The Resolution Complexity of Random Constraint Satisfaction Problems

Michael Molloy
Department of Computer Science
University of Toronto
and Microsoft Research
molloy@cs.toronto.edu

Mohammad Salavatipour
Department of Computer Science
University of Toronto
mreza@cs.toronto.edu

Abstract

We consider random instances of constraint satisfaction problems where each variable has domain size d, and each constraint contains t restrictions on k variables. For each (d, k, t) we determine whether the resolution complexity is a.s. constant, polynomial or exponential in the number of variables. For a particular range of (d, k, t), we determine a sharp threshold for resolution complexity where the resolution complexity drops from a.s. exponential to a.s. polynomial when the clause density passes a specific value.

1. Introduction

A constraint satisfaction problem (CSP) is a generalized form of satisfiability which is widely studied in the Artificial Intelligence community. For example, the journal *Constraints* is devoted to these problems. Roughly speaking, a CSP generalizes SAT in the sense that variables can draw their values from a more general domain than simply $\{T, F\}$, and each clause (a.k.a. constraint) forms a set of restrictions on the values that the variables in the clause may jointly take.

Random instances of k-SAT have been extremely well-studied over the past few decades (see [1] for many references). More recently, the interest in this area has expanded into random instances of various generalizations of k-SAT, such as NAE-SAT[3], XOR-SAT[12, 14, 15], $(2+p)$-SAT [27, 28, 29, 4, 2] and many others. All of these can be expressed as CSP's. It was natural for this interest to eventually spread to random instances of CSP's, rigorously in [5, 13, 26, 16, 24, 25, 31] and experimentally even earlier (see [19] for a good survey).

One of the most important results regarding random k-SAT is that of Chvatal and Szemeredi[11], who showed that for any $k \geq 3$ and $c > 0$, a random instance of k-SAT with n variables and cn clauses will almost surely (a.s.)[1] have no resolution proof of unsatisfiabilty of length less than $2^{\Theta(n)}$. It is easy to show that for large values of c, such random instances are almost surely unsatisfiable. This immediately implied that for sufficiently large values of c, any Davis-Putnam style algorithm will take exponential time on such an input. Furthermore, it provided an astoundingly vast and rich class to the, beforehand rather sparse[21, 30], list of unsatisfiable instances of k-SAT for which there is no polytime resolution proof of unsatisfiability. Such instances are of great interest since their existence can be viewed as a step towards proving that there are some unsatisfiable instances with no polytime proof of unsatisfiability of any kind, i.e. that $NP \neq co-NP$. Chvatal and Szemeredi's paper spawned numerous extensions and generalizations, eg. [7, 8, 2, 6], including a general framework for proving lower bounds on resolution complexity by Ben-Sasson and Wigderson[9].

Mitchell [24, 25] extended the framework of Ben-Sasson and Wigderson to the setting of CSP's. He then used this framework to prove exponential lower bounds on the resolution complexity of a very natural class of random CSP's - one where the number of restrictions per constraint is fixed. Specifically, he considered random CSP's with domain size $d \geq 2$, and every constraint containing precisely t restrictions on $k \geq 2$ variables.[2] Note that these CSP's are trivial if either d or k is equal to 1, and that they are the well-studied 2-SAT when $d = k = 2$. Mitchell showed that for $t \leq (d-1)/2, k = 2$ and for $t \leq d-1, k \geq 3$, and for any constant $c > 0$, such a random instance with cn constraints will almost surely have no subexponential proof of

[1] Formal definitions of these and other terms will appear in the next section

[2] This natural model was, historically, one of the first two random models of random CSP to be studied; the other turned out to be problematic and a.s. has $O(1)$ length resolution proofs of unsatisfiablity for any non-trivial number of constraints. See [26] or [19] for more details; the latter reference contains more than 30 references to the study of the model considered here.

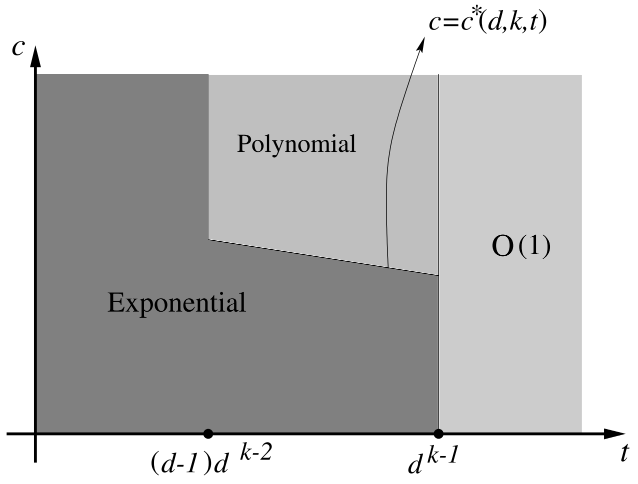

Figure 1. Resolution complexities

unsatisfiability.[3] Again, it is easy to see that for sufficiently large c, these instances are a.s. unsatisfiable (for $t > 0$). In contrast, Achlioptas et al[5] showed that for $t \geq d^{k-1}$, and any $c > 0$, such a random instance will a.s. have an unsatisfiable subproblem of size $O(1)$, and thus will have a $O(1)$-length resolution proof of unsatisfiability. In this paper, we fill in the gap between $d-1$ and d^{k-1}. Using $\mathcal{F}_{n,M}^{d,k,t}$ to denote such a random CSP with M clauses, we prove the following theorems which are summarized in Figure 1.

Theorem 1 *For any constants $d, k \geq 2$ and $1 \leq t < (d-1)d^{k-2}$, and for every constant $c > 0$, $\mathcal{F}_{n,M=cn}^{d,k,t}$ a.s. has resolution complexity at least $2^{\Theta(n)}$.*

For $d, k \geq 2$ and $t \geq (d-1)d^{k-2}$, we define

$$c^*(d,k,t) = \tfrac{1}{dk(k-1)} \binom{d^k}{t} / \binom{d^k - (d-1)d^{k-2}}{t - (d-1)d^{k-2}}.$$

(An explanation of the derivation of this expression will have to wait until the end of the next section - see Lemma 8.)

Theorem 2 *For any constants $d, k \geq 2$ and $(d-1)d^{k-2} \leq t < d^{k-1}$, and for every $c < c^*(d,k,t)$, $\mathcal{F}_{n,M=cn}^{d,k,t}$ a.s. has resolution complexity at least $2^{\Theta(n)}$.*

Theorem 3 *For any constants $d, k \geq 2$ and $(d-1)d^{k-2} \leq t < d^{k-1}$, and for every constant $c > c^*(d,k,t)$, $\mathcal{F}_{n,M=cn}^{d,k,t}$ a.s. has resolution complexity poly(n).*

[3] In [24], Mitchell claims to prove that this holds for $t \leq (d-1)(k-1)$ so long as d, k are not both 2. But there is an unfortunate error in his Lemmas 8 and 10, and his proof only holds for $t \leq (d-1)/2$, $k = 2$ and $t \leq d-1, k \geq 3$.

