The solution space structure of planted constraint satisfaction problems with growing domains

Wei Xu and Zhe Zhang

School of Mathematics and Physics, University of Science and Technology Beijing, Beijing 100083, People’s Republic of China
E-mail: s20190769@xs.ustb.edu.cn

Received 3 August 2021
Accepted for publication 22 January 2022
Published 3 March 2022

Online at stacks.iop.org/JSTAT/2022/033401
https://doi.org/10.1088/1742-5468/ac4e86

Abstract. Planting a solution into the random revised B (RB) model, which is a prototype of the random constraint satisfaction problem (CSP) with growing domains, can generate very hard satisfiable CSP benchmarks. We study the solution space structure of the planted RB model. In the thermodynamic limit, by the first moment method we find that this model goes through four phase transitions as the constraint density increases. In the replica symmetric phase, what we call the independent phase transition occurs, after which the planted cluster (cluster containing the planted solution) is separated from the giant cluster. From then on, the solution space except the planted cluster goes through the same clustering phase transition and the same satisfiability phase transition as on the random RB model. The planted cluster goes through the isolated phase transition, after which the planted cluster contains only one solution. When the instances have about $10^2$ variables, experiments show that the last three phase transitions have already appeared as in the thermodynamic limit. This phase diagram provides strong evidence that this model can generate very hard satisfiable CSP benchmarks. Finally, we find when the constraint density goes to infinity, there is a single smooth valley in the energy landscape.

Keywords: optimization over networks, message-passing algorithms, random graphs, networks

*Author to whom any correspondence should be addressed.
1. Introduction

Constraint satisfaction problems (CSPs) play a significant role in computer science, statistical physics and mathematics. The average computational complexity of the CSP has been studied from the point of view of the spin glass theory [1], and fruitful results have been achieved. The planted CSP is also in the category of the inference problem, which has been intensely studied [2].

Timetabling, hardware configuration, factory scheduling, floor-planning and many other tasks can be solved by a unified process called ‘constraint programming’: firstly those tasks are modeled as CSPs and secondly the CPSs are solved by computer algorithms [3]. For a better understanding of this unified process, it is necessary to generate CSP instances randomly, and for this purpose the classical models A, B, C and D are proposed [4, 5]. But Achlioptas et al [6] found that those models suffer trivial asymptotic insolubility: asymptotically almost all of the instances they generate have no solutions. To overcome this deficiency, two main approaches are tried: one is incorporating some structures (e.g. arc-consistent), another is controlling the way parameters change as the problem size increases [7]. The random revised B (RB) model [8] with growing domains follows the second approach.

To benchmark algorithms, not only are pure random models required, but also models that only generate satisfiable instances. If an incomplete algorithm does not find a
solution for a satisfiable instance, there is no doubt that the algorithm fails. Many efforts have been done to generate satisfiable CSP instances, including that from the physical point of view [9–11]. Planting is a basic way to hide a solution: a solution $\vec{\sigma}_P$ to be hidden is chosen first, then only the constraints that are compatible with $\vec{\sigma}_P$ are chosen. In cryptographic application, hard planted CSP instances can serve as one-way functions.

Planting into the random RB model can generate very hard satisfiable CSP benchmarks [12]. They have been used in various kinds of algorithm competitions (e.g. CSP, SAT and MaxSAT), and the results confirmed the intrinsic hardness of these benchmarks. Based on this model, Xu [13] proposed an instance ‘frb100-40’ with 100 variables in 2005 and challenged that the instance cannot be solved on a PC in less than 24 h in 20 years. The challenge is still continuing after several researchers have tried [14, 15].

CSPs are related to the spin glass theory naturally. A CSP instance contains a set of discrete variables and a collection of constraints. A constraint restricts the joint values of some variables, so a constraint acts as an interaction among those variables. A Gibbs measure can be defined, where the energy of a configuration (an assignment of all variables) represents the number of constraints that the configuration violates. At zero temperature, the partition function is the number of solutions (configurations whose energy is zero). The distribution of the interactions is given in a CSP model.

From the point of view of statistical physics, the average computational hardness is related to the solution space structure. Using the cavity method, fruitful results on the phase diagrams have been achieved [16–19], and efficient algorithms are designed [20–23]. Usually a CSP model goes through the replica symmetric (RS) phase, where almost all solutions belong to a giant cluster; the dynamic one step replica symmetry breaking phase (or so called the clustering phase), where the solutions shatter into exponentially many clusters; the static one step replica symmetry breaking phase, etc.

In this paper, the first moment method is frequently used. Our results on solution space structure are established on typical instances, i.e. the results hold ‘with high probability (w.h.p.)’, which means ‘with probability $1 - o(1)$ when the problem size goes to infinity’. As we always study properties on typical instances, ‘w.h.p.’ is omitted if there is no ambiguity. We use the first moment method to locate phase transition points of the model based on two observations: firstly, the number of solutions are sharply concentrated (or to say has small fluctuations); secondly, experiments show that the solutions (before clustering transition) and the clusters (after clustering transition) distribute widely and relatively uniformly in the solution space. Those observations suggest that the instances under the same parameters are quite similar to each other, and the first moment is useful in locating the transition points.

In this paper, we find that in the thermodynamic limit the planted RB model goes through the independent, clustering, satisfiability (by configurations outside the planted cluster), and isolated phase transitions. The phase diagram is drawn in figure 4. The small-size instances are studied by experiments, which show that the last three transitions occur at the same locations as in the thermodynamic limit. The rest of the paper is organized as follows: we give definitions of the random RB model and the planted RB model in section 2, then study the independent phase transition and the satisfiability (by configurations outside the planted cluster) phase transition in section 3, study the clustering phase transition in section 4, study the isolated phase transition in section 5,
2. Definitions of the random and planted RB model

An instance of the RB model is comprised of \( n \) variables and \( t \) constraints. All the \( n \) variables take values from a domain \( D = \{1, 2, \ldots, d\} \), where \( d = n^\alpha, \alpha > 0 \). Each of the \( t \) constraints involves \( k \) variables and restricts the tuples of values of the \( k \) variables. For a constraint, the set of compatible tuples of values is a subset of \( D^k \). A tuple of values of all of the \( n \) variables \( \vec{\sigma} = (\sigma_1, \ldots, \sigma_n) \) is called an assignment (configuration). An assignment is a solution if it satisfies all of the \( t \) constraints. We give the definitions of the (pure) random RB model and the planted RB model by their steps to generate instances. A random RB instance is generated by the following two steps [8]:

(a) Select with repetition \( t = rn \ln n \) random constraints. Each constraint is formed by randomly selecting without repetition \( k \) of \( n \) variables.

(b) For each constraint, randomly select without repetition \( (1 - p)d^k \) (\( 0 < p < 1 \) measures the tightness of the constraint) compatible tuples of values.

A planted RB instance is generated by the following three steps:

(a) Choose a random assignment \( \vec{\sigma}_P \) to be the planted solution.

(b) This step is as same as step (a) of the random RB model.

(c) For each constraint, randomly select without repetition \( (1 - p)d^k \) compatible tuples of values, where the tuple of values in \( \vec{\sigma}_P \) must be included.

For the random RB model, when \( p < 1 - \frac{1}{k} \) and \( \alpha > 1/k \), the satisfiability transition occurs at \( r_s \) [8]:

\[
r_s = -\frac{\alpha}{\ln(1 - p)}.
\]

In the following, we let \( r = r_0r_s \), then the satisfiability transition occurs at \( r_0 = 1 \). In the following, we only consider models under the conditions: \( p < 1 - \frac{1}{k} \) and \( \alpha > 1/k \).

We should list some definitions which follow [24]. The (Hamming) distance between two assignments (configurations) \( \vec{\sigma}_A \) and \( \vec{\sigma}_B \) is the number of variables where \( \vec{\sigma}_A \) and \( \vec{\sigma}_B \) take different values. \( \vec{\sigma}_A \) and \( \vec{\sigma}_B \) are connected if and only if the distance between them is 1. A cluster is a connected component of solutions. A cluster-region is a set of clusters. The diameter of a cluster-region is the biggest distance between two solutions in the region.

In section 3, we will find that after the independent phase transition, a set of solutions centered on \( \vec{\sigma}_P \) is separated from other solutions. It only includes solutions that are at a distance smaller than \( \epsilon n \) from \( \vec{\sigma}_P \), where \( \epsilon \) is an arbitrarily small positive constant. Because the diameter of this set of solutions is very small, we might as well call the set a cluster: the planted cluster.
3. The independent phase transition and the transition at $r_s$

In this section, we will show that, in the RS phase, what we call the independent phase transition occurs, before which the planted solution belongs to the giant cluster, and after which the planted cluster (cluster containing the planted solution) is separated from the giant cluster.

