The satisfiability threshold for randomly generated binary constraint satisfaction problems

Alan Frieze[1] and Michael Molloy[2]

[1] Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh PA15213, USA. ***
[2] Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4 and Microsoft Research, Redmond WA.

Abstract. We study two natural models of randomly generated constraint satisfaction problems. We determine how quickly the domain size must grow with n to ensure that these models are robust in the sense that they exhibit a non-trivial threshold of satisfiability, and we determine the asymptotic order of that threshold. We also provide resolution complexity lower bounds for these models.

1 Introduction

The *Constraint Satisfaction Problem* (CSP) is a fundamental problem in Artificial Intelligence, with applications ranging from scene labeling to scheduling and knowledge representation. See for example Dechter [12], Mackworth [18] and Waltz [25]. An instance of the CSP comprises a set of n *variables*, each taking a value in some given domain, and a set of *constraint relations*, each of which determines the permitted joint values of a given subset of the variables. The problem is either to determine any set of values for the variables which respects all the constraint relations, or determine that none exists. In recent years, there has been a strong interest in studying the relationship between the input parameters that define an instance of CSP (e.g. number of variables, domain sizes, tightness of constraints) and certain solution characteristics, such as the likelihood that the instance has a solution or the difficulty with which a solution may be discovered. An extensive account of relevant results, both experimental and theoretical, can be found in Hogg, Hubermann and Williams [15].

One of the most commonly used practices for conducting experiments with CSP is to generate a large set of random instances, all with the same defining parameters, and then for each instance in the set to use heuristics for deciding if a solution exists. Note that, in general CSP is NP-complete. The proportion of random instances that have a solution is used as an indication of the likelihood that an instance will be soluble, and the average time taken per instance (by some standard algorithm) gives some measure of the hardness of such instances. A characteristic of many of these experiments is that the fraction of assignments of values that are permissible for each constraint is kept *constant* as the number of variables increases. The very active experimental study of random models of CSP has necessitated a rigorous analysis of such models. Various models of random CSP's for which m, the domain-size, is constant have been studied in several papers, for example [2, 21, 11, 22, 23]. One of the earliest such studies, [2] discovered that the most natural models suffer a fatal flaw (described below). The first study of the case where m grows with n was [13], where one of these most natural models was studied. Implicit in that study was the fact that for certain settings of the relevant parameters, the fatal flaw did not occur and we had a rich random model to study. One the main contributions of this paper is to determine which parameter settings avoid that fatal flaw, and thus provide random models that are both natural and robust.

In this paper we consider only *binary* CSPs (BCSPs). These can be succinctly described in the following way: A graph $G = (V, E)$ is given, where $V = \{x_1, x_2, \ldots, x_n\}$ denotes the set of variables of the problem, and E the set of binary relations of the instance. We assume, without loss of generality, that each variable can take values in the same set $[m] = \{1, 2, \ldots, m\}$. For each

*** Supported in part by NSF grant CCR0200945. Research carried out during a visit to the Microsoft Research, Theory Group

edge $e = \{x_i, x_j\} \in E$, the relation can then be represented by an $m \times m$ 0-1 matrix M_e, where 0 indicates that the pair of values is forbidden and 1 that it is allowed. A solution to the associated BCSP is an assignment $f : V \to [m]$ of values to the variables, such that $M_e(f(x_i), f(x_j)) = 1$ for all $e = \{x_i, x_j\} \in E$.

The aim of this paper is to conduct a probabilistic analysis of some aspects of the following simple random models of BCSP:

Model A: The underlying graph G is G_{n,p_1} for some $p_1 = p_1(n) < 1$ where $p_1 \neq o(1/n)$. (This means that, with $V = \{x_1, x_2, \ldots, x_n\}$, we let each of the $\binom{n}{2}$ possible edges occur independently in E with probability p_1.) We let $d = np_1$. For each edge e of G there is a random $m \times m$ constraint matrix M_e where $M_e(i,j) = 1$ or 0 independently with probability p_2 or $q_2 = 1 - p_2$ respectively, for some constant $0 < p_2 < 1$. (In the final paper we will consider $p_2 \to 0$ and $p_2 \to 1$ as well.)

For $p_1 = o(1/n)$, the graph G_{n,p_1} is very sparse, and consists of a collection of small vertex-disjoint trees in which all but $o(n)$ of the vertices have degree 0. This is why we restrict our attention to $p_1 \neq o(1/n)$.

Given m, p_2 we wish to know: for what values of p_1 is our random CSP almost surely satisfiable? This question has been asked for many similar models of CSP, SAT and other problems. Traditionally, one of the first steps is to determine some values of p_1 for which it is not satisfiable as follows:

Fact: For $p_1 \geq \frac{2 \ln m}{q_2 n}$, the random CSP is unsatisfiable **whp**.

The proof follows easily from the fact that the expected number of satisfying solutions is $m^n \left(1 - p_1 q_2\right)^{\binom{n}{2}}$.

Inspired by a familiar pattern of similar random models, it is tempting to assume that $\frac{\ln m}{n}$ is the asymptotic order of a so-called "satisfiability threshold" and so hypothesize that:

Hypothesis A: There is some constant $c > 0$ so that for $p_1 \leq c\frac{\ln m}{n}$, the random CSP is satisfiable **whp**.

See [16] for a lengthy list of papers in which the authors fell to the temptation of assuming an equivalent hypothesis. In [2], it was observed that for most of those papers, and in fact whenever m, p_2 are both constants, the hypothesis is wrong. In fact, if $p_1 \geq \omega(n)/n^2$ for any $\omega(n)$ that tends to infinity with n, then almost surely the random CSP is trivially unsatisfiable in the sense that it has an edge whose constraint forbids every pair of values; we call such an edge a *blocked edge*

In this paper we asymptotically determine which values of m meet Hypothesis A.