Trivially, the resolution complexity of a satisfiable CSP is infinite, so Theorem 2 is of no interest if for all $(d-1)d^{k-2} \leq t < d^{k-1}$ and $c < c^*(d,k,t)$, $\mathcal{F}_{n,M}^{d,k,t}$ is a.s. satisfiable. This is, in fact, well-known to be the case for $d, k = 2$ (i.e. 2-SAT), but it is easy to prove that it is not the case for all other d, k:

Theorem 4 *(a) For any constants $d, k \geq 2$ and for every $c > \ln d / \ln[d^k/(d^k - t)]$, $\mathcal{F}_{n,M=cn}^{d,k,t}$ is a.s. unsatisfiable.*

(b) For any constants $d, k \geq 2, d + k > 4$ and $(d-1)d^{k-2} \leq t < d^{k-1}$:

$$\frac{\ln d}{\ln[d^k/(d^k - t)]} < c^*(d,k,t).$$

As mentioned above, when $t \geq d^{k-1}$, the resolution complexity of $\mathcal{F}_{n,M=cn}^{d,k,t}$ is a.s. $O(1)$. So these theorems completely characterize the resolution complexity of $\mathcal{F}_{n,M=cn}^{d,k,t}$ for every constant d, k, t, c except for $(d-1)d^{k-2} \leq t < d^{k-1}$ and $c = c^*(d,k,t)$. Here we have a sharp threshold for resolution complexity, similar to that found in [2], where the main technical result was:

Theorem 5 *For any $\Delta, \epsilon > 0$, consider a random CNF-formula F on n variables with Δn 3-clauses and $(1-\epsilon)n$ 2-clauses where every such formula is equally likely. F a.s. has resolution complexity at least $2^{\Theta(n)}$.*

(The other side of the "sharp threshold", i.e. that if the number of 2-clauses is $(1+\epsilon)n$ for some $\epsilon > 0$ then a.s. the resolution complexity of F is poly(n), was previously known to follow from the work in [10].)

Theorems 2 and 3 are at heart very similar to Theorem 5. For $(d-1)d^{k-2} \leq t < d^{k-1}$, a certain type of constraint called a *forcer* arises. Forcers play, essentially, the same role that 2-clauses play in random CNF-formulas. We show that if $c > c^*$ then the forcers alone provide a unsatisfiable CSP with low resolution complexity, while if $c < c^*$ then, even along with the additional non-forcer constraints, the CSP has high resolution complexity.

Independently, Gao and Culberson [18] proved Theorem 3 for the special case $d = 2$. Essentially, they showed that in this case the forcers imply 2-clauses and for $c > c^*$ these 2-clauses for a random instance of 2-SAT which is above the satisfiability threshold. It is well-known that such an instance will have low resolution complexity. We remark more on this at the end of Subsection 2.3.

At first, we tried to adapt the lengthy proof of Theorem 5 to the setting of Theorem 2, but we were unsuccessful. Fortunately, we found an alternate proof technique, and to our pleasant surprise, it produced a proof of Theorem 2 which was dramatically shorter than the proof of Theorem 5 from [2]. In fact, our technique yields a short proof of Theorem

5 which we will provide in a full version of the paper. This technique looks like it will be of great value to those who wish to prove future similar theorems.

2. Preliminaries

Here we give formal definitions of some of the concepts discussed in the introduction, along with other concepts required for the remainder of the paper.

2.1. The random model

In our setting, the variables of our problem all have the same domain of permissable values, $\mathcal{D} = \{1, ..., d\}$, and all constraints will be on k variables, for some fixed integers $d, k \geq 2$. Given a k-tuple of variables, $(x_1, ..., x_k)$, a *restriction* on $(x_1, ..., x_k)$ is a k-tuple of values $R = (\delta_1, ...\delta_k)$ where each $\delta_i \in \mathcal{D}$. A set of restrictions on a k-tuple $(x_1, ..., x_k)$ is called a *constraint*. An assignment of values to the variables of a constraint C *satisfies* C if that assignment is not one of the restrictions in C. An assignment of values to all variables in a CSP satisfies that CSP if every constraint is simultaneously satisfied. A CSP is *satisfiable* if there is at least one such satisfying assignment. The *degree* of a variable is the number of constraints that it lies in.

A *subproblem* of a CSP $\mathcal{I}$ is a CSP which is obtained by removing some of the variables and some of the constraints from $\mathcal{I}$, where of course, if a variable x is removed then every constraint containing x is also removed. When there is no possibility of confusion we often use, for example, $\mathcal{I} - \{C_1, C_2\}$ to denote the subproblem obtained by deleting the constraints C_1, C_2 from $\mathcal{I}$ and $\mathcal{I} - \{x_1, x_2\}$ to denote the subproblem obtained by deleting the variables x_1, x_2 from $\mathcal{I}$, along with any constraints containing them.

Recall that a k-uniform hypergraph is a generalization of a graph, where each edge contains k vertices. The *constraint hypergraph* of a CSP is the k-uniform hypergraph whose vertices correspond to the variables, and whose edges correspond to the k-tuples of variables which have (non-empty) *constraints*. Of course, when $k = 2$, the constraint hypergraph is simply a graph, and so we often call it the *constraint graph*.

We define $\Omega^{d,k,t}$ to be the set of CSP's in which every variable has domain $\{1, ..., d\}$ and every constraint has k variables and t restrictions.

The Random Model: Specify M, n, d, k, t. First choose a random constraint hypergraph with n vertices and M edges of size k, where each such hypergraph is equally likely. Next, for each edge e, we choose a random constraint on the k variables of e, with domains $\mathcal{D} = \{1, ..., d\}$, uniformly from amongst all constraints with exactly t restrictions.

Note that every member of $\Omega^{d,k,t}$ with n variables and M clauses is equally likely to be chosen. We use $\mathcal{F}^{d,k,t}_{n,M}$ to denote a random CSP drawn from this model. We say that a property holds *almost surely* (a.s.) if the probability that it holds tends to 1 as n tends to infinity.

Remark: Alternatively, we could have chosen the constraint hypergraph by making an independent choice for each potential edge, deciding to put it in the hypergraph with probability $p = \frac{c \times k!}{n^{k-1}}$. The resulting random CSP is, in many senses, equivalent to the model described above. In particular, it is easy to show that all the theorems in this paper translate to this alternate model. We will make use of the equivalence of these models in the proofs of Lemmas 14 and 15.

2.2. Resolution complexity

For a boolean CNF-formula F, a *resolution refutation* of F with length r is a sequence of clauses $C_1, ..., C_r = \emptyset$ such that each C_i is either a clause of F, or is derived from two earlier clauses $C_j, C_{j'}$ for $j, j' < i$ by the following rule: $C_j = (A \vee x), C_{j'} = (B \vee \overline{x})$ and $C_i = (A \vee B)$, for some variable x. The *resolution complexity* of F, denoted $\mathbf{RES}(F)$, is the length of the shortest resolution refutation of F. (If F is satisfiable then $\mathbf{RES}(F) = \infty$.)

Mitchell[25] discusses two natural ways to extend the notion of resolution complexity to the setting of a CSP. These two measures of resolution complexity are denoted $\mathbf{C-RES}$ and $\mathbf{NG-RES}$. The latter appears on the surface to be the most natural extention in that it extends resolution rules to the setting of a CSP and then carries them out. $\mathbf{C-RES}$, on the other hand, converts a CSP to a boolean CNF-formula and then carries out CNF-resolution on that formula. Mitchell shows that for every CSP $\mathcal{I}$, $\mathbf{C-RES}(\mathcal{I}) \leq \text{poly}(\mathbf{NG-RES}(\mathcal{I}))$ whereas there are many choices for $\mathcal{I}$ for which the converse it not true. Furthermore, all commonly used resolution-type CSP algorithms correspond nicely to the $\mathbf{C-RES}$ complexity of the input, but there are some that do not correspond to the $\mathbf{NG-RES}$. For that reason, we focus in this paper on the $\mathbf{C-RES}$ complexity, as did Mitchell in [24].

