J Stat Phys (2008) 130: 251–293
DOI 10.1007/s10955-007-9417-7

On the Freezing of Variables in Random Constraint Satisfaction Problems

Guilhem Semerjian

Received: 16 May 2007 / Accepted: 27 August 2007 / Published online: 3 October 2007

Abstract The set of solutions of random constraint satisfaction problems (zero energy groundstates of mean-field diluted spin glasses) undergoes several structural phase transitions as the amount of constraints is increased. This set first breaks down into a large number of well separated clusters. At the freezing transition, which is in general distinct from the clustering one, some variables (spins) take the same value in all solutions of a given cluster. In this paper we introduce and study a message passing procedure that allows to compute, for generic constraint satisfaction problems, the sizes of the rearrangements induced in response to the modification of a variable. These sizes diverge at the freezing transition, with a critical behavior which is also investigated in details. We apply the generic formalism in particular to the random satisfiability of boolean formulas and to the coloring of random graphs. The computation is first performed in random tree ensembles, for which we underline a connection with percolation models and with the reconstruction problem of information theory. The validity of these results for the original random ensembles is then discussed in the framework of the cavity method.

Keywords Random constraint satisfaction problems · Freezing transition · Cavity method

1 Introduction

The theory of computational complexity [1] establishes a classification of constraint satisfaction problems (CSP) according to their difficulty in the worst case. For concreteness let us introduce the three problems we shall use as running examples in the paper:

- k-XORSAT. Find a vector $\vec{x}$ of boolean variables satisfying the linear equations $A\vec{x} = \vec{b} \pmod 2$, where each row of the $0/1$ matrix A contains exactly k non-null elements, and $\vec{b}$ is a given boolean vector.

G. Semerjian (✉)
LPTENS, Unité Mixte de Recherche (UMR 8549) du CNRS et de l'ENS associée à l'Université Pierre et Marie Curie, 24 Rue Lhomond, 75231, Paris Cedex 05, France
e-mail: guilhem@lpt.ens.fr

- q-Coloring (q-COL). Given a graph, assign one of q colors to each of its vertices, without giving the same color to the two extremities of an edge.
- k-Satisfiability (k-SAT). Find a solution of a boolean formula made of the conjunction (logical AND) of clauses, each made of the disjunction (logical OR) of k literals (a variable or its logical negation).

Each of these problems admits several variants. In the decision version one has to assert the existence or not of a solution, for instance a proper coloring of a given graph. More elaborate questions are the estimation of the number of such solutions, or, in the absence of solution, the discovery of optimal configurations, for instance colorings minimizing the number of monochromatic edges. The decision variant of the three examples stated above fall into two distinct complexity classes: k-XORSAT is in the P class, while the two others are NP-complete for $k, q \geq 3$ (see [2] for a classification of generic boolean CSPs). This means that the existence of a solution of the XORSAT problem can be decided in a time growing polynomially with the number of variables, for any instance of the problem; one can indeed use the Gaussian elimination algorithm. On the contrary no fast algorithm able of solving every coloring or satisfiability problem is known, and the existence of such a polynomial time algorithm is considered as highly improbable.

This notion of computational complexity, being based on worst-case considerations, could overlook the possibility that "most" of the instances of an NP problem are in fact easy and that the difficult cases are very rare. Random ensembles of problems have thus been introduced in order to give a quantitative content to this notion of typical instances; a property of a problem will be considered as typical if its probability (with respect to the random choice of the instance) goes to one in the limit of large problem sizes. Most random ensembles depend on an external parameter that can be varied continuously. In the coloring problem one can for instance consider the traditional Erdös–Rényi random graphs [3] which are parameterized by their mean connectivity c. For (XOR)SAT instances this role is played by the ratio α of the number of constraints (clauses for SAT or rows in the matrix for XORSAT) to the number of variables. A remarkable threshold phenomenon, first observed numerically [4], occurs when this parameter is varied: when a particular value c_s, α_s is crossed from below, the instances go from typically satisfiable to typically unsatisfiable. This statement has been rigorously proven for XORSAT [5, 6] and for 2-SAT [7], in the other cases it is only a largely accepted conjecture, with sharpness condition on the width of the transition window [8] and bounds on its possible location [9, 10].

Threshold phenomena are largely studied in statistical mechanics under the name of phase transitions. There is moreover a natural analogy between optimization problems and statistical mechanics; if one defines the energy as the number of violated constraints, for instance the number of monochromatic edges, the optimal configurations of a problem coincide with the groundstates of the associated physical system, an antiferromagnetic Potts model in the coloring case. This analogy triggered a large amount of research, relying on methods of statistical mechanics of disordered systems originally devised for the study of mean-field spin-glasses [11]. Early examples of this approach for the satisfiability and coloring problems can be found in [12, 13].

One of the most interesting outcomes of this line of research [14–16] has been the suggestion that other structural threshold phenomena take place before the satisfiability one.[1]

[1] It was of course already known that the algorithms rigorously studied to derive lower bounds on the satisfiability threshold work only up to values of α smaller than α_s [9]. These values are however largely algorithm-dependent and not directly related to a change of structure in the configuration space.

The set of solutions of a random CSP, viewed as a subset of the whole configuration space, is smooth at low values of the constraint ratio but becomes fragmented into clusters of solutions for intermediate values of the control parameter, $\alpha \in [\alpha_d, \alpha_s]$. This clustering transition has been rigorously demonstrated in the XORSAT case [5, 6], for which it has a simple geometric interpretation. α_d is indeed the threshold for the percolation of the 2-core of the hypergraph underlying the CSP; between α_d and α_s there is typically a finite fraction of the variables and constraints in a peculiar sub-formula known as the backbone. Every solution of the backbone gives birth to a cluster of the complete formula. The variables of the backbone are said to be frozen in a given cluster, i.e. they take the same value in all the solutions belonging to a cluster; this is merely a consequence of the definition of a cluster in this case.

Establishing a precise and generic definition of the clusters is not an easy task, not to speak about proving tight rigorous results on their existence or properties (for recent results in this direction see [17–20]). Even at the heuristic level, it was recently argued [21–23] that the computation of α_d for random satisfiability (or c_d for coloring) by previous statistical mechanics studies [24, 25] was incorrect. Roughly speaking, in these two models, the sizes of the clusters can have large fluctuations [26] that must be taken into consideration. In [21] the existence of yet another threshold (for $k, q \geq 4$) $\alpha_c \in [\alpha_d, \alpha_s]$ was also pointed out; this condensation (or replica symmetry breaking) threshold separates two clustered regimes, one where the relevant clusters are exponentially numerous (for smaller values of α) and the other where there is only a sub-exponential number of them.

The clustering transition of XORSAT, because of its geometric interpretation, is certainly a good example on which developing one's intuition of the clustering phenomenon. There are however at least two aspects in which XORSAT departs from other CSP and where the intuitive picture must be taken with a grain of salt. The first is that the clusters of XORSAT all have the same size, because of the linear algebra structure of its set of solutions. For this reason the condensation phenomenon is not present in XORSAT. The second point is that clusters of XORSAT have frozen variables, by definition. There is however no obvious reason that this should be true for any CSP. On the contrary we shall argue in this paper that in general frozen variables appear at another value α_f of the control parameter, with generically $\alpha_f \in [\alpha_d, \alpha_s]$. This was one of the results of [22, 23], here we shall develop this point and quantify the precursors of the transition before α_f. For this we build upon the study of XORSAT presented in [27] and extend it to generic CSPs, in particular satisfiability and coloring. The central notion studied here is the one of rearrangement (to some extent related to the long-range frustration of [28]): given an initial solution of a CSP and a variable i that one would like to modify, a rearrangement is a path in configuration space that starts from the initial solution and leads to another solution where the value of the ith variable is changed with respect to the initial one. The minimal length of such a path is a measure of how constrained was the variable i in the initial configuration. In intuitive terms this length diverges with the system size when the variable was frozen in the initial cluster. More formally we shall indeed find an identification between the probability a rearrangement has a diverging size and the fraction of variables submitted to a "hard field" in the more usual cavity point of view.

The paper is organized as follows. In Sect. 2 we introduce a generic class of CSPs and precise the definition of the rearrangements. Sections 3 and 4 are devoted to modified (tree) random ensembles in which the approach is essentially rigorous; the former presents detailed computations in a rather generic setting and its application to the three selected examples, while the latter presents the numerical results and discuss the generic phenomenology at the approach of the freezing transition in the tree ensembles, with some more technical details deferred to Appendix 1. The computation is reconsidered in the perspective of the

reconstruction problem in Sect. 5. The applicability of these results to the original ensembles is discussed in Sect. 6, through a precise statement of the hypotheses of the cavity method. Conclusions and perspectives for future work are presented in Sect. 7.

2 Definitions

We introduce here some notations and definitions for a class of problems that encompasses the three examples we shall treat in more details. The degrees of freedom of the CSP will be N variables σ_i taking values in a discrete alphabet $\mathcal{X}$; global configurations are denoted $\underline{\sigma} = (\sigma_1, \ldots, \sigma_N)$. An instance (or formula) F of the CSP is a set of M constraints between the variables σ_i. The ath constraint is defined by a function $\psi_a(\underline{\sigma}_a) \to \{0, 1\}$, which depends on the configuration of a subset of the variables $\underline{\sigma}_a$ and is equal to 1 if the constraint is satisfied, 0 otherwise. The set $\mathcal{S}_F \subset \mathcal{X}^N$ of solutions of F is composed of the configurations satisfying simultaneously all the constraints. It can thus be formally defined as $\mathcal{S}_F = \{\underline{\sigma} | \psi_F(\underline{\sigma}) = 1\}$, where the indicator function ψ_F is

$$\psi_F(\underline{\sigma}) = \prod_{a=1}^{M} \psi_a(\underline{\sigma}_a). \tag{1}$$

When the formula admits a positive number of solutions, call it Z_F, the uniform measure over the solutions is denoted $\mu_F(\underline{\sigma}) = \psi_F(\underline{\sigma})/Z_F$.

Factor graphs [29] provide an useful representation of a CSP. These graphs (see Fig. 1 for an example) have two kind of nodes. Variable nodes (filled circles on the figure) are associated to the degrees of freedom σ_i, while constraint nodes (empty squares) represent the clauses ψ_a. An edge between constraint a and variable i is drawn whenever ψ_a depends on σ_i. The neighborhood ∂a of a constraint node is the set of variable nodes that appear in $\underline{\sigma}_a$. Conversely ∂i is the set of constraints that depend on σ_i. We shall conventionally use the indices $i, j, \ldots$ for the variable nodes, $a, b, \ldots$ for the constraints, and denote by $\setminus$ the subtraction from a set. Two variable nodes are called adjacent if they appear in a common constraint. The graph distance between two variable nodes i and j is the number of constraint nodes encountered on a shortest path linking i and j (formally infinite if the two variables are not in the same connected component of the graph).

The three illustrative examples presented above admits a simple representation in this formalism:

- k-XORSAT. The degrees of freedom of this CSP are boolean variables that we shall represent, following the physics conventions, by Ising spins, $\mathcal{X} = \{-1, +1\}$. Each constraint

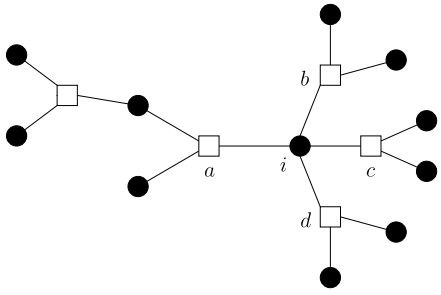

Fig. 1 An example of factor graph. The neighborhoods are for instance $\partial i = \{a, b, c, d\}$ and $\partial i \setminus a = \{b, c, d\}$

involves a subset of k variables, $\underline{\sigma}_a = (\sigma_{i_a^1}, \ldots, \sigma_{i_a^k})$, and reads $\psi_a(\underline{\sigma}_a) = \mathbb{I}(\sigma_{i_a^1}, \ldots, \sigma_{i_a^k} = J_a)$, where here and in the following $\mathbb{I}(\cdot)$ denotes the indicator function of an event and $J_a \in \{-1, +1\}$ is a given constant. This is equivalent to the definition given in the introduction: defining $x_i, b_a \in \{0, 1\}$ such that $\sigma_i = (-1)^{x_i}$ and $J_a = (-1)^{b_a}$, the constraint imposed by ψ_a reads $x_{i_a^1} + \cdots + x_{i_a^k} = b_a \pmod 2$, which is nothing but the ath row of the matrix equation $A\vec{x} = \vec{b}$. The addition modulo 2 of boolean variables can also be read as the binary exclusive OR operation, hence the name XORSAT used for this problem.

- q-COL. Here $\mathcal{X} = \{1, \ldots, q\}$ is the set of allowed colors on the N vertices of a graph. Each edge a connecting the vertices i and j prevents them from being of the same color: $\psi_a(\sigma_i, \sigma_j) = \mathbb{I}(\sigma_i \neq \sigma_j)$.
- k-SAT. As in the XORSAT problem one deals with Ising represented boolean variables, but in each clause the XOR operation between variables is replaced by an OR between literals (i.e. a variable or its negation). In other words a constraint a is unsatisfied only when all literals evaluate to false, or in Ising terms when all spins σ_i involved in the constraint take their wrong value that we denote J_a^i: $\psi_a(\underline{\sigma}_a) = 1 - \mathbb{I}(\sigma_i = J_a^i \ \forall i \in \partial a)$.

The random ensembles of CSPs instances we shall use are defined as follows:

- k-XORSAT. For each of the M clauses a a k-uplet of distinct variable indices $(i_a^1, \ldots, i_a^k)$ is chosen uniformly at random among the $\binom{N}{k}$ possible ones, and the constant J_a is taken to be ± 1 with probability one-half.
- q-coloring. A set of M among the $\binom{N}{2}$ possible edges $a = \{i, j\}$ is chosen uniformly at random.
- k-SAT. The variables i_a^j are chosen as in the XORSAT ensemble, and the J_a^i are independently taken to be ± 1 with equal probability.

For the coloring problem this construction is the classical Erdös–Rényi random graph $G(N, M)$, the two other cases are its random hypergraph generalization. We are interested in the thermodynamic limit of large instances where N and M both diverge with a fixed ratio[2] $\alpha = M/N$. Random (hyper)graphs have many interesting properties in this limit [3]. For instance the degree of a variable node of the factor graph converges to a Poisson law of average αk for the XORSAT and SAT cases, and 2α for the coloring ensemble. For clarity in the latter case we shall use the notation $c = 2\alpha$ for the average connectivity. Moreover, picking at random one variable node i and isolating the subgraph induced by the variable nodes at a graph distance smaller than a given constant L yields, with a probability going to one in the thermodynamic limit, a (random) tree. This tree can be described by a Galton–Watson branching process: the root i belongs to l constraints, where l is a Poisson random variable of parameter αk (c in the coloring case). The variable nodes adjacent to i give themselves birth to new constraints, in numbers which are independently Poisson distributed with the same parameter. This reproduction process is iterated on L generations, until the variable nodes at graph distance L from the initial root i have been generated.

