SIAM J. Discrete Math.
Vol. 25, No. 2, pp. 771–808

RECONSTRUCTION AND CLUSTERING IN RANDOM CONSTRAINT SATISFACTION PROBLEMS*

ANDREA MONTANARI†, RICARDO RESTREPO‡, AND PRASAD TETALI§

Abstract. Random instances of constraint satisfaction problems (CSPs) appear to be hard for all known algorithms when the number of constraints per variable lies in a certain interval. Contributing to the general understanding of the structure of the solution space of a CSP in the satisfiable regime, we formulate a set of technical conditions on a large family of random CSPs and prove bounds on three most interesting thresholds for the density of such an ensemble: namely, the *satisfiability* threshold, the threshold for *clustering* of the solution space, and the threshold for an appropriate *reconstruction* problem on the CSPs. The bounds become asymptoticlally tight as the number of degrees of freedom in each clause diverges. The families are general enough to include commonly studied problems such as random instances of Not-All-Equal SAT, k-XOR formulae, hypergraph 2-coloring, and graph k-coloring. An important new ingredient is a condition involving the Fourier expansion of clauses, which characterizes the class of problems with a similar threshold structure.

Key words. random SAT, sharp threshold, message passing algorithms

AMS subject classifications. 68Q87, 82B26, 05C80, 05C15

DOI. 10.1137/090755862

1. Introduction.

Given a set of n variables taking values in a finite alphabet, and a collection of m constraints, each restricting a subset of variables, a constraint satisfaction problem (CSP) requires finding an assignment to the variables that satisfies the constraints. A celebrated example is k-satisfiability (k-SAT), whereby variables are binary and each constraint forbids a subset of k variables to take a specific k-uple of values. Other examples include Not-All-Equal-SAT, hypergraph bicoloring, and graph (vertex) coloring with k colors.

An instance of a CSP can be conveniently described through a factor graph. This is a bipartite graph with m "factor nodes," corresponding to constraints, and n "variable nodes," corresponding to variables. An edge connects variable node $i \in [n] \equiv \{1, \dots, n\}$ to factor node $a \in [m] \equiv \{1, \dots, m\}$ if and only if the ith variable participates in the ath constraint (see Figure 1). The locality structure conveyed by the factor graph plays a key role in our work as well as in statistical mechanics approaches to CSPs [MM06].

In this paper we will study *random* CSP instances, where the number of constraints scales linearly with the number of variables. A precise definition of the distribution of the instances is provided in section 2. An important qualitative feature of these random instances is that the resulting factor graph is a sparse random graph. In particular, such

*Received by the editors April 13, 2009; accepted for publication (in revised form) March 7, 2011; published electronically July 1, 2011.

http://www.siam.org/journals/sidma/25-2/75586.html

†Departments of Electrical Engineering and Statistics, Stanford University, Stanford, CA 94305 (montanari@stanford.edu). This author's research was funded in part by NSF grants CCF-0743978 and DMS-0806211.

‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, and Departamento de Matematicas, Universidad de Antioquia, Medellin, Colombia (restrepo@math.gatech.edu). Scholarship "200 years."

§Schools of Mathematics and Computer Science, Georgia Institute of Technology, Atlanta, GA 30332-0160. This author's research was funded in part by NSF grants DMS-0701043 and CCF-0910584.

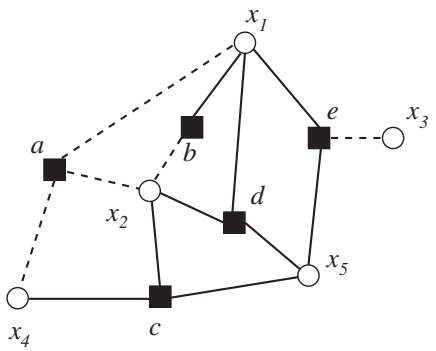

FIG. 1. *Factor graph of a SAT formula: circles represent variable nodes and squares factor nodes.*

graphs are locally tree-like: any neighborhood of bounded depth of a uniformly random vertex converges in distribution to a well-defined random tree.

For several distributions over CSP instances, the probability that a random instance is satisfiable goes *sharply* from 1 to 0 when the number of constraints per variable (the constraint density) crosses a critical threshold [F99], [F05]. This is known as the "satisfiability threshold" or the "satisfiability phase transition." A significant effort has been devoted to the characterization of this phenomenon, and good bounds on the threshold have been proved in some regimes. The most successful approach exploits the sharp concentration of a properly weighted number of solutions. It turns out that this quantity can be controlled using the second moment method when the number of constraints is sufficiently small (see, e.g., [ANP05]), thus proving that the random instances are satisfiable with high probability. For a significantly larger number of constraints, computing the expected number of solutions is sufficient to prove unsatisfiability. While the resulting upper and lower bounds do not coincide, in several cases their ratio converges to 1 as the number of variables per constraint[1] gets large.

This proof technique is nonalgorithmic in the sense that it does not provide any efficient algorithm to construct solutions of random CSP instances. A significant effort has been devoted to the mathematical analysis of polynomial-time algorithms for solving random CSPs. All algorithms studied so far are able to find a solution with probability bounded away from zero, provided the constraint density is smaller than an (algorithm-dependent) threshold. Unfortunately, this threshold appears to be much smaller than the satisfiability threshold. In summary, for a large interval of the constraint density, we know that random CSPs have exponentially many soloutions, but we do not have any efficient algorithm that finds them.

The attempt to understand this universal failure led to studying the geometry of the set of solutions of random CSPs [MPZ02], [AC08] (see also [Sem08]) as well as the emergence of strong correlations among variables in random satisfying assignments [KM+07]. These research directions are motivated by two heuristic explanations of the failure of polynomial algorithms: (1) The space of solutions becomes increasingly complicated as the number of constraints increases and is not captured correctly by simple algorithms; (2) when drawing a uniformly random solution, the induced joint distribution on disjoint subsets of variables becomes increasingly dependent. Local algorithms cannot unveil such dependencies.

[1]As an example, in the case of k-satisfiability, the difference between upper and lower bounds is $O(k)$.

With respect to the geometry of the space of solutions, nonrigorous statistical mechanics analyses conjectured that this be disconnected (in a sense that will be made more precise in the next sections) above a certain threshold in the constraint density. This phenomenon is referred to as the "clustering phase transition" [MPZ02]. Several aspects of it were subsequently proven [AC08].

The emergence of strong correlations is instead defined in connection with the distance structure defined (in the usual way) by the factor graph. Given a satisfiable CSP instance, consider a uniformly random solution of this instance. One can then ask whether the value taken by the ith variable in this solution is correlated or not with the values taken by "far apart" variables (whereby distance is defined with respect to the factor graph). It was conjectured in [KM+07] that, for random CSP instances, correlations vanish asymptotically if and only if the constraint density is below a certain threshold. A precise prediction was provided for this threshold based on statistical mechanics methods. Further, this phase transition was conjectured to coincide with the one in the geometry of the solutions space mentioned above.

The *strength* of correlations mentioned here can be quantified in many equivalent ways, an interesting one being provided by the following thought experiment (also known as "reconstruction problem"). Imagine that a solution of the CSP instance is sampled uniformly at random and that the values of all variables are revealed, except for those that are within distance t (in the factor graph) from the ith one. Does this information allow us to guess the value of the ith variable with success probability significantly larger than in the absence of the same information? This reconstruction problem was studied in some detail in the context of Gibbs measures on trees [MP03] but not for the random CSPs of interest here (the only exception being proper colorings of trees).

A first step towards understanding the relation between clustering and reconstruction was taken in [GM07]. This paper provided an approach to the computation of reconstruction thresholds on sparse random graphs. In the following we will demonstrate that this approach can be successfully applied to random CSPs, thus providing a rigorous foundation for the statistical mechanics picture.

(1) We consider CSPs whose factor graph is a (random) tree. In the case of binary variables and k-ary constraints, we prove bounds on the reconstruction threshold that are optimal to first order, as k goes to infinity.
(2) For these models, we verify the sufficient condition of [GM07], which enables us to transfer the reconstruction result from trees to the same on *sparse* random graphs.
(3) We establish, for the same class of problems, a concentration result for the number of solutions. This allows us to determine the clustering threshold to the same order for large k. We verify that the clustering and reconstruction threshold coincide to this accuracy.

We further prove analogous results for graph coloring with k colors (in this case point (1) was carried out in [Sly09] and point (3) in [AC08]). Our analysis holds for a broad class of CSPs with binary variables, which is characterized through a series of easy-to-check assumptions on the Fourier transform of the constraints. As illustrative examples, we will present specific bounds (on various thresholds) that follow for some standard models, such as the NAE k-SAT, k-XOR formulae, and hypergraph bicoloring.

These results provide rigorous support for the conjectured identity between clustering and reconstruction phase transitions [KM+07]. It further validates the general

methodology of statistical mechanics approaches that—roughly speaking—reduce questions on the geometry of the space of solutions to tree calculations.

Finally, as a by-product, we extend the applicability of the second moment method [ANP05] to a rich class of binary CSPs, thereby showing its genericity. Via "planting" [AC08], this considerably facilitates the study of clustering.

1.1. Related work. As mentioned above, the role played by the geometry of the set of solutions was put forward by statistical physicists [BMW00], [MPZ02], [MZ02]. In particular, these papers unveiled the *clustering* phase transition preceding the satisfiability phase transitions at smaller constraint density. This result motivated the development of surprisingly efficient message passing algorithms to solve random CSPs. For instance, survey propagation has been shown empirically to find solutions of random 3-SAT extremely close to the SAT-UNSAT transition. Rigorous studies confirmed—in a certain interval of constraint density—the emergence of an exponential number of sets (or clusters) of solutions, where solutions within a cluster are closer (in the sense of Hamming distance, say) compared to the intracluster distance [MMZ05], [AR11], [AC08]. Although these results hold only for k-SAT with $k \geq 8$, the resulting bounds on the clustering threshold converge to the statistical physics prediction as $k \to \infty$.

The fact that solutions within a cluster impose long-range correlations among assignments of variables motivated the study of the so-called reconstruction problem in the context of random CSPs. As mentioned, nonrigorous statistical mechanics calculations imply that the clustering and reconstruction thresholds coincide [MM06], [KM+07].

Finally, understanding the threshold for (non)reconstruction is also becoming relevant, if not crucial, to understanding the limit of the Glauber dynamics to sample from the set of solutions of a CSP. Indeed, nonreconstructibility was proved in [BK+05] to be a necessary condition for fast mixing and is expected to be sufficient for a large class of "sufficiently random" problems. Reconstruction problems were intensively studied on trees (see, e.g., [MP03]). A recent paper [GM07] provides sufficient conditions under which the reconstruction problem on locally tree-like graphs is solvable if and only if it is solvable on the associated random trees.

1.2. Plan of the paper. The organization of the paper is as follows. In section 2, we give the formal definitions and assumptions of our models. We state our main results in section 3. In section 4, we state and prove the optimal bounds for the tree reconstruction problem. In section 5, we verify the sufficient condition (from [GM07]) for the specific problem of proper graph q-coloring, thus proving one of our main results—optimal bounds on the (sparse) random graph reconstruction problem for colorings. In Appendix A, we derive a certain technical second moment bound that is needed to prove our theorem on the satisfiability threshold. In Appendix B, we prove various technical results needed to complete the proof of the clustering threshold. In Appendix C, certain sharp threshold results are derived making use of recent results of [AC08], [CD09] so that we can extend the high-probability statements derived in the previous appendices to hold with probability tending to one. Further details on what is proved in these appendices appear in section 3.3, after the precise statement of our main results.

2. Definitions. In this section we define a family of random CSP ensembles: problems with constraints involving k-tuples of binary variables. We further define q-ary ensembles as a natural extension of the latter. We finally introduce some analytic definitions that will be necessary in order to present our results.

Binary k-CSP ensemble. Given an integer n, $\alpha \in \mathbb{R}_+$, and a distribution $p = \{p(\varphi)\}$ over Boolean functions $\varphi : \{+1, -1\}^k \to \{0, 1\}$, $\mathrm{CSP}(n, \alpha, p)$ is the ensemble of random CSPs over n Boolean variables $\underline{x} = (x_1, \ldots, x_n)$ defined as follows. For each $a \in \{1, \ldots, m = n\alpha\}$, draw k indices $i_a(1), \ldots, i_a(k)$ independently and uniformly at random in $[n]$ and a function φ_a with distribution $p(\varphi)$. An assignment $\underline{x}$ satisfies the resulting instance if $\varphi_a(x_{i_a(1)}, \ldots, x_{i_a(k)}) = 1$ for each $a \in [m]$. A CSP instance can be naturally described by a bipartite graph G (often referred to in the literature as a "factor graph"), including a node for each clause $a \in [m]$ and for each variable $i \in [n]$, and an edge (i, a) whenever variable x_i appears in the ath clause.

q-ary ensembles. A q-ary ensemble is the natural generalization of a binary ensemble to the case in which variables take q values. For the sake of simplicity, we restrict our discussion here to the case of pairwise constraints (i.e., $k = 2$ in the language of the previous paragraph).

Given an integer n, $\alpha \in \mathbb{R}_+$, and a distribution $p = \{p(\varphi)\}$ over Boolean functions $\varphi : [q] \times [q] \to \{0, 1\}$, $\mathrm{CSP}_q(n, \alpha, p)$ is the collection of random CSPs over q-ary variables x_i for $i = 1, 2, \ldots, n$ defined as follows. For each $a \in \{1, \ldots, m = n\alpha\}$, draw 2 indices i_a, j_a independently and uniformly at random in $[n]$, and a function φ_a with distribution $p(\varphi)$. An assignment $\underline{x} = (x_1, \ldots, x_n)$ satisfies the resulting instance if $\varphi_a(x_{i_a}, x_{j_a}) = 1$ for each $a \in [m]$.

In this paper, by way of illustrating how the results for binary ensembles could be (purportedly) extended to q-ary ensembles, we will study the q-coloring model which consists of ensembles with the single clause $\varphi(x, y) = \mathbb{I}(x \neq y)$. This model corresponds to proper colorings with q colors of a random sparse graph with an edge-to-vertex density of $\alpha > 0$.

3. Main results. As mentioned in the introduction, our goal is estimating the thresholds for satisfiability, clustering, and reconstruction in random CSPs. In general, one should speak of threshold functions depending on the problem size n. With a slight abuse of notation, we shall leave implicit the dependence on n of threshold functions unless necessary.

3.1. Binary *k*-CSP ensembles.

3.1.1. Assumptions. We will always assume the following basic conditions on the CSP ensemble.

1. *Permutation symmetry.* If φ^π is the Boolean function obtained from φ by permuting its arguments, we require $p(\varphi^\pi) = p(\varphi)$. (Notice that this assumption does not imply any loss of generality in this context. Indeed, in the definition of the ensemble $\mathrm{CSP}(n, \alpha, p)$ the indexes of the arguments of clause $\varphi_a(x_{i_a(1)}, \ldots, x_{i_a(k)})$ are independent and uniformly random.)
2. *Balance.* The distribution p is supported on Boolean functions such that $\varphi(x_1, \ldots, x_k) = \varphi(-x_1, \ldots, -x_k)$. This condition implies that the odd Fourier coefficients of φ are zero. Indeed, this condition can be regarded as the most restrictive in a structural sense. By introducing it, we rule out well-studied models such as k-SAT.
3. *Feasibility.* For each Boolean function φ in the support of p, every partial assignment $(x_1, \ldots, x_{k-1})$ can be extended to a satisfying assignment $(x_0, x_1, \ldots, x_{k-1})$ of φ. This condition implies that $\|\varphi\|^2 \geq 1/2$. (See section 3.4 for the definition of norm.)

We will also make further assumptions that are more conveniently formulated in terms of the Fourier spectrum of the constraints φ. In order to simplify the exposition,

we postpone these conditions to section 3.5. These assumptions will be denoted as per the following definition.

DEFINITION 3.1. *We say that the probability distribution* $p = \{p(\varphi)\}$ *over Boolean functions* $\varphi : \{+1, -1\}^k \to \{0, 1\}$ *has the property of* dominance of balanced assignments *if it satisfies condition* 4 *in section* 3.5.

We say that $p = \{p(\varphi)\}$ *is* consistent *if it has properties* 1–3 *and dominance of balanced assignments and further satisfies conditions* (a) *and* (b) *in section* 3.5.

We finally say that $p = \{p(\varphi)\}$ *is* clustering-consistent *if it further satisfies conditions* (a′) *and* (b′) *in section* 3.5.

Intuitively, the condition of dominance of balanced assignments ensures that most of the assignments satisfying a typical instance from the ensemble are "balanced." By the latter we mean that they have roughly half of the variables taking value $+1$ and half taking -1.

The condition of being clustering-consistent is instead related to the fact that each constraint does not depend mostly on a small subset of its k arguments. Finally, the condition of being clustering-consistent amounts to a strengthening of the above.

3.1.2. Results. An ensemble of binary k-CSPs will be characterized by the following quantities:

$$\frac{1}{\boldsymbol{\Omega}_k} \stackrel{\text{def}}{=} \mathbb{E}_\varphi \frac{2\mathrm{I}_1(\varphi)}{1 - 2\mathrm{I}_1(\varphi)}, \qquad \frac{1}{\hat{\boldsymbol{\Omega}}_k} \stackrel{\text{def}}{=} -\mathbb{E}_\varphi \log(1 - 2\mathrm{I}_1(\varphi)), \qquad \frac{1}{\tilde{\boldsymbol{\Omega}}_k} \stackrel{\text{def}}{=} \frac{2\mathbb{E}_\varphi \mathrm{I}_1(\varphi)}{1 - 2\mathbb{E}_\varphi \mathrm{I}_1(\varphi)}.$$

Here $I_1(\varphi)$ is the *influence* of constraint φ. "Influence" is a basic notion in discrete Fourier analysis that describes how much the value of φ is sensitive to any single argument. For a formal definition we refer the reader to section 3.4.

Notice that $\boldsymbol{\Omega}_k \le \hat{\boldsymbol{\Omega}}_k$ and that $\boldsymbol{\Omega}_k \le \tilde{\boldsymbol{\Omega}}_k$. Indeed, the first inequality follows by using the inequality $\log(z) \le z - 1$ with $z = 1/(1 - 2I_1)$, and the second follows by Jensen's, noting the convexity of $x \mapsto (2x)/(1 - 2x)$. Moreover, $\hat{\boldsymbol{\Omega}}_k \approx (e^{1/\hat{\boldsymbol{\Omega}}_k} - 1)^{-1} \le \tilde{\boldsymbol{\Omega}}_k$; indeed, denoting $1/\hat{\boldsymbol{\Omega}}_k$ as $\mathbb{E}(X)$ and using Jensen's, we have

$$\frac{1}{\tilde{\boldsymbol{\Omega}}_k} = \frac{\mathbb{E}(1 - e^{-X})}{\mathbb{E}e^{-X}} = \frac{1}{\mathbb{E}e^{-X}} - 1 \le e^{\mathbb{E}(X)} - 1 = e^{1/\hat{\boldsymbol{\Omega}}_k} - 1.$$

PROPOSITION 3.2. *A random binary constraint satisfaction instance from the* consistent ensemble $\mathrm{CSP}(n, \boldsymbol{\alpha}, p)$ *is satisfiable, with high probability, if* $\boldsymbol{\alpha} < \underline{\alpha}_{\mathrm{s}}(k)(1 - o_n(1))$, *where*

$$\boldsymbol{\Omega}_k \log 2\{1 + o_k(1)\} \le \underline{\boldsymbol{\alpha}}_{\mathbf{s}}(k, n).$$

Vice versa, if $\boldsymbol{\alpha} > \bar{\boldsymbol{\alpha}}_{\mathrm{s}}(k)(1 + o_n(1))$, *where*

$$\bar{\boldsymbol{\alpha}}_{\mathrm{s}}(k, n) \le \hat{\boldsymbol{\Omega}}_k \log 2\{1 + o_k(1)\},$$

then, with high probability, a $\mathrm{CSP}(n, \boldsymbol{\alpha}, p)$ *instance is unsatisfiable. Further* $|\boldsymbol{\Omega}_k^{-1} - \hat{\boldsymbol{\Omega}}_k^{-1}| \le 8\mathbb{E}_\varphi\{\mathrm{I}_1(\varphi)^2\}$.

