = 0$. One can imagine that such a modification creates a gradient in the vicinity of σ, a "crater" with σ at its bottom. What we prove is that as long as H_F already had an exponential number of craters and the number of craters is concentrated, adding one more crater does not make a big difference. Of course, if the density is increased further, the opened crater becomes increasingly obvious, as it takes a larger and larger cone to get from the typical values of H_F down to 0. This observation also relates to the ease with which algorithms solve planted instances of high density.

To prove our transfer theorems we instantiate the above idea for random graph k-coloring, random k-uniform hypergraph 2-coloring, and random k-SAT. A crucial step for this is deriving a lower bound on the number of solutions of a random instance. For example, in the case of random graph k-coloring, we prove that the number of k-colorings, $|\mathcal{S}(I_{n,m})|$, for a random graph with n vertices and m edges is "concentrated" around its expectation in the sense that w.h.p.

$$n^{-1}\ |\ln|\mathcal{S}(I_{n,m})| - \ln \mathbf{E}|\mathcal{S}(I_{n,m})|\ | = o(1)\ . \quad (4)$$

To prove this, we use the upper bound on the second moment $\mathbf{E}\left[|\mathcal{S}(I_{n,m})|^2\right]$ from [4] to show that w.u.p.p. $|\mathcal{S}(I_{n,m})| = \Omega(\mathbf{E}|\mathcal{S}(I_{n,m})|)$. Then, we perform a sharp threshold analysis using theorems of Friedgut [15], to prove that (4) holds, in fact, with *high* probability. A similar approach applies to hypegraph 2-coloring.

The situation for random k-SAT is more involved. Indeed, we can prove that the number of satisfying assignments is *not* concentrated around its expectation in the sense of (4). This problem is mirrored by the fact that the second moment of the number of satisfying assignments exceeds the square of the first moment by an exponential factor (for any constraint density). Nonetheless, letting $F_k(n,m)$ denote a uniformly random k-CNF formula with n variables and m clauses, combining techniques from [6] with a sharp threshold analysis, we can derive a lower bound on the number of satisfying assignments that holds w.h.p., namely $n^{-1}\ln|\mathcal{S}(F_k(n,m))| \geq n^{-1}\ln \mathbf{E}|\mathcal{S}(F_k(n,m))| - \phi(k)$, where $\phi(k) \to 0$ exponentially with k. This estimates allows us to approximate the uniform model by the planted model sufficiently well in order to establish Theorems 2 and 4.

5 Proof sketches

In this section we give proof sketches of our results for k-coloring to offer a feel of the transfer theorems and of the style of the arguments one can employ given those theorems. The proofs for hypergraph 2-coloring are relatively similar, as it is also a "symmetric" CSP and the second moment methods works on its number of solutions. For k-SAT, though, a significant amount of additional work is needed, as properties must be established with exponentially small error probability to overcome the large deviations in the number of satisfying assignments.

5.1 The Transfer for Random Graph Coloring

We consider a fixed number $\varepsilon > 0$ and assume that $k \geq k_0$ for some sufficiently large $k_0 = k_0(\varepsilon)$. We denote $\{1, \ldots, k\}$ as $[k]$. We are interested in the probability distribution $\mathcal{U}_{n,m}$ on $\Lambda_{n,m}$ resulting from first choosing a random graph $G = G(n, m)$ and then a random k-coloring of G (if one exists). To analyze this distribution, we consider the distribution $\mathcal{P}_{n,m}$ on $\Lambda_{n,m}$ induced by following expermient.

P1. Generate a uniformly random k-partition $\sigma \in [k]^n$.

P2. Generate a graph G with m edges chosen uniformly at random among the edges bicolored under σ.

P3. Output the pair (G, σ).

The distribution $\mathcal{P}_{n,m}$ is known as the *planted model*.

Theorem 6 *Suppose that $d = 2m/n \leq (2-\varepsilon)k \ln k$. There exists a function $f(n) = o(n)$ such that the following is true. Let $\mathcal{D}$ be any graph property such that $G(n, m)$ has $\mathcal{D}$ with probability $1 - o(1)$, and let $\mathcal{E}$ be any property of pairs $(G, \sigma) \in \Lambda_{n,m}$. If for all sufficiently large n,*

$$\Pr_{\mathcal{P}_{n,m}}[(G, \sigma) \text{ has } \mathcal{E} | G \text{ has } \mathcal{D}] \geq 1 - \exp(-f(n)), \quad (5)$$

then $\Pr_{\mathcal{U}_{n,m}}[(G, \sigma) \text{ has } \mathcal{E}] = 1 - o(1)$.