Theorem 1. *(a) If $m \leq (1-\epsilon)\sqrt{\ln nd/\ln(1/q_2)}$ for some constant $\epsilon > 0$, then provided $nd \to \infty$, the random CSP has a blocked edge* **whp**

(b) If $m \geq (1+\epsilon)\sqrt{\ln nd/\ln(1/q_2)}$ then there is some constant $c > 0$ so that for $p_1 \leq c\frac{\ln m}{n}$, the random CSP is satisfiable **whp**. *Furthermore, an assignment can be found in $O(mn)$ time* **whp**.

For m, p_2 as in case (b), Hypothesis A holds, and so $\frac{\ln m}{n}$ is, indeed, the order of the satisfiability threshold. In case (a), **whp** the fact that the random CSP is unsatisfiable can be demonstrated easily by examining a single edge. We show that for $m \geq (\ln n)^{1+\epsilon}$ for any $\epsilon > 0$, this is far from the case. In particular, we show that **whp** there is. no short resolution proof of unsatisfiability when p_1 is of the same asymptotic order as the threshold of satisfiability.

Theorem 2. *If $m \geq (\ln n)^{1+\epsilon}$, $d = c\ln m$, for any constants $\epsilon, c > 0$, then* **whp** *the resolution complexity of the random CSP is $2^{\Omega(n/m)}$.*

The resolution complexity of various models of random boolean formula has been well-studied, starting with [10], and continuing through [4],[5],[3] and other papers. This line of inquiry was first extended to random models of CSP in [20, 19] and was then continued in [23]. In both of those studies, the domain-size was constant. Our Theorem 2 is the first result on the resolution complexity for a model of random CSP where the domain-size grows with n.

We now consider another model.

Model B: Here we generate a random $m \times m$ *symmetric matrix* M with density p_2 and put $M_e = M$ for every edge of $G = G_{n,p_1}$.

Theorem 3. *Let ϵ be a small positive constant, and consider a random CSP from Model B.*

(a) *If $d \leq (4-\epsilon)(\ln(1/q_2))^{-1} \ln m \ln\ln m$ then* **whp** *the CSP is satisfiable* **whp**.
(b) *If $d \leq (1-\epsilon)(\ln(1/q_2))^{-1} \ln m \ln\ln m$ then an assignment can be found in polynomial time* **whp**.
(c) *If $0 < q_2 < 1$ is constant and if $d \geq K \ln m \ln\ln m$ for sufficiently large K then* **whp** *the CSP is unsatisfiable.*

We can prove high resolution complexity in a restricted range of d, m, p_2.

Theorem 4. *If $m \to \infty$ and $d = c \ln m \ln\ln m$ for some constant $c > 0$, then* **whp** *the resolution complexity of a random CSP from Model B is $2^{\Omega(n/(d^3 m))}$.*

2 Some inequalities

We start with the basic Chernoff bounds for the binomial random variable $Bin(N, p)$ viz: Assume that $0 \leq \epsilon \leq 1$.

$$\mathbf{Pr}(Bin(N,p) \leq (1-\epsilon)Np) \leq e^{-\epsilon^2 NP/2}$$
$$\mathbf{Pr}(Bin(N,p) \geq (1+\epsilon)Np) \leq e^{-\epsilon^2 NP/3}$$

In Theorem 5 below we will have a random variable $Z = Z(Y_1, Y_2, \ldots, Y_N)$ where $Y_i \in \Omega_i$ are independent so that Z is defined on $\Omega = \Omega_1 \times \cdots \Omega_N$.

Assumption 1

Suppose that $Y, Y' \in \Omega$ and there exists i such that $Y_j = Y'_j$ for $j \neq i$. Our assumption is that in such a case we have $|Z(Y) - Z(Y')| \leq a$.

Assumption 2

Suppose that, in addition, for any ξ, if $Z(Y) \geq \xi$ then there exist $c(\xi)$ indices $j_1, j_2, \ldots, j_{c(\xi)}$ such that if $Y'_{j_t} = Y_{j_t}$ for $t = 1, 2, \ldots, c(\xi)$ then $Z(Y') \geq \xi$ also.

Let $M = MED(Z)$ denote a *median* of Z i.e. $\mathbf{Pr}(Z \geq M) \geq \frac{1}{2}, \mathbf{Pr}(Z \leq M) \geq \frac{1}{2}$.

Theorem 5. *(***Talagrand's Inequality***) If the random variable Z satisfies Assumptions 1 and 2 then*

$$\mathbf{Pr}(|Z - M| \geq tcM^{1/2}) \leq 2e^{-t^2/(4a^2)}. \tag{1}$$

Proofs of these inequalities can be found for example in Janson, Łuczak and Ruciński [17].

3 Model A: Unsatisfiable Region

3.1 Blocked Edges and Vertices

Let an edge $e = (x, y)$ of G be *blocked* if $M_e = \mathbf{O}$ (the matrix with all zero entries). Of course, any CSP with a blocked edge is unsatisfiable, since there is no possible consistent assignment to x, y. We start with a simple lemma:

Lemma 1. *Let $\epsilon > 0$ be a small positive constant and assume that $nd \to \infty$ (so that* **whp** *G has edges). Let $m_0 = \sqrt{(\ln n + \ln d)/\ln(1/q_2)}$. Then*

(a) *$m \geq (1+\epsilon)m_0$ implies that there are no blocked edges,* **whp**.
(b) *$m \leq (1-\epsilon)m_0$ implies that there are blocked edges,* **whp**.

Proof Let Z be the number of blocked edges in our instance. Given the graph G, the distribution of Z is $Bin(|E|, q_2^{m^2})$.

$$\mathbf{E}(Z) = \binom{n}{2} p_1 q_2^{m^2} \tag{2}$$

If $m \geq (1+\epsilon)m_0$ then (2) implies that

$$\mathbf{E}(Z) \leq (nd)^{-\epsilon} \to 0$$

and then $Z = 0$ **whp** and (a) follows.

If $m \leq (1-\epsilon)m_0$ then (2) implies that

$$\mathbf{E}(Z) \geq \frac{1}{3}(nd)^{\epsilon} \to \infty.$$

Part (b) now follows from the Chernoff bounds.