Let $E(Y(x))$ be the average number of solutions being at distance $xn$ from the planted solution $\vec{\sigma}_P$, then by the definition of the planted RB model,

$$E(Y(x)) = \binom{n}{x_n, n-x_n} (d-1)^{x_n} r^n \ln n,$$

where

$$\hat{p} = \frac{\binom{n-x_n}{k, n-x_n-k}}{\binom{n}{k, n-k}} + \frac{(1-p)d^k - 1}{d^k - 1} \left(1 - \frac{\binom{n-x_n}{k, n-x_n-k}}{\binom{n}{k, n-k}}\right),$$

and $\binom{n}{m, n-m}$ represents the combination formula that equals to $\frac{n!}{m!(n-m)!}$. Let

$$h(x) = \lim_{n \to \infty} \frac{\ln E(Y(x))}{\alpha n \ln n}.$$ 

For constant $x$, $0 < x < 1$, with $n \to \infty$, we have

$$\frac{(1-p)d^k - 1}{d^k - 1} \to 1 - p,$$

$$\frac{\binom{n-x_n}{k, n-x_n-k}}{\binom{n}{k, n-k}} \to (1-x)^k,$$

and

$$\ln \left(\frac{n}{x_n, n-x_n}\right) \to -\ln(x^x(1-x)^{1-x}),$$

where the last one is from the Stirling formula. Then we have

$$h(x) = \lim_{n \to \infty} \left[\frac{-\ln(x^x(1-x)^{1-x})}{\alpha \ln n} + \frac{x \ln(d-1)}{\alpha \ln n} + \frac{r_0}{-\ln(1-p)} \ln(1-p + p(1-x)^k)\right]$$

$$= x + \frac{r_0}{-\ln(1-p)} \ln(1-p + p(1-x)^k).$$

https://doi.org/10.1088/1742-5468/ac4e86
The solution space structure of planted constraint satisfaction problems with growing domains

Figure 1. Function images of $h(x)$ when $k = 2$, $p = 0.4$ and $r = 0.6r_s$, $r = 0.8r_s$, $r = 1r_s$, $r = 1.2r_s$.

In figure 1, we draw the function images of $h(x)$ for $k = 2$, $p = 0.4$ and different $r_0$. The first and second order derivatives of $h(x)$ is

$$\frac{dh(x)}{dx} = 1 + \frac{r_0 pk(1 - x)^{k-1}}{\ln(1 - p)(1 - p + p(1 - x)^k)},$$

$$\frac{d^2h(x)}{dx^2} = \frac{r_0 pk(1 - x)^{k-2}[(k - 1)(1 - p) - p(1 - x)^k]}{(-\ln(1 - p))(1 - p + p(1 - x)^k)^2}.$$ 

Under the condition $k \geq \frac{1}{1-p}$, we have $(k - 1)(1 - p) - p \geq 0$ and $(k - 1)(1 - p) - p(1 - x)^k > 0$ ($0 < x < 1$), then $h(x)$ is concave:

$$\frac{d^2h(x)}{dx^2} > 0.$$ 

We study the following three situations:

(a) $\left.\frac{dh(x)}{dx}\right|_{x=0} > 0$. It is easy to verify that $h(0) = 0$. If $\left.\frac{dh(x)}{dx}\right|_{x=0} > 0$, then $h(x) > 0$ for all constant $x$, $0 < x < 1$.

(b) $\left.\frac{dh(x)}{dx}\right|_{x=0} < 0$ and $r < r_s$ ($r_0 < 1$). We know: $h(0) = 0$, $h(1) = 1 - r_0 > 0$, $\left.\frac{dh(x)}{dx}\right|_{x=0} < 0$, and $h(x)$ is concave. The function image of $h(x)$ will be like the $r = 0.8r_s$ line in figure 1. Let $\epsilon$ be an arbitrarily small positive constant; let $a$ be the solution of $h(x) = 0$ for $0 < x < 1$; let $b$ be a little smaller than $a$; and let $h_0 = \max\{h(\epsilon), h(b)\}$. We have $h_0 < 0$ and $h(x) \leq h_0$ for $\epsilon \leq x \leq b$. Then

$$\lim_{n \to \infty} E \left( \sum_{\epsilon n \leq x_n \leq bn} Y(x) \right) = \lim_{n \to \infty} \sum_{\epsilon n \leq x_n \leq bn} E(Y(x)) \leq \lim_{n \to \infty} n e^{h_0 \alpha n \ln n} = 0.$$ 

By the first moment method we have w.h.p. $\sum_{\epsilon n \leq x_n \leq bn} Y(x) = 0$. 

https://doi.org/10.1088/1742-5468/ac4e86
The solution space structure of planted constraint satisfaction problems with growing domains

(c) \( r > r_s \) \((r_0 > 1)\). We know: \( h(0) = 0, h(1) = 1 - r_0 < 0, \) and \( h(x) \) is concave. For an arbitrarily small positive constant \( \epsilon \), we have \( h(x) \leq h_1 \) for \( \epsilon \leq x \leq 1 \), where \( h_1 = \max\{h(\epsilon), h(1)\} < 0 \). Then

\[
\lim_{n \to \infty} E \left( \sum_{\epsilon n \leq x \leq n} Y(x) \right) = \lim_{n \to \infty} \sum_{\epsilon n \leq x \leq n} E(Y(x)) \leq \lim_{n \to \infty} n e^{h_{1 \text{con}} \ln n} = 0.
\]

By the first moment method, we have that w.h.p. \( \sum_{\epsilon n \leq x \leq n} Y(x) = 0 \).

**The independent phase transition.** In situation (b), we find for typical instances, the solutions being at distance \( x n (\epsilon n \leq x n \leq bn) \) from the planted solution \( \vec{\sigma}_P \) do not exist, where \( \epsilon \) is an arbitrarily small positive constant. The solution space can be split into two parts: the planted cluster and other solutions. The planted cluster is very small because \( \epsilon \) can take an arbitrarily small positive value.

Equation \( \frac{d h}{d x} \big|_{x=0} = 0 \) defines the independent transition point, denoted by \( r_{id} \). When \( r < r_{id} \), the planted solution is in the giant cluster (this is a reasonable but not a mathematically proven conclusion in this paper). When \( r > r_{id} \), the planted cluster is separated from the giant cluster. If the gap between the planted cluster and the giant cluster is the order parameter, the transition is of the second order. Solving equation \( \frac{d h}{d x} \big|_{x=0} = 0 \), we have \( r_0 = \frac{-\ln(1-p)}{kp} \), so

\[
r_{id} = r_0 r_s = \frac{-\ln(1-p)}{kp} r_s = \frac{\alpha}{kp}.
\]

At the end of section 4, we show that this transition occurs before the clustering transition.

**The satisfiability (by configurations outside the planted cluster) phase transition.** In situation (c), we find for typical instances, no solutions are at a distance bigger than \( cn \) from \( \vec{\sigma}_P \), where \( \epsilon \) is an arbitrarily small positive constant. It is to say \( r_s \) is a transition point: before which, the model is in the clustering phase, seeing section 4; after which, (for typical instances) no solutions exist except that in the planted cluster.

### 4. The clustering phase transition

In this section, we will show that the planted RB model has the same clustering phase (where the solutions shatter into exponentially many clusters) as the random RB model. The method that we will apply has no differences to what has been used on the random RB model in [25]. The method focuses on the number of solution-pairs at certain distances. If there are certain \( a_0 \) and \( b_0 \), such that solution-pairs at Hamming distances \( x n \) do not exist for any \( a_0 n < x n < b_0 n \), then by the method the solution space shatters into cluster-regions, where the diameter of a cluster-region is at most \( a_0 n \) and the distance between two cluster-regions is at least \( b_0 n \). This method has been applied on random graph coloring, random k-SAT, random hypergraph two-coloring, random RB model, and random \( d\)-\( k \)-CSP model, etc [24–29].

https://doi.org/10.1088/1742-5468/ac4e86
Denote the expectation of the number of solution-pairs at distance $x_n$ by $\mathbb{E}(Z(x))$, and define its normalised version as

$$f(x) = \lim_{n \to \infty} \frac{\ln(\mathbb{E}(Z(x)))}{n \ln n}. \quad (4)$$

As long as $f(x) < 0$, we can conclude that the typical instances do not have solution-pairs at a distance $x_n$. In the appendix, we show that

$$f(x) = \max(f_1(x), f_2(x)), \quad (5)$$

where

$$f_1(x) = \alpha(1 + x) + r \ln[(1 - p)^2 + (1 - p)p(1 - x)^k], \quad (6)$$

$$f_2(x) = 2\alpha x + r \ln[(1 - p)^2 + (2p - p^2)(1 - x)^k]. \quad (7)$$

We take $k = 2$, $p = 0.47$, $r = 0.98r_s$ as an example, and draw the function images of $f_1(x)/\alpha$ and $f_2(x)/\alpha$ in figure 2. In the figure, we find that when $x \in (0.038, 0.958)$ we have $f(x) < 0$, which means that the typical instances do not have solution-pairs at distances between $0.038n$ and $0.958n$. The solution space can be divided into cluster-regions, where the diameter of a cluster-region is at most $0.038n$ and the distance between two cluster-regions is at least $0.958n$ [25].