5.2 Loose Variables Below the Transition

Suppose that $d \leq (1 - \varepsilon)k \ln k$. Recall that a graph with vertex set V is said to be ζ-choosable if for any assignment of color lists of length at least ζ to the elements of V, there is a proper coloring in which every vertex receives a color from its list. To prove Theorem 5, we consider the property $\mathcal{E}$ that all vertices are $o(n)$-loose and the following condition $\mathcal{D}$:

> For any set $S \subset V$ of size $|S| \leq g(n)$ the subgraph induced on S is 3-choosable.

Here $g(n)$ is some function such that $f(n) \ll g(n) = o(n)$, where $f(n)$ is the function from Theorem 6. A standard argument shows that a random graph $G(n, m)$ with $m = O(n)$ satisfies $\mathcal{D}$ w.h.p.

By Theorem 6, we are thus left to establish (5). Let $\sigma \in [k]^n$ be a uniformly random k-partition, and let G be a random graph with m edges such that σ is a k-coloring of G. Since σ is uniformly random, we may assume that the color classes $V_i = \sigma^{-1}(i)$ satisfy $|V_i| \sim n/k$. Let $v_0 \in V$ be any vertex, and let $l \neq \sigma(v_0)$ be the "target color" for v_0. Our goal is to find a coloring τ such that $\tau(v_0) = l$ and $\text{dist}(\sigma, \tau) \leq g(n)$.

If v_0 has no neighbor in V_l, then we can just assign color l to v_0. Otherwise, we run the following process. In the course of the process, every vertex is either *awake*, *dead*, or *asleep*. Initially, all the neighbors of v_0 in V_l are awake, v_0 is dead, and all other vertices are asleep. In each step of the process, pick an awake vertex w arbitrarily and declare it dead (if there is no awake vertex, the process terminates). If there are at least four colors $c_1(w), \ldots, c_4(w)$ available such that w has no asleep neighbor in $V_{c_i(w)}$, then we do nothing. Otherwise, we pick four colors $c_1(w), \ldots, c_4(w)$ randomly and declare all asleep neighbors of w in $V_{c_j(w)}$ awake for $1 \leq j \leq 4$.

Lemma 1 *With probability at least $1 - \exp(-f(n))$ there are at most $g(n)$ dead vertices when the process terminates.*

Proof sketch. We relate the above process to a subcritical branching process. To this end, we bound the expected "offspring" (i.e., number of new awake vertices) generated in any step of the process. Suppose that in some step of the process an awake vertex w is chosen. To bound the expected offspring, we basically observe that when $d < (1 - \varepsilon)k \ln k$ it is very likely that a vertex w has four immediately available colors, and thus no offspring will be generated at all. More precisely, the number of neighbors of w in any class V_i with $i \neq \sigma(w)$ is asymptotically Poisson with mean

$$\frac{2m}{(k-1)n} \leq (1 - \varepsilon)\frac{k \ln k}{k - 1}.$$

Hence, the probability that w does *not* have a neighbor in V_i is asymptotically equal to

$$\pi_k = \exp\left(-\frac{2m}{(k-1)n}\right) \geq k^{\varepsilon/2-1}$$

(for sufficiently large k). As there are k colors in total, the expected number of colors $i \neq \sigma(w)$ such that w has no neighbor in V_i is asymptotically Poisson with mean $(k - 1)\pi_k \geq k^{\varepsilon/3}$. Consequently, the probability that w has at least than four available colors is at least

$$1 - \Pr\left[\text{Poisson}(k^{\varepsilon/3}) < 4\right] \geq 1 - \exp(-k^{\varepsilon/4}), \quad (6)$$

and in this case w does not produce any offspring at all. If w has fewer than four available colors, then the number of neighbors of w in a randomly chosen color class is stochastically dominated by a Poisson distribution with mean $k \ln k$ conditioned on being at least one. Hence, in this case we can bound the expected number of newly awake vertices by $4 \cdot 2k \ln k$. Thus, (6) entails that the expected number of offspring vertices is at most $8k \ln k \cdot \exp(-k^{\varepsilon/4}) < 1$.

