

# On the Solution-Space Geometry of Random Constraint Satisfaction Problems

Dimitris Achlioptas

Department of Computer Science, University of California Santa Cruz  
optas@cs.ucsc.edu

Amin Coja-Oghlan

School of Informatics, University of Edinburgh  
acoghlan@inf.ed.ac.uk

Federico Ricci-Tersenghi

Physics Department, University of Rome “La Sapienza”  
federico.ricci@roma1.infn.it

## Abstract

For various random constraint satisfaction problems there is a significant gap between the largest constraint density for which solutions exist and the largest density for which any polynomial time algorithm is known to find solutions. Examples of this phenomenon include random  $k$ -SAT, random graph coloring, and a number of other random Constraint Satisfaction Problems. To understand this gap, we study the structure of the solution space of random  $k$ -SAT (i.e., the set of all satisfying assignments viewed as a subgraph of the Hamming cube). We prove that for densities well below the satisfiability threshold, the solution space decomposes into an exponential number of connected components and give quantitative bounds for the diameter, volume and number.

## 1 Introduction

For a number of random Constraint Satisfaction Problems (CSP), by now very good estimates are available for the largest constraint density (ratio of constraints to variables) for which typical problems have solutions. For instance, in the random  $k$ -SAT problem one asks if a random  $k$ -CNF formula,  $F_k(n, m)$ , with  $n$  variables and  $m$  clauses is satisfiable. It is widely believed that the probability that such a formula is satisfiable exhibits a sharp threshold. Specifically, the *Satisfiability Threshold Conjecture* asserts that  $r_k = r_k^*$  for all  $k \geq 3$ , where

$$\begin{aligned} r_k &\equiv \sup\{r : F_k(n, rn) \text{ is satisfiable w.h.p.}\} , \\ r_k^* &\equiv \inf\{r : F_k(n, rn) \text{ is unsatisfiable w.h.p.}\} . \end{aligned}$$

(As usual, we will say that a sequence of events  $\mathcal{E}_n$  occurs with high probability (w.h.p.) if  $\lim_{n \rightarrow \infty} \Pr[\mathcal{E}_n] = 1$ .)

A simple first moment argument shows that  $r_k^* \leq 2^k \ln 2$ . Moreover, it was shown in [3] via the second moment method that random  $k$ -CNF formulas have satisfying assignments for densities very close to this upper bound: for all  $k \geq 3$ ,

$$r_k > 2^k \ln 2 - \frac{(k+1) \ln 2 + 3}{2} . \tag{1}$$

At the same time, however, there is a significant gap between the lower bound (1) for the existence of satisfying truth assignments and the best algorithmic result: no polynomial algorithm is known that finds satisfying assignments in random  $k$ -CNF formulas when  $r = \omega(k) 2^k / k$ , for some function  $\omega(k) \rightarrow \infty$  (arbitrarily slowly). Table 1 illustrates this gap for some small values of  $k$ . For  $k = 3$ , the upper bound on  $r_k^*$  comes from [11], while for  $k > 3$  from [10, 18].

$k$	3	4	7	10	20	21
Best known upper bound for $r_k^*$	4.508	10.23	87.88	708.94	726,817	1,453,635
Best known lower bound for $r_k$	3.52	7.91	84.82	704.94	726,809	1,453,626
Best known algorithmic lower bound	3.52	5.54	33.23	172.65	95,263	181,453

The best algorithmic lower bound for  $k = 3$  is from [17], while for  $k > 3$  it is from [14]. Similar huge gaps exist for a number of other constraint satisfaction problems, such as random NAE  $k$ -SAT or random graph coloring (for example, see [2]).

Sparse random CSPs have also been studied by physicists under the name “mean-field diluted spin-glasses”. In mathematical terms, “spins” corresponds to the fact that the variables are discrete and have small domain, while “glass” to the fact that different constraints prefer different values for the variables. The term “diluted” refers to the sparsity of the bipartite graph in which each constraint is adjacent to the variables it binds, i.e., the factor graph of the instance. Finally, the term “mean field” refers to the fact that the factor graph is random, i.e., there is no underlying spatial structure mandating which variables interact. The physical interest in mean-field systems stems partly from the fact that for many statistical mechanics problems in which the variables lie on a lattice such as  $\mathbb{Z}^d$ , the effect of the underlying geometry vanishes for all  $d \geq d_u$ , for some upper critical dimension  $d_u$ .

In the last few years, motivated by ideas developed for the study of spin glasses, physicists have put forward a hypothesis for the origin of the aforementioned algorithmic gap in random CSPs and, most remarkably, a method for overcoming it. Specifically, Mézard, Parisi, and Zecchina [23] developed an extremely efficient algorithm, called Survey Propagation (SP), for finding satisfying assignments of random formulas in the satisfiable regime. For example, their algorithm typically finds a satisfying truth assignment of a random 3-CNF formula with  $n = 10^6$  variables and  $4.25n$  clauses in minutes (and appears to scale as  $O(n \log n)$ ). No other algorithm practically solves formulas of such density with  $n = 10^4$ .

The SP algorithm is based on a hypothesis for the solution-space geometry which, in turn, is motivated by a mathematically sophisticated but non-rigorous analysis that uses techniques of statistical physics. In the present paper we make progress towards establishing this hypothesis mathematically. In particular, we prove that already much below the satisfiability threshold, the set of satisfying assignments fragments into exponentially many connected components. Moreover, we prove that these components are relatively small in size and far apart from one another and that as the formula density is increased, these components decrease in volume and grow further apart from one another. We emphasize that while both the discussion and the results we present refer to  $k$ -SAT, this is not strictly necessary: our ideas and proofs are quite generic, and should generalize readily to many other random CSP, e.g., graph coloring.

## 2 Statement of Results

We first need to introduce some definitions. Throughout, we assume that we are dealing with a CNF formula  $F$ , defined over variables  $X = x_1, \dots, x_n$ , and we let  $\mathcal{S}(F) \subseteq \{0, 1\}^n$  denote the satisfying assignments of  $F$ .

**Definition 1.** The *diameter* of an arbitrary set  $X \subseteq \{0, 1\}^n$  is the largest Hamming distance between any two elements of  $X$ . The *distance* between two arbitrary sets  $X, Y \subseteq \{0, 1\}^n$ , is the minimum Hamming distance between any  $x \in X$  and any  $y \in Y$ . The *clusters* of a formula  $F$  are the connected components of  $\mathcal{S}(F)$  when  $x, y \in \{0, 1\}^n$  are considered adjacent if they have Hamming distance 1. A *cluster-region* is a non-empty set of clusters.

**Theorem 2.** For every  $k \geq 8$ , there exists a value of  $r < r_k$  and constants  $\alpha_k < \beta_k < 1/2$  and  $\epsilon_k > 0$  such that w.h.p. the set of satisfying assignments of  $F_k(n, rn)$  consists of  $2^{\epsilon_k n}$  non-empty cluster-regions, such that

1. The diameter of each cluster-region is at most  $\alpha_k n$ .
2. The distance between every pair of cluster-regions is at least  $\beta_k n$ .

In other words, for all  $k \geq 8$ , at some point below the satisfiability threshold, the set of satisfying assignments consists of exponentially many, well-separated cluster-regions. The picture suggested by Theorem 2 comes in sharper

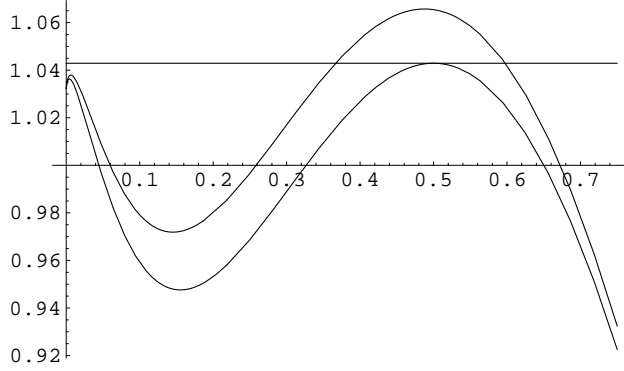


Figure 1: Upper curve  $\Lambda(\alpha, 8, 169)$  and lower curve  $\Lambda_b(\alpha, 8, 169)$  for  $\alpha \in [0, 3/4]$ .

focus for large  $k$ . In particular, for sufficiently large  $k$ , sufficiently close to the threshold, the cluster regions become arbitrarily small and maximally far apart (while remaining exponentially many). The following result gives a quantitative version of this fact.

