Satisfiability Thresholds for Regular Occupation Problems

Konstantinos Panagiotou
LMU München, Germany
kpanagio@math.lmu.de

Matija Pasch
LMU München, Germany
pasch@math.lmu.de

Abstract

In the last two decades the study of random instances of constraint satisfaction problems (CSPs) has flourished across several disciplines, including computer science, mathematics and physics. The diversity of the developed methods, on the rigorous and non-rigorous side, has led to major advances regarding both the theoretical as well as the applied viewpoints. The two most popular types of such CSPs are the Erdős-Rényi and the random regular CSPs.

Based on a ceteris paribus approach in terms of the density evolution equations known from statistical physics, we focus on a specific prominent class of problems of the latter type, the so-called *occupation problems*. The regular r-in-k occupation problems resemble a basis of this class. By now, out of these CSPs only the satisfiability threshold – the largest degree for which the problem admits asymptotically a solution – for the 1-in-k occupation problem has been rigorously established. In the present work we take a general approach towards a systematic analysis of occupation problems. In particular, we discover a surprising and explicit connection between the 2-in-k occupation problem satisfiability threshold and the determination of *contraction coefficients*, an important quantity in information theory measuring the loss of information that occurs when communicating through a noisy channel. We present methods to facilitate the computation of these coefficients and use them to establish explicitly the threshold for the 2-in-k occupation problem for $k = 4$. Based on this result, for general $k \geq 5$ we formulate a conjecture that pins down the exact value of the corresponding coefficient, which, if true, is shown to determine the threshold in all these cases.

2012 ACM Subject Classification Theory of computation → Randomness, geometry and discrete structures; Mathematics of computing → Discrete mathematics; Mathematics of computing → Graph theory; Mathematics of computing → Probability and statistics; Mathematics of computing → Probabilistic inference problems; Mathematics of computing → Probabilistic reasoning algorithms; Mathematics of computing → Information theory

Keywords and phrases Constraint satisfaction problem, replica symmetric, contraction coefficient, first moment, second moment, small subgraph conditioning

Digital Object Identifier 10.4230/LIPIcs.ICALP.2019.90

Category Track A: Algorithms, Complexity and Games

Related Version A full version of the paper is available at `https://arxiv.org/abs/1811.00991`.

Funding *Konstantinos Panagiotou*: The research leading to these results has received funding from the European Research Council, ERC Grant Agreement 772606–PTRCSP.

1 Introduction

Inspired by the pioneering work [22] of Erdős and Rényi in 1960, random discrete structures have been systematically studied in literally thousands of contributions. The initial motivation of this research was to study open problems in graph theory and combinatorics. In the following decades, however, the application of such models proved useful as a unified approach

46th International Colloquium on Automata, Languages, and Programming (ICALP 2019).
Editors: Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi;
Article No. 90; pp. 90:1–90:14
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

to treat a variety of problems in several fields. To mention just a few, random graphs turned out to be valuable in solving fundametal theoretical and practical problems, such as the development of error correcting codes [34], the study of statistical inference through the stochastic block model [1], and the establishment of lower bounds in complexity theory [27, 24].

The results of the past years of research suggest the existence of *phase transitions* in many classes of random discrete structures, i.e. a specific value of a given model parameter at which the properties of the system in question change dramatically. Constraint satisfaction problems are one specific type of such structures that tend to exhibit this remarkable property and that are of particular interest in too many areas to mention, covering complexity theory, combinatorics, statistical mechanics, artificial intelligence, biology, engineering and economics. An instance of a CSP is defined by a set of variables that take values in – typically finite – domains and a set of constraints, where each constraint is satisfied for specific assignments of the subset of variables it involves. A major computational challenge is to determine whether such an instance is satisfiable, i.e. to determine if there is an assignment of all variables that satisfies all constraints.

Since the 1980s non-rigorous methods have been introduced in statistical physics that are targeted at the analysis of phase transitions in *random* CSPs [37, 36, 33]. Within this line of research, a variety of exciting and unexpected phenomena were discovered, as for example the existence of multiple phase transitions with respect to the structure of the solution space in random CSPs; these transitions may have a significant impact on the hardness of the underlying instances. Since then these methods and the description of the conjectured regimes have been heavily supported by several findings, including the astounding empirical success of randomized algorithms like *belief* and *survey propagation* [9], as well as rigorous verifications, most prominently the phase transition in k-SAT [19] (for sufficiently large k) and the condensation phase transition in many important models [14]. However, a complete rigorous study is still a big challenge for computer science and mathematics.

Usually, the relevant model parameter of a random CSP is a certain problem specific *density* as illustrated below. The main focus of research is to study the occurrence of phase transitions in the solution space structure and in particular the existence of *(sharp) satisfiability thresholds*, i.e. critical values of the density such that the probability that a random CSP admits a solution tends to one as the number of variables tends to infinity for densities below the threshold, while this limiting probability tends to zero for densities above the threshold.

Random CSPs. The two most popular types of random CSPs are Erdős Rényi (ER) type CSPs and random regular CSPs. In both cases the number $n = |V|$ of variables and the number k of variables involved in each constraint is fixed. In ER type CSPs we further fix the number $m = |F|$ of constraints and thereby the *density* $\alpha = m/n$, i.e. the average number of constraints that a variable is involved in. In random regular CSPs we only consider instances where each variable is involved in the same number d of constraints, which fixes the *density* d as well as the number $m = dn/k$ of constraints. In a second step we randomly choose the sets of satisfying assignments for each constraint depending on the problem. For example, in the prominent k-SAT one forbidden assignment is chosen uniformly at random from all possible assignments of the involved binary variables for each constraint independently. Another famous example is the coloring of hypergraphs, where the constraints are attached to the hyperedges and the variables to the vertices of the hypergraph, i.e. the variables involved in a constraint correspond to the vertices incident to a hyperedge. In this case the satisfying assignments are determined since each constraint is violated iff all involved vertices take the same color.