We now define the main object of our study. First recall the well-known definition of the Hamming distance between two configurations, $d(\underline{\sigma}, \underline{\tau}) = \sum_{i=1}^N \mathbb{I}(\sigma_i \neq \tau_i)$. Consider an initial solution of the formula, $\underline{\sigma} \in \mathcal{S}_F$, and imagine one wants to modify the value of

[2] In this limit the quantities studied in this paper are not affected by some variations around these models. For instance in the coloring case $G(N, M)$ can be replaced by the ensemble $G(N, p)$ where each edge is present independently with probability $p = 2\alpha/N$, such that the average number of edges is close to M. The choice of the (hyper)edges with or without replacement is also irrelevant.

the variable i. A rearranged solution is a new configuration $\underline{\tau} \in \mathcal{S}_F$ such that $\tau_i \neq \sigma_i$. The minimal size of a rearrangement (m.s.r.) for variable i starting from $\underline{\sigma} \in \mathcal{S}_F$ is defined as

$$n_i(\underline{\sigma}, F) = \min_{\underline{\tau}} \{ d(\underline{\sigma}, \underline{\tau}) \mid \underline{\tau} \in \mathcal{S}_F, \tau_i \neq \sigma_i \}, \tag{2}$$

and measures how costly (in terms of Hamming distance) it is to perturb the solution at variable i.[3] It can also be viewed as the minimal length of a path in configuration space, modifying one variable at a time, between $\underline{\sigma}$ and another solution with a different value of variable i, thus providing a quantification of how much constrained was initially this variable. We shall also speak of the support of a rearrangement as the set of variables which differ in the initial and final configurations, the size of the rearrangement being the cardinality of its support.

In general the m.s.r. will depend on the starting configuration, we thus define its distribution with respect to an uniform choice of $\underline{\sigma}$ (in abbreviation m.s.r.d.),

$$q_n^{(i,F)} = \sum_{\underline{\sigma}} \mu_F(\underline{\sigma}) \delta_{n, n_i(\underline{\sigma}, F)}. \tag{3}$$

There should be no possibility of confusion between the distribution q_n and the number q of allowed colors in the q-COL problem. When dealing with random CSPs we shall study the average of this distribution,

$$q_n = \mathbb{E} q_n^{(i,F)}, \tag{4}$$

where the expectation is taken with respect to the instance ensemble (in the cases considered here all variable nodes are equivalent on average). Its behavior in the thermodynamic limit will drastically change with the connectivity parameter α (or c for the coloring). We shall indeed define the threshold α_{f} (c_{f}) as the value above which a finite fraction of the distribution q_n is supported on sizes n that diverge with the number of variables. In pictorial terms clusters acquire frozen variables at this point, their rearrangements must be of diverging size and thus lead to a final solution outside the initial cluster.

The computation of the average m.s.r.d. will be first undertaken in a random tree ensemble, mimicking the tree neighborhoods of the random graphs. The threshold for the freezing transition in these tree instances will be computed, along with a set of exponents characterizing the behavior of the average m.s.r.d. when the transition is approached from the unfrozen phase. For clarity we shall denote α_{p} instead of α_{f} the thresholds in the tree ensembles. We shall then argue in Sect. 6, on the basis of the non-rigorous cavity method, that for some values of α and k the properties of the random graphs instances are correctly described by the computations in the tree ensemble. In particular for large enough values of k we shall conjecture that $\alpha_{\mathrm{p}} = \alpha_{\mathrm{f}}$. We will also explain how the computation has to be amended to handle the more elaborated version of the cavity method (with replica-symmetry breaking), and what are the expectedly universal characteristics of the critical behavior at the freezing transition.

3 Minimal Size Rearrangements in Random Tree Ensembles

In this and the next section all the instances of CSP encountered have an underlying factor graph which is a finite tree. Given such a formula F (or equivalently its factor graph) and an

[3] If σ_i takes the same value in every solution we formally define $n_i = N + 1$.

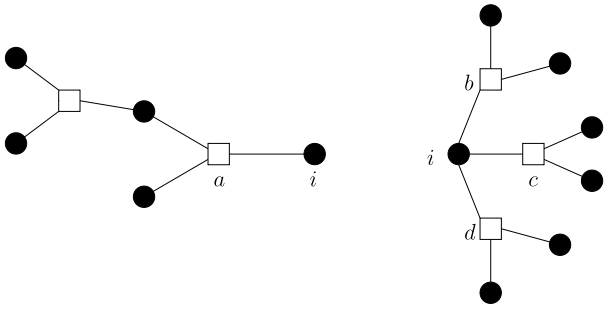

Fig. 2 The cavity graphs $F_{a\to i}$ and $F_{i\to a}$ obtained from the example of Fig. 1

edge $i-a$ between a variable node i and an adjacent constraint node a, we define two sub formulas (cavity graphs) $F_{i\to a}$ and $F_{a\to i}$. $F_{i\to a}$ is obtained from F by deleting the branch of the formula rooted at i starting with constraint a. Conversely $F_{a\to i}$ is obtained by keeping only this branch (see Fig. 2). We also decompose the configuration $\underline{\sigma}$ as $(\underline{\sigma}_{a\to i}, \sigma_i, \underline{\sigma}_{i\to a})$, where $\underline{\sigma}_{a\to i}$ (resp. $\underline{\sigma}_{i\to a}$) is the configurations of the variable nodes in $F_{a\to i}$ (resp. $F_{i\to a}$) distinct from i. The notation $\underline{\sigma}_{\setminus i}$ will be used for the configuration of all variables except i. The computation, based on the natural recursive structure of trees, will be performed in three steps: we shall first see how to obtain $n_i(\underline{\sigma}, F)$, then its distribution with respect to $\underline{\sigma}$, $q_n^{(i,F)}$, which shall finally be averaged over a random tree ensemble. For notational simplicity F will often be kept implicit. This approach is presented in a general setting before the three specific cases of XORSAT, COL and SAT are treated.

3.1 General Case

3.1.1 Given Tree, Given $\underline{\sigma}$

The computation of the m.s.r. n_i on a tree factor graph can be performed in a recursive way. One has to determine, for each value of $\tau_i \neq \sigma_i$, the cost, in terms of Hamming distance, of the modification $\sigma_i \to \tau_i$. This can be done by computing separately these costs in the factor graphs $F_{a\to i}$ for all the constraint nodes a around i and then patching together the rearrangements of the sub-formulae. Rearranging a factor graph $F_{a\to i}$ amounts to looking for a configuration of the variables $j \in \partial a \setminus i$ which satisfies the interaction a and which provokes a minimal propagation of the rearrangement in the branches $F_{j\to a}$.

To formalize this reasoning we introduce a q-component vectorial notation, $\vec{n}$, where the rows of the vectors are indexed by a spin value in $\mathcal{X}$, and we shall denote $[\vec{n}]_\tau$ the τth component of $\vec{n}$. We define $\vec{n}_i(\underline{\sigma})$ as the m.s.r. for i starting from the initial configuration $\underline{\sigma}$, and with the final value τ_i encoded in the row of the vector:

$$[\vec{n}_i(\underline{\sigma})]_{\tau_i} = \min_{\underline{\tau}_{\setminus i}} \{ d(\underline{\sigma}, \underline{\tau} = (\tau_i, \underline{\tau}_{\setminus i})) \mid \underline{\tau} \in S_F \}. \tag{5}$$

The original quantity $n_i(\underline{\sigma})$ is obtained from this more detailed one as $n_i(\underline{\sigma}) = \min_{\tau_i \neq \sigma_i} [\vec{n}_i(\underline{\sigma})]_{\tau_i}$. The recursive computation of $\vec{n}_i$ is performed in terms of vectorial messages on the directed edges of the factor graph, $\vec{n}_{i\to a}$ and $\vec{n}_{a\to i}$. The former, $\vec{n}_{i\to a}(\sigma_i, \underline{\sigma}_{i\to a})$ is defined exactly as $\vec{n}_i$ with the cavity graph $F_{i\to a}$ replacing the original formula F. The latter reads

$$[\vec{n}_{a\to i}(\underline{\sigma}_{a\to i})]_{\tau_i} = \min_{\underline{\tau}_{a\to i}} \{ d(\underline{\sigma}_{a\to i}, \underline{\tau}_{a\to i}) \mid (\tau_i, \underline{\tau}_{a\to i}) \in \mathcal{S}_{F_{a\to i}} \}. \tag{6}$$

Note that here one does not count the cost of flipping the root variable, which avoids overcounting when gluing together the cavity graphs. A moment of thought reveals that these messages obey the following recursive equations:

$$\begin{aligned}\vec{n}_{a\to i}(\underline{\sigma}_{a\to i}) &= \widetilde{f}(\{\vec{n}_{j\to a}(\sigma_j, \underline{\sigma}_{j\to a})\}_{j\in\partial a\setminus i}),\\ \vec{n}_{i\to a}(\sigma_i, \underline{\sigma}_{i\to a}) &= \widetilde{g}_{\sigma_i}(\{\vec{n}_{b\to i}(\underline{\sigma}_{b\to i})\}_{b\in\partial i\setminus a}),\end{aligned} \tag{7}$$

where the functions $\widetilde{f}$ and $\widetilde{g}$ are given by

$$\left[\widetilde{f}(\{\vec{n}_{j\to a}\}_{j\in\partial a\setminus i})\right]_{\tau_i} \equiv \min_{\underline{\tau}_{a\setminus i}}\left\{\sum_{j\in\partial a\setminus i}[\vec{n}_{j\to a}]_{\tau_j} \;\middle|\; \psi_a(\tau_i, \underline{\tau}_{a\setminus i}) = 1\right\}, \tag{8}$$

$$[\widetilde{g}_\sigma(\vec{n}_1,\ldots,\vec{n}_l)]_\tau \equiv \mathbb{I}(\tau\neq\sigma) + [\vec{n}_1]_\tau + \cdots + [\vec{n}_l]_\tau. \tag{9}$$

To lighten the notations we keep implicit the dependence of the functions $\widetilde{f}$ and $\widetilde{g}$ on the edges of the factor graph. These equations can be easily solved, for a given initial satisfying assignment $\underline{\sigma}$, noting that the messages from the leaf variable nodes i satisfy the boundary condition $\vec{n}_{i\to a}(\sigma_i) = \vec{o}(\sigma_i)$, where we define $[\vec{o}(\sigma)]_\tau = \mathbb{I}(\sigma\neq\tau)$. The recursions (7) can then be successively applied to determine the value of all messages in a single sweep from the exterior of the graph towards its center. When this is done the m.s.r. for a variable i is obtained from

$$\vec{n}_i(\underline{\sigma}) = \widetilde{g}_{\sigma_i}(\{\vec{n}_{a\to i}(\sigma_i, \underline{\sigma}_{a\to i})\}_{a\in\partial i}). \tag{10}$$

Note that this recursive approach provides not only the size of a minimal rearrangement, but also a final configuration achieving this bound. One just has to bookkeep, along with the size informations encoded in the messages $\vec{n}$, the configuration reaching the minimum in (8) (if there are several of them one is chosen arbitrarily). By construction the support of these optimal rearrangements is connected.

3.1.2 Given Tree, Distribution with Respect to $\underline{\sigma}$

Following the program sketched above, we introduce now a probability distribution μ for the initial solution $\underline{\sigma}$ of the formula:

$$\mu(\underline{\sigma}) = \frac{1}{Z}\prod_a \psi_a(\underline{\sigma}_a)\prod_{i\in B}\eta_{\mathrm{ext},i}(\sigma_i), \tag{11}$$

where Z is a normalization constant, B is a subset of the leaves of the factor graph, and the η_{ext} are probability laws on $\mathcal{X}$ that, by analogy with magnetic systems, we shall call fields. μ vanishes for configurations which do not satisfy the formula; if $B=\emptyset$ it is uniform on the set of solutions, otherwise the external fields η_{ext} can introduce a bias in the law (this possibility will reveal useful in the following). We shall only assume the external fields to be "permissive" enough for the above expression to remain well defined, i.e. they do not put a vanishing weight on the solutions of the formula (even if η_{ext} can be in principle $\{0,1\}$ valued).

The absence of cycles in the factor graph induces a Markovian property of the measure μ which greatly simplifies its characterization. One can indeed compute recursively the marginals of the law on any subset of variable nodes, introducing on each directed edge of the factor graph another family of messages (cavity measures) $\nu_{a\to i}(\sigma_i)$ (resp. $\eta_{i\to a}(\sigma_i)$).

These are the law of σ_i in the measure associated to the cavity factor graph $F_{a\to i}$ (resp. $F_{i\to a}$), and are solutions of

$$\begin{aligned}
&\nu_{a\to i} = f(\{\eta_{j\to a}\}_{j\in\partial a\setminus i}),\\
&f(\{\eta_{j\to a}\}_{j\in\partial a\setminus i})(\sigma_i) = \frac{1}{z(\{\eta_{j\to a}\}_{j\in\partial a\setminus i})}\sum_{\underline{\sigma}_{a\setminus i}}\psi_a(\sigma_i,\underline{\sigma}_{a\setminus i})\prod_{j\in\partial a\setminus i}\eta_{j\to a}(\sigma_j),
\end{aligned} \tag{12}$$

$$\begin{aligned}
&\eta_{i\to a} = g(\{\nu_{b\to i}\}_{b\in\partial i\setminus a}),\\
&g(\{\nu_{b\to i}\}_{b\in\partial i\setminus a})(\sigma_i) = \frac{1}{z(\{\nu_{b\to i}\}_{b\in\partial i\setminus a})}\prod_{b\in\partial i\setminus a}\nu_{b\to i}(\sigma_i),
\end{aligned} \tag{13}$$

where the functions z are defined by normalization. Again for clarity we do not indicate explicitly the dependence of the functions f, g and z on the edges. The boundary conditions are $\eta_{i\to a} = \eta_{\mathrm{ext},i}$ when i is a leaf in B, $\eta_{i\to a} = \overline{\eta}$ (the uniform law on $\mathcal{X}$) if i is a leaf not in B. This set of equations enjoys the same structure as the one on the $\vec{n}$'s (see (7)), and can also be solved in a sweep from the leaves of the factor graph. The marginals of μ for any connected subset of variables can be easily expressed in terms of the solution of this set of equations. For instance the marginal of a single variable reads

$$\mu(\sigma_i) = g(\{\nu_{a\to i}\}_{a\in\partial i})(\sigma_i), \tag{14}$$

while the variables of a constraint, conditioned to the value of one of them, are drawn according to

$$\mu(\underline{\sigma}_{a\setminus i}|\sigma_i;\{\eta_{j\to a}\}_{j\in\partial a\setminus i}) = \frac{1}{z(\sigma_i,\{\eta_{j\to a}\}_{j\in\partial a\setminus i})}\psi_a(\sigma_i,\underline{\sigma}_{a\setminus i})\prod_{j\in\partial a\setminus i}\eta_{j\to a}(\sigma_j), \tag{15}$$

where again z is a normalizing factor.