This proves Theorem 1(a). □

We now consider another simple cause of unsatisfiability that [2] also discovered to be prevalent amongst the models commonly used for experimentation. We say that a vertex (variable) x is *blocked* if for every possible assignment $i \in [m]$ there is some neighbour y which blocks the assignment of i to x, because the ith row of M_e, $e = (x, y)$ is all zero.

Lemma 2. *Let ϵ be a small positive constant, and suppose that $m - \sqrt{\ln n / \ln(1/q_2)} \to \infty$. Then*

(a) *$m \geq (1+\epsilon)\sqrt{(\ln n + m \ln d)/\ln(1/q_2)}$ implies that there are no blocked vertices,* **whp**.
(b) *$m \leq (1-\epsilon)\sqrt{(\ln n + m \ln d)/\ln(1/q_2)}$ implies that there are blocked vertices,* **whp**.

Remark: Note that $m = \sqrt{(\ln n + m \ln d)/\ln(1/q_2)}$, for m slightly smaller than m_0 from Lemma 1.

Proof If the graph G is given and vertex v has degree d_v then

$$\mathbf{Pr}(v \text{ is blocked} \mid G) = (1 - (1 - q_2^m)^{d_v})^m.$$

This is because for $i \in [m]$, $(1-q_2^m)^{d_v}$ is the probability that no neighbour w of v is such that row i of $M_{(v,w)}$ is all zero.

Part (a) now follows from an easy first moment calculation, which we omit. We turn our attention to proving part (b). Rearranging our assumption yields $\ln d \geq (1-\epsilon)^{-1}(m \ln(1/q_2) - \frac{1}{m}\ln n)$. So we choose d such that $\ln d = (1-\epsilon)^{-1}(m\ln(1/q_2) - \frac{1}{m}\ln n)$, i.e. $d = (q_2^{-m^2} n)^{1/m(1-\epsilon)}$ as proving the result for that value of d clearly implies that it holds for all larger values.

Our assumption implies that $d \to \infty$ and so **whp** $n - o(n)$ vertices v have $d_v \in I = [(1-\epsilon)d, (1+\epsilon)d]$. Thus if Z is the number of blocked vertices with $d_v \in I$ then

$$\begin{aligned}
\mathbf{E}(Z) &\geq (n - o(n))(1 - (1-q_2^m)^{d(1-\epsilon)})^m \geq (n-o(n))(d(1-\epsilon)q_2^m)^m \\
&\geq (1-o(1))\left(q_2^{-m^2} n\right)^{\epsilon/(1-\epsilon)}(1-\epsilon)^m \\
&\geq (1-o(1))n^{\epsilon/(1-\epsilon)}(1-\epsilon)^{m_0} \qquad \text{(see the Remark preceding this proof)} \\
&\geq n^{\epsilon/2} \to \infty.
\end{aligned}$$

To show that $Z \neq 0$ **whp** we use Talagrand's inequality (1). We condition on G. Then we let each $\Omega_e, e \in E$ be an independent copy of $\{0,1\}^{m^2}$ (the set of $m \times m$ 0-1 matrices). Now changing a single M_e can change z by at most 2 and so Assumption 1 holds with $a = 2$. Then to show that a vertex v is blocked we only have to expose M_e for e incident with v. Thus Assumption 2 holds with $c(\xi) = (1+\epsilon)d\xi$. Thus if $M = Med(Z)$, ((1) gives

$$\mathbf{Pr}(|Z - M| \geq t(1+\epsilon)dM^{1/2}) \leq 2e^{-t^2/16} \tag{3}$$

for any $t > 0$.

Our assumptions imply that $d^2 = o(\mathbf{E}(Z))$ and so (3) implies the result. □

4 Model A: Satisfiable Region

We assume for this section that

$$m = (1+\epsilon)\left(\frac{\ln n}{\ln q_2^{-1}}\right)^{1/2},\ d = c \ln m \text{ and } p_2 \text{ is constant}$$

where c, ϵ are small. (Note that this also implies the result for larger m).

Now let a vertex v be *troublesome* if it has degree $\geq D = 10d$ or there are assignments to its neighbours which leave v without a consistent assignment. Let $\mathcal{T}$ denote the set of troublesome vertices. A subset of $\mathcal{T}$ is called a troublesome set.

Let $\mathcal{A}$ be the event that every set of k_0 vertices contains at most k_0 edges where

$$k_0 = \left\lceil \frac{2\ln n}{d} \right\rceil.$$

Then

Lemma 3.

$$\mathbf{Pr}(\mathcal{A}) = 1 - o(1).$$

Proof

$$\mathbf{Pr}(\overline{\mathcal{A}}) \leq \binom{n}{k_0}\binom{\binom{k_0}{2}}{k_0+1}\left(\frac{d}{n}\right)^{k_0+1} \leq \left(\frac{ne}{k_0}\right)^{k_0} \cdot \left(\frac{d}{n}\right)^{k_0+1} \cdot \left(\frac{k_0 e}{2}\right)^{k_0+1}$$
$$= \frac{k_0 e^{2k_0+1} d^{k_0+1}}{2^{k_0+1}} \cdot \frac{d}{n} = o(1).$$

□

We show next that **whp** the sub-graph induced by $\mathcal{T}$ has no large trees.

Lemma 4. Whp *there are no troublesome trees with* $\geq k_0$ *vertices.*

Proof If $\mathcal{T}$ contains a tree of size greater than k_0 then it contains one of size k_0. Let Z be the number of troublesome trees with k_0 vertices. Let Ω be the set of trees/unicyclic graphs spanning $[k_0]$. Then for any subset J of $[k_0]$ we may write

$$\mathbf{E}(Z \cdot 1_{\mathcal{A}}) \leq \binom{n}{k_0} \sum_{T \in \Omega} \left(\frac{d}{n}\right)^{k_0-1} \prod_{i \in J} \mathbf{Pr}(x_i \in \mathcal{T} \mid \mathcal{G}_T, x_j \in \mathcal{T}, \forall j \in J, j < i). \tag{4}$$

Here $\mathcal{G}_T$ is the event that the sub-graph of G induced by$[k_0]$ is T.