In our work we focus on the class of random regular CSPs where the choice of satisfying assignments per constraint is determined, i.e. a class that covers the regular occupation problems and the coloring of (d-regular k-uniform) hypergraphs amongst others, sparing problems with random constraints like k-SAT and XORSAT. A unique feature of this class is, intuitively speaking, that the local structure of almost all instances is fixed almost everywhere for sufficiently large n. The lack of randomness makes this class particularly accessible for an analysis of the asymptotic solution space structure and significantly simplifies simulations based on the well-known *population dynamics*. Using such simulations, non-rigorous results for this class have been mostly established for the case where the variables are binary valued, so called occupation problems, or restricted to variants of hypergraph coloring for non-binary variables. Besides the extensive studies on the coloring of simple graphs, i.e. $k = 2$, the only rigorous results derived so far consider the arguably most simple type of occupation problems where each constraint is satisfied if exactly one involved variable evaluates to true, which we refer to as d-regular 1-in-k occupation problem. In our current work we strive to extend these results to general d-regular r-in-k occupation problems, i.e. problems where each constraint is satisfied if r out of the k involved variables evaluate to true.

1.1 Occupation Problems

We continue with the formal definition of the class of problems we consider. Let k, $d \in \mathbb{Z}_{>1}$ and $r \in [k-1] := \{1, \dots, k-1\}$ be fixed. Additionally, we are given non-empty sets V of variables and constraints F. We will use the convention to index elements of V with the letter i and elements of F with the letter a (and subsequent letters) in the remainder. Then an instance o of the d-regular r-in-k occupation problem is specified by a sequence $o = (v(a))_{a \in F}$ of $m = |F|$ subsets $v(a) \subseteq V$ of size k such that each of the $n = |V|$ variables is contained in d of the subsets. In graph theory the instance o has a natural interpretation as a (d, k)-biregular graph (or d-regular k-factor graph) with node sets $V \dot\cup F$ and edges $\{i, a\} \in E$ if $i \in v(a)$.

Given an instance o as just described, we say that an assignment $x \in \{0, 1\}^V$ *satisfies* a constraint $a \in F$ if $\sum_{i \in v(a)} x_i = r$, otherwise x *violates* a. If x satisfies all constraints $a \in F$, then x is a *solution* of o. We write $z(o)$ for the number of solutions of o. An example of a 4-regular 2-in-3 occupation problem is shown in Figure 1a.

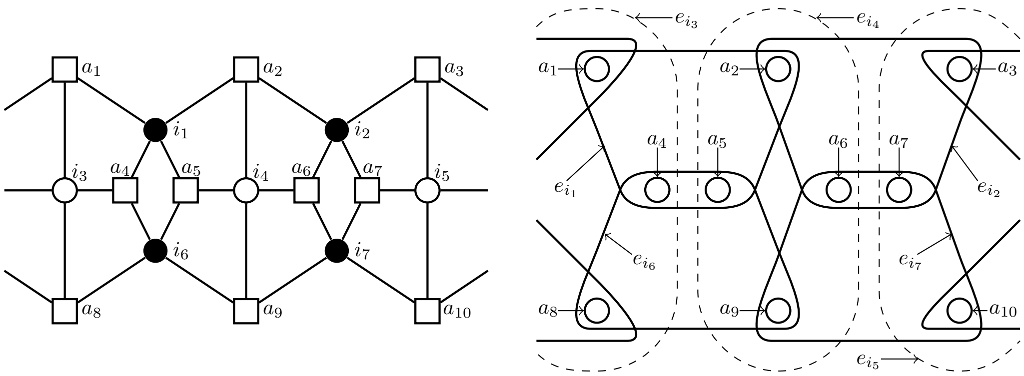
Further, for given m, $n \in \mathbb{Z}_{>0}$ let $\mathcal{O} = \mathcal{O}(k, d, n, m)$ denote the set of all instances o with variables $V = [n]$ and constraints $F = [m]$. If $\mathcal{O}$ is not empty, then the random d-regular r-in-k occupation problem O is the random variable O equipped with the uniform distribution $P = P_{\mathcal{O}}$ on $\mathcal{O}$ and $Z = z(O)$ the number of solutions of O.

1.2 Examples and Related Problems

A problem that is closely related and can be reduced to the d-regular r-in-k occupation problem is the d-regular positive r-in-k SAT problem, a variant of k-SAT introduced above. In this case, we consider a boolean formula

$$f = \bigwedge_{a \in F} c_a, \quad c_a = \bigvee_{i \in v(a)} i, \, a \in F,$$

in conjunctive normal form with m clauses over n variables $i \in V$, such that no literal appears negated (hence *positive* r-in-k SAT), and where each clause c_a is the disjunction of k literals and each variable appears in exactly d clauses (hence d-regular). The decision problem is to determine if there exists an assignment x such that exactly r literals in each



(a) Solution of the 2-occupation problem.

(b) A 2-factor in a hypergraph.

Figure 1 On the left we see a solution of the 4-regular 2-in-3 occupation problem on a 4-regular 3-factor graph, where the rectangles and circles depict the constraints (factors) and variables (filled if they take the value one in the solution). The figure on the right shows a 2-factor in a 3-regular 4-uniform hypergraph, where the circles, solid and dashed shapes represent the vertices, hyperedges in the 2-factor and the other hyperedges respectively.

clause evaluate to true (hence r-in-k SAT). In [39] the satisfiability threshold for this problem was determined for $r = 1$, i.e. the case where exactly one literal in each clause evaluates to true. One of our main results, Theorem 1.1, solves this problem when $r = 2$ and $k = 4$.

Our second example deals with a prominent problem related to graph theory. A k-regular d-uniform hypergraph h is a pair $h = (F, E)$ with $m = |F|$ vertices and $n = |E|$ (hyper-)edges such that each edge contains d vertices and the degree of each vertex is k. An *r-factor E'* is a subset of the hyperedges such that each vertex $a \in F$ is incident to r hyperedges $e_i \in E'$. In this case the problem is to determine if h has an r-factor. For example, the case $r = 1$ is the well-known perfect matching problem and the threshold was determined in [16]. An example of a 2-factor in a hypergraph is shown in Figure 1b. Theorem 1.1 solves also this problem for $r = 2$ and $k = 4$.