As clarified by the last part of the statement, the upper and lower bound approach each other when the influence of a single variable in a clause becomes smaller.

Given a measure $\mu(\underline{x})$ over variable assignments in $\{+1, -1\}^V$, the reconstruction problem is said to be unsolvable if correlations with respect to μ decay rapidly with the

distance r on G. More precisely, if $\mu_{i,\sim r}$ denotes the joint distribution of x_i and $\{x_j : d_G(i,j) \geq r\}$, then $\lim_{r\to\infty} \limsup_{n\to\infty} \mathbf{E}\|\mu_{i,\sim r} - \mu_i \mu_{\sim r}\|_{\mathrm{TV}} = 0$.

THEOREM 3.3. *Let $\mu(\underline{x})$ be the uniform measure over solutions of an instance from the* consistent ensemble $\mathrm{CSP}(n,\alpha,p)$. *The reconstruction problem is solvable for μ if $\alpha > \bar{\alpha}_{\mathrm{r}}(k)$, and it is unsolvable for μ if $\alpha < \underline{\alpha}_{\mathrm{r}}(k)$, where*

$$\bar{\alpha}_{\mathrm{r}}(k) = \frac{\Omega_k}{k}\{\log\, k + o(\log\, k)\}, \qquad \underline{\alpha}_{\mathrm{r}}(k) = \frac{\Omega_k}{k}\{\log\, k - o(\log\, k)\}.$$

Given an instance of $\mathrm{CSP}(n,\alpha,p)$, a $d_{\max}$-cluster of solutions is any equivalence class of solutions under the (closure of the) relation $\underline{x} \simeq \underline{x}'$ if $d_{\mathrm{Hamming}}(\underline{x},\underline{x}') \leq d_{\max}$. We say that the set of solutions is *clustered* if it is partitioned into exponentially many clusters for some function $d_{\max} = d_{\max}(n)$ with $d_{\max}(n) \uparrow \infty$ as $n \to \infty$.

THEOREM 3.4. *Consider a* clustering-consistent ensemble $\mathrm{CSP}(n,\alpha,p)$. *The set of solutions of a random instance from this ensemble is clustered, with high probability, if $\alpha > \alpha_{\mathrm{d}}(k)$, where*

$$\alpha_{\mathrm{d}}(k) = \frac{\tilde{\Omega}_k}{k}\{\log\, k + \mathrm{o}(\log\, k)\}.$$

Further $|\tilde{\Omega}_k^{-1} - \Omega_k^{-1}| \leq 8\mathbb{E}_\varphi\{I_1(\varphi)^2\}$.

Thus, a key result of the present paper is that, for a large number of ensembles, $\alpha_{\mathrm{d}}(k)$ and $\underline{\alpha}_r(k)$ (as well as $\bar{\alpha}_{\mathrm{r}}(k)$) differ at most by a quantity whose relative size is negligible for large k.

3.2. q-ary ensembles: Graph coloring. The following results concerning the colorability and clustering of proper colorings were proved by Achlioptas and Naor [AN05] and Achlioptas and Coja-Oghlan [AC08], respectively.

THEOREM 3.5 (graph q-colorability [AN05]). *A random graph with n vertices and $n\alpha$ edges is satisfiable, with high probability, if $\alpha < \alpha_{\mathrm{s}}(q)$, where*

$$\alpha_{\mathrm{s}}(q) = q[\log\, q + o_q(1)].$$

Vice versa, if $\alpha > \alpha_{\mathrm{s}}(q)(1 + o_q(1))$, such a graph is with high probability uncolorable.

THEOREM 3.6 (clustering of q-colorings [AC08]). *The set of proper q-colorings of a random graph with n vertices and $n\alpha$ edges is clustered with high probability, if $\alpha > \alpha_{\mathrm{d}}(q)$, where*

$$\alpha_{\mathrm{d}}(q) = \frac{q}{2}[\log\, q + o(\log\, q)].$$

One of our main results is to prove a corresponding reconstruction theorem for this model as follows.

THEOREM 3.7 (graph q-coloring reconstruction). *Let $\mu(\underline{x})$ be the uniform measure over of proper q-colorings of random graph with n vertices and $n\alpha$ edges. For q large enough, the reconstruction problem is solvable for μ if $\alpha > \alpha_{\mathrm{r}}(q)$, where*

$$\alpha_{\mathrm{r}}(q) = \frac{q}{2}[\log\, q + \log\log\, q + O(1)].$$

Vice versa, the reconstruction problem is unsolvable, with high probability, if $\alpha < \alpha_{\mathrm{r}}(q)$.

3.3. General strategy. The results described in the previous section are of three types: bounds on the satisfiability thresholds (cf. Proposition 3.2 and Theorem 3.5);

bounds on the clustering threshold (cf. Theorems 3.4 and 3.6); and bounds on the reconstruction threshold (cf. Theorems 3.3 and 3.7). The proof strategy is as follows.

The satisfiability threshold can be upper-bounded using the first moment of the number of solutions and lower-bounded using the second moment method. This technique is discussed in detail in [AM02], [AN05], [ANP05]; we describe its application to the general CSP(n, α, p) ensemble in Appendix A.

The clustering threshold can be upper-bounded through an analysis of the recursive "whitening" process that associates to each cluster a single configuration in an extended space [AR11]. This naive estimate of the clustering threshold is, however, far from tight. Significantly better upper bounds on this threshold were obtained in [AC08] by approximating the CSP ensemble with an appropriate "planted ensemble." Theorems 3.4 and 3.6 use this approach.

The proof of Theorem 3.4 is presented in Appendix B. While the general approach is the same developed in [AC08], several technical steps are new and potentially useful in other contexts: (i) We show that the Fourier spectrum of clauses and the Bonami–Beckner operator are natural tools for the relevant calculations; (ii) we use a recent result by Creignou and Daude [CD09] to prove that the property of having more than e^{an} solutions has a sharp threshold for any constant a (such a sharp threshold result was established earlier for specific cases [AR11]).

The reconstruction threshold is characterized via a three-step procedure.

(1) Bound the reconstruction threshold for an appropriate ensemble of (infinite) tree instances, i.e., CSP instances for which the associated factor graph is an infinite Galton–Watson tree. In the case of proper q-colorings, a sharp characterization was obtained independently by two groups in the past year [BVV11], [Sly09]. In section 4 we prove sharp bounds on tree reconstruction for binary CSPs. The proof amounts to deriving an exact distributional recursion for the so-called belief process and carefully bounding its asymptotic behavior.

(2) Call a solution *balanced* if each possible variable value is taken on the same number of vertices. Given two balanced solutions $\underline{x}^{(1)}, \underline{x}^{(2)}$, define their *joint type* $\nu(x, y)$ as the matrix such that the fraction of vertices i with $x_i^{(1)} = x$ and $x_i^{(2)} = y$ is equal to $\nu(x, y)$. Consider the number $Z_{\rm b}(\nu)$ of balanced solution pairs $\{\underline{x}_1, \underline{x}_2\}$, with joint type ν. One has to show that $\mathbb{E}Z_{\rm b}(\nu)$ is exponentially dominated by its value at the uniform type $\bar{\nu}(x, y) = 1/q^2$ (with $q = 2$ for binary CSPs). More precisely, $\mathbb{E}Z_{\rm b}(\nu) \doteq \exp\{n\Phi(\nu)\}$ with Φ achieving its unique maximum at $\bar{\nu}$.
This is also a crucial step in the second moment method. It was accomplished in [AN05] for proper q-colorings of random graphs. In the case of binary CSPs, we prove this estimate in Appendix A

(3) Prove that step (2) above implies, for the model in consideration, that the set of solutions of a random instance is, with high probability, *roughly spherical*. By this we mean that the joint type ν_{12} of two uniformly random solutions $\underline{x}^{(1)}, \underline{x}^{(2)}$ satisfies $\|\nu_{12} - \bar{\nu}\|_{\rm TV} \le \delta$ with high probability for all $\delta > 0$. Notice that this implication requires bounding the expected ratio of $Z_{\rm b}(\nu)$ to the total number of solution pairs. We prove that the implication nevertheless holds in section 5 for q-colorings. The argument for binary CSPs is completely analogous, and we omit it.

Finally, it was proved in [GM07] that, under such a sphericity condition, graph reconstruction and tree reconstruction are equivalent, which finishes the proof of Theorems 3.3 and 3.7.

Notice that the techniques used for the clustering and reconstruction thresholds are very different. Thus it is a surprising (and arguably deep) phenomenon that they do coincide as far as the present techniques can tell.

3.4. Fourier analysis of constraints. In this section, we briefly review some well-known definitions in discrete Fourier analysis. For general background on this material, the reader may consult any classical textbook on (discrete) Fourier analysis or the lecture notes by Diaconis [Dia88]; for a more breezy introduction and a summary of some key tools one may also find the recent survey [Odo08] useful.

Functional analysis of clauses. We denote by v_θ the measure defined over $\{-1,+1\}^k$ such that

$$v_\theta(x) = \prod_{i=1}^{k}\left(\frac{1+x_i\theta}{2}\right) \tag{1}$$

for every $x \in \{-1,+1\}^k$. This is just the measure induced by choosing k independent copies of a random variable that takes values ± 1 and has expectation θ. Notice that when $\theta = 0$, v_θ corresponds to the uniform measure over $\{-1,+1\}^k$.

The inner product induced by this measure, on the space of real functions defined on $\{-1,+1\}^k$, is denoted by $(\cdot,\cdot)_\theta$, and the corresponding norm is denoted by $\|\cdot\|_\theta$. If $\theta = 0$, we drop the subindex and just use $(\cdot,\cdot)$ and $\|\cdot\|$, respectively. Thus, if f, $g\colon\{-1,+1\}^k \to \mathbb{R}$, then

$$(f,g)_\theta = \sum_{x\in\{-1,+1\}^k} f(x)g(x)v_\theta(x), \qquad \|f\|_\theta^2 = \sum_{x\in\{-1,+1\}^k} f^2(x)v_\theta(x),$$
$$(f,g) = \frac{1}{2^k}\sum_{x\in\{-1,+1\}^k} f(x)g(x), \qquad \|f\|^2 = \frac{1}{2^k}\sum_{x\in\{-1,+1\}^k} f^2(x).$$

We denote the Hilbert space of functions $\{-1,+1\}^k \to \mathbb{R}$ under the inner product $(\cdot,\cdot)$ by J_k.

Fourier transform of clauses. For any $Q \subseteq [k] \equiv \{1,\ldots,k\}$, let $\gamma_Q(x) \stackrel{\text{def}}{=} \prod_{i\in Q} x_i$. Under the scalar product defined above (with $\theta = 0$), the functions $\{\gamma_S\}_{S\subseteq[k]}$ form an orthonormal basis for J_k. Moreover, they are exactly the algebraic characters of $\{-1,1\}^k$ with the group operation of pointwise multiplication. Thus, we define the Fourier transform of a function $f \in J_k$ by letting, for any $Q \subseteq [k]$,

$$f_Q \stackrel{\text{def}}{=} (\gamma_Q, f) = 2^{-k}\sum_{x\in\{-1,+1\}^k} f(x)\gamma_Q(x).$$

Noise operator. Given $\theta \in [-1,1]$, we recall the *Bonami–Beckner* operator $\mathrm{T}_\theta\colon J_k \to J_k$ [Bon70], [Bec75], by

$$(\mathrm{T}_\theta f)(x) \stackrel{\text{def}}{=} \sum_{y\in\{-1,1\}^k} f(xy)v_\theta(y),$$

where $xy = (x_1y_1,\ldots,x_ky_k)$. Notice that $(\mathrm{T}_\theta f)(x)$ corresponds to the expected value of $f(\mathbf{x}_\theta)$, where $\mathbf{x}_\theta$ is obtained from x by flipping each coordinate independently with

probability $(1-\theta)/2$. Notice that T_1 is just the identity operator and T_0 sends f to the constant function $(f, \gamma_\varnothing)$.

The Bonami–Beckner operator diagonalizes with respect to the Fourier basis in the sense that $(\mathrm{T}_\theta \gamma_Q)(x) = \theta^{|Q|}\gamma_Q(x)$ for any $Q \subseteq [k]$.

More generally, given $h \in [-1,1]^k$, we define $(\mathrm{T}_h f)(x) \stackrel{\text{def}}{=} \mathbb{E}[f(\mathbf{x}_h)]$, where $\mathbf{x}_h$ is obtained from x by flipping the ith coordinate independently and with probability $\frac{1-h_i}{2}$. Since T_h also diagonalizes with respect to the Fourier basis, one gets $(\mathrm{T}_h \gamma_S)(x) = \gamma_S(h)\gamma_S(x)$.

Discrete derivative and influence. Given a function $f \in J_k$, we define its *discrete derivative* $f^{(1)} \in J_{k-1}$ as $f^{(1)}(x) = \frac{1}{2}[f(1,x) - f(-1,x)]$. We define analogously $f^{(i)}$ for any other variable index. Finally, the influence of the ith variable on f is the norm of the derivative

$$\mathrm{I}_i(f) \stackrel{\text{def}}{=} \|f^{(i)}\|^2.$$

For any $Q \subseteq [k]$, $f^{(i)}_Q = f_{Q \cup \{i\}}$.

3.5. Assumptions on the Fourier spectrum. We now formally state the conditions for *consistent* and *clustering-consistent* ensembles. We start with the notion of dominance of balanced assignments.

4. *Dominance of balanced assignments.* For every $\theta \in [-1,1]$,

$$\mathbb{E}_\varphi \, \log \|\varphi\|_\theta \le \mathbb{E}_\varphi \, \log \|\varphi\|$$

with equality if and only if $\theta = 0$. This condition implies that, in a typical random instance, most solutions are balanced in the sense that they have almost as many $+1$'s as -1's.

While our ultimate goal is to exhibit results as $k \to \infty$, the probability distribution p over the functions $\varphi : \{-1,1\}^k \to \{0,1\}$ must be defined for *every* k, and some agreement should exist between such probability distributions for different k's. In our work this agreement is given by two conditions concerning the derivative of the clauses in the support of p.

(a) *ℓ_1 norm of the Fourier transform grows at most polynomially in k.* That is, for every $\varphi \in \mathrm{supp}(p)$,

$$\sum_Q |\varphi^{(i)}_Q| \le k^a \tag{2}$$

for some constant a not depending on k, and recall that $\varphi^{(i)}_Q = (\gamma_Q, \varphi^{(i)})$.

(b) "*Small-weight*" *Fourier coefficients are small.* There is a constant $C > 0$ (not depending on k) such that for every $\varphi \in \mathrm{supp}(p)$,

$$\|\mathrm{T}_\theta \varphi^{(i)}\|^2 \le e^{-Ck(1-\theta)} \|\varphi^{(i)}\|^2, \qquad \theta \in [0,1]. \tag{3}$$

Notice that the feasibility condition implies that all the variables of φ have the same influence, namely,

$$\mathrm{I}_i(\varphi) = \frac{1 - \|\varphi\|^2}{2}. \tag{4}$$

In order to prove this, consider, say, $i = 1$, and let $N_{ab}(\varphi)$, $a, b \in \{0,1\}$, be the number of partial assignments $x_1, \ldots, x_{k-1}$ such that $\varphi(+1, x_1, \ldots, x_{k-1}) = a$ and $\varphi(-1, x_1, \ldots, x_{k-1}) = b$. Then, by definition we have

$$\|\varphi\|^2 = \frac{1}{2^k}[N_{01}(\varphi) + N_{10}(\varphi) + 2N_{11}(\varphi)], \tag{5}$$

$$\mathrm{I}_1(\varphi) = \frac{1}{2^{k+1}}[N_{01}(\varphi) + N_{10}(\varphi)], \tag{6}$$

whence our claim (4) follows using $N_{01}(\varphi) + N_{10}(\varphi) + 2N_{11}(\varphi) = 2^{k-1}$.

Condition (a) above on the ℓ_1 norm of the Fourier transform implies, in particular, that for any fixed l, there exists $A_l > 0$ (independent of k) such that

$$\sum_{1\le|Q|\le l} |\varphi_Q|^2 \le A_l e^{-Ck/2} \sum_{|Q|\ge 1} |\varphi_Q|^2. \tag{7}$$

An equivalent formulation of (3), with a possibly different constant C, is

$$(\mathrm{T}_\theta \varphi^{(i)}, \varphi^{(i)}) \le e^{-Ck(1-\theta)} \|\varphi^{(i)}\|^2, \qquad \theta \in [0,1]. \tag{8}$$

In order to establish clustering, we require two more conditions:

(a′) First, we have a slightly stronger form of *dominance of balanced assignments*:

$$\mathbb{E}_\varphi\{\|\varphi\|_\theta^2\} \le \mathbb{E}_\varphi\{\|\varphi\|^2\}.$$

(b′) Next we have the following condition on the Fourier transform of clauses:

$$\sum_{Q_1\subseteq Q_2} \mathbb{E}_\varphi\{\varphi_{Q_1}\varphi_{Q_2}\}\theta^{|Q_1|}\delta^{|Q_2|-|Q_1|} \le \sum_Q \mathbb{E}_\varphi\{\varphi_Q^2\}\theta^{|Q|}$$

holding for all $\theta \in [-1,+1]$, $\delta \in [0, 1-|\theta|]$. In particular, the latter condition holds whenever $p(\varphi^{(s)}) = p(\varphi)$ for all $s = (s_1, \ldots, s_k) \in \{+1,-1\}^k$, where $\varphi^{(s)}(x_1, \ldots, x_k) = \varphi(s_1 x_1, \ldots, s_k x_k)$, that is, when the ensemble is closed under *polarization* [CD04].

3.6. Examples. In this section, we apply our results to a few concrete examples.

Example 1: 2-*coloring hypergraphs.* Let us consider the ensemble of CSPs consisting of clauses of the type φ, where $\varphi(x_1, \ldots, x_k) = \mathbb{I}(\sum x_i \notin \{-k, k\})$. The $\mathrm{CSP}(n, \alpha, p)$ in this case corresponds to the distribution of 2-colorings of a random hypergraph on n vertices and αn edges, with edge size k, and each edge chosen independently and uniformly at random.

The conditions 1–3 (permutation symmetry, balance, and feasibility) clearly hold for this model. The dominance of balanced assignments, in its weak and strong form, follows after checking that $\|\varphi\|_\theta^2 = 1 - (\frac{1+\theta}{2})^k - (\frac{1-\theta}{2})^k$ is maximized at $\theta = 0$. To establish condition (a) (cf. (2)), notice that

$$\varphi_Q^{(i)} = -\frac{1}{2^k}[1 - (-1)^{|Q|}],$$

which clearly implies that the ℓ_1 norm of the Fourier transform is bounded. In order to check condition (b) (cf. (3)), notice that

$$\frac{(\mathrm{T}_\theta \varphi^{(i)}, \varphi^{(i)})}{\|\varphi^{(i)}\|^2} = \left(\frac{1+\theta}{2}\right)^{k-1} - \left(\frac{1-\theta}{2}\right)^{k-1} \le e^{-k(1-\theta)/2}$$

for all $\theta \in [0,1]$. On the other hand, we have that

$$\sum_{Q_1\subseteq Q_2}\mathbb{E}_\varphi\{\varphi_{Q_1}\varphi_{Q_2}\}\theta^{|Q_1|}\delta^{|Q_2|-|Q_1|}$$
$$=\left(1-\frac{1}{2^{k-1}}\right)-\frac{1}{2^k}[(1+\delta)^k+(1-\delta)^k]$$
$$+\left(\frac{1}{2^k}\right)^2[(1+(\delta+\theta))^k+(1-(\delta+\theta))^k+(1+(\delta-\theta))^k+(1-(\delta-\theta))^k],$$

and the previous expression reaches its maximum for $\delta=0$. Thus,

$$\sum_{Q_1\subseteq Q_2}\mathbb{E}_\varphi\{\varphi_{Q_1}\varphi_{Q_2}\}\theta^{|Q_1|}\delta^{|Q_2|-|Q_1|}\le\left(1-\frac{1}{2^{k-2}}\right)+\left(\frac{1}{2^k}\right)^2\left[\frac{(1+\theta)^k+(1-\theta)^k}{2}\right],$$

and the right-hand side of the previous formula is equal to $\sum_Q\mathbb{E}_\varphi\{\varphi_Q^2\}\theta^{|Q|}$, proving condition (b′).