**Theorem 3.** *For any  $0 < \delta < 1/3$ , if  $r = (1 - \delta)2^k \ln 2$ , then for all  $k \geq k_0(\delta)$ , Theorem 2 holds with*

$$\alpha_k = \frac{1}{k}, \quad \beta_k = \frac{1}{2} - \frac{5}{6}\sqrt{\delta}, \quad \epsilon_k = \frac{\delta}{2} - 3k^{-2}.$$

It is worth noting that, as we will show shortly,

**Remark 4.** *Theorems 2 and 3 remain valid for any definition of clusters in which a pair of assignments are deemed adjacent whenever their distance is at most  $f(n)$  where  $f(n) = o(n)$ .*

### 3 Proof Outline and Related Work

There are two main ingredients for proving Theorems 2 and 3. The first one excludes the possibility of pairs of truth assignments at certain Hamming distances, implying an upper bound on the diameter (and thus the volume) of every cluster.

#### 3.1 Forbidden distances and their implications for clustering

It is easy to show, see e.g., [1], that the expected number of pairs of satisfying assignments in  $F_k(n, rn)$  with Hamming distance  $z$  is at most  $\Lambda(z/n, k, r)^n$ , where

$$\Lambda(\alpha, k, r) = \frac{2(1 - 2^{1-k} + 2^{-k}(1 - \alpha)^k)^r}{\alpha^\alpha(1 - \alpha)^{1-\alpha}}.$$

Therefore, if for some  $k, r$  and  $z = \alpha n$  we have  $\Lambda(\alpha, k, r) < 1$ , it immediately follows by the union bound that w.h.p. in  $F_k(n, rn)$  no pair of satisfying assignments has distance  $z$ . This observation was first made and used in [21]. In Figure 1 we draw the function  $\Lambda$  (upper curve), and a related function  $\Lambda_b$  (lower curve, to be discussed shortly), for  $\alpha \in [0, 3/4]$  with  $k = 8$  and  $r = 169$ . Recall that by the results of [3],  $F_8(n, 169n)$  is w.h.p. satisfiable and, thus, excluding the possibility of satisfying pairs at certain distances is a non-vacuous statement. Letting  $I \equiv [0.06, 0.26] \cup [0.68, 1]$  we see that  $\Lambda(\alpha, 8, 169) < 1$  for  $\alpha \in I$ , implying that w.h.p. in  $F_8(n, 169n)$  there is no pair of satisfying assignments with Hamming distance  $\alpha n$ , where  $\alpha \in I$ .

Knowing that there exists a distance  $z$  such that there are no pairs of assignments at distance  $z$  immediately implies an upper bound on the diameter of every cluster. This is because if a cluster  $C$  has diameter  $d$ , then it must contain

pairs of solutions at every distance  $1 \leq t \leq d$ . To see this, take any pair  $\sigma_1, \sigma_2 \in C$  that have distance  $d$ , any path from  $\sigma_1$  to  $\sigma_2$  in  $C$ , and observe that the sequence of distances from  $\sigma_1$  along the vertices of the path must contain every integer in  $\{1, \dots, d\}$ . Therefore, if  $\Delta = \Delta_{k,r} \equiv \inf\{\alpha : \Lambda(\alpha, k, r) < 1\}$ , then w.h.p. every cluster in  $F_k(n, rn)$  has diameter at most  $\Delta n$ .

If we can further prove that  $\Lambda(\alpha, k, r) < 1$  in an interval  $(\alpha, \beta)$ , then we can immediately partition the set of satisfying assignments into well-separated regions, as follows. Start with any satisfying assignment  $\sigma$ , let  $C$  be its cluster, and consider the set  $R(C) \subseteq \{0, 1\}^n$  of truth assignments that have distance at most  $\alpha n$  from  $C$  and the set  $B(C) \subseteq \{0, 1\}^n$  of truth assignments that have distance at most  $\beta n$  from  $R(C)$ . Observe now that the set  $B(C) \setminus R(C)$  cannot contain any satisfying truth assignments, as any such assignment would be at distance  $\alpha n < d < \beta n$  from some assignment in  $C$ . Thus, the set of satisfying assignments in  $R(C)$  is a union of clusters (cluster-region), all of which have distance at least  $\beta n$  from any cluster not in the region. Repeating this process until all satisfying assignments have been assigned to a cluster region gives us exactly the subsets of Theorems 2 and 3 and note that that this argument actually bounds the diameter of each cluster-region by  $\alpha n$ , not just of each cluster.

It is straightforward to see that

**Remark 5.** *The arguments above remains valid even if assignments are deemed adjacent whenever their distance is bounded by  $f(n)$ , for any  $f(n) = o(n)$ . As a result, Theorems 2 and 3 remain valid as stated for any definition of clusters in which assignments are deemed to belong in the same cluster if their distance is  $o(n)$ .*

### 3.2 Establishing exponentially many clusters

Proving the existence of exponentially many non-empty cluster regions requires greater sophistication and leverages in a strong way the results of [3]. This is because having  $\Lambda(\alpha, k, r) > 1$  for some  $\alpha, k, r$  does *not* imply that pairs of satisfying assignments exist for such  $\alpha, k, r$ : in principle, the behavior of  $\Lambda$  could be determined by a tiny minority of solution-rich formulas. Hence the need for the second moment method [1, 3]. Specifically, say that a satisfying assignment is *balanced* if its number of satisfied literal occurrences is in the range  $km/2 \pm \sqrt{n}$ , and let  $X$  be the number of balanced assignments in  $F_k(n, rn)$ . In [3], it was shown that  $\mathbb{E}[X]^2 = \Lambda_b(1/2, k, r)^n$  and

$$\mathbb{E}[X^2] < C \times \max_{\alpha \in [0,1]} \Lambda_b(\alpha, k, r)^n, \quad (2)$$

for some explicit function  $\Lambda_b$  and constant  $C = C(k) > 0$ . It was also shown that for all  $r$  smaller than the r.h.s. of (1), the maximum of  $\Lambda_b$  occurs at  $\alpha = 1/2$ , implying that for such  $k, r$  we have  $\mathbb{E}[X^2] < C \times \mathbb{E}[X]^2$ . By the Paley-Zigmund inequality, this last fact implies that for any  $t \leq \mathbb{E}[X]$ ,

$$\Pr[X > t] \geq \frac{(\mathbb{E}[X] - t)^2}{\mathbb{E}[X^2]}. \quad (3)$$

Inequality (3) was applied with  $t = 0$  in [3], i.e., per the ‘‘second moment method’’, thus establishing that  $F_k(n, rn)$  has at least one (balanced) satisfying assignment with probability at least  $1/C$ . Since the probability of having at least one satisfying assignment, i.e., of being satisfiable, exhibits a sharp threshold [13] this implies that, in fact, for all  $r$  smaller than the r.h.s. of (1),  $F_k(n, rn)$  is satisfiable w.h.p.