Let $g(x)$ be the counterpart of $f(x)$ on the random RB model, i.e. the logarithm of the expectation of the number of solutions pairs at distance $x_n$ divided by $n \ln n$ on the random RB model, then equation (5) in [25] shows that

$$g(x) = f_1(x). \quad (8)$$

We draw a lot of function images of $f_1(x)$ and $f_2(x)$ for different parameters $k \geq 2$, $0 < p < 1 - 1/k$ and $0 < r < r_s$. Numerical calculation shows that $f_1(x)$ and $f(x)$ have
The solution space structure of planted constraint satisfaction problems with growing domains

Figure 3. Function images of \( f_1(x)/\alpha \) and \( f_2(x)/\alpha \) when \( k = 2, p = 0.4 \) and \( r = 0.98r_s \).

the same positive part, i.e. \( f_1^+(x) = f^+(x) \), then

\[
g^+(x) = f^+(x). \tag{9}\]

In the most cases, \( f_1(x) > f_2(x) \) for \( 0 \leq x \leq 1 \), as shown in figure 3. In a few cases, \( f_1(x) < f_2(x) \) for some \( x \) but for those \( x \) we have \( f_1(x) < f_2(x) < 0 \), as shown in figure 2. In all cases, we have that \( f_1^+(x) > f_2^+(x) \) then \( f_1^+(x) = f^+(x) \) and \( g^+(x) = f^+(x) \).

Equation (9) means that inequalities \( f(x) < 0 \) and \( g(x) < 0 \) hold in the same interval of \( x \). By the method splitting the solution space, it means that the planted RB model has the same clustering phase as the random model. The clustering phase shows up when \( r \in (r_d, r_s) \), where \( r_d \) is the smallest value of \( r \) for which \( g(x) = 0 \) has at least one solution for \( x \in [0, 1] \).

The number of solutions in a cluster-region is limited by the diameter of the cluster-region. In [25], it is found that the typical instances of random RB model have a lot solutions (when \( r < r_s \)), but one cluster-region contains only an exponentially small proportion of them. For the planted model, the diameter of cluster-region remains unchanged. But the instances of the planted model are generated with probabilities proportional to the numbers of their solutions, so they tend to have more solutions than that are generated from the random model. We conclude that for the planted model, the number of cluster-regions is exponential and one cluster-region contains an exponentially small proportion of solutions.

It is easy to verify that \( g(x) = \alpha(h(x) + (1 - r_0)) \), where \( h(x) \) is defined in (3). We have \( g(x) > \alpha h(x) \) for \( r_0 < 1 \). Because of this relation and the properties of \( h(x) \), as \( r_0 \) grows, \( \left. \frac{dh}{dx} \right|_{x=0} = 0 \) occurs before ‘\( g(x) = 0 \) has at least one solution for \( x \in [0, 1] \)’. This means that the independent phase transition is before the clustering phase transition.
The solution space structure of planted constraint satisfaction problems with growing domains

5. The isolated phase transition

In this section, we show that as $r$ grows, the planted cluster goes through a transition, after which the planted cluster contains only the planted solution. We call it the isolated transition.

The average number of solutions being at a distance $x_n$ from the planted solution $\bar{\sigma}_P$, $E(Y(x))$, was given in equation (1). For any $x_n = 1, 2, \ldots, n$, we have

$$
\left( \frac{n - x_n}{k, n - x_n - k} \right) \frac{n}{k, n - k} = \frac{(n - x_n) \ldots (n - x_n - k + 1)}{n \ldots (n - k + 1)} < (1 - x)^k.
$$

Combining

$$
\frac{(1 - p)d^k - 1}{d^k - 1} < 1 - p,
$$

we have for any $x_n = 1, 2, \ldots, n$,

$$
E(Y(x)) < n^{rn}d^{rn}(1 - p + p(1 - x)^k)^{rn \ln n},
$$

and

$$
\ln E(Y(x)) < 1 + \alpha + r \ln (1 - p + p(1 - x)^k) / x.
$$

Because

$$
\lim_{x \to 0} \frac{\ln (1 - p + p(1 - x)^k)}{x} = -kp,
$$

we have

$$
\lim_{x \to 0} \frac{\ln E(Y(x))}{x_n \ln n} \leq 1 + \alpha - rkp.
$$

Let $\omega = 1 + \alpha - rkp$. When $\omega < 0$, there is a constant number $\epsilon$, such that for $0 < x \leq \epsilon$ we have

$$
\ln E(Y(x)) < \frac{1}{2} \omega.
$$

Then we have

$$
E \left( \sum_{1 \leq x_n \leq cn} Y(x) \right) = \sum_{1 \leq x_n \leq cn} E(Y(x)) \leq \sum_{x_n = 1, 2, \ldots} E(Y(x)) < n^{\frac{1}{2} \omega} + n^{2^{\frac{1}{2} \omega}} + \ldots = \frac{n^{\frac{1}{2} \omega}}{1 - n^{\frac{1}{2} \omega}} \xrightarrow{n \to \infty} 0.
$$

https://doi.org/10.1088/1742-5468/ac4e86
The solution space structure of planted constraint satisfaction problems with growing domains

Figure 4. Phase diagram of the planted RB model in the \( r_{is} > r_s \) case. The planted RB model goes through the independent transition at \( r_{id} = \frac{\alpha}{kp} \), the clustering transition at \( r_d \), the satisfiability (by configurations outside the planted cluster) transition at \( r_s = -\frac{\alpha}{\ln(1-p)} \), and the isolated transition at \( r_{is} = \frac{1+\alpha}{kp} \). \( r_d \) is the smallest value of \( r \) for which \( g(x) = 0 \) has at least a solution in \( x \in [0,1] \), where \( g(x) \) is defined in equation (8). The plant solution is the red point. When \( r > r_{id} \), the diameter of the planted cluster is smaller than \( \epsilon n \) for any constant \( \epsilon > 0 \).

By the first moment method, as long as \( \omega < 0 \) (i.e. \( r > \frac{1+\alpha}{kp} \)), we have w.h.p. \( \sum_{1 \leq x_n \leq \epsilon n} Y(x) = 0 \). It means that, for typical instances, there are no solutions at a distance smaller than \( \epsilon n \) (\( \epsilon \) is a positive constant number depending on \( \omega \)) from the planted solution, except the planted solution itself. We call \( r_{is} = 1 + \frac{\alpha}{kp} \) the isolated transition point, after which the planted cluster shrinks into a point (the planted solution). \( r_{is} = \frac{1+\alpha}{\alpha r_{id}} \), so the isolated transition occurs after the independent transition. \( r_d \) and \( r_{is} \) do not have a certain order. If \( r > r_s \) and \( r > \frac{1+\alpha}{kp} \) hold simultaneously, for typical instances, the planted solution is the only solution.

6. Phase diagram in the thermodynamic limit

By the analyses in the above three sections, in the thermodynamic limit, we find four phase transitions of the planted RB model: the independent transition at \( r_{id} = \frac{\alpha}{kp} \), the clustering transition at \( r_d \), the satisfiability (by configurations outside the planted cluster) transition at \( r_s = -\frac{\alpha}{\ln(1-p)} \), and the isolated transition at \( r_{is} = \frac{1+\alpha}{kp} \). \( r_d \) is the smallest value of \( r \) for which \( g(x) = 0 \) has at least one solution in \( x \in [0,1] \), where \( g(x) \) is defined in equations (6) and (8).

The phase diagram in the \( r_{is} > r_s \) case is drawn in figure 4. Below \( r = r_{id} \), most of the solutions, including the planted solution, belong to a giant cluster. Above \( r = r_{id} \), the planted cluster is separated from the giant cluster, and the distance between them is linear to \( n \). The diameter of the planted cluster is smaller than \( \epsilon n \), where \( \epsilon \) is an arbitrarily small positive constant. From then on, the planted cluster and other solutions go through different transitions. Above \( r = r_{id} \), solutions except that in the planted cluster shatter into exponentially many clusters, and each cluster contains only
The solution space structure of planted constraint satisfaction problems with growing domains

Figure 5. Phase diagram of the planted RB model around $r_s$ in the $r_{is} < r_s$ case. The planted solution (the red point) is isolated.

an exponentially small proportion of solutions. Above $r = r_s$, no solutions exist except that in the planted cluster. Above $r = r_{is}$, the planted cluster has only the planted solution left.

From the above three sections, we have the relations: $r_{id} < r_{ds}, r_{is} < r_{ds}$ and $r_{ds} < r_{is}$ are all possible, depending on the parameters $\alpha, p, k$. In figure 5, we show the phase diagram around $r = r_s$ in the $r_{is} < r_s$ case. In this case, the planted solution is the only solution when $r > r_s$.

The phase diagram provides strong evidence that this model can generate very hard satisfiable CSP benchmarks at $r = r_s$. (1) At $r = r_s$, it seems that there are exponentially many clusters that are distributed widely in the solution space, so there are many places that are worth searching. (2) At $r = r_s$, it also seems that only one cluster (the planted cluster) remains, so many of the above ‘clusters’ could only be traps. (3) The planted cluster and other clusters (if they exist) are all very small, so they are hard to find.

We propose a condition under which the problem should be even harder (at $r = r_s$). The condition is $r_{is} < r_s$ or equivalently

\[
\frac{1 + \alpha}{kp} < -\frac{\alpha}{\ln(1 - p)}.
\]

Under this condition, the phase diagram around $r = r_s$ has been shown in figure 5. The planted solution is isolated, so it should be difficult to find.