Now, an easy computation shows that $\boldsymbol{\Omega}_k=\widetilde{\boldsymbol{\Omega}}_k=2^{k-1}-1$ and $\hat{\boldsymbol{\Omega}}_k^{-1}=-\log(1-2^{-k+1})$; therefore we have the following:

	Reconstruction–clustering	Lower bound satisfiability	Upper bound satisfiability
2-coloring	$(2^{k-1}/k)[\log\ k+\mathrm{o}(\log\ k)]$	$2^{k-1}\ \log\ 2[1+\mathrm{o}(1)]$	$2^{k-1}\ \log\ 2[1+\mathrm{o}(1)]$

Example 2: *Not-All-Equal-k*–SAT. Let us consider now an ensemble of CSPs consisting of clauses of type $\{\varphi_s\}_{s\in\{+1,-1\}^k}$, where $\varphi_s(x_1,\ldots,x_k)=\mathbb{I}(\sum x_is_i\notin\{-k,k\})$ and $p(\varphi_s)=2^{-k}$ for each $s\in\{+1,-1\}^k$. In this case, the CSP(n,α,p) model corresponds to the distribution of Not-All-Equal-k–SAT instances for a random formula in n variables, consisting of αn random clauses, each with k literals.

For this model, the conditions 1–3 are easily verified. The dominance of balanced assignments in its strong form follows from the fact that

$$\mathbb{E}_s\|\varphi\|_\theta^2=\mathbb{E}_s\left(1-\prod_{i=1}^k\frac{1+s_i\theta}{2}-\prod_{i=1}^k\frac{1-s_i\theta}{2}\right)=\mathbb{E}_s\|\varphi\|^2,$$

which, for instance, implies also the dominance of balanced assignments in this weak form:

$$2\mathbb{E}_s\ \log\|\varphi\|_\theta\le\log\ \mathbb{E}_s\|\varphi\|_\theta^2=\log\ \mathbb{E}_s\|\varphi\|^2=2\mathbb{E}_s\ \log\|\varphi\|.$$

On the other hand, the Fourier expansion of φ_s is given by $\varphi_{s,Q}=-2^{-k}[\gamma_Q(s)+\gamma_Q(-s)]$ (for $Q\neq\varnothing$) and $\varphi_{s,Q}^{(i)}=-2^{-k}\gamma_Q(s)[1-(-1)^{|Q|}]$. In particular, $|\varphi_{s,Q}^{(i)}|=2^{-k}[1-(-1)^{|Q|}]$ so that both (2) and (3) hold along the same lines as the previous example, while the condition (b′) follows from the closure under *polarization* of this model. Indeed, in this case we get the same values for $\boldsymbol{\Omega}_k$, $\widetilde{\boldsymbol{\Omega}}_k$, and $\hat{\boldsymbol{\Omega}}_k$, so that we have the following:

	Reconstruction—clustering	Lower bound satisfiability	Upper bound satisfiability
NAE-SAT	$(2^{k-1}/k)[\log\ k+\mathrm{o}(\log\ k)]$	$2^{k-1}\ \log\ 2[1+\mathrm{o}(1)]$	$2^{k-1}\ \log\ 2[1+\mathrm{o}(1)]$

Example 3: k-XOR *formulas.* For an even integer k, the k-XOR ensemble (k even) consists of clauses of type $\{\varphi_\epsilon\}_{\epsilon\in\{+1,-1\}}$, where $\varphi_\epsilon=\frac{1}{2}(\gamma_\varnothing+\epsilon\gamma_{[k]})$. This set of clauses is endowed with the uniform probability distribution $p(\varphi_{+1})=p(\varphi_{-1})=1/2$. In this

case, the CSP(n, α, p) model corresponds to a system of αn random linear equations in $\mathbb{Z}_2$, in which every equation involves k randomly chosen variables (with replacement) from a total of n possible variables.

Conditions 1–3 hold for k even, and the dominance of the balanced assignments condition in its weak and strong form follows from the fact that $\mathbb{E}_\varphi \|\varphi\|_\theta^2 = \mathbb{E}_\varphi \|\varphi\|^2$. The condition on Fourier expansion of clauses for this model is straightforward: The Fourier expansion of φ_ϵ is concentrated at $\varnothing$ and $[k]$, so that (2) holds with $a = 0$ and (2) holds with $C = 1$. Also, condition (b′) follows from the following calculation:

$$\sum_{Q_1 \subseteq Q_2} \mathbb{E}_\varphi\{\varphi_{Q_1}\varphi_{Q_2}\}\theta^{|Q_1|}\delta^{|Q_2|-|Q_1|} = \frac{1}{4} + \frac{1}{4}\theta^k = \sum_Q \mathbb{E}_\varphi\{\varphi_Q^2\}\theta^{|Q|}.$$

In this case, we have that $\Omega_k = 1$, while $\hat{\Omega}_k = 1/\log 2$. Therefore, we have the following:

	Reconstruction—clustering	Lower bound satisfiability	Upper bound satisfiability
XOR-SAT	$\frac{1}{k}[\log k + \mathrm{o}(\log k)]$	$\log 2 + \mathrm{o}(1)$	$1 + \mathrm{o}(1)$

We remark here that, in the case of XOR-SAT, the clustering and satisfiability thresholds can be determined *exactly* by exploiting the underlying group structure [MRZ03], [CD+03] (see [MM09] for a discussion of the reconstruction problem in XOR-SAT).

4. Tree ensembles and tree reconstruction for binary k-CSP ensembles.

In this section we define tree ensembles and prove estimates about the corresponding tree reconstruction thresholds.

4.1. The tCSP(α, p) ensemble.

The ensemble tCSP(α, p) is defined by $\alpha \in \mathbb{R}_+$ and a distribution p over Boolean functions $\varphi : \{-1, +1\}^k \to \{0, 1\}$. We assume the conditions on the distribution p introduced in section 3.1. An (infinite) instance from this ensemble is generated starting by a root variable node ϕ, drawing an integer $\eta \stackrel{\mathcal{D}}{=} \mathrm{Poisson}(k\alpha)$ and connecting ϕ to η function nodes $\{1, \ldots, \eta\}$. Each function node has degree k, and each of its $k - 1$ descendants is the root of an independent infinite tree. Finally, each function node a is associated independently with a random clause φ drawn according to p.

A uniform solution for such an instance is sampled by drawing the root value $\mathbf{x}_\phi \in \{-1, +1\}$ uniformly at random. The values of descendants of each variable node i are then drawn recursively. If the function node a connects i to $i_1, \ldots, i_{k-1}$, then the values $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{k-1}}$ are sampled uniformly from those that satisfy the clause associated with a, that is, such that the quantity $\varphi(x_i, x_{i_1}, \ldots, x_{i_{k-1}})$ is equal to 1.

By the *balance* condition, this procedure can be shown to be equivalent to sampling a solution according to the "free boundary Gibbs measure." The latter is a distribution over solutions of the entire (infinite) tCSP formula defined by considering the uniform distribution over solutions of the first ℓ generations of the tree, and then letting $\ell \to \infty$.

We notice in passing that the above simplification does in fact hold under a weaker balance condition as well. Namely, it is sufficient that (for each $i \in \{1, \ldots k\}$) the number of truth assignments $(x_1, \ldots, x_k)$ that make $\varphi(x_1, \ldots, x_k) = 1$ and such that

$x_i = +1$ is equal to the number of assignments that make $\varphi(x_1, \ldots, x_k) = 1$ and such that $x_i = -1$.

4.2. Reconstruction. Given any fixed tree ensemble T, let $\mathbf{x}$ be a random satisfying assignment for T according to the distribution described previously. We denote by $\mathbf{x}_\ell$ the value of $\mathbf{x}$ at the variables at generation ℓ, and in the case that the root degree is 1, we denote by $\mathbf{x}_{0,1}, \ldots, \mathbf{x}_{0,k-1}$ the values at the variable nodes connected to the unique child of the root. Also, we use η_0 for the root degree of T. If the tree ensemble T has root degree $\eta_0 = d$, we denote by T_i, $i = 1, \ldots, d$, the subtree generated by the root, its ith child, and the child's descendants. If $\eta_0 = 1$, we denote by T'_i, $i = 1, \ldots, k-1$, the subtree generated by the ith child of the root's child and its descendants.

Finally, because the tree ensemble T could be random (for instance, we denote by $\mathbf{T}$ a random tCSP(α, p)), we will use $\mathbf{E}$ for expectation with respect to $\mathbf{T}$ and $\langle\cdot\rangle_T$ for expectation with respect to $\mathbf{x}$ (given $\mathbf{T} = T$) and $\mathbb{E}$ for expectation with respect to any other independent random variable (adding, if not in context, a subindex to indicate such random variable).

Reconstruction: For a fixed tree ensemble T, let $\mu_{\emptyset,\ell}$ be the joint distribution of $(\mathbf{x}_0, \mathbf{x}_\ell)$, and let $\mu_\emptyset$, μ_ℓ be the marginal distribution of $\mathbf{x}_0$ and $\mathbf{x}_\ell$, respectively. The reconstruction rate for T is defined as the quantity $\|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\mu_\ell(\cdot)\|_{\mathrm{TV}}$. We say that the reconstruction problem for T is *tree-solvable* if

$$\liminf_{\ell\to\infty} \|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\mu_\ell(\cdot)\|_{\mathrm{TV}} > 0.$$

Analogously, if $\mathbf{T}$ is a random tCSP(α, p), we define the reconstruction rate of $\mathbf{T}$ as

$$\mathbf{E}\|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\mu_\ell(\cdot)\|_{\mathrm{TV}},$$

and we say that the reconstruction problem for $\mathbf{T}$ is *tree-solvable*:

$$\liminf_{\ell\to\infty} \mathbf{E}\|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\mu_\ell(\cdot)\|_{\mathrm{TV}} > 0.$$

Bias, compatibility: Given a satisfying assignment x_ℓ for the variables at generation ℓ, define the "bias" of the root, restricted to the value of the variables at level ℓ, as

$$h_T(x_\ell) \stackrel{\mathrm{def}}{=} \langle \mathbf{x}_0 | \mathbf{x}_\ell = x_\ell \rangle_T.$$

Throughout the forthcoming proofs we will study $h_T(x_\ell)$ for random x_l, subject to different kinds of distributions. Notice that under the balance condition $\|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\mu_\ell(\cdot)\|_{\mathrm{TV}} = \frac{1}{2}\langle |h_T(\mathbf{x}_\ell)| \rangle_T$. In fact, it is the case that

$$|h_T(x_\ell)|\mu_\ell(x_\ell) = |\mu_{\emptyset,\ell}(1, x_\ell) - \mu_{\emptyset,\ell}(-1, x_\ell)| = 2\left|\mu_{\emptyset,\ell}(1, x_\ell) - \frac{1}{2}\mu_\ell(x_\ell)\right|,$$

and similarly,

$$|h_T(x_\ell)|\mu_\ell(x_\ell) = 2\left|\mu_{\emptyset,\ell}(-1, x_\ell) - \frac{1}{2}\mu_\ell(x_\ell)\right|.$$

By the balance condition, $\mu_{\varnothing}(1) = \mu_{\varnothing}(-1) = 1/2$. Therefore,

$$\begin{aligned}\langle |h_T(\mathbf{x}_\ell)| \rangle_T &= \sum_{x_\ell} (|\mu_{\varnothing,\ell}(1, x_\ell) - \mu_{\varnothing}(1)\mu_\ell(x_\ell)| + |\mu_{\varnothing,\ell}(-1, x_\ell) - \mu_{\varnothing}(-1)\mu_\ell(x_\ell)|) \\ &= 2\|\mu_{\varnothing,\ell}(\cdot,\cdot) - \mu_{\varnothing}(\cdot)\mu_\ell(\cdot)\|_{\mathrm{TV}}.\end{aligned}$$

Now, let $D_T(x_\ell) \stackrel{\text{def}}{=} \{x\}$ if $h_T(x_\ell) = x$ and $D_T(x_\ell) \stackrel{\text{def}}{=} \{-1, 1\}$ if $|h_T(x_\ell)| < 1$. Observe that $D_T(x_\ell)$ consists of the values of the root that are compatible with the assignment x_ℓ for the variables at generation l.

Domain of clauses: Given a binary function $\varphi(x_0, \ldots, x_{k-1})$, define the partial solution sets

$$\begin{aligned}S^+(\varphi) &\stackrel{\text{def}}{=} \{(x_1, , x_{k-1}) : \varphi(1, x_1, \ldots, x_{k-1}) = 1\}, \\ S^-(\varphi) &\stackrel{\text{def}}{=} \{(x_1, , x_{k-1}) : \varphi(-1, x_1, \ldots, x_{k-1}) = 1\}, \\ \Lambda^+(\varphi) &\stackrel{\text{def}}{=} S^+(\varphi) \setminus S^-(\varphi), \qquad \Lambda^-(\varphi) \stackrel{\text{def}}{=} S^-(\varphi) \setminus S^+(\varphi).\end{aligned}$$

If the clause φ is balanced and feasible, we have that $|S^+(\varphi)| = |S^-(\varphi)| = 2^{k-1}\|\varphi\|^2$ and $|\Lambda^+(\varphi)| = |\Lambda^-(\varphi)| = 2^k \mathrm{I}_1(\varphi)$.

THEOREM 4.1. *The reconstruction problem for the ensemble $tCSP(\alpha, p)$ is* tree-solvable *if and only if $\alpha > \alpha_{\mathrm{tree}}(k)$, where*

$$\alpha_{\mathrm{tree}}(k) = \frac{\Omega_k}{k}\{\log k + o(\log k)\}.$$

Proof. Upper bound: Given a tree ensemble T, the rate of "naive reconstruction" for T is defined as

$$z_\ell(T) \stackrel{\text{def}}{=} \langle \mathbb{I}[h_T(\mathbf{x}_\ell) = 1] \rangle_T \qquad (= \langle \mathbb{I}[h_T(\mathbf{x}_\ell) = -1] \rangle_T \text{ by the balance condition}),$$

which indicates the probability that a random assignment for the variables at generation ℓ, distributed as $\mathbf{x}_\ell$, fixes the root to be equal to 1 (or -1). We notice in passing that "naive reconstructibility" (i.e., the property that $z_\ell(T)$ does not vanish as $\ell \to \infty$) is likely to be related to the appearance of "frozen variables" in random CSPs (see, e.g., [AR11]). In particular, it is not hard to realize that the naive reconstruction threshold is a lower bound on the threshold for the appearance of $\Theta(n)$ frozen variables. It is natural to conjecture that the two thresholds do indeed coincide.

It is easy to see that $\langle |h_T(\mathbf{x}_\ell)| \rangle_T \geq z_\ell(T)$. Observe also that for any $x, y \in \{-1, 1\}$,

$$\langle \mathbb{I}[h_T(\mathbf{x}_\ell) = x] | \mathbf{x}_0 = y \rangle_T = 2 z_\ell(T) \delta_{x,y}. \tag{9}$$

Thus, our objective is to show that in an appropriate regime of the parameter α, the quantity $\mathbf{E}[z_\ell(\mathbf{T})]$ remains bounded away from zero as $\ell \to \infty$, implying tree-solvability of the reconstruction problem in such regime. Indeed, this implies tree-solvability by "naive reconstruction," i.e., by the procedure that assigns to the root any value compatible with the values at generation ℓ. By notational convenience, define

$$z_\ell(\alpha) = 2\mathbf{E}[z_\ell(\mathbf{T})] \quad \text{and} \quad \widehat{z_\ell}(\alpha) = 2\mathbf{E}[z_\ell(\mathbf{T}) | \eta_0 = 1].$$

Now, notice that for a tree ensemble T with root degree $\eta_0 = d$, and any assignment x_ℓ for the variables at generation ℓ, $h_T(x_\ell) = 1$ if and only if $h_{T_i}(x_\ell \restriction T_i) = 1$ for some $i = 1, \dots, d$, so that

$$\begin{aligned} 2z_\ell(T) &= \left\langle 1 - \prod_{i=1}^{d} (1 - \mathbb{I}[h_{T_i}(\mathbf{x}_\ell \restriction T_i) = 1]) | \mathbf{x}_0 = 1 \right\rangle_T \\ &= 1 - \prod_{i=1}^{d} \langle (1 - \mathbb{I}[h_{T_i}(\mathbf{x}_\ell) = 1]) | \mathbf{x}_0 = 1 \rangle_{T_i} \quad \text{(by the tree Markov property)} \\ &= 1 - \prod_{i=1}^{d} (1 - 2z_\ell(T_i)). \end{aligned}$$

Therefore, averaging over T, we get

$$\begin{aligned} z_\ell(\alpha) &= \mathbb{E}_\eta \left[1 - \prod_{i=1}^{\eta} (1 - \widehat{z_\ell}(\alpha)) \right], \qquad \eta \sim \text{Poisson}(k\alpha) \\ &= 1 - \exp\left(-k\alpha\, \widehat{z_\ell}(\alpha)\right) \end{aligned}$$

On the other hand, given a tree ensemble T with root degree $\eta_0 = 1$ and with the clause φ assigned to the root's child, we have that for any satisfying assignment x_ℓ for the variables at generation ℓ, $h_T(x_\ell) = 1$ if and only if

$$\prod_{i=1}^{k-1} D_{T'_i}(x^{(i)}_{\ell-1}) \subseteq \Lambda^+(\varphi), \tag{10}$$

where $x^{(i)}_{\ell-1}$ is the assignment $x_\ell \restriction T'_i$ for the variables at generation $\ell - 1$ in the subtree T'_i. Observe that (10) holds, in particular, if for some $a = (a_1, \dots, a_{k-1}) \in \Lambda^+(\varphi)$, $h_{T'_i}(x^{(i)}_{\ell-1}) = a_i$ for $i = 1, \dots, k-1$. Therefore, if $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_{k-1})$ denotes a random uniform vector from $S^+(\varphi)$, we have

$$\begin{aligned} z_\ell(T) &\geq \frac{1}{2} \sum_{a \in \Lambda^+(\varphi)} \left\langle \prod_{i=1}^{k-1} \mathbb{I}[h_{T'_i}(\mathbf{x}^{(i)}_{\ell-1}) = a_i] | \mathbf{x}_0 = 1 \right\rangle_T \\ &= \frac{1}{2} \sum_{a \in \Lambda^+(\varphi)} \mathbb{E}_{\mathbf{y}} \prod_{i=1}^{k-1} \langle \mathbb{I}[h_{T'_i}(\mathbf{x}_{\ell-1}) = a_i] | \mathbf{x}_0 = y_i \rangle_{T'_i} \quad \text{(by the tree Markov property)} \\ &= \frac{1}{2} \frac{|\Lambda^+(\varphi)|}{|S^+(\varphi)|} \prod_{i=1}^{k-1} 2z_{\ell-1}(T'_i) \quad \text{(by (9))}. \end{aligned}$$

This in turn implies, after averaging over T, that

$$\widehat{z_\ell}(\alpha) \geq \mathbb{E}_{\boldsymbol{\varphi}} \left[\frac{2\mathrm{I}_1(\varphi)}{\|\varphi\|^2} \right] (z_{\ell-1}(\alpha))^{k-1} = \frac{(z_{\ell-1}(\alpha))^{k-1}}{\Omega_k},$$

which leads to the recursion $z_\ell(\alpha) \geq 1 - \exp\left(-k\alpha (z_{\ell-1}(\alpha))^{k-1}/\Omega_k\right)$. Now, it is standard to verify that this recursion implies that $z_\ell(\alpha)$ is, for all ℓ, greater than or equal to the maximum of the fixed points of the function $g(z) = 1 - \exp\left(-k\alpha z^{k-1}/\Omega_k\right)$ in the interval $[0, 1]$. The minimum value of α for which such fixed point is positive is given by

$$\alpha^* = \frac{\Omega_k(1+u(1+\frac{1}{u})^{k-2})}{k(k-1)},$$

where u is the unique solution of the equation $u = (k-1)\log(1+u)$. In particular, asymptotically in k, we have that $\alpha^* = \frac{\Omega_k}{k}(\log k + \mathrm{o}(\log k))$, which implies the upper bound for α_{tree}.