Given an instance $\mathcal{I}$ of a CSP in which every variable has domain $\{1, ..., d\}$, we construct a boolean CNF-formula CNF($\mathcal{I}$) as follows. For each variable x of $\mathcal{I}$, there are d variables in CNF($\mathcal{I}$), denoted $x : 1, x : 2, ..., x : d$, and there is a *domain clause* $(x : 1 \vee ... \vee x : d)$. For each restriction $(\delta_1, ..., \delta_k)$ on variables $(x_1, ..., x_k)$ in any constraint of $\mathcal{I}$, CNF($\mathcal{I}$) has a *conflict clause* $(\overline{x_1 : \delta_1} \vee ... \vee \overline{x_k : \delta_k})$. It is easy to see that CNF($\mathcal{I}$) has a satisfying assignment iff $\mathcal{I}$ does - if $\mathcal{I}$ has a satisfying assignment, then we produce one for CNF($\mathcal{I}$) by setting $x : \delta$ to True iff $x = \delta$; if CNF($\mathcal{I}$) has a satisfying assignment, then we produce one for $\mathcal{I}$ by setting $x = \delta$ where δ is any one of the values for which

$x : \delta$ is True.

Remark: It is natural to consider adding an extra set of constraints for each variable x which specify that $x : \delta$ can be true for at most one value of δ. But it is easily verified that each of the results in this paper (in particular, Lemma 6) holds regardless of whether we include these clauses; to be specific, we do not include them.

We define the resolution complexity of $\mathcal{I}$, denoted $\mathbf{C}-\mathbf{RES}(\mathcal{I})$ to be equal to $\mathbf{RES}(\text{CNF}(\mathcal{I}))$.

In most previous papers bounding the resolution complexity of random instances of SAT or CSP for $k \geq 3$, a key lemma has been to establish that the following two conditions hold almost surely for some constants $\alpha, \zeta > 0$:

(A) Every subproblem on at most αn variables is satisfiable.

(B) Every subproblem on v variables where $\frac{1}{2}\alpha n \leq v \leq \alpha n$ has at least ζn variables of degree 1.

For SAT, these two facts imply that a.s. the resolution complexity is exponential in n using principles introduced in [10] and refined to easily applied tools in [8] and [9]. For more general instances of CSP, one needs to establish an additional fact (which is trivially true for SAT):

(C) If x is a variable of degree 1 in a CSP f then, letting f' be the subproblem obtained by removing x and its constraint, any satisfying assignment of f' can be extended to a satisfying assignment of f by assigning some value to x.

In our setting, (C) holds if $t < d$, but for $t \geq d$, it is easy to see that it fails: Suppose that the constraint x lies in contains the d restrictions: $(1,1,1,...,1,1), (1,1,1,...,1,2), ..., (1,1,1,...,1,d)$, where x is the last variable in the constraint. Then any satisfying assignment for f' in which all the other variables of the clause receive 1 cannot be extended to f. For $k \geq 3$ Mitchell's proof[24] applies precisely to the range of t for which (C) holds. For $k = 2$, Mitchell modifies the conditions, replacing "degree 1" by "degree 2" in (B) and (C); this revised condition (C) holds precisely when $t \leq (d-1)/2$.

For higher values of t, we need to replace "degree 1" in condition (B) by a more complicated notion, and then prove that something similar to condition (C) still holds. We describe how to do this in the next section, after presenting several necessary definitions.

2.3. Some new boundaries

A constraint C on variables $x_1, ..., x_k$ *forbids* $x_i : \delta$ if each of the d^{k-1} possible k-tuples $(\delta_1, ..., \delta_k)$ with $\delta_i = \delta$ is a restriction of C. Such a C is called a *forbidder*. As explained in [5] (and expanded on in [26]), it is the presence of forbidders that causes $\mathbf{C}-\mathbf{RES}(\mathcal{F}^{d,k,t}_{n,M=cn}) = O(1)$ a.s. for all c when $t \geq d^{k-1}$. C *permits* $(x_i : \delta, x_j : \gamma)$ if at least one of the d^{k-2} possible k-tuples $(\delta_1, ..., \delta_k)$ with $\delta_i = \delta$ and $\delta_j = \gamma$ is not a restriction of C. C is a $(x_i : \delta) \rightarrow (x_j : \gamma)$ *forcer* if C does not permit $(x_i : \delta, x_j : \gamma')$ for any $\gamma' \neq \gamma$; i.e. if each of the $(d-1)d^{k-2}$ possible k-tuples $(\delta_1, ..., \delta_k)$ with $\delta_i = \delta$ and $\delta_j \neq \gamma$ is a restriction of C. In this case, we say that the forcer C *starts* at x_i and *finishes* at x_j. As predicted by Mitchell[25], it is the presence of forcers that causes $\mathbf{C}-\mathbf{RES}(\mathcal{F}^{d,k,t}_{n,M=cn}) = \text{poly}(n)$ a.s. for large c when $(d-1)d^{k-2} \leq t < d^{k-1}$.

A *path* of length r in a k-uniform hypergraph H is a sequence of r edges $e_1, e_2, ..., e_r$ such that:

- for $1 \leq i \leq r-1$, $e_i \cap e_{i+1} = x_i$ - this is called a *connecting vertex*;
- for all $1 \leq i \leq r-2$ and $j > i+1$, $e_i \cap e_j = \emptyset$.
- there are specified vertices $x_0 \in e_1$ and $x_r \in e_r$, called the *endpoints* of the path.

A *pendant path* is a path in which no vertices other than the endpoints lie in any edges of H off the path. In other words, there is no restriction on the degrees of the endpoints, each connecting vertex has degree 2 in H, and every other vertex in the path has degree 1 in H.

A *(pendant) path* of length r in a CSP is a sequence of r constraints whose underlying edges form a (pendant) path of length r in the underlying hypergraph. If there are values $\delta_0, ..., \delta_r$ such that the constraint on each e_i is a $(x_{i-1} : \delta_{i-1}) \rightarrow (x_i : \delta_i)$ forcer, then we say that the (pendant) path is a $(x_0 : \delta_0) \rightarrow (x_r : \delta_r)$ *forcing (pendant) path*.

A constraint on the edge e_i of a path is a *P-forcer* if it is a $(x_{i-1} : \delta) \rightarrow (x_i, \gamma)$ forcer or a $(x_i : \gamma) \rightarrow (x_{i-1}, \delta)$ forcer for some δ, γ.

For any $\mathcal{I} \in \Omega^{d,k,t}$:

- The *first boundary* of $\mathcal{I}$, denoted by $\mathcal{B}^1(\mathcal{I})$, is the set of non-forbidding constraints of $\mathcal{I}$ which contain at most one variable of degree greater than 1.
- The *second boundary* of $\mathcal{I}$, denoted by $\mathcal{B}^2(\mathcal{I})$, is the set of pendant paths of length 4 in $\mathcal{I}$ which have no P-forcers.
- The *third boundary* of $\mathcal{I}$, denoted by $\mathcal{B}^3(\mathcal{I})$, is the set of pendant paths of length 2 where one of the two constraints is a P-forcer that starts at the connecting vertex, and the other is not a P-forcer that finishes at the connecting vertex.

The *boundary* of $\mathcal{I}$ is $\mathcal{B}(\mathcal{I}) = \mathcal{B}^1(\mathcal{I}) \cup \mathcal{B}^2(\mathcal{I}) \cup \mathcal{B}^3(\mathcal{I})$.

Our main lemma corresponds to conditions (A) and (B) from Section 2.2:

Lemma 6 *Consider any $\mathcal{I} \in \Omega^{d,k,t}$ on n variables, where $t < d^{k-1}$. If for some $\alpha, \zeta > 0$, we have:*

(a) every subproblem on at most αn variables is satisfiable, and

(b) *every subproblem $\mathcal{I}'$ on v variables where $\frac{1}{2}\alpha n \leq v \leq \alpha n$ has $|\mathcal{B}(\mathcal{I})| \geq \zeta n$.*

then $\mathbf{C}-\mathbf{RES}(\mathcal{I}) \geq 2^{\Theta(n)}$.