Fix $T \in \Omega$ and let I_1 be the set of vertices of T with degree at most 4 in T. Then $|I_1| \geq k_0/2$. Note next that I_1 contains an independent set I of size at least $k_0/10$.

Now if $i \in I$ then

$$\mathbf{Pr}(x_i \in \mathcal{T} \mid \mathcal{G}_T, x_1, x_2, \ldots, x_{i-1} \in \mathcal{T}) \leq \binom{n}{D-4}\left(\frac{d}{n}\right)^{D-4} + \sum_{t=0}^{D} m^t(1-p_2^t)^m.$$

The first term bounds the probability that x_i has at least $D-4$ neighbours outside the tree and assuming the degree of x_i is at most D, the second term bounds the probability that the $\leq D$ neighbours have an assignment which can not be extended to x_i. We use the fact that I is an independent set to gain the stochastic independence we need.

Thus, applying (4) with $J = I$ we obtain

$$\mathbf{E}(Z \cdot 1_{\mathcal{A}}) \leq \binom{n}{k_0} k_0^{k_0-2} k_0^2 \left(\frac{d}{n}\right)^{k_0-1} \left(\binom{n}{D-4}\left(\frac{d}{n}\right)^{D-4} + \sum_{t=0}^{D} m^t(1-p_2^t)^m\right)^{k_0/10} \tag{5}$$
$$\leq n(de)^{k_0} \left(\left(\frac{de}{D-4}\right)^{D-4} + Dm^D e^{-mp_2^D}\right)^{k_0/10} = o(1).$$

□

Now we deal with troublesome cycles in a similar manner.

Lemma 5. Whp *there are no troublesome cycles.*

Proof It follows from Lemma 4 that we need only consider cycles of length k_0 or less. If Z now denotes the number of troublesome cycles of length k_0 or less then arguing as in (4), (5) we see that

$$\mathbf{E}(Z) \leq \sum_{k=3}^{k_0} \binom{n}{k} \frac{(k-1)!}{2} \left(\frac{d}{n}\right)^k \left(\binom{n}{D-2}\left(\frac{d}{n}\right)^{D-2} + \sum_{t=0}^{D} m^t(1-p_2^t)^m\right)^{\lfloor k/2 \rfloor} = o(1).$$

□

Let a tree be *small* if it contains at most k_0 vertices.

We have therefore shown that **whp** the troublesome vertices $\mathcal{T}$ induce a forest of small trees.

We show next that **whp** there at most $n^{1+o(1)}$ small trees.

Lemma 6. **Whp** *there are at most $n^{1+o(1)}$ small trees.*

Proof Let σ_T denote the number of small trees. Then

$$\mathbf{E}(\sigma_T) = \sum_{k=1}^{k_0} \binom{n}{k} k^{k-2} \left(\frac{d}{n}\right)^{k-1} \leq \sum_{k=1}^{k_0} n(de)^k = n^{1+o(1)}.$$

The result now follows from the Markov inequality. □

Our method of finding an assignment to our CSP is to (i) make a consistent assignment to the vertices of $\mathcal{T}$ first and then (ii) extend this assignment "greedily" to the non-troublesome vertices.

It is clear from the definition of troublesome that it is possible to carry out Step (ii). We wish to show that (i) can be carried out successfully **whp**. For this purpose we show that **whp** G does not contain a small tree which cannot be given a consistent assignment.

So we fix a small tree T and a vertex $v \in T$ and root T at v. Then let $X_i, 0 \leq i \leq k_0$ denote the vertices at distance i from v in T. Then let d_ℓ be the maximum number of descendants of a vertex in X_ℓ and let L denote the depth of T.

For $u \in X_\ell$ let $S_\ell(u)$ be the the set of values δ such that there is a consistent assignment to the sub-tree of T rooted at u in which u receives δ. We let $t = \lceil 10/\epsilon \rceil$ and define the events

$$\mathcal{B}_{u,i}^\ell = \left\{ \frac{(i-1)m}{t} \leq |S_\ell(u)| \leq \frac{im}{t} \right\}.$$

Then for $1 \leq i \leq t$ let

$$\pi_{i,\ell} = \max_{u \in X_\ell} \mathbf{Pr}\left(\bigcup_{j=1}^{i} \mathcal{B}_{u,j}\right).$$

Note that $\pi_{t,\ell} = 1$.

We claim that for $\ell > 1$,

$$\pi_{i,\ell} \leq \sum_{k_1+\cdots+k_t=d_\ell} \sum_{r=\frac{t-i}{t}m}^{\frac{t-i+1}{t}m} \binom{m}{r} \prod_{j=1}^{t} (1-(1-q_2^{\frac{j-1}{t}})^{k_j})^r \pi_{j,\ell-1}^{k_j} \tag{6}$$

$$\leq \sum_{j=1}^{t} t^{d_\ell} 2^m (1-(1-q_2^{\frac{j-1}{t}m})^{d_\ell})^{\frac{t-i}{t}m} \pi_{j,\ell-1}.$$

$$\leq \sum_{j=1}^{t} t^{d_\ell} 2^m (d_\ell q_2^{\frac{j-1}{t}m})^{\frac{t-i}{t}m} \pi_{j,\ell-1}. \tag{7}$$

Explanation of (6)**:** Suppose that there are k_j descendants w of u for which $\mathcal{B}_{w,j}^{\ell-1}$ occurs. If $u \in \mathcal{B}_{u,i}^\ell$ then r assignment values will be forbidden to it, $\frac{t-i}{t}m \leq r \leq \frac{t-i+1}{t}m$. The product bounds the probability that these values are forbidden and that $\mathcal{B}_{w,j}^{\ell-1}$ occurs for the corresponding descendants.