Lower bound: The matching lower bound on $\alpha_{\mathrm{tree}}(k)$ requires a more elaborate proof; we first prove three lemmas before returning to complete the lower bound proof. ∎

Given a tree ensemble T, let $\mathbf{x}_\ell^+ \stackrel{\mathcal{D}}{=} (\mathbf{x}_\ell | \mathbf{x}_0 = 1)$ and $\mathbf{x}_\ell^- \stackrel{\mathcal{D}}{=} (\mathbf{x}_\ell | \mathbf{x}_0 = -1)$. When the tree ensemble is not clear in the definition of $\mathbf{x}_\ell^+$ (or $\mathbf{x}_\ell^-$), we add a subindex indicating the tree ensemble from where it is defined. Notice that, if μ^+ and μ^- are the distributions of $\mathbf{x}_\ell^+$ and $\mathbf{x}_\ell^-$, respectively, then

$$\frac{d\mu^-}{d\mu^+} = \frac{1-h_T(x_\ell)}{1+h_T(x_\ell)}. \tag{11}$$

By the balance condition, it is clear that

$$h_T(\mathbf{x}_\ell^+) \stackrel{\mathcal{D}}{=} -h_T(\mathbf{x}_\ell^-). \tag{12}$$

Also, it is easy to show that $\langle h_T(\mathbf{x}_\ell^+)\rangle_T = \langle [h_T(\mathbf{x}_\ell)]^2\rangle_T$ (and therefore $[R_l(T)]^2 \le \langle h_T(\mathbf{x}_\ell^+)\rangle_T \le R_l(T)$), so that nonreconstructibility for T is equivalent to the condition $\lim_{\ell\to\infty}\langle h_T(\mathbf{x}_\ell^+)\rangle_T = 0$ (see [MP03]). Similarly, if $\mathbf{T}$ is a random tCSP(α, p) ensemble, nonreconstructibility for $\mathbf{T}$ is equivalent to the condition $\lim_{\ell\to\infty}\mathbf{E}[\langle h_{\mathbf{T}}(\mathbf{x}_\ell^+)\rangle_{\mathbf{T}}] = 0$.

LEMMA 4.2.

(a) *Given a tree ensemble T with root degree $\eta_0 = d$, we have*

$$\left[\frac{1-h_T(\mathbf{x}_\ell^+)}{1+h_T(\mathbf{x}_\ell^+)}\right] \stackrel{\mathcal{D}}{=} \prod_{i=1}^{d}\left[\frac{1-h_{l,i}}{1+h_{l,i}}\right], \tag{13}$$

where $(h_{l,i})_{i=1}^d$ are independent random variables such that $h_{l,i} \stackrel{\mathcal{D}}{=} h_{T_i}(\mathbf{x}_\ell^+)$.

(b) *Given a tree ensemble T with root degree $\eta_0 = 1$ and with the clause φ assigned to the unique child of the root, we have that*

$$\left[\frac{1-h_T(\mathbf{x}_{\ell+1}^+)}{1+h_T(\mathbf{x}_{\ell+1}^+)}\right] \stackrel{\mathcal{D}}{=} \frac{T_{h_l}\varphi(-1,\mathbf{s})}{T_{h_l}\varphi(1,\mathbf{s})}, \tag{14}$$

where $\mathbf{s} \sim \mathrm{Unif}(S^+(\varphi))$ and $h_l = (h_{l,i})_{i=1}^{k-1}$ are independent random variables such that $h_{l,i} \stackrel{\mathcal{D}}{=} h_{T_i'}(\mathbf{x}_l^+)$.

Proof. This recursion follows straightforwardly from the recursive definition of tree formulae. The balance condition on clauses implies

$$\frac{1-h_T(\mathbf{x}_l^+)}{1+h_T(\mathbf{x}_l^+)} = \frac{\langle \mathbb{I}[\mathbf{x}_l = \mathbf{x}_l^+]|\mathbf{x}_0 = -1\rangle_T}{\langle \mathbb{I}[\mathbf{x}_l = \mathbf{x}_l^+]|\mathbf{x}_0 = 1\rangle_T}.$$

Therefore, if the root degree of T is $\eta_0 = d$, we have by the tree Markov property that

$$\frac{1-h_T(\mathbf{x}_l^+)}{1+h_T(\mathbf{x}_l^+)} = \prod_{i=1}^{d}\frac{\langle \mathbb{I}[\mathbf{x}_l = \mathbf{x}_l^+ \restriction T_i]|\mathbf{x}_0 = -1\rangle_{T_i}}{\langle \mathbb{I}[\mathbf{x}_l = \mathbf{x}_l^+ \restriction T_i]|\mathbf{x}_0 = 1\rangle_{T_i}},$$

and the last expression has the same distribution as $\prod_{i=1}^{d} \frac{1-u_{l,i}}{1+u_{l,i}}$, due to the fact that $(\mathbf{x}_l^+ \restriction T_i)_{i=1}^d$ are independent random assignments for the variables at generation l of T_i, such that $\mathbf{x}_l^+ \restriction T_i \stackrel{\mathcal{D}}{=} \mathbf{x}_{l,T_i}^+$. This proves (13). Now, if the root degree of T is $\eta_0 = 1$, define $(\tilde{\mathbf{x}}_{l,i}^+)_{i=1}^{k-1}$ to be independent random assignments for the variables at generation l of the subtrees T_i', such that $\tilde{\mathbf{x}}_{l,i}^+ \stackrel{\mathcal{D}}{=} \mathbf{x}_{l,T_i'}^+$. By the tree Markov property, we have that $(\mathbf{x}_{l+1}^+ \restriction T_i')_{i=1}^{k-1} \stackrel{\mathcal{D}}{=} (\mathbf{s}_i \tilde{\mathbf{x}}_{l,i}^+)_{i=1}^{k-1}$, where $\mathbf{s} \sim \mathrm{Unif} S^+(\varphi)$. Using the tree Markov property once more, we get

$$\begin{aligned}\left[\frac{1-h_T(\mathbf{x}_{\ell+1}^+)}{1+h_T(\mathbf{x}_{\ell+1}^+)}\right] &= \frac{\sum_y \varphi(-1,y) \prod_{i=1}^{k-1} \langle \mathbb{I}[\mathbf{x}_l = \mathbf{s}_i \tilde{\mathbf{x}}_{l,i}^+] | \mathbf{x}_0 = y_i \rangle_{T_i'}}{\sum_y \varphi(-1,y) \prod_{i=1}^{k-1} \langle \mathbb{I}[\mathbf{x}_l = \mathbf{s}_i \tilde{\mathbf{x}}_{l,i}^+] | \mathbf{x}_0 = y_i \rangle_{T_i'}} \\ &= \frac{\mathrm{T}_{h_l} \varphi(-1,\mathbf{s})}{\mathrm{T}_{h_l} \varphi(1,\mathbf{s})},\end{aligned}$$

which is precisely (14). ∎

The first step of the above recursion can be analyzed precisely, in terms of its distribution.

LEMMA 4.3. *If* $\mathbf{T}$ *is a random* $\mathrm{tCSP}(\alpha, p)$ *ensemble, then the random variable* $h_{\mathbf{T}}(\mathbf{x}_1^+)$ *takes values in* $\{0,1\}$ *and, if* $\alpha < (1-\delta)(\Omega_k \log k)/k$, *we have* $\mathbf{E} h_{\mathbf{T}}(\mathbf{x}_1^+) \le 1 - k^{-1+\delta}$.

Proof. If T is a tree ensemble with root degree $\eta_0 = 1$ and clause φ assigned to the root's child, from part (b) of Lemma 4.2, we have that $\frac{1-h_T(\mathbf{x}_1^+)}{1+h_T(\mathbf{x}_1^+)} \stackrel{\mathcal{D}}{=} \varphi(-1,\mathbf{s})$, where $\mathbf{s} \sim \mathrm{Unif}(S^+(\varphi))$. Recall that $h_{0,i} \equiv 1$. Therefore, it follows that $h_T(\mathbf{x}_1^+) = 1$ with probability $\frac{|\Lambda^+(\varphi)|}{|S^+(\varphi)|} = 1/\Omega_k$ and $h_T(\mathbf{x}_1^+) = 0$ otherwise. Similarly, if T is a tree ensemble with root degree $\eta_0 = d$, it follows from part (a) of Lemma 4.2 that $h_T(\mathbf{x}_1^+) = 1$ with probability $1 - (1 - 1/\Omega_k)^d$ and $h_T(\mathbf{x}_1^+) = 0$ otherwise. This implies then that $h_{\mathbf{T}}(\mathbf{x}_1^+)$ has support in $\{0,1\}$ and that $\mathbf{E} h_{\mathbf{T}}(\mathbf{x}_1^+) = 1 - \exp(-k\alpha(1 - 1/\Omega_k))$. The conclusion follows straightforwardly. ∎

For subsequent steps we track the averages, $h_\ell^{\mathrm{ave}} \stackrel{\mathrm{def}}{=} \mathbf{E}\langle h_{\mathbf{T}}(\mathbf{x}_l^+)\rangle_{\mathbf{T}}$ and $\hat{h}_\ell^{\mathrm{ave}} \stackrel{\mathrm{def}}{=} \mathbf{E}[\langle h_{\mathbf{T}}(\mathbf{x}_l^+)\rangle_{\mathbf{T}} | \eta_0 = 1]$, using the following bounds.

LEMMA 4.4. *For any* $\ell \ge 0$ *we have*

$$h_\ell^{\mathrm{ave}} \le 1 - e^{-2k\alpha \hat{h}_\ell^{\mathrm{ave}}}, \qquad \hat{h}_{\ell+1}^{\mathrm{ave}} \le \frac{1}{2} F_k(h_\ell^{\mathrm{ave}}) + \frac{1}{2} R_k(\sqrt{h_\ell^{\mathrm{ave}}}), \tag{15}$$

$$F_k(\theta) \stackrel{\mathrm{def}}{=} 2\mathbb{E}_\varphi\left[\frac{(\varphi^{(1)}, \mathrm{T}_\theta \varphi^{(1)})}{\|\varphi\|^2}\right], \quad R_k(\theta) \stackrel{\mathrm{def}}{=} 2\mathbb{E}_\varphi\left[\frac{2\mathrm{I}_1(\varphi)}{\|\varphi\|^2} \sum_{Q \subseteq [k-1]} |(\varphi^{(1)}, \gamma_Q)| \theta^{\max(|Q|,2)}\right]. \tag{16}$$

Finally, if h_ℓ *is supported on nonnegative values, then*

$$\hat{h}_\ell^{\mathrm{ave}} \le F_k(h_\ell^{\mathrm{ave}}). \tag{17}$$

Proof. We will say that a random variable $\mathbf{X} \in [-1,+1]$ is "consistent" if $\mathbf{E} f(-\mathbf{X}) = \mathbf{E}[(\frac{1-\mathbf{X}}{1+\mathbf{X}}) f(\mathbf{X})]$ for every function f such that the expectation values exist. A useful preliminary remark [MM06] is that the random variable $h_T(\mathbf{x}_l^+)$ is consistent (no matter the tree ensemble). In fact, this follows directly from (11) and (12):

$$\mathbf{E}f(-h_T(\mathbf{x}_l^+)) = \sum_{x_l} f(-h_T(x_l))\mu^+(x_l) = \sum_{x_l} f(-h_T(x_l))\frac{1+h_T(x_\ell)}{1-h_T(x_\ell)}\mu^-(x_l)$$
$$= \mathbf{E}\left[f(-h_T(\mathbf{x}_l^-))\frac{1+h_T(\mathbf{x}_l^-)}{1-h_T(\mathbf{x}_l^-)}\right] = \mathbf{E}\left[f(h_T(\mathbf{x}_l^+))\frac{1-h_T(\mathbf{x}_l^+)}{1+h_T(\mathbf{x}_l^+)}\right].$$

A number of properties of consistent random variables can be found in [RU08]. Let us now consider the first inequality. If T is a tree ensemble with root degree $\eta_0 = d$, then it is immediate from (13) that

$$\left\langle \left(\frac{1-h_T(\mathbf{x}_l^+)}{1+h_T(\mathbf{x}_l^+)}\right)^{1/2}\right\rangle_T = \prod_{i=1}^d \left\langle \left(\frac{1-h_{Ti}(\mathbf{x}_l^+)}{1+h_{T_i}(\mathbf{x}_l^+)}\right)^{1/2}\right\rangle_{Ti}. \tag{18}$$

It is possible to show that consistency implies that $\mathbf{E}X = \mathbf{E}X^2$ and $\mathbf{E}(\frac{1-X}{1+X})^{1/2} = \mathbf{E}\sqrt{1-X^2}$ (through the test functions $f(x) = x(1+x)$ and $f(x) = x(1+x)^{1/2}(1-x)^{-1/2}$); we thus have

$$\sqrt{1-\langle h_T(\mathbf{x}_l^+)\rangle_T} = \sqrt{1-\langle [h_T(\mathbf{x}_l^+)]^2\rangle_T}$$
$$\geq \left\langle \sqrt{1-[h_T(\mathbf{x}_l^+)]^2}\right\rangle_T \quad \text{(by Jensen's ineq.)}$$
$$= \left\langle \left(\frac{1-h_T(\mathbf{x}_l^+)}{1+h_T(\mathbf{x}_l^+)}\right)^{1/2}\right\rangle_T = \prod_{i=1}^d \left\langle \left(\frac{1-h_{T_i}(\mathbf{x}_l^+)}{1+h_{T_i}(\mathbf{x}_l^+)}\right)^{1/2}\right\rangle_{T_i}$$
$$= \prod_{i=1}^d \left\langle \sqrt{1-[h_{T_i}(\mathbf{x}_l^+)]^2}\right\rangle_{T_i}$$
$$\geq \prod_{i=1}^d (1-\langle h_{T_i}(\mathbf{x}_l^+)\rangle_{T_i}) \quad (\text{using} \sqrt{x} \geq x, \text{for } x \in [0,1]).$$

This implies, in particular, that if $\mathbf{T}$ is a random tCSP(α, p), then

$$\sqrt{1-\mathbf{E}\langle h_{\mathbf{T}}(\mathbf{x}_l^+)\rangle_{\mathbf{T}}} \geq \mathbb{E}_\eta\left[\prod_{i=1}^\eta (1-\mathbf{E}[\langle h_{\mathbf{T}}(\mathbf{x}_l^+)\rangle_{\mathbf{T}}|\eta_0 = 1])\right], \qquad \eta \sim \text{Poisson}(k\alpha),$$

whence the first inequality follows.

Now, from the recursion equation (14), we have for a tree ensemble T with root degree $\eta_0 = 1$ and random clause φ assigned to the child of the root,

$$h_T(\mathbf{x}_{l+1}^+) = \frac{2\mathrm{T}_{h_l}\varphi^{(1)}(\mathbf{s})}{1+\mathrm{T}_{h_l}\psi(\mathbf{s})}, \qquad \psi(s) \stackrel{\text{def}}{=} \varphi(1,s)\varphi(-1,s),$$

or alternatively

$$h_T(\mathbf{x}_{l+1}^+) = \mathrm{T}_{h_1}\varphi^{(1)}(\mathrm{s}) + (\mathrm{T}_{h_1}\varphi^{(1)}(\mathrm{s}))\mathcal{G}_k(h_l, \mathbf{s}), \qquad \mathcal{G}_k(h_l, s) \stackrel{\text{def}}{=} \left[\frac{1-\mathrm{T}_{h_l}\psi(s)}{1+\mathrm{T}_{h_l}\psi(s)}\right],$$

where $\mathbf{s} \sim \text{Unif}\, S^+(\varphi)$. Notice that for any antisymmetric function $f(s)$, we have that $\mathbb{E}_{\mathbf{s}} f(\mathbf{s}) = \frac{(\varphi^{(1)}, f)}{\|\varphi\|^2}$. Therefore, due to the fact that $\mathrm{T}_{h_1}\varphi^{(1)}(s)$ is antisymmetric and $\mathcal{G}_k(h_l, s)$ is symmetric (both in s and h_l, actually), we have the formulas

$$\langle h_T(\mathbf{x}_{l+1}^+)\rangle_T = \frac{2}{\|\varphi\|^2}\left\langle \left(\varphi^{(1)}, \frac{\mathrm{T}_{h_l}\varphi^{(1)}(\mathbf{s})}{1+\mathrm{T}_{h_l}\psi(\mathbf{s})}\right)\right\rangle_T \tag{19}$$

and

$$\langle h_T(\mathbf{x}_{l+1}^+)\rangle_T = \left\langle \frac{(\varphi^{(1)}, \mathrm{T}_{h_l}\varphi^{(1)})}{\|\varphi\|^2}\right\rangle_T + \left\langle \frac{(\varphi^{(1)}, (\mathrm{T}_{h_l}\varphi^{(1)})\mathcal{G}_k(h_l,\cdot))}{\|\varphi\|^2}\right\rangle_T. \tag{20}$$

In the last expression, the first term is equal to $\frac{(\varphi^{(1)}, T_{\langle h_l\rangle_T}\varphi^{(1)})}{\|\varphi\|^2}$, while the second term can be written, using Fourier expansion, as

$$\frac{1}{\|\varphi\|^2}\sum_{\substack{Q\subseteq[k-1]\\ |Q| \text{ odd}}} (\varphi^{(1)}, \gamma_Q \mathbb{E}_{h_l}[\gamma_Q(h_l)\mathcal{G}_k(h_l,\cdot)])(\varphi^{(1)}, \gamma_Q).$$

Using the fact that $\mathbb{E}|\mathbf{X}| \le (\mathbb{E}\mathbf{X})^{1/2}$ for consistent random variables, we can bound the terms with $|Q| \ge 3$ by

$$\frac{|(\varphi^{(1)},1)|}{\|\varphi\|^2}\sum_{\substack{Q\subseteq[k-1]\\ |Q|\ge 3 \text{ odd}}} |(\varphi^{(1)}, \gamma_Q)|\left(\prod\nolimits_{i\in Q}\langle h_{T_i}(\mathbf{x}_l^+)\rangle_{T_i}\right)^{1/2}.$$

Also, using the fact that for any even function $f(x)$ with $0 \le f(x) \le 1$ and a consistent random variable $\mathbf{X}$, we have

$$|\mathbb{E}[\mathbf{X}f(\mathbf{X})]| = |\mathbb{E}[2\mathbf{X}^2 f(\mathbf{X})/(1+\mathbf{X})\mathbb{I}_{\{\mathbf{X}\ge 0\}}]| \le |\mathbb{E}[2\mathbf{X}^2/(1+\mathbf{X})\mathbb{I}_{\{\mathbf{X}\ge 0\}}]| = |\mathbb{E}[\mathbf{X}]|;$$

we can bound the terms with $|Q| = 1$ by

$$\frac{|(\varphi^{(1)},1)|}{\|\varphi\|^2}\sum_{i=1}^{k-1} (\varphi^{(1)}, \gamma_{\{i\}})|\langle h_{T_i}(\mathbf{x}_l^+)\rangle_{T_i}|.$$

Therefore, for a random tCSP(α, p) with root degree $\eta_0 = 1$, we obtain after averaging

$$\hat{h}_{l+1}^{\text{ave}} \le \mathbb{E}_\varphi \frac{(\varphi^{(1)}, \mathrm{T}_{h_l^{\text{ave}}}\varphi^{(1)})}{\|\varphi\|^2} + \mathbb{E}_\varphi\left[\frac{2\mathrm{I}_1(\varphi)}{\|\varphi\|^2}\sum_{\substack{Q\subseteq[k-1]\\ |Q|\ge 3 \text{ odd}}} |(\varphi^{(1)}, \gamma_Q)|(\sqrt{h_l^{\text{ave}}})^{\max\{|Q|,2\}}\right],$$

which is precisely the second inequality in the lemma.