We require the following lemma, which corresponds to condition (C) from Section 2.2. With it, the proof of Lemma 6 is straightforward, following Mitchell's framework [24]; we omit the details.

Lemma 7 *Consider any $\mathcal{I} \in \Omega^{d,k,t}$, where $t < d^{k-1}$ and any $X \in \mathcal{B}(\mathcal{I})$. Any satisfying assignment of $\mathcal{I} - X$ can be extended to a satisfying assignment of $\mathcal{I}$.*

Proof: Suppose $X \in \mathcal{B}^1(\mathcal{I})$. Then the lemma follows from the fact that since $t < d^{k-1}$, X cannot be a forbidding constraint.

Suppose $X \in \mathcal{B}^2(\mathcal{I})$, and consider any satisfying assignment of $\mathcal{I} - X$ where x_0, x_4 are assigned δ_0, δ_4. ($\mathcal{I} - X$ is the subproblem obtained by removing all clauses of X and all variables of X other than the endpoints.) Since e_1 is not an (x_0, x_1)-forcer, there are at least two choices for δ_1 such that e_1 permits $(x_0 : \delta_0, x_1 : \delta_1)$. Similarly, there are at least two choices for δ_3 which can be assigned to x_3 so e_4 permits $(x_3 : \delta_3, x_4 : \delta_4)$. We will show that for at least one of these four choices for the pair (δ_1, δ_3), there is a value δ_2 such that e_2 permits $(x_1 : \delta_1, x_2 : \delta_2)$ and e_3 permits $(x_2 : \delta_2, x_1 : \delta_3)$. If this were not the case, then for every $\delta_2 \in \{1, ..., d\}$ either (i) e_2 does not permit $(x_1 : \delta_1, x_2 : \delta_2)$ for either of the two choices of δ_1 (this requires $2d^{k-2}$ restrictions) or (ii) e_3 does not permit $(x_2 : \delta_2, x_3 : \delta_3)$ for either of the two choices of δ_3 (this also requires $2d^{k-2}$ restrictions). Thus e_2, e_3 would have a total of at least $2d^{k-1}$ restrictions which is not possible by hypothesis.

Suppose $X \in \mathcal{B}^3(H)$. Let the endpoints of X be x_0, x_2, the connecting variable of X be x_1 and the constraints of X be C_1, C_2 where C_1 is a x_1, x_0-forcer and C_2 is not a x_2, x_1-forcer. Consider any satisfying assignment of $\mathcal{I} - X$ where x_0, x_2 are assigned δ_0, δ_2. There are at least $d-1$ choices for δ_1 such that C_1 permits $(x_0 : \delta_0, x_1 : \delta_1)$ and there are at least 2 choices for δ_1 such that C_2 permits $(x_1 : \delta_1, x_2 : \delta_2)$. At least one choice for δ_1 lies in the intersection of these sets and so the lemma follows. ∎

We close this section with a lemma explaining the significance of $c^*(d, k, t)$. We omit the easy proof.

Lemma 8 *For $d, k \geq 2$ and $t \geq (d-1)d^{k-2}$, let $c = c^*(d, k, t) = \frac{1}{dk(k-1)}\binom{d^k}{t}/\binom{d^k-(d-1)d^{k-2}}{t-(d-1)d^{k-2}}$. Specify any variable x and value $\delta \in \{1, ..., d\}$. The expected number of forcers in $\mathcal{F}^{d,k,t}_{n,M=cn}$ starting with $x : \delta$ is 1.*

It is instructive to consider the case $k = d = 2$ and $t = (d-1)d^{k-2} = 1$ which is the more familiar random 2-SAT. Here, every 2-clause can be viewed as the union of two forcers, eg. $(x_1 \vee x_2)$ is equivalent to the conjunction of the two forcers $(x_1 : F) \to (x_2 : T)$ and $(x_2 : F) \to (x_1 : F)$. (Of course, we are considering the domain to be {T(rue),F(alse)} rather than $\{1, 2\}$. Note that $c^* = 1$ which is the satisfiability threshold for 2-SAT. The reader who is familiar with random 2-SAT, will recognize that the property guaranteed by Lemma 8 corresponds very closely to what happens to cause the random 2-SAT to be unsatisfiable. Thus, it is not surprising that for general d, k, at $c > c^*$ the forcers alone produce an unsatisfiable formula and that it, like random 2-SAT, has small resolution complexity.

For $d = 2$ and general k, it is easy to see that an $(x : F) \to (y : T)$-forcer is also an $(y : F) \to (x : T)$ forcer. Thus, such a forcer implies the 2-clause $(x \vee y)$. Extending this reasoning shows that for $c > c^*$, the forcers alone will contain a random instance of 2-SAT where the number of 2-clauses is above the satisfiability threshold. As mentioned earlier, this was discovered independently by Gao and Culberson [18].

3. Proof of Theorem 1

We begin with a lemma of a type that has become standard in papers on the resolution complexity of random formulae. It says that a.s. every subproblem of $\mathcal{F}^{d,k,t}_{n,M}$ with at most αn variables has a very low clause-vertex ratio. Thus, to prove that the conditions of Lemma 6 hold, it suffices to prove that certain types of subproblems must have high clause-vertex ratio.

Lemma 9 *Let $c > 0$ and $k \geq 2$, and let H be the random k-uniform hypergraph with n vertices and $m = cn$ edges. Then for any $\delta > 0$, there exists $\alpha = \alpha(\sigma, k, \delta) > 0$ such that a.s. H has no subgraph with $0 < h \leq \lfloor \alpha n \rfloor$ vertices and at least $\left(\frac{1+\delta}{k-1}\right) h$ edges.*

The proof is straightforward, and we include it in the Appendix.

The next lemma will allow us to show that subproblems with small boundaries must have high clause-variable ratio, and hence, by the previous lemma, must be large.

Lemma 10 *Let $r \geq 2$ be a constant, H a k-uniform hypergraph on n vertices and m edges that does not have any component which is a cycle. Let B_1 be the set of edges which have at most one vertex of degree greater than 1, and B_2 be the set of pendant paths of length r. If $|B_1| + |B_2| \leq n/(72r^2k^3)$. Then for $\delta = \frac{1}{3rk^2}$: $m \geq n\left(\frac{1+\delta}{k-1}\right)$.*

Intuitively, the lemma is clear: If $B_1 = 0$ and if no vertex has degree greater than 2, then it is easy to see that we would have $m = n/(k-1)$. So in order for the average degree to not be much bigger, either B_1 must be large or there must be very few vertices of degree greater than 2. If the latter is true and B_1 is small, then H must contain long pendant paths and so B_2 will be big. Our formal proof is a bit lengthy and so we must omit it; it will appear in the full version of the paper.

These two lemmas are enough to prove the first of our main theorems:

Proof of Theorem 1: It suffices to prove that a.s. conditions (a) and (b) of Lemma 6 hold for $\mathcal{F}^{d,k,t}_{n,M=cn}$, where $\alpha = \alpha(c, k, \delta = 1/(12k^2)$ from Lemma 9 and $\zeta = \min(1/(72 \times 16k^3), \alpha/3)$.

Since $t < (d-1)d^{k-2}$, $\mathcal{F}^{d,k,t}_{n,M}$ has no forcer constraints. Using this fact, our proof would follow immediately from Lemmas 9 and 10, if it were not for the fact that Lemma 10 only applies to hypergraphs with no cycle components.