There are several other problems in complexity and graph theory that are closely related to the examples above. The satisfiability threshold in Theorem 1.1 also applies to a variant of the vertex cover problem (or hitting set problem from set theory perspective), where we choose a subset of the vertices (variables with value one) in a d-regular 4-uniform hypergraph such that each hyperedge is incident to exactly two vertices in the subset. Analogously, Theorem 1.1 also establishes the threshold for a variant of the set cover problem in set theory corresponding to 2-factors in hypergraphs, i.e. given a family of d-subsets (hyperedges) and a universe (vertices) with each element contained in four subsets, the problem is to find a subfamily of the subsets such that each element of the universe is contained in exactly two subsets of the subfamily. Further, Theorem 1.1 can e.g. also be used to give sufficient conditions for the (asymptotic) existence of Euler families in regular uniform hypergraphs as discussed in [6].

1.3 Main Results

The d-regular 1-in-k occupation problem has been completely solved in [39, 16], which also covers the d-regular 2-in-3 occupation problem due to color symmetry. Our first result addresses the next non-trivial case, namely the location of the satisfiability threshold of the random d-regular 2-in-4 occupation problem. For $k \in \mathbb{Z}_{>3}$ let

$$w_1^* = w_1^*(k) = \frac{2}{k},\ w_2^* = w_2^*(k) = \binom{k}{2}^{-1} \text{ and } d^* = d^*(k) = \frac{kH(w_1^*)}{kH(w_1^*) + \ln(w_2^*)}, \tag{1}$$

where $H(p) = -p\ln(p) - (1-p)\ln(1-p)$ is the binary entropy of $p \in [0,1]$. The following theorem establishes the location of the threshold at $d^*(4) \approx 2.83$ for $k = 4$.

▶ **Theorem 1.1** (2-in-4 Occupation Satisfiability Threshold). *Let $k = 4$, $d \in \mathbb{Z}_{>1}$, and $\mathcal{O} = \mathcal{O}(k,d,n,m)$, $Z = Z_{k,d,n,m}$ be as in Section 1.1.*

(a) *The set $\mathcal{O}$ is non-empty iff $m = m(n) = \frac{dn}{k}$. Then, the number Z of solutions is zero almost surely if k does not divide $2n$, i.e. $\mathbb{P}[Z = 0] = 1$. Further, the threshold d^* is not an integer.*

(b) *There exists a sharp satisfiability threshold at d^*, i.e. for any increasing sequence $(n_i)_{i\in\mathbb{Z}_{>0}} \subseteq \mathcal{N} = \{n : dk^{-1}n,\ 2k^{-1}n \in \mathbb{Z}_{>0},\}$ and $m_i = m(n_i)$ we have*

$$\lim_{i\to\infty} \mathbb{P}[Z > 0] = \begin{cases} 1 & , d < d^* \\ 0 & , d > d^* \end{cases}.$$

We prove Theorem 1.1 using the second moment method for Z and the small subgraph conditioning method to boost the probability asymptotically to one below the threshold d^*. However, an important question remains at this point, namely what happens when $k > 4$ or $r > 2$.

Our second main result in this paper addresses the behavior for $k > 4$, which can be directly extended to $r > 2$. In particular, a main technical contribution in proving Theorem 1.1 is the optimization of a certain multivariate function that appears in the computation of the second moment, which encodes the interplay between the "similarity" of various assignments and the change in the corresponding probability of being satisfying that they induce. A similar but more complex function appears in the computation of the second moment for $k > 4$, but there we are unfortunately not able to pin down the maximizer. However, apart from that, we discover a surprising connection between this optimization problem and a seemingly unrelated fundamental problem in information theory. In particular, we find that the optimization problem is equivalent to developing a so-called *strong data processing inequality (SDPI)*, which, roughly speaking, encodes the minimum amount of loss in the process of communication through a noisy channel. Such inequalities are of particular importance in the analysis of noisy channels.

We postpone the formal definitions and more relevant background to the next section. Further, we show that our anticipated forms of the corresponding SDPIs directly yield the locations of the global extrema required for our satisfiability threshold proof and thereby imply the following theorem based on Conjecture 1.3.

▶ **Theorem 1.2** (2-in-k Occupation Satisfiability Threshold). *Assume that Conjecture 1.3 is true. Then Theorem 1.1 holds for any $k \in \mathbb{Z}_{>3}$.*

We are confident that this surprising connection does not only apply to 2-in-k occupation problems, but to all r-in-k occupation problems and we believe that it also extends to other classes of random CSPs. Hence, this bridge facilitates the combination of the methods that have been devised in information theory and the study of random graphs, ultimately relating the second moment method to the hypercontractivity ribbon.

1.4 Contraction Coefficients

One central concept in information theory [28, 17] is the notion of a communication channel. Let us assume for concreteness that we have sets $[m]$ and $[n]$ of input and output symbols respectively. We consider the communication through a noisy channel, that is, for a given input $x \in [m]$ the output is $y \in [n]$ with a certain fixed probability $W_{y,x}$. Thus, the channel is completely characterized by its column stochastic transition probability matrix $W = (W_{y,x})_{y\in[n],x\in[m]} \in [0,1]^{n\times m}$.

In a second step, let us consider a distribution P on $[m]$ with probability mass function (pmf) $p \in [0,1]^m$, i.e. a distribution on the inputs. Then the corresponding distribution Q on the received outputs is given by the pmf $q = Wp \in [0,1]^n$. The study of the properties of such channels involves the quantification of the communicated information and further a channel capacity, i.e. the maximum amount of transmittable information. The *data processing inequality (DPI)* is a fundamental result stating that information can only *decrease* when communicated through a noisy channel.

The version of the DPI, see e.g. Lemma 3.11 in [17], relevant here is as follows. Fix a reference input distribution P^* with pmf $p^* \in \mathbb{R}^m$, i.e. the reference output distribution Q^* has the pmf $q^* = Wp^* \in \mathbb{R}^n$. If we then consider an input distribution P with pmf p and the corresponding output distribution Q with pmf $q = Wp$, it is easier to distinguish the distributions P and P^* *before* the transmission. This suggests a loss of information in the process of communication; formally, this means that

$$D_{\mathrm{KL}}(P \parallel P^*) \geq D_{\mathrm{KL}}(Q \parallel Q^*), \quad \text{where} \quad D_{\mathrm{KL}}(P \parallel P^*) = \sum_{x\in[m]} p_x \ln\left(\frac{p_x}{p_x^*}\right).$$

The quantity $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ is the well-known K(ullback)–L(eibler) divergence and one of the most important means of measuring the similarity between given distributions.