We generalize the result of Friedgut [13] to prove that the probability that  $F_k(n, rn)$  has at least a certain number of satisfying assignments exhibits a sharp threshold. Combined with (3) this will allow us to prove that

**Theorem 6.** *For all  $k \geq 3$ , and all*

$$r < 2^k \ln 2 - \frac{(k+1) \ln 2 + 3}{2},$$

*w.h.p.  $F_k(n, rn)$  has at least  $[\Lambda_b(1/2, k, r) - o(1)]^{n/2}$  satisfying assignments.*

Armed with Theorem 6, we establish the existence of exponentially many clusters by dividing the lower bound it provides for the total number of satisfying assignments with the following upper bound for the number of truth assignments in each cluster-region. Recall that  $\Delta = \Delta_{k,r} \equiv \inf\{\alpha : \Lambda(\alpha, k, r) < 1\}$  and let

$$g(k, r) = \max_{\alpha \in [0, \Delta]} \Lambda(\alpha, k, r).$$

If  $B$  is the expected number of pairs of truth assignments with distance at most  $\Delta n$  in  $F_k(n, rn)$ , it follows that  $B < \text{poly}(n) \times g(k, r)^n$ , since the expected number of pairs at each distance is at most  $\Lambda(\alpha, k, r)^n$  and there are no more than  $n + 1$  possible distances. By Markov's inequality, this implies that w.h.p. the number of pairs of truth assignments in  $F_k(n, rn)$  that have distance at most  $\Delta n$  is  $\text{poly}(n) \times g(k, r)^n$ . Recall now that w.h.p. every cluster-region in  $F_k(n, rn)$  has diameter at most  $\Delta n$ . Therefore, w.h.p. the total number of pairs of truth assignments in each cluster-region is at most  $\text{poly}(n) \times g(k, r)^n$  and so the number of satisfying assignments in each cluster-region is at most  $\text{poly}(n) \times g(k, r)^{n/2}$ . Thus, if  $g(k, r) < \Lambda_b(1/2, k, r)$ , we can conclude that  $F_k(n, rn)$  has at least

$$\left( \frac{\Lambda_b(1/2, k, r) - o(1)}{g(k, r)} \right)^{n/2}$$

cluster-regions. Indeed, the higher of the two horizontal lines in Figure 1 highlights that  $g(8, 169) < \Lambda_b(1/2, 8, 169)$ .

Thus, we see that to establish Theorems 2 and 3 it suffices to prove the following analytical fact. We prove the claims regarding  $\alpha_k, \beta_k$  in Theorem 7 in Section 4, while in Section 5 we prove the claim regarding  $\epsilon_k$ .

**Theorem 7.** *For every  $k \geq 8$ , there exists a value of  $r < r_k$  and constants  $\alpha_k < \beta_k < 1/2$  and  $\epsilon_k > 0$  such that  $\Lambda(\alpha, k, r) < 1$  for all  $\alpha \in (\alpha_k, \beta_k)$  and*

$$\log_2 \left[ \left( \frac{\Lambda_b(1/2, k, r)}{g(k, r)} \right)^{1/2} \right] > \epsilon_k .$$

*In particular, for any  $0 < \delta < 1/3$  and all  $k \geq k_0(\delta)$ , if  $r = (1 - \delta)2^k \ln 2$ , we can take*

$$\alpha_k = \frac{1}{k} , \quad \beta_k = \frac{1}{2} - \frac{5}{6}\sqrt{\delta} , \quad \epsilon_k = \frac{\delta}{2} - 3k^{-2} . \quad (4)$$

Finally, we note that for  $r = (1 - \delta)r_k$ , where  $\delta \in (0, 1/5)$  and  $k \geq k_0(\delta)$ , it is possible to prove the existence of exponentially many clusters by leveraging the following result of [4] regarding the existence of frozen variables in random formulas (a variable is frozen in a cluster if it takes the same value in all truth assignments in the cluster).

**Theorem 8.** *For every  $k \geq 9$ , there exists  $c_k < r_k$  such that for all  $r \geq c_k$ , w.h.p. every cluster of  $F_k(n, rn)$  has at least  $(1 - 2/k) \cdot n$  frozen variables. As  $k$  grows,*

$$\frac{c_k}{2^k \ln 2} \rightarrow \frac{4}{5} .$$

To see how Theorem 8 implies the existence of exponentially many clusters, consider  $r$  and  $k$  such that  $c_k < r < r_k - \epsilon$ , for some  $\epsilon > 0$ . By Theorem 8, every cluster of  $F_k(n, rn)$  has  $(1 - 2/k) \cdot n$  frozen variables. Therefore, the probability that any given cluster will contain at least one satisfying assignment if we add another  $\zeta n$  random  $k$ -clauses to the formula is at most

$$\left[ 1 - \left( \frac{k-2}{2k} \right)^k \right]^{\zeta n} .$$

As a result, we see that unless  $F_k(n, rn)$  contains exponentially many clusters w.h.p., then for any  $0 < \zeta < \epsilon$ , the formula  $F_k(n, (r + \zeta)n)$  will be unsatisfiable w.h.p., a contradiction.

As the presence of  $\Omega(n)$  frozen variables implies the existence of  $2^{\Omega(n)}$  clusters by the above argument, it turns out we can establish clustering for densities lower than those in [4] for frozen variables. That said, recent numerical studies suggest that hardness in finding solutions is more probably connected to the existence of frozen variables than to the splitting of solutions in many clusters.

### 3.3 Related Work

The observation that if  $\Lambda(\alpha, k, r) < 1$ , then w.h.p.  $F_k(n, rn)$  has no pairs of satisfying assignments at distance  $\alpha n$  was first made in [21] and was related to “clustering”, even though there was no concrete definition of clusters or cluster-regions, the latter a seemingly necessary notion if one is to exploit the fact  $\Lambda(\alpha, k, r) < 1$ . More importantly, while the fact  $\Lambda(\alpha, k, r) < 1$  implies the absence of pairs of satisfying assignments at distance  $\alpha n$ , it falls far short of proving the existence of multiple clusters. In an attempt to show that there exist more than one clusters, in [21, 22] the authors derived an expression for the second moment of the number of *pairs of* balanced assignments at distance  $\alpha n$ , for each  $\alpha \in [0, 1]$ . If  $\alpha, k, r$ , are such that the dominant contribution to this second moment comes from uncorrelated pairs of pairs (of balanced assignments), this implies that with constant probability  $F_k(n, rn)$  contains at least one (balanced) pair of assignments at distance  $\alpha n$ . The authors further prove that the property “has a pair of satisfying assignments at distance  $q$ ” has a sharp threshold, thus boosting this constant probability to a high one.

Unfortunately, determining the dominant contribution to the above second moment for given  $\alpha, k, r$ , is a highly non-trivial problem. In particular, this “fourth moment” optimization problem is *much* harder than the already complicated second moment analysis of [3]. The authors address it numerically for small  $k$  (with no guarantee that the true maximizer has been found), and completely heuristically for general  $k$ , i.e., by simply guessing the locus of the local maximizer corresponding to correlated pairs and comparing it to the contribution of uncorrelated pairs. But even if the maximizer in this second moment computation could be determined rigorously and turned out to coincide with the numeric/heuristic estimate of [22], the strongest conclusion one could draw from these considerations is that for every  $k \geq 8$ , there is  $r < r_k$  and constants  $\alpha_k < \beta_k < c_k < 1/2 < d_k$ , such that in  $F_k(n, rn)$ :

- W.h.p. every pair of satisfying assignments has distance either less than  $\alpha_k n$  or more than  $\beta_k n$ .
- For every  $d \in [c_k, d_k] \cdot n$ , w.h.p. there is a pair of truth assignments that have distance  $d$ .

In particular, these two assertions above are completely consistent with the possibility that for every  $k \geq 8$ , w.h.p. the set  $\mathcal{S}(F_k(n, rn))$  consists of no more than *two* clusters.

In contrast, not only we *prove* that  $\mathcal{S}(F_k(n, rn))$  exhibits clustering, but that the number of clusters is *exponential*. Moreover, we give explicit, quantitative bounds for the diameter, the volume and the separation of these clusters.