Now, suppose that h_l is supported on nonnegative values, and let $A_s = \{h_l : \mathrm{T}_{h_l}\varphi^{(1)}(s) > 0\}$. Notice that the complement of A_s is $-A_s$ (due to the antisymmetry of $\mathrm{T}_{h_l}\varphi^{(1)}(s)$ with respect to h_l). Therefore, using the consistency of the random variables $h_{l,i}$, from (19) we get

$$\begin{aligned}\langle h_T(\mathbf{x}_{l+1}^+)\rangle_T &= \frac{2}{\|\varphi\|^2}\left\langle \left(\varphi^{(1)}, \frac{\mathrm{T}_{h_1}\varphi^{(1)}(\mathbf{s})}{1+\mathrm{T}_{h_l}\psi(\mathbf{s})}\right)\mathbb{I}(h_l\in A_{\mathbf{s}})\right.\\ &\quad \left. - \left(\varphi^{(1)}, \frac{\mathrm{T}_{-h_1}\varphi^{(1)}(\mathbf{s})}{1+\mathrm{T}_{-h_l}\psi(\mathbf{s})}\right)\mathbb{I}(-h_l\in A_{\mathbf{s}})\right\rangle_T\\ &= \frac{2}{\|\varphi\|^2}\left\langle \left(\varphi^{(1)}, \frac{\mathrm{T}_{h_1}\varphi^{(1)}(\mathbf{s})}{1+\mathrm{T}_{h_l}\psi(\mathbf{s})}\right)\mathbb{I}(h_l\in A_{\mathbf{s}})\left[1-\prod_{i=1}^{k-1}\frac{1-h_{l,i}}{1+h_{l,i}}\right]\right\rangle_T\\ &\le \frac{2}{\|\varphi\|^2}\left\langle (\varphi^{(1)}, \mathrm{T}_{h_1}\varphi^{(1)}(\mathbf{s}))\mathbb{I}(h_l\in A_{\mathbf{s}})\left[1-\prod_{i=1}^{k-1}\frac{1-h_{l,i}}{1+h_{l,i}}\right]\right\rangle_T\\ &= \frac{2(\varphi^{(1)}, \mathrm{T}_{\langle h_1\rangle}\varphi^{(1)}(\mathbf{s}))}{\|\varphi\|^2}.\end{aligned}$$

Therefore, for a random tCSP(α, p) with root degree $\eta_0 = 1$, we obtain after averaging that

$$\hat{h}_{l+1}^{\mathrm{ave}} \le 2\mathbb{E}_\varphi \frac{(\varphi^{(1)}, \mathrm{T}_{h_l^{\mathrm{ave}}}\varphi^{(1)})}{\|\varphi\|^2},$$

which corresponds to the last inequality of the lemma. ∎

We now return to completing the proof of Theorem 4.1.

Proof of Theorem 4.1*, lower bound.* If $\theta = 1$, T_1 is the identity operator whence $(\varphi^{(1)}, \mathrm{T}_1\varphi^{(1)}) = \mathrm{I}_1(\varphi)$. We have therefore $F_k(1) = 1/\boldsymbol{\Omega}_k$. Now, expanding in Fourier series we get

$$(\varphi^{(1)}, \mathrm{T}_1\varphi^{(1)}) = \sum_{Q\subseteq[k-1]} |(\varphi^{(1)}, \gamma_Q)|^2\, \theta^{|Q|} = \sum_{Q\subseteq[k], Q\ni\{i\}} |(\varphi^{(1)}, \gamma_Q)|^2\, \theta^{|Q|-1}.$$

By the *Fourier expansion condition*,

$$F_k(\theta) \le e^{-Ck(1-\theta)}/\boldsymbol{\Omega}_k. \tag{21}$$

Now fix $\alpha = (1-\delta)(\boldsymbol{\Omega}_k \log k)/k$, whence, by Lemma 4.3, $h_1^{\mathrm{ave}} \le 1 - k^{-1+\delta}$, and h_1 is supported on nonnegative reals. Using (17), we get $\hat{h}_2^{\mathrm{av}} \le e^{-Ck^\delta}/\boldsymbol{\Omega}_k$, and therefore,

$$h_2^{\mathrm{av}} \le 1 - \exp\{-2(1-\delta)e^{-Ck^\delta}\log k\} \le e^{-Ck^\delta/2}.$$

On the other hand, from (7), we obtain the following bounds for $F_k(\theta)$, $R_k(\theta)$:

$$F_k(\theta) \le 2\mathbb{E}_\varphi\left[\frac{\sum_{i=1}^{k-1}|(\varphi^{(1)}, \gamma_{\{i\}})|^2}{\|\varphi\|^2}\right]\theta + 2\mathbb{E}_\varphi\left[\frac{\mathrm{I}_1(\varphi)}{\|\varphi\|^2}\right]\theta^2 \le \frac{Ae^{-Ck/2}\theta + \theta^2}{\boldsymbol{\Omega}_k},$$

$$\begin{aligned}R_k(\theta) &\le 2\mathbb{E}_{\varphi\mathrm{i}}\left[\frac{2\mathrm{I}_1(\varphi)}{\|\varphi\|^2}\sum_{i=1}^{k-1}|(\varphi^{(1)}, \gamma_{\{i\}})|^2\right]\theta^2 + 2\mathbb{E}_\varphi\left[\frac{2\mathrm{I}_1(\varphi)}{\|\varphi\|^2}\sum_{Q\subseteq[k-1]}|(\varphi^{(1)}, \gamma_Q)|\right]\theta^3\\ &\le \frac{Ae^{-Ck/2}\theta^2 + k^a\theta^3}{\boldsymbol{\Omega}_k}.\end{aligned}$$

Therefore, for all ℓ we have

$$h^{\rm av}_{\ell+1} \le 1 - e^{-k\alpha[F_k(h^{\rm av}_\ell)+R_k(h^{\rm av}_\ell)]} \le (1-\delta)\log\, k(2Ae^{-Ck/2}h^{\rm av}_\ell + 2k^a(h^{\rm av}_\ell)^{3/2}),$$

which implies $h^{\rm av}_\ell \to 0$ if, for some $\ell > 0$, $h^{\rm av}_\ell \le k^{-5a}$, thus finishing the proof. ∎

5. Reconstruction on trees to graphs: The case of proper q colorings. In this section we prove that the set of solutions of the proper q-coloring ensemble satisfies the *sphericity* condition described in section 3.3. Recall that this in turn implies the equivalence of (sparse random) graph reconstruction and tree reconstruction for the proper q-coloring model.

Given two assignments (as in two proper colorings) $\underline{x}^{(1)}$, $\underline{x}^{(2)}$ of the variables $x_1, \dots, x_n$, their joint type $v_{\underline{x}^{(1)},\underline{x}^{(2)}}$ is the $q \times q$ matrix with $v_{\underline{x}^{(1)},\underline{x}^{(2)}}(i,j) \stackrel{\rm def}{=} \frac{1}{n}\#\{t \in G\colon \underline{x}^{(1)}(t) = i \text{ and } \underline{x}^{(2)}(t) = j\}$. We consider random assignments $\underline{\mathbf{x}}^{(1)}$, $\underline{\mathbf{x}}^{(2)}$ taken uniformly and independently over all the satisfying assignments of a random instance of the q-coloring model with edge-variable density α. Our purpose is to prove that for all $\epsilon > 0$, $\|v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} - \bar{v}\|_{\rm TV} \le \epsilon$ with high probability, where $\bar{v}$ is the matrix with all entries equal to $1/q^2$. More exactly, we have the following.

THEOREM 5.1. *Let $\underline{\mathbf{x}}^{(1)}$, $\underline{\mathbf{x}}^{(2)}$ be random assignments taken uniformly and independently over all satisfying assignments of a random instance of the q-coloring model with edge-variable density α. If $\alpha < (q-1)\log(q-1)$, then for any $\epsilon > 0$,*

$$\text{Prob}(\|v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} - \bar{v}\|^2 > \epsilon) \to 0 \quad \textit{as } n \to \infty.$$

The statistic $v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}}$ samples the correlation between the colors of two random vertices of the graph. The main result in [GM07] was that concentration of this statistic implies equivalence of tree and random graph reconstruction (in the diluted regime).

At this point we should recall the so-called *transfer theorem* introduced in [AC08], which says that with the edge-variable density $\alpha < q\log q$, the set of events that hold with high probability at exponential rate in the planted model hold with high probability in the uniform model as well; the planted model here is induced by choosing a uniform random q-partition of the vertices and then a graph with m edges chosen uniformly at random from among the edges properly colored (as in nonmonochromatic) by the partition. In particular, the transfer theorem implies that most of the colorings of a random graph (at edge-variable density $\alpha < q\log q$) are "balanced" in the sense that, for any $\epsilon > 0$,

$$\text{Prob}(\|w_{\underline{\mathbf{x}}} - \bar{w}\|^2 > \epsilon) \to 0 \quad \text{as } n \to \infty, \tag{22}$$

where w is the vector with q entries such that $w_{\underline{\mathbf{x}}}(i) = \frac{1}{n}\#\{v \in G\colon \underline{\mathbf{x}}_v = i\}$ and $\bar{w}$ is the vector with all entries equal to $1/q$. Notice that a similar transfer theorem for *pairs* of colorings would imply the result stated in Theorem 5.1. Although we believe that such a transfer holds in the appropriate regime, rather than proving it in full, we prove instead just the conclusion that we need in Theorem 5.1. Our argument makes crucial use of the following estimate for the partition function, also from [AC08].

LEMMA 5.2. *Let Z be the number of satisfying assignments of a random instance of the q-coloring model with edge-variable density $\alpha < q\log q$. Then, for some function $f(n)$ of order $o(n)$, we have*

$$\text{Prob}(Z < e^{-f(n)}\mathbf{E}[Z]) \to 0 \quad \textit{as } n \to \infty.$$

We should point out also the following estimate for the expected value:

$$\mathbf{E}\, Z \geq \Omega\left(\frac{1}{n^{(q-1)/2}}\right)\left[q\left(1-\frac{1}{q}\right)^{\alpha}\right]^{n}. \tag{23}$$

Let us fix some notation first. If v is a $q \times q$ matrix, let $\mathcal{H}$ and $\mathcal{E}$ denote their *entropy* and *energy*, respectively, where

$$\mathcal{H}(v) = -\sum_{i,j} v(i,j) \log\, v(i,j),$$

$$\mathcal{E}(v) = \log\left(1 - \sum_i \left(\sum_j v(i,j)\right)^2 - \sum_j \left(\sum_i v(i,j)\right)^2 + \sum_{i,j} v(i,j)^2\right).$$

Also, given ϵ, $\delta > 0$, let $S(\delta, \epsilon)$ denote the set of all $q \times q$ matrices v with nonnegative entries such that

$$\|(v - \bar{v})\mathbb{1}\|^2 \leq \delta, \quad \|\mathbb{1}^T(v - \bar{v})\|^2 \leq \delta \quad \text{and} \quad (v - \bar{v})\|^2 \geq \epsilon,$$

where $\mathbb{1}$ is the $q \times 1$ vector of all 1's. Before returning to the proof of Theorem 5.1, we introduce estimates concerning an additive functional depending on the energy and entropy of matrices in $S(\delta, \epsilon)$; for this purpose, we define $\kappa(\delta, \epsilon)$ as the upper limit of the interval (indeed, it is easy to see that this is an interval) consisting of the values c such that

$$\sup_{v \in S(\delta,\epsilon)} \mathcal{H}(v) + c\mathcal{E}(v) \leq \mathcal{H}(\bar{v}) + \alpha\mathcal{E}(\bar{v}).$$

To motivate, let us recall that an important part of the second moment argument of Achlioptas and Naor [AN05, Theorem 7] (in showing that the chromatic number $\chi[G(n, d/n)]$ concentrated on two possible values) relied on an optimization of the expression $\mathcal{H}(v) + \alpha\mathcal{E}(v)$ over the Birkoff polytope $\mathcal{B}_{q\times q}$ of the $q \times q$ doubly stochastic matrices. In particular, they proved that, as long as $\alpha \leq (q-1)\log(q-1)$, one has

$$\sup_{v \in \mathcal{B}_{q\times q}} \mathcal{H}(v) + \alpha\mathcal{E}(v) = \mathcal{H}(\bar{v}) + \alpha\mathcal{E}(\bar{v}). \tag{24}$$

In particular, since $S(0, \epsilon) \subseteq \mathcal{B}_{q\times q}$, we have $\kappa(0, \epsilon) \geq \alpha_q = (q-1)\log(q-1)$. This implies also, due to the continuity of $\kappa(\delta, \epsilon)$, that whenever $\alpha < \alpha_q$, for every $\epsilon > 0$ there is some $\delta > 0$ such that $\kappa(\delta, \epsilon) > \alpha$.

LEMMA 5.3. *Suppose that $v \in S(\delta, \epsilon)$, where $\epsilon > 2\delta$; then if $\kappa(\delta, \epsilon) > \alpha$, we have that*

$$[\mathcal{H}(v) + \alpha\mathcal{E}(v)] \leq [\mathcal{H}(\bar{v}) + \alpha\mathcal{E}(\bar{v})] - \frac{(\kappa(\delta,\epsilon) - \alpha)(\epsilon - 2\delta)}{2(1 - 1/q)^2}.$$

Proof. Indeed,

$$\begin{aligned}
&[\mathcal{H}(\bar{v}) + \alpha\mathcal{E}(\bar{v})] - [\mathcal{H}(v) + \alpha\mathcal{E}(v)] \\
&\quad = [\mathcal{H}(\bar{v}) + \kappa(\delta,\epsilon)\mathcal{E}(\bar{v})] - [\mathcal{H}(v) + \kappa(\delta,\epsilon)\mathcal{E}(v)] + (\kappa(\delta,\epsilon) - \alpha)[\mathcal{E}(v) - \mathcal{E}(\bar{v})] \\
&\quad \geq (\kappa(\delta,\epsilon) - \alpha)\left[\log\left(1 + \frac{1}{(1-1/q)^2}[\|v - \bar{v}\|^2 - \|v - \bar{v}\|^2 - \|1^t(v - \bar{v})\|^2]\right)\right] \\
&\quad \geq \frac{(\kappa(\delta,\epsilon) - \alpha)(\epsilon - 2\delta)}{2(1-1/q)^2}. \qquad \square
\end{aligned}$$

Proof of Theorem 5.1. Given a property P, denote by $Z^{(2)}(P)$ the number of pairs of satisfying assignments for which P holds. Now, choose $\delta < \epsilon/2$, such that $\kappa(\delta,\epsilon) > \alpha$ (see the comment previous to Lemma 5.3), and let $\xi = \frac{(\kappa(\delta,\epsilon)-\alpha)(\epsilon-2\delta)}{2(1-1/q)^2}$. We have that

$$\mathrm{Prob}(\|v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} - \bar{v}\|^2 > \epsilon) = \mathbf{E}\left[\frac{Z^{(2)}(\|v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} - \bar{v}\|^2 > \epsilon)}{Z^2}\right].$$

Now, according to Lemma 5.2 and Eq. (22), the events $Z < e^{-n\xi}\mathbf{E}[Z]$, $\|(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} - \bar{v})1\|^2 > \epsilon$, and $\|1^t(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} - \bar{v})\|^2 > \epsilon$ are negligible. Therefore, to show that $\mathrm{Prob}(\|v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} - \bar{v}\|^2 > \epsilon) \to 0$ is sufficient to prove that the term

$$\frac{\mathbf{E}[Z^{(2)}(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} \in \mathcal{B}^{\delta,\epsilon}_{q\times q})]}{e^{-2n\xi}\mathbf{E}[Z]^2}$$

vanishes. Now, consider the set $\mathcal{G}_{\epsilon,\delta}$ of $q \times q$ matrices L, with nonnegative integer entries, such that $L/n \in S(\delta,\epsilon)$, and denote by $\mathbf{\Omega}_v$ the set of pairs of colorings x_1, x_2 such that v_{x_1,x_2} is equal to the matrix v; then

$$\begin{aligned}
&\mathbf{E}[Z^{(2)}(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} \in \mathcal{B}^{\delta,\epsilon}_{q\times q})]\\
&= \sum_{L\in\mathcal{G}_{\epsilon,\delta}} \sum_{x_1,x_2\in\mathbf{\Omega}_{L/n}} \mathrm{Prob}(x_1 \text{ and } x_2 \text{ are satisfying assignments})\\
&= \sum_{L\in\mathcal{G}_e} \frac{n!}{\prod_{i,j} L_{ij}!}\left[\frac{n}{n-1}\right]^{\alpha n}\left(1 - \sum_i\left(\sum_j L_{ij}/n\right)^2 - \sum_j\left(\sum_i L_{ij}/n\right)^2 + \sum_{i,j}(L_{ij}/n)^2\right)^{\alpha n}\\
&\le \sum_{L\in\mathcal{G}_{\epsilon,\delta}} 3q^{2q}\sqrt{n}\,\exp\left(n[\mathcal{H}(L/n) + \alpha\mathcal{E}(L/n)]\right).
\end{aligned}$$

Now, we can invoke Lemma 5.3 to get that

$$[\mathcal{H}(L/n) + \alpha\mathcal{E}(L/n)] \le [\mathcal{H}(\bar{v}) + \alpha\mathcal{E}(\bar{v})] - \xi.$$

Therefore,

$$\mathbf{E}[Z^{(2)}(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} \in \mathcal{B}^{\delta,\epsilon}_{q\times q})] \le \mathrm{poly}(n) \times [q(1-1/q)^{\alpha}]^{2n}\,\exp(-n\xi),$$

so by applying (23) we get that

$$\frac{\mathbf{E}[Z^{(2)}(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} \in \mathcal{B}^{\delta,\epsilon}_{q\times q})]}{e^{-2n\xi}\mathbf{E}[Z]^2} \le \mathrm{poly}(n) \times \exp(-n\xi).$$

The result follows. ∎

Appendix A. Proof of Proposition 3.2. Given a random instance from the ensemble $\mathrm{CSP}(n,p,\alpha)$, let $\{\varphi_a\}_{a=1}^{\alpha n}$ be its set of clauses and consider the symmetrized statistic

$$L_n(\varphi) = \frac{1}{n\alpha\, k!}\sum_{\sigma\in S_k} \#\{\, a \in [n\alpha] : \varphi_a = \varphi^\sigma\}. \tag{25}$$

It is convenient to introduce two slightly modified ensembles. We denote by $\mathrm{CSP}(n,p,\alpha;\tilde{p}_n)$ the ensemble $\mathrm{CSP}(n,p,\alpha)$ conditioned on $L_n=\tilde{p}_n$.

A binary configuration $\underline{x}$ is said to be balanced if $|\underline{x}\cdot\underline{1}|\le 1$. We will use Z and Z_b to denote the variable that counts the number of satisfying assignments and balanced satisfying assignments, respectively, of a random CSP ensemble. Given two binary assignments $\underline{x}^{(1)},\underline{x}^{(2)}$, we define their overlap as

$$Q_{12}\stackrel{\text{def}}{=}\frac{1}{n}\underline{x}^{(1)}\cdot\underline{x}^{(2)}=\frac{1}{n}\sum_{i=1}^{n}x_i^{(1)}x_i^{(2)}. \tag{26}$$

In other words, $(1-Q_{12})/2$ is the normalized Hamming distance of $\underline{x}^{(1)}$ and $\underline{x}^{(2)}$.