We have now to compute the distribution of the minimal size rearrangements when the starting configuration $\underline{\sigma}$ is drawn from μ. The generation of $\underline{\sigma}$ can be performed in a recursive broadcasting way: one first draws an arbitrarily chosen root variable σ_i according to its marginal $\mu(\sigma_i)$. Because the factor graph is a tree, the law of the remaining variables factorizes on the different branches around i,

$$\mu(\underline{\sigma}_{\setminus i}|\sigma_i) = \prod_{a\in\partial i}\mu(\underline{\sigma}_{a\to i}|\sigma_i). \tag{16}$$

For each branch $F_{a\to i}$ one proceeds by drawing the variables of $\underline{\sigma}_{a\setminus i}$, conditioned on σ_i (see (15)). Then the value of σ_j for each $j\in\partial a\setminus i$ conditions the generation of $\underline{\sigma}_{j\to a}$, which can itself be broken in subtrees as in (16). This process is repeated outwards until the leaves of the tree are reached.

This observation leads us to introduce the distribution of the $\vec{n}$'s messages with respect to the conditional distributions of the initial configuration,

$$\begin{aligned}
q_{\vec{n}}^{(i\to a,\sigma_i)} &= \sum_{\underline{\sigma}_{i\to a}}\mu(\underline{\sigma}_{i\to a}|\sigma_i)\delta_{\vec{n},\vec{n}_{i\to a}(\sigma_i,\underline{\sigma}_{i\to a})},\\
\widehat{q}_{\vec{n}}^{(a\to i,\sigma_i)} &= \sum_{\underline{\sigma}_{a\to i}}\mu(\underline{\sigma}_{a\to i}|\sigma_i)\delta_{\vec{n},\vec{n}_{a\to i}(\underline{\sigma}_{a\to i})}.
\end{aligned} \tag{17}$$

Combining the recursive computations of the messages $\vec{n}$ expressed in (7) and the recursive generation of the initial configuration $\underline{\sigma}$ leads to

$$\widehat{q}_{\vec{n}}^{(a\to i,\sigma_i)} = \sum_{\underline{\sigma}_{a\setminus i}} \mu(\underline{\sigma}_{a\setminus i}|\sigma_i;\{\eta_{j\to a}\}) \prod_{j\in\partial a\setminus i} \sum_{\vec{n}_{j\to a}} q_{\vec{n}_{j\to a}}^{(j\to a,\sigma_j)} \delta_{\vec{n},\widetilde{f}(\{\vec{n}_{j\to a}\})}, \tag{18}$$

$$q_{\vec{n}}^{(i\to a,\sigma_i)} = \prod_{b\in\partial i\setminus a} \sum_{\vec{n}_{b\to i}} \widehat{q}_{\vec{n}_{b\to i}}^{(b\to i,\sigma_i)} \delta_{\vec{n},\widetilde{g}_{\sigma_i}(\{\vec{n}_{b\to i}\})}, \tag{19}$$

with the boundary condition given by $q_{\vec{n}}^{(i\to a,\sigma_i)} = \delta_{\vec{n},\vec{o}(\sigma_i)}$ for the leaves i. The distribution of the m.s.r. for i when $\underline{\sigma}$ is drawn from μ can then be obtained from the distributions on the edges neighboring i,

$$q_n^{(i)} = \sum_{\sigma_i} \mu(\sigma_i) \sum_{\vec{n}} q_{\vec{n}}^{(i,\sigma_i)} \delta_{n,\min\limits_{\tau_i\neq\sigma_i}[\vec{n}]_{\tau_i}}, \qquad q_{\vec{n}}^{(i,\sigma_i)} = \prod_{a\in\partial i} \sum_{\vec{n}_{a\to i}} \widehat{q}_{\vec{n}_{a\to i}}^{(a\to i,\sigma_i)} \delta_{\vec{n},\widetilde{g}_{\sigma_i}(\{\vec{n}_{a\to i}\})}. \tag{20}$$

3.1.3 Average over the Choice of the Tree

At this point we define an ensemble of random rooted tree factor graphs on which we shall perform the average of the m.s.r. distribution. The ingredients of the definition are p_l, a distribution on the positive integers, $\rho(\psi)$ a distribution on the 0/1 constraint functions (with possibly a random degree k), and a distribution of fields $\mathcal{P}(\eta)$. Let us denote $\mathbb{T}_L$ a random tree of the ensemble of depth L, and for notational simplicity $\widehat{\mathbb{T}}_L$ the elements of this ensemble conditioned on their root being of degree one. $\mathbb{T}_L$ is defined by induction on L as a (Galton–Watson like) branching process. $\mathbb{T}_0$ is made of a single variable node (the root) to which is applied an external field η drawn from $\mathcal{P}$. $\widehat{\mathbb{T}}_L$ is generated by introducing a root variable node i, connected to a single interaction node a whose constraint function ψ_a is drawn from ρ. Then each variable node in $\partial a\setminus i$ is taken to be the root of an independently generated $\mathbb{T}_L$. Conversely $\mathbb{T}_{L+1}$ is made by identifying the roots of l (a random integer drawn from p_l) independent copies of $\widehat{\mathbb{T}}_L$.

For each tree drawn from this ensemble the two recursive computations yield a set of messages on each edge of the factor graph directed towards the root, $(\eta,\{q_{\vec{n}}^{(\sigma)}\}_{\sigma=1}^q)$ for an edge from a variable to a constraint, $(\nu,\{\widehat{q}_{\vec{n}}^{(\sigma)}\}_{\sigma=1}^q)$ from a constraint to a variable. The randomness in the definition of the tree turn these objects into random variables, whose distribution depends only on the distance between the considered edge and the leaves. To be more precise, let us call $\mathfrak{P}_L(\eta,\{q_{\vec{n}}^{(\sigma)}\})$ the distribution of $(\mu(\sigma_i),\{q_{\vec{n}}^{(i,\sigma_i)}\})$ when i is the root of a random $\mathbb{T}_L$ tree, and similarly $\widehat{\mathfrak{P}}_L(\nu,\{\widehat{q}_{\vec{n}}^{(\sigma)}\})$ for the distribution of the messages directed to the root variable node of $\widehat{\mathbb{T}}_L$.

One can first notice that the recursion between the messages η,ν do not involve the size distributions $q_{\vec{n}}$ and $\widehat{q}_{\vec{n}}$, and thus define $\mathcal{P}_L(\eta)$ as the marginal of $\mathfrak{P}_L$ disregarding the $q_{\vec{n}}$'s, and similarly $\widehat{\mathcal{P}}_L(\nu)$ from $\widehat{\mathfrak{P}}_L$. $\mathcal{P}_L$ and $\widehat{\mathcal{P}}_L$ obey functional equations of the form $\widehat{\mathcal{P}}_L = F[\mathcal{P}_L]$, $\mathcal{P}_{L+1} = G[\widehat{\mathcal{P}}_L]$, with $\mathcal{P}_{L=0} = \mathcal{P}$, and where the functionals F and G have a compact distributional writing,

$$\nu \overset{\mathrm{d}}{=} f(\eta_1,\ldots,\eta_{k-1},\psi), \qquad \eta \overset{\mathrm{d}}{=} g(\nu_1,\ldots,\nu_l). \tag{21}$$

The first equation means that drawing a variable ν from $\widehat{\mathcal{P}}_L$ amounts to drawing a constraint function ψ from ρ, $k-1$ i.i.d. variables η_i from $\mathcal{P}_L$ and computing ν from (12). Similarly $\mathcal{P}_{L+1}$ is obtained from $\widehat{\mathcal{P}}_L$ thanks to (13), with the branching number l drawn from p_l. In the

following we shall assume that the distribution $\mathcal{P}$ on the boundary of the tree is a solution of the fixed point functional equation $\mathcal{P} = G[F[\mathcal{P}]]$. This implies a stationarity property with respect to the number of generation L, $\mathcal{P}_L = \mathcal{P}$, $\widehat{\mathcal{P}}_L = \widehat{\mathcal{P}} = F[\mathcal{P}]$. This justifies a posteriori the choice we made of including non-trivial biases at the boundary in the law (11): in generic models unbiased boundary conditions represented by $\mathcal{P}(\eta) = \delta(\eta - \overline{\eta})$ do not satisfy this stationary property, this will be in particular the case for the random k-SAT problem studied below.

The evolution of the size distributions when iterating the tree construction is coupled, through the term $\mu(\underline{\sigma}_{a\setminus i}|\sigma_i)$ of (18), to the η, ν messages. We are however interested in a rather simple quantity, the average of the m.s.r. distribution of the root (see (20)) with respect to the random tree. It is thus possible to compute an average of the $q_{\vec{n}}^{(i\to a,\sigma_i)}$ on an edge of depth L, provided this average is *conditioned* on the value of the associated message $\eta_{i\to a}$. This conditional average, denoted $q_{\vec{n}}^{(\sigma,L)}(\eta)$, and its counterpart $\widehat{q}_{\vec{n}}^{(\sigma,L)}(\nu)$, are then found to obey the following equations,

$$\begin{aligned}\widehat{q}_{\vec{n}}^{(\sigma,L)}(\nu)\widehat{\mathcal{P}}(\nu) = \mathbb{E}_\psi \int d\mathcal{P}(\eta_1)\dots d\mathcal{P}(\eta_{k-1})\delta(\nu - f(\eta_1,\dots,\eta_{k-1},\psi)) \\ \times \sum_{\sigma_1,\dots,\sigma_{k-1}} \mu(\sigma_1,\dots,\sigma_{k-1}|\sigma,\eta_1,\dots,\eta_{k-1},\psi) \\ \times \sum_{\vec{n}_1,\dots,\vec{n}_{k-1}} q_{\vec{n}_1}^{(\sigma_1,L)}(\eta_1)\dots q_{\vec{n}_{k-1}}^{(\sigma_{k-1},L)}(\eta_{k-1})\delta_{\vec{n},\widetilde{f}(\vec{n}_1,\dots,\vec{n}_{k-1},\psi)}, \qquad (22)\end{aligned}$$

$$\begin{aligned}q_{\vec{n}}^{(\sigma,L+1)}(\eta)\mathcal{P}(\eta) = \sum_l p_l \int d\widehat{\mathcal{P}}(\nu_1)\dots d\widehat{\mathcal{P}}(\nu_l)\delta(\eta - g(\nu_1,\dots,\nu_l)) \\ \times \sum_{\vec{n}_1,\dots,\vec{n}_l} \widehat{q}_{\vec{n}_1}^{(\sigma,L)}(\nu_1)\dots\widehat{q}_{\vec{n}_l}^{(\sigma,L)}(\nu_l)\delta_{\vec{n},\widetilde{g}_\sigma(\vec{n}_1,\dots,\vec{n}_l)}, \qquad (23)\end{aligned}$$

with the boundary condition $q_{\vec{n}}^{(\sigma,L=0)}(\eta) = \delta_{\vec{n},\vec{o}(\sigma)}$. Finally the sought-for average m.s.r.d. for the root of a random tree of depth L reads:

$$q_n^{(L)} = \int d\mathcal{P}(\eta)\sum_\sigma \eta(\sigma)\sum_{\vec{n}} q_{\vec{n}}^{(\sigma,L)}(\eta)\delta_{n,\min_{\tau\neq\sigma}[\vec{n}]_\tau}. \qquad (24)$$

The numerical resolution of (22, 23) could at first sight seem rather difficult, as they involve, for each value of the random variable η (or ν), q distributions of vectors $\vec{n}$. One can however devise a simple method, generalizing the population dynamics algorithm of [30]. The important point is to notice that for a given value of σ, $q_{\vec{n}}^{(\sigma,L)}(\eta)\mathcal{P}(\eta)$ can be viewed as a joint distribution of variables $(\eta, \vec{n}^{(\sigma)})$, which can be numerically represented by a population of a large number $\mathcal{N}$ of couples $\{(\eta_i, \vec{n}_i^{(\sigma)})\}_{i=1}^{\mathcal{N}}$. The empirical distribution of these couples is taken as an approximation (known as a particle approximation in the statistics literature) of $q_{\vec{n}}^{(\sigma,L)}(\eta)\mathcal{P}(\eta)$. This suggests the following algorithm. Initialize a population $\{\eta_i\}_{i=1}^{\mathcal{N}}$ drawn i.i.d. from $\mathcal{P}$ (this shall be itself performed by a standard population dynamics approach), and associate to each of them q vectors, $\vec{n}_i^{(\sigma)} = \vec{o}(\sigma)$. We thus have, for trees of depth $L = 0$, a population $\{(\eta_i, \vec{n}_i^{(1)},\dots,\vec{n}_i^{(q)})\}_{i=1}^{\mathcal{N}}$. To take this population from depth L to depth $L+1$ one has to:

– generate in an i.i.d. way $\mathcal{N}$ elements $(\nu_j, \vec{n}_j^{(1)},\dots,\vec{n}_j^{(q)})$, with $j \in [\mathcal{N}+1, 2\mathcal{N}]$ to avoid notational confusion, by:

- choosing randomly a constraint function ψ from ρ, and $k-1$ indices $i_1,\dots,i_{k-1}$ uniformly at random in $[1,\mathcal{N}]$;
- computing $\nu_j = f(\eta_{i_1},\dots,\eta_{i_{k-1}},\psi)$;
- for each $\sigma\in[1,q]$:
  - ∗ generating a configuration $(\sigma_1,\dots,\sigma_{k-1})$ according to the law $\mu(\cdot|\sigma,\eta_{i_1},\dots,\eta_{i_{k-1}},\psi)$;
  - ∗ computing $\vec{n}_j^{(\sigma)} = \widetilde{f}(\vec{n}_{i_1}^{(\sigma_1)},\dots,\vec{n}_{i_{k-1}}^{(\sigma_{k-1})},\psi)$.

– then generate a new population $\{(\eta_i,\vec{n}_i^{(1)},\dots,\vec{n}_i^{(q)})\}_{i=1}^{\mathcal{N}}$, repeating for each $i\in[1,\mathcal{N}]$ independently the following steps:
- Choose randomly a degree l from p_l and l indices $j_1,\dots,j_l$ uniformly at random in $[\mathcal{N}+1,2\mathcal{N}]$.
- Compute $\eta_i = g(\nu_{j_1},\dots,\nu_{j_l})$.
- For each $\sigma\in[1,q]$, compute $\vec{n}_i^{(\sigma)} = \widetilde{g}_\sigma(\vec{n}_{j_1}^{(\sigma)},\dots,\vec{n}_{j_l}^{(\sigma)})$.