7. Energy landscape in the $r \to \infty, n \to \infty$ limit

If instances have very large constraint densities, it is known that simple message passing algorithms can find solutions in polynomial time [30, 31]. Experiments show that on the planted RB model, if $r$ has exceeded $r_s$ and keeps growing, both message passing algorithms and local search algorithms can succeed again. In this section, we show that in the $r \to \infty, n \to \infty$ limit there is a single smooth valley in the energy landscape. This energy landscape means that the instances are tractable by local search.
We study the energy of configurations, which is the number of constraints that the configurations violate. For the planted RB model, let $E^{x,\lambda}(X)$ be the average number of configurations with energy $\lambda t$ and at a distance $xn$ from $\tilde{\sigma}_P$, then

$$E^{x,\lambda}(X) = \left(\frac{n}{x_n, n-x_n}\right)(d-1)^{nx} \left(\frac{t}{\lambda t - \lambda t}\right) \hat{p}^{(1-\lambda)p}(1-\hat{p})^\lambda,$$

where $\hat{p}$ is defined in (2) and $t = rn \ln n$. Let

$$f(x, \lambda) = \lim_{n \to \infty} \frac{\ln E^{x,\lambda}(X)}{n \ln d},$$

we can simplify the value of $f(x, \lambda)$ for the constant number $x$:

$$f(x, \lambda) = x + \phi(\lambda)r_0,$$

where

$$\phi(\lambda) = \frac{1}{\ln(1-p)} \ln(\lambda^\lambda(1-\lambda)^{1-\lambda}) - (1-\lambda)\frac{\ln(1-p+p(1-x)^k)}{\ln(1-p)} - \frac{\lambda \ln(p-p(1-x)^k)}{\ln(1-p)}.$$ 

Calculating the first and the second order derivatives, we have

$$\frac{\partial \phi(\lambda)}{\partial \lambda} = \frac{1}{\ln(1-p)} [\ln \lambda - \ln(1-\lambda)] + \frac{\ln(1-p+p(1-x)^k)}{\ln(1-p)} - \frac{\ln(p-p(1-x)^k)}{\ln(1-p)},$$

$$\frac{\partial^2 \phi(\lambda)}{\partial \lambda^2} = \frac{1}{\ln(1-p)} \left(\frac{1}{\lambda} + \frac{1}{1-\lambda}\right) < 0.$$ 

We find that $\phi(\lambda)$ is convex, and $\phi(\lambda)$ achieves its maximum value 0 at $\lambda = p - p(1-x)^k$.

For the constant number $0 \leq x \leq 1$, let $\lambda^* = p - p(1-x)^k$ and $r^* = \frac{\lambda^*}{\lambda}$. When $\lambda \neq \lambda^*$ and $r > r^*r_s$, we have $f(x, \lambda) < 0$. By the first moment method, $f(x, \lambda) < 0$ means that (w.h.p.) configurations with energy $\lambda t$ and at a distance $xn$ from $\tilde{\sigma}_P$ do not exist.

So when $r$ is big enough, the configurations at a distance $xn$ from the planted solution $\tilde{\sigma}_P$ have a certain energy

$$\lambda^* t = (p - p(1-x)^k)t,$$

which equals $(2x - x^2)pt$ when $k = 2$. The energy equals 0 when $x = 0$, increases as $x$ grows, and equals $pt$ when $x = 1$. In the $r \to \infty$, $n \to \infty$ limit, there is a single smooth valley in the energy landscape.

In the left picture of figure 6, a schematic picture for energy landscape when $k = 2$, $r \to \infty$, $n \to \infty$ is shown. In the right picture of figure 6, we use pure min-conflicts heuristic to find a solution, and only reassign a variable if the energy can be reduced. In each step, the energy is reduced, so the algorithm finishes soon whether or not a solution can been found. We find that, when $n = 200$, $p = 0.4$, $\alpha = 0.6$, $k = 2$, the probabilities of finding a solution increase to about 1 at about $r_0 = 14$, and the numbers of steps to
find a solution decrease as \( r_0 \) grows. It indicates that, as \( r_0 \) grows, the energy landscape is becoming more and more like that in the \( r \to \infty, n \to \infty \) limit.

8. Phase diagram of small-size instances

The above discussions are in the thermodynamic limit, but to benchmark algorithms the sizes of the planted RB instances only need to be small, for example \( n = 10^2 \). The small-size instances are already very hard to solve (when \( r \approx r_s \)) because of that the domain size \( d \) grows with \( n \). In this section, we study the small-size instances (e.g. \( n = 50, n = 100, n = 200 \)) by experiments, and find that they have the similar clustering, satisfiability (by configurations outside the planted cluster) and isolated phase transitions as in the thermodynamic limit, but the independent transition at \( r_{id} = \frac{-\ln(1-p)}{pk} r_s \) does not occur.

The instances that we will use fall into three cases, where \( k \) is always set to be 2. The domain sizes \( d \) and the constraint numbers \( t \) of the three cases are listed in table 1. The transition points in the \( n \to \infty \) limit are listed in table 2. For the same \( r_0 \), instances in the first case have more than twice as many constraints as in the other two cases. For the same \( n, n \geq 50 \), variables in the third case have more than twice as many values in their domains as in the other two cases. We use different cases of instances in order to be more representative.

We use the reinforced belief propagation (RBP, referring to [32, 33]) algorithm and the WMCH (min conflicts heuristic with random walk, referring to chapter 5 of [3]).
algorithm to find solutions. RBP is a modified version of belief propagation decimation (BPD), and the idea is to build solutions according to marginals estimated by belief propagation (BP) iterations. BP iterations are

\[ \nu_{i \rightarrow a}(\sigma_i) = \frac{1}{Z_{i \rightarrow a}} \prod_{b \in \partial i \setminus a} \hat{\nu}_{b \rightarrow i}(\sigma_i), \]  

(10)

\[ \hat{\nu}_{a \rightarrow i}(\sigma_i) = \frac{1}{Z_{a \rightarrow i}} \sum_{\sigma_{a \setminus i}} \psi_a(\sigma_{a \setminus i}) \prod_{k \in \partial a \setminus i} \nu_{k \rightarrow a}(\sigma_k), \]  

(11)

where \( \nu_{i \rightarrow a}(\sigma_i) \) is the message from variable \( i \) to constraint \( a \), \( \hat{\nu}_{a \rightarrow i}(\sigma_i) \) is the message from the constraint \( a \) to variable \( i \), at step \( \eta \). \( \psi_a \) equals 1 if constraint \( a \) is satisfied by \( \sigma_{a \setminus i} \), and equals 0 otherwise. \( \partial a \) denotes the set of variables connected to constraint \( a \), and \( \partial a \setminus i \) is the same set removing \( i \). \( Z_{i \rightarrow a}^{(\eta)} \) and \( Z_{a \rightarrow i}^{(\eta)} \) are normalization constants, and here if they equal to 0, we set the messages equal to 1/\( d \) for all \( \sigma_i = 1, \ldots, d \). In the RBP, an external field is introduced enforcing the marginals toward a solution. The same as that in [34], here the update rule (10) is modified with probability \( 1 - \xi \) into

\[ \nu_{i \rightarrow a}^{(\eta+1)}(\sigma_i) = \frac{1}{Z_{i \rightarrow a}^{(\eta)}} \mu_i^{(\eta)}(\sigma_i) \prod_{b \in \partial i \setminus a} \hat{\nu}_{b \rightarrow i}^{(\eta)}(\sigma_i), \]  

(12)

\[ \mu_i^{(\eta)}(\sigma_i) = \frac{1}{Z_i^{(\eta)}} \prod_{b \in \partial i} \hat{\nu}_{b \rightarrow i}^{(\eta-1)}(\sigma_i), \]  

(13)

where \( Z_{i \rightarrow a}^{(\eta)} \) and \( Z_i^{(\eta)} \) are normalization constants. At step \( \eta \), the iterations follow update rules (10) and (11) with probability \( \xi \); and follow update rules (11)–(13) with probability

Table 1. The sizes of domain \( d \) and the numbers of constraints \( t \) (when \( r = r_s \)) for different \( p, \alpha \) and \( n \).

| Cases          | \( d \) when \( n = 50, n = 100, n = 200 \) | \( t \) (\( r = r_s \)) when \( n = 50, n = 100, n = 200 \) |
|---------------|------------------------------------------|--------------------------------------------------|
| \( p = 0.2, \alpha = 0.6 \) | 10, 16, 24                               | 516, 1243, 2848                                  |
| \( p = 0.4, \alpha = 0.6 \) | 10, 16, 24                               | 225, 543, 1244                                  |
| \( p = 0.4, \alpha = 0.8 \) | 23, 40, 69                                | 307, 722, 1658                                  |

Table 2. Transition points for different \( p \) and \( \alpha \) when \( k = 2 \), in the \( n \to \infty \) limit.

| Cases          | \( r_{id} = \frac{-\ln(1-p)}{pk} r_s \) | \( r_d \) | \( r_s = -\frac{\alpha}{\ln(1-p)} \) | \( r_{ls} = \frac{-(1+\alpha)\ln(1-p)}{akp} r_s \) |
|---------------|------------------------------------------|-----------|--------------------------------|-------------------------------------------|
| \( p = 0.2, \alpha = 0.6 \) | 0.558r_s                                | 0.866r_s  | 2.689r_s                      | 1.488r_s                                  |
| \( p = 0.4, \alpha = 0.6 \) | 0.639r_s                                | 0.883r_s  | 1.174r_s                      | 1.703r_s                                  |
| \( p = 0.4, \alpha = 0.8 \) | 0.639r_s                                | 0.883r_s  | 1.5661r_s                     | 1.437r_s                                  |
The solution space structure of planted constraint satisfaction problems with growing domains

Here, we set $\xi = \eta^{-0.1}$. At step $\eta$, the algorithm estimates marginals by

$$\nu_i^{(\eta)}(\sigma_i) = \frac{1}{\tilde{Z}_i^{(\eta)}} \prod_{b \in \partial i} \tilde{\nu}_b^{(\eta)}(\sigma_i),$$

where $\tilde{Z}_i^{(\eta)}$ is the normalization constant, then verifies whether the marginals guided assignment $\tilde{\sigma} = \{\arg \max_{\sigma_i} \nu_i^{(\eta)}(\sigma_i)\}$ is a solution. If the marginals guided assignment is a solution, the RBP succeeds. If the maximum number of iterations (which is set to be 1000 here) is reached and no solution is found, the RBP fails.

