Many Hard Examples in Exact Phase Transitions[1]

Ke Xu and Wei Li

National Lab of Software Development Environment
Department of Computer Science
Beihang University, Beijing 100083, China
Email:{kexu,liwei}@nlsde.buaa.edu.cn

Abstract. This paper analyzes the resolution complexity of two random CSP models (i.e. Model RB/RD) for which we can establish the existence of phase transitions and identify the threshold points exactly. By encoding CSPs into CNF formulas, it is proved that almost all instances of Model RB/RD have no tree-like resolution proofs of less than exponential size. Thus, we not only introduce new families of CSPs and CNF formulas hard to solve, which can be useful in the experimental evaluation of CSP and SAT algorithms, but also propose models with both many hard instances and exact phase transitions. Finally, conclusions are presented, as well as a detailed comparison of Model RB/RD with the Hamiltonian cycle problem and random 3-SAT, which, respectively, exhibit three different kinds of phase transition behavior in NP-complete problems.

Keywords: constraint satisfaction problem (CSP), random problems, resolution complexity, phase transitions, SAT.

1. Introduction

Since the seminal paper by Cheeseman et al. in 1991, the study of phase transition and threshold phenomena has been among the most vibrant areas in artificial intelligence, and is emerging as an active interdisciplinary research field of computer science, discrete mathematics and statistical physics. It is shown experimentally (and partly supported theoretically) that for many NP-complete problems, as a parameter is varied, there is a sharp transition from 1 to 0 at a threshold point with respect to the probability of a random instance being soluble. More interestingly, the hardest instances to solve are concentrated in the sharp transition region. As well known, finding ways to generate hard instances for a problem is important both for understanding the complexity of the problem and for providing challenging benchmarks for experimental evaluation of algorithms [11]. So the finding of phase transition phenomena in computer science not only gives a new method to generate hard instances but also provides useful insights into the study of computational complexity from a new perspective.

Although tremendous progress has been made in the study of phase transitions, there is still some lack of research about the connections between the threshold phenomena and the generation of hard instances, especially from a theoretical point of view. For example, some problems can be used to generate hard instances but the existence of phase transitions in such

[1] This research was partially supported by the National 973 Program of China (Grant Nos. G1999032701 and 2005CB321901), the National Natural Science Foundation of China (Grant No. 60403003) and FANEDD (Grant No. 200241). An earlier version of this paper appeared as CoRR Report cs.CC/0302001 in Feb. 2003.

problems has not been proved. One such an example is the well-studied random 3-SAT. A theoretical result by Chvátal and Szemerédi [9] shows that for random 3-SAT, no short resolution proofs exists in general, which means that almost all proofs for this problem require exponential resolution lengths. Experimental results further indicate that instances from the phase transition region of random 3-SAT tend to be particularly hard to solve [23]. Since the early 1990's, considerable efforts have been put into random 3-SAT, but until now, the existence of the phase transition phenomenon in this problem has not been established, although Friedgut [15] made tremendous progress in proving that the width of the phase transition region narrows as the number of variables increases. On the other hand, for some problems with proved phase transitions, it was found either theoretically or experimentally that instances generated by these problems are easy to solve or easy in general. Such examples include random 2-SAT, Hamiltonian cycle problem and random 2+p-SAT ($0 < p \leq 0.4$). For random 2-SAT, Chvátal and Reed [9] and Goerdt [21] proved that the phase transition phenomenon will occur when the ratio of clauses to variables is 1. But we know that 2-SAT is in P class which can be solved in polynomial time, implying that random 2-SAT can not be used to generate hard instances. For the Hamiltonian cycle problem which is NP-compete, Komlós and Szemerédi [22] not only proved the existence of the phase transition in this problem but also gave the exact location of the transition point. However, both theoretical results [7] and experimental results [30] suggest that generally, the instances produced by this problem are not hard to solve. Different from the above two problems, random 2+p-SAT [28] was first proposed as an attempt to interpolate between the polynomial time problem random 2-SAT with $p = 0$ and the NP-complete problem random 3-SAT with $p = 1$. It is not hard to see that random 2+p-SAT is in fact NP-compelte for $p > 0$. The phase transition behavior in this problem with $0 < p \leq 0.4$ was established by Achlioptas et al. and the exact location of the threshold point was also obtained [1]. But it was further shown that random 2+p-SAT is essentially similar to random 2-SAT when $0 < p \leq 0.4$ with the typical computational cost scaling linearly with the number of variables [27].

As mentioned before, from a computational theory point of view, what attracts people most in the study of phase transitions is the finding of many hard instances in the phase transition region. Hence, starting from this point, we can say that the problem models which can not be used to generate random hard instances are not so interesting for study as random 3-SAT. However, until now, for the models with many hard instances, e.g. random 3-SAT, the existence of phase transitions has not been established, not even the exact location of the threshold points. So, from a theoretical perspective, we still do not have sufficient evidence to support the long-standing observation that there exists a close relation between the generation of many hard instances and the threshold phenomena, although this observation opened the door for, and has greatly advanced the study of phase transitions in the last decade. From the discussion above, an interesting question naturally arises: *whether there exist models with both proved phase transitions and many hard instances.*

Recently, to overcome the trivial asymptotic insolubility of the previous random CSP models, Xu and Li [31] proposed a new CSP model, i.e. Model RB, which is a revision to the standard Model B. It was proved that the phase transitions from solubility to insolubility do exist for Model RB as the number of variables approaches infinity. Moreover, the threshold points at which the phase transitions occur are also known exactly. Based on previous experiments and by relating the hardness of Model RB to Model B, it has already been shown that Model RB abounds with hard instances in the phase transition region. In this paper, by encoding CSPs into CNF formulas, we will prove that almost all instances of Model RB have no tree-like resolution proofs of less than exponential size. Thus, we give a positive answer to the question above.

The rest of this paper is organized as follows. Section 2 will introduce some basic definitions. In Section 3, we will first give an overview of Model RB and then propose a random CSP model, called Model RD, along the same line as for Model RB. Section 4 will give the resolution complexity result for Model RB and Model RD while the proof of this result will be detailed in Section 5. Finally, we conclude in Section 6 by discussing the phase transition behavior in NP-complete problems.