After L iterations of these two steps, for a given value of σ, an element $(\eta_i,\vec{n}_i^{(\sigma)})$ with i uniformly chosen in $[1,\mathcal{N}]$ is distributed with the joint law $q_{\vec{n}}^{(\sigma,L)}(\eta)\mathcal{P}(\eta)$.[4] We can thus complete the computation of $q_n^{(L)}$ in terms of a weighted histogram,

$$q_n^{(L)} = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}\sum_{\sigma=1}^{q}\eta_i(\sigma)\delta_{n,\min\limits_{\tau\neq\sigma}[\vec{n}_i^{(\sigma)}]_\tau}. \tag{25}$$

We shall now examine how this general formalism can be applied to the three exemplar problems of XORSAT, COL and SAT.

3.2 k-XORSAT

3.2.1 On a Given Tree Factor Graph

Let us recall the factor graph representation of a k-XORSAT formula we use: the variables are Ising spins $\sigma_i=\pm1$, and each constraint node a is satisfied if and only if the product of its k neighboring variables $\prod_{i\in\partial a}\sigma_i$ is equal to a given constant $J_a=\pm1$. The computation of the m.s.r., already performed in [27], is much simpler than the general case presented above. Note first that for any CSP where variable can only take two values, a rearrangement $\underline{\sigma}\to\underline{\tau}$ is completely specified by its support, the set R of variables which are different in the initial and final configurations. A second simplification is specific to the XORSAT problem. Consider an initial solution $\underline{\sigma}$ and the configuration $\underline{\tau}$ obtained by flipping the variables in R. This second configuration is also a solution if and only if for each constraint a, an even (possibly null) number of variables of ∂a are in R. A rearrangement for the variable i is hence a set R verifying this condition and containing i. The m.s.r. n_i is the minimal cardinality of such a set of variables; on a tree this minimum can always be achieved requiring that each a contains either zero or two (and not an higher even value) variables of R. The recursive strategy for the computation of n_i and the construction of a rearrangement of this size amounts to constructing a m.s.r. $R_{a\to i}$ for all the branches $F_{a\to i}$ around i (their sizes being denoted $1+n_{a\to i}$) and to combining the rearrangements of the sub-factor graphs,

[4] We do not claim that $(\eta_i,\vec{n}_i^{(1)},\dots,\vec{n}_i^{(q)})$ is drawn according to $\mathcal{P}(\eta)q_{\vec{n}^{(1)}}^{(1,L)}\dots q_{\vec{n}^{(q)}}^{(q,L)}$, i.e. that the $\vec{n}_i^{(\sigma)}$ are independent conditionally on η_i, which is not true. The algorithm induces correlations between the various values of σ, yet these are irrelevant for the linear averages we compute.

$R = \{i\} \bigcup_{a \in \partial i} R_{a \to i}$. To construct $R_{a \to i}$ one has to choose exactly one variable $j \in \partial a \setminus i$ that minimizes the cost $n_{j \to a}$ of the rearrangement in the branch $F_{j \to a}$. Summarizing this reasoning in formulas, we obtain:

$$n_{a \to i} = \min_{j \in \partial a \setminus i} n_{j \to a}, \qquad n_{i \to a} = 1 + \sum_{b \in \partial i \setminus a} n_{b \to i}, \qquad n_i = 1 + \sum_{a \in \partial i} n_{a \to i}. \tag{26}$$

The reader will easily verify that (7, 8, 9, 10) of the general formalism reduce indeed to this simple form, noting in particular that the m.s.r. is here independent of the initial configuration, as appears clearly from the geometric characterization of the optimal supports R.

3.2.2 Random Tree

This independence with respect to the initial configuration allows to skip the second step of the general formalism, as for a given tree the distribution of the m.s.r. is trivially concentrated on a single integer, and to study directly the ensemble of random tree formula. We shall follow the general definition of $\mathbb{T}_L$ given above, with a Poisson law of parameter αk for the branching probability p_l, and all constraint nodes of degree k. For definiteness one can assume that the boundary condition is free (no bias on the leaves of the tree) and that $J_a = \pm 1$ with probability one half; these last two choices are in fact irrelevant, as the m.s.r. depends only on the geometry of the factor graph.

This random ensemble induces a probability law $q_n^{(L)}$ for the m.s.r. of the root of $\mathbb{T}_L$, and an associated law $\widehat{q}_n^{(L)}$ for the message sent to the root of $\widehat{\mathbb{T}}_L$. Simplifying (22, 23, 24) of the general formalism, or interpreting the specific ones (26) in a distributional sense, leads to

$$\widehat{q}_n^{(L)} = \sum_{n_1, \ldots, n_{k-1}} q_{n_1}^{(L)} \cdots q_{n_{k-1}}^{(L)} \delta_{n, \min[n_1, \ldots, n_{k-1}]}, \tag{27}$$

$$q_n^{(L+1)} = \sum_{l=0}^{\infty} \frac{e^{-\alpha k} (\alpha k)^l}{l!} \sum_{n_1, \ldots, n_l} \widehat{q}_{n_1}^{(L)} \cdots \widehat{q}_{n_l}^{(L)} \delta_{n, 1 + n_1 + \cdots + n_l}, \tag{28}$$

with the initial condition $q_n^{(L=0)} = \delta_{n,1}$.

These equations can be solved by a simplified version of the population dynamics algorithm introduced in the general case. The distributions $q_n^{(L)}$ and $\widehat{q}_n^{(L)}$ are represented by samples of integers $\{n_i\}$, each element of the population associated to $q_n^{(L+1)}$ is generated by drawing a Poisson distributed integer l, extracting at random l elements of the sample representing $\widehat{q}_n^{(L)}$ and computing their sum plus one. Conversely the elements of $\widehat{q}_n^{(L)}$ are the minimum of $k-1$ randomly chosen integers drawn from the population encoding $q_n^{(L)}$. In the following we shall be interested in the $L \to \infty$ limit, which is the counterpart of the $N \to \infty$ thermodynamic limit of the original random graph ensembles. One could reach it numerically by repeated iterations of the population dynamics step. There is however a simpler numerical method which allows to perform analytically this limit.

Let us first define the integrated version of the m.s.r.d.,

$$Q_n^{(L)} = \sum_{n' \geq n} q_{n'}^{(L)}, \tag{29}$$

which gives the probability of a m.s.r. being larger than n. A few simple properties follow from this definition,

$$q_n^{(L)} = Q_n^{(L)} - Q_{n+1}^{(L)}, \qquad Q_n^{(L)} = 1 - \sum_{n'<n} q_{n'}^{(L)}, \qquad \lim_{n\to\infty} Q_n^{(L)} = 0. \tag{30}$$

A slightly less obvious property is that, for a fixed value of n, $Q_n^{(L)}$ is monotonously increasing with L. This arises from the fact that larger trees have larger rearrangements, and can be proven from (27, 28) via a standard stochastic domination argument [31]. Being moreover bounded from above by 1, $Q_n^{(L)}$ converges as L goes to infinity, to a limit we shall denote Q_n. By continuity in the first equality of (30) the limit q_n of $q_n^{(L)}$ also exists; same statements apply to $\widehat{Q}_n$ and $\widehat{q}_n$. Equation (27) can be rewritten as $\widehat{Q}_n^{(L)} = (Q_n^{(L)})^{k-1}$, in the infinite L limit we thus obtain:

$$\widehat{Q}_n = Q_n^{k-1}, \tag{31}$$

$$q_n = \sum_{l=0}^{\infty} \frac{e^{-\alpha k}(\alpha k)^l}{l!} \sum_{n_1,\ldots,n_l} \widehat{q}_{n_1} \ldots \widehat{q}_{n_l} \delta_{n,1+n_1+\cdots+n_l}. \tag{32}$$

These limit distributions can now be determined by a recursion on n. Equation (28) implies that $q_1^{(L)} = e^{-\alpha k}$ for all $L \geq 1$; hence $q_1 = e^{-\alpha k}$, which fixes the starting point of the recursion. Assume q_n has been computed up to rank m. This means that $Q_n = 1 - \sum_{n'<n} q_{n'}$ is known up to rank $m+1$, and the same is true for $\widehat{Q}_n$ because of (31). We thus have at our disposal the values of $\widehat{q}_n$ up to $n = m$, which allows the computation of q_{m+1} through (32). We defer the presentation of the numerical results obtained in this way until Sect. 4, in order to confront them with the COL and SAT problems.

Let us only anticipate one feature by emphasizing that the limit $L \to \infty$ was taken here at a fixed value of n. We shall see that for some values of α the limits $L, n \to \infty$ do not commute, a situation reminiscent of a percolating regime. In such cases Q_n tends for large n to a strictly positive value ϕ, q_n is not normalized anymore and cannot be directly considered as the distribution of an integer random variable n. It will be however convenient to formally consider n as an extended integer, with a probability ϕ of being infinite.

3.3 q-COL

3.3.1 Given Tree, Given $\underline{\sigma}$

The second example of CSP we shall consider is the q-coloring problem. The variables σ_i can take one of the q values (colors) in $\{1, \ldots, q\}$, and the constraint node a linking two variables i and j forbids the configurations with $\sigma_i = \sigma_j$. The solutions of this CSP are thus the proper colorings of the underlying graph.

At variance with the XORSAT problem, the m.s.r. does depend on the initial satisfying assignment: take for instance a small graph made of a central site i with $q-1$ neighbors. If in the initial coloring all the peripheral sites have distinct colors, the minimal size to rearrange i is two. Otherwise, if at least two peripheral sites have the same color, there is one color available for the central site to be rearranged without modifying its neighborhood.

There is however room for simplifications with respect to the general formalism. Consider the constraint a between two adjacent vertices i and j. The vectorial message $\vec{n}_{a\to i}(\underline{\sigma}_{a\to i})$ has only one non-zero component, corresponding to the perturbation $\sigma_i \to \tau_i = \sigma_j$. This is a formal consequence of (8), but has a very intuitive meaning: in the

cavity graph $F_{a\to i}$ the root σ_i can be given any value $\tau_i \neq \sigma_j$ without having to propagate the rearrangement. We can thus get rid of the vectorial character of the messages. Note also that the information contained in the messages $\vec{n}_{a\to i}$ and $\vec{n}_{j\to a}$ is redundant, as each constraint node involves only two variables. We shall thus eliminate the variable to constraint messages, and rename $n_{j\to i}(\underline{\sigma}_{j\to i})$ what was denoted in the general formalism $[\vec{n}_{a\to i}(\underline{\sigma}_{a\to i})]_{\sigma_j}$. Simplifying (7–10) with these new notations, we obtain

$$n_{j\to i}(\underline{\sigma}_{j\to i}) = 1 + \min_{\tau_j\neq\sigma_j}\left\{\sum_{k\in\partial j\setminus i}\delta_{\sigma_k,\tau_j} n_{k\to j}(\underline{\sigma}_{k\to j})\right\}, \tag{33}$$

$$n_i(\underline{\sigma}) = 1 + \min_{\tau_i\neq\sigma_i}\left\{\sum_{j\in\partial i}\delta_{\sigma_j,\tau_i} n_{j\to i}(\underline{\sigma}_{j\to i})\right\}, \tag{34}$$

with $n_{j\to i}(\sigma_j) = 1$ if j is a leaf of the tree. The interpretation of these equations is clear: to modify the color σ_i of a vertex i in a coloring $\underline{\sigma}$ one has to probe the $q-1$ possibilities of $\tau_i \neq \sigma_i$, and follow the effect of this modification in the branches $F_{j\to i}$ that become unsatisfied, i.e. those who had $\sigma_j = \tau_i$ before the modification.

3.3.2 Given Tree, Distribution with Respect to $\underline{\sigma}$

We shall study in the coloring case the distribution of the m.s.r. with respect to the measure $\mu(\underline{\sigma})$ uniform on the proper colorings. In other words we use a free boundary condition and do not impose any external field on the leaves. This choice preserves the permutation symmetry among colors, which implies that the marginal distribution $\mu(\sigma_i)$ of any variable i is uniform over the q possible values. Once the color of an arbitrary root variable i has been chosen, the generation of the remaining sites can be done in a recursive way: the colors of the neighbors of i are drawn independently, uniformly over the $q-1$ colors distinct from σ_i, and this process is repeated from i outwards. Exploiting this symmetry and the recursions (33, 34), one finds that the distributions of the m.s.r. with respect to the uniform choice of the initial proper coloring is given by

$$q_n^{(i)} = \frac{1}{(q-1)^{|\partial i|}}\prod_{j\in\partial i}\sum_{\substack{\sigma_j\neq 1\\ n_{j\to i}}} q_{n_{j\to i}}^{(j\to i)}\mathbb{I}\left(n = 1 + \min_{\sigma\neq 1}\left[\sum_{j\in\partial i}\delta_{\sigma,\sigma_j} n_{j\to i}\right]\right), \tag{35}$$

where the distributions of the messages on the edges of the tree are solutions of

$$q_n^{(j\to i)} = \frac{1}{(q-1)^{|\partial j|-1}}\prod_{k\in\partial j\setminus i}\sum_{\substack{\sigma_k\neq 1\\ n_{k\to j}}} q_{n_{k\to j}}^{(k\to j)}\mathbb{I}\left(n = 1 + \min_{\sigma\neq 1}\left[\sum_{k\in\partial j\setminus i}\delta_{\sigma,\sigma_k} n_{k\to j}\right]\right), \tag{36}$$

with the boundary condition $q_n^{(j\to i)} = \delta_{n,1}$ when j is a leaf.

3.3.3 Average over the Choice of the Tree

We now consider the ensemble of random trees $\mathbb{T}_L$ where the variable nodes have a Poissonian branching probability of mean c, and all constraint nodes are identical, $\psi(\sigma_i,\sigma_j) = \mathbb{I}(\sigma_i \neq \sigma_j)$. One can easily show from (35, 36) that the m.s.r.d. for uniformly distributed

initial proper colorings, averaged over this random tree ensemble is given by

$$q_n^{(L+1)} = \sum_{l=0}^{\infty} \frac{e^{-c}c^l}{l!} \frac{1}{(q-1)^l} \sum_{\sigma_1,\ldots,\sigma_l=2}^{q} \sum_{n_1,\ldots,n_l} q_{n_1}^{(L)} \ldots q_{n_l}^{(L)} \mathbb{I}\left(n = 1 + \min_{\sigma=2,\ldots,q}\left[\sum_{i=1}^{l} \delta_{\sigma,\sigma_i} n_i\right]\right), \tag{37}$$

with $q_n^{(L=0)} = \delta_{n,1}$. This equation could be solved following the population dynamics approach explained above. One can however unveil a formal equivalence with the computation performed for the XORSAT problem. Consider indeed the random variables l_σ which counts in (37) the number of σ_i's assigned to the value σ. Conditional on l the l_σ's are multinomially distributed; as l is itself a Poisson random variable the l_σ turn out to be independent Poisson random variables. This allows to rewrite (37) as

$$q_n^{(L+1)} = \sum_{l_2,\ldots,l_q=0}^{\infty} \frac{e^{-c}(c/(q-1))^{l_2+\cdots+l_q}}{l_2!\ldots l_q!} \sum_{m_2,\ldots,m_q} \delta_{n,\min[m_2,\ldots,m_q]} \times \prod_{\sigma=2}^{q}\left(\sum_{n_\sigma^1,\ldots,n_\sigma^{l_\sigma}} q_{n_\sigma^1}^{(L)} \ldots q_{n_\sigma^{l_\sigma}}^{(L)} \delta_{m_\sigma,1+n_\sigma^1+\cdots+n_\sigma^{l_\sigma}}\right). \tag{38}$$

Comparing with (27, 28) one realizes that the solution of the coloring case can be directly read off from the study of the XORSAT one with a simple translation of the parameters,

$$q_n^{(L,\mathrm{COL})}[q,c] = \widehat{q}_n^{(L,\mathrm{XORSAT})}\left[k = q, \alpha = \frac{c}{q(q-1)}\right]. \tag{39}$$

In particular the simple recursion on n to solve directly in the $L \to \infty$ limit is still applicable to the coloring problem.