This fundamental DPI can be further improved by introducing the optimal ratio $d_* = d_*(P^*, W)$ of $D_{\mathrm{KL}}(Q \parallel Q^*)$ and $D_{\mathrm{KL}}(P \parallel P^*)$ and deriving the tight bound

$$d_* D_{\mathrm{KL}}(P \parallel P^*) \geq D_{\mathrm{KL}}(Q \parallel Q^*) \quad \text{with} \quad d_* = \sup_{P\neq P^*} \frac{D_{\mathrm{KL}}(Q \parallel Q^*)}{D_{\mathrm{KL}}(P \parallel P^*)}.$$

In particular d_* is independent of the input distribution P and the output distribution $Q = Q(P)$. A data processing inequality of this type is referred to as a strong data processing inequality (SDPI) with *contraction coefficient* d_* [2, 4, 3]. In this sense the contraction coefficient $d^*(P^*, W) = d^*(X;Y)$ can be regarded as an alternative measure for the mutual information $I(X;Y)$, i.e. the KL divergence of the distribution of (X,Y) with respect to the distribution of X and Y assuming independence, where the distribution of (X,Y) has the pmf $(W_{yx}p_x^*)_{x,y} \in \mathbb{R}^{n\times m}$. This quantity is of great importance in the analysis of noisy channels and hence not only of interest in theory building, but also in many applications covering image and audio processing, biology, economics and engineering.

1.5 The Conjecture

Let $k \in \mathbb{Z}_{>3}$, $w_1^* = \frac{2}{k}$, $w_2^* = \binom{k}{2}^{-1}$ as defined in (1),

$$W = (W_{(s-1),(t-1)})_{s,t\in[3]} = \begin{pmatrix} 1-2w_1^* & 1-\frac{3}{2}w_1^* & 1-w_1^* \\ 2w_1^* & w_1^* & 0 \\ 0 & \frac{1}{2}w_1^* & w_1^* \end{pmatrix}, \tag{2}$$

and for $w \in \mathcal{W} = \{w \in [0,1]^2 : 2w_1 - 1 \leq w_2 \leq w_1\}$ let

$$p = (p_{s-1})_{s\in[3]} = \begin{pmatrix} 1 - 2w_1 + w_2 \\ 2(w_1 - w_2) \\ w_2 \end{pmatrix}, \quad q = Wp = (q_{s-1})_{s\in[3]} = \begin{pmatrix} 1 - 2w_1^* + w_1^* w_1 \\ 2w_1^*(1 - w_1) \\ w_1^* w_1 \end{pmatrix}. \quad (3)$$

Notice that W is the transition probability matrix of a (fixed) channel for fixed k, that p, q are pmfs for all $w \in \mathcal{W}$ and further any pmf on $\{0,1,2\}$ can be attained by p. As discussed in Section 1.3, $w \in \mathcal{W}$, p and q quantify the similarity of two random satisfying assignments in the following sense. Intuitively and due to symmetry, a given variable i involved in any given constraint takes the value one with probability $w_1^* = \mathbb{P}[X_i = 1]$, while two given variables i, j involved in the constraint both take the value one with probability $w_2^* = \mathbb{P}[X_i + X_j = 2]$ under a random satisfying assignment X. The parameter $w_1 = \mathbb{P}[Y_i = 1|X_i = 1]$ gives the conditional probability that i takes the value one under a second satisfying assignment Y given that i takes the value one under X, while $w_2 = \mathbb{P}[Y_i + Y_j = 2|X_i + X_j = 2]$ gives the conditional probability that both i, j take the value one under Y assuming that they both take the value one under X. Further, p is the pmf of $(Y_i + Y_j)|(X_i + X_j) = 2$, i.e. of the distribution of the number $(Y_i + Y_j)$ of ones taken by i and j under Y given that i, j take one under X, while q is the pmf of the distribution of the number $(X_i + Y_i)$ of ones taken by i under X and Y. In this sense, w_1 and w_2 quantify the similarity of two satisfying assignments X and Y. For example, the choice $w_1 = w_2 = 1$ of parameters implies that Y is determined by X and hence, intuitively, corresponds to a minimum loss of information.

Let P^* be the reference input distribution with pmf $p^* = p(w^*)$, then by the discussion above and in Section 1.4 we can employ the contraction coefficient $d_* = d_*(k) = d_*(P^*, W)$ to quantify the loss of information in a communication through the channel W, and further expect that d_* is attained at $w = (1,1)$.

▶ **Conjecture 1.3** (Contraction Coefficient Conjecture). *The contraction coefficient d_* is attained for the degenerate input pmf p at two, that is,*

$$w_1 = w_2 = 1 \quad \textit{and} \quad d_* = \frac{H(w_1^*)}{-\ln(w_2^*)}.$$

In our contribution, we do not only show that the computation of d_* is equivalent to the optimization problem in the second moment method, but that Conjecture 1.3 is actually equivalent to the applicability of the second moment method.

1.6 Related Work

The regular version of the random 1-in-k occupation problem (and related problems) has been completely solved in [16, 39] using the first and second moment method with small subgraph conditioning. The paper [41] shows that $d^*(k) \geq 2$ for $k \in \mathbb{Z}_{>1}$ in the d-regular 2-in-k occupation problem, i.e. the existence of 2-factors in k-regular simple graphs. A recent discussion of 2-factors (and the related Euler families) that does not rely on the probabilistic method is presented in [6]. Further, randomized polynomial time algorithms for the generation and approximate counting of 2-factors in random regular simple graphs have been introduced in [25].

The study of Erdős Rényi (hyper-)graphs was initiated by the ground breaking publication [22] in 1960 and turned into a fruitful field of research with many applications, including early results on 1-factors in simple graphs [23]. On the contrary, results for the random d-regular k-uniform (hyper-)graph ensemble were rare before the introduction of the configuration (or pairing) model by Bollobás [8] and the development of the small subgraph conditioning method [30, 31] thereafter, see also [44]. While the derived proof scheme facilitated rigorous arguments to establish the existence and location of satisfiability thresholds of random regular CSPs [38, 7, 32, 12, 15, 20, 21, 5], the problems are treated on a case by case basis, while results on entire classes of random regular CSPs are still outstanding.