2. Preliminaries

A CNF formula F is a conjunction ($\wedge$) of *clauses*, where each clause is a disjunction ($\vee$) of *literals* and a literal is a propositional variable or its negation ($\neg$). A CNF formula is *satisfiable* if there is an assignment of truth values to the variables which makes the formula true; otherwise it is *unsatisfiable.* The problem of determining whether a CNF formula is satisfiable is known as the *propositional satisfiability problem* (SAT).

Resolution is a simple and complete proof system for proving unsatisfiability of CNF formulas, which is based on the following rule: if $(A \vee x)$ and $(B \vee \neg x)$ are two clauses, then we can derive the clause $(A \vee B)$, called the *resolvent.* A *resolution derivation* of a clause C from a CNF formula F is a sequence of clauses $\pi = C_1, C_2, \cdots, C_m$ where $C_m = C$ and every C_i is either a clause of F or the resolvent of two clauses C_j and C_k with $j, k < i$. The size of π is the number of clauses in it. A derivation of the empty clause, denoted by $\Box$, from F is called a *refutation* or *proof* of F. A derivation π is called *tree-like* if each non-empty derived clause is used exactly once in π.

A *constraint satisfaction problem* (CSP) is a generalization of SAT, which consists of a finite set $U = \{u_1, \cdots, u_n\}$ of n variables and a set of constraints defining the values that the variables can simultaneously take. More specifically, each variable u_i is associated with a *domain* $D(u_i)$ which specifies the possible values of that variable. A *constraint* $C_{i1,i2,\cdots,ik}$ consists of a subset $\{u_{i1}, u_{i2}, \cdots, u_{ik}\}$ of U and a relation $R_{i1,i2,\cdots,ik} \subseteq D(u_{i1}) \times \cdots \times D(u_{ik})$, where $i1, i2, \cdots, ik$ are distinct and $R_{i1,i2,\cdots,ik}$ specifies the compatible tuples of values for the variables $u_{i1}, \cdots, u_{ik}$. The incompatible tuples are called *nogoods.* The number of variables bounded by a constraint is called its *arity.* A constraint is called binary if its arity $k = 2$ and non-binary if $k \geq 3$. A CSP is called binary if the constraints of this CSP are binary. A *solution* to a CSP is an assignment of a value to each variable from its domain such that every constraint is satisfied. A constraint $C_{i1,i2,\cdots,ik}$ is satisfied if the tuple of values assigned to the variables $u_{i1}, \cdots, u_{ik}$ is compatible. A CSP that has a solution is called *satisfiable*; otherwise it is *unsatisfiable.*

There are two natural ways to discuss the resolution complexity of a CSP. One way is to directly encode a CSP instance into a CNF formula and define the resolution complexity of the CSP instance to be the resolution complexity of the corresponding CNF formula. As indicated in [24], this approach is stronger in simulating popular CSP algorithms than the other one and so we adopt it in this paper. Given a CSP instance P, we directly encode it into a CNF formula, denoted by $\phi(P)$, as follows. For each value j of each CSP variable u_i, we introduce a propositional variable x_{ij}, called a *domain variable* of u_i. If $x_{ij} = T$, then it means that the value j is assigned to the variable u_i. There are three sets of clauses needed in the encoding.

The *domain clause* ensures that each variable must be assigned a value from its domain. For example, if u_i is a CSP variable whose domain has d elements, then there is a domain clause: $x_{i1} \vee x_{i2} \vee \cdots \vee x_{id}$. The *at-most-one clause* asserts that each variable is assigned at most one value from its domain. For example, if $j_1, j_2 \in D(u_i)$ and $j_1 \neq j_2$, then there is an at-most-one clause: $\neg x_{ij_1} \vee \neg x_{ij_2}$. Finally, the *conflict clause* excludes any nogoods of each constraint. For example, if $u_1 = 2$ and $u_2 = 1$ is a nogood, then there is a conflict clause: $\neg x_{12} \vee \neg x_{21}$.

3. Model RB and Model RD

The CSP is a fundamental problem in Artificial Intelligence, with a distinguished history and many applications, such as in knowledge representation, scheduling and pattern recognition. To compare the efficiency of different CSP algorithms, some standard random CSP models have been widely used experimentally to generate benchmark instances in the past decade. Among these models, the most commonly used one is Model B which is defined by four parameters $< n, d, p_1, p_2 >$, where n is the number of variables, d is the uniform domain size, p_1 (called *constraint density*) is the proportion of constraints selected at random from a set of $n(n-1)/2$ possible binary constraints and p_2 (called *constraint tightness*) is the proportion of nogoods in each constraint selected at random from a set of d^2 possible nogoods. For Model B, Achlioptas et al. [2] proved that except for a small range of values of the constraint tightness, almost all instances generated are unsatisfiable as the number of variables approaches infinity. This result, as shown in [20], implies that most previous experimental results about random CSPs are asymptotically uninteresting. However, it should be noted that Achlioptas et al.'s result holds under the condition of fixed domain size and so is applicable only when the number of variables is overwhelmingly larger than the domain size. But in fact, it can be observed that the domain size, compared to the number of variables, is not very small in most experimental CSP studies. This, in turn, explains why there is a big gap between Achlioptas et al.'s theoretical result and the experimental findings about the phase transition behavior in random CSPs. Motivated by the observation above, and to overcome the trivial asymptotic insolubility of the previous random CSP models, Xu and Li [31] proposed an alternative CSP model as follows.

Model RB: Given a set U of n variables, first, we select with repetition $m = rn \ln n$ random constraints. Each random constraint is formed by selecting without repetition k of n variables, where $k \geq 2$ is an integer. Next, for each constraint we select uniformly at random without repetition $q = p \cdot d^k$ nogoods, i.e., each constraint contains exactly $(1-p) \cdot d^k$ compatible tuples of values, where $d = n^\alpha$ is the domain size of each variable and $\alpha > 0$ is a constant.

Note that the way of generating random instances for Model RB is almost the same as that for Model B. However, like the N-queens problem and Latin square, the domain size of Model RB is not fixed but polynomial in the number of variables. It is proved that Model RB not only avoids the trivial asymptotic behavior but also has exact phase transitions. More precisely, the following theorems hold for Model RB, where $\Pr(Sat)$ denotes the probability that a random CSP instance generated by Model RB is satisfiable.