As a consequence, the total number of dead vertices at the end of the process is dominated by the total number of offspring generated by a branching process with rate at most

$8k \ln k \cdot \exp(-k^{\varepsilon/4}) < 1$. Therefore, standard tail bounds for branching processes imply the assertion. □

To obtain a new coloring τ in which v_0 takes color l we consider the set D of all dead vertices. We let $\tau(u) = \sigma(u)$ for all $u \in V \setminus D$. Moreover, conditioning on the event $\mathcal{D}$, we can assign color l to v_0 and a color from the list $\{c_1(w), \ldots, c_5(w)\} \setminus \{l\}$ to each $w \in D \setminus \{v_0\}$. Thus, the new coloring τ differs from σ on at most $|D| \leq g(n) = o(n)$ vertices.

5.3 Rigid Variables Above the Transition

Suppose that $d \geq (1+\varepsilon)k \ln k$. To prove Theorem 4 for coloring we apply Theorem 6 as follows. We let $\alpha, \beta > 0$ be sufficiently small numbers and denote by $\mathcal{E}$ the following property of a pair $(G, \sigma) \in \Lambda_{n,m}$:

> There is a subgraph $G_* \subset G$ of size $|V(G_*)| \geq (1-\alpha)n$ such that for every vertex v of G_* and each color $i \neq \sigma(v)$ there are at least $\beta \ln k$ vertices w in G_* that are adjacent to v such that $\sigma(w) = i$. (7)

Also, we let $\mathcal{D}$ be the property that the maximum degree is at most $(\ln n)^2$.

Lemma 2 *Condition (5) holds for $\mathcal{D}$ and $\mathcal{E}$ as above.*

Proof sketch. Let $(G, \sigma) \in \Lambda_{n,m}$ be a random pair chosen from the distribution $\mathcal{P}_{n,m}$. We may assume that $|\sigma^{-1}(i)| \sim n/k$ for all i. To obtain the graph G_*, we perform a "stripping process". As a first step, we obtain a subgraph H by removing from G all vertices that have fewer than $\gamma \ln k$ neighbors in any color class other than their own. If $\gamma = \gamma(\varepsilon)$ is sufficiently small, then the expected number of vertices removed in this way is less than $nk^{-\delta}$ for some fixed $\delta > 0$, because for each vertex w the expected number of neighbors in another color class is asymptotically Poisson with mean at least $(1+\varepsilon)\ln k$. Then, we keep removing vertices from H that have "a lot" of neighbors outside of H. Given the event $\mathcal{D}$, we can show that with probabiltiy $1-\exp(-\Omega(n))$ the final result of this process is a subgraph G_* that satisfies (7). □

Furthermore, the following lemma follows from a standard "first moment" argument.

Lemma 3 *W.h.p. the random graph $G = G_{n,m}$ has the following property.*

> *There is no set S of vertices of size $|S| \leq n/(k \ln k)$ such that S spans at least $\frac{1}{2}|S|\beta \ln k$ edges.* (8)

To complete the proof of Theorem 4 for coloring, consider a pair (G, σ) chosen from $\mathcal{U}_{n,m}$. By Lemma 2 and Theorem 6 there is a subgraph G_* as in (7) w.h.p. Moreover, by Lemma 3 we may assume that G satisfies (8). Now, assume for contradiction that G has a k-coloring $\tau \neq \sigma$ such that the set $U = \{v \in G_* : \sigma(v) \neq \tau(v)\}$ has size $|U| \leq n/(k \ln k)$. Let

$$\begin{aligned} U_i^+ &= \{v \in G_* : \tau(v) = i \neq \sigma(v)\}, \\ U_i^- &= \{v \in G_* : \sigma(v) = i \neq \tau(v)\} \end{aligned}$$

for $1 \leq i \leq k$. Then

$$|U| = \sum_{i=1}^{k} |U_i^+| = \sum_{i=1}^{k} |U_i^-|. \tag{9}$$

Every vertex $v \in G_* \setminus \sigma^{-1}(i)$ has at least $\beta \ln k$ neighbors in $G_* \cap \sigma^{-1}(i)$. Hence, if $v \in U_i^+$, then all of these neighbors lie inside of U_i^-. We claim that this implies that $|U_i^+| < |U_i^-|$; for assume that $U_i^+ \geq U_i^-$. Set $S = U_i^+ \cup U_i^-$. Then $|S| \leq |U| \leq n/(k \ln k)$, and S spans at least $|S|\frac{\beta}{2} \ln k$ edges, in contradiction to (8). Thus, we conclude that $|U_i^+| < |U_i^-|$ for *all* i, in contradiction to (9). Hence, all the vertices in G_* are $\frac{n}{k \ln k}$-rigid.