We begin with condition (a). Suppose that $\mathcal{J}$ is a minimally unsatisfiable subproblem of $\mathcal{F}^{d,k,t}_{n,M}$. Thus, the underlying hypergraph of $\mathcal{J}$ is connected. Furthermore, since $t < (d-1)d^{k-2}$, it is easily verified that the underlying hypergraph of $\mathcal{J}$ cannot be a single cycle. Finally, Lemma 7 implies that $|\mathcal{B}^1(\mathcal{J})| = |\mathcal{B}^2(\mathcal{J})| = 0$. Therefore, since $\mathcal{J}$ has no forcers, Lemma 9 with $r = 4$ applies to the underlying hypergraph of $\mathcal{J}$ and so $\mathcal{J}$ has clause-variable ratio at least $(1+\delta)/(k+1)$. Thus Lemma 10 implies that a.s. $\mathcal{F}^{d,k,t}_{n,M}$ has no minimally unsatisfiable subproblems of size at most αn. Therefore a.s. $\mathcal{F}^{d,k,t}_{n,M}$ has no unsatisfiable subproblems of size at most αn.

Next, condition (b). We will use the well-known fact that a.s. the underlying random hypergraph of $\mathcal{F}^{d,k,t}_{n,M}$ has fewer than $\log n$ cycles of length at most 4. Suppose, by contradiction, that $\mathcal{J}$ is a subproblem of $\mathcal{F}^{d,k,t}_{n,M}$ with v variables where $\frac{1}{2}\alpha n \le v \le \alpha n$, and with $|\mathcal{B}^1(\mathcal{J})| + |\mathcal{B}^2(\mathcal{J})| \le \zeta n$. (Since there are no forcers, $|B^3(\mathcal{J})| = 0$.) Let H' be the subhypergraph obtained by removing all the cycle components from the underlying hypergraph of $\mathcal{J}$. By Lemmas 9 and 10 (with $r = 4$), a.s. for every such subproblem $\mathcal{J}$, H' is empty, and so every component in the underlying hypergraph of $\mathcal{J}$ is a cycle. But any such cycle of length $l > 4$ will contain at least l members of $\mathcal{B}^2(\mathcal{J})$ and there are at most ζn vertices in those cycles. And, as mentioned above, a.s. there are at most $4 \log n$ vertices which lie in cycles of length at most 4 in $\mathcal{F}^{d,k,t}_{n,M}$. Since $\zeta n + 4 \log n < \frac{1}{2}\alpha n \le |\mathcal{J}|$, we have a contradiction.

■

4. Proof of Theorem 2

A Z_q *configuration* is a collection of q vertex-disjoint forcing paths in $\mathcal{I}$, each with possibly length zero (i.e. a single vertex), plus $q\left(\frac{1+\gamma}{k-1}\right)$ other edges, each containing k endpoints of the paths, where $\gamma = 1/(300k^2)$.

For $\mathcal{I} \in \Omega^{d,k,t}$ for $t \ge (d-1)k^{d-2}$, let $\mathcal{P} = p_1, \ldots, p_q$ be a collection of forcing paths of $\mathcal{I}$ such that: (i) every vertex lies in exactly one of these paths, (ii) it is not possible to transform $\mathcal{P}$ into a collection of $q-1$ paths meeting condition (i) by adding another forcer from $\mathcal{I}$. Obviously a collection of paths of length 0, one for each variable of $\mathcal{I}$, satisfies (i), and so some collection exists which satisfies (i) and (ii).

Lemma 11 *If $|\mathcal{B}(\mathcal{I})| < q/(72000k^3)$ and if the underlying hypergraph of $\mathcal{I}$ has no cycle components then $\mathcal{I}$ contains a Z_q configuration.*

Proof: Suppose that among the paths in $\mathcal{P}$, exactly $p_1, \ldots, p_r$ ($r \le q$) have at least one edge each; the others have length 0. Consider a path p_i, $(1 \le i \le r)$, and suppose it has $l \ge 1$ edges. Let x_0 and x_l be the start and end points of this path; so p_i is a $(x_0 : \delta_0) \to (x_l : \delta_l)$ forcer, for some values δ_0, δ_l. Remove from $\mathcal{I}$ all of the clauses and variables of p_i other than x_0, x_l. Then add $k-2$ new variables and a $(x_0 : \delta_0) \to (x_l : \delta_l)$ forcer. The new forcer is called a *reduced* forcer. We do this for every p_i, $1 \le i \le r$. The new CSP obtained after these operations is denoted by $\mathcal{I}'$. Note that n', the number of variables in $\mathcal{I}'$, is $q + r(k-1)$. Note also that since $p_1, ..., p_q$ are vertex-disjoint, no two reduced forcers in $\mathcal{I}'$ share a vertex.

Claim 12 *There is no forcer path of length at least 4 in $\mathcal{I}'$.*

Proof: By contradiction, assume that $p = e_1 e_2 e_3 e_4$ is a forcer path in $\mathcal{I}'$, with u_{i-1}, u_i being the start and end points of e_i. If e_2 is not a reduced forcer, then adding the forcer e_2 to $\mathcal{P}$ would concatenate the path in $\mathcal{P}$ containing u_1 with the one containing u_2 without violating condition (i). This contradicts condition (ii). If e_2 is a reduced forcer, then e_3 cannot be (since no two reduced forcers share a variable). Thus a similar contradiction arises when we consider adding the forcer e_3 to $\mathcal{P}$. ■

Claim 13 *Every pendant path of length 10 in $\mathcal{I}'$ has a subpath which is in $\mathcal{B}^2(\mathcal{I}) \cup \mathcal{B}^3(\mathcal{I})$.*

Proof: Let P' be a pendant path in $\mathcal{I}'$. By replacing each reduced edge of P' by its corresponding path from $\mathcal{P}$, we obtain a path P in $\mathcal{I}$. Every non-forcer in P' is a non-forcer in P. Assume $P = e_1, \ldots, e_l$ where $x_i = e_i \cap e_{i+1}$, and that e_a and e_b, $b > a$, are two non-forcers in P', and hence in P, such that there is no other non-forcer between them. If $b - a > 1$ then e_{a+1} must be a forcer starting at x_{a+1}, for otherwise $\{e_a, e_{a+1}\} \in \mathcal{B}^3(\mathcal{I})$ and we are done. Similarly, e_{b-1} must be a forcer starting at x_{b-2}, otherwise $\{e_{b-1}, e_b\} \in \mathcal{B}^3(\mathcal{I})$. But these two imply that along the path $e_{a+1}, \ldots, e_{b-1}$ there is a member of $\mathcal{B}^3(\mathcal{I})$. Thus, we can assume that $b - a = 1$, i.e. that the non-forcers in P' are consecutive. Also, if i is the largest index for which e_i

is a non-forcer in P, then $e_l e_{l-1} \ldots e_{i+1}$ must be a forcing path going into x_i, for otherwise there is a member of $\mathcal{B}^3(\mathcal{I})$ along this path. Therefore the portions of P' on the sides of these non-forcers form forcing paths. Similar arguments show that if P' has no non-forcers, then it contains at most two forcing paths starting at the end points of P', or else P' contains a member of $\mathcal{B}^3(\mathcal{I})$.