One of the main reasons responsible for the complexity of a rigorous analysis of random (regular) CSPs seems to be a conjectured structural change of the solution space for increasing densities. This hypothesis has been put forward by physicists, verified in parts and mostly for ER ensembles, further led to new rigorous proof techniques [19, 15, 13] and to randomized algorithms [9, 35] for NP-hard problems that are not only of great value in practice, but can also be employed for precise numerical (though non-rigorous) estimates of satisfiability thresholds An excellent introduction to this *replica theory* can be found in [36, 33, 43]. Specifically, numerical results indicating the satisfiability thresholds for d-regular r-in-k occupation problems (more general variants, and for ER type hypergraphs) based on this conjecture were discussed in various publications [10, 18, 42, 26, 29, 46, 45], where *occupation problems* were introduced for the first time in [40].

Another fundamental obstacle in the rigorous analysis is of a very technical nature and directly related to the second moment method as discussed in detail in our current presentation. In the case of regular r-in-k occupation problems (amongst others) this optimization problem is closely related to the computation of the *contraction coefficient* (for fixed channels and reference distributions) known from information theory. For a general introduction to information theory we recommend [17], while profound discussions and applications of contraction coefficients can be found in [3, 4] and references therein.

1.7 Open Problems

As mentioned in Section 1.2, we focus on the analysis of random regular CSPs with determined constraints. The starting point for this systematic study are r-in-k occupation problems, where we rigorously established the threshold for $r = 2$ and $k = 4$. However, apart from the optimization step in the second moment calculation our proof canonically extends to the general case. A rigorous proof of this step for general r and k is involved, but further assumptions may significantly simplify the analysis. For example, as an extension of the current work one may focus on r-in-$2r$ occupation problems, where the constraints are symmetric in the colors. As can be seen from our proof, this yields useful symmetry properties of the objective function $\frac{D_{\mathrm{KL}}(Q \parallel Q^*)}{D_{\mathrm{KL}}(P \parallel P^*)}$. Further, as suggested by the literature [11, 13, 14] such *balanced* problems [45, 46] are usually more accessible to a rigorous study. On the other hand, the optimization usually also significantly simplifies if only carried out for $k \geq k_0(r)$ for some large $k_0(r)$, as this pushes the minimum to the boundary of the function domain.

Apart from the generalizations discussed above, results for the r-in-k occupation problems are also still outstanding for Erdős-Rényi type CSPs. An analysis of this related problem might allow to tackle the crucial optimization step from a different perspective and thereby also help to establish the thresholds for the regular version.

From the algorithmic perspective, although some methods have been developed for simple graphs [25], we are not aware of algorithms designed specifically to identify solutions of the regular occupation problems (like WalkSAT for the k-SAT problem), only general methods like belief propagation based decimiation. However, problem specific obstacles for the design of such algorithms were discussed in [46].

2 Proof Techniques

In this section we give a high-level overview of our proof, in particular we present the major steps that lead to the main results. We make heavy use of the so-called *configuration model* for the generation of random instances in the form used by Moore [39].

2.1 The Configuration Model

Working with the uniform distribution on d-regular k-uniform hypergraphs directly is challenging. Instead, we show Theorems 1.2 and 1.1 for occupation problems on configurations. A d-regular k-configuration is simply a bijection $g : [n] \times [d] \to [m] \times [k]$, where the v-edges $(i, h) \in \mathrm{dom}(g)$ represent pairs of variables $i \in [n]$ and i-edges, i.e. half-edge indices $h \in [d]$. The image $(a, h') = g(i, h)$ is an f-edge, i.e. a pair of a constraint (factor) $a \in [m]$ and an a-edge (or half-edge) $h' \in [k]$, indicating that the i-edge h of the variable i is wired to the a-edge h' of a and thereby suggesting that i is connected to a in the corresponding d-regular k-factor graph. The number of such d-regular k-configurations on n variables can be easily determined and is given by $(dn)! = (km)!$, hence the uniform distribution on configurations is suitable for combinatorial arguments. Further, the occupation problem on factor graphs directly translates to configurations, which allows to introduce the number Z of solutions of the occupation problem on the random configuration G. In the following we discuss the proof of the analogues to Theorems 1.2 and 1.1 for configurations and further the translation of these results back to factor graphs and hypergraphs.

2.2 The First Moment Method

In the first step we apply the first moment method to the occupation problem on configurations, yielding the following result.

▶ **Lemma 2.1** (First Moment Method). *Let $k \in \mathbb{Z}_{>3}$, $d \in \mathbb{Z}_{>1}$. For $n \in \mathcal{N}$ tending to infinity we have*

$$\mathbb{E}[Z] \sim \sqrt{d}e^{n\phi_1}, \text{ where } \phi_1 = \frac{d}{k}(-\ln(w_2^*)) - (d-1)H(w_1^*).$$

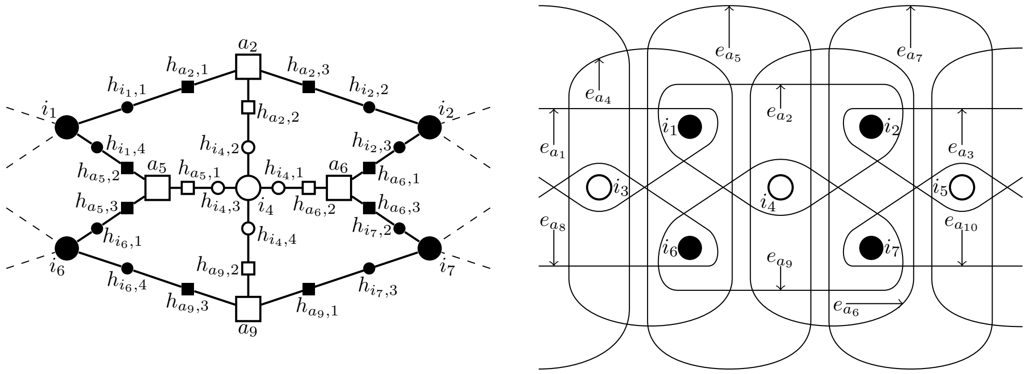
In particular this implies that $\mathbb{E}[Z] \to \infty$ for $d < d^*$ and $\mathbb{E}[Z] \to 0$ for $d > d^*$ with d^* as defined in (1). With an application of Markov's inequality we see that $\mathbb{P}[Z > 0] \to 0$ for $d > d^*$. The map ϕ_1 is known as annealed free entropy density. While the domain of ϕ_1 is trivial in this case (and further in any r-in-k occupation problem), it is non-trivial for the vast majority of CSPs, also covering the general occupation problem.