Now let us prove by induction on ℓ that for $\eta = \epsilon/3$ and for $1 \leq j \leq t$ we have

$$\pi_{j,\ell} \leq t^{\ell} n^{-(1+\eta)\frac{t-j}{t}}. \tag{8}$$

This is clearly true for $\ell = 0$ since $\pi_{j,0} = 0$ for $j < t$ and $\pi_{t,0} = 1$. Then from (7) we obtain

$$\pi_{i,\ell} \leq t^{\ell-1} \sum_{j=1}^{t} t^{d_\ell} 2^m d_\ell^{\frac{t-i}{t}m} q_2^{\frac{(j-1)(t-i)}{t^2}m^2} n^{-(1+\eta)\frac{t-j}{t}} \leq t^{\ell-1} \sum_{j=1}^{t} n^{-\frac{(j-1)(t-i)}{t^2}(1+\frac{\epsilon}{2})-\frac{t-j}{t}(1+\eta)}.$$

Notice that in going from the first to second inequality we use the fact that since $\ell, d_\ell \leq k_0$ we find that $2^m t^{d_\ell} d_\ell^{\frac{t-i}{t}m} = n^{o(1)}$. This term is then absorbed by using $1+\epsilon/2$ in place of $1+\epsilon$.

Now consider the expression

$$\Delta = \frac{(j-1)(t-i)}{t^2}(1+\frac{\epsilon}{2}) + \frac{t-j}{t}(1+\eta) - \frac{t-i}{t}(1+\eta) = \frac{(j-1)(t-i)}{t^2}(1+\frac{\epsilon}{2}) + \frac{i-j}{t}(1+\eta).$$

To complete the inductive proof of (8) we have only to show that it is non-negative.

Now Δ is clearly non-negative if $i \geq j$ and so assume that $j > i$. Now for a fixed j, Δ can be thought of as a linear function of i and so we need only check non-negativity for $i = 1$ or $i = j-1$.

For $i = 1$ we need

$$(j-1)(t-1)(1+\frac{\epsilon}{2}) \geq (j-1)t(1+\eta) \tag{9}$$

and this holds for $\epsilon \leq 1$.

For $i = j-1$ we need

$$(j-1)(t-j+1)(1+\frac{\epsilon}{2}) \geq t(1+\eta).$$

But here $j \geq 2$ and the LHS is at least $(t-1)(1+\frac{\epsilon}{2})$ and the inequality reduces to (9) (after dividing through by $j-1$). This competes the proof of (8). In particular

$$\pi_{1,k_0} \leq t^{k_0} n^{-(1+\eta)(t-1)/t}.$$

$$\mathbf{Pr}(\exists \text{a troublesome tree which cannot be consistently assigned}) \leq o(1) + n^{1+o(1)} t^{k_0} n^{-(1+\eta)(t-1)/t} = o(1)$$

which implies that Step (i) can be completed **whp**. This proves the satisfiability claim in Theorem 1(b).

It only remains to discuss the time to find an assignment. Once we have assigned values to $\mathcal{T}$ then we can fill in an assignment in $O(mn)$ time. So let us now fix a small tree T of troublesome vertices. Choose a root $v \in T$ arbitrarily. Starting at the lowest levels we compute the set of values $S_\ell(u)$ available to a vertex $u \in X_\ell$. For each descendant w of u we compute $T_\ell(w) = \{a \in S_{\ell+1}(w) : M_{(u,w)}(a) = 1\}$ and then we have $S_\ell(u) = \bigcap_w T_\ell(w)$. At the leaves, $S_L = [m]$ and so in this way we can assign a value to the root and then work back down the tree to the leaves giving an assignment to the whole of T. Thus the whole algorithm takes $O(mn)$ time as claimed. □

5 Model A: 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[20] 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 $\mathbf{C}-\mathbf{RES}$ and $\mathbf{NG}-\mathbf{RES}$. Here, our focus will be on the $\mathbf{C}-\mathbf{RES}$ measure, as it was in [19] and in [23].

Given an instance $\mathcal{I}$ of a CSP in which every variable has domain $\{1, ..., m\}$, we construct a boolean CNF-formula CNF($\mathcal{I}$) as follows. For each variable x of $\mathcal{I}$, there are m variables in CNF($\mathcal{I}$), denoted $x : 1, x : 2, ..., x : m$, and there is a *domain clause* $(x : 1 \vee ... \vee x : m)$. For each pair of variables x, y and each *restriction* (i, j) such that $M_{(x,y)}(i, j) = 0$, CNF($\mathcal{I}$) has a *conflict clause* $(\overline{x : i} \vee \overline{y : j})$. We also add $\binom{m}{2}$ 2-clauses for each x which specify that $x : i$ can be true for at most one value of i. It is easy to see that CNF($\mathcal{I}$) has a satisfying assignment iff $\mathcal{I}$ does. We define the resolution complexity of $\mathcal{I}$, denoted $\mathbf{C} - \mathbf{RES}(\mathcal{I})$ to be equal to $\mathbf{RES}(\text{CNF}(\mathcal{I}))$.

A variable x is *free* if any assignment which satisfies $\mathcal{I} - x$ can be extended to a satisfying assignment of $\mathcal{I}$. The *boundary* $\mathcal{B}(\mathcal{I})$ is the set of *free* variables. We extend a key result from [20] to the case where m grows with n:

Lemma 7. *Suppose that there exist $s, \zeta > 0$ such that*

(a) Every subproblem on at most s variables is satisfiable, and
(b) Every subproblem $\mathcal{I}'$ on v variables where $\frac{1}{2}s \le v \le s$ has $|\mathcal{B}(\mathcal{I}')| \ge \zeta n$.

then $\mathbf{C} - \mathbf{RES}(\mathcal{I}) \ge 2^{\Omega(\zeta^2 n/m)}$.

The proof is a straightforward adaptation of the proof of the corresponding work in [20] and so we omit it.

We assume now that ϵ is a small positive constant and

$$m \ge (\ln n)^{1+\epsilon},\ d = c \ln m \text{ and } p_2 \text{ is constant.} \tag{10}$$

Let γ be a sufficiently small constant. Let $\mathcal{T}_1$ denote the set of vertices v for which there are γd neighbours W and a set of assignments of values to W for which v has no consistent assignment.