By Claim 12, the length of each forcing path is at most 3. Therefore, if P' has length 10, then P' has at least 4 consecutive non-forcers and so that subpath of P' is in $\mathcal{B}^2(\mathcal{I})$. ∎

Let B_1 be the set of clauses which have at most 1 variable of degree greater than 1 in $\mathcal{I}'$ and B_2 be the set of pendant paths of length 10 in $\mathcal{I}'$. Note that $|\mathcal{B}^1(\mathcal{I})| \geq |B_1|$. Since no subpath can lie in more than 10 members of B_2, Claim 13, implies that $|B_2| \leq 10(|\mathcal{B}^2(\mathcal{I})+\mathcal{B}^3(\mathcal{I})|)$. Therefore, $|B_1|+|B_2| \leq 10|\mathcal{B}(\mathcal{I})| \leq q/(7200k^3)$. So applying Lemma 10 with $r = 10$ to the underlying hypergraph of $\mathcal{I}'$, the number of clauses in $\mathcal{I}'$ is at least $n'\left(\frac{1+\gamma}{k-1}\right)$, and at least $n'\left(\frac{1+\gamma}{k-1}\right) - r \geq q\left(\frac{1+\gamma}{k-1}\right)$ are not reduced forcers. Those clauses and the paths in $\mathcal{P}$ will form a Z_q configuration in $\mathcal{I}$. ∎

Lemma 14 *For any constants $d, k \geq 2$ and $(d-1)d^{k-2} \leq t < d^{k-1}$, and for every $c < c^*(d,k,t)$, there exists $\sigma > 0$, such that a.s. $\mathcal{F}^{d,k,t}_{n,M=cn}$ has no Z_q configuration with $q \leq \sigma n$.*

Proof: For this proof, it will be convenient to work in the alternate model described in Section 2.1 where each of the $\binom{n}{k}$ potential edges is chosen for the constraint hypergraph with probability $ck!/n^{k-1}$. A standard argument shows that proving this model a.s. has no Z_q configuration with $q \leq \sigma n$ will imply that a.s. neither does $\mathcal{F}^{d,k,t}_{n,M=cn}$.

We compute the expected number of Z_q configurations. For each i, let $a_i \geq 0$ be the number of forcers in p_i. The number of ways to choose the end-points of the paths is at most n^{2q} (fewer if some of the path lengths are 0). For each p_i, there are at most n^{a_i-1} choices for the connecting variables, and d^{a_i+1} choices of values to use on the variables to form a forcing path. Also, for each edge of each p_i, there are there are $\binom{n}{k-2}$ choices for the set of degree 1 variables. We multiply by the probability that all these specified forcers exist for each p_i. We divide by the $q!$ ways of permuting the paths. There are at most $\binom{2q}{k}$ potential edges containing only endpoints of the paths, and we must choose $B = q\left(\frac{1+\gamma}{k-1}\right)$ of them. We multiply by the probability that all of these edges are present in H. Letting $A = \sum_{i=1}^{q} a_i$, and $c = (1-\epsilon)c^*$ and applying similar calculations to those in Lemma 8, this yields:

$$
\begin{aligned}
&\mathbf{E}(|Z_q|) \\
\leq\ & \frac{n^{2q}}{q!}\binom{\binom{2q}{k}}{B}\left(\frac{ck!}{n^{k-1}}\right)^B \\
&\times \sum_{a_1,\ldots,a_q \geq 0} \binom{n}{k-2}^A n^{A-q} d^{A+q}\left[\frac{(1-\epsilon)(k-2)!}{dn^{k-1}}\right]^A \\
\leq\ & \frac{n^{2q}}{q!}\left(\frac{(2q)^k e}{k!B}\frac{ck!}{n^{k-1}}\right)^B \\
&\times \sum_{a_1,\ldots,a_q \geq 0} \frac{n^{(k-2)A+A-q} d^{A+q}}{(k-2)!^A}\left[\frac{(1-\epsilon)(k-2)!}{dn^{k-1}}\right]^A \\
\leq\ & \frac{n^{2q}}{q!}\left(\frac{(2q)^k ec}{n^{k-1}(q/k-1)}\right)^{q(1+\gamma)/(k-1)} \\
&\times \frac{d^q}{n^q}\sum_{a_1,\ldots,a_q \geq 0}(1-\epsilon)^A \\
\leq\ & \left(\frac{\psi q}{n}\right)^{\gamma q}\left[\sum_{i \geq 0}(1-\epsilon)^i\right]^q \qquad \text{for some } \psi > 0 \\
\leq\ & \left(\frac{\psi' q}{n}\right)^{\gamma q} \qquad \text{for some } \psi' > 0.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sum_{q:|q|\geq 1}^{\sigma n} \mathbf{E}(|Z_q|) \\
\leq\ & \sum_{q:|q|=1}^{\lfloor \log n \rfloor}\left(\frac{\psi' q}{n}\right)^{\gamma q} + \sum_{q:|q|=\lfloor \log n \rfloor+1}^{\sigma n}\left(\frac{\psi' q}{n}\right)^{\gamma q} \\
\leq\ & O\left(\frac{\log^2 n}{n}\right) + \sum_{q:|q|=\lfloor \log n \rfloor+1}^{\sigma n} O(n^{-2}) = o(1),
\end{aligned}
$$

for sufficiently small $\sigma > 0$. ∎

Lemma 15 *For $c < c^*(d,k,t)$, there exists $\pi = \pi(c) > 0$ such that for every $\theta > 0$, a.s. the number of variables of $\mathcal{F}^{d,k,t}_{n,M=cn}$ that are part of a maximal forcing path of length at least θ is at most $e^{-\pi\theta}n$.*

The proof is a very standard branching process argument, of the sort that is used to prove well-known analagous results for random graphs[22], and random 2-SAT[1]. We omit the details.

Proof of Theorem 2: The proof is nearly identical to that of Theorem 1, this time using Lemmas 11 and 14 rather than Lemmas 10 and 9. The only significant difference is in proving that condition (b) of Lemma 6 holds:

Consider a subproblem $\mathcal{J}$ on $\frac{1}{2}\alpha n \leq v \leq \alpha n$ variables. Lemmas 11 and 14 only imply that $\mathcal{B}(\mathcal{J}) \geq \zeta q$ for some

$\zeta > 0$ where q is the number of paths in a minimal set of vertex-disjoint forcing paths which cover all the variables of $\mathcal{J}$. If the average path length in this set is sufficiently large, then our lower bound on the boundary size is not $\Theta(n)$ as required by Lemma 6. We use Lemma 15 to prove that this is not the case.

After specifying α, choose θ such that $e^{-\pi\theta} \leq \alpha/4$. Therefore, a.s. for every $\mathcal{J}$ on at least $\frac{1}{2}\alpha n$ variables, at least $\frac{1}{4}\alpha n$ of them do not lie on any forcer paths of length at least θ. Thus, any collection of forcer paths which covers all the variables of $\mathcal{J}$ must contain at least $\frac{\alpha}{4\theta}n$ paths. Hence, upon using Lemmas 11, 14 to prove that $\mathcal{B}(\mathcal{J}) \geq \zeta_1 q$ for some $\zeta_1 > 0$, we show that $\mathcal{B}(\mathcal{J}) \geq \zeta n$ for $\zeta = \zeta_1\alpha/4\theta > 0$. ■

5. Proof of Theorem 3

We will show that a.s. $\mathcal{F}_{n,M}^{d,k,t}$ contains a small unsatisfiable subproblem with a structure that is inspired by the snakes of [10]. Our proof is similar to the corresponding proof in [10].

A *forbidding cycle* is a $x : \delta \rightarrow x' : \delta'$ forcing path along with a $x' : \delta' \rightarrow x : \delta''$ forcer where $\delta \neq \delta''$. Thus, there is no satisfying assignment where $x = \delta$. We say that the cycle *forbids* $x : \delta$. Consecutive clauses in the cycle intersect in exactly one variable; such variables are called *connecting variables*.