3.4 k-SAT

3.4.1 Given Tree, Given $\underline{\sigma}$

We consider now the third example of CSP, in which the factor graph encodes a k-satisfiability formula. The boolean variables are represented by Ising spins $\sigma_i = \pm 1$; each constraint node a is linked to k variable nodes, and is unsatisfied if and only if these k variable all takes their unsatisfying value, $\sigma_i = J_i^a$ for all $i \in \partial a$. We shall denote $\partial_+ i(a)$ (resp. $\partial_- i(a)$) the set of clauses in $\partial i \setminus a$ agreeing (resp. disagreeing) with a on the satisfying value of σ_i. We also denote $\partial_\sigma i$ the set of clauses in ∂i which are satisfied by $\sigma_i = \sigma$.

Because of the boolean nature of the variables a rearrangement is specified by the set of variables to be flipped (recall the discussion of the XORSAT problem), we can get rid of the vectorial character of the general formalism and denote, for instance, $n_i(\underline{\sigma})$ for the m.s.r. of the variable i under the perturbation $\sigma_i \to \tau_i = -\sigma_i$. This quantity does depend on the initial satisfying assignment. In the simplest case where there is one single constraint node a in the factor graph, $n_i(\underline{\sigma}) = 2$ if a was satisfied only by i before its flip, $n_i(\underline{\sigma}) = 1$ for all the other satisfying assignments. Generalizing this observation to generic factor graphs, one reduces the recursion relations of the general formalism (see (7–10)) to:

$$n_{a\to i}(\sigma_i, \underline{\sigma}_{a\to i}) = \begin{cases} \min_{j\in\partial a\setminus i} n_{j\to a}(\sigma_j, \underline{\sigma}_{j\to a}) & \text{if } \sigma_i = -J_i^a \text{ and } \sigma_j = J_j^a \ \forall j \in \partial a \setminus i, \\ 0 & \text{otherwise,} \end{cases} \tag{40}$$

$$n_{i\to a}(\sigma_i, \underline{\sigma}_{i\to a}) = 1 + \sum_{b\in\partial i\setminus a} n_{b\to i}(\sigma_i, \underline{\sigma}_{b\to i}), \tag{41}$$

$$n_i(\underline{\sigma}) = 1 + \sum_{a\in\partial i} n_{a\to i}(\sigma_i, \underline{\sigma}_{a\to i}), \tag{42}$$

with again $n_{i\to a}(\sigma_i) = 1$ for the leaves of the graph.

3.4.2 Given Tree, Distribution with Respect to $\underline{\sigma}$

We now consider the probability law $\mu(\underline{\sigma})$ on the initial satisfying assignments, with external fields on some of the leaves of the graph. More precisely, we use the form (11), with the biases on a subset B of the leaves parameterized by a real $h_{\mathrm{ext},i}$:

$$\eta_{\mathrm{ext},i}(\sigma_i) = \frac{1+\sigma_i \tanh h_{\mathrm{ext},i}}{2}. \tag{43}$$

The messages $\nu_{a\to i}$ and $\eta_{i\to a}$ are probability laws of Ising spins and can thus be parameterized by a single real. To simplify the notations we make a gauge transformation with respect to the value of the variable satisfying the clause and define

$$\nu_{a\to i}(\sigma_i) = \frac{1-J_a^i\sigma_i \tanh u_{a\to i}}{2}, \qquad \eta_{i\to a}(\sigma_i) = \frac{1-J_a^i\sigma_i \tanh h_{i\to a}}{2}. \tag{44}$$

With these conventions (12, 13) become

$$u_{a\to i} = f(\{h_{j\to a}\}_{j\in\partial a\setminus i}), \qquad f(h_1,\dots,h_{k-1}) = -\frac{1}{2}\ln\left(1-\prod_{i=1}^{k-1}\frac{1-\tanh h_i}{2}\right), \tag{45}$$

$$h_{i\to a} = \sum_{b\in\partial_+ i(a)} u_{b\to i} - \sum_{b\in\partial_- i(a)} u_{b\to i}, \tag{46}$$

with $h_{i\to a} = -J_i^a h_{\mathrm{ext},i}$ if i is a leaf in B, 0 if it is a leaf not in B. The solution of these equations allows to compute the two quantities that we shall need below:

- the marginal law of σ_i,

$$\mu(\sigma_i) = \frac{1+\sigma_i \tanh h_i}{2}, \qquad h_i = \sum_{a\in\partial_+ i} u_{a\to i} - \sum_{a\in\partial_- i} u_{a\to i}; \tag{47}$$

- the probability that, conditional on σ_i satisfying the constraint a, all other variables in ∂a take their wrong values,

$$\mu(\sigma_j = J_j^a\ \forall j\in\partial a\setminus i | \sigma_i = -J_i^a) = \prod_{j\in\partial a\setminus i}\frac{1-\tanh h_{j\to a}}{2}. \tag{48}$$

We now proceed with the introduction of the distributions $\widehat{q}_n^{(a\to i,\sigma_i)}$ (resp. $q_n^{(i\to a,\sigma_i)}$) of the messages $n_{a\to i}(\sigma_i, \underline{\sigma}_{a\to i})$ (resp. $n_{i\to a}(\sigma_i, \underline{\sigma}_{i\to a})$) when $\underline{\sigma}_{a\to i}$ (resp. $\underline{\sigma}_{i\to a}$) is drawn conditionally on σ_i. In fact for each directed edge the distribution corresponding to one of the two values of σ_i can be discarded. Consider first the cavity factor graph $F_{a\to i}$. If $\underline{\sigma}_{a\to i}$ is drawn conditionally on σ_i not satisfying constraint a, necessarily one of the $k-1$ other variables of a will satisfy it so that σ_i can be flipped without propagating the rearrangement further in the

branch. This is translated in formula as $\widehat{q}_n^{(a\to i,\sigma_i=J_i^a)}=\delta_{n,0}$, we shall thus simplify notation and write $\widehat{q}_n^{(a\to i)}$ instead of $\widehat{q}_n^{(a\to i,\sigma_i=-J_i^a)}$ for the only non-trivial size distribution born by the edge $a\to i$. This last quantity, in virtue of (18), has to be expressed in terms of the distributions $q_n^{(j\to a,\sigma_j)}$ for $j\in\partial a\setminus i$. However the rearrangement has to be propagated only if none of these variables were satisfying constraint a, we can thus rename $q_n^{(j\to a)}\equiv q_n^{(j\to a,\sigma_j=J_j^a)}$ and forget about $q_n^{(j\to a,\sigma_j=-J_j^a)}$. Collecting these various observations we obtain

$$\widehat{q}_n^{(a\to i)}=\prod_{j\in\partial a\setminus i}\sum_{n_{j\to a}}q_{n_{j\to a}}^{(j\to a)}\Bigg[\Bigg(1-\prod_{j\in\partial a\setminus i}\frac{1-\tanh h_{j\to a}}{2}\Bigg)\delta_{n,0}+\Bigg(\prod_{j\in\partial a\setminus i}\frac{1-\tanh h_{j\to a}}{2}\Bigg)\delta_{n,\min\{n_{j\to a}\}}\Bigg],\tag{49}$$

$$q_n^{(i\to a)}=\prod_{b\in\partial_- i(a)}\sum_{n_{b\to i}}\widehat{q}_{n_{b\to i}}^{(b\to i)}\delta_{n,1+\sum n_{b\to i}},\tag{50}$$

with $q_n^{(i\to a)}=\delta_{n,1}$ on the leaves. Finally the law of the m.s.r. for i is given by

$$q_n^{(i)}=\sum_{\sigma}\frac{1+\sigma\tanh h_i}{2}\prod_{a\in\partial_\sigma i}\sum_{n_{a\to i}}\widehat{q}_{n_{a\to i}}^{(a\to i)}\delta_{n,1+\sum n_{a\to i}}.\tag{51}$$

3.4.3 Average over the Choice of the Tree

We shall study random trees $\mathbb{T}_L$ with a Poissonian law of mean αk for the branching probability p_l of variable nodes. The constraint nodes are all of degree k with the signs J_i^a of the unsatisfying literals i.i.d. random variables equal to ±1 with equal probability. This implies that the cardinality of the neighborhoods $\partial_+ i$ and $\partial_- i$ of the root are two independent Poisson random variables of mean $\alpha k/2$, whose law shall be denoted p_{l_+,l_-}. The same statement is true for $\partial_+ i(a)$ and $\partial_- i(a)$ in the bulk of the tree. The last element defining $\mathbb{T}_L$ is the distribution $\mathcal{P}(h)$ for the biases on the leaves of depth L of the tree. Following the general formalism we assume this distribution to be stationary under the iterations

$$u\stackrel{\mathrm{d}}{=}f(h_1,\dots,h_{k-1}),\qquad h\stackrel{\mathrm{d}}{=}\sum_{i=1}^{l_+}u_i^+-\sum_{i=1}^{l_-}u_i^-,\tag{52}$$

where $l_\pm$ are drawn from p_{l_+,l_-} and the h_i (resp. the $u_i^\pm$) are independent copies drawn from $\mathcal{P}(h)$ (resp. $\widehat{\mathcal{P}}(u)$). The computation proceeds with the introduction of $q_n^{(L)}(h)$ (resp. $\widehat{q}_n^{(L)}(u)$), the average of the $q_n^{(i\to a)}$ (resp. $\widehat{q}_n^{(a\to i)}$) conditioned by the event $h_{i\to a}=h$ (resp. $u_{a\to i}=u$). The generic equations (22, 23) translate into

$$\widehat{q}_n^{(L)}(u)\widehat{\mathcal{P}}(u)=\int\prod_{i=1}^{k-1}d\mathcal{P}(h_i)\delta(u-f(h_1,\dots,h_{k-1}))\times\sum_{n_1,\dots,n_{k-1}}\prod_{i=1}^{k-1}q_{n_i}^{(L)}(h_i)\Bigg[\Bigg(1-\prod_{i=1}^{k-1}\frac{1-\tanh h_i}{2}\Bigg)\delta_{n,0}+\Bigg(\prod_{i=1}^{k-1}\frac{1-\tanh h_i}{2}\Bigg)\delta_{n,\min[n_1,\dots,n_{k-1}]}\Bigg],\tag{53}$$

$$q_n^{(L+1)}(h)\mathcal{P}(h) = \sum_{l_+,l_-=0}^{\infty} p_{l_+,l_-} \int \prod_{i=1}^{l_+} d\widehat{\mathcal{P}}(u_i^+) \prod_{i=1}^{l_-} d\widehat{\mathcal{P}}(u_i^-) \delta\left(h - \sum_{i=1}^{l_+} u_i^+ + \sum_{i=1}^{l_-} u_i^-\right)$$

$$\times \sum_{n_1,\ldots,n_{l_-}} \prod_{i=1}^{l_-} \widehat{q}_{n_i}^{(L)}(u_i^-)\delta_{n,1+n_1+\cdots+n_{l_-}}, \tag{54}$$

with $q_n^{(L=0)}(h) = \delta_{n,1}$. Finally the sought-for average m.s.r.d. for the root of $\mathbb{T}_L$ reads

$$q_n^{(L)} = \int d\mathcal{P}(h)(1 - \tanh h) q_n^{(L)}(h), \tag{55}$$

which is obtained from (51) by using the statistical equivalence between positive and negative literals. This implies in particular that h has a symmetric distribution, so that $q_n^{(L)}$ is well normalized.

The adaptation of the general population dynamics algorithm to this case is simple. The joint distribution $q_n^{(L)}(h)\mathcal{P}(h)$ is represented by a sample of couples $\{(h_i, n_i)\}_{i=1}^{\mathcal{N}}$, initialized with $n_i = 1$ and the h_i's distributed according to $\mathcal{P}(h)$ (thanks to preliminary population dynamics steps). The recursion over L amounts to generating a sample $\{(u_j, n_j)\}$, where for each j one selects $k-1$ indices $i_1, \ldots, i_{k-1}$ in $[1, \mathcal{N}]$. u_j is set to $f(h_{i_1}, \ldots, h_{i_{k-1}})$, n_j to the minimum of $\{n_{i_1}, \ldots, n_{i_{k-1}}\}$ with probability $1 - \exp[-2u_j]$, to 0 otherwise. In the converse step for each new element i two Poisson integers $l_\pm$ of mean $\alpha k/2$ are independently drawn, then two sets of indices J_+ and J_- of cardinalities l_+ and l_- are generated. h_i is given by $\sum_{j\in J_+} u_j - \sum_{j\in J_-} u_j$, while n_i reads $1 + \sum_{j\in J_-} n_j$. From (55) we obtain $q_n^{(L)}$ as a weighted histogram of the population,

$$q_n^{(L)} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} (1 - \tanh h_i)\delta_{n,n_i}. \tag{56}$$

The large L limit is obtained by repeating a sufficient number of these steps to achieve convergence within numerical precision.

4 The Freezing Transition in Random Tree Ensembles

In the previous section we have established numerical procedures to compute the average m.s.r.d. q_n for the various random tree ensembles, based either on a simple recurrence over n for the XORSAT and COL case, or on a more elaborate population dynamics algorithm for SAT. We now discuss the outcome of these computations, the limit of infinite depth trees ($L \to \infty$) being kept implicit.