Theorem 1 (Xu and Li [31]) Let $r_{cr} = -\frac{\alpha}{\ln(1-p)}$. If $\alpha > \frac{1}{k}$, $0 < p < 1$ are two constants and k, p satisfy the inequality $k \geq \frac{1}{1-p}$, then

$$\begin{aligned} \lim_{n\to\infty} \Pr(Sat) &= 1 \text{ for any constant } r < r_{cr}, \\ \lim_{n\to\infty} \Pr(Sat) &= 0 \text{ for any constant } r > r_{cr}. \end{aligned}$$

Theorem 2 (Xu and Li [31]) Let $p_{cr} = 1 - e^{-\frac{\alpha}{r}}$. If $\alpha > \frac{1}{k}$, $r > 0$ are two constants and k, α and r satisfy the inequality $ke^{-\frac{\alpha}{r}} \geq 1$, then

$$\begin{aligned} \lim_{n\to\infty} \Pr(Sat) &= 1 \text{ for any constant } p < p_{cr}, \\ \lim_{n\to\infty} \Pr(Sat) &= 0 \text{ for any constant } p > p_{cr}. \end{aligned}$$

As shown in [31], many instances generated following Model B in previous experiments can also be viewed as instances of Model RB, and more importantly, the experimental results for these instances agree well with the theoretical predictions for Model RB. Therefore, in this sense, we can say that Model B can still be used experimentally to generate benchmark instances with non-trivial threshold behaviors. However, to achieve this, a natural and convenient way is to vary the values of CSP parameters under the framework of Model RB. For more discussions on the experimental aspects of Model RB, please see [33]. Note that another standard CSP Model, i.e. Model D, is almost the same as Model B except that for every constraint, each tuple of values is selected to be incompatible with probability p. Similarly, we can make a revision to Model D and then get a new Model as follows.

Model RD: Given a set U of n variables, first, we select with repetition $m = rn \ln n$ random constraints. Each random constraint is formed by selecting without repetition k of n variables, where $k \geq 2$ is an integer. Next, for each constraint, from d^k possible tuples of values, each tuple is selected to be incompatible with probability p, where $d = n^\alpha$ is the domain size of each variable and $\alpha > 0$ is a constant.

Along the same line as in the proof for Model RB [31], we can easily prove that exact phase transitions also exist for Mode RD. More precisely, Theorem 1 and Theorem 2 hold for Model RD too. In fact, it is exactly because the differences between Model RB and Model RD are very small that many properties hold for both of them and the proof techniques are also almost the same. So in this paper, we will discuss both models, denoted by Model RB/RD.

Recently, there has been a growing theoretical interest in random CSPs, especially with respect to their phase transition behaviors [10, 12-14, 17, 18, 29, 32] and resolution complexity [18, 19, 24, 26]. In what follows, we will discuss the resolution complexity of Model RB/RD.

4. Main Results

In this paper, we prove the following result.

Theorem 3 Let P be a random CSP instance generated following Model RB/RD. Then, **whp**[2] P has no tree-like resolutions of length less than $2^{\Omega(n)}$ and no general resolutions of length less than $2^{\Omega(n/d)}$.

It should be noted that unlike Theorems 1 and 2, there is no restriction on the values of CSP parameters to make the above theorem hold. As far as we know, this theorem is also the first resolution complexity result for a general (non-binary) CSP model with growing domains. A similar resolution result was proved in [18] for a binary CSP model where the domain size $d \geq (\ln n)^{1+\epsilon}$ for any constant $\epsilon > 0$. This also leaves an open question whether the lower bound for general resolutions of CSPs with growing domains can be improved to be as large as

[2] When we say that a property holds **whp** (with high probability) it means that this property holds with probability tending to 1 as the number of variables approaches infinity.

that for tree-like resolutions, which seems to be worth further investigation in future studies. Combining Theorems 1, 2 and 3, Model RB/RD provide a general framework for generating asymptotically hard CSPs with non-trivial threshold behaviors, which especially can help to satisfy an increasing interest in the study of non-binary CSPs.

5. Proof of Theorem 3

The core of the proof for Theorem 3 is to show that **whp** there exists a clause with large width in every refutation. The width of a clause C, denoted by $w(C)$, is the number of variables appearing in it. The width of a set of clauses is the maximal width of a clause in the set. The width of deriving a clause C from the formula F, denoted by $w(F \vdash C)$ is defined as the minimum of the widths of all derivations of C from F. So, the width of refutations for F can be denoted by $w(F \vdash 0)$. Ben-Sasson and Wigderson [6] gave the following theorem on size-width relations and proposed a general strategy for proving width lower bounds for CNF formulas.

Theorem 4 (Ben-Sasson and Wigderson [6]) Let F be a CNF formula with n variables and $S(F)$ ($S_T(F)$) be the minimal size of a (tree-like) refutation. Then we have

$$S_T(F) \geq 2^{(w(F \vdash 0) - w(F))}, \ S(F) = \exp\left(\Omega\left(\frac{(w(F \vdash 0) - w(F))^2}{n}\right)\right).$$

By extending Ben-Sasson and Wigderson's strategy, Mitchell [24] proved exponential resolution lower bounds for some random CSPs of fixed domain size. In what follows, to obtain lower bounds on width for RB/RD, we will basically use the same strategy as in [24], but extend it to handle random CSPs with growing domains. First, we prove the following local sparse property for RB/RD.

Lemma 1 Let P be a random CSP instance of Model RB/RD. There is constant $c > 0$ such that **whp** every sub-problem of P with size $s \leq cn$ has at most $b = \beta s \ln n$ constraints, where $\beta = \frac{\alpha}{6k \ln \frac{1}{1-p}}$.