5.4 Proof of Theorem 1

Theorem 1 concerns the "view" from a random coloring σ of $G(n,m)$. Basically, our goal is to show that only a tiny fraction of all possible colorings are "visible" from σ, i.e., σ lives in a small, isolated valley. To establish the theorem, we need a way to measure how "close" two colorings σ, τ are. The Hamming distance is inappropriate here because two colorings σ, τ can be at Hamming distance n, although τ simply results from permuting the color classes of σ, i.e., although σ and τ are essentially identical. Instead, we shall use the following concept. Given two coloring σ, τ, we let $M_{\sigma,\tau} = (M_{\sigma,\tau}^{ij})_{1 \leq i,j \leq k}$ be the matrix with entries

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

To measure how close τ is to σ we let

$$f_\sigma(\tau) = \|M_{\sigma,\tau}\|_F^2 = \sum_{i,j=1}^{k} (M_{\sigma,\tau}^{ij})^2 \ ,$$

be the squared Frobenius norm of $M_{\sigma,\tau}$. Observe that this quantity reflects the probability that a single random edge is monochromatic under both σ and τ, i.e., the correlation of σ and τ, precisely as desired. Hence, f_σ is a map from the set $[k]^n$ of k-partitions to the interval $\left[k^{-2}, f_\sigma(\sigma)\right]$, where $f_\sigma(\sigma) \geq k^{-1}$. Thus, the *larger* $f_\sigma(\tau)$, the more τ resembles σ. Furthermore, for a fixed $\sigma \in \mathcal{S}(G)$ and a number $\lambda > 0$ we let

$$g_{\sigma,G,\lambda}(x) = |\{\tau \in [k]^n : f_\sigma(\tau) = x \wedge H(\tau) \leq \lambda n\}|.$$

In order to show that $\mathcal{S}(G_{n,m})$ with $m = dn/2$ decomposes into exponentially many regions, we employ the following lemma.

Lemma 4 *Suppose that $d > (1+\varepsilon_k)k\ln k$. There are numbers $k^{-2} < y_1 < y_2 < k^{-1}$ and $\lambda, \gamma > 0$ such that with high probability a pair $(G,\sigma) \in \Lambda_{n,m}$ chosen from the distributoin $\mathcal{U}_{n,m}$ has the following two properties.*

1. *For all $x \in [y_1, y_2]$ we have $g_{\sigma,G,\lambda}(x) = 0$.*
2. *The number of colorings $\tau \in \mathcal{S}(G)$ such that $f_\sigma(\tau) > y_2$ is at most $\exp(-\gamma n) \cdot |\mathcal{S}(G)|$.*

Let $G = G_{n,m}$ be a random graph and call $\sigma \in \mathcal{S}(G)$ *good* if both (1) and (2) hold. Then Lemma 4 states that w.h.p. a $1 - o(1)$-fraction of all $\sigma \in \mathcal{S}(G)$ are good. Hence, to decompose $\mathcal{S}(G)$ into regions, we proceed as follows. For each $\sigma \in \mathcal{S}(G)$ we let $\mathcal{C}_\sigma = \{\tau \in \mathcal{S}(G) : f_\sigma(\tau) > y_2\}$. Then starting with the set $S = \mathcal{S}(G)$ and removing iteratively some $\mathcal{C}_\sigma$ for a good $\sigma \in S$ yields an exponential number of regions. Furthermore, each such region $\mathcal{C}_\sigma$ is separated by a linear Hamming distance from the set $\mathcal{S}(G) \setminus \mathcal{C}_\sigma$. This is because f_σ is "continuous" with respect to $n^{-1}\times$Hamming distance: for any $\xi > 0$ there is $\eta > 0$ such that for any two colorings τ, τ' with $\mathrm{dist}(\tau, \tau') < \eta n$ we have $|f_\sigma(\tau) - f_\sigma(\tau')| < \xi$. Thus, Theorem 1 follows from Lemma 4.

Finally, by Theorem 6, to prove Lemma 4 it is sufficient to show the following.