## 4 The existence of cluster regions

In this section we prove the existence of  $\alpha_k, \beta_k$  as in Theorem 7. Let

$$\begin{aligned} h(x) &\equiv -x \ln x - (1-x) \ln(1-x) \\ &\leq \ln 2 - 2(1/2 - x)^2, \quad \text{for any } x \in [0, 1]. \end{aligned}$$

We begin by bounding  $\ln \Lambda$  from above as follows,

$$\begin{aligned} \ln \Lambda(\alpha, k, \gamma 2^k \ln 2) &= \ln 2 + h(\alpha) + \gamma 2^k \ln 2 \ln [1 - 2^{1-k} + 2^{-k}(1-\alpha)^k] \\ &< 2 \ln 2 - 2(1/2 - \alpha)^2 - \gamma \ln 2 [2 - (1-\alpha)^k] \\ &\equiv w(\alpha, k, \gamma). \end{aligned}$$

We note that for any fixed  $k, \gamma$ , the function  $w(\alpha, k, \gamma)$  is non-increasing in  $k$  and decreasing in  $\gamma$ . Moreover,

$$\frac{\partial^3 w}{\partial \alpha^3} = -\gamma \ln 2 k(k-1)(k-2)(1-\alpha)^{k-3} < 0, \quad (5)$$

implying that for any fixed  $k, \gamma$ , the equation  $w(\alpha, k, \gamma) = 0$  can have at most three roots for  $\alpha \in (0, 1)$ . To bound the location of these roots we observe that for any  $k \geq 8$  and  $\gamma > 2/3$ ,

$$w(0, k, \gamma) = (2 - \gamma) \ln 2 - \frac{1}{2} > 0, \quad (6)$$

$$w(1/2, k, \gamma) = [2 - (2 - 2^{-k})\gamma] \ln 2 > 0, \quad (7)$$

$$w(99/100, k, \gamma) < w(99/100, 8, 2/3) = -0.0181019... < 0, \quad (8)$$

where the inequality in (8) relies on the monotonicity of  $w$  in  $k, \gamma$ . Therefore, from (6)–(8) we can conclude that for every  $k \geq 8$  and  $\gamma > 2/3$ , if there exist  $\alpha_k, \beta_k \in (0, 1/2)$  such that  $w(\alpha_k, k, \gamma) < 0$  and  $w(\beta_k, k, \gamma) < 0$ , then  $\Lambda(\alpha, k, \gamma 2^k \ln 2) < 1$  for all  $\alpha \in [\alpha_k, \beta_k]$ . Below we first prove that such  $\alpha_k, \beta_k$  exist for all  $k \geq 8$  and then prove that for sufficiently large  $k$ , we can take  $\alpha_k, \beta_k$  as in (4).

- For  $k = 8$  it is enough to consider the plot of  $\Lambda(\alpha, 8, 169)$  in Figure 1. For  $k \geq 9$  we take  $\gamma = 0.985 > 2/3$ . Note that  $0.985 \cdot 2^k \ln 2$  is smaller than the lower bound for  $r_k$  given in (1), for all  $k \geq 9$ .

- We take  $\alpha_k = 1/k$ . We note that  $w(1/9, 9, 0.985) = -0.0451\dots < 0$  and prove that  $w(1/k, k, \gamma)$  is decreasing in  $k$  for any  $k \geq 4$  and  $\gamma < 1$  as follows,

$$\begin{aligned} \frac{\partial w(1/k, k, \gamma)}{\partial k} &= \gamma \ln 2 \left(1 - \frac{1}{k}\right)^{k-1} \left[ \frac{1}{k} + \left(1 - \frac{1}{k}\right) \ln \left(1 - \frac{1}{k}\right) \right] - \frac{4}{k^2} \left(\frac{1}{2} - \frac{1}{k}\right) \\ &< \gamma \ln 2 \left(1 - \frac{1}{k}\right)^{k-1} \frac{1}{k^2} - \frac{4}{k^2} \left(\frac{1}{2} - \frac{1}{k}\right) \\ &< \frac{1}{k^2} \left(\ln 2 - 2 + \frac{4}{k}\right) \\ &< 0 . \end{aligned}$$

- We take  $\beta_k = 3/8$ . We note that  $w(\alpha, k, \gamma)$  is non-increasing in  $k$  when  $\alpha$  and  $\gamma$  are fixed and that  $w(3/8, 9, 0.985) = -0.000520265\dots < 0$ .

- For the setting where  $r = (1 - \delta)2^k \ln 2$ , we will additionally use that  $-2x \ln 2 < \ln(1 - x) < -x$  for all  $0 < x < 1/2$  to establish that for any  $1 \leq c < k/2$ ,

$$\begin{aligned} \ln \Lambda(c/k, k, r) &= \ln 2 + h(c/k) + r \ln(1 - 2^{1-k} + 2^{-k}(1 - c/k)^k) \\ &< \ln 2 + (c/k)(\ln k + 2 \ln 2) - r(2^{1-k} - 2^{-k}(1 - c/k)^k) . \end{aligned} \quad (9)$$

Substituting  $r = \gamma 2^k \ln 2$  into (9) we get

$$\ln \Lambda(c/k, k, \gamma 2^k \ln 2) < \ln 2(1 - 2\gamma + \gamma e^{-c}) + (c/k)(\ln k + 2 \ln 2) . \quad (10)$$

- If  $c = 1$  and  $\gamma > \frac{1}{2-1/e} = 0.612\dots$ , then (10) implies that  $\ln \Lambda(1/k, k, \gamma) < 0$  for all sufficiently large  $k$ .
- If  $\gamma = (1 - \delta) > 2/3$ , then for any  $1 < \lambda \leq 3/(4 \ln 2) = 1.082\dots$

$$w(1/2 - \sqrt{\lambda \delta \ln 2}, k, 1 - \delta) = -2(\lambda - 1)\delta \ln 2 + (1 - \delta) \ln 2 \left(\frac{1}{2} + \sqrt{\lambda \delta \ln 2}\right)^k ,$$

which is negative for all sufficiently large  $k$ . The choice  $\beta_k = 1/2 - (5/6)\sqrt{\delta}$  corresponds to  $\lambda = (5/6)^2 / \ln 2 = 1.00187\dots$ , which is a valid value. For  $k$  large enough we have  $\alpha_k = 1/k < \beta_k = 1/2 - 5\sqrt{\delta}/6$  for any  $\delta \in (0, 1/3)$ .

## 5 The existence of exponentially many cluster regions

We will use the following two lemmata.

**Lemma 9.** *If  $\gamma \geq 49/50$  and  $k > 11$ , or  $\gamma \in (2/3, 1)$  and  $k > 15$ ,*

$$\ln g(k, \gamma 2^k \ln 2) \leq (1 - \gamma) \ln 2 + \left(1 + \frac{9 \ln 2}{16}\right) k^{-2} . \quad (11)$$

**Lemma 10.** For all  $k \geq 8$ ,

$$\ln \Lambda_b(1/2, k, \gamma 2^k \ln 2) \geq 2 \ln 2 [1 - \gamma m(k)] ,$$

where

$$m(k) = 1 + \frac{2k+3}{2} 2^{-k} + \frac{3k^2+6k-4}{2} 2^{-2k} + \frac{13k^2-12k+1}{2} 2^{-3k} \\ + (6k^3-13k^2+2k) 2^{-4k} + \frac{9k^4-24k^3+10k^2}{2} 2^{-5k} + (9k^4-6k^3) 2^{-6k} + \frac{9}{2} k^4 2^{-7k} .$$

Combining the two lemmata above we get that if  $r = \gamma 2^k \ln 2$  and either  $\gamma = 49/50$  and  $k > 11$ , or  $\gamma \in (2/3, 1)$  and  $k > 15$ , then

$$\log_2 \left[ \left( \frac{\Lambda_b(1/2, k, r)}{g(k, r)} \right)^{1/2} \right] > \frac{1}{2 \ln 2} \left[ \ln 2 (1 + \gamma - 2\gamma m(k)) - \left( 1 + \frac{9 \ln 2}{16} \right) k^{-2} \right] , \quad (12)$$

where  $m(k)$  is as in Lemma 10. It is not hard to check that  $m(k)$  is decreasing in  $k$ .

- For  $8 \leq k \leq 12$ , the existence of  $\epsilon_k > 0$  can be verified by plotting  $\Lambda$  and  $\Lambda_b$  and noting that

$$\Lambda_b(1/2, k, r) > \max_{\alpha \in [0, \Delta]} \Lambda(\alpha, k, r) ,$$

both when  $k = 8$  and  $r = 169$  and when  $9 \leq k \leq 12$  and  $r = 0.985 \cdot 2^k \ln 2$ . For  $k > 12$  and  $\gamma = 0.985$ , the existence of  $\epsilon_k > 0$  follows from the fact that the expression inside the square brackets in (12) is positive when  $k = 13$  and  $\gamma = 0.985$  and  $m(k)$  is decreasing in  $k$ .