An r*-flower* is the union of d forcing cycles $\mathcal{C}_1, \ldots, \mathcal{C}_d$ such that: (i) each has exactly r forcers; (ii) each cycle contains a particular variable x; (iii) no other variable lies in more than one of the cycles; (iv) each cycle C_i forbids $x : i$. We call x the *center variable*. Thus, any r-flower is unsatisfiable.

Lemma 16 *For any constants $d, k \geq 2, d+k > 4$ and $(d-1)d^{k-2} \leq t < d^{k-1}$, and for every constant $c > c^*(d,k,t)$, $\mathcal{F}_{n,M=cn}^{d,k,t}$ a.s. contains all constraints of an r-flower, where $r = \lambda \log n$, for some sufficiently large constant λ.*

Theorem 3 follows immediately since if $\mathcal{F}_{n,M}^{d,k,t}$ contains an r-flower with $r = \lambda \log n$, then we can use exhaustive search to prove that the r-flower, and hence $\mathcal{F}_{n,M}^{d,k,t}$, is unsatisfiable in $exp(r) = \text{poly}(n)$ steps. Such a proof can be simulated by a resolution proof with only a polynomial increase in length, so a.s. $\mathbf{RES}(\mathcal{F}_{n,M}^{d,k,t}) = \text{poly}(n)$. (The case $k = d = 2$, i.e. 2-SAT, is already known[10].)

Proof: For any potential flower A, we let X_A be the indicator variable for the event that the clauses of A all appear in the random CSP. With $X = \sum_A X_A$, it is enough to show that $\mathbf{E}(X^2) \leq \mathbf{E}(X)^2(1 + o(1))$. Then the theorem follows easily from the Chebyshev inequality.

First, we compute $\mathbf{E}(X)$. We must choose $s = dr - d + 1$ connecting variables (including the center variable) and the $(k-2)dr$ other variables. There are $s!$ ways to arrange the connecting variables and then $[(k-2)dr]!/(k-2)!^{dr}$ ways to arrange the other variables into the flower. For each C_i, we need to choose some value other than i for the center variable and an arbitrary value for each of the other connecting variables. Then we multiply by the probability of all our forcers being present. We have $c = (1+\epsilon)c^*$ for some $\epsilon > 0$. For fixed variables x, y and values $u, v \in D$, the probability that there exists a $(x\!:\!u) \longrightarrow (y\!:\!v)$ forcer is $p = (1+\epsilon)\frac{(k-2)!}{dn^{k-1}}$, by the same calculations as in Lemma 8.

$$\begin{aligned}
\mathbf{E}(X) &= \binom{n}{s} s! \binom{n-s}{(k-2)dr} \frac{[(k-2)dr]!}{(k-2)!^{dr}} \\
&\quad \times (d-1)^d d^{(r-1)d} p^{dr} \\
&= (1+o(1))(1+\epsilon)^{dr} \left(\frac{d-1}{d}\right)^r n^{1-d} \\
&= \Omega(n^2)
\end{aligned}$$

for sufficiently large λ. Next we compute an upper bound for $\mathbf{E}(X^2)$. Consider a fixed r-flower A and its underlying hypergraph H_A. For each $i, j \geq 1$ we will upper bound the number of potential r-flowers B that have exactly i constraints in common with A where these constraints form j connected components in H_A. (A very loose upper bound will suffice.) First, we consider choosing the j components. At most one component contains the center variable - for each C_i, such a component contains either all of C_i, none of C_i, or the portion of C_i between 2 variables. So there are at most $(2+r^2)^d$ choices for such a component. Each of the other components is specified by 2 variables on the same cycle. Thus, the number of choices for the components of H_A is at most $(2+r^2)^d(dr^2)^{j-1}$. To obtain a very loose upper bound on the number of ways that these components can fit into B, we simply multiply by the number of ways to choose j components from B and then multiply by $j!$ for the number of ways to pair them up with the components of A. Note that, since $t < d^{k-1}$ and d, k are not both 2, no constraint can be a $a : \delta \rightarrow b : \gamma$ forcer for more than one choice of a, b, δ, γ. Therefore, if an edge lies in the underlying hypergraph of both A and B, then its forcer in B must be identical to its forcer in A. Therefore, to choose the rest of B, we choose the remaining at most $s - i - j$ variables, a value for each of them (d values if one of them is the center variable) and then choose $k-2$ non-connecting variables for each of the remaining clauses. Thus, the total number of potential such r-flowers is at most $\left((2+r^2)^d(dr^2)^{j-1}\right)^2 j! n^{s-i-j} d^{s-i-j+d-1} \binom{n}{k-2}^{dr-i}$. Therefore, $\mathbf{E}(X^2)$ is:

$$
\begin{aligned}
&\sum_{A,B} \mathbf{Pr}(X_A = 1 \wedge X_B = 1) \\
= \;& \sum_{A} \sum_{B: A \cap B = \emptyset} \mathbf{Pr}(X_A = 1)\mathbf{Pr}(X_B = 1) \\
&+ \sum_{A} \sum_{B: H_A \cap H_B \neq \emptyset} \mathbf{Pr}(X_A = 1 \wedge X_B = 1) \\
< \;& \mathbf{E}(X)^2 + \sum_{A} \mathbf{Pr}(X_A = 1) \\
&\times \sum_{i=1}^{s} \sum_{j=1}^{i} \Big[\left((2+r^2)^d (dr^2)^{j-1}\right)^2 \\
&\times j! n^{s-i-j} d^{s-i-j+d-1} \binom{n}{k-2}^{dr-i} p^{dr-i} \Big] \\
= \;& \mathbf{E}(X)^2 + \mathbf{E}(X) \sum_{i=1}^{s} \Big[(2+r^2)^{2d} r^{-4} n^{s-i} d^{s-i+d-3} \\
&\times \binom{n}{k-2}^{dr-i} p^{dr-i} \sum_{j=1}^{i} \left(\frac{dr^4 j}{n}\right)^j \Big] \\
\leq \;& \mathbf{E}(X)^2 + \mathbf{E}(X) \times \mathbf{E}(X) \times O(r^{4d-4}) \\
&\times \sum_{i=1}^{s} (1+\epsilon)^{-i} O\left(\frac{r^4}{n}\right) \\
\leq \;& \mathbf{E}(X)^2 \left(1 + O\left(\frac{r^{4d}}{n}\right)\right),
\end{aligned}
$$

which completes the proof since $r = O(\log n)$. ■

6. Proof of Theorem 4

We close the paper with the short proof of Theorem 4. Part (a) uses a very standard technique whereby we compute the expected number of satisfying assignments. There are other standard techniques around which will improve this theorem (eg. the technique from [23]); we made no attempt to do so as our only goal was to show that $\mathcal{F}_{n,M=cn}^{d,k,t}$ is a.s. unsatisfiable for some $c < c^*(d,k,t)$.

Proof of Theorem 4:

Part (a): Consider any instance $\mathcal{I}$ chosen from $\mathcal{F}_{n,M=cn}^{d,k,t}$. There are d^n assignments to the variables of $\mathcal{I}$ and for each such assignment, the probability that all constraints are satisfied is

$$
\left(\frac{\binom{d^k - 1}{t}}{\binom{d^k}{t}} \right)^{cn} = \left(\frac{d^k - t}{d^k} \right)^{cn}.
$$

Therefore, the expected number of assignments that satisfy $\mathcal{I}$ is

$$
d^n \left(\frac{d^k - t}{d^k} \right)^{cn} = e^{(\ln d - c \ln(d^k/d^k - t))n},
$$

which is $o(1)$, if $c > \frac{\ln d}{\ln(d^k/d^k - t)}$. This implies that for such c, a.s. $\mathcal{I}$ is unsatisfiable.