Proof of Proposition 3.2*, upper bound.* The upper bound in Proposition 3.2 follows from a first moment calculation. Let Z be the number of solutions of a random instance from the ensemble $\overline{\mathrm{CSP}}(n,p,\alpha)$. We will show that, for $\alpha>(1+\epsilon)\hat{\Omega}_k\log 2$, $\mathbf{E}[Z]\to 0$ as $n\to\infty$. First fix $\tilde{p}_n$ such that $\|\tilde{p}_n-p\|_{\mathrm{TV}}\le 1/n^{1/2-\gamma}$. Notice that the probability that a random clause of type φ is satisfied by the assignment x with $x\cdot 1=n\theta$ is $\|\varphi\|_\theta^2$. This implies

$$\begin{aligned}
\mathbf{E}[Z|L_n=\tilde{p}_n]&=\sum_{x\in\{-1,1\}^n}\mathbf{P}\left(x\text{ is a satisfying assignment}|L_n=\tilde{p}_n\right)\\
&\le n\sup_{\theta\in[-1,1]}\sum_{x\cdot 1=n\theta}\mathbf{P}\left(x\text{ is a satisfying assignment}|L_n=\tilde{p}_n\right)\\
&\le n2^n\prod_{\varphi}\|\varphi\|_\theta^{2\tilde{p}_n(\varphi)\alpha n}\\
&\le n\ \exp\left(n\left\{\log 2+\alpha\sum_{\varphi}p(\varphi)\log\|\varphi\|_\theta^2+O(n^{-1/2+\gamma})\right\}\right)\\
&\le n\ \exp\left(n\left\{\log 2+\alpha\sum_{\varphi}p(\varphi)\log\|\varphi\|^2+O(n^{-1/2+\gamma})\right\}\right),
\end{aligned}$$

where in the last step we used the condition of dominance of balanced assignments. By taking expectation over $\tilde{p}_n$, we obtain $\mathbf{E}[Z]\to 0$ whenever $\alpha>(1+\epsilon)\hat{\Omega}_k\log 2$, as claimed. ∎

To establish the corresponding lower bound, we use the second moment method, but first we need a few preliminary lemmas.

We define by $\mathcal{K}_n(p;a,A,\gamma)$ to be the set of probability distributions $\{\tilde{p}(\varphi)\}$ over clauses $\varphi:\{+1,-1\}\to\{0,1\}$ such that

(i) $\mathrm{supp}(\tilde{p})=\mathrm{supp}(p)$;
(ii) $\tilde{p}$ satisfies conditions 1–4 and (a), (b) stated in section 3 with constants a, A; and finally,
(iii) $\|\tilde{p}_n-p\|_{\mathrm{TV}}\le 1/n^{1/2-\gamma}$ for some $\gamma>0$. Then we have the following.

Lemma A.1. *Let L_n be the statistics defined in* (25) *for a random formula from the* $\mathrm{CSP}(n,p,\alpha)$ *ensemble. Then there exists constants a, A such that for any $\gamma>0$, with high probability*

$$L_n\in\mathcal{K}_n(p;a,A,\gamma). \tag{27}$$

Proof. Notice that for each permutation $\pi L_n(\varphi^\pi) = L_n(\varphi)$ and that, for each $\varphi\{-1,+1\}^k \to \{0,1\}$, $k!L_n(\varphi)$ is distributed as a binomial with parameters $n\alpha$, and $k!p(\varphi)$. In particular, $L_n(\varphi) = 0$ if $p(\varphi) = 0$ and $L_n(\varphi) > 0$ with high probability otherwise. This implies item (i) in the definition of $\mathcal{K}_n(p; a, A, \gamma)$.

Item (iii), that $\|L_n - p\|_{\mathrm{TV}} \le 1/n^{1/2-\gamma}$, follows immediately from the central limit theorem.

Consider finally item (ii). Condition 1 is enforced by the symmetrization procedure in (25). Conditions 2 and 3 depend only on $\mathrm{supp}(L_n)$ and thus hold with high probability by the above argument.

Dominance of balanced assignments (condition 4) is the statement that

$$\mathbb{E}_\varphi \log\|\varphi\|_\theta - \mathbb{E}_\varphi \log\|\varphi\| < 0 \tag{28}$$

for all $\theta \neq 0$, $\theta \in [-1,1]$. Notice that the left-hand side is a polynomial in θ whose coefficients are continuous function of the quantities $\{L_n(\varphi)\}$. Hence this condition is of the form $L_n \in \mathcal{A}$ for $\mathcal{A}$, an open set in $\mathbb{R}^D$, $D = 2^{2^k}$. Since $p \in \mathcal{A}$ and $\|L_n - p\|_{\mathrm{TV}} \le n^{-1/2+\gamma}$ with high probability, we conclude $L_n \in \mathcal{A}$.

Finally conditions (a) and (b) depend only on $\mathrm{supp}(L_n)$ and therefore follow from the above. ∎

LEMMA A.2. *Given $\tilde{p}_n \in \mathcal{K}_n(p; a, A, \gamma)$, consider a random instance from the* $\mathrm{CSP}(n, p, \alpha; \tilde{p}_n)$ *ensemble. For $\theta \in \{-1, -1+2/n, \dots, 1-2/n, 1\}$, let $Z_{\mathrm{b}}(Q_{12} = \theta)$ be the number of balanced solution pairs $\underline{x}^{(1)}, \underline{x}^{(2)} \in \{+1,-1\}^n$ with overlap θ. Then,*

$$\frac{\mathbf{E}[Z_{\mathrm{b}}(Q_{12} = \theta)]}{[\mathbf{E}Z_{\mathrm{b}}]^2} \le Cn^{-1/2} \exp\{n\Phi(\theta)\},$$

where C is bounded uniformly in θ and

$$\Phi(\theta) \stackrel{\mathrm{def}}{=} H(\theta) + \alpha\mathbb{E}_{\varphi\sim\tilde{p}_n} \log\left\{\frac{(\varphi, T_\theta\varphi)}{\|\varphi\|^4}\right\}.$$

Here $H(\theta) \equiv -\frac{1+\theta}{2}\log(1+\theta) - \frac{1-\theta}{2}\log(1-\theta)$ is the binary entropy function.

Proof. For simplicity take n to be even (the argument is analogous for n odd). Let φ be a Boolean function, and let $i\colon[k] \to [n]$ be a uniform random choice of the indexes of the variables in φ (i.e., $i(1), \dots, i(k)$ are independent and uniform in $[n]$). Given two *balanced* vectors $\underline{x}^{(1)}, \underline{x}^{(2)} \in \{+1,-1\}^n$, with $Q_{12} = \theta$, we have

$$\mathbb{E}_\pi[\varphi(x^{(1)}_{i(1)}, \dots, x^{(1)}_{i(k)})\varphi(x^{(2)}_{i(1)}, \dots, x^{(2)}_{i(k)})] = (\varphi, \mathrm{T}_\theta\varphi).$$

Therefore,

$$\begin{aligned}
\mathbf{E}Z_{\mathrm{b}}(|Q_{12}| = \theta) &= \sum_{\underline{x}^{(1)}\cdot\underline{x}^{(2)} = n\theta} \mathbf{P}\ (\underline{x}^{(1)}, \underline{x}^{(2)} \text{ are satisfying assignments}) \\
&\le \sum_{\underline{x}^{(1)}\cdot\underline{x}^{(2)} = n\theta} \prod_\varphi (\varphi, \mathrm{T}_\theta\varphi)^{\tilde{p}_n(\varphi)n\alpha} \\
&\le \frac{C}{n^{3/2}} \exp\left(n\left\{\mathcal{H}\left(\frac{1+\theta}{4}, \frac{1+\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}\right) + \alpha\sum_\varphi \tilde{p}_n(\varphi)\log(\varphi, \mathrm{T}_\theta\varphi)\right\}\right),
\end{aligned}$$

where $\mathcal{H}$ is the entropy function

$$\mathcal{H}(\theta_1, \dots, \theta_d) = -\sum_{i=1}^{d} \theta_i \log \theta_i \tag{29}$$

and we used the following bound on binomial coefficients (valid for $\theta_i \geq 0$, $\theta_1 + \cdots + \theta_d = 1$):

$$\frac{n!}{\prod_{i=1}^{d}(n\theta_i!)} \leq \frac{C}{n^{(d-1)/2}} \ \exp\{\mathcal{H}(\theta_1, \dots, \theta_d)\}. \tag{30}$$

By the very same argument, for some positive C',

$$\begin{aligned}\mathbf{E}Z_{\mathrm{b}} &= \sum_{\underline{x}\,\text{balanced}} \prod_{\varphi} \|\varphi\|^{2\tilde{p}_n(\varphi)\alpha n} \\ &> \frac{C'}{n^{1/2}} \exp\left(n\left\{\mathcal{H}\left(\frac{1}{2},\frac{1}{2}\right) + \alpha\sum_{\varphi}\tilde{p}_n(\varphi)\log\|\varphi\|^2\right\}\right).\end{aligned}$$

It is straightforward now to check that

$$\frac{\mathbf{E}Z_{\mathrm{b}}(Q_{12}=\theta)}{(\mathbf{E}Z_{\mathrm{b}})^2} \leq \frac{C''}{n^{1/2}} \ \exp\{n\Phi(\theta)\}, \tag{31}$$

which implies the claim. ∎

LEMMA A.3. *Given $\tilde{p}_n \in \mathcal{K}_n(p; a, A, \gamma)$, consider a random instance from the* $\mathrm{CSP}(n, p, \alpha; \tilde{p}_n)$ *ensemble, and define*

$$\boldsymbol{\Omega}_{k,\tilde{p}_n} \stackrel{\text{def}}{=} \mathbb{E}_{\varphi\sim\tilde{p}_n} \frac{2\mathrm{I}_1(\varphi)}{1-2\mathrm{I}_1(\varphi)}. \tag{32}$$

If $\alpha \leq (1-\varepsilon)\boldsymbol{\Omega}_{k,\tilde{p}_n} \log 2$, then there exists a constant $C_0 = C_0(p; a, A, \gamma, \varepsilon) > 0$ (independent of $\tilde{p}_n \in \mathcal{K}_n(p; a, A, \gamma)$) and an absolute constant C such that for any $\theta \in \{-1, -1+2/n, \dots, 1-2/n, 1\}$

$$\frac{\mathbf{E}[Z_{\mathrm{b}}(Q_{12}=\theta)]}{(\mathbf{E}Z_{\mathrm{b}})^2} \leq \frac{C}{n^{1/2}}\, e^{-nC_0\theta^2}. \tag{33}$$

Proof. In view of the previous lemma, it is sufficient to prove that there exists a constant $C_0 = C_0(p; a, A, \gamma, \varepsilon) > 0$ (independent of $\tilde{p}_n \in \mathcal{K}_n(p; a, A, \gamma)$) such that

$$\Phi(\theta) \leq -C_0\theta^2. \tag{34}$$

Since throughout this proof $\tilde{p}_n$ is fixed, it will be understood that $\varphi \sim \tilde{p}_n$ whenever we take expectation over the clause distribution. Also, dependence of $\boldsymbol{\Omega}_{k,\tilde{p}_n}$ and $\hat{\boldsymbol{\Omega}}_{k,\tilde{p}_n}$ (defined analogously) upon $\tilde{p}_n$ will be dropped.

Fix $\alpha \leq (1-\varepsilon)\boldsymbol{\Omega}_k \log 2 \leq (1-\varepsilon)\hat{\boldsymbol{\Omega}}_k \log 2$. We will prove the thesis claim by considering three different regimes for θ: $0 < \theta \leq e^{-ck}$, $e^{-ck} \leq \theta \leq 1-\varepsilon^{1/2}$, and $1-\varepsilon^{1/2} \leq \theta \leq 1$, where c is a small constant. In the first two intervals we will prove that the derivative of $\Phi(\theta)$ with respect to θ is strictly negative. Recalling that $\|\varphi\|^2 \geq 1/2$, we have

$$\begin{aligned}\frac{\mathrm{d}\Phi}{\mathrm{d}\theta} &\le -\mathrm{atanh}\theta + k\alpha\mathbb{E}_\varphi \frac{(\varphi^{(1)}, \mathrm{T}_\theta\varphi^{(1)})}{\|\varphi\|^4} \\ &\le -\theta + 2k\alpha\mathbb{E}_\varphi \frac{\sum_{i=1}^{k-1}|\varphi^{(1)}_{\{i\}}|^2}{\|\varphi\|^2}\,\theta + 2k\alpha\mathbb{E}_\varphi \frac{\|\varphi^{(1)}\|^2}{\|\varphi\|^2}\theta^3 \\ &\le -\theta + Ae^{-Ck}\frac{\alpha}{\Omega_k}\theta + 2k\frac{\alpha}{\Omega_k}\theta^2 \\ &\le -\frac{1}{2}\theta + 4k\theta^2,\end{aligned}$$

where we used (from (3)) the hypothesis on low weight Fourier coefficients. The last expression is strictly negative if $0 < \theta < e^{-ck}$ for any $c > 0$ and all k large enough. Integrating the last expression over θ, we get $\Phi(\theta) \le -C_0\theta^2$.

Next assume $e^{-ck} \le \theta \le 1-\varepsilon$. Using the hypothesis $(\varphi^{(1)}, T_\theta\varphi^{(1)}) \le e^{-Ck(1-\theta)}\|\varphi^{(1)}\|^2$, we have

$$\begin{aligned}\frac{\mathrm{d}\Phi}{\mathrm{d}\theta} &\le -\mathrm{atanh}\theta + 4k\alpha\mathbb{E}_\varphi \frac{\|\varphi^{(1)}\|^2}{\|\varphi\|^4}\,e^{-Ck\epsilon} \\ &\le -\mathrm{atanh}\theta + 2k\frac{\alpha}{\Omega_k}\,e^{-Ck\sqrt{\epsilon}} \le -\mathrm{atanh}\theta + 2(\log 2)ke^{-Ck\epsilon},\end{aligned}$$

which is strictly negative if $\theta > c^{-ak}$ with, say, $c = (C\varepsilon^2)/2$.

Finally, we notice that, for $1-\varepsilon^2 \le \theta \le 1$, any ε small enough we have $H(\theta) \le -\log 2 + \varepsilon/10$. Further, using the fact that $(\varphi, T_\theta\varphi) = \|T_{\theta^{1/2}}\varphi\|^2$ is nondecreasing in θ

$$\Phi(\theta) \le -\log 2 + \frac{\varepsilon}{10} - \alpha\mathbb{E}_\varphi \log\|\varphi\|^2 = -\log 2 + \frac{\varepsilon}{10} + \frac{\alpha}{\hat{\Omega}_k} \le -\varepsilon\frac{\log 2}{2},$$

which finishes the proof. ∎

Proof of Proposition 3.2, *lower bound.* Fix $\tilde{p}_n \in \mathcal{K}_n(p; a, A, \gamma)$, $\alpha \le (1-\varepsilon)\Omega_{k,\tilde{p}_n}\log 2$, and let Z_{b} be the number of balanced solutions of a random instance from the CSP$(n, p, \alpha; \tilde{p}_n)$ ensemble. From Lemma A.3 we have that, for $U_n \equiv \{-1, -1+2/n, \dots, 1-2/n, 1\}$,

$$\frac{\mathbf{E}\{Z_{\mathrm{b}}^2\}}{\{\mathbf{E}Z_{\mathrm{b}}\}^2} = \sum_{\theta\in U_n}\frac{\mathbf{E}\{Z_{\mathrm{b}}(Q_{12}=\theta)\}}{\{\mathbf{E}Z_{\mathrm{b}}\}^2} \tag{35}$$

$$\le \frac{C}{n^{1/2}}\sum_{\theta\in U_n} e^{-C_0 n\theta^2} \tag{36}$$

$$\le \frac{C'}{n^{1/2}}\,n\int_{-\infty}^{\infty} e^{-C_0 n\theta^2}\mathrm{d}\theta \le C_0' \tag{37}$$

for some new constant $C_0' = C_0'(p; a, A, \gamma, \varepsilon) > 0$.

For $\tilde{p}_n \in \mathcal{K}_n(p; a, A, \gamma)$, we gave $\Omega_{k,\tilde{p}_n} = \Omega_k(1 + O(n^{-1/2+\gamma}))$. Let $\mathcal{F}_n$ be a random instance from the CSP(n, p, α) ensemble, $\tilde{p}_n \in \mathcal{K}_n(p; a, A, \gamma)$, $\alpha \le (1-2\varepsilon)\Omega_k\log 2$, whence $\alpha \le (1-\varepsilon)\Omega_{k,\tilde{p}_n}\log 2$. By the Paley–Zygmund inequality

$$\mathbf{P}(\mathcal{F}_n \text{ is sat} | L_n = \tilde{p}_n) \geq \frac{\mathbf{E}\{Z_{\rm b}^2\}}{2\{\mathbf{E}\, Z_{\rm b}\}^2} > C_0'/2. \tag{38}$$

By Lemma A.1 we have $\mathbf{P}(\mathcal{F}_n \text{ is sat}) \geq C_0'/4$. Finally, the fact that the satisfiability property (of our CSP ensembles) exhibits a sharp transition, thanks to the theorem of Creignou and Daude [CD09] (see Theorem C.1 in Appendix C here) implies $\mathbf{P}(F_n \text{is sat}) \to 1$ as $n \to \infty$. ∎

Appendix B. Proof of Theorem 3.4. In this appendix we introduce the planted CSP ensemble, clarify its connection to the original ensemble, and use it to prove Theorem 3.4. Throughout the section, we denote a CSP instance with $n\alpha$ clauses by $F = (F_1, F_2, \dots, F_{n\alpha})$. Here

$$F_a = (\varphi_a; i_a(1), \dots, i_a(k)) \tag{39}$$

denotes the clause labeled a, which is completely specified by the Boolean function $\varphi_a : \{+1, -1\}^k \to \{0, 1\}$ and by the choice of k indices $i_a(1), \dots, i_a(k)$. The number of solutions of the instance F is denoted by $Z(F)$.

Given a distribution $p = \{p(\varphi)\}$, it is also convenient to define the "average clause" $\bar{\varphi} : \{+1, -1\}^n \to \mathbb{R}_+$:

$$\bar{\varphi}(\underline{x}) = \frac{1}{n^k} \sum_{i(1),\dots,i(k)\in[n]} \sum_{\varphi} p(\varphi)\varphi(x_{i(1)}, \dots, i(k)). \tag{40}$$

Throughout this section, we will assume that the strong balance condition (condition (a′) in section 3.5) holds. We think that this condition can be refined at the price of a more careful analysis.

B.1. The planted ensemble and a transfer theroem. Given $n \in \mathbb{N}, \alpha \geq 0$, and a distribution $p = \{p(\varphi)\}$ over k-clauses, the planted ensemble $\text{pCSP}(n, \alpha, p)$ is a joint distribution over binary assignments $\underline{x}^* = (x_1^*, x_2^*, \dots, x_n^*) \in \{0, 1\}^n$ and random CSP formulas F defined as follows. The assignment $\underline{x}^*$ is drawn with distribution

$$\mathbf{P}_{\rm p}(\underline{x}) \equiv \frac{1}{\mathbf{E}\, Z(F)} \bar{\varphi}(\underline{x})^{n\alpha}. \tag{41}$$

It is easy to check that this is normalized, i.e., that $\mathbf{E}\, Z(F) = \sum_x \bar{\varphi}(\underline{x})^{n\alpha}$.

We will use $\mathbf{P}_{\rm p}, \mathbf{E}_{\rm p}$ to denote probability and expectation with respect to the planted model. Sampling $\underline{x}$ from this distribution is straightforward, since $\mathbf{P}_{\rm p}(\underline{x})$ is uniform once we condition on the weight of $\underline{x}$ (i.e., on $\underline{x} \cdot \underline{1}$).

Conditional on $\underline{x}^*$, the clauses F_a, $a = 1, 2, \dots, n\alpha$, are independent and distributed according to

$$\mathbf{P}_{\rm p}\{F_a = (\varphi_a, i_a(1), \dots, i_a(k)) | \underline{x}^*\} \equiv \frac{1}{n^k \bar{\varphi}(\underline{x}^*)} p(\varphi_a)\varphi_a(x^*_{i_a(1)}, \dots, x^*_{i_a(k)}), \tag{42}$$

where the indices $i_a(1), \dots, i_a(k) \in [n]$ are drawn independently and uniformly at random. Notice that this is indeed a well-defined distribution over clauses, and in particular it is normalized thanks to (40). In order to sample from the above clause distribution, one can proceed as follows. Sample indices $i_a(1), \dots, i_a(k) \in [n]$ independently and uniformly at random and a Boolean function φ_a with distribution $p(\cdot)$. If $\varphi_a(x^*_{i_a(1)}, \dots, x^*_{i_a(k)}) = 1$, accept this choice; otherwise reject it and repeat the sampling.