Proof: A *sub-problem* is defined by a subset of variables with all the constraints where only the variables in the subset occur. We consider the number of sub-problems on s variables with $b = \beta s \ln n$ constraints for $0 < s \leq cn$. There are $\binom{n}{s}$ possible choices for the variables and $\binom{m}{b}$ for the constraints. Given such choices, the probability that all the b constraints are in the s variables is not greater than $\left(\frac{s}{n}\right)^{kb}$. So, the number of such sub-problems is at most

$$\begin{aligned}
\binom{n}{s}\binom{m}{b}\left(\frac{s}{n}\right)^{kb} &\leq \left(\frac{en}{s}\right)^{s}\left(\frac{em}{b}\right)^{b}\left(\frac{s}{n}\right)^{kb} \\
&= \left(\frac{en}{s}\right)^{s}\left(\frac{ern\ln n}{\beta s \ln n}\right)^{\beta s \ln n}\left(\frac{s}{n}\right)^{k\beta s \ln n} \\
&= \left[\frac{e^{1+\beta \ln n} r^{\beta \ln n}}{\beta^{\beta \ln n}}\left(\frac{s}{n}\right)^{(k-1)\beta \ln n - 1}\right]^{s}.
\end{aligned}$$

For sufficiently large n, there exists a constant $c_1 > 0$ such that

$$\frac{e^{1+\beta \ln n} r^{\beta \ln n}}{\beta^{\beta \ln n}} < n^{c_1}.$$

Thus we get

$$\binom{n}{s}\binom{m}{b}\left(\frac{s}{n}\right)^{kb} < \left[n^{c_1}\left(\frac{s}{n}\right)^{(k-1)\beta\ln n-1}\right]^{s}.$$

Let $c < \frac{1}{2}\exp\left(-\frac{2+c_1}{(k-1)\beta}\right)$ be a positive constant. For $0 < s \le cn$, it follows from the above inequality that

$$\binom{n}{s}\binom{m}{b}\left(\frac{s}{n}\right)^{kb} < \left(\frac{1}{n^2}\right)^{s} \le \frac{1}{n^2}.$$

Thus the expected number of such sub-problems with $s \le cn$ is at most

$$\sum_{s=1}^{cn}\binom{n}{s}\binom{m}{b}\left(\frac{s}{n}\right)^{kb} < \frac{1}{n^2}cn = o(1).$$

This finishes the proof. □

The basic idea behind the proofs in this paper and [24] is to show that **whp** small sub-problems of a CSP do not have too many constraints (i.e. the constraint graph of the CSP is locally sparse) and are satisfiable. From the above lemma and the corresponding one in [24], we can see that the constraint graph discussed in this paper is much denser than that in [24] where the domain size is taken to be constant. This is because that in the former case, the number of constraints is super-linear in the number of variables while in the latter case such a relationship is linear. Mainly due to this reason, most results in [24] are not directly applicable here. To prove that for instances of Model RB/RD, small sub-problems are satisfiable **whp**, we need to introduce the following two definitions.

Definition 1 Consider a variable u and a set of i constraints each containing u. In the set of i constraints, there is an assignment of values to all the variables except u. We call this an *i-constraint assignment tuple*, denoted by $T_{i,u}$.

Definition 2 Given a variable u and an i-constraint assignment tuple $T_{i,u}$. We assign a value v to u from its domain. So, all the variables in the i constraints of $T_{i,u}$ have been assigned values. If at least one constraint in $T_{i,u}$ is violated by these values, then we say that *the value v of u is flawed by $T_{i,u}$*. If all the values of u in its domain are flawed by $T_{i,u}$, then we say that *the variable u is flawed by $T_{i,u}$*, and $T_{i,u}$ is called a *flawed* i-constraint assignment tuple.

Lemma 2 Let P be a random CSP instance of Model RB/RD. Then, **whp** no assignment of values to any subset of variables can produce a flawed i-constraint assignment tuple $T_{i,u}$ in P with $i \le 3k\beta\ln n$.

Proof: Now consider an i-constraint assignment tuple $T_{i,u}$ with $i \le 3k\beta\ln n$. It is easy to see that the probability that $T_{i,u}$ is flawed increases with the number of constraints i. Recall that in Model RD, for every constraint, each tuple of values is selected to be incompatible with probability p. So, given a value v of u, the probability that v is flawed by $T_{i,u}$ is

$$1-(1-p)^{i}.$$

Thus, since each assignment becomes a nogood independently, the probability that all the $d = n^\alpha$ values of u are flawed by $T_{i,u}$, i.e. the probability of $T_{i,u}$ being flawed is

$$\left[1-(1-p)^i\right]^d.$$

Note that $\beta = \frac{\alpha}{6k\ln\frac{1}{1-p}}$. Thus for $0 < i \leq 3k\beta\ln n$, we have

$$\begin{aligned}\Pr(T_{i,u} \text{ is flawed})|_{i\leq 3k\beta\ln n} &\leq \left[1-(1-p)^{3k\beta\ln n}\right]^{n^\alpha}\\ &= [1-\frac{1}{n^{\frac{\alpha}{2}}}]^{n^\alpha} \leq e^{-n^{\frac{\alpha}{2}}}.\end{aligned}$$

The above analysis only applies to Model RD. For Model RB, such an analysis is much more complicated, and so we leave it in the appendix. Recall that there are n variables and $m = rn\ln n$ constraints. So the number of possible choices for i-constraint assignment tuples is at most

$$n\binom{m}{i}d^{(k-1)i}.$$

For $i \leq 3k\beta\ln n$, when n is sufficiently large, there exists a constant $c_2 > 0$ such that

$$\begin{aligned}n\binom{m}{i}d^{(k-1)i} &= n\binom{rn\ln n}{i}n^{(k-1)\alpha i} \leq n\binom{rn\ln n}{3k\beta\ln n}n^{3(k-1)\alpha k\beta\ln n}\\ &\leq n\left(\frac{ern\ln n}{3k\beta\ln n}\right)^{3k\beta\ln n}n^{3(k-1)\alpha k\beta\ln n} < e^{c_2\ln^2 n}.\end{aligned}$$

Thus the expected number of flawed i-constraint assignment tuples with $i \leq 3k\beta\ln n$ is at most

$$\begin{aligned}\sum_{i=1}^{3k\beta\ln n} n\binom{m}{i}d^{(k-1)i}\Pr(T_{i,u} \text{ is flawed}) &< e^{c_2\ln^2 n}\sum_{i=1}^{3k\beta\ln n}\Pr(T_{i,u} \text{ is flawed})\\ &\leq e^{c_2\ln^2 n}\cdot e^{-n^{\frac{\alpha}{2}}}\cdot 3k\beta\ln n\\ &= o(1).\end{aligned}$$

This implies that **whp** there does not exist a variable u and an i-constraint assignment tuple $T_{i,u}$ with $i \leq 3k\beta\ln n$ such that u is flawed by $T_{i,u}$. This is exactly what we need and so we are done. □

Lemma 3 Let P be a random CSP instance of Model RB/RD. Then, **whp** every sub-problem of P with at most cn variables is satisfiable.