- For the setting where  $r = (1 - \delta) 2^k \ln 2$ , we note that the limit of the expression inside the square brackets in (12) as  $k \rightarrow \infty$  is  $(1 - \gamma)/2$ . In particular, writing  $r = (1 - \delta) 2^k \ln 2$ , it is not hard to show that the right hand side of (12) is greater than  $\delta/2 - 3/k^2$  for all  $k \geq k_0(\delta)$ .

## 5.1 Proof of Lemma 9: The volume of the largest cluster

Below, we consider  $k$  and  $r$  to be fixed, so that all derivatives are with respect to  $\alpha$ . Specifically, we will give i) a value  $\alpha_M$  such that  $\Lambda$  is non-increasing in  $(\alpha_M, \alpha_k)$  and ii) a function  $u$  which is non-decreasing in  $[0, \alpha_M)$  and for which  $\Lambda(\alpha, k, r) \leq u(\alpha, k, r)$ . Thus, we will conclude  $g(k, r) \leq u(\alpha_M, k, r)$ .

We begin by getting an upper bound for  $\Lambda'$ , as follows:

$$\begin{aligned} \Lambda'(\alpha, k, r) &= -\ln \alpha + \ln(1 - \alpha) - r \frac{k(1 - \alpha)^{k-1}}{2^k + (1 - \alpha)^k - 2} \\ &\leq -\ln \alpha - \alpha - 2^{-k} r k (1 - \alpha)^{k-1} \\ &< -\ln \alpha - 2^{-k} r k (1 - \alpha)^{k-1} \\ &\leq -\ln \alpha - 2^{-k} r k (1 - k\alpha) \\ &\equiv \hat{u}(\alpha, k, r) . \end{aligned} \quad (13)$$

**Lemma 11.** If  $r = \gamma 2^k \ln 2$ , then for all  $k \geq 8$  and  $\gamma > 3k^{-1} \log_2 k$ , there exists

$$\alpha_M \leq 2^{-\gamma k} (1 + 4\gamma k^2 2^{-\gamma k} \ln 2) , \quad (14)$$

such that  $\hat{u}(\alpha_M, k, r) = 0$ .

*Proof of Lemma 11.* Let

$$q(\alpha) = 2^{-\gamma k} 2^{\gamma k^2 \alpha} .$$



We begin by noting that if  $\alpha_M$  is such that  $q(\alpha_M) = \alpha_M$  then  $\hat{u}(\alpha_M, k, r) = 0$ . Now, let us define

$$s(\alpha) = 2^{-\gamma k}(1 + 2\alpha\gamma k^2 \ln 2) .$$

Observe that the unique solution of  $s(\alpha) = \alpha$  is

$$\alpha^* = \frac{2^{-\gamma k}}{1 - 2\gamma k^2 2^{-\gamma k} \ln 2} \quad (15)$$

and that  $s(\alpha) > \alpha$  for all  $\alpha \in [0, \alpha^*)$ .

Recall that  $e^x \leq 1 + 2x$  for all  $0 \leq x \leq 1$ . Therefore,  $q(\alpha) < s(\alpha)$  for all  $\alpha$  such that  $\gamma k^2 \alpha \ln 2 \leq 1$ . In particular, if  $\gamma k^2 \alpha^* \ln 2 \leq 1$ , then since  $s(\alpha) > \alpha$  for all  $\alpha \in [0, \alpha^*)$ , we can conclude that the equation  $q(\alpha) = \alpha$  has at least one root  $\alpha_M \leq \alpha^*$ , as desired.

By (15), the condition  $\gamma k^2 \alpha^* \ln 2 \leq 1$  is equivalent to

$$\gamma k^2 2^{-\gamma k} \leq \frac{1}{3 \ln(2)} = 0.4808\dots \quad (16)$$

To establish that (16) holds we note that for any  $\gamma > 3k^{-1} \log_2 k$  the quantity  $\gamma k^2 2^{-\gamma k}$  is decreasing in  $\gamma$  and, therefore, it is bounded by  $z(k) = 3k^{-2} \log k$ . As  $z(k)$  is decreasing for  $k \geq 2$ , for all  $k \geq 8$  we have  $\gamma k^2 2^{-\gamma k} \leq z(8) = 9/64 = 0.1406\dots < 0.4808\dots$ , as desired. The fact  $\gamma k^2 2^{-\gamma k} \leq 0.1406\dots$  along with the inequality  $1/(1-x) \leq 1 + 2x$  valid for  $x \leq 1/2$ , gives us  $\alpha_M \leq \alpha^* \leq 2^{-\gamma k}(1 + 4\gamma k^2 2^{-\gamma k} \ln 2)$ .  $\square$

To bound  $\Lambda$  by an non-decreasing function we note

$$\ln \Lambda(\alpha, k, r) \leq \ln 2 - \alpha \ln \alpha + \alpha - r2^{-k}(1 + \alpha) \equiv u(\alpha, k, r) . \quad (17)$$

**Lemma 12.** *If  $r = \gamma 2^k \ln 2$ , then for every  $k \geq 8$  and  $\gamma \in (3k^{-1} \log_2 k, 1]$ ,*

$$u(\alpha_M, k, r) \leq (1 - \gamma) \ln 2 + \left(1 + \frac{9 \ln 2}{16}\right) k^{-2} .$$

*Proof.* Using Lemma 11 to pass from (18) to (19), we see that for every  $k \geq 8$  and  $\gamma \in (3k^{-1} \log_2 k, 1]$ ,

$$\begin{aligned} u(\alpha_M, k, r) &= \ln 2 + \alpha_M(\gamma k \ln 2 - \gamma k^2 \alpha_M \ln 2) + \alpha_M - \gamma \ln 2(1 + \alpha_M) \\ &\leq (1 - \gamma) \ln 2 + \alpha_M[1 + \gamma(k - 1) \ln 2] \end{aligned} \quad (18)$$

$$\leq (1 - \gamma) \ln 2 + 2^{-\gamma k}(1 + 4\gamma k^2 2^{-\gamma k} \ln 2)(\gamma k \ln 2 + 1) . \quad (19)$$

Recalling that (16) holds for all  $k \geq 8$  and  $\gamma > 3k^{-1} \log_2 k$ , we conclude

$$\begin{aligned} u(\alpha_M, k, r) &\leq (1 - \gamma) \ln 2 + k^{-3} \left(1 + \frac{9 \ln 2}{16}\right) (k \ln 2 + 1) \\ &\leq (1 - \gamma) \ln 2 + \left(1 + \frac{9 \ln 2}{16}\right) k^{-2} . \end{aligned}$$

$\square$

We can now prove Lemma 9.

*Proof of Lemma 9.* Recall the definition of the function  $u$  from (17) and note that, since  $u'(\alpha) = -\ln \alpha - r2^{-k}$ , it is non-decreasing for  $r \leq 2^k$  and  $\alpha \leq 1/e$ . From (14) we see that  $\alpha_M < 1/e$  and therefore we can conclude that  $\Lambda(\alpha, k, r) < u(\alpha_M, k, r)$  for all  $\alpha \in [0, \alpha_M)$ . To complete the proof it thus suffices to prove that  $\Lambda$  is non-increasing in the interval  $(\alpha_M, 1/k)$  since, by our results in the previous section, we know that  $\Delta \leq 1/k$  both when  $\gamma \geq 49/50$  and  $k > 11$ , and when  $\gamma \in (2/3, 1)$  and  $k > 15$ . For that we first observe that

$$\hat{u}'(\alpha, k, r) = -\frac{1}{\alpha} + 2^{-k} r k^2 < -\frac{1}{\alpha} + k^2 .$$