Lemma 5 *Suppose that $d > (1+\varepsilon_k)k\ln k$. There are $k^{-2} < y_1 < y_2 < k^{-1}$ and $\lambda, \gamma > 0$ such that with probability at least $1-\exp(-\Omega(n))$ a pair $(G,\sigma) \in \Lambda_{n,m}$ chosen from the distributoin $\mathcal{P}_{n,m}$ has the two properties stated in Lemma 4.*

The proof of Lemma 5 is based on the "first moment method". That is, for any $k^{-2} < y < k^{-1}$ we compute the *expected* number of assignments $\tau \in [k]^n$ such that $f_\sigma(\tau) = y$ and $H(\tau) \le \lambda n$. This computation is feasible in the planted model and yields similar expressions as encountered in [4] in the course of computing the second moment of the number of k-colorings. Therefore, we can show that the expected number of such assignments τ is exponentially small for a regime $y_1 < y < y_2$, whence Lemma 5 follows from Markov's inequality.

References

[1] D. Achlioptas, E. Friedgut. *A sharp threshold for k-colorability*, Random Struct. Algorithms **14**, 63–70, 1999.

[2] D. Achlioptas and M. Molloy, *The analysis of a list-coloring algorithm on a random graph*, in Proc. of FOCS 1997, 204–212.

[3] D. Achlioptas and C. Moore, *Random k-SAT: two moments suffice to cross a sharp threshold*, SIAM Journal on Computing, **36** (2006), 740–762.

[4] D. Achlioptas and A. Naor, *The two possible values of the chromatic number of a random graph*, Annals of Mathematics, **162** (2005), 1333–1349.

[5] D. Achlioptas, A. Naor, and Y. Peres, *Rigorous location of phase transitions in hard optimization problems*, Nature, **435** (2005), 759–764.

[6] D. Achlioptas and Y. Peres, *The threshold for random k-SAT is $2^k \ln 2 - O(k)$*, Journal of the American Mathematical Society **17** (2004), 947–973.

[7] D. Achlioptas, F. Ricci-Tersenghi, *On the solution space geometry of random constraint satisfaction problems*, in Proc. 38th ACM Symp. on Theory of Computing (2006), 130–139.

[8] D. Battaglia, M. Kolar, R. Zecchina, *Minimizing energy below the glass thresholds*, Phys. Rev. E. **70** (2004), 036107.

[9] A. Braunstein, R. Mulet, A. Pagnani, M. Weigt, R. Zecchina, *Polynomial iterative algorithms for coloring and analyzing random graphs*, Phys. Rev. E. **68** (2004), 036702.

[10] M.-T. Chao and J. Franco, *Probabilistic analysis of two heuristics for the 3-satisfiability problem*, SIAM J. Comput. **15** (1986), 1106–1118.

[11] V. Chvátal and B. Reed, *Mick gets some (the odds are on his side)*, in Proc. 33th Annual Symposium on Foundations of Computer Science (1992), 620–627.

[12] A. Coja-Oghlan, M. Krivelevich, D. Vilenchik, *Why almost all k-colorable graphs are easy*, in Proc. 24th STACS (2007) 121–132.

[13] A. Coja-Oghlan, M. Krivelevich, D. Vilenchik, *Why almost all k-CNF formulas are easy*, in Proc. 13th International Conference on Analysis of Algorithms.

[14] E. Friedgut, *Sharp thresholds of graph proprties, and the k-SAT problem.* J. Amer. Math. Soc. **12** (1999), 1017–1054.

[15] E. Friedgut, *Hunting for sharp thresholds.* Random Struct. Algorithms **26** (2005) 37–51

[16] A. M. Frieze and S. Suen, *Analysis of two simple heuristics on a random instance of k-SAT*, Journal of Algorithms **20** (1996), 312–355.

[17] A. Gerschenfeld, A. Montanari. *Reconstruction for models on random graphs.* in Proc. FOCS 2007, 194–204.

[18] G.R. Grimmett, C.J.H. McDiarmid, *On colouring random graphs*, Math. Proc. Cambridge Philos. Soc., **77** (1975), 313–324.

[19] A. Juels, M. Peinado: *Hiding cliques for cryptographic security.* in Proc. SODA 1998, 678–684.

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

[21] M. Mézard, T. Mora, and R. Zecchina, *Clustering of solutions in the random satisfiability problem*, Phys. Rev. Lett. **94** (2005), 197205.

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

[23] A. Montanari, F. Ricci-Tersenghi, G. Semerjian. *Solving constraint satisfaction problems through Belief Propagation-guided decimation.* in Proc. 45th Allerton (2007).