The joint distribution of the planted assignment and the CSP instance is then

$$\mathbf{P}_{\rm p}(F,\underline{x}^*) = \frac{1}{n^{nk\alpha}\mathbf{E}Z(F)}\prod_{a=1}^{n\alpha} p(\varphi_a)\varphi_a(x^*_{i_a(1)},\dots,x^*_{i_a(k)}). \tag{43}$$

By construction, the assignment $\underline{x}^*$ satisfies F. It is convenient to compare the planted distribution with the uniform distribution we have been considering so far. In this case, an instance is drawn according to the ensemble $\mathrm{CSP}(n,\alpha,p)$, and an assignment $\underline{x}^*$ is drawn uniformly at random from among the ones satisfying F. The joint distribution is then

$$\mathbf{P}_{\rm p}(F,\underline{x}) = \frac{1}{n^{nk\alpha}Z(F)}\prod_{a=1}^{n\alpha} p(\varphi_a)\varphi_a(x^*_{i_a(1)},\dots,x^*_{i_a(k)}). \tag{44}$$

By taking the ratio of the above probabilities, we immediately get the following lemma.

LEMMA B.1. *Let $\mathcal{F}\colon(F,\underline{x}^*)\to\mathbb{R}$ be a function of an instance–solution pair. Its expectations with respect to the planted and uniform model are related as follows:*

$$\mathbf{E}_{\rm p}\mathcal{F}(F,\underline{x}^*) = \mathbf{E}\left\{\frac{Z(F)}{\mathbf{E}Z(F)}\mathcal{F}(F,\underline{x}^*)\right\}. \tag{45}$$

Proof. By a standard change-of-measure argument $\mathbf{E}_{\rm p}\mathcal{F}(F,\underline{x}^*)$ is equal to

$$\begin{aligned}\sum_{(F,\underline{x}^*)}\mathbf{P}_p(F,\underline{x}^*)\mathcal{F}(F,\underline{x}^*) &= \sum_{(F,\underline{x}^*)}\mathbf{P}(F,\underline{x}^*)\left\{\frac{\mathbf{P}_{\rm p}(F,\underline{x}^*)}{\mathbf{P}(F,\underline{x}^*)}\mathcal{F}(F,\underline{x}^*)\right\}\\ &= \sum_{(F,\underline{x}^*)}\mathbf{P}(F,\underline{x}^*)\left\{\frac{Z(F)}{\mathbf{E}Z(F)}\mathcal{F}(F,\underline{x}^*)\right\},\end{aligned} \tag{46}$$

which is nothing but our claim. ∎

It is clear that the planted and uniform models are strictly related as soon as $Z(F)$ concentrates around its expectation $\mathbf{E}Z(F)$.

LEMMA B.2. *Fix $\alpha < \Omega_k \log 2\{1+o_k(1)$ and let $Z(F)$ be the number of solutions of a random instance F from the $\mathrm{CSP}(n,\alpha,p)$ ensemble. Then, for any $\varepsilon>0$, $Z(F) > e^{-n\varepsilon}\mathbf{E}Z(F)$ with high probability.*

Proof. For any constant A, the property $Z(F) > e^{nA}$ is monotone over the space of CSP instances (regarded as a product space). Applying, as in [AC08], a sharp threshold result (which we prove as Lemma C.2 in Appendix C), it is sufficient to prove that $Z(F) > e^{-n\varepsilon}\mathbf{E}Z(F)$ with probability bounded away from 0 as $n\to\infty$.

Let $Z_{\rm b}(F)$ be the number of balanced solutions (i.e., the number of solutions such that $|\underline{x}\cdot\underline{1}|\le 1$). Obviously, $Z(F)\ge Z_{\rm b}(F)$. On the other hand, by an argument already employed in Appendix A (here $U_n \equiv \{-1,-1+2/n,\dots,1-2/n,1\}$),

$$\begin{aligned}\mathbf{E}\{Z(F)\} &= \sum_{x\in\{-1,1\}^k} \mathbf{P}\,(x \text{ is a satisfying assignment})\\ &\le \sum_{\theta\in U_n} \binom{n}{n(1+\theta)/2} \mathbb{E}_\varphi\{\|\varphi\|_\theta\}^2\\ &\le \sum_{\theta\in U_n} \binom{n}{n(1+\theta)/2} \mathbb{E}_\varphi\{\|\varphi\|\}^2\\ &\le n\binom{n}{n/2} \mathbb{E}_\varphi\{\|\varphi\|\}^2 = n\mathbf{E}\{Zb_{\mathrm{b}}(F)\}.\end{aligned}$$

That is, $\mathbf{E}\{Z(F)\}$ and $\mathbf{E}\{Z_b(F)\}$ differ at most by a polynomial factor. It is therefore sufficient to prove that $Z_{\mathrm{b}}(F) > e^{-n\varepsilon}\mathbf{E}Z_{\mathrm{b}}(F)$ with probability bounded away from 0 as $n\to\infty$.

This follows from the Paley–Zygmund inequality, since

$$\mathbf{P}\left\{Z_{\mathrm{b}}(F) \ge \frac{1}{2}\mathbf{E}Z_{\mathrm{b}}(F)\right\} \ge \frac{\mathbf{E}\{Z_{\mathrm{b}}(F)\}^2}{4\mathbf{E}\{Z_{\mathrm{b}}(F)^2\}} \ge \frac{1}{4C} \tag{47}$$

for some uniformly bounded $C>0$ by (37). ∎

THEOREM B.3. *Given a sequence of events $\{A_n\}$ and a constant $c>0$, assume that $(\underline{x}^*, F)\in A_n$ with probability larger than $1-e^{-cn}$ under the planted model* $\mathrm{pCSP}(n,\boldsymbol{\alpha},p)$. *Then $(\underline{x}^*,F)\in A$ with high probability under the uniform model.*

Proof. Consider the complement of A_n, denoted by A_n^{c}. By Lemma B.1, we have

$$\begin{aligned}\mathbf{P}_{\mathrm{p}}\{(\underline{x}^*,F)\in A_n^c\} &= \mathbf{E}\left\{\frac{Z(F)}{\mathbf{E}Z(F)}\mathbb{I}_{(\underline{x}^*,F)\in A_n^{\mathrm{c}}}\right\}\\ &\ge \mathbf{E}\left\{\frac{Z(F)}{\mathbf{E}Z(F)}\mathbb{I}_{(\underline{x}^*,F)\in A_n^{\mathrm{c}}}\mathbb{I}_{Z(F)\ge e^{-cn/2}}\mathbf{E}Z(F)\right\}\\ &\ge e^{-cn/2}\{\mathbf{P}\{(\underline{x}^*,F)\in A_n^{\mathrm{c}}\} - \mathbf{P}\{(\underline{x}^*,F)\in A_n^{\mathrm{c}}, Z< e^{-cn/2}\mathbf{E}Z(F)\}\}.\end{aligned}$$

By solving for $\mathbf{P}\{(\underline{x}^*,F)\in A_n^{\mathrm{c}}\}$, we get

$$\mathbf{P}\{(\underline{x}^*,F)\in A_n^c\} \le e^{cn/2}\mathbf{P}_{\mathrm{p}}\{(\underline{x}^*,F)\in A_n^{\mathrm{c}}\} + \mathbf{P}\{Z< e^{-cn/2}\mathbf{E}Z(F)\}.$$

The first term vanishes by assumption, and the second by Lemma B.2. ∎

B.2. Clustering. The proof of Theorem 3.4 proceeds in two steps. First we consider a pair $(\underline{x}^*,F)$ drawn according to the planted model and show that the planted solution is isolated from most of the other solutions. Next, we use Theorem B.3 to transfer this statement to the uniform ensemble.

In order to establish the first result, we need the following estimate.

LEMMA B.4. *Let $(\underline{x}^*,F)$ be a solution/instance pair distributed according to the planted model, and denote by $Z^{(2)}(\theta)$ the number of solutions $\underline{x}$ of F such that $\underline{x}^*\cdot\underline{x} = n\theta$. Then, for any $a<1$,*

$$\mathbf{E}_{\mathrm{p}}\{Z^{(2)}(\theta)\,|\,|\underline{x}^*\cdot\underline{1}|\le n^a\} = \exp\{n\Psi(\theta)+o(n)\}, \tag{48}$$

$$\Psi(\theta) \equiv H(\theta) + \boldsymbol{\alpha}\,\log\left\{\frac{\mathbb{E}_\varphi(\varphi,\mathrm{T}_\theta\varphi)}{\mathbb{E}_\varphi\{\|\varphi\|^2\}}\right\}. \tag{49}$$

Proof. For the sake of simplicity we shall focus on the case $\underline{x}^* \cdot \underline{1} = 0$ (i.e., n is even and the planted solution is perfectly balanced). It should be clear from the derivation that allowing for $|\underline{x}^* \cdot \underline{1}| \le n^a$ produces only a change of order $O(n^{-1+a})$ in the exponent.

Fix such a planted solution $\underline{x}^*$, and let $\underline{x}$ be such that

$$\sum_{i:x_i^*=+1} x_i^* x_i = \frac{n}{2}\theta_+, \qquad \sum_{i:x_i^*=+1} x_i^* x_i = \frac{n}{2}\theta_-, \tag{50}$$

with $(\theta_+ + \theta_-)/2 = \theta$ (whence $\underline{x}^* \cdot \underline{x} = n\theta$). Then

$$\mathbf{P}_{\mathrm{p}}(\underline{x} \text{ is a solution } |\underline{x}^*) = [\mathbf{P}_{\mathrm{p}}(\varphi_a(x_{i_a(1)}, \dots, x_{i_a(k)}) = 1 | \underline{x}^*)]^{n\alpha}, \tag{51}$$

and by the definition of planted ensemble

$$\begin{aligned}
&\mathbf{P}_{\mathrm{p}}(\varphi_a(x_{i_a(1)}, \dots, x_{i_a(k)}) = 1 | \underline{x}^*) \\
&\quad = \frac{1}{n^k \bar{\varphi}(\underline{x}^*)} \sum_{i_a(1),\dots,i_a(k)} \sum_{\varphi} p(\varphi)\varphi(x^*_{i_a(1)}, \dots, x^*_{i_a(k)})\varphi(x_{i_a(1)}, \dots, x_{i_a(k)}) \\
&\quad = \frac{1}{\bar{\varphi}(\underline{x}^*)} \mathbb{E}_{\varphi}(\varphi, S_{\theta_+,\theta_-}\varphi),
\end{aligned}$$

where we introduced the operator S_{θ_+,θ_-} acting as follows:

$$S_{\theta_+,\theta_-}\varphi(x_1, \dots, x_k) \equiv \sum_{y\in\{+1,-1\}^k} \prod_{i=1}^{k} \frac{1+\theta_{x_i} x_i y_i}{2} \varphi(y_1, \dots, y_k). \tag{52}$$

Further

$$\mathbf{P}_{\mathrm{p}}(\underline{x}^* \cdot \underline{1} = 0) = \frac{1}{\mathbf{E}Z(F)} \bar{\varphi}(\underline{x}^*)^{n\alpha} \binom{n}{n/2}.$$

Combining the above, and after a few algebraic manipulations, we get

$$\begin{aligned}
\mathbf{E}_p\{Z^{(2)}(\theta) | \underline{x}^* \cdot 1 = 0\} = \frac{1}{\bar{\varphi}(\underline{x}^*)^{n\alpha}} \sum_{\theta_+ + \theta_- = 2\theta} &\binom{n/2}{n(1+\theta_+)/4} \binom{n/2}{n(1+\theta_-)/4} \\
&\times [\mathbb{E}_{\varphi}(\varphi, S_{\theta_+,\theta_-}\varphi)]^{n\alpha},
\end{aligned}$$

where the sum runs over $\theta_+, \theta_- \in \{-1, -1+4/n, \dots, 1-4/n, 1\}$. Now letting $\delta = (\theta_+ - \theta_-)/2$ and passing to the Fourier transform, we get

$$\mathbb{E}_{\varphi}(\varphi, S_{\theta_+,\theta_-}\varphi) = \sum_{Q_1 \subseteq Q_2} \mathbb{E}_{\varphi}\{\varphi_{Q_1}\varphi_{Q_2}\}\theta^{|Q_1|}\delta^{|Q_2|-|Q_1|} \le \sum_{Q} \mathbb{E}_{\varphi}\{\varphi_Q^2\}\theta^{|Q|} = \mathbb{E}_{\varphi}(\varphi, S_{\theta,\theta}\varphi),$$

where we used (2). Also notice that $(\varphi, S_{\theta,\theta}\varphi) = (\varphi, T_{\theta}\varphi)$. Therefore, the sum over θ_+, θ_- can be estimated by the $\theta_+ = \theta_-$ term, up to a polynomial factor

$$\mathbf{E}_p\{Z^{(2)}(\theta) | \underline{x}^* \cdot 1 = 0\} = \frac{1}{\bar{\varphi}(\underline{x}^*)^{n\alpha}} n^{O(1)} \binom{n/2}{n(1+\theta)/4}^2 [E_{\varphi}(\varphi, T_{\theta}\varphi)]^{n\alpha}.$$

The statement follows by noticing that $\bar{\varphi}(\underline{x}^*) = \mathbb{E}\|\varphi\|^2$ for $\underline{x}^*$ balanced. ∎

LEMMA B.5. *Let $(\underline{x}^*, F)$ be a solution/instance pair distributed according to the planted model* $\mathrm{pCSP}(n, \alpha, p)$ *and assume*

$$\frac{\tilde{\Omega}_k}{k}(\log\, k)(1+\varepsilon) \le \alpha \le \Omega_k(\log\, 2)(1-\varepsilon). \tag{53}$$

Then there exists constants $0 < \theta_1 < \theta_2 < 1$ and $c,\ c' > 0$ such that, with probability at least $1 - e^{-cn}$, the following happens. The instance F does not admit any solution $\underline{x}$ with $n\theta_1 \le \underline{x} \cdot \underline{x}^* \le n\theta_2$, and the number of solutions with $\underline{x} \cdot \underline{x}^* \ge n\theta_2$ is at most $e^{-nc'}\mathbb{E}Z(F)$ (expectation is here with respect to the uniform model).

Proof. In view of Lemma B.4 it is sufficient to show that $\theta_* \in (0,1)$ such that the following hold:

(a) $\Psi(\theta_*) < 0$.

(b) $\sup_{\theta\in[\theta_*,1]}\Psi(\theta) < \log\, 2 + \alpha\, \log\, \mathbb{E}\|\varphi\|^2$.

In order to prove (a), we first notice that for any $\varepsilon \in (0, 1/2)$,

$$\frac{\mathbb{E}_\varphi(\varphi, \mathrm{T}_\theta\varphi)}{\mathbb{E}_\varphi\|\varphi\|^2} \le 1 - \frac{1}{(1+\varepsilon)\tilde{\Omega}_k} + \frac{1}{(1+\varepsilon)\tilde{\Omega}_k}\, e^{-k(1+\varepsilon)(1-\theta)}, \tag{54}$$

provided $\theta > 1 - \varepsilon$. Indeed, both sides equal 1 at $\theta = 1$. Further, the derivative of the left-hand side can be estimated as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\mathbb{E}_\varphi(\varphi, \mathrm{T}_\theta\varphi)}{\mathbb{E}_\varphi\|\varphi\|^2} &= 2k\frac{\mathbb{E}_\varphi(\varphi^{(1)}, \mathrm{T}_\theta\varphi^{(1)})}{\mathbb{E}_\varphi\|\varphi\|^2} \ge 2k\frac{e^{-k(1+\varepsilon)(1-\theta)}\mathbb{E}_\varphi\|\varphi^{(1)}\|^2}{\mathbb{E}_\varphi\|\varphi\|^2} \\
&= ke^{-k(1+\varepsilon)(1-\theta)}\frac{2\mathbb{E}_\varphi \mathrm{I}_1(\varphi)}{1 - 2\mathbb{E}_\varphi \mathrm{I}_1(\varphi)} \\
&\ge \frac{\mathrm{d}}{\mathrm{d}\theta}\left\{1 - \frac{1}{(1+\varepsilon)\tilde{\Omega}_k} + \frac{1}{(1+\varepsilon)\tilde{\Omega}_k}\, e^{-k(1+\varepsilon)(1-\theta)}\right\}.
\end{aligned}$$

Here we used the following inequality, valid for any $f\colon\{+1,-1\}^k \to \{0,1\}$, provided $\theta > 1 - \varepsilon$:

$$(f, \mathrm{T}_\theta f) = \sum_Q |f_Q|^2\theta^{|Q|} \ge \|f\|^2\theta^k \ge \|f\|^2 e^{-k(1+\varepsilon)(1-\theta)}. \tag{55}$$

Let $\alpha = (1+\varepsilon)(\tilde{\Omega}_k/k)\gamma\, \log\, k/k$ and $\theta_* = 1 - \omega_*/k$. Equation (54) implies

$$\Psi(\theta_* = 1 - \omega_*/k) \le H(\omega_*/k) - \frac{\gamma}{k}(\log\, k) + \frac{\gamma}{k}(\log\, k)e^{-(1+\varepsilon)\omega_*} \tag{56}$$

for all $\varepsilon > \omega_*/k$. If we fix $\varepsilon = \omega_{\max}/k$ and let $k \to \infty$, we finally obtain (for $\omega \in (0, \omega_{\max})$)

$$\Psi(\theta_* = 1 - \omega_*/k) \le \{\omega_* - \gamma + \gamma e^{-\omega_*}\}\frac{\log\, k}{k} + O(k^{-1}). \tag{57}$$

As soon as $\gamma > 1$, we can find ω_* such that $\omega_* - \gamma + \gamma e^{-\omega_*} < 1$ (just take $\omega_* = \log\, \gamma$). Further, $\sup_{\theta\in[\theta_*,1]}\Psi(\theta) = O(1/k)$, which is smaller than $\log\, 2 + \alpha\, \log\, \mathbb{E}\|\varphi\|^2$ for k large enough and $\alpha < \Omega_k(\log\, 2)(1-\varepsilon)$. ∎

Proof of Theorem 3.4 Consider a random instance from the $\mathrm{CSP}(n, \alpha, p)$ ensemble, and sample a solution $\underline{x}^*$ uniformly at random. By Lemma B.5 and Theorem B.3, with

high probability there is no solution $\underline{x}$ such that $\underline{x} \cdot \underline{x}^* \in [n\theta_1, n\theta_2]$. Declare the cluster of $\underline{x}^*$, $\mathcal{C}(\underline{x}^*)$ to be the set of solutions $\underline{x}$ such that $\underline{x} \cdot \underline{x}^* \geq n\theta_2$. It will contain an exponentially small fraction of solutions.

The same operation can be repeated $e^{n\delta}$ times. Since each cluster thus constructed is exponentially small, for δ small enough the probability that any of the two clusters intersects is exponentially small. □

Appendix C. Sharp threshold results for CSPs. Recall that in the previous section, we appealed crucially in two places to certain sharp transition behavior of the CSPs under consideration. We furnish the requisite references and details here.

Since we are interested in the behavior of binary k-CSPs for large k, in what follows we may safely assume that $k \geq 3$. Once again for simplicity, let $F = F_k(n, \alpha n)$ denote a random binary CSP(n, α, p) on n variables and αn clauses, and the distribution p over clauses satisfying the main conditions 1–4 mentioned in section 3. As is customary, for the SAT-UNSAT threshold to be meaningful, we also assume that p satisfies the following elementary condition.

5. *Unsatisfiability of the ensemble.* For every $\epsilon = \pm 1$, there is at least one clause g with $p_g > 0$ such that $g(\epsilon, \ldots, \epsilon) = 0$. (Note that by the balance condition 2, necessarily $g(-\epsilon, \ldots, -\epsilon) = 0$.)

Building on their previous work, Creignou and Daude recently showed [CD09] that the satisfiability of $F_k(n, \alpha n)$ undergoes a sharp transition, except when the formula contains a function of one of the following two types.