WMCH is a local search algorithm, which combines random walk and min-conflicts heuristic. Initializing in a random assignment (configuration), at each search step WMCH runs a random walk step with probability $\tau$, and runs a min-conflicts heuristic step with probability $1 - \tau$. Here we set $\tau = 0.1$. At each step, the conflict set is updated, which is the set of variables that appears in at least one unsatisfied constraint under the current assignment. At the random walk step, a randomly selected variable from the conflict set is reassigned a random value from its domain. At the min-conflicts heuristic step, a randomly selected variable from the conflict set is reassigned a value that minimizes the number of unsatisfied constraints. If the maximum number of iterations (which is set to be $200^*n$ here) is reached and no solution is found, the WMCH fails.

Firstly, the probabilities of finding a solution by RBP and WMCH are shown in figure 7, where the instances are in three cases: $\alpha = 0.6, p = 0.4$; $\alpha = 0.6, p = 0.2$; and $\alpha = 0.8, p = 0.4$, as in tables 1 and 2. From figure 7, we can find that:

(a) RBP tends to fail at about $r_d$, WMCH tends to fail before $r_d$. This indicates that the solution space has a rapid change at $r_d$, and we suppose that it is the clustering transition as in the $n \to \infty$ limit.

(b) RBP is not suitable for small $r$. When $r$ is small, marginals from BP are almost uniform, and the intensity of the external field added in RBP is small, so the marginals guided assignment will not always be a solution. But experiments show that BPD can find a solution for small $r$.

(c) If $r$ has exceeded $r_s$ and keeps growing, both algorithms can succeed again. If the constraint densities are very large, the instances are tractable, as discussed in section 7.

Secondly, the average overlaps between the found solutions and the planted solutions divided by $n$ are shown in figure 8. For more details, we take $n = 50, \alpha = 0.6, p = 0.2$ as an example, and draw the histograms of the overlaps divided by $n$ for $r_0 = 0.85$ and $r_0 = 0.95$ in figure 9. We can find that:

(a) Before $r_d$ the found solution is far away from the planted solution $\tilde{\sigma}_P$.

(b) Between $r_d$ and $r_s$ the overlap grows; at $r_s$ the found solution becomes very close to $\tilde{\sigma}_P$.

(c) At about $r_u$ the found solution totally overlaps with $\tilde{\sigma}_P$.

(d) In figure 9, the overlaps are centered on about $0.2n$ before $r_d$, and reach the maximum at about $0.8n$ after $r_d$.
The solution space structure of planted constraint satisfaction problems with growing domains

Figure 7. Probabilities of finding a solution by RBP and WMCH. (Top left) $\alpha = 0.6, p = 0.4, n = 50, 100, 200$, RBP. (Top right) $\alpha = 0.6, p = 0.4, n = 50, 100$, WMCH. (Bottom left) $\alpha = 0.6, p = 0.2, n = 50, 100$, RBP and WMCH. (Bottom right) $\alpha = 0.8, p = 0.4, n = 50, 100$, RBP and WMCH. 100–300 instances are used for each point. $r_d, r_s, r_{is}$ are the phase transition points in $n \to \infty$ limit.

This indicates that, before $r_d$, the found solution belongs to the giant cluster where most solutions are far away from $\vec{\sigma}_P$. After $r_d$ the solution space shatters into many small clusters (including the planted cluster), and the found solution belongs to those small clusters. The planted cluster is a relatively big cluster, so in the right histogram of figure 9, it is found the most times (where the overlap is about 0.8$n$). The small clusters are biased to $\vec{\sigma}_P$: the closer the clusters are to the $\vec{\sigma}_P$, the larger they tend to be and the easier they are to find. After $r_s$, only the planted cluster remains, so the found solution is in the planted cluster and the found solution is very close to $\vec{\sigma}_P$. After $r_{is}$, only the planted solution $\vec{\sigma}_P$ remains, so the found solution totally overlaps with $\vec{\sigma}_P$. The transitions at $r_d, r_s$, and $r_{is}$ should be similar to that in figure 4, except that the planted cluster is much bigger and the small clusters are biased to $\vec{\sigma}_P$.

Thirdly, the average percentages of frozen variables, which are obtained by running the whitening procedure starting from the found solutions, are shown in

https://doi.org/10.1088/1742-5468/ac4e86
The solution space structure of planted constraint satisfaction problems with growing domains

Figure 8. The average overlaps between the found solutions and the planted solutions divided by \( n \). (Top left) \( \alpha = 0.6, p = 0.4, n = 50, 100, 200 \), RBP. (Top right) \( \alpha = 0.6, p = 0.4, n = 50, 100 \), WMCH. (Bottom left) \( \alpha = 0.6, p = 0.2, n = 50, 100 \), RBP and WMCH. (Bottom right) \( \alpha = 0.8, p = 0.4, n = 50, 100 \), RBP and WMCH. 100–300 instances are used for each point. \( r_{\text{id}}, r_{\text{d}}, r_{\text{s}}, r_{\text{is}} \) are the phase transition points in \( n \rightarrow \infty \) limit.

Figure 10. The whitening procedure [29, 35–37] starts from a solution where each variable is assigned a value. If each variable has a set of values, we have a Cartesian product of the sets of values, which is a set of assignments. Starting from a solution (which can be seen as a Cartesian product), we expand the Cartesian product in such a way: if a neighbor of the Cartesian product is a solution then expand the Cartesian product minimally to include the neighbor. A neighbor means that the Hamming distance between an assignment in the Cartesian product and the neighbor is 1. When the expansion finishes, the Cartesian product contains the cluster to which the starting solution belongs. The variables frozen in the Cartesian product, which only take one value in the Cartesian product, are also frozen in the cluster.

From figure 10, we find that:

(a) The found solutions have no frozen variables until about \( r_{\text{d}} \).
Figure 9. Histograms of overlaps between the found solutions and the planted solutions divided by \( n \). (Left) \( r_0 = 0.85, n = 50, \alpha = 0.6, p = 0.2 \), over 793 data. (Right) \( r_0 = 0.95, n = 50, \alpha = 0.6, p = 0.2 \), over 592 data.

(b) The average percentage of frozen variables increases rapidly at about \( r_d \), then increases slowly until 1 is reached at about \( r_s \).

(c) Solutions found by WMCH have more frozen variables than that by RBP.

It indicates that once the solution space shatters into small clusters, many variables in the small clusters are frozen immediately. \( r_d \) seems to be the freezing transition point. The percentage of frozen variables reflects the size of a cluster. The planted cluster has about 70%–80% frozen variables (at \( r_s \)), which is consistent with the overlaps at \( r_s \) shown in figure 8. RBP ends up into bigger clusters compared with WMCH. As shown in figure 7, once variables become frozen, both algorithms tend to be invalid. The results indicate that \( r_d \) should be the clustering transition point, the freezing transition point, and the easy-hard transition point.

Fourthly, we run the BP iterations initialized in a found solution, and the results should be affected by the solution space structure. By ‘initialized in a found solution’ we mean that \( \nu^{(0)}_{i \to a} (\sigma_i) \) equals 1 if \( \sigma_i \) is the value of variable \( i \) in the found solution, and \( \nu^{(0)}_{i \to a} (\sigma_i) = 0 \) otherwise. If the BP iterations (10) and (11) converge, the marginal probabilities of variable \( i \) can be obtained from the converging messages \( \tilde{\nu}_{a \to i} (\sigma_i) \):

\[
\nu_i (\sigma_i) = \frac{1}{\tilde{Z}_i} \prod_{b \in \partial i} \tilde{\nu}_{b \to i} (\sigma_i),
\]

where \( \tilde{Z}_i \) is normalization constant. We define the normalised information entropy of the marginals:

\[
\beta_1 = -\frac{1}{n \ln d} \sum_{i=1}^{d} \sum_{\sigma_i=1}^{d} \nu_i (\sigma_i) \ln (\nu_i (\sigma_i)).
\]
Figure 10. Average percentages of frozen variables obtained by running the whitening procedure on the solutions found by RBP and WMCH. (Top left) $\alpha = 0.6, p = 0.4, n = 50, 100, 200$, RBP. (Top right) $\alpha = 0.6, p = 0.4, n = 50, 100$, WMCH. (Bottom left) $\alpha = 0.6, p = 0.2, n = 50, 100$, RBP and WMCH. (Bottom right) $\alpha = 0.8, p = 0.4, n = 50, 100$, RBP and WMCH. 100–300 instances are used for each point. $r_{ld}, r_{rd}, r_s, r_{ls}$ are the phase transition points in $n \to \infty$ limit.