Lemma 8.
$$\mathbf{Pr}(\mathcal{T}_1 \ne \emptyset) = o(1).$$

Proof

$$\begin{aligned}\mathbf{E}(|\mathcal{T}_1|) &\le n \sum_{t=\gamma d}^{n-1} \binom{n}{t} \left(\frac{d}{n}\right)^t \binom{t}{\gamma d} m^{\gamma d} (1 - p_2^{\gamma d})^m \le n \sum_{t=\gamma d}^{n-1} \left(\frac{de}{t}\right)^t \left(\frac{tem}{\gamma d}\right)^{\gamma d} e^{-mp_2^{\gamma d}} \\ &\le ne^{-m^{1-\epsilon/2}} \left(\sum_{t=\gamma d}^{10d} (de)^{10d} (10e\gamma^{-1} m)^{\gamma d} + \sum_{10d}^{n-1} (mn)^{\gamma d} \right) = o(1).\end{aligned}$$

□

Now we show that **whp** every set of $s \le s_0 = \alpha n$ vertices, $\alpha = \gamma/3$ has less than $\gamma ds/2$ edges. Let $\mathcal{B}$ denote this event.

Lemma 9.
$$\mathbf{Pr}(\mathcal{B}) = 1 - o(1).$$

Proof

$$\mathbf{Pr}(\overline{\mathcal{B}}) \le \sum_{s=\gamma d}^{\alpha n} \binom{n}{s} \binom{\binom{s_0}{2}}{\gamma ds/2} \left(\frac{d}{n}\right)^{\gamma ds/2} \le \sum_{s=\gamma d}^{\alpha n} \left(\left(\frac{se}{\gamma n}\right)^{-1+\gamma d/2} \cdot \frac{e^2}{\gamma} \right)^s = o(1).$$

□

Let us now check the conditions of Lemma 7. Condition (a) holds because Lemma 9 implies that if $s = |S| \le \alpha n$ then we can order S as $v_1, v_2, \ldots, v_s$ so that v_j has less than αd neighbours among $v_1, v_2, \ldots, v_{j-1}$ for $1 \le j \le s$. Because we can assume that $\mathcal{T}_1 = \emptyset$ (Lemma 8) we see that it will be possible to sequentially assign values to $v_1, v_2, \ldots, v_s$ in order. Lemma 9 implies that at least $\frac{1}{2}$ the vertices of S have degree $\le \alpha d$ in S and now $\mathcal{T}_1 = \emptyset$ implies that (b) holds with $\zeta = 1/2$.

We conclude that with the parameters as stated in (10), $\mathbf{C} - \mathbf{RES}(\mathcal{I})$ is **whp** as large as is claimed by Theorem 2.

6 Model B: Satisfiability

We have a blocked edge iff $M = \mathbf{O}$ and this happens with probability $q_2^{m(m-1)}$ and so there is not much more to say on this point.

Secondly, if $M \neq \mathbf{O}$ then there are two values x, y which can be assigned to adjacent vertices. This implies that for any bipartite subgraph H of G there is a satisfying assignment for H just using x, y. So, in particular there will be no blocked vertices.

Let us now consider Theorem 3. Let H be the graph defined by treating M as its adjacency matrix. Thus $H = G_{m,p_2}$. As such it has a clique I of size $(2 - o(1)) \ln m/(\ln 1/q_2)$.

If we can properly colour G with I (i.e. give adjacent vertices different values in I) then we will have a satisfying assignment for our CSP. Now the chromatic number of G is $(1 + o(1))d/(2 \ln d)$ **whp**. So the CSP is satisfiable **whp** if

$$(2 - o(1)) \ln m/(\ln 1/q_2) \geq (1 + o(1))d/(2 \ln d)$$

and this holds under assumption (a).

For (b) we observe that we can find a clique of size $(1 - o(1)) \ln m/(\ln 1/q_2)$ in polynomial time and we can colour G with $(1 + o(1))d/ \ln d$ colours in polynomial time.

We now prove part (c) of Theorem 3. We first observe

Lemma 10. *There exists a constant ϵ_0 such that for $\epsilon \leq \epsilon_0$ there exist $R_0 = R_0(\epsilon), Q_0 = Q_0(\epsilon)$ such that if $Q \geq Q_0, R \geq R_0$ and $s_0 = R \ln m$ then*

(a) **whp** *every pair of disjoint sets $S_1, S_2 \subseteq [m]$, $|S_1| = s_1 \geq s_0, |S_2| = s_2 \geq s_0$ contains at most $(1 - \epsilon)s_1 s_2$ $S_1 : S_2$ edges of H;*

(b) **whp** *every $S \subseteq [m]$, $|S| = s \geq s_0$ contains at most $Q \ln m$ members with degree greater than $(1 - \epsilon)s$ in the subgraph of H induced by S.*

Proof

(a) We can bound the probability that there are sets S_1, S_2 with more than the stated number of $S_1 : S_2$ edges by

$$\sum_{s_1=s_0}^{m} \sum_{s_2=s_0}^{m} \binom{m}{s_1}\binom{m}{s_2}\binom{s_1 s_2}{\epsilon s_1 s_2} p_2^{(1-\epsilon)s_1 s_2}$$
$$\leq \sum_{s_1=s_0}^{m} \sum_{s_2=s_0}^{m} \left(\frac{me}{s_1}\right)^{s_1} \left(\frac{me}{s_2}\right)^{s_2} \left(\left(\frac{e}{\epsilon}\right)^{\epsilon} p_2^{1-\epsilon}\right)^{s_1 s_2} = o(1).$$