Since, by definition,  $\hat{u}(\alpha_M, k, r) = 0$  this implies  $\hat{u} \leq 0$  for all  $\alpha \in [\alpha_M, 1/k^2]$  and since  $\Lambda' \leq \hat{u}$ , it follows that  $\Lambda' \leq 0$  also for such  $\alpha$ . Using (13), it is straightforward to check that for  $\alpha \in [1/k^2, 1/k]$ , the derivative of  $\Lambda$  is negative both when i)  $\gamma \geq 49/50$  and  $k > 11$ , and when ii)  $2/3 < \gamma < 1$  and  $k > 15$ , thus concluding the proof.  $\square$

## 5.2 Proof of Lemma 10: A lower bound on the number of balanced assignments

*Proof of Lemma 10.* Recalling the definition of  $\Lambda_b$  from [3] we have

$$\ln \Lambda_b(1/2, k, r) = 2 \ln 2 + r \ln \left[ \frac{((1 - \epsilon/2)^k - 2^{-k})^2}{(1 - \epsilon)^k} \right], \quad (20)$$

where  $\epsilon$  satisfies

$$\epsilon(2 - \epsilon)^{k-1} = 1. \quad (21)$$

We note for later use that, as shown in [3], if  $\epsilon$  satisfies (21) then

$$2^{1-k} + k4^{-k} < \epsilon < 2^{1-k} + 3k4^{-k}. \quad (22)$$

Since all coefficients in the binomial expansion of  $(1 - \epsilon)^{-k}$  are positive,

$$(1 - \epsilon)^{-k} \geq 1 + k\epsilon + \frac{k(k+1)}{2}\epsilon^2. \quad (23)$$

To get a lower bound for the numerator inside the logarithm in (20) we consider the binomial expansion of  $(1 - \epsilon/2)^k$ . We observe that the sum of a pair of successive terms where the lower term corresponds to an even power equals

$$\binom{k}{j}(\epsilon/2)^j - \binom{k}{j+1}(\epsilon/2)^{j+1} = \binom{k}{j}(\epsilon/2)^j \left[ 1 - \frac{(k-j)\epsilon}{2(j+1)} \right]. \quad (24)$$

For  $k \geq 8, j \geq 4$  and  $\epsilon \leq 5/2$  the expression in (24) is positive. Moreover, when  $k$  is even the last term in the binomial expansion has a positive coefficient and can be safely discarded. Therefore, for all  $k \geq 8$  and  $\epsilon \leq 5/2$ ,

$$(1 - \epsilon/2)^k \geq 1 - \frac{k\epsilon}{2} + \frac{k(k-1)\epsilon^2}{8} - \frac{k(k-1)(k-2)\epsilon^3}{48}. \quad (25)$$

Substituting (23) and (25) into (20) we get a lower bound of the form  $\ln \Lambda_b \geq c_0 + c_1\epsilon + c_2\epsilon^2 \dots + c_8\epsilon^8$ . It is not hard to check directly that  $c_8 \geq 0$  for all  $k \geq 8$ . Similarly, using the upper bound for  $\epsilon$  from (22), it is not hard to check that for  $i = 2, 4, 6$ , we have  $c_i + c_{i+1}\epsilon \geq 0$  for all  $k \geq 8$ . Therefore, we can conclude

$$\begin{aligned} \ln \Lambda_b(1/2, k, r) &\geq 2 \ln 2 + r \ln [1 - 2^{1-k} + 2^{-2k} - \epsilon k 2^{-k}(1 - 2^{-k})] \\ &\geq 2 \ln 2 + r \ln [1 - 2^{1-k} + 2^{-2k} - k 2^{-k}(1 - 2^{-k})(2^{1-k} + 3k 2^{-2k})], \end{aligned} \quad (26)$$

where in (26) we have replaced  $\epsilon$  with its upper bound from (22).

The argument of the logarithm in (26) is increasing in  $k$  for all  $k \geq 3$  (a fact that can be easily established by considering its derivative). As a result, we have that for all  $k \geq 8$ , it is at least equal to its value for  $k = 8$  which is  $1 - 0.00805183\dots > 1/2$ . Thus, using the inequality  $\ln(1+x) > x - x^2$  valid for all  $x > -1/2$ , we can finally write

$$\ln \Lambda_b(1/2, k, \gamma 2^k \ln 2) \geq 2 \ln 2 [1 - \gamma m(k)],$$

where

$$\begin{aligned} m(k) = 1 + \frac{2k+3}{2}2^{-k} + \frac{3k^2+6k-4}{2}2^{-2k} + \frac{13k^2-12k+1}{2}2^{-3k} \\ + (6k^3-13k^2+2k)2^{-4k} + \frac{9k^4-24k^3+10k^2}{2}2^{-5k} + (9k^4-6k^3)2^{-6k} + \frac{9}{2}k^4 2^{-7k} \end{aligned} \quad (27)$$

□

## 6 Proof of Theorem 6

Recall that  $F_k(n, m)$  denotes a random  $k$ -CNF formula with  $n$  variables and  $m$  clauses. For a fixed number  $B > 1$  we let  $\mathcal{A}_B$  denote the property that a  $k$ -CNF formula  $F$  has fewer than  $\frac{1}{2}B^n$  satisfying assignments.

**Lemma 13.** *For any  $B > 1$  there is a sequence  $T_n^B$  such that for any  $\epsilon > 0$ ,*

$$\begin{aligned} \lim_{n \rightarrow \infty} \Pr(F_k(n, (1 - \epsilon)T_n^B) \text{ has property } \mathcal{A}_B) &= 0, \text{ and} \\ \lim_{n \rightarrow \infty} \Pr(F_k(n, (1 + \epsilon)T_n^B) \text{ has property } \mathcal{A}_B) &= 1. \end{aligned}$$

We will prove Lemma 13 shortly. Theorem 6 follows easily from 13, as follows.

*Proof of Theorem 6.* Equations (20) and (21) show that  $\rho \mapsto \Lambda_b(1/2, k, \rho)$  is a continuous function. Therefore, for every  $\epsilon > 0$  there is  $\delta > 0$  such that if  $r' = (1 + \delta)^2 r$ , then

$$\Lambda_b(1/2, k, r') > \Lambda_b(1/2, k, r) - \epsilon.$$

Fix  $\epsilon > 0$ , let  $r$  be smaller than the right hand side of (1) and let  $B = \sqrt{\Lambda_b(1/2, k, r')}$ . Taking  $t = \frac{1}{2}B^n$  in (3) and using (2) we obtain

$$\liminf_{n \rightarrow \infty} \Pr[F_k(n, r'n) \text{ does not satisfy } \mathcal{A}_B] > 0.$$

By Lemma 13, for all sufficiently large  $n$ , it follows that  $r'n < (1 + \delta)T_n^B$  and, thus,  $rn = (1 + \delta)^{-2}r'n < (1 + \delta)^{-1}T_n^B$ , implying

$$\lim_{n \rightarrow \infty} \Pr[F_k(n, rn) \text{ does not satisfy } \mathcal{A}_B] = 1.$$

Thus, w.h.p. the number  $Z$  of satisfying assignments of  $F_k(n, rn)$  satisfies

$$Z \geq \frac{1}{2}B^n = \frac{1}{2}\Lambda_b(1/2, k, r')^{n/2} \geq \frac{1}{2}(\Lambda_b(1/2, k, r) - \epsilon)^{n/2}.$$

Since this is true for any  $\epsilon > 0$ , the lemma follows.  $\square$

To prove Lemma 13, we introduce a bit of notation and build upon [13, Section 3.3].

Let  $\Phi$  be a CNF formula on variables  $y_1, \dots, y_l$  and let  $X = \{x_1, \dots, x_n\}$  be a disjoint set of  $n$  Boolean variables. We let  $\Phi_n$  denote the set of all  $\binom{n}{l}$  formulas that can result from  $\Phi$  by selecting  $l$  distinct variables from  $X$  and replacing them for  $y_1, \dots, y_l$  in  $\Phi$ . Given a CNF formula on  $X$ , we let  $F \oplus \Phi = F \wedge \Phi^*$ , where  $\Phi^*$  is chosen uniformly at random among all formulas in  $\Phi_n$ .