Part (b): It is straightforward to verify that the statement holds for $d = 2$ and $3 \leq k \leq 5$. Also, $c^*(d,k,t) > \frac{1}{dk(k-1)} \times \left(\frac{d^k}{t}\right)^{(d-1)d^{k-2}} > \frac{d^{(d-1)d^{k-2}}}{dk(k-1)}$. So it is enough to show that for $d \geq 2$, $k \geq 6$, and for $d \geq 3$, $3 \leq k \leq 5$:

$$
\frac{\ln d}{\ln[d^k/(d^k - t)]} < \frac{d^{(d-1)d^{k-2}}}{dk(k-1)} \tag{1}
$$

A simple inductive argument shows that for $k \geq 6$: $k(k-1) < 2^k$ and $k < 2^{k-2} - 4$, which implies $2k(k-1) < 2^{2^{k-2}-3} \leq d^{(d-1)d^{k-2}-3}$. Similarly, for $d \geq 3$ and $k \geq 3$: $2k(k-1) < 3^k$ and $k \leq 2 \times 3^{k-2} - 3$, which implies $2k(k-1) < 3^{2 \times 3^{k-2} - 3} \leq d^{(d-1)d^{k-2}-3}$. Therefore, in both cases:

$$
\begin{aligned}
2dk(k-1)\ln d &< d^{(d-1)d^{k-2}-2} \ln d \\
\frac{2dk(k-1)\ln d}{d^{(d-1)d^{k-2}}} &< \frac{\ln d}{d^2} < \frac{(d-1)d^{k-2}}{d^k - (d-1)d^{k-2}}
\end{aligned} \tag{2}
$$

Using (2) and the fact that for $0 < x < 1$, $e^x < 1 + 2x$:

$$
\begin{aligned}
\exp\left(\frac{dk(k-1)\ln d}{d^{(d-1)d^{k-2}}}\right) &\leq 1 + \frac{(d-1)d^{k-2}}{d^k - (d-1)d^{k-2}} \\
&\leq \frac{d^k}{d^k - t} \\
\frac{dk(k-1)\ln d}{d^{(d-1)d^{k-2}}} &< \ln\left(\frac{d^k}{d^k - t}\right) \\
\frac{\ln d}{\ln[d^k/(d^k - t)]} &< \frac{d^{(d-1)d^{k-2}}}{dk(k-1)},
\end{aligned}
$$

thus establishing (1) as required.

■

Acknowledgements

We would like to thank 3 anonymous referees for suggestions that improved the presentation. We are grateful for the support of NSERC, a Sloan Research Fellowship, and a Premier's Research Excellence Award.

References

[1] D. Achlioptas. *Lower bounds for random 3-SAT via differential equations.* Theoretical Computer Science, 265 (1-2), (2001), p.159-185.

[2] D. Achlioptas, P. Beame and M. Molloy. *A sharp threshold in proof complexity.* Proceedings of STOC 2001, 337 - 346.

[3] D. Achlioptas, A. Chtcherba, G. Istrate and C. Moore. *The phase transition in NAESAT and 1-in-k SAT*. Proceedings of SODA 01, 721 - 722.

[4] D. Achlioptas, L. Kirousis, E. Kranakis, and D. Krizanc *Rigorous results for random $(2+p)$-SAT.* Proceedings of RALCOM '97 (1997), and Theoretical Computer Science, 265 (1-2), (2001), p.109-129.

[5] D. Achlioptas, L. Kirousis, E. Kranakis, D. Krizanc, M. Molloy, and Y. Stamatiou. *Random constraint satisfaction: a more accurate picture.* Constraints **6**, 329 - 324 (2001). Conference version in Proceedings of CP 97, 107 - 120.

[6] P. Beame, J. Culberson and D. Mitchell. *The resolution complexity of random graph k-colourability.* In preparation.

[7] P. Beame and T. Pitassi. *Simplified and improved resolution lower bounds.* Proceedings of FOCS 1996, 274 - 282.

[8] P. Beame, R. Karp, T. Pitassi and M. Saks. *The efficiency of resolution and Davis-Putnam procedures.* Proceedings of STOC 1998 and SIAM Journal on Computing, **31**, 1048 - 1075 (2002).

[9] E. Ben-Sasson and A. Wigderson. *Short proofs are narrow - resolution made simple.* Proceedings of STOC 1999 and Journal of the ACM **48** (2001)

[10] V. Chvátal and B. Reed. Mick gets some (the odds are on his side). In *Proceedings 33rd Annual Symposium on Foundations of Computer Science*, pages 620–627, Pittsburgh, PA, October 1992. IEEE.

[11] V. Chvatal and E. Szemeredi. *Many hard examples for resolution.* Journal of the ACM **35** (1988) 759 - 768.

[12] N. Creignou and H. Daude. *Satisfiability threshold for random XOR-CNF formulas.* Discrete Applied Mathematics, **96-97**, 4153 (1999)

[13] N. Creignou and H. Daude. *Random generalized satisfiability problems.* Proceedings of SAT 2002.

[14] N. Creignou, H. Daude and O. Dubois. *Approximating the satisfiability threshold of random k-XOR-formulas.* Preprint.

[15] O. Dubois and J. Mandler. *The 3-XORSAT threshold (strong evidence in favour of the replica method of statistical physics).* Proceedings of FOCS 2002.

[16] M. Dyer, A. Frieze and M. Molloy. *A probabilistic analysis of randomly generated binary constraint satisfaction problems.*

[17] W. Fernandex de la Vega. *On random 2-SAT.* manuscript (1992).

[18] Y. Gao and J. Culberson. *Resolution complexity of random constraint satisfaction problems: another half of the story.* Proceedings of LICS'03, Workshop on Typical Case Complexity and Phase Transitions.

[19] I. Gent, E. MacIntyre, P. Prosser, B. Smith and T. Walsh. *Random constraint satisfaction: flaws and structure.* Constraints **6**, 345 - 372 (2001).

[20] A. Goerdt. A threshold for unsatisfiability. *Journal of Computer and System Sciences*, 53:469–486, 1996.

[21] A. Haken. *The intractability of resolution.* Theoretical Computer Science **39**, 297 - 305 (1985).

[22] S. Janson, T. Łuczak and A. Ruciński. Random Graphs. Wiley, New York (2000).

[23] L. Kirousis, E. Kranakis, D. Krizanc and Y. Stamatiou. *Approximating the unsatisfiability threshold of random formulas.* Random Structures and Algorithms **12** (1998), 253 - 269.

[24] D. Mitchell. *Resolution complexity of random constraints.* Proceedings of Principles and Practices of Constraint Programming - CP 2002.

[25] D. Mitchell *The Resolution Complexity of Constraint Satisfaction.* Ph.D. Thesis, University of Toronto, 2002.

[26] M. Molloy, *Models for Random Constraint Satisfaction Problems*. Proceedings of STOC 2002, 209 - 217. Longer version to appear in SIAM J. Computing.

[27] R. Monasson and R. Zecchina. *Tricritical point in the random 2+p-SAT problem.* J. Phys. A **31**, 9209 (1998).

[28] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman and L. Troyansky. *2+p-SAT: Relation of typical-case complexity to the nature of the phase transition.* Random Structure and Algorithms **15**, 414 (1999).

[29] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman and L. Troyansky. *Phase transition and search cost in the 2+p-sat problem.* Proceedings of PhysComp 96, T. Toffoli, M. Biafore, J. Leao eds., Boston (1996).

[30] A. Urquhart. *Hard examples for resolution.* Journal of the ACM, **34** 209 - 219 (1987).

[31] K. Xu and W. Li. *Exact Phase transitions in random constraint satisfaction problems.* J. Artificial Intelligence Research, 12(2000):93-103.