(b) We choose $\epsilon > 0$ so that $p_2 < 1 - 3\epsilon$. Given S, we consider a set $L \subset S$ of size $Q \ln m$. For $R > Q\epsilon^{-1}$ we have $|L| < \epsilon|S|$ and so if each $i \in L$ has at least $(1 - \epsilon)s$ neighbours in S then it has at least $(1 - 2\epsilon)s$ neighbours in $S - L$. By the Chernoff bound, this occurs with probability at most $\left(e^{-\zeta s}\right)^{|L|}$, for some $\zeta > 0$ and this is less than m^{-2s} for Q sufficiently high. Therefore, the expected number of S, L violating part (b) is at most

$$\sum_{s=s_0}^{m} \binom{m}{s}\binom{s}{Q \ln m} m^{-2s} < \sum_{s=s_0}^{m} \left(\frac{em}{s}\right)^s 2^s m^{-2s} < \sum_{s \geq s_0} m^{-s} = o(1).$$

□

Now consider an assignment σ for our CSP and let N_i be the set of variables that are assigned the value i by σ. We observe that if σ is consistent then each N_i is an independent set in G and so **whp** G is such that we must have

$$|N_i| \leq \frac{3n \ln d}{d} < \frac{4n}{K \ln m} \qquad \text{for } i = 1, 2, \ldots, m. \tag{11}$$

Thus, we will restrict our attention to assignments which satisfy (11). We will prove that the expected number of such assignments that are consistent is $o(1)$, thus proving part (c) of Theorem 3.

We say that a pair of vertices is *forbidden* by σ if that pair cannot form an edge of G without violating σ. Note that every pair in the same set N_i is forbidden, and a pair in $N_i \times N_j$ is forbidden iff ij is not an edge of H. We will show that the number of forbidden pairs is at least $n^2/\ln\ln m$. It follows that

$$\mathbf{Pr}(\sigma \text{ is consistent}) \leq (1-p_1)^{n^2/\ln\ln m} \leq e^{-nd/\ln\ln m} = o(m^{-n}),$$

assuming that $d \geq K \ln m \ln\ln m$ for sufficiently large K. Since this probability is $o(m^{-n})$ we can multiply by m^n, which is an overcount of the number of assignments satisfying (11), and so obtain the desired first moment bound.

Let $n_i = |N_i|$ and let $I = \{i: \; n_i \geq n/(2m)\}$. Now

$$\sum_{i \in I} n_i = n - \sum_{i \notin I} n_i \geq n - m \cdot \frac{n}{2m} = \frac{n}{2}. \tag{12}$$

For the following analysis we choose constants:

$$\epsilon, \qquad Q = \max\{Q_0, 100\epsilon^{-1}\}, \qquad K_1 = 100R_0, \qquad K = 100K_1Q$$

where $\epsilon \leq \epsilon_0, Q_0, R_0$ are from Lemma 10.

We partition I into 3 parts:

- $I_1 = \{i: \; n/(K_1 \ln m \ln\ln m) \leq n_i < 4n/K\ln m\}$
- $I_2 = \{i: \; n/(K_1 \ln m)^2 \leq n_i < n/(K_1 \ln m \ln\ln m)\}$
- $I_3 = \{i: \; n/(2m) \leq n_i < n/(K_1 \ln m)^2\}$

Case 1: $\sum_{i\in I_1} n_i \geq \frac{n}{6}$ Let H_1 be the subgraph of H induced by I_1, and for each $i \in I_1$, we let $\overline{d}(i)$ be the degree of i in $\overline{H_1}$. Note that the total number of forbidden pairs of vertices for G is at least

$$\frac{1}{2} \sum_{i \in I_1} \overline{d}(i) n_i \times \frac{n}{K_1 \ln m \ln\ln m}, \tag{13}$$

since for all $i' \in I_1, n_{i'} \geq n/(K_1 \ln m \ln\ln m)$.

By (11), we have $|I_1| \geq (K\ln m)/24$, so $(K \ln m)/Q < \epsilon|I_1|$. Thus, by Lemma 10(b) then there are at most $Q \ln m$ members $i \in I_1$ with $\overline{d}(i) < (K\ln m)/Q$. Again using (11), these members contribute at most $4Qn/K < n/12$ to $\sum_{i\in I_1} n_i$. Therefore, the sum in (13) is at least

$$\frac{1}{2} \times \frac{K \ln m}{Q} \times \frac{n}{12} \times \frac{n}{K_1 \ln m \ln\ln m} \geq \frac{n^2}{\ln\ln m}.$$

Case 2: $\sum_{i\in I_2} n_i \geq \frac{n}{6}$ We let $I(j) = \{i \in I_2: \; n/2^j \leq n_i \leq n/2^{j-1}\}$, for $\log_2(K_1 \ln m \ln\ln m) \leq j \leq 2\log_2(K_1 \ln m)$. We set $t_j = \sum_{i\in I(j)} n_i$ and $s_j = |I(j)| \geq t_j \times (K_1 \ln m \ln\ln m/n)$. We set $J = \{j : t_j \geq n/(100 \ln\ln m)\}$ and note that $s_j \geq s_0$ (from Lemma 10) for each $j \in J$. Note also that

$$\sum_{j\in J} t_j \geq \frac{n}{6} - 2\log_2(K_1 \ln m) \times \frac{n}{100 \ln\ln m} \geq \frac{n}{8}.$$

Consider $I(j)$ for any $j \in J$. By Lemma 10, there are at least $\epsilon\binom{s_j}{2}$ pairs $i, i' \in I(j)$ such that every pair of vertices in $N_i \times N_{i'}$ is forbidden. Also, for any i, every pair in $N_i \times N_i$ is forbidden. Since the sizes of the sets $N_i, i \in I(j)$ differ by at most a factor of 2, this implies that the number of forbidden pairs in $\cup_{i\in I(j)} N_i$ is at least $\frac{\epsilon}{8} t_j^2$. Now consider any pair $I(j), I(j')$ with $j, j' \in J$. By Lemma 10(a), there are at least $\epsilon s_j s_{j'}$ pairs $i \in I(j), i' \in I(j')$ such that every pair of vertices in $N_i \times N_{i'}$ is forbidden, and this implies that the number of forbidden pairs in $\cup_{i\in I(j)} N_i \times \cup_{i\in I(j')} N_i$ is at least $\frac{\epsilon}{4} t_j t_{j'}$. Thus, the total number of forbidden pairs is at least

$$\frac{\epsilon}{8}\left(\sum_{j\in J} t_j^2 + \sum_{j,j'\in J; j<j'} 2t_j t_{j'}\right) = \frac{\epsilon}{8}\left(\sum_{j\in J} t_j\right)^2 \geq \frac{\epsilon n^2}{8^3} > \frac{n^2}{\ln\ln m}.$$