$\beta_1$ can measure how uniform the marginals are. If $\nu_i(\sigma_i) = 1/d$ for all $i$ and $\sigma_i$, we have $\beta_1 = 1$. If marginals bias totally to a value $\sigma_i$ for all $i$, we have $\beta_1 = 0$. The marginals guided assignment is defined to be $\bar{\sigma} = \{\arg\max_{\sigma_i} \nu_i(\sigma_i)\}$. To measure how biased the marginals are to the initialized solution, we define the bias intensity of the marginals to the initialized solution:

$$\beta_2 = \frac{\text{the overlap between the initialized solution and the marginals guided assignment}}{n}$$

Bethe free entropy [1] is $S_{\text{Bethe}} = \sum_a S_a + \sum_i S_i - \sum_{(i,a)} S_{ia}$, where $S_a = \log \left[ \sum_{\hat{\nu}_a} \psi_\sigma(\sum_{\partial_a} \hat{\nu}_a \nu_i \to a(\sigma_i)) \right]$, $S_i = \log \left[ \sum_{\hat{\nu}_i} \prod_{b \in \partial_i} \hat{\nu}_b \nu_i (\sigma_i) \right]$, $S_{ia} = \log \left[ \sum_{\hat{\nu}_i} \nu_i \to a(\sigma_i) \hat{\nu}_a \nu_i (\sigma_i) \right]$.
Figure 11. Belief propagation initialized in solutions found by RBP and WMCH. $\beta_1$ is the normalised information entropy of the marginals. $\beta_2$ is the bias intensity of the marginals to the initialized solution. $\beta_3$ is the normalised Bethe free entropy. 100 instances are used for each point. $r_d, r_d, r_s, r_is$ are the phase transition points in $n \to \infty$ limit.

We define the normalised Bethe free entropy:

$$\beta_3 = \frac{S_{\text{Bethe}}}{n \ln d}$$

From figure 11, we find that:

(a) $\beta_1$ is almost 1 for small $r_0$, which means that the BP marginals are almost uniform. $\beta_1$ drops rapidly at $r_d$, then decreases slowly until 0 is reached at about $r_is$.

(b) Before $r_d$, $\beta_2$ corresponding to WMCH increases; $\beta_2$ corresponding to RBP decreases. Between $r_d$ and $r_s$, $\beta_2$ increases rapidly. After $r_s$, $\beta_2$ increases slowly until 1 is reached at about $r_is$.

(c) $\beta_3$ drops linearly before $r_d$, then decreases slowly until 0 is reached at about $r_is$. 

https://doi.org/10.1088/1742-5468/ac4e86
The solution space structure of planted constraint satisfaction problems with growing domains

Figure 12. Average percentages of frozen variables obtained by running the whitening procedure starting from the planted solution, over 100 instances. (Left) $\alpha = 0.6, p = 0.4, n = 50, 100, 200, 300$. (Middle) $\alpha = 0.6, p = 0.2, n = 50, 100, 200$. (Right) $\alpha = 0.8, p = 0.4, n = 50, 100, 200$. $r_{\text{id}}, r_{\text{d}}, r_{\text{s}}, r_{\text{is}}$ are the phase transition points in the $n \to \infty$ limit.

Table 3. The average percentages of frozen variables at $r_s$ obtained by running the whitening procedure starting from the planted solution, over 100 instances.

|               | $\alpha = 0.6, p = 0.4$ | $\alpha = 0.6, p = 0.2$ | $\alpha = 0.8, p = 0.4$ |
|---------------|-------------------------|-------------------------|-------------------------|
| $n = 50$      | 0.7660                  | 0.7784                  | 0.8528                  |
| $n = 100$     | 0.8122                  | 0.8536                  | 0.8890                  |
| $n = 200$     | 0.8531                  | 0.8998                  | 0.9183                  |

It indicates that the BP initialized in a solution correctly recovers the marginals of the cluster to which the initialized solution belongs. Before $r_{\text{id}}$, it is initialized in the giant cluster, so the marginals are almost uniform. Before $r_{\text{d}}$, the location of the solution found by WMCH is weakly related to the marginals. But the location of the solution found by RBP is strongly connected to the marginals for small $r_0$. $\beta_2$ corresponding to RBP decreases before $r_{\text{d}}$ means that the external field of RBP plays a more and more important role with $r_0$ increasing. After $r_{\text{d}}$, BP is initialized in a small cluster, so the marginals are no longer uniform, but bias to the initialized solution. Between $r_{\text{d}}$ and $r_{\text{s}}$, the $\beta_2$ corresponding to WMCH tends to be bigger than the $\beta_2$ corresponding to RBP, because WMCH ends up in smaller clusters than RBP. After $r_{\text{is}}$, BP is initialized in the isolated $\vec{\sigma}_P$, so the BP marginals recover $\vec{\sigma}_P$ totally. Bethe free entropy reflects the size of the initialized cluster, which dramatically changes at $r_{\text{d}}$. The performances of BP initialized in a found solution are consistent with the phase diagram revealed in figures 7, 8, and 10.

Fifthly, we run the whitening procedure starting from $\vec{\sigma}_P$, and run the BP iterations initialized in $\vec{\sigma}_P$, in order to study the planted cluster. The average percentages of frozen variables obtained by running the whitening procedure starting from $\vec{\sigma}_P$ are shown in figure 12 and table 3. The percentages of converging of BP initialized in $\vec{\sigma}_P$, and corresponding $\beta_1$, $\beta_2$, $\beta_3$ are shown in figure 13.

https://doi.org/10.1088/1742-5468/ac4e86
The solution space structure of planted constraint satisfaction problems with growing domains

Figure 13. Belief propagation initialized in the planted solution, over 100 instances. \(\beta_1\) is the normalised information entropy of the marginals. \(\beta_2\) is the bias intensity of the marginals to the planted solution. \(\beta_3\) is the normalised Bethe free entropy. \(r_{id}, r_d, r_s, r_{is}\) are the phase transition points in the \(n \to \infty\) limit.

From figure 12, table 3 and figure 13, we find that:

(a) The average percentages of frozen variables increase rapidly at about \(r_d\). As \(n\) grows the percentages increase, when \(r > r_d\). In the left and right pictures of figure 12, many variables are already frozen at \(r_0 = 0.85 < r_d\).

(b) We can find that from table 3: the average percentage of frozen variables at \(r_s\) increases with \(n\), increases with \(\alpha\), and decreases with \(p\).

(c) \(\beta_1, \beta_2\) and \(\beta_3\) change dramatically at about \(r_d\).

This indicates that the planted cluster of small-size instance is different from that in the \(n \to \infty\) limit. Firstly, instead of \(r_{id}\), the planted cluster is separated from other solutions at about \(r_d\). Secondly, the size of the planted cluster is bigger than that in the \(n \to \infty\) limit (where the cluster diameter is smaller than \(en\) for any positive constant \(e\)). If \(n\) keeps growing, the planted cluster should be more like that in the \(n \to \infty\) limit. There are two clues: (1) after \(r_d\), as \(n\) grows, the percentage of frozen variables increases; (2) in the left and right pictures of figure 12, variables become frozen a little earlier than \(r_d\).

To sum up, experiments show that the small-size instances have similar transitions at \(r_d, r_s, r_{is}\) as in the \(n \to \infty\) limit. \(r_d\) seems not only to be the clustering transition point, but also the freezing transition point and the easy-hard transition point. The independent transition at \(r_{id} = -\ln(1-p)p/k\) does not occur on small-size instances. After the solution space shatters into small clusters at \(r_d\), the planted cluster is bigger than that in the \(n \to \infty\) limit, and the small clusters are biased to (tend to be close to) \(\vec{\sigma}_P\). The bigger the planted cluster is, the easier it is to be found. From table 3 and experiments in figure 7, we find that when the percentages of frozen variables (at \(r_s\)) reach about 0.9 the problems will be extremely hard, so we suggest a set of parameters for generating extremely hard instances: \(k = 2, n = 100, \alpha = 0.8, p = 0.2\).
9. Conclusion

We find four phase transitions of the planted RB model in the thermodynamic limit. When the independent phase transition (at $r = r_{id}$) occurs, the planted cluster is separated from the majority of solutions. From then on, the majority and the planted cluster have different developments: the majority of solutions will go through the clustering transition (at $r = r_d$) and the satisfiability transition by configurations outside the planted cluster (at $r = r_s$); the planted cluster will go through the isolated transition (at $r = r_{is}$). $r_{is} < r_d$, $r_d < r_{is} < r_s$ and $r_s < r_{is}$ are all possible, depending on parameters $\alpha, p, k$.

Through experiments, we studied small-size instances that have about $10^2$ variables. They show that the last three phase transitions (except the independent phase transition) show up as in the thermodynamic limit. When $r$ is smaller than but close to $r_s$, there are many very small clusters distributed widely in the solution space, and a large part of the variables are frozen in those clusters. After $r_s$, only the planted cluster remains in the solution space. This guarantees that this model can generate very hard satisfiable CSP benchmarks. If $r_{is}$ is smaller than $r_s$, i.e. $(1 + \alpha)/(kp) < -(\alpha)/\ln(1 - p)$, the problem should be even harder (at $r = r_s$), because the planted cluster has shrunk into a point at $r = r_s$.