(i) A Boolean function f *strongly depends on one component* if there exist $\epsilon \in \{+1, -1\}$ and i with $1 \leq i \leq k$ such that $(x_1, \ldots, x_n) \in \{+1, -1\}^n$ and $f(x_1, \ldots, x_n) = 1$ imply that $x_i = \epsilon$.

(ii) A Boolean function f *strongly depends on a 2-XOR-relation* if there exist i, j with $1 \leq i \neq j \leq k$ such that $(x_1, \ldots, x_n) \in \{+1, -1\}^n$ and $f(x_1, \ldots, x_n) = 1$ imply that $x_i \oplus x_j = 1$.

THEOREM C.1 (see [CD09]). *With $F = F_k(n, \alpha n)$ and p satisfying 5 above, the transition from SAT(F) to UNSAT(F) is sharp if and only if F contains no function strongly dependent on one component and no function strongly dependent on a 2-XOR-relation.*

Note that we had used this result in completing the proof of the lower bound in Proposition 3.2 in Appendix A.

We now furnish various details needed to justify that the property of having an exponential number of solutions has a sharp threshold. Recall that this was needed to boost Lemma B.2 (see Appendix B) in the proof of the clustering threshold, to show that the probability once bounded away from 0, is actually tending to 1, as the problem size n went to infinity.

Let Φ be a formula on the variables $y_1, \ldots, y_l$ that can be constructed from our ensemble, let $X = \{x_1, \ldots, x_n\}$ be a set of n variables (disjoint from $\{y_1, \ldots, y_l\}$), and let Φ_n denote the set of all formulas that results after substituting l distinct variables from X and replacing them in Φ. Given a CSP ensemble F on n variables, let $F \oplus \Phi$ be equal to $F \wedge \Phi^*$, where Φ^* is a random formula chosen uniformly from Φ_n.

We say a random ensemble F has the property $\mathcal{A}_B = \mathcal{A}_B(F)$ if F has fewer than $\frac{1}{2} B^n$ satisfying assignments. We want to prove the following.

LEMMA C.2. *For any $B \in [1, 2)$ there is a sequence t_n^B such that for any $\epsilon > 0$,*

$$\lim_{n \to +\infty} \mathbf{P}(F_k(n, (1-\epsilon) t_n^B) \text{ has property } \mathcal{A}_B) = 0, \quad \text{and}$$
$$\lim_{n \to +\infty} \mathbf{P}(F_k(n, (1+\epsilon) t_n^B) \text{ has property } \mathcal{A}_B) = 1. \tag{58}$$

Note that $\mathcal{A}_B$ is a monotone property, since whenever F has the property, then $F \wedge F'$ will have the property for any formula F' on the variables $\{x_1, \ldots, x_n\}$. We will use the following theorem of Friedgut [F99], [F05] to prove that $\mathcal{A}_B$ has a "sharp threshold," in the sense of Lemma C.2].

THEOREM C.3 (see [F05]). *Suppose that $\mathcal{A}_B$ does not have a sharp threshold. Then, there exists $\alpha > 0$, a formula Φ, and for any $n_0 > 0$, there exist $n > n_0$, $m > 0$, and a formula F with variables $x_1, \ldots, x_n$ such that all of the following hold:*

T1. $\mathbf{P}(F \oplus \Phi$ *has the property* $\mathcal{A}_B) > 1 - \alpha$.

T2. $\alpha < \mathbf{P}(F_k(n, m)$ *has the property* $\mathcal{A}_B) < 1 - 3\alpha$.

T3. *With probability at least α, a random formula $F_k(n, m)$ contains an element of Φ_n as a subformula.*

T4. $\mathbf{P}(F \wedge F_k(n, 2 \log n)$ *has the property* $\mathcal{A}_B) < 1 - 2\alpha$.

A first observation is the subtle fact that Theorem C.3 is originally stated in terms of a parametric Bernoulli model, while our model is binomial. But it is the case, by standard arguments, that we can translate results concerning the existence of a sharp threshold of monotone properties from one model to other, provided that m is of order $\Omega(n)$. We will prove that this is the case in step (1) below.

An important fact that we will use throughout is that, because of the feasibility condition, a *pure literal* reduction scheme exists: Suppose that x_l is a variable that appears *only once* in a formula $F = C_1 \wedge \cdots \wedge C_m$, say, in the clause $C_1 = f(x_l, x_{i_1}, \ldots, x_{i_{k-1}})$. Then, any satisfying assignment $\chi : [n] \setminus \{l\} \to \{\pm 1\}$ of $C_2 \wedge \cdots \wedge C_m$ can be extended to a satisfying assignment $\bar{\chi} : [n] \to \{\pm 1\}$ of $C_1 \wedge C_2 \wedge \cdots C_m$ by setting $\bar{\chi}(l)$ to the appropriate value (due to feasibility), such that $f(\bar{\chi}(l), \chi(i_1), \ldots, \chi(i_{k-1})) = 1$.

Notice that using iteratively a pure literal reduction scheme, we can find a satisfying assignment for the formula if we can iteratively find a variable contained once in the formula, eliminate the clause containing the variable, and proceed again with the new formula, until obtaining an empty formula. This procedure is equivalent to that of finding the 2-core of the associated hypergraph [M05], and, in fact, it is the case that if the associated hypergraph has an empty 2-core, then this pure literal reduction scheme will be successful in finding a satisfying assignment.

The approach we will use to prove Lemma C.2 follows that of [AC08], with some variations that follow the work of Creignou and Daude in [CD02], [CD04], and [CD09]. As is standard in these proofs, in what follows we will assume the existence of α, Φ, n, and m satisfying T1–T3, and to conclude that the property $\mathcal{A}_B$ has a sharp threshold, we will prove that T4 cannot hold. Notice that we can always assume that n is large enough by choosing n_0 appropriately. We will divide the core of the proof into three steps. In the first step, we determine the correct scaling of m. In the second step, we prove that the small formula Φ is indeed satisfiable. And, in the last step, we proceed to conclude that T4 does not hold, completing the contradiction argument.

(1) *Scaling of m: Lower bound:* Notice that for $m \equiv \epsilon n/k$, necessarily $(1 - \epsilon)n$ variables do not appear in $F_k(n, m)$, so that if $F_k(n, m)$ is satisfiable, it contains at least $2^{(1-\epsilon)n}$ satisfying assignments. But, following [M05], there is a constant c^* such that if $m < c^* n$, then the hypergraph associated to $F_k(n, m)$ with high probability does not have a 2-core, and as mentioned before, the pure literal reduction is successful in finding a satisfying assignment. This proves, by choosing ϵ small enough, that for $m \equiv \epsilon n/k$, with high probability, $F_k(n, m)$ has at least $2^{(1-\epsilon)n} \geq \frac{1}{2} B^n$ satisfying assignments. Therefore, by T2, it should be the case that $m = \Omega(n)$.

Upper bound: From the first moment estimates in the present paper, we have that there is a constant C_p (depending only on p), such that with high probability, a random formula $F_k(n, C_p n)$ is not satisfiable. Therefore (by T2), due to the monotonicity of $\mathcal{A}_B$, it should be the case that $m = \mathrm{O}(n)$.

(2) *Satisfiability of* $\mathbf{\Phi}$*:* Given a formula $\mathbf{\Phi}$, define $\mathbf{v}(\mathbf{\Phi})$ to be the number of variables in $\mathbf{\Phi}$, and $\mathbf{w}(\mathbf{\Phi})$ to be the number of clauses in $\mathbf{\Phi}$. By an easy counting, for any $t \geq 1$, if $m = \mathrm{O}(n)$, then the probability that a random formula $F_k(n, m)$ contains a subformula $\mathbf{\Phi}$ with $w(\mathbf{\Phi}) \leq t$ and such that $\mathbf{v}(\mathbf{\Phi}) \leq (k-1)\mathbf{w}(\mathbf{\Phi}) - 1$ goes to zero as $n \to +\infty$. Now, if $\mathbf{\Phi}$ is unsatisfiable, then it contains a minimal unsatisfiable formula ψ with $\mathbf{w}(\psi) \leq t$, and therefore, by the previous conclusion, by T2 and T3, we have that $\mathbf{v}(\psi) > (k-1)\mathbf{w}(\psi)$ with high probability. Then, using [CD02, Lemma 5.2], ψ has either a constraint with $k-1$ variables appearing only once, or it is unicyclic. In either case, for $k \geq 3$, there is at least one variable appearing only once in the formula; therefore, the pure literal reduction operates, contradicting the minimality of ψ.

(3) *Contradicting* T4*:*

Step 3a: By T3 and the conclusion of step (1), $\mathbf{\Phi}$ is with high probability satisfiable. Let $\{y_1, \ldots, y_l\}$ be the variables appearing in $\mathbf{\Phi}$, and let $\sigma : \{1, \ldots, l\} \to \{\pm 1\}$ be a fixed satisfying assignment of $\mathbf{\Phi}$. We say that a satisfying assignment χ of F is compatible with a tuple $(z_1, \ldots, z_l) \in [n]^l$ if $\chi(z_i) = \sigma(i)$ for all $i = 1, \ldots, l$. Furthermore, we say that the tuple $(z_1, \ldots, z_k)$ is *bad* if F has fewer than $\frac{1}{2}B^n$ satisfying assignments compatible with $(z_1, \ldots, z_l)$. Notice that by T1, there are at least $(1-\alpha)n^l$ bad tuples.

Step 3b: By the Erdös–Simonovits theorem [ES82], if l k-tuples $(w_1^1, \ldots, w_1^k), \ldots, (w_l^1, \ldots, w_l^k)$ are chosen uniformly at random and independently from n^k, then with probability at least γ', for every function $f : [l] \to [k]$, the tuple $(w_1^{f(1)}, \ldots, w_l^{f(l)})$ is a bad tuple. In particular, we have that with probability at most $(1 - p_g^l \gamma')^{(\log n)/l}$, a random formula $F_k(n, \log n)$ will not contain l clauses $C_1, \ldots, C_l$ satisfying the following:

(i) $C_i = g(v_i^1, \ldots, v_i^k)$ for $i = 1, \ldots, l$, where g is the Boolean function whose existence is implied by condition 5.

(ii) For every function $f : [l] \to [k]$, the l-tuple $(v_1^{f(1)}, \ldots, v_l^{f(l)})$ is bad.

Therefore, by choosing n large enough, the probability that a random formula $F_k(n, \log n)$ contains clauses satisfying (i) and (ii) is at least $1 - \alpha$.

Step 3c: Let $C_1, \ldots, C_l$ be clauses satisfying (i) and (ii), and let $\chi : [n] \to \{\pm 1\}$ be a satisfying assignment of $F \wedge C_1 \wedge \ldots \wedge C_l$. Then note that for every $i = 1, \ldots, l$, there exists an $f(i)$ such that $\chi(v_i^{f(i)}) = \sigma(i)$. Otherwise, for some i, and all $j = 1, \ldots, k$, $\chi(v_i^j) = -\sigma(i)$, which implies that χ does not satisfy C_i, which is a contradiction. It now follows that χ is compatible with $(v_1^{f(1)}, \ldots, v_l^{f(l)})$. Therefore, we conclude that every satisfying assignment of $F \wedge C_1 \wedge \cdots \wedge C_l$ is compatible with $(v_1^{f(1)}, \ldots, v_l^{f(l)})$ for some function $f : [l] \to [k]$. But, by condition (ii), every one of these l-tuples is bad, and therefore, each one does not have more than $\frac{1}{2}B^n$ satisfying assignments compatible with them. As a result, $F \wedge C_1 \wedge \cdots \wedge C_l$ does not have more than $\frac{1}{2}k^l B^n$ satisfying assignments. Moreover, combining step 2b and step 3c, we conclude that with probability at least $1 - \alpha$, $F \wedge F^*$ contains at most $\frac{1}{2}k^l B^n$ satisfying assignments, where F^* is a random $F_k(n, \log n)$ formula.

Step 3d: Given a satisfying assignment $\chi : [n] \to \{\pm 1\}$, with probability at least 2^{1-k}, the clause $g(v_1, \ldots v_k)$, where $(v_1, \ldots, v_k)$ is chosen uniformly at random from $[n]^k$, will not be satisfied by χ. In particular, a random clause will be satisfied by χ with probability at most $1 - p_g 2^{1-k}$. More generally, a random $F_k(n, \log n)$ will be satisfied by χ with probability at most $(1 - p_g 2^{1-k})^{\log n} \leq \frac{1}{n^{c_k}}$, where $c_k = p_g 2^{1-k}$. Therefore, if F^{**} is a $F_k(n, \log n)$ random formula independent of F^*, we have that

$$\mathbf{E}\left[\#\text{sat. assign. of } F \wedge F^* \wedge F^{**} \mid \#\text{sat. assign. of } F \wedge F^* \leq \frac{1}{2} k^l B^n\right] \leq \frac{1}{2n^{c_k}} k^l B^n,$$

and therefore, by Markov's inequality,

$$\mathbf{P}\left[\#\text{sat. assign.of } F \wedge F^* \wedge F^{**} \geq \frac{1}{2} B^n \mid \#\text{sat. assign. of } F \wedge F^* \leq \frac{1}{2} k^l B^n\right] \leq \frac{k^l}{n^{c_k}},$$

which is less than $\alpha/2$ for n large enough. Thus, combining the conclusion of step 2c and the previous formula, we obtain

$$\mathbf{P}\left[\#\text{sat. assign. of } F \wedge F^* \wedge F^{**} \geq \frac{1}{2} B^n\right] \geq 3\alpha/2,$$

and this contradicts T4, thereby proving that property $\mathcal{A}_B$ has a sharp threshold. ∎

Acknowledgments. The last two authors are grateful to Eric Vigoda and Linji Yang for many insightful discussions on reconstruction problems, and for their role in the early development of this project. The authors also gratefully acknowledge the support and the hospitality of BIRS (Canada) and DIMACS (USA), which provided ideal environs for carrying out a significant part of this research collaboration. Finally, the authors thank the anonymous referees for several helpful remarks which resulted in an improved presentation.

REFERENCES

[AC08] D. Achlioptas and A. Coja-Oghlan, *Algorithmic barriers from phase transitions*, in Proceedings of the IEEE FOCS, IEEE, Washington, DC, 2008, pp. 793–802.

[AM02] D. Achlioptas and C. Moore, *The asymptotic order of the random k-SAT threshold*, in Proceedings of the IEEE FOCS, IEEE, Washington, DC, 2002, pp. 779–788.

[AN05] D. Achlioptas and A. Naor, *The two possible values of the chromatic number of a random graph*, Ann. of Math., 162 (2005), pp. 1333–1349.

[ANP05] D. Achlioptas, A. Naor, and Y. Peres, *Rigorous location of phase transitions in hard optimization problems*, Nature, 435 (2005), pp. 759–764.

[AR11] D. Achlioptas, A. Coja-Oghlan, and F. Ricci-Tersenghi, *On the solution-space geometry of random formulas*, Random Structures Algorithms, 38 (2011), pp. 251–268.

[AS04] D. Aldous and J. M. Steele, *The objective method: Probabilistic combinatorial optimization and local weak convergence*, in Probability on Discrete Structures, H. Kesten, ed., Springer, New York, 2004, pp. 1–72.

[Bec75] W. Beckner, *Inequalities in Fourier analysis*, Ann. of Math., 102 (1975), pp. 159–182.
[BK+05] N. Berger, C. Kenyon, E. Mossel, and Y. Peres, *Glauber dynamics on trees and hyperbolic graphs*, Probab. Theory Related Fields, 131 (2005), pp. 311–340.
[BVV11] N. Bhatnagar, J. Vera, and E. Vigoda, *Reconstruction for colorings on trees*, SIAM J. Discrete Math., (2011), to appear.
[BMW00] G. Biroli, R. Monasson, and M. Weigt, *A variational description of the ground state structure in random satisfiability problems*, Eur. Phys. J. B, 14 (2000), pp. 551–568.
[Bon70] A. Bonami, *Études des coefficients Fourier des fonctiones de $L^p(G)$*, Ann. Inst. Fourier, 20 (1970), pp. 335–402.
[CD+03] S. Cocco, O. Dubois, J. Mandler, and R. Monasson, *Rigorous decimation-based construction of ground states for spin-glass models on random lattices*, Phys. Rev. Lett., 90 (2003), 047205.
[CD02] N. Creignou and H. Daude, *Random generalized satisfiability problems*, in Proceedings of SAT, Cincinnati, OH, 2002, pp. 17–26.
[CD04] N. Creignou and H. Daude, *Combinatorial sharpness criterion and phase transition classification for random CSPs*, Inform. Comput., 190 (2004), pp. 220–238.
[CD09] N. Creignou and H. Daude, *The SAT–UNSAT transition for random constraint satisfaction problems*, Discrete Math., 309 (2009), pp. 2085–2099.
[Dia88] P. Diaconis, *Group representations in probability and statistics*, Institute of Mathematical Statistics Lecture Notes 11, Institute of Mathematical Statistics, Hayward, CA, 1988.
[ES82] P. Erdös and M. Simonovits, *Supersaturated graphs and hypergraphs*, Combinatorica, 3 (1982), pp. 181–192.
[F99] E. Friedgut, *Sharp thresholds of graph properties, and the k-SAT problem*, J. Amer. Math. Soc., 12 (1999), pp. 1017–1054
[F05] E. Friedgut, *Hunting for sharp thresholds*, Random Structures Algorithms, 26 (2005), pp. 37–51.
[Geo88] H.-O. Georgii, *Gibbs Measures and Phase Transitions*, de Gruyter, Berlin, 1988.
[GM07] A. Gerschenfeld and A. Montanari, *Reconstruction for models on random graphs*, in Proceedings of the IEEE FOCS, IEEE, Washington, DC, 2007, pp. 194–204.
[HPT08] J. Hartigan, D. Pollard, and S. Tatikonda, *Conditioned Poisson Distributions and the Concentration of Chromatic Numbers*, http://www.stat.yale.edu/~pollard/Papers/chromatic.30june08.pdf.
[KM+07] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborova, *Gibbs states and the set of solutions of random constraint satisfaction problems*, Proc. Natl. Acad. Sci. USA, 104 (2007), 10318–10323
[MPZ02] M. Mézard, G. Parisi, and R. Zecchina, *Analytic and algorithmic solution of random satisfiability problems*, Science, 297 (2002), pp. 812–815.
[MRZ03] M. Mézard, F. Ricci-Tersenghi, and R. Zecchina, *Alternative solutions to diluted p-spin models and XORSAT problems*, J. Statist. Phys., 111 (2003), pp. 505–553.
[MZ02] M. Mézard and R. Zecchina, *Random K-satisfiability problem: From an analytic solution to an efficient algorithm*, Phys. Rev. E, 66 (2002), pp. 056–126.
[MM06] M. Mézard and A. Montanari, *Reconstruction on trees and spin glass transition*, J. Statist. Phys., 124 (2006), pp. 1317–1350.
[MM09] M. Mézard and A. Montanari, *Information, Physics, and Computation*, Oxford University Press, Oxford, UK, 2009.
[MMZ05] M. Mézard, T. Mora, and R. Zecchina, *Clustering of solutions in the random satisfiability problem*, Phys. Rev. Lett., 94 (2005), pp. 197–205.
[M05] M. Molloy, *Cores in random hypergraphs and Boolean formulas*, Random Structures Algorithms, 27 (2005), pp. 124–135.
[MP03] E. Mossel and Y. Peres, *Information flow on trees*, Ann. Appl. Probab., 13 (2003), pp. 817–844.
[Odo08] R. O'Donnell, *Some topics in analysis of Boolean functions*, in Proceedings of ACM STOC, ACM, New York, 2008, pp. 569–578.
[RU08] T. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, Cambridge, UK, 2008.
[Sem08] G. Semerjian, *On the freezing of variables in random constraint satisfaction problems*, J. Statist. Phys., 130 (2008), pp. 130–251.
[Sly09] A. Sly, *Reconstruction of random colourings*, Comm. Math. Phys., 288 (2009), pp. 943–961.