Case 3: $\sum_{i \in I_3} n_i \geq \frac{n}{6}$. Here we follow essentially the same argument as in Case 2. Again, let $I(j) = \{i \in I : \ n/2^j \leq n_i \leq n/2^{j-1}\}$, but this time we consider $2\log_2(K_1 \ln m) < j \leq \log_2(2m)$. Again, $t_j = \sum_{i \in I(j)} n_i$ and $s_j = |I(j)|$, but note that this time we have

$$s_j \geq \frac{t_j}{n/(K_1 \ln m)^2}.$$

Here, we set $J = \{j : t_j \geq n/K_1 \ln m\}$ and so again we have $s_j \geq s_0$ for every $j \in J$.

$$\sum_{j \in J} t_j \geq \frac{n}{4} - \log_2(2m) \times \frac{n}{K_1 \ln m} \geq \frac{n}{8}.$$

The same argument as in Case 2 now goes through to imply that the total number of forbidden pairs is at least

$$\frac{\epsilon}{8}\left(\sum_{j \in J} t_j\right)^2 > \frac{n^2}{\ln \ln m}.$$

□

7 Model B: Resolution complexity

First note that **whp** every set of 10 vertices in H has a common neighbour, since the probability of at least one such set not having a common neighbour is less than $\binom{m}{10} q_2^{m-10} = o(1)$. Assuming that H has this property, every vertex of degree at most 10 in G will be in the boundary.

A straightforward first moment argument shows that a.s. every subgraph G' of G with at most $n/d^{3/2}$ vertices has at most $5|G'|$ edges. (We omit the standard calculation.) Therefore, every such G' has at least $|G'|/11$ vertices of degree at most 10. This implies both conditions of Lemma 7 with $s = n/d^{3/2}$ and $\zeta = 1/(22d^{3/2})$ and thus implies Theorem 4. □

We remark that the exponent "3" of d in the statement of Theorem 4 can be replaced by values arbitrarily close to 2 by replacing "10" with a larger value in this proof.

References

1. D. Achlioptas, P. Beame and M. Molloy. *A sharp threshold in proof complexity.* Proceedings of STOC 2001, 337 - 346.
2. 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.
3. P. Beame, J. Culberson and D. Mitchell. *The resolution complexity of random graph k-colourability.* In preparation.
4. P. Beame and T. Pitassi. *Simplified and improved resolution lower bounds.* Proceedings of FOCS 1996, 274 - 282.
5. 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).
6. E. Ben-Sasson and A. Wigderson. *Short proofs are narrow - resolution made simple.* Proceedings of STOC 1999 and Journal of the ACM **48** (2001)
7. B. Bollobás, Random graphs, Second Edition, Cabridge University Press, 2001.
8. B. Bollobás, *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs*, European Journal on Combinatorics **1** (1980) 311–316.
9. E. A. Bender and E. R. Canfield, *The asymptotic number of labelled graphs with given degree sequence*, Journal of Combinatorial Theory (A) **24** (1978) 296–307.
10. V. Chvatal and E. Szemeredi. *Many hard examples for resolution.* Journal of the ACM **35** (1988) 759 - 768.
11. N. Creignou and H. Daude. *Random generalized satisfiability problems.* Proceedings of SAT 2002.
12. R. Dechter, *Constraint networks*, in Encyclopedia of Artificial Intelligence, S. Shapiro (ed.), Wiley, New York, 2nd ed. (1992) 276–285.

13. M. Dyer, A. Frieze and M. Molloy, *A probabilistic analysis of randomly generated binary constraint satisfaction problems.* Theoretical Computer Scince **290**, 1815 - 1828 (2003).
14. E. C. Freuder, *A sufficient condition for backtrack-free search*, Journal of the ACM **29** (1982) 24–32.
15. D. G. Bobrow and M. Brady, eds., Special Volume on Frontiers in Problem Solving: Phase Transitions and Complexity, Guest Editors: T. Hogg, B. A. Hubermann, and C. P. Williams, *Artificial Intelligence* **81** (1996), nos. 1 and 2.
16. I. Gent, E. MacIntyre, P. Prosser, B. Smith and T. Walsh. *Random constraint satisfaction: flaws and structure.* Constraints **6**, 345 - 372 (2001).
17. S. Janson, T. Łuczak and A. Ruciński, *Random Graphs*, Wiley, 2000.
18. A. K. Mackworth, *Constraint satisfaction*, in Encyclopedia of Artificial Intelligence, S. Shapiro (ed.), Wiley, New York, 2nd ed. (1992) 285-293.
19. D. Mitchell, *The Resolution complexity of random constraints.* Proceedings of Principles and Practices of Constraint Programming - CP 2002.
20. D. Mitchell, *The Resolution Complexity of Constraint Satisfaction.* Ph.D. Thesis, University of Toronto, 2002.
21. M. Molloy, *Models for Random Constraint Satisfaction Problems.* Proceedings of STOC 2002, 209 - 217. Longer version to appear in SIAM J. Computing.
22. M. Molloy, *When does the giant component bring unsatisfiability?* Submitted.
23. M. Molloy and M. Salavatipour, *The resolution complexity of random constraint satisfaction problems.* Submitted.
24. B. Pittel, J. Spencer and N. Wormald, *Sudden emergence of a giant k-core in a random graph*, Journal of Combinatorial Theory (B) **67** (1996) 111–151.
25. D. Waltz, *Understanding line drawings of scenes with shadows*, The Psychology of Computer Vision, McGraw-Hill, New York, (1975) 19-91.