Proof: Here we define the size of a problem as the number of variables in this problem. We will prove this lemma by contradiction. Assume that we have an unsatisfiable sub-problem of size at most cn. Thus we can get a minimum sized unsatisfiable sub-problem with size $s \leq cn$, denoted by P_1. From Lemma 1 we know that P_1 has at most $\beta s\ln n$ constraints **whp**. Thus there exists a variable u in P_1 with degree at most $k\beta\ln n$, i.e. the number of constraints in P_1 associated with u is not greater than $k\beta\ln n$. Removing u and the constraints associated with u

from P_1, we get a sub-problem P_2. By minimality of P_1, we know that P_2 is satisfiable, and so there exists an assignment satisfying P_2. Suppose that the variables in P_2 have been assigned values by such an assignment. Now consider the variable u and the i constraints associated with u, where $i \leq k\beta \ln n$. By Definition 2 this constitutes an i-constraint assignment tuple for u, denoted by $T_{i,u}$. Recall that P_1 is unsatisfiable. This means that no value of u can satisfy all the i constraints. That is to say, the variable u is flawed by $T_{i,u}$. Therefore, if a sub-problem of size at most cn is unsatisfiable, then, **whp** there is a variable u and an i-constraint assignment tuple $T_{i,u}$ such that u is flawed by $T_{i,u}$, where $i \leq k\beta \ln n$. This is in contradiction with Lemma 2 and so finishes the proof. □

Now we will prove that there **whp** exist a complex clause in the refutation proofs of Model RB/RD. The complexity of a clause was defined in [24] by Mitchell as follows. For a CSP instance P and a clause C over the literals in $\phi(P)$, the *complexity* of C with respect P, denoted by $\mu(C)$, is the size of the smallest sub-problem Π such that C can be derived by resolution from $\phi(\Pi)$. Along the same line as in the proof of [24], we have the following lemma.

Lemma 4 Let P be a random CSP instance of Model RB/RD. Then, **whp** every refutation π of $\phi(P)$ has a clause C of complexity $\frac{cn}{2} \leq \mu(C) \leq cn$.

Proof: The proof for the corresponding lemma in [24] can be directly applied here. For the convenience of readers, we give the sketch of the proof as follows. Construct a graph G_π for π where the nodes are the clauses in π and the parent of two clauses is their resolvent. The root of G_π is the empty clause □ and the leaves are the input clauses. By Lemma 3, **whp** any sub-problem with size at most cn is satisfiable, and thus $\mu(\Box) > cn$. We claim that there must be a clause B in G_π with complexity no less than cn and its children (denoted by C_1 and C_2) have complexity no greater than cn. This is because that in the path from the root to every node the complexity of clauses is non-increasing and the complexity of each input clause is 1. It follows that $\mu(C_1), \mu(C_2) \leq cn$ and $\mu(C_1) + \mu(C_2) \geq \mu(B) \geq cn$. So, one of C_1 and C_2 must be the clause satisfying the condition in Lemma 4. □

Lemma 5 Given a random CSP instance P of Model RB/RD, let C be a clause of complexity $\frac{cn}{2} \leq \mu(C) \leq cn$ with respect to P. Then, **whp** C has at least $\frac{c}{6}n$ literals, i.e. $w(C) \geq \frac{c}{6}n$.

Proof: We will prove this by contradiction. Let P_1 be the smallest sub-problem such that $\phi(P_1) \models C$. Hence, the size of P_1 is at least $\frac{c}{2}n$ and at most cn. By Lemma 1, there are at most $\beta cn \ln n$ constraints in P_1. So, there are at most $\frac{c}{3}n$ variables with degree greater than $3k\beta \ln n$. Then, there are at least $\frac{c}{2}n - \frac{c}{3}n = \frac{c}{6}n$ variables in P_1 with degree at most $3k\beta \ln n$. We will prove that for these variables, **whp** there does not exist a variable such that no domain variable of it appears in C. Now assume that we have a variable u in P_1 with degree $i \leq 3k\beta \ln n$ and no domain variable of it appears in C. Removing u and the constraints associated with it from P_1, we get a sub-problem P_2. By minimality of P_1, we know that $\phi(P_2) \not\models C$. So we can find an assignment satisfying P_2 but not satisfying C. Suppose that the propositional variables in P_2 and C have been assigned values by such an assignment. Now consider the variable u and the constraints associated with it. By Definition 2, this constitutes an i-constraint assignment tuple for u, denoted by $T_{i,u}$. By assumption, no domain variable of u appears in C. So, assigning any value to u will not affect the truth value of C. Recall that $\phi(P_1) \models C$ and C is false under the current assignment. Therefore, no value of u can satisfy $\phi(P_1)$, i.e. setting any value to u will violate at least one constraint associated with it. It follows that u is flawed by $T_{i,u}$, i.e. there exists a flawed i-constraint assignment tuple with $i \leq 3k\beta \ln n$. This is in contradiction with Lemma 2 and so we are done. □

Combining Lemma 4 and Lemma 5, we have that, for a random CSP instance P of Model RB/RD, **whp** $w(\phi(P) \vdash 0) \geq \frac{c}{6}n$. Now, by applying Theorem 4 and noting that the number of propositional variables is nd, we finish the proof. One point worth mentioning is that when $d = n^{\alpha} \geq n$, the initial width of domain clauses is not less than the number of variables. In such a case, to make Theorem 4 applicable, we need to make the following extensions. For each domain clause $x_1 \vee x_2 \vee \cdots \vee x_{n^{\alpha}}$, we introduce extension variables $y_1, y_2, \cdots, y_{n^{\alpha-0.99}}$ and replace the original domain clause with new clauses: $y_1 \vee y_2 \vee \cdots \vee y_{n^{\alpha-0.99}}$, $\neg y_1 \vee x_1 \vee x_2 \vee \cdots \vee x_{n^{0.99}}$, $\neg y_2 \vee x_{n^{0.99}+1} \vee x_{n^{0.99}+2} \vee \cdots \vee x_{2n^{0.99}}$, $\cdots$, $\neg y_{n^{\alpha-0.99}} \vee x_{n^{\alpha}-n^{0.99}+1} \vee x_{n^{\alpha}-n^{0.99}+2} \vee \cdots \vee x_{n^{\alpha}}$. If $\alpha \geq 1.99$, then we need to make a similar extension to the clause $y_1 \vee y_2 \vee \cdots \vee y_{n^{\alpha-0.99}}$ which (if necessary) will be repeated finite times such that every new clause will finally have at most $o(n)$ literals. It is easy to see that such extensions have no effect on the main results in this paper. More precisely, **whp** the minimum size of tree-like resolutions with new clauses is still exponential and the total number of propositional variables (including extension variables) is still $O(nd)$. On the other hand, it is straightforward to see that every domain clause can be derived by resolutions from its associated new clauses in polynomial steps. Based on the above two points, it follows that the minimum size of resolutions with original domain clauses is (up to a polynomial size) greater than or equal to that with new clauses, which directly implies the result we desire.

6. Conclusions

In this paper, by encoding CSPs into CNF formulas, we proved exponential lower bounds for tree-like resolution proofs of two general random CSP models with exact phase transitions, i.e. Model RB/RD. This result suggests that we not only introduce new families of hard instances for CSP and SAT, which is of importance for experimentally evaluating CSP and SAT algorithms, but also propose models with both many hard instances and exact phase transitions.

As mentioned before, there are some other NP-complete problems with proved exact phase transitions, e.g. Hamiltonian cycle problem and random 2+p-SAT ($0 < p \leq 0.4$). However, it has been shown either experimentally or theoretically that the instances produced by these problems are generally easy to solve. So one would naturally ask what the main difference between these "easy" NP-complete problems and RB/RD is. It seems that for these "easy" NP-complete problems with exact phase transitions, they usually have some kind of local property which can be used to design polynomial time algorithms working with high probability, and the exact phase transitions are, in fact, obtained by probabilistic analysis of such algorithms. So, it appears that if a problem has exact phase transitions obtained algorithmically, then it also means that the problem is not hard to solve. For RB/RD, the situation is, however, completely different. More specifically, the exact phase transitions of RB/RD are not obtained algorithmically, but by use of the first and the second moment methods which say nothing about the local property of the problem and are, therefore, unlikely to be useful for designing more efficient algorithms. Thus, it seems that RB/RD, unlike the "easy" NP-complete problems, can indeed provide a reliable source to generate hard instances. When talking about the hardness of solving combinatorial problems, we should mention that using concepts from statistical physics, people [3, 27] in the past few years have made remarkable progress in providing deep insights into it, which lies at the frontier between statistical physics and computer science.

Recently, Frieze and Wormald [16] studied random k-SAT for moderately growing k, i.e. $k = k(n)$ satisfies $k - \log_2 n \to \infty$ where n is the number of variables. For this model, they established similarly, by use of the first and the second moment methods, that there exists a satisfiability threshold at which the number of clauses is $m = 2^k n \ln 2$. From Beame et al's earlier work on the complexity of unsatisfiability proofs for random k-SAT formulas [4, 5], we know that the size of resolution refutations for this model is exponential with high probability. So, the variant of random k-SAT studied by Frieze and Wormald is also a model with both exact phase transitions and many hard instances. Note that in all of these studies, either the domain size or the constraint size grows with the number of variables. More recently, a similar result was shown for a CSP model with constant sized domains and constraints [10].

To gain a better understanding of Model RB/RD, we now make a comparison of them with the well-studied random 3-SAT of similar proof complexity. First, we think that the exact phase transitions should be one advantage of RB/RD, which can help us to locate the hardest instances more precisely and conveniently when implementing large-scale computational experiments. As for the theoretical aspect, it seems that RB/RD, intrinsically, are much mathematically easier to analyze than random 3-SAT, such as in the derivation of thresholds. From a personal perspective, we think that such mathematical tractability should be another advantage of RB/RD, making it possible to obtain some interesting results which do not hold or can not be easily obtained for random 3-SAT.

In summary, the Hamiltonian cycle problem, random 3-SAT and Model RB/RD, respectively, exhibit three different kinds of phase transition behavior in NP-complete problems. More specifically: 1) The Hamiltonian cycle problem has a known threshold but its instances are generally easy to solve; 2) Random 3-SAT can generate hard instances but its threshold seems intrinsically hard to derive; 3) Model RB/RD have both known thresholds and hard instances. Compared with the former two that have been extensively explored in the past decade, the third one, due to various reasons, has not received much attention so far. It is hoped that more investigations, either experimental or theoretical, will be carried out on this behavior, and we also believe that such studies will lead to fresh insights and new discoveries in this active area of research (i.e. on phase transitions and computational complexity).

Acknowledgements

We would like to thank two anonymous referees for their careful reading and valuable suggestions, which helped improve the presentation of this paper.

References

1. D. Achlioptas, L. Kirousis, E. Kranakis and D. Krizanc, Rigorous results for random (2+p)-SAT, In: *Proceedings of RALCOM-97*, pp.1-10.

2. D. Achlioptas, LM Kirousis, E. Kranakis, D. Krizanc, M. SO Molloy, and YC. Stamatiou, Random Constraint Satisfaction: A More Accurate Picture, In: *Proc. Third International Conference on Principles and Practice of Constraint Programming* (CP 97), LNCS 1330, pp.107-120, 1997.

3. D. Aldous and A. Percus, Scaling and Universality in Continuous length Combinatorial Optimization, *Proc. Natl. Acad. Sci. USA*, 100:11211-11215, 2003.

4. P. Beame, R. Karp, T. Pitassi, and M. Saks. On the complexity of unsatisfiability proofs for random k-CNF formulas. In: *Proceeding of STOC-98*, pp.561-571.

5. P. Beame, R. Karp, T. Pitassi, and M. Saks. The efficiency of resolution and Davis-Putnam procedures. *SIAM Journal on Computing*, 31(4):1048-1075, 2002.

6. E. Ben-Sasson and A. Wigderson. Short proofs are narrow - resolution made simple. *Journal of the ACM*, 48(2):149-169, 2001.

7. B. Bollobás, T.I. Fenner and A.M. Frieze. An algorithm for finding Hamilton paths and cycles in random graphs. *Combinatorica* 7(4):327-341, 1987.

8. V. Chvátal and E. Szemerédi. Many hard examples for resolution. *Journal of the ACM*, 35(4) (1988) 759-208.

9. V. Chvátal and B. Reed. Miks gets some (the odds are on his side). In: *Proceedings of the 33rd IEEE Symp. on Foundations of Computer Science*, pages 620-627, 1992.

10. H. Connamacher and M. Molloy. The Exact Satisfiability Threshold for a Potentially Intractable Random Constraint Satisfaction Problem. In: *Proceedings of FOCS 2004*.

11. S. Cook and D. Mitchell. Finding Hard Instances of the Satisfiability Problem: A Survey, In: *Satisfiability Problem: Theory and Applications.* Du, Gu and Pardalos (Eds). DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Volume 35, 1997.

12. N. Creignou, H. Daudé. Combinatorial Sharpness Criterion and Phase Transition Classification for Random CSPs. *Information and Computation*, 190(2), pp.220-238, 2004.

13. O. Dubois and J. Mandler. The 3-XORSAT threshold. In: *Proc. FOCS 2002.*

14. M. Dyer, A. Frieze and M. Molloy. A probabilistic analysis of randomly generated binary constraint satisfaction problems. *Theoretical Computer Science* 290 (2003) 1815-1828.

15. E. Friedgut, Sharp thresholds of graph properties, and the k-sat problem. With an appendix by Jean Bourgain. *Journal of the American Mathematical Society* 12 (1999) 1017-1054.

16. A.M. Frieze and N.C. Wormald. Random k-SAT: A tight threshold for moderately growing k, In: *Proceedings of the Fifth International Symposium on Theory and Applications of Satisfiability Testing*, pp.1-6, 2002.

17. A. Flaxman. A sharp threshold for a random constraint satisfaction problem, preprint.

18. A. Frieze and M. Molloy. The satisfiability threshold for randomly generated binary constraint satisfaction problems. In: *Proceedings of RANDOM-03*, 2003.

19. Y. Gao and J. Culberson. Resolution Complexity of Random Constraint Satisfaction Problems: Another Half of the Story. In: *Proc. of LICS-03, Workshop on Typical Case Complexity and Phase Transitions*, Ottawa, Canada, June, 2003.

20. I.P. Gent, E. MacIntyre, P. Prosser, B.M. Smith and T. Walsh, Random Constraint Satisfaction: flaws and structures. *Journal of Constraints* 6(4), 345-372, 2001.

21. A. Goerdt. A threshold for unsatisfiability. In: *17th International Symposium of Mathematical Foundations of Computer Science*, Springer LNCS 629 (1992), pp.264-275.

22. M. Komlós and E. Szemerédi. Limit distribution for the existence of a Hamilton cycle in a random graph. *Discrete Mathematics*, 43, pp.55-63, 1983.

23. D. Mitchell, B. Selman, and H. Levesque. Hard and easy distributions of sat problems. In: *Proceedings of 10th National Conf. on Artificial Intelligence* (AAAI-92), pp.459-465, 1992.

24. D. Mitchell. Resolution Complexity of Random Constraints, In: *Proceedings of CP 2002*, LNCS 2470, pp.295-309.

25. M. Molloy. Models for Random Constraint Satisfaction Problems, submitted. Conference version in *Proceedings of STOC 2002*.

26. M. Molloy and M. Salavatipour. The resolution complexity of random constraint satisfaction problems. In: *Proc. FOCS-03*, 2003.

27. R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman and L. Troyansky. Determining computational complexity from characteristic phase transitions. *Nature*, 400(8):133-137, 1999.

28. R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman and L. Troyansky, Phase transition and search Cost in the 2+p-SAT problem, In: *4th Workshop on Physics and Computation*, Boston University 22-24 November 1996, (PhysComp96).

29. B.M. Smith. Constructing an Asymptotic Phase Transition in Random Binary Constraint Satisfaction Problems. *Theoretical Computer Science*, vol. 265, pp. 265-283 (Special Issue on NP-Hardness and Phase Transitions), 2001.

30. B. Vandegriend and J. Culberson. The $G_{n,m}$ phase transition is not hard for the Hamiltonian Cycle problem. *Journal of Artificial Intelligence Research*, 9:219-245, 1998.

31. K. Xu and W. Li. Exact Phase Transitions in Random Constraint Satisfaction Problems. *Journal of Artificial Intelligence Research*, 12:93-103, 2000.

32. K. Xu and W. Li. On the Average Similarity Degree between Solutions of Random k-SAT and Random CSPs. *Discrete Applied Mathematics*, 136(2004):125-149.

33. K. Xu, F. Boussemart, F. Hemery and C. Lecoutre. A Simple Model to Generate Hard Satisfiable Instances. In: *Proc. of 19th International Joint Conference on Artificial Intelligence* (IJCAI), pp.337-342, Edinburgh, Scotland, 2005.

Appendix

Now we consider the proof of Lemma 2 for Model RB. Given a variable u an i-constraint assignment tuple $T_{i,u}$. It is easy to see that the probability that u is flawed by $T_{i,u}$ increases with the number of constraints i. Thus we have

$$\Pr(T_{i,u} \text{ is flawed})|_{i\leq 3k\beta \ln n} \leq \Pr(T_{i,u} \text{ is flawed})|_{i=3k\beta \ln n}.$$

For the variable u, there are $d = n^{\alpha}$ values in its domain, denoted by $v_1, v_2, \cdots, v_d$. Let $\Pr(A_j)$ denote the probability that v_j is not flawed by $T_{i,u}$. Thus the probability that at least one value is not flawed by $T_{i,u}$, i.e. the probability that the variable u is not flawed by $T_{i,u}$ is

$$\begin{aligned}\Pr(A_1 \cup A_2 \cup \cdots \cup A_d) &= \sum_{1\leq p\leq d} \Pr(A_p) - \sum_{1\leq p,q\leq d, p\neq q} \Pr(A_p A_q) \\ &\quad + \cdots + (-1)^{d-1} \Pr(A_1 A_2 \cdots A_d).\end{aligned}$$

Then

$$\begin{aligned}\Pr(T_{i,u} \text{ is flawed}) &= 1 - \Pr(A_1 \cup A_2 \cup \cdots \cup A_d) \\ &= 1 + \sum_{j=1}^{d} (-1)^j \binom{d}{j} \Pr(A_1 A_2 \cdots A_j).\end{aligned}$$

Recall that in Model RB, for each constraint, we uniformly select without repetition pd^k incompatible tuples of values and each constraint is generated independently. So we have

$$\begin{aligned}
\Pr(A_1A_2\cdots A_j) &= \left[\frac{\binom{d^k-j}{pd^k}}{\binom{d^k}{pd^k}}\right]^i \\
&= \left[\frac{(d^k-pd^k)(d^k-pd^k-1)\cdots(d^k-pd^k-j+1)}{d^k(d^k-1)\cdots(d^k-j+1)}\right]^i.
\end{aligned}$$

Note that $j \le d = n^\alpha$ and $k \ge 2$. Now consider the case of $i = 3k\beta \ln n$, where $\beta = \frac{\alpha}{6k\ln\frac{1}{1-p}}$. By asymptotic analysis, we have

$$\begin{aligned}
&\Pr(A_1A_2\cdots A_j)|_{i=3k\beta\ln n} \\
=\ &[(1-p)(\frac{1-p-\frac{1}{n^{k\alpha}}}{1-\frac{1}{n^{k\alpha}}})(\frac{1-p-\frac{2}{n^{k\alpha}}}{1-\frac{2}{n^{k\alpha}}})\cdots(\frac{1-p-\frac{j-1}{n^{k\alpha}}}{1-\frac{j-1}{n^{k\alpha}}})]^{3k\beta\ln n} \\
=\ &[(1-p)^{3k\beta\ln n}]^j[1-\frac{p}{1-p}\frac{(j-1)j}{2n^{k\alpha}}+O(\frac{j^4}{n^{2k\alpha}})]^{3k\beta\ln n} \\
=\ &(n^{-\frac{\alpha}{2}})^j[1-\frac{p}{1-p}\frac{(j-1)j}{2n^{k\alpha}}+O(\frac{j^4}{n^{2k\alpha}})]^{3k\beta\ln n}.
\end{aligned}$$

Let $H(j) = [1-\frac{p}{1-p}\frac{(j-1)j}{2n^{k\alpha}}+O(\frac{j^4}{n^{2k\alpha}})]^{3k\beta\ln n}$. Then we get

$$\begin{aligned}
\Pr(T_{i,u}\text{ is flawed})|_{i=3k\beta\ln n} &= 1+\sum_{j=1}^{n^\alpha}(-1)^j\binom{n^\alpha}{j}\Pr(A_1A_2\cdots A_j)|_{i=3k\beta\ln n} \\
&= 1+\sum_{j=1}^{n^\alpha}(-1)^j\binom{n^\alpha}{j}(n^{-\frac{\alpha}{2}})^jH(j).
\end{aligned}$$

For $0 \le j \le n^{\frac{4}{5}\alpha}$, we can easily show that $H(j) = 1+o(1)$. Therefore,

$$\begin{aligned}
&\Pr(T_{i,u}\text{ is flawed})|_{i=3k\beta\ln n} \\
\approx\ &1+\sum_{j=1}^{n^\alpha}(-1)^j\binom{n^\alpha}{j}(n^{-\frac{\alpha}{2}})^j+\sum_{j=n^{\frac{4}{5}\alpha}}^{n^\alpha}(-1)^j\binom{n^\alpha}{j}(n^{-\frac{\alpha}{2}})^j(H(j)-1) \\
=\ &(1-\frac{1}{n^{\frac{\alpha}{2}}})^{n^\alpha}+\sum_{j=n^{\frac{4}{5}\alpha}}^{n^\alpha}(-1)^j\binom{n^\alpha}{j}(n^{-\frac{\alpha}{2}})^j(H(j)-1) \\
\approx\ &e^{-n^{\frac{\alpha}{2}}}+\sum_{j=n^{\frac{4}{5}\alpha}}^{n^\alpha}(-1)^j\binom{n^\alpha}{j}(n^{-\frac{\alpha}{2}})^j(H(j)-1).
\end{aligned}$$

It is easy to verify that

$$\binom{n^\alpha}{j}(n^{-\frac{\alpha}{2}})^j \leq (\frac{en^\alpha}{j})^j(n^{-\frac{\alpha}{2}})^j = e^{j-j\ln j+\frac{\alpha}{2}j\ln n}.$$

Let $B(j) = j - j\ln j + \frac{\alpha}{2}j\ln n$. Differentiating $B(j)$ with respect to j, we obtain

$$B'(j) = \frac{\alpha}{2}\ln n - \ln j < 0 \text{ when } j \geq n^{\frac{4}{5}\alpha}.$$

So for $n^{\frac{4}{5}\alpha} \leq j \leq n^\alpha$, we have

$$\binom{n^\alpha}{j}(n^{-\frac{\alpha}{2}})^j \leq e^{B(n^{\frac{4}{5}\alpha})} = (\frac{e}{n^{\frac{3}{10}\alpha}})^{n^{\frac{4}{5}\alpha}} = o(e^{-n^{\frac{4}{5}\alpha}}).$$

Note that $H(j) = O(n^{c_2})$ for $n^{\frac{4}{5}\alpha} \leq j \leq n^\alpha$, where $c_2 > 0$ is a constant. Hence,

$$\begin{aligned} |\sum_{j=n^{\frac{4}{5}\alpha}}^{n^\alpha}(-1)^j\binom{n^\alpha}{j}(n^{-\frac{\alpha}{2}})^j(H(j)-1)| &\leq \sum_{j=n^{\frac{4}{5}\alpha}}^{n^\alpha}\binom{n^\alpha}{j}(n^{-\frac{\alpha}{2}})^j|H(j)-1| \\ &= O(n^\alpha)O(n^{c_2})o(e^{-n^{\frac{4}{5}\alpha}}) = o(e^{-n^{\frac{\alpha}{2}}}). \end{aligned}$$

Thus we get

$$\Pr(T_{i,u} \text{ is flawed})|_{i\leq 3k\beta\ln n} \leq \Pr(T_{i,u} \text{ is flawed})|_{i=3k\beta\ln n} \approx e^{-n^{\frac{\alpha}{2}}}.$$

The remaining part of the proof is identical to that of Lemma 2 for Model RD, and so we are done.