2.3 The Second Moment Method

Let $k \in \mathbb{Z}_{>3}$, $d \in \mathbb{Z}_{>1}$, further let p, q and $\mathcal{W}$ be the notions from Section 1.5, the distributions P, Q be given by the pmfs p, q and let $\phi_2 : \mathcal{W} \to \mathbb{R}$ be given by

$$\phi_2(w) = \frac{d}{k} D_{\mathrm{KL}}(P \parallel P^*) - (d-1) D_{\mathrm{KL}}(Q \parallel Q^*) \text{ for } w \in \mathcal{W}. \tag{4}$$

Conjecture 1.3 can be used to show that ϕ_2 attains its global minimum at zero iff $w = w^*$ and $d < d^*$. The proof for the specific case $k = 4$ will be presented later in this work. This conclusion then allows to derive the following result using Laplace's method for sums.



(a) occupation problem on configurations. **(b)** 4-regular 2-in-3 vertex cover.

Figure 2 The figure on the left shows the solution on a configuration corresponding to the solution in Figure 1. We only denoted a-edges (small boxes, filled if they the a-edge takes the value one) and i-edges (small circles, filled if the i-edge takes the value one) instead of f-edges and v-edges for brevity (e.g. $h_{a_1,1}$ instead of $(a_1, h_{a_1,1})$). The figure on the right illustrates the corresponding 2-in-3 vertex cover (given by the filled circles).

▶ **Lemma 2.2** (Second Moment Method). *Assume that Conjecture 1.3 holds. Then we have*

$$\frac{\mathbb{E}[Z^2]}{\mathbb{E}[Z]^2} \sim \sqrt{\frac{2}{(2\pi)^2 \prod_{i=0}^{2} p_i(w^*)}} \sqrt{\frac{(2\pi)^2}{\det\left(\frac{k}{\sqrt{2d}} H\right)}} = \sqrt{\frac{k-1}{k-d}},$$

for $n \in \mathcal{N}$ tending to infinity and where H denotes the Hessian of ϕ_2 at $w = w^*$.

Using Lemma 2.2 and Chebyshev's inequality we see that $\mathbb{P}[Z = 0] \leq \sqrt{\frac{k-1}{k-d}} - 1$. While this bound suggests a threshold exists, we need to show that the threshold at d^* is *sharp*.

2.4 Small Subgraph Conditioning

We conclude the proof of the theorem (for configurations) by applying the small subgraph conditioning method to establish that the satisfiability threshold d^* is *sharp*.

▶ **Theorem 2.3** (Small Subgraph Conditioning). *Let Z and $X_1, X_2, \dots$ be non-negative integer-valued random variables. Suppose that $\mathbb{E}[Z] > 0$ and that for each $\ell \in \mathbb{Z}_{>0}$ there are constants $\lambda_\ell \in \mathbb{R}_{>0}$, $\delta_\ell \in \mathbb{R}_{>-1}$ such that*

(a) *for any $\bar{\ell}$ the variables $X_\ell, \dots, X_{\bar{\ell}}$ are asymptotically independent and Poisson distributed with $\mathbb{E}[X_\ell] \sim \lambda_\ell$,*

(b) *for any sequence $r_1, \dots, r_{\bar{\ell}}$ of non-negative integers,*

$$\frac{\mathbb{E}\left[Z \prod_{\ell=1}^{\bar{\ell}} (X_\ell)_{r_\ell}\right]}{\mathbb{E}[Z]} \sim \prod_{\ell=1}^{\bar{\ell}} \mu_\ell^{r_\ell},\ \mu_\ell = \lambda_\ell(1 + \delta_\ell),$$

(c) *we explain the variance, i.e.*

$$\frac{\mathbb{E}[Z^2]}{\mathbb{E}[Z]^2} \sim \exp\left(\sum_{\ell=1}^{\infty} \lambda_\ell \delta_\ell^2\right),\ \left|\sum_{\ell=1}^{\infty} \lambda_\ell \delta_\ell^2\right| < \infty.$$

Then we have $\lim_{n\to\infty} \mathbb{P}[Z > 0] = 1$.

The discussion of the factorial moments in Theorem 2.3 (b) is performed in detail, which requires additional concepts and complex combinatorial arguments. To facilitate the presentation we also give a self-contained proof of the following well-known theorem on the expected number of small cycles (the variables X_ℓ in Theorem 2.3), which can then be extended to a proof of Theorem 2.3 (b). In order to understand what a cycle in a configuration is, we notice that we can represent a configuration g by an equivalent graph with (disjoint) vertex sets given by the variables $V = [n]$, constraints (factors) $F = [m]$, v-edges $H_1 = [n] \times [d]$ and f-edges $H_2 = [m] \times [k]$, where each variable $i \in [n]$ connects to all its v-edges $(i, h_1) \in H_1$, each constraint $a \in [m]$ to all its f-edges $(a, h_2) \in H_2$ and a v-edge (i, h_1) connects to an f-edge (a, h_2) if $g(i, h_1) = (a, h_2)$. Since we are mostly interested in the factor graph associated with a configuration we divide lengths of paths by three, e.g. a cycle of length four in a configuration is actually a cycle of length twelve in its equivalent graph representation. Figures 1a and 2a show an example of a factor graph and the corresponding configuration in its graph representation.

▶ **Theorem 2.4** (Number of Small Cycles). *For $\ell \in \mathbb{Z}_{>0}$ let X_ℓ be the number of 2ℓ-cycles in G, further*

$$\lambda_\ell = \frac{[(k-1)(d-1)]^\ell}{2\ell},$$

and $Z_\ell \sim \mathrm{Po}(\lambda_\ell)$ be independent Poisson distributed random variables. Then the random variables X_ℓ converge in distribution to Z_ℓ for $n \to \infty$, jointly for all $\ell \in \mathbb{Z}_{>0}$.

Using Theorem 2.4 we determine μ_ℓ, δ_ℓ for $\ell \in \mathbb{Z}_{>0}$ and use these results to establish the remaining parts of Theorem 2.3.