In Fig. 3 we have plotted the integrated distribution Q_n, for various values of α, in the XORSAT and SAT case. These two families of curves present the same striking feature: when α is increased Q_n develops a plateau, in other words q_n becomes bimodal with a positive fraction of rearrangements shifting towards larger and larger values. When a critical value $\alpha_{\rm p}$ is reached the length of the plateau becomes infinite. This transition is thus described by the order parameter $\phi = \lim_{n\to\infty} Q_n$, which represents the fraction of percolating optimal rearrangements whose size diverge with L. From the point of view of the order

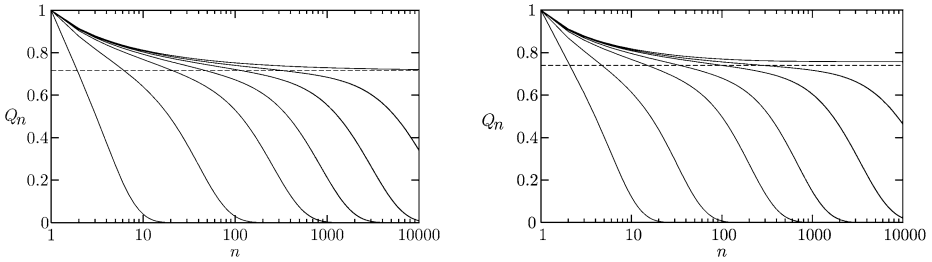

Fig. 3 Integrated average distribution of minimal size rearrangements in tree ensembles. *Left*: random 3-XORSAT, from *left* to *right* $\alpha = 0.4, 0.7, 0.78, 0.8, 0.81, 0.815, \alpha_{\rm p}$. The *dashed horizontal line* is the order parameter at the transition, $\phi_{\rm p} \approx 0.715332$. *Right*: random 3-SAT, from *left* to *right*, $\alpha = 3, 4, 4.3, 4.36, 4.39, 4.40, 4.41$. The *dashed line* indicates $\phi_{\rm p} \approx 0.74$

parameter the transition is discontinuous, ϕ jumps from 0 to a positive value $\phi_{\rm p}$ when the threshold $\alpha_{\rm p}$ is crossed.

Let us follow the interpretation suggested at the end of Sect. 3.2.2 of Q_n being the distribution of an extended integer which has probability ϕ of being infinite. With the rules that the minimum of several such extended integers is infinite if and only if each of them is infinite, while their sum is infinite as soon as one of them is so, (31, 32) imply in the XORSAT case

$$\widehat{\phi} = \phi^{k-1}, \qquad \phi = 1 - \exp[-\alpha k \widehat{\phi}], \tag{57}$$

where we denoted $\widehat{\phi} = \lim \widehat{Q}_n$. This can be closed under the form $\phi = 1 - \exp[-\alpha k \phi^{k-1}]$, with $\alpha_{\rm p}$ being the smallest value of α for which there exists a non-trivial solution. At $\alpha_{\rm p}$ this solution appears discontinuously, with the positive value $\phi_{\rm p}$ corresponding to the height of the plateau in the curves of Q_n. For larger values of α there are two non-trivial solutions, the relevant one being the largest. Numerical values of $\alpha_{\rm p}$ and $\phi_{\rm p}$ are given in Table 1 for a few values of k.

Thanks to the formal equivalence between XORSAT and COL summarized in (39) we immediately obtain the equation on the order parameter of the COL freezing transition and the critical value c_p (see also Table 1 for their numerical values),

$$\phi = \left(1 - \exp\left[-\frac{c\phi}{(q-1)}\right]\right)^{q-1}, \qquad c_{\rm p}^{\rm (COL)}[q] = q(q-1)\alpha_{\rm p}^{\rm (XORSAT)}[k=q]. \tag{58}$$

The initiated reader will recognize the order parameter as the fraction of hard fields in the solution of the 1RSB equations at $m = 1$ given in [22]; we shall come back on this point later on.

The determination of the threshold $\alpha_{\rm p}$ is slightly more involved in the SAT problem. We have indeed a family of distributions $q_n(h), \widehat{q}_n(u)$ indexed by a real h, u; it is thus necessary to define for each of them an order parameter $\phi(h), \widehat{\phi}(u)$, as the fraction of infinite values of n born by $q_n(h), \widehat{q}_n(u)$. The equivalent of (57) takes now a functional form easily derived from (53, 54),

$$\widehat{\phi}(u)\widehat{\mathcal{P}}(u) = \int \prod_{i=1}^{k-1} d\mathcal{P}(h_i)\delta(u - f(h_1, \ldots, h_{k-1})) \prod_{i=1}^{k-1} \frac{1 - \tanh h_i}{2} \phi(h_i), \tag{59}$$

$$\phi(h)\mathcal{P}(h) = \sum_{l_+,l_-=0}^{\infty} p_{l_+,l_-} \int \prod_{i=1}^{l_+} d\widehat{\mathcal{P}}(u_i^+) \prod_{i=1}^{l_-} d\widehat{\mathcal{P}}(u_i^-) \times \delta\left(h - \sum_{i=1}^{l_+} u_i^+ + \sum_{i=1}^{l_-} u_i^-\right)\left(1 - \prod_{i=1}^{l_-}(1-\widehat{\phi}(u_i^-))\right). \tag{60}$$

From the solution of these equations the order parameter of the average m.s.r.d. is obtained (see (55)) as $\phi = \int d\mathcal{P}(h)(1-\tanh h)\phi(h)$. Again, ϕ is the fraction of hard fields in the $m = 1$ 1RSB equations of [23], this connection shall be discussed further in Sect. 6.3 and Appendix 2. A solution of the functional equation on $\phi(h)$ can be sought by a population dynamics algorithm: the distribution $\mathcal{P}(h)$ being represented by a sample $\{h_i\}$, we associate to each of them an estimation ϕ_i of $\phi(h_i)$ and consider a population of couples $\{(h_i,\phi_i)\}_{i=1}^{\mathcal{N}}$. From this a new population $\{u_j, \widehat{\phi}_j\}_{j=\mathcal{N}+1}^{2\mathcal{N}}$ is generated according to (59): for each element of the new population $k-1$ indices $i_1,\ldots,i_{k-1}$ are chosen uniformly at random in $[1,\mathcal{N}]$ and the new couple $(u_j,\widehat{\phi}_j)$ is computed as

$$(u_j,\widehat{\phi}_j) = \left(f(h_{i_1},\ldots,h_{i_{k-1}}), \prod_{m=1}^{k-1} \frac{1-\tanh h_{i_m}}{2}\phi_{i_m}\right). \tag{61}$$

In turns the sample $\{(h_i,\phi_i)\}$ is generated from the $\{u_i,\widehat{\phi}_i\}$'s according to (60), and an estimation for the order parameter is computed as

$$\phi = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}(1-\tanh h_i)\,\phi_i. \tag{62}$$

These two steps are iterated a large number of times, starting with the initial condition $\phi(h) = 1$, i.e. $\phi_i = 1$ for all elements of the initial population. For small values of α the function $\phi(h)$ converges to 0 upon these iterations, while for larger values a non-trivial fixed point is reached. The numerical estimation of the threshold α_{p} separating these two regimes, along with the deduced order parameter at the transition, are presented in Table 1. The precision on these numbers is rather low; indeed, strong finite $\mathcal{N}$ effects make difficult a precise determination of the discontinuous disappearance of the non-trivial solution. Moreover the numerical method becomes difficult for large values of k, hence the limitation of the results presented to $k \in [3,6]$. For $k = 3$ ϕ_{p} can also be successfully compared on the right part of Fig. 3 with the plateau in the numerically determined Q_n.

The discontinuous character of the transition exhibited by the jump of the order parameter should not hide the strong precursor effects, usually associated to continuous transitions, present in the low connectivity phase. The existence of a diverging scale of rearrangement sizes is indeed obvious on Fig. 3. One can for instance define $n_\epsilon(\alpha)$ as the point where Q_n crosses a threshold ϵ. This scale $n_\epsilon(\alpha)$ diverges at α_{p} (as long as $0 < \epsilon < \phi_{\mathrm{p}}$), in other words arbitrary large rearrangements are present with positive probability sufficiently close to the transition. A detailed study of the XORSAT problem [27], drawing on a formal analogy with the mode-coupling theory of supercooled liquids [32], revealed that the divergence of n_ϵ is algebraic, $n_\epsilon \sim (\alpha_{\mathrm{p}} - \alpha)^{-\nu}$. This exponent ν is the solution of an universal type of relations,

$$\nu = \frac{1}{2a} + \frac{1}{2b}, \qquad \frac{\Gamma^2(1-a)}{\Gamma(1-2a)} = \frac{\Gamma^2(1+b)}{\Gamma(1+2b)} = \lambda, \tag{63}$$

Table 1 Threshold, order parameter and critical exponents for the freezing transition in random tree ensembles

k,q	XORSAT		COL		XORSAT and COL			
	α_p	ϕ_p	c_p	ϕ_p	λ	a	b	ν
3	0.818469	0.715332	4.910815	0.511700	0.397953	0.422096	1.221834	1.593787
4	0.772280	0.851001	9.267358	0.616297	0.350174	0.433412	1.341647	1.526313
5	0.701780	0.903350	14.035605	0.665924	0.320971	0.439997	1.421808	1.488035
6	0.637081	0.930080	19.112434	0.695986	0.300707	0.444431	1.481191	1.462601
7	0.581775	0.945975	24.434557	0.716600	0.285554	0.447677	1.527913	1.444121
8	0.534997	0.956381	29.959848	0.731841	0.273649	0.450187	1.566174	1.429899
9	0.495255	0.963661	35.658363	0.743697	0.263961	0.452205	1.598411	1.418505
10	0.461197	0.969008	41.507763	0.753261	0.255868	0.453873	1.626162	1.409102

k,q	SAT					
	α_p	ϕ_p	λ	a	b	ν
3	4.40	0.74	0.55	0.38	0.90	1.87
4	10.55	0.86	0.40	0.42	1.22	1.60
5	21.22	0.91	0.33	0.44	1.40	1.50
6	39.87	0.93	0.31	0.44	1.45	1.47

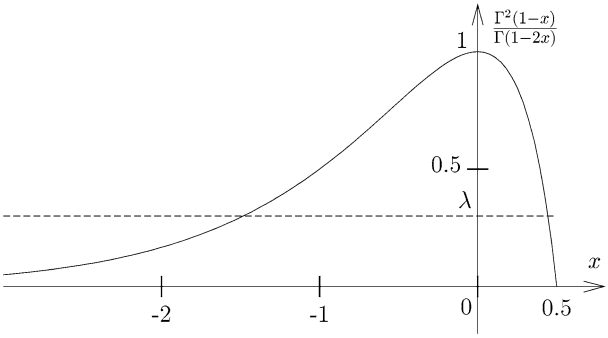

Fig. 4 The exponent a (respectively $-b$) is the positive (respectively negative) root of the equation represented here, see (63)

where Γ denotes Euler's special function (see Fig. 4) and λ a k-dependent parameter in $[0, 1]$. In fact a and b are also critical exponents governing the asymptotic behavior of Q_n around its plateau, see Appendix 1 for details. The non-universal parameter λ was found [27] to be, in the XORSAT case,

$$\lambda^{(\mathrm{XORSAT})} = \frac{k-2}{\alpha_p k(k-1)\phi_p^{k-1}}. \tag{64}$$

Numerical values of this parameter and the associated exponents a, b, ν can be found in Table 1. Because of (39) the exponents for the q-coloring are exactly the same as the one of k-XORSAT, provided one identifies k and q. It will be useful for future discussion to rewrite the parameter λ under the form

$$\lambda^{(\mathrm{COL})} = (q-2)\frac{1-\phi_p^{1/(q-1)}}{\phi_p^{1/(q-1)}}. \tag{65}$$

The asymptotic behavior of the distribution q_n for SAT could be a priori more complicated, because of the underlying infinity of distributions $q_n(h)$. We shall however argue in Appendix 1 that the phenomenology remains the same, in particular the exponents a, b and ν are still given by (63). The parameter λ is now

$$\lambda^{(\mathrm{SAT})} = \frac{2^k(k-2)}{\alpha_{\mathrm{p}} k(k-1)\phi_{\mathrm{p}}^{k-1}}, \tag{66}$$

the expression (64) being only modified by a scale factor 2^k on the connectivity. The value of λ can thus be determined from the numerical evaluation of α_{p} and ϕ_{p} explained above (see Table 1 for the results). The technical details of the analysis, along with numerical evidence supporting it, can be found in Appendix 1.

5 A Digression about the Reconstruction Problem

It is instructive, and shall be useful for the discussion of the following section, to reconsider the freezing transition from a slightly different perspective, namely the problem of tree reconstruction [33]. For simplicity we consider first the q-colorings of regular trees with $L+1$ generations, where every vertex has degree $l+1$ (apart from the root which has degree l and from the leaves of degree 1). The generation of an uniform proper coloring can be seen as a broadcasting process: the color of the root being chosen, each of its sons has a color uniformly chosen among the $q-1$ other ones, and this is propagated until the leaves of depth L have been reached. In an information theoretic vision the color of the root is an information transmitted through a noisy channel, the tree. The reconstruction problem consists in inferring the color of the root given the observation of the colors of the leaves, while the rest of the coloring is hidden to the observer. Depending on the values of (l, q) a correlation between the color of the root and the one of the leaves survives or not the limit $L \to \infty$. An optimal algorithm will be able to infer the value of the root from the observation of the leaves with a probability of success larger than the one of a random uniform guess if and only if this correlation remains positive. In this case the reconstruction problem is said to be solvable, which can also be formulated as the non-extremality (or impurity) of the free-boundary Gibbs measure [34] on the infinite tree. On general grounds one expects a critical value $l_{\mathrm{d}}(q)$ separating a solvable regime for $l \geq l_{\mathrm{d}}(q)$ and an unsolvable one when $l < l_{\mathrm{d}}(q)$. The values $l_{\mathrm{d}}(3) = 5$, $l_{\mathrm{d}}(4) = 8$, $l_{\mathrm{d}}(5) = 13$ and $l_{\mathrm{d}}(6) = 17$ have been conjectured in [35], along with rigorous bounds $l_{\mathrm{d}}(3) \leq 5$, $l_{\mathrm{d}}(4) \leq 9$, $l_{\mathrm{d}}(5) \leq 13$ and $l_{\mathrm{d}}(6) \leq 17$.