Note that  $\mathcal{A}_B$  is a *monotone* property, i.e., if  $F$  has the property  $\mathcal{A}_B$  and  $F'$  is another formula on the variables  $x_1, \dots, x_n$ , then  $F \wedge F'$  has the property  $\mathcal{A}_B$  as well. Therefore, we can use the following theorem from Friedgut [13] to prove by contradiction that  $\mathcal{A}_B$  has a sharp threshold. Let  $\omega(n) = \lceil \log n \rceil$  for concreteness.

**Theorem 14.** *Suppose that  $\mathcal{A}_B$  does not have a sharp threshold. Then there exist a number  $\alpha > 0$ , a formula  $\Phi$ , and for any  $n_0 > 0$  numbers  $N > n_0$ ,  $M > 0$  and a formula  $F$  with variables  $x_1, \dots, x_N$  such that all of the following hold:*

- T1.**  $\Pr(F \oplus \Phi \text{ has the property } \mathcal{A}_B) > 1 - \alpha$ .
- T2.**  $\alpha < \Pr(F_k(N, M) \text{ has the property } \mathcal{A}_B) < 1 - 3\alpha$ .
- T3.** *With probability at least  $\alpha$  a random formula  $F_k(N, M)$  contains an element of  $\Phi_N$  as a subformula.*
- T4.**  $\Pr(F \wedge F_k(N, 2\omega(N)) \text{ has the property } \mathcal{A}_B) < 1 - 2\alpha$ .

In the sequel we assume the existence of  $\alpha$ ,  $\Phi$ ,  $N$ ,  $M$ , and  $F$  satisfying conditions **T1–T3**. To conclude that  $\mathcal{A}_B$  has a sharp threshold, we shall show that condition **T4** cannot then hold. Clearly, we may assume that  $N$  is sufficiently large (by choosing  $n_0$  appropriately). Let  $V = \{x_1, \dots, x_N\}$ .

**Lemma 15.** *The formula  $\Phi$  is satisfiable.*

*Proof.* We shall prove that the probability  $Q$  that  $F_k(N, M)$  contains a subformula on  $l' \leq l$  variables without a pure literal is smaller than  $\alpha$ . This implies that with probability bigger than  $1 - \alpha$  the pure literal rule finds a satisfying assignment of any subformula on  $l$  variables; then the assertion follows from **T3**.

To estimate  $Q$ , we employ the union bound. Let  $k \leq l' \leq l$ . Any subformula on  $l'$  variables without a pure literal contains at least  $l'' = \lceil 2l'/k \rceil$  clauses. There are  $\binom{N}{l'}$  ways to choose a set of  $l'$  variables, and  $\binom{M}{l''}$  ways to choose slots for the  $l''$  clauses of the subformula. Furthermore, the probability that the random clauses in these  $l'$  slots contain only the chosen variables is at most  $(l'/N)^{kl''}$ . Hence, the probability that  $F_k(N, M)$  has  $l'$  variables that span a subformula with at least  $l''$  clauses is at most

$$Q \leq \binom{N}{l'} \binom{M}{l''} (l'/N)^{kl''} \leq \left( \frac{el'}{N} \cdot \left( \frac{ekM}{2l'} \right)^{2/k} \right)^{l'} \quad (28)$$

Further, **T2** implies that  $M/N \leq 2^k$ , because for  $M/N > 2^k$  the expected number of satisfying assignments of  $F_k(N, M)$  is less than 1. Thus, assuming that  $N$  is sufficiently large, we see that (28) implies  $Q < \alpha$ , as claimed.  $\square$

Now that we know that  $\Phi$  is satisfiable, let us fix a satisfying assignment  $\sigma : \{y_1, \dots, y_l\} \rightarrow \{0, 1\}$  of  $\Phi$ . We say that a satisfying assignment  $\chi$  of  $F$  is *compatible* with a tuple  $(z_1, \dots, z_l) \in V^l$  if  $\chi(z_i) = \sigma(y_i)$  for all  $1 \leq i \leq l$ . Furthermore, we call a tuple  $(z_1, \dots, z_l) \in V^l$  *bad* if  $F$  has fewer than  $\frac{1}{2}B^N$  satisfying assignments compatible with  $(z_1, \dots, z_l)$ .

**Lemma 16.** *There are at least  $(1 - \alpha)N^l$  bad tuples.*

*Proof.* The formula  $F \oplus \Phi$  is obtained by substituting  $l$  randomly chosen variables  $(z_1, \dots, z_l) \in V^l$  for the variables  $y_1, \dots, y_l$  of  $\Phi$  and adding the resulting clauses to  $F$ . Since by **T1** with probability at least  $1 - \alpha$  the resulting formula has at most  $\frac{1}{2}B^N$  satisfying assignments, a uniformly chosen tuple  $(z_1, \dots, z_l) \in V^l$  is bad with probability at least  $1 - \alpha$ . Thus, there are at least  $(1 - \alpha)N^l$  bad tuples.  $\square$

**Lemma 17.** *With probability at least  $1 - \alpha$ , a random formula  $F_k(N, \omega(N))$  contains  $l$  clauses  $C_1, \dots, C_l$  with the following two properties.*

**B1.** *For each  $1 \leq i \leq l$  there is a  $k$ -tuple of variables  $(v_i^1, \dots, v_i^k) \in V^k$  such that  $C_i = v_i^1 \vee \dots \vee v_i^k$  if  $\sigma(i) = 1$ , and  $C_i = \neg v_i^1 \vee \dots \vee \neg v_i^k$  if  $\sigma(i) = 0$ .*

**B2.** *For any function  $f : [l] \rightarrow [k]$  the  $l$ -tuple  $(v_1^{f(1)}, \dots, v_l^{f(l)})$  is bad.*

The proof of Lemma 17 is based on the following version of the Erdős-Simonovits theorem [12] (cf. [13, Proposition 3.5]).

**Theorem 18.** *For any  $\gamma > 0$  there are numbers  $\gamma' > 0$  and  $\nu_0 > 0$  such that for any  $\nu > \nu_0$  and any set  $H \subset [v]^l$  of size  $|H| \geq \gamma v^l$  the following is true. If  $l$   $k$ -tuples  $(w_1^1, \dots, w_1^k), \dots, (w_l^1, \dots, w_l^k) \in [v]^k$  are chosen uniformly at random and independently, then with probability at least  $\gamma'$  for any function  $f : [l] \rightarrow [k]$  the tuple  $(w_1^{f(1)}, \dots, w_l^{f(l)})$  belongs to  $H$ .*

*Proof of Lemma 17.* Assuming that  $N$  is sufficiently large, we apply Theorem 18 to  $\gamma = 1 - \alpha$ ,  $\nu = N$ , and the set  $H \subset [N]^l$  of bad  $l$ -tuples. Then by Lemma 16 we have  $|H| \geq \gamma N^l$ . Now, consider  $l$  random  $k$ -clauses  $C_1, \dots, C_l$  over the variable set  $V$  chosen uniformly and independently. Let  $V_1, \dots, V_l$  be the  $k$ -tuples of variables underlying  $C_1, \dots, C_l$ . Then Theorem 18 entails that  $V_1, \dots, V_l$  satisfy condition **B2** with probability at least  $\gamma'$ . Moreover, given that this is the case, condition **B1** is satisfied with probability  $2^{-kl}$ . Therefore, the clauses  $C_1, \dots, C_l$  satisfy both **B1** and **B2** with probability at least  $\gamma' 2^{-kl}$ . Hence, the probability that  $F_k(N, \omega(N))$  does not feature an  $l$ -tuple of clauses satisfying **B1** and **B2** is at most  $(1 - \gamma' 2^{-kl})^{\lfloor \omega(N)/l \rfloor}$ . Since  $\omega(N) = \lceil \log N \rceil$ , we can ensure that this expression is less than  $\alpha$  by choosing  $N$  large enough.  $\square$

**Corollary 19.** *With probability at least  $1 - \alpha$ , the formula  $F \wedge F_k(N, \omega(N))$  has at most  $\frac{1}{2}k^l \cdot B^N$  satisfying assignments.*

*Proof.* We will show that if  $C_1, \dots, C_l$  are clauses satisfying the two conditions from Lemma 17, then  $F \wedge C_1 \wedge \dots \wedge C_l$  has at most  $\frac{1}{2}k^l B^N$  satisfying assignments. Then the assertion follows from Lemma 17.