In the $r \to \infty$, $n \to \infty$ limit, we find that there is a single smooth valley in the energy landscape of the planted RB model. The sizes of instances on which we do the experiments are limited by the algorithms to find solutions. In the future, we will try to design more efficient algorithms, such as a message passing algorithm with backtracking [37] and a quasi-greedy algorithm [38].

Acknowledgments

This project was supported by the National Natural Science Foundation of China (Grant No. 11801028).

Appendix A

In appendix A, we will prove (5). First of all, from the definition of the planted RB model we will give the expression of $\mathbb{E}(Z(x))$ in the following formula (14). Secondly, the normalized version $f(x)$ will be given in (15). Thirdly the operation of finding the maximum value in (15) will be carried out and then (5) will be proven.

Let $\vec{\sigma}_A$ and $\vec{\sigma}_B$ be two assignments at a Hamming distance $xn$. The number of such assignment-pairs is

$$\binom{n}{xn, n-xn} d^n d^{-1}^{xn}.$$

Because of the planted solution, not all of those assignment-pairs have the same opportunities to be solution-pairs. We have to consider the distance between $\vec{\sigma}_A$ and $\vec{\sigma}_P$, and
The solution space structure of planted constraint satisfaction problems with growing domains

the distance between $\vec{\sigma}_B$ and $\vec{\sigma}_P$, where $\vec{\sigma}_P$ is the solution we planted. We denote by $\sigma_i(A), \sigma_i(B), \sigma_i(S)$ the values of variable $i$ in assignments $\vec{\sigma}_A, \vec{\sigma}_B, \vec{\sigma}_P$. The $n$ variables $1, 2, \ldots, n$ can be divided into five sets:

(a) $\{i|\sigma_i(A) = \sigma_i(S), \sigma_i(B) \neq \sigma_i(S), \sigma_i(A) \neq \sigma_i(B), i = 1, \ldots, n\}$;
(b) $\{i|\sigma_i(A) \neq \sigma_i(S), \sigma_i(B) = \sigma_i(S), \sigma_i(A) \neq \sigma_i(B), i = 1, \ldots, n\}$;
(c) $\{i|\sigma_i(A) \neq \sigma_i(S), \sigma_i(B) \neq \sigma_i(S), \sigma_i(A) \neq \sigma_i(B), i = 1, \ldots, n\}$;
(d) $\{i|\sigma_i(A) \neq \sigma_i(S), \sigma_i(B) \neq \sigma_i(S), \sigma_i(A) = \sigma_i(B), i = 1, \ldots, n\}$;
(e) $\{i|\sigma_i(A) = \sigma_i(S), \sigma_i(B) = \sigma_i(S), \sigma_i(A) = \sigma_i(B), i = 1, \ldots, n\}$.

The sizes of the sets are $an, bn, cn, dn, en$ respectively, then $a + b + c = x, d + e = 1 - x$. In such a way, we classify assignment-pairs into types labeled by a tuple $(a, b, c, d, e)$. The number of assignment-pairs of type $(a, b, c, d, e), a + b + c = x, d + e = 1 - x$ is

$$F_1 = \binom{n}{xn, n-xn} \binom{xn}{an, bn, cn} (n-xn)(dn, en)(d-1) \binom{n}{x-a-b}.$$ 

We will show that assignment-pairs of the same type have the same opportunity of being solution-pairs. According to the steps of generating a planted RB instances, $t = rn \ln n$ constraints are chosen randomly. Constraint $a$ involves $k$ variables that are chosen randomly, denoted by $\vec{\sigma}_a$. Let the values of those $k$ variables in assignments $\vec{\sigma}_A, \vec{\sigma}_B, \vec{\sigma}_P$ be $\vec{\sigma}_a(A), \vec{\sigma}_a(B), \vec{\sigma}_a(S)$, respectively. The relations among $\vec{\sigma}_a(A), \vec{\sigma}_a(B)$ and $\vec{\sigma}_a(S)$ are in three sorts:

(a) $\vec{\sigma}_a(A) = \vec{\sigma}_a(B) = \vec{\sigma}_a(S)$;
(b) $\vec{\sigma}_a(A) = \vec{\sigma}_a(B) \neq \vec{\sigma}_a(S)$ or $\vec{\sigma}_a(A) \neq \vec{\sigma}_a(B) = \vec{\sigma}_a(S)$ or $\vec{\sigma}_a(B) \neq \vec{\sigma}_a(A) = \vec{\sigma}_a(S)$;
(c) $\vec{\sigma}_a(A) \neq \vec{\sigma}_a(B)$ and $\vec{\sigma}_a(A) \neq \vec{\sigma}_a(S)$ and $\vec{\sigma}_a(B) \neq \vec{\sigma}_a(S)$.

For constraint $a$, $(1 - p)^d$ compatible tuples of values are selected, including $\vec{\sigma}_a(S)$. For the first sort, because $\vec{\sigma}_a(S)$ satisfies the constraint $a$, so $\vec{\sigma}_a(A)$ and $\vec{\sigma}_a(B)$ both satisfy $a$. The probability of the occurrence of the first sort is the probability that the randomly chosen $k$ variables all belong to the above fifth set (whose size is $en$), and the probability is the following $s_0$. For the second and the third sorts, the probabilities of occurrences are the following $s_1, s_2$; the probabilities that both $\vec{\sigma}_a(A)$ and $\vec{\sigma}_a(B)$ satisfy the constraint $a$ are the following $p_1, p_2$. The opportunity that an assignment-pair of type $(a, b, c, d, e)$ is a solution-pair is

$$F_2 = (s_0 + s_1p_1 + s_2p_2)^f,$$

where

$$s_0 = \binom{en}{k, en-k} / \binom{n}{k, n-k},$$

$$s_1 = s_3 + s_4 + s_5 - 3s_0,$$

$$s_2 = 1 - s_3 - s_4 - s_5 + 2s_0,$$

https://doi.org/10.1088/1742-5468/ac4e86
The solution space structure of planted constraint satisfaction problems with growing domains

\[ p_1 = \frac{d^k - 2}{(1-p)d^k - 2, pd^k} / \frac{d^k - 1}{(1-p)d^k - 1, pd^k}, \]

\[ p_2 = \frac{d^k - 3}{(1-p)d^k - 3, pd^k} / \frac{d^k - 1}{(1-p)d^k - 1, pd^k}, \]

and

\[ s_3 = \frac{(a+e)n}{k, (a+e)n - k} / \frac{n}{k, n - k}, \]

\[ s_4 = \frac{(b+e)n}{k, (b+e)n - k} / \frac{n}{k, n - k}, \]

\[ s_5 = \frac{n - xn}{k, n - xn - k} / \frac{n}{k, n - k}. \]

The expectation of the number of solutions pairs at distance \( xn \) is

\[ \mathbb{E}(Z(x)) = \sum_{an + bn + cn = xn, dn + en = n - xn} F_1 \cdot F_2. \tag{14} \]

This summation operation contains a polynomial number of items, so \( \lim_{n \to \infty} \frac{\ln(\mathbb{E}(Z(x)))}{\ln n} \) is determined by the largest term. With \( n \to \infty \), we have \( p_1 \to 1 - p, p_2 \to (1 - p)^2, s_0 \to e^k, s_3 \to (a + e)^k, s_4 \to (b + e)^k, s_5 \to (1 - x)^k \), and by the Stirling formula, we have

\[ \frac{\ln \left( \frac{n}{xn, n - xn} \right)}{n} \to - \ln(x^r(1 - x)^{1-x}). \]

The formula \( f(x) \) defined in (4) can be simplified:

\[ f(x) = \lim_{n \to \infty} \frac{\ln(\mathbb{E}(Z(x)))}{\ln n} \]

\[ = \lim_{n \to \infty} \frac{\ln \left( \max_{an + bn + cn = xn, dn + en = n - xn} F_1 \cdot F_2 \right)}{\ln n} \]

\[ = \max_{0 \leq a,b \leq x, 0 \leq c \leq 1-x, 0 \leq e \leq 1-x} \left[ a(1 + x - e - a - b) + r \ln A(x, a, b, e) \right], \tag{15} \]

where

\[ A(x, a, b, e) = e^k + [(1 - x)^k + (a + e)^k + (b + e)^k - 3e^k] (1 - p) \]

\[ + [1 - (1 - x)^k - (a + e)^k - (b + e)^k + 2e^k] (1 - p)^2 \]

\[ = (1 - p)^2 + [1 - 3(1 - p) + 2(1 - p)^2] e^k + [(1 - p) - (1 - p)^2] \]

\[ \times [(1 - x)^k + (a + e)^k + (b + e)^k]. \]
The solution space structure of planted constraint satisfaction problems with growing domains

Denoting \( m = a + b, (a + e)^k + (b + e)^k \) reaches its maximum at \( a = 0, b = m \) or \( a = m, b = 0 \), so

\[
\max_{0 \leq a, b \leq x, \ 0 \leq a + b \leq x, \ 0 \leq e \leq 1 - x} A(x, a, b, e) = \max_{0 \leq m \leq x, \ 0 \leq e \leq 1 - x} (1 - p)^2 + p^2 e^k + [(1 - p) - (1 - p)^2] ((1 - x)^k + (m + e)^k)
\]  