A very naive, suboptimal algorithm to perform this inference proceeds from the leaves towards the root, according to the following rule: if the set of colors on the descendants of a vertex i contains $q-1$ distinct elements in $[1, q]$, the remaining color is assigned to the vertex σ_i. Otherwise it is assigned a neutral color, say white, $\sigma_i = 0$. It is easy to realize that at the end of the execution of this algorithm, starting from the observation of the leaves of a proper coloring, the vertices in the interior of the tree are either white or have been assigned the correct value they had in the initial coloring. What is the probability ϕ_L (with respect to the choice of the initial coloring) that the root is correctly reconstructed in such a way? For this to be possible, $q-1$ distinct colors had to be assigned to its sons in the initial coloring, and for each color at least one of them had to be correctly reconstructed. ϕ_L can thus be

determined by recurrence according to $\phi_{L+1} = V(\phi_L)$, with

$$V(\phi) = \frac{1}{(q-1)^l} \sum_{\substack{l_1,\ldots,l_{q-1} \\ \sum l_i = l}} \frac{l!}{l_1! \ldots l_{q-1}!} \prod_{i=1}^{q-1} (1-(1-\phi)^{l_i})$$

$$= \sum_{p=0}^{q-1} \binom{q-1}{p} (-1)^p \left(1 - \frac{p}{q-1}\phi\right)^l, \tag{67}$$

and $\phi_{L=0} = 1$. The limit of ϕ_L for large L is the largest solution of the fixed-point equation $\phi = V(\phi)$ on the interval $[0, 1]$. Depending on the values of q and l this limit is either zero (for instance V vanishes identically if $l < q-1$ as there are not enough descendants for the root to be fully constrained) or strictly positive. By numerical inspection of (67) we found the latter case to happen when $l \geq l_{\mathrm{p}}(q)$, with $l_{\mathrm{p}}(3) = 5$, $l_{\mathrm{p}}(4) = 9$, $l_{\mathrm{p}}(5) = 14$ and $l_{\mathrm{p}}(6) = 19$. This means that the algorithm has a positive probability of guessing correctly the root from the observation of arbitrarily distant leaves when $l \geq l_{\mathrm{p}}(q)$, whereas it is doomed to fail if $l < l_{\mathrm{p}}(q)$. The reasoning presented here is essentially a constructive proof of the bound $l_{\mathrm{d}}(q) \leq l_{\mathrm{p}}(q)$, weaker yet conceptually much simpler than the rigorous bound of [35]. Let us underline that such a reconstruction procedure is far from optimal; we only retain the information given by a drastic event, when the color of a vertex is unambiguously determined by its descendants, and discard the cases where one color is only more probable than the others.

This naive reconstruction algorithm is in a sense dual with the main subject of the paper: it correctly infers the color of the root if and only in all proper colorings with the observed assignment of the leaves the root takes always the same color. In other words in all rearrangements (not necessarily of minimum size) for the root starting from the initial coloring at least one site on the boundary of the tree has to be rearranged. This can be determined using the recursion relation on the sizes of the minimal rearrangements (for instance (33, 34) in the case of the coloring) with a different boundary condition, $n_{i\to j} = \infty$ when i is a leaf of the tree. The value n_i computed with this boundary condition will be infinite if there are no rearrangements of the root which can avoid rearranging the leaves (the algorithm is successful), finite otherwise (the root is white at the end of the algorithm). This difference in the boundary condition ($n_{i\to j} = \infty$ vs 1) is irrelevant in the large L limit: m.s.r. of finite size have supports of finite depth, hence are not affected by the boundary when L gets larger than this depth, while m.s.r. of sizes growing with L are correctly assigned their formal infinite size in this way.

To summarize the connection between this section and the rest of the paper, the constraints that imply large rearrangements are precisely the information exploited in the naive reconstruction procedure. The probability of success of the algorithm on arbitrarily large trees can thus be identified with the order parameter of the freezing transition introduced in the previous section. This identification holds for generic CSPs on random trees, provided one averages the success probability of the naive reconstruction over the ensemble of trees. Another suggestive perspective on the problem is given in terms of percolation. The order parameter can indeed be viewed as the probability of percolation of the support of the rearrangement from the root to an infinitely distant boundary. In the case of XORSAT this percolation is purely geometrical and corresponds to the existence of an infinite subtree where all variable nodes have degree greater than two. For COL and SAT the object which percolates is subtler: the rearrangements depend both on the geometry of the factor graph and on its initial solution.

6 From Random Trees to Random Graphs

6.1 Local and Global Aspects of the Cavity Method

We turn now to the more delicate issue of the validity of the results derived in the random tree ensembles for the original random graphs. As mentioned in Sect. 2 the latter have a local tree structure, with high probability in the thermodynamic limit. The point thus amounts to giving a description of the boundary condition induced by the rest of the factor graph. We shall handle this problem in the framework of the cavity method [11, 36] for sparse random graphs [30] (see also [35, 37] for related discussions) that we briefly survey below.

Consider a $G(N, M)$[5] random factor graph F with N variable nodes and $M = \alpha N$ constraint nodes of degree k, the associated random measure on $\mathcal{X}^N$,

$$\mu(\underline{\sigma}) = \frac{1}{Z} \prod_{a=1}^{M} \psi_a(\underline{\sigma}_a), \tag{68}$$

and suppose the weights ψ_a are i.i.d. positive random functions on $\mathcal{X}^k$ (not necessarily $\{0, 1\}$ valued).

Two kind of intertwined properties of the model can be investigated: thermodynamic (global) ones, with the characterization of the random variable Z, and local ones, concerning the behavior of the measure μ itself. Because of the self-averaging properties of $\ln Z$ for large graphs the central thermodynamic quantity is the quenched free-entropy density,

$$\Phi = \lim_{N\to\infty} \frac{1}{N} \mathbb{E} \ln Z. \tag{69}$$

The latter aspect of the problem, which is the important one for our present concerns, can be formulated as follows. Call F_L the sub-factor graph induced in F by the variable nodes at a graph distance smaller than L from an arbitrary root i, and $\underline{\sigma}_L$ the configuration of these variable nodes. As we are interested in the thermodynamic limit with L finite we can assume without harm that F_L is a tree. The marginalization of (68) leads to a law μ_L for $\underline{\sigma}_L$; it can be seen as a random measure, conditioning F on a given realization of F_L, because of the choices in $F \setminus F_L$. At this point a question arises naturally: what is the (weak) limit of μ_L when the thermodynamic limit is taken?

The cavity method provides a series of possible answers to this question, and an heuristic to choose the right one. Let us introduce some notations: we denote b the number of sites in the boundary B made of the sites at distance exactly L from i, $B = \{i_1, \ldots, i_b\}$, and define a measure on $\underline{\sigma}_L$ with external fields η_j (probability measures on $\mathcal{X}$) acting on this boundary:

$$\mu^{(0)}(\underline{\sigma}_L; \eta_1, \ldots, \eta_b) = \frac{1}{Z_0(\eta_1, \ldots, \eta_b)} \prod_{a\in F_L} \psi_a(\underline{\sigma}_a) \prod_{j=1}^{b} \eta_j(\sigma_{i_j}), \tag{70}$$

where Z_0 ensures the normalization of the law.

The statement of the simplest (so-called Replica-Symmetric, RS) situation described by the cavity method is

$$\mu_L(\cdot) \xrightarrow{\mathrm{d}} \mu^{(0)}(\cdot; \eta_1, \ldots, \eta_b), \tag{71}$$

[5] Random hypergraphs with arbitrary degree distributions can be studied similarly.

where the η_i are i.i.d. from a distribution $\mathcal{P}_{(0)}$. Roughly speaking, this is true when μ is a (finite size) pure state, so that the effect of $F \setminus F_L$ on the boundary variables can be factorized. In more complicated situations there is a large number of pure states on which the Gibbs measure has to be decomposed for this factorization to be possible.[6] We shall thus introduce a new measure on $\underline{\sigma}_L$ as a weighted superposition of different $\mu^{(0)}$,

$$\mu^{(1)}(\underline{\sigma}_L; P_1^{(1)}, \dots, P_b^{(1)}, m) = \frac{1}{Z_1[P_1^{(1)}, \dots, P_b^{(1)}; m]} \int \prod_{i=1}^{b} dP_i^{(1)}(\eta_i) \mu^{(0)}(\underline{\sigma}_L; \eta_1, \dots, \eta_b) Z_0(\eta_1, \dots, \eta_b)^m. \quad (72)$$

In this definition $m \in [0, 1]$ is known as the Parisi breaking parameter, the $P_i^{(1)}$'s are distributions of fields, and again Z_1 is a normalization. The hypothesis of the cavity method at the level of one step of Replica-Symmetry Breaking (1RSB) reads

$$\mu_L(\cdot) \stackrel{\mathrm{d}}{\to} \mu^{(1)}(\cdot; P_1^{(1)}, \dots, P_b^{(1)}, m), \quad (73)$$

where the $P_i^{(1)}$ are i.i.d. from a distribution $\mathcal{P}_{(1)}$. In some cases the 1RSB description coincides with the RS one, for instance whenever the $P_i^{(1)}$ in the support of $\mathcal{P}_{(1)}$ are concentrated on a single value of the field (in the following we shall call this a trivial 1RSB solution). A less obvious reduction happens when the parameter m is equal to 1: from (70, 72) one realizes that in this case $\mu^{(1)}$ is indistinguishable from $\mu^{(0)}$ with properly averaged values of the external fields, more precisely

$$\mu^{(1)}(\underline{\sigma}_L; P_1^{(1)}, \dots, P_b^{(1)}, m = 1) = \mu^{(0)}(\underline{\sigma}_L; \overline{\eta}_1, \dots, \overline{\eta}_b) \quad \text{with } \overline{\eta}_i = \int dP_i^{(1)}(\eta)\eta. \quad (74)$$

This 1RSB formalism can be promoted to an arbitrary level of symmetry breaking by a recursive construction. Let us call $\mathcal{M}_0$ the set of probability laws on $\mathcal{X}$, and define by recurrence $\mathcal{M}_{K+1}$ as the set of probability laws on $\mathcal{M}_K$. The measure $\mu^{(K)}$ with K steps of replica symmetry breaking is parameterized by K reals $0 \le m_1 \le \cdots \le m_K \le 1$ and b elements $P_i^{(K)}$ of $\mathcal{M}_K$, and can be expressed recursively as

$$\mu^{(K+1)}\big(\underline{\sigma}_L; \{P_i^{(K+1)}\}_{i=1}^b, m_1, \dots, m_{K+1}\big) = \frac{1}{Z_{K+1}[\{P_i^{(K+1)}\}; m_1, \dots, m_{K+1}]} \int \prod_{i=1}^{b} dP_i^{(K+1)}\big(P_i^{(K)}\big) \times \mu^{(K)}\big(\underline{\sigma}_L; \{P_i^{(K)}\}, m_2, \dots, m_{K+1}\big) Z_K\big(\{P_i^{(K)}\}, m_2, \dots, m_{K+1}\big)^{m_1/m_2}. \quad (75)$$

The K-RSB assumption of the cavity method reads

$$\mu_L(\cdot) \stackrel{\mathrm{d}}{\to} \mu^{(K)}(\cdot; P_1^{(K)}, \dots, P_b^{(K)}, m_1, \dots, m_K), \quad (76)$$

with the $P_i^{(K)}$ i.i.d. from $\mathcal{P}_{(K)}$, a given element of $\mathcal{M}_{K+1}$. Eventually the limit of an infinite number of steps of replica symmetry breaking ($K \to \infty$) can be formally taken. Note that,

[6] We skip the intermediate case of a finite number of pure states; for instance the low temperature phase of an Ising ferromagnet should be described by the superposition of the two $\mu^{(0)}$ of positive and negative magnetization.

as discussed in the 1RSB case, $\mu^{(K)}$ incorporates as special cases (when the distributions concentrates on a single value, or when the m_i's are degenerate) all possible descriptions at a smaller level of RSB.

We face now the problem of choosing, among all these possible assumptions, which is the correct one. A first condition on the allowed values of $\mathcal{P}_{(K)}$ arises from a simple consistency requirement. μ_L can indeed be obtained in two ways: from a direct application of the statement (76), or by considering a larger neighborhood of depth $L' > L$ and making a partial marginalization of $\mu_{L'}$. As $F_{L'} \setminus F_L$ is distributed according to a Galton–Watson branching process, the consistency of these various ways of obtaining μ_L induces conditions restricting the possible values of $\mathcal{P}_{(K)}$. At the RS ($K = 0$) level this is nothing but the stationarity property stated in (21). The heuristic for the choice of K and the values of the breaking parameters m_i arises from the global aspect of the cavity method, namely the computation of the typical value of the free-entropy density Φ. More precisely, for each level of the RSB hierarchy there is a functional $\Phi_{(K)}[\mathcal{P}_{(K)}, m_1, \ldots, m_K]$ whose minimum is taken as an estimation of Φ. The bounds $\Phi \leq \Phi_{(K)}$ have indeed been rigorously proven in some cases [38–40], and are expected to hold with a certain generality. The best estimation of Φ, which is presumably exact in mean-field models (this has been proven in one case [41]), should thus be sought through the minimization of $\Phi_{(K)}$ in the formal $K \to \infty$ limit which encompasses all possible levels of RSB. The limit of μ_L is expected to be described by the set of parameters achieving this minimum (note that the extremization of $\Phi_{(K)}$ with respect to $\mathcal{P}_{(K)}$ corresponds to the consistency requirement explained above). This minimization is obviously a formidable task which seems out of reach in its full generality for models on sparse random graphs. There are however partial arguments which can be used to assess the validity of the simplest RS and 1RSB hypothesis. The decay of point-to-set correlations at large distance (in other words the purity of the Gibbs measure, or the non reconstructibility of the value of a spin from the observation of distant sites) is indeed related to the absence of a non-trivial solution of the 1RSB consistency equations at $m = 1$ [35], and suggests the RS hypothesis to be correct. A test of the plausibility of the 1RSB description is usually performed via a local stability analysis [42]: one checks in this way the absence of a non-trivial solution of the 2RSB consistency equations in the vicinity of a 1RSB solution $\mathcal{P}_{(1)}$.

Let us finally underline the deep connection between these issues and the local weak convergence method developed by Aldous (see [43, 44] for reviews) on related optimization problems. Recently the above stated local properties of the RS cavity method were rigorously proven in some discrete models (cf. for instance [45–47]), under a priori non-optimal conditions (worst-case vs typical decay of correlations, i.e. uniqueness vs extremality conditions [21]).

6.2 Minimal Size Rearrangements in the Random Graph Ensembles

We shall now reconsider the computations of the m.s.r.d. performed in the random tree ensembles in the light of the above presented cavity method. It should be clear that the thermodynamic limit ($N \to \infty$) of the average distribution q_n defined in (4) for the original random graph ensembles coincide with the infinite L limit of their tree counterpart whenever the RS assumption stated in (71) is valid. The probability measure on the initial configuration we used for the computation of the m.s.r. in finite tree formulae (cf. (11)) corresponds indeed to the limit measure $\mu^{(0)}$ on the finite neighborhood of the random graphs. The validity of this RS scenario depends on the particular model and on the value of the connectivity parameter α (c for coloring).

In the case of XORSAT [5, 6] the local properties of the uniform measure over the set of solutions are well described by the RS assumption up to the satisfiability threshold α_s,

for all values of k. In consequence the computation of q_n performed in the random tree ensemble extends to random graphs throughout the satisfiable phase $\alpha \leq \alpha_s$, the threshold for the freezing transition in random graphs (α_f) and in random trees (α_p) are equal, and the exponents governing the divergence of the m.s.r. in the limit $\alpha \to \alpha_f^-$ are correctly obtained from (64). In fact α_p corresponds also to the clustering transition α_d due to the appearance of an extensive 2-core: a rearrangement for a variable in the 2-core (more precisely in the backbone [5, 6]) is necessarily of extensive size. In agreement with this correspondence, the order parameter of the freezing transition solution of (57) is precisely the fraction of vertices in the backbone.