▶ **Lemma 2.5.** *The constants μ_ℓ and δ_ℓ for $\ell \in \mathbb{Z}_{>0}$ in Theorem 2.3 are given by*

$$\delta_\ell = \left(-\frac{1}{k-1}\right)^\ell .$$

2.5 Translation of the Results

We first translate the results for configurations to factor graphs using Theorem 2.4, i.e. the contiguity of the factor graph model with respect to the configuration model. For completeness we then also provide self-contained proofs to establish the application to hypergraphs with labeled and unlabeled hyperedges (where the constraints may be attached to either the vertices or to the hyperedges). This establishes our claims in Sections 1.2 and 1.3 except for the verification of Conjecture 1.3 for $k = 4$.

2.6 Contraction Coefficient for $k = 4$

Finally, we prove Conjecture 1.3 for $k = 4$, i.e. we derive Theorem 1.1 from Theorem 1.2. Using a slightly different parametrization and simplifying the KL divergence in the nominator yields

$$\begin{aligned}
d_*(4) &= \sup_{w \in \mathcal{W} \setminus \{w^*\}} R(w),\ R(w) = \frac{D_2(w_1)}{D_1(w)},\\
D_1(w) &= (w_1 - w_2)\ln(6(w_1 - w_2)) + 2w_2 \ln(3w_2) + (1 - w_1 - w_2)\ln(6(1 - w_1 - w_2)),\\
D_2(w_1) &= w_1 \ln(2w_1) + (1 - w_1)\ln(2(1 - w_1)),\\
\mathcal{W} &= \{w \in [0,1]^2 : w_2 \le w_1,\ w_2 \le 1 - w_1\}.
\end{aligned}$$

We focus on suitable lower bounds for D_1, therefore we minimize D_1 with respect to w_2, yielding

$$D_{\min}(w_1) = w_1 \ln(6(w_1 - w_2)) + (1 - w_1)\ln(6(1 - w_1 - w_2)),$$
$$w_2 = w_2(w_1) = \frac{1}{3}\left(2 - \sqrt{12\left(w_1 - \frac{1}{2}\right)^2 + 1}\right), \; w_1 \in [0,1].$$

Since R is symmetric to $w_1 = \frac{1}{2}$, it is sufficient to show that $R \le d_*$ for $w_1 \le \frac{1}{2}$. On this interval we lower bound $D_{\min}$ using the functions

$$D_-(w_1) = 2w_1 \ln\left(\frac{12}{5}w_1\right) + (1 - 2w_1)\ln(6(1 - 2w_1)), \; w_1 \in [0, \bar{w}_1], \text{ and}$$
$$D_+(w_1) = 6\left(\frac{1}{2} - w_1\right)^2, \; w_1 \in [\bar{w}_1, 0.5],$$

where $\bar{w}_1 \approx 0.10831$ is an intersection point of D_- and D_+ that we determined numerically. Finally, we use monotonicity arguments for the corresponding upper bounds of R to derive $R \le d^*$.

References

1 E. Abbe. Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research*, 18:1–86, 2018.

2 R. Ahlswede, P. Gács, and J. Körner. Bounds on conditional probabilities with applications in multi-user communication. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 34(2):157–177, 1976. `doi:10.1007/BF00535682`.

3 V. Anantharam, A. Gohari, S. Kamath, and C. Nair. On hypercontractivity and the mutual information between Boolean functions. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 13–19, October 2013. `doi:10.1109/Allerton.2013.6736499`.

4 V. Anantharam, A. Gohari, S. Kamath, and C. Nair. On hypercontractivity and a data processing inequality. In *2014 IEEE International Symposium on Information Theory*, pages 3022–3026, June 2014. `doi:10.1109/ISIT.2014.6875389`.

5 H. Assiyatun and N. C. Wormald. 3-star factors in random d-regular graphs. *European J. Combin.*, 27(8):1249–1262, 2006. `doi:10.1016/j.ejc.2006.05.003`.

6 M. A. Bahmanian and M. Šajna. Quasi-Eulerian hypergraphs. *Electron. J. Combin.*, 24(3):Paper 3.30, 12, 2017.

7 V. Bapst, A. Coja-Oghlan, and C. Efthymiou. Planting colourings silently. *Combin. Probab. Comput.*, 26(3):338–366, 2017. `doi:10.1017/S0963548316000390`.

8 B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European J. Combin.*, 1(4):311–316, 1980. `doi:10.1016/S0195-6698(80)80030-8`.

9 A. Braunstein, M. Mézard, and R. Zecchina. Survey propagation: an algorithm for satisfiability. *Random Structures Algorithms*, 27(2):201–226, 2005. `doi:10.1002/rsa.20057`.

10 T. Castellani, V. Napolano, F. Ricci-Tersenghi, and R. Zecchina. Bicolouring random hypergraphs. *J. Phys. A*, 36(43):11037–11053, 2003. `doi:10.1088/0305-4470/36/43/026`.

11 A. Coja-Oghla, T. Kapetanopoulos, and N. Müller. The replica symmetric phase of random constraint satisfaction problems. *arXiv e-prints*, page arXiv:1802.09311, February 2018. `arXiv:1802.09311`.

12 A. Coja-Oghlan. Constraint satisfaction: random regular k-SAT. In *Statistical physics, optimization, inference, and message-passing algorithms*, pages 231–251. Oxford Univ. Press, Oxford, 2016.

13 A. Coja-Oghlan, C. Efthymiou, N. Jaafari, M. Kang, and T. Kapetanopoulos. Charting the replica symmetric phase. *Comm. Math. Phys.*, 359(2):603–698, 2018. `doi:10.1007/s00220-018-3096-x`.

14 A. Coja-Oghlan, F. Krzakala, W. Perkins, and L. Zdeborova. Information-theoretic thresholds from the cavity method. *Adv. Math.*, 333:694–795, 2018.

15 A. Coja-Oghlan and K. Panagiotou. The asymptotic k-SAT threshold. *Adv. Math.*, 288:985–1068, 2016. `doi:10.1016/j.aim.2015.11.007`.

16 C. Cooper, A. Frieze, M. Molloy, and B. Reed. Perfect matchings in random r-regular, s-uniform hypergraphs. *Combin. Probab. Comput.*, 5(1):1–14, 1996. `doi:10.1017/S0963548300001796`.