Thus, let  $\chi$  be a satisfying assignment of  $F \wedge C_1 \wedge \dots \wedge C_l$ . Then by the **B1** for each  $1 \leq i \leq l$  there is an index  $f_\chi(i) \in [k]$  such that  $\chi(v_i^{f_\chi(i)}) = \sigma(i)$ . Moreover, by **B2** the tuple  $(v_1^{f_\chi(1)}, \dots, v_l^{f_\chi(l)})$  is bad. Hence, the map  $\chi \mapsto f_\chi \in [k]^l$  yields a bad tuple  $(v_i^{f_\chi(i)})_{1 \leq i \leq l}$  for each satisfying assignment. Therefore, the number of satisfying assignments mapped to any tuple in  $[k]^l$  is at most  $\frac{1}{2}B^n$ . Consequently,  $F \wedge C_1 \wedge \dots \wedge C_l$  has at most  $\frac{1}{2}k^l \cdot B^n$  satisfying assignments in total.  $\square$

**Corollary 20.** *With probability at least  $1 - \frac{3}{2}\alpha$  the formula  $F \wedge F_k(N, 2\omega)$  satisfies  $\mathcal{A}_B$ .*

*Proof.* The formula  $F^{**} = F \wedge F_k(N, 2\omega)$  is obtained from  $F$  by attaching  $2\omega(N)$  random clauses. Let  $F^* = F \wedge F_k(N, \omega(N))$  be the formula resulting by attaching the first  $\omega(N)$  random clauses. Then by Corollary 19 with probability at least  $1 - \alpha$  the formula  $F^*$  has at most  $\frac{1}{2}k^l \cdot B^N$  satisfying assignments. Conditioning on this event, we form  $F^{**}$  by attaching another  $\omega(N)$  random clauses to  $F^*$ . Since for any satisfying assignment of  $F^*$  the probability that these additional  $\omega(N)$  clauses are satisfied as well is  $(1 - 2^{-k})^{\omega(N)}$ , the expected number of satisfying assignments of  $F^{**}$  is at most

$$\frac{1}{2}k^l \cdot B^N \cdot (1 - 2^{-k})^{\omega(N)} \leq \frac{\alpha}{4} \cdot B^N,$$

provided that  $N$  is sufficiently large. Therefore, Markov's inequality entails that

$$\Pr(F^{**} \text{ violates } \mathcal{A}_B | F^* \text{ has at most } \frac{1}{2}k^l \cdot B^N \text{ satisfying assignments}) \leq \alpha/2.$$

Thus, we obtain

$$\begin{aligned} \Pr(F^{**} \text{ violates } \mathcal{A}_B) &\leq \Pr(F^* \text{ has more than } \frac{1}{2}k^l \cdot B^N \text{ satisfying assignments}) \\ &\quad + \Pr(F^{**} \text{ violates } \mathcal{A}_B | F^* \text{ has at most } \frac{1}{2}k^l \cdot B^N \text{ satisfying assignments}) \leq 3\alpha/2, \end{aligned}$$

as desired.  $\square$

Combining Theorem 14 and Corollary 20, we conclude that  $\mathcal{A}_B$  has a sharp threshold, thereby completing the proof of Lemma 13.

**Acknowledgment.** We thank Ehud Friedgut for his helpful comments on [13], which have greatly facilitated the proof of Theorem 6.

## References

- [1] D. Achlioptas and C. Moore, *The asymptotic order of the random  $k$ -SAT threshold*, in Proc. 43th Annual Symposium on Foundations of Computer Science (2002), 126–127.
- [2] D. Achlioptas, A. Naor, and Y. Peres, *Rigorous location of phase transitions in hard optimization problems*, Nature **435** (2005), 759–764.
- [3] 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.
- [4] D. Achlioptas and F. Ricci-Tersenghi, *Random formulas have frozen variables*, Preprint (2007).
- [5] M. Alava, J. Ardelius, E. Aurell, P. Kaski, S. Krishnamurthy, P. Orponen, and S. Seitz, *Circumspect descent prevails in solving random constraint satisfaction problems*, arXiv:0711.4902v1, November 30th 2007.
- [6] M. Alekhovich and E. Ben-Sasson, *Linear Upper Bounds for Random Walk on Small Density Random 3CNFs*, in Proceedings of FOCS 2003, 352–361.

- [7] M.-T. Chao and J. Franco, *Probabilistic analysis of two heuristics for the 3-satisfiability problem*, SIAM J. Comput. **15** (1986), 1106–1118.
- [8] 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.
- [9] L. Dall’Asta, A. Ramezani, and R. Zecchina, *Entropy landscape and non-Gibbs solutions in constraint satisfaction problems*, Phys. Rev. E **77** (2008), 031118.
- [10] O. Dubois and Y. Boufkhad, *A general upper bound for the satisfiability threshold of random  $r$ -SAT formulae*, Journal of Algorithms **24** (1997), 395–420.
- [11] O. Dubois, Y. Boufkhad, and J. Mandler, *Typical random 3-SAT formulae and the satisfiability threshold*, Electronic Colloquium on Computational Complexity **10** (2003).
- [12] P. Erdős, M. Simonovits, *Supersaturated graphs and hypergraphs*. Combinatorica **7** (1987) 35–38.
- [13] E. Friedgut, *Hunting for sharp thresholds*, Random Structures Algorithms **26** (2005), 37–51.
- [14] 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.
- [15] A. Kamath, R. Motwani, K. Palem, and P. Spirakis, *Tail bounds for occupancy and the satisfiability threshold conjecture*, Random Structures & Algorithms **7** (1995), 59–80.
- [16] A. Kaporis, L. M. Kirousis, and E. G. Lalas, *The probabilistic analysis of a greedy satisfiability algorithm*, in Proc. 10th Annual European Symposium on Algorithms, volume 2461 of *Lecture Notes in Computer Science*, Springer (2002), 574–585.
- [17] A. Kaporis, L. M. Kirousis, and E. G. Lalas, *Selecting complementary pairs of literals*, in Proc. LICS’03 Workshop on Typical Case Complexity and Phase Transitions, 2003.
- [18] L. M. Kirousis, E. Kranakis, D. Krizanc, and Y. Stamatidou, *Approximating the unsatisfiability threshold of random formulas*, Random Structures & Algorithms **12** (1998), 253–269.
- [19] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborova, *Gibbs States and the Set of Solutions of Random Constraint Satisfaction Problems*, Proc. Natl. Acad. Sci. USA **104** (2007) 10318.
- [20] E. Maneva, E. Mossel, and M. J. Wainwright, *A New look at Survey Propagation and its Generalizations*, in Proc. of SODA 2005, 1089–1098.
- [21] M. Mézard, T. Mora, and R. Zecchina, *Clustering of Solutions in the Random Satisfiability Problem*, Phys. Rev. Lett. **94** (2005), 197205. Also [arxiv:cond-mat/0504070](https://arxiv.org/abs/cond-mat/0504070), April 4th 2005.
- [22] M. Mézard, T. Mora, and R. Zecchina, *Pairs of SAT Assignments and Clustering in Random Boolean Formulae*, [arxiv:cond-mat/0506053](https://arxiv.org/abs/cond-mat/0506053), June 2nd 2005.
- [23] M. Mézard, G. Parisi, and R. Zecchina, *Analytic and Algorithmic Solution of Random Satisfiability Problems*, Science **297** (2002), 812–815.
- [24] A. Montanari, F. Ricci-Tersenghi, and G. Semerjian *Clusters of solutions and replica symmetry breaking in random  $k$ -satisfiability*, J. Stat. Mech. (2008) P04004.
- [25] G. Parisi, *Some remarks on the survey decimation algorithm for  $K$ -satisfiability*, [arXiv:cs/0301015v1](https://arxiv.org/abs/cs/0301015v1), January 16th 2003.