The picture of the satisfiable phase of random k-SAT and q-COL advocated in [21–23] is richer. Let us first describe it on the example of SAT. At low values of the connectivities, $\alpha < \alpha_d(k)$, one expects a plain RS description to hold. The clustering transition $\alpha_d(k)$ corresponds to the appearance of long-range point-to-set correlations, in other words to a non-trivial solution of the 1RSB equations with $m = 1$. In an intermediate regime $[\alpha_d(k), \alpha_c(k)]$ the thermodynamics of the system is described by a 1RSB scenario with $m = 1$, the dominant clusters of solutions are exponentially numerous (their complexity is strictly positive).[7] At $\alpha_c(k)$ a condensation phenomenon occurs, the degeneracy of the thermodynamically relevant clusters becomes sub-exponential, and the 1RSB breaking parameter m decreases from 1 to 0 as α increases from $\alpha_c(k)$ to the satisfiability threshold $\alpha_s(k)$. Higher levels of RSB might be necessary to describe the condensated regime $[\alpha_c(k), \alpha_s(k)]$; we shall in the following make the hypothesis (partly supported by [48]) that this is not the case for $\alpha \leq \alpha_c(k)$. Because of the equivalence, for the local properties of the measure, of an RS description and a 1RSB with $m = 1$ (cf. (74)), we thus expect the computation of the minimal size rearrangements performed on the tree to be correct for random SAT formulae with $\alpha \leq \alpha_c(k)$. For the sake of readability we reproduce in Table 2 the values of $\alpha_d(k)$ and $\alpha_c(k)$ obtained in [21, 23], along with the satisfiability threshold $\alpha_s(k)$ of [24].

Depending on the values of k the freezing threshold $\alpha_p(k)$ for the random tree ensemble is, or not, smaller than the condensation one. For $k \in [3, 5]$ one finds $\alpha_p(k) > \alpha_c(k)$: for these values of k the computation in the tree ensemble does not allow the determination of the freezing threshold of the original ensembles $\alpha_f(k)$ (at this point we can just say that $\alpha_f(k) > \alpha_c(k)$). For $k = 6$ the situation is reversed, $\alpha_p(6) < \alpha_c(6)$, we thus conclude that $\alpha_f(6) = \alpha_p(6)$, and that the exponents a, b, ν describing the precursors of the freezing transition can be safely computed from (66). We expect the ordering of the various thresholds, and hence the validity of the conclusions just stated for $k = 6$, to remain the same for all greater values of k. This is corroborated by an analysis of the large k limit presented in Appendix 1: the asymptotic behavior of $\alpha_p(k)$ is much smaller than the one of $\alpha_c(k)$ [21, 23],

$$\alpha_f(k) = \alpha_p(k) = \frac{2^k}{k}(\ln k + O(\ln \ln k)) \ll \alpha_c(k) = 2^k \ln 2 - O(1). \tag{77}$$

In fact the SAT problem in the limit of large k becomes similar to the XORSAT problem: the threshold $\alpha_f(k) = \alpha_p(k)$ is equivalent to 2^k times the corresponding value for XORSAT, the order parameter at the transitions are equivalent in both problems, hence the parameter λ governing the critical exponents becomes the same in the large k limit. Moreover from the results of [21, 23] on the behavior of the clustering threshold one realizes that the regime

[7] The case $k = 3$ is special from this point of view, one finds indeed $\alpha_d(3) = \alpha_c(3)$ and no intermediate phase with an exponential number of relevant clusters.

Table 2 Thresholds for the original random ensembles, from left to right clustering, freezing, freezing of the tree ensemble, condensation and satisfiability. The COL values are from [22, 25] for the satisfiability threshold c_s, the SAT ones from [21, 23, 24] for α_s. For $q \in [3, 8]$ the freezing threshold of [22] is computed at the 1RSB level

k, q	COL					SAT				
	c_d	c_f	c_p	c_c	c_s	α_d	α_f	α_p	α_c	α_s
3	4	4.6	4.911	4	4.68	3.86		4.40	3.86	4.267
4	8.35	8.8	9.267	8.4	8.90	9.38		10.55	9.547	9.931
5	12.83	13.5	14.036	13.2	13.67	19.16		21.22	20.80	21.117
6	17.64	18.6	19.112	18.4	18.88	36.53	39.87		43.08	43.37
7	22.70	24.1	24.435	24.0	24.45					
8	27.95	29.93	29.960	29.90	30.33					
9	33.45	35.658		36.0	36.49					
10	39.0	41.508		42.5	42.9					

$[\alpha_d(k), \alpha_f(k)]$ where clusters are present yet do not have frozen variables is of vanishing width in this limit.

The picture of the satisfiable regime for the q-coloring of random graphs presented in [22] is essentially the same as the one of SAT we just described. The dynamical, condensation and satisfiability thresholds obtained in [22] are recalled in Table 2 (the last two are denoted $c_g(q)$ and $c_q(q)$ in [22]). As argued above the computation performed in the random tree ensemble should be correct for Poissonian random graphs of mean connectivity $c \leq c_c(q)$; for $q \in [3, 8]$ this regime does not include the tree freezing transition $c_p(q)$ (called $c_r(m = 1)$ in [22]). Conversely for $q \geq 9$ we have $c_f(q) = c_p(q)$, which is given exactly by $q(q-1)$ times the threshold of XORSAT (recall the formal equivalence between XORSAT and the free boundary COL problem stated in (39)), and the exponents a, b, ν are the same as in XORSAT (identifying q and k). This ordering of the thresholds is confirmed by the analysis at large q,

$$c_f(q) = q(\ln q + O(\ln \ln q)) \ll c_c(q) = 2q \ln q - O(\ln q), \tag{78}$$

the behavior of $c_f(q)$ being justified in Appendix 1 while the one of the condensation threshold was given in [22].

6.3 Dealing with RSB

We have thus reached the frustrating conclusion that the computations performed up to now were not able to determine the average m.s.r.d. in the condensated phase of SAT and COL, and in particular for $k \in [3, 5]$, $q \in [3, 8]$, to locate the freezing transition and describe its critical behavior. The presentation of the cavity method of Sect. 6.1 indicates clearly what has to be done to remedy this insufficiency: one should reproduce the computations of the m.s.r.d. on finite trees, taking for the probability law on the initial configurations $\mu^{(K)}$ instead of the $\mu^{(0)}$ we initially considered. This generalized computation can in fact be performed in a similar way, at the price of some technical complications, and is sketched for the $K = 1$ level of replica symmetry breaking in Appendix 2. The resulting equations become rather difficult to solve and we leave the complete determination of the distribution q_n as an open problem. One can however draw some general observations that we want to underline here.

The order parameter of the freezing transition, i.e. the fraction of rearrangements of diverging size, corresponds to the probability (over the pure states distribution) of a variable being acted on by an hard field which constrains it to a single value. This was found above in the three CSP we considered when the freezing transition happens in a 1RSB phase with $m = 1$, and will be shown in Appendix 2 to hold in non trivial situations with $m < 1$. This should remain true for any CSP and any further level of RSB. Another universality statement concerns the critical behavior of the distribution q_n around the freezing transition $\alpha_{\rm f}$. The phenomenology described by the exponents a, b, ν can indeed be argued to persist even when $\alpha_{\rm f}$ belongs to the condensated regime $[\alpha_{\rm c}, \alpha_{\rm s}]$. Moreover the parameter λ fixing the value of the exponents can be expressed from the standard RSB computation. The reader will find in Appendix 2 the technical details leading to this conclusion for SAT and COL at the 1RSB level, which is also expected to hold for other CSPs and higher levels of RSB.

7 Conclusions and Perspectives

One of the main themes of the paper was the distinction that has to be made between the clustering and freezing transitions. These can coincide in sufficiently symmetric problems like XORSAT, yet in general the solution space gets clustered without variables taking the same value in all elements of the clusters. A definition of the clustering threshold $\alpha_{\rm d}$ was put forward in [21] as the smallest connectivity such that the long-range point-to-set correlation

$$\lim_{L\to\infty}\lim_{N\to\infty}\mathbb{E}\sum_{\underline{\sigma}_{\partial L}}\mu(\underline{\sigma}_{\partial L})\sum_{\sigma_i}|\mu(\sigma_i|\underline{\sigma}_{\partial L})-\mu(\sigma_i)| \tag{79}$$

remains positive, where i is an arbitrary variable node and $\underline{\sigma}_{\partial L}$ the configuration of the nodes at graph distance exactly L from i. A similar definition of the freezing transition $\alpha_{\rm f}$ can be given in terms of the stronger notion of correlation

$$\lim_{L\to\infty}\lim_{N\to\infty}\mathbb{E}\sum_{\underline{\sigma}_{\partial L}}\mu(\underline{\sigma}_{\partial L})\sum_{\sigma_i}\mathbb{I}(\mu(\sigma_i|\underline{\sigma}_{\partial L})=1), \tag{80}$$

hence $\alpha_{\rm f} \geq \alpha_{\rm d}$. The sub-optimality of the naive reconstruction algorithm given in Sect. 5 should clarify why this inequality is in general strict.

In this paper we concentrated on the rearrangements of finite sizes in the thermodynamic limit, i.e. we computed the limit $N \to \infty$ (or $L \to \infty$ in tree ensembles) of the distributions q_n at a fixed value of the sizes n. The percolating rearrangements thus appeared as formally infinite values of n which had to be included to ensure the normalization of the limiting q_n. It should be an interesting research problem to describe more precisely these diverging size rearrangements by taking a scaling limit of q_n, letting n grows with N. The leading order is expected to be linear in N, as are the minimal Hamming distances between clusters studied for instance in [49]. This investigation might in particular clarify the equality found, in the three particular cases and up to the first level of replica symmetry breaking, between the order parameter of the freezing transition and the fraction of hard fields in the usual cavity computations.

The divergence of the minimal size of rearrangements can be viewed as a percolation phenomenon of their supports. In the case of XORSAT this is nothing but the classical 2-core percolation of random hypergraphs; for general CSP, in particular SAT and COL, the percolating structure is defined in two steps, the factor graph being equipped with a measure on the set of initial configurations. The universality of their critical behavior described by the

exponents a, b, ν and the relations (63) between them is shared by other similar problems, for instance rigidity [50] and q-core [51] percolation when defined on Bethe lattices. The latter problem is strongly related to kinetically constrained models [52], for which minimal size rearrangements can be also computed and have the same critical behavior [53].

The recursion relations (7) could form the basis of new investigations on the structure of a single formula, following the line of research pioneered in [15, 16]. Though there is no guarantee of convergence in the presence of cycles in the factor graph, they can be turned into an heuristic message passing algorithm that will provide informations on a solution of a given instance of CSP. This solution should be found by an independent solver algorithm, or, as was proposed in [54], in an incremental way. Starting from an empty formula and an arbitrary assignment of the variables constraints are introduced one by one. Whenever the new constraint is violated by the current assignment one rearranges it; in [54] this step was performed by a local search algorithm, that could be replaced by the single sample m.s.r. message passing heuristic.

The study of the rearrangements of XORSAT performed in [27] addressed further issues left apart in the present work. One was the characterization of the geometrical properties of the m.s.r., through the distribution of their average depths and a measure of their cooperativity by a geometrical susceptibility. We expect some of these geometrical results to extend from XORSAT to arbitrary CSPs, in particular the value of the critical exponents $\zeta = 1/2$, $\eta = 1$ (see [27] for their definitions). Another aspect should on the contrary be much more problem dependent, namely the structure of the energy barriers between rearranged configurations. Given a pair of satisfying assignments $\underline{\sigma}$, $\underline{\tau}$ one can define the set of paths in the configuration space which leads from one to the other by modifying one variable at a time, each variable being modified at most once. The barrier between $\underline{\sigma}$ and $\underline{\tau}$ can be defined as the minimum over this set of paths of the maximum along the path of the number of violated constraints. One can then study the rearrangements which modify a given variable i and achieves a minimal value of the barrier between the initial and final configurations. The structure of XORSAT is such that minimal barrier and minimal size rearrangements do not coincide, and that energy barriers are always strictly positive (unless the variable appears in no constraint, otherwise flipping a variable always makes at least one constraint unsatisfied). On the contrary for SAT a finite size rearrangement can always be performed remaining in the set of satisfying configurations: one just has to flip the variables in decreasing order with respect to the distance from the root of the rearrangement.

Let us finally mention that the general formalism can be applied to several CSPs besides the three examples on which we concentrated. For instance the bicoloring of random hypergraphs [55], which admits a stationary free boundary, is easily seen to have a freezing transition in random tree ensembles with branching ratio

$$\alpha_{\mathrm{p}}^{(\mathrm{BICOL})}(k) = (2^{k-1} - 1)\alpha_{\mathrm{p}}^{(\mathrm{XORSAT})}(k). \tag{81}$$

Acknowledgements I warmly thank Florent Krzakala, Andrea Montanari, Federico Ricci-Tersenghi and Lenka Zdeborová for a fruitful collaboration, in particular A. M. with whom some of the techniques were developed in [27] and for enlightening discussions on the issue of Sect. 6.1, and Jorge Kurchan for interesting exchanges about [54].

The work was partially supported by EVERGROW, integrated project No. 1935 in the complex systems initiative of the Future and Emerging Technologies directorate of the IST Priority, EU Sixth Framework.

Appendix 1: Critical Behavior around the Freezing Transition

XORSAT

In this appendix we shall give some details on the asymptotic behavior of the average m.s.r.d. in the neighborhood of the freezing transition in the random tree ensembles. The case of XORSAT was treated in [27], the main interest will thus be in the extension of these results to the SAT problem. For the sake of clarity we first recall briefly some of the key points of Appendix C in [27].

Let us define the generating functions of q_n and $\widehat{q}_n$ as

$$R(x) = \sum_{n=1}^{\infty} q_n x^n, \qquad \widehat{R}(x) = \sum_{n=1}^{\infty} \widehat{q}_n x^n. \tag{82}$$

Equations (28, 27) can be rewritten as

$$\widehat{Q}_n = Q_n^{k-1}, \tag{83}$$

$$R(x) = x \exp[-\alpha k + \alpha k \widehat{R}(x)]. \tag{84}$$

The order parameter $\phi = \lim_{n\to\infty} Q_n$ can also be expressed as $R(x = 1)$; the equation determining ϕ is formally written as $\phi = V(\phi, \alpha)$ with $V(\phi, \alpha) = 1 - \exp[-\alpha k \phi^{k-1}]$. At the transition point $(\alpha_{\mathrm{p}}, \phi_{\mathrm{p}})$ we have $\partial_\phi V = 1$: the two curves become tangent at this point. More explicitly,

$$\phi_{\mathrm{p}} = 1 - \exp[-\alpha_{\mathrm{p}} k \phi_{\mathrm{p}}^{k-1}], \tag{85}$$