17 I. Csiszár and J. Körner. *Information theory: coding theorems for discrete memoryless systems*. Probability and mathematical statistics. Academic Press, 1981. URL: `https://books.google.de/books?id=wiTvAAAAMAAJ`.

18 L. Dall'Asta, A. Ramezanpour, and R. Zecchina. Entropy landscape and non-Gibbs solutions in constraint satisfaction problems. *Phys. Rev. E (3)*, 77(3):031118, 16, 2008. `doi:10.1103/PhysRevE.77.031118`.

19 J. Ding, A. Sly, and N. Sun. Proof of the satisfiability conjecture for large k. In *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*, pages 59–68. ACM, New York, 2015.

20 J. Ding, A. Sly, and N. Sun. Maximum independent sets on random regular graphs. *Acta Math.*, 217(2):263–340, 2016. `doi:10.1007/s11511-017-0145-9`.

21 J. Ding, A. Sly, and N. Sun. Satisfiability threshold for random regular NAE-SAT. *Comm. Math. Phys.*, 341(2):435–489, 2016. `doi:10.1007/s00220-015-2492-8`.

22 P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

23 P. Erdős and A. Rényi. On the existence of a factor of degree one of a connected random graph. *Acta Math. Acad. Sci. Hungar.*, 17:359–368, 1966. `doi:10.1007/BF01894879`.

24 U. Feige. Relations Between Average Case Complexity and Approximation Complexity. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 534–543, New York, NY, USA, 2002. ACM.

25 A. Frieze, M. Jerrum, M. Molloy, R. W. Robinson, and N. C. Wormald. Generating and counting Hamilton cycles in random regular graphs. *J. Algorithms*, 21(1):176–198, 1996. `doi:10.1006/jagm.1996.0042`.

26 M. Gabrié, V. Dani, G. Semerjian, and L. Zdeborová. Phase transitions in the q-coloring of random hypergraphs. *J. Phys. A*, 50(50):505002, 44, 2017. `doi:10.1088/1751-8121/aa9529`.

27 A. Galanis, D. Štefankovič, and E. Vigoda. Inapproximability for Antiferromagnetic Spin Systems in the Tree Non-uniqueness Region. In *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pages 823–831, New York, NY, USA, 2014.

28 R. G. Gallager. *Information theory and reliable communication*. Wiley, 1968.

29 S. Higuchi and M. Mézard. Correlation-based decimation in constraint satisfaction problems. *Journal of Physics: Conference Series*, 233(1):012003, 2010. URL: `http://stacks.iop.org/1742-6596/233/i=1/a=012003`.

30 S. Janson. Random regular graphs: asymptotic distributions and contiguity. *Combin. Probab. Comput.*, 4(4):369–405, 1995. `doi:10.1017/S0963548300001735`.

31 S. Janson, T. Łuczak, and A. Rucinski. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000. `doi:10.1002/9781118032718`.

32 G. Kemkes, X. Pérez-Giménez, and N. C. Wormald. On the chromatic number of random d-regular graphs. *Adv. Math.*, 223(1):300–328, 2010. `doi:10.1016/j.aim.2009.08.006`.

33 F. Krzakala, F. Ricci-Tersenghi, L. Zdeborová, R. Zecchina, E. W. Tramel, and L. F. Cugliandolo, editors. *Statistical physics, optimization, inference, and message-passing algorithms*. Oxford University Press, Oxford, 2016.

34 S. Kudekar, T. Richardson, and R. Urbanke. Spatially coupled ensembles universally achieve capacity under belief propagation. *IEEE Trans. Inform. Theory*, 59:7761–7813, 2013.

35 E. Maneva, E. Mossel, and M. J. Wainwright. A new look at survey propagation and its generalizations. *J. ACM*, 54(4):Art. 17, 41, 2007. `doi:10.1145/1255443.1255445`.

36 M. Mézard and A. Montanari. *Information, physics, and computation*. Oxford Graduate Texts. Oxford University Press, Oxford, 2009. `doi:10.1093/acprof:oso/9780198570837.001.0001`.

37 M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond*, volume 9 of *World Scientific Lecture Notes in Physics*. World Scientific Publishing Co., Inc., Teaneck, NJ, 1987.

38 M. Molloy, H. D. Robalewska, R. W. Robinson, and N. C. Wormald. 1-factorizations of random regular graphs. *Random Structures Algorithms*, 10(3):305–321, 1997. `doi:10.1002/(SICI)1098-2418(199705)10:3<305::AID-RSA1>3.0.CO;2-#`.

39 C. Moore. The phase transition in random regular exact cover. *Ann. Inst. Henri Poincaré D*, 3(3):349–362, 2016. `doi:10.4171/AIHPD/31`.

40 T. Mora. *Geometry and Inference in Optimization and in Information Theory*. Theses, Université Paris Sud - Paris XI, September 2007. URL: `https://tel.archives-ouvertes.fr/tel-00175221`.

41 H. D. Robalewska. 2-factors in random regular graphs. *J. Graph Theory*, 23(3):215–224, 1996. `doi:10.1002/(SICI)1097-0118(199611)23:3<215::AID-JGT1>3.3.CO;2-K`.

42 C. Schmidt, N. Guenther, and L. Zdeborová. Circular coloring of random graphs: statistical physics investigation. *J. Stat. Mech. Theory Exp.*, 2016(8):083303, 28, 2016. `doi:10.1088/1742-5468/2016/08/083303`.

43 M. Talagrand. *Spin glasses: a challenge for mathematicians*, volume 46 of *Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics*. Springer-Verlag, Berlin, 2003.

44 N. C. Wormald. Models of random regular graphs. In *Surveys in combinatorics, 1999 (Canterbury)*, volume 267 of *London Math. Soc. Lecture Note Ser.*, pages 239–298. Cambridge Univ. Press, Cambridge, 1999.

45 L. Zdeborová and F. Krzakala. Quiet planting in the locked constraint satisfaction problems. *SIAM J. Discrete Math.*, 25(2):750–770, 2011. `doi:10.1137/090750755`.

46 L. Zdeborová and M. Mézard. Constraint satisfaction problems with isolated solutions are hard. *J. Stat. Mech. Theory Exp.*, 2008(12):P12004, 2008. URL: `http://stacks.iop.org/1742-5468/2008/i=12/a=P12004`.