(16)

Let \( m + e = l \). If \( l < 1 - x \), the above formula reaches maximum at \( e = l \); if \( l > 1 - x \), the above formula reaches maximum at \( e = 1 - x \); then

\[
(16) = \max\{A_1, A_2\},
\]

where

\[
A_1 = \max_{0 \leq l \leq 1 - x} r \ln[(1 - p)^2 + (1 - p)p(1 - x)^k + pl^k],
\]

\[
A_2 = \max_{1 - x < l \leq 1} r \ln[(1 - p)^2 + p(1 - x)^k + (1 - p)pl^k].
\]

Combining (15) and (16), we have

\[
f(x) = \max\{\max_{0 \leq l \leq 1 - x} f_3, \max_{1 - x < l \leq 1} f_4\},
\]

where

\[
f_3 = \alpha(1 + x - l) + r \ln[(1 - p)^2 + (1 - p)p(1 - x)^k + pl^k],
\]

\[
f_4 = \alpha(1 + x - l) + r \ln[(1 - p)^2 + p(1 - x)^k + (1 - p)pl^k].
\]

Calculate the second order differentials of \( f_3(x, l) \) and \( f_4(x, l) \),

\[
\frac{\partial^2 f_3}{\partial l^2} = \frac{rplk^{k-2} [(k - 1) ((1 - p)^2 + (1 - p)p(1 - x)^k + pl^k) - pl^k]}{[(1 - p)^2 + (1 - p)p(1 - x)^k + pl^k]^2},
\]

\[
\frac{\partial^2 f_4}{\partial l^2} = \frac{r(1 - p)plk^{k-2} [(k - 1) ((1 - p)^2 + p(1 - x)^k + (1 - p)pl^k) - (1 - p)pl^k]}{[(1 - p)^2 + p(1 - x)^k + (1 - p)pl^k]^2},
\]

then we find \( \frac{\partial^2 f_3}{\partial l^2} \geq 0 (0 \leq l \leq 1 - x), \frac{\partial^2 f_4}{\partial l^2} \geq 0 (1 - x < l \leq 1) \) under the condition \( p < 1 - \frac{1}{k} \).

The maximum values \( \max_{0 \leq l \leq 1 - x} f_3 \) and \( \max_{1 - x < l \leq 1} f_4 \) can only be achieved on the endpoints, i.e.

\[
f(x) = \max\{f_3(l = 0), f_3(l = 1 - x), f_4(l = 1 - x), f_4(l = 1)\},
\]

and further calculating shows

\[
f(x) = \max\{f_1(x), f_2(x)\},
\]

where \( f_1(x) \) and \( f_2(x) \) are defined in (6) and (7).

https://doi.org/10.1088/1742-5468/ac4e86
The solution space structure of planted constraint satisfaction problems with growing domains

References

[1] M´ezard M and Montanari A 2009 Information, Physics, and Computation (Oxford: Oxford University Press)
[2] Zdeborová L and Krzakala F 2016 Statistical physics of inference: thresholds and algorithms Adv. Phys. 65 453–552
[3] Rossi F, Beek P V and Walsh T 2006 Handbook of Constraint Programming (New York: Elsevier)
[4] Gent I P, Macintyre E, Prosser P, Smith B M and Walsh T 2001 Random constraint satisfaction: flaws and structure Constraints 6 345–72
[5] Smith B M and Dyer M E 1996 Locating the phase transition in binary constraint satisfaction problems Artif. Intell. 81 155–81
[6] Achlioptas D, Kirousis L M, Kranakis E, Krizanc D, Molloy M S O and Stamatiou Y C 1997 Random constraint satisfaction: a more accurate picture Int. Conf. on Principles and Practice of Constraint Programming (Berlin: Springer) pp 107–20
[7] Lecoupe C 2009 Constraint Networks: Techniques and Algorithms (New York: Wiley)
[8] Xu K and Li W 2000 Exact phase transitions in random constraint satisfaction problems J. Artif. Intell. Res. 12 93–103
[9] Barthel W, Hartmann A K, Leone M, Ricci-Tersenghi F, Weigt M and Zecchina R 2002 Hiding solutions in random satisfiability problems: a statistical mechanics approach Phys. Rev. Lett. 88 188701
[10] Jia H, Moore C and Selman B 2004 From spin glasses to hard satisfiable formulas Int. Conf. on Theory and Applications of Satisfiability Testing (Berlin: Springer) pp 199–210
[11] Zdeborová L and Krzakala F 2011 Quiet planting in the locked constraint satisfaction problems SIAM J. Discrete Math. 25 750–70
[12] Xu K, Boussemart F, Hemery F and Lecoupe C 2007 Random constraint satisfaction: easy generation of hard (satisfiable) instances Artif. Intell. 171 514–34
[13] Xu K 2003 Forced satisfiable CSP and SAT benchmarks of Model RB (http://www.nlsde.buaa.edu.cn/~kexu/benchmarks/benchmarks.htm)
[14] Cai S, Su K and Sattar A 2011 Local search with edge weighting and configuration checking heuristics for minimum vertex cover Artif. Intell. 175 1672–96
[15] Rosin C D 2014 Unweighted stochastic local search can be effective for random CSP benchmarks (arXiv:1411.7480)
[16] Montanari A, Ricci-Tersenghi F and Semerjian G 2008 Clusters of solutions and replica symmetry breaking in random k-satisfiability J. Stat. Mech. P04004
[17] M´ezard M, Ricci-Tersenghi F and Zecchina R 2003 Two solutions to diluted p-spin models and XORSAT problems J. Stat. Phys. 111 505
[18] Krzakala F, Montanari A, Ricci-Tersenghi F, Semerjian G and Zdeborová L 2007 Gibbs states and the set of solutions of random constraint satisfaction problems Proc. Natl Acad. Sci. 104 25
[19] Krzakala F and Zdeborová L 2009 Hiding quiet solutions in random constraint satisfaction problems Phys. Rev. Lett. 102 238701
[20] Ricci-Tersenghi F and Semerjian G 2009 On the cavity method for decimated random constraint satisfaction problems and the analysis of belief propagation guided decimation algorithms J. Stat. Mech. 355–71
[21] M´ezard M, Parisi G and Zecchina R 2002 Analytic and algorithmic solution of random satisfiability problems Science 297 812–5
[22] M´ezard M and Zecchina R 2002 The random K-satisfiability problem: from an analytic solution to an efficient algorithm Phys. Rev. E 66 056126
[23] Braunstein A, M´ezard M and Zecchina R 2005 Survey propagation: an algorithm for satisfiability Random Struct. Algorithms 27 201–26
[24] Achlioptas D, Coja-Oghlan A and Ricci-Tersenghi F 2011 On the solution-space geometry of random constraint satisfaction problems Random Struct. Algorithms 38 251
[25] Xu W, Zhang P, Liu T and Gong F 2015 The solution space structure of random constraint satisfaction problems with growing domains J. Stat. Mech. P12006
[26] Achlioptas D and Ricci-Tersenghi F 2006 On the solution-space geometry of random constraint satisfaction problems Proc. STOC ‘06 pp 130–9
[27] Achlioptas D 2008 Solution clustering in random satisfiability Eur. Phys. J. B 64 395–402
[28] M´ezard M, Mora T and Zecchina R 2005 Clustering of solutions in the random satisfiability problem Phys. Rev. Lett. 94 197205

https://doi.org/10.1088/1742-5468/ac4e86
The solution space structure of planted constraint satisfaction problems with growing domains

[29] Xu W, Gong F and Zhou G 2020 Clustering phase of a general constraint satisfaction problem model $d$-$k$-CSP *Physica A* **537** 122708

[30] Coja-Oghlan A, Mossel E and Vilenchik D 2009 A spectral approach to analyzing belief propagation for 3-coloring *Comb. Probab. Comput.* **18** 881–912

[31] Coja-Oghlan A, Krivelevich M and Vilenchik D 2010 Why almost all $k$-colorable graphs are easy to color *Theory Comput. Syst.* **46** 523–65

[32] Chavas J, Furtlehner C, Mézard M and Zecchina R 2005 Survey-propagation decimation through distributed local computations *J. Stat. Mech.* P11016

[33] Braunstein A and Zecchina R 2005 Learning by message-passing in networks of discrete synapses *Phys. Rev. Lett.* **96** 030201

[34] Zhao C, Zhang P, Zheng Z and Xu K 2012 Analytical and belief-propagation studies of random constraint satisfaction problems with growing domains *Phys. Rev. E* **85** 016106

[35] Parisi G 2008 On the survey-propagation equations in random constraint satisfiability problems *J. Math. Phys.* **49** 812–695

[36] Braunstein A and Zecchina R 2004 Survey propagation as local equilibrium equations *J. Stat. Mech.* P06007

[37] Marino R, Parisi G and Ricci-Tersenghi F 2016 The backtracking survey propagation algorithm for solving random K-SAT problems *Nat. Commun.* **7** 12996

[38] Bernaschi M, Bisson M, Fatica M, Marinari E, Martin-Mayor V, Parisi G and Ricci-Tersenghi F 2021 How we are leading a 3-XORSAT challenge: from the energy landscape to the algorithm and its efficient implementation on GPUs *Europhys. Lett.* **133** 60005

https://doi.org/10.1088/1742-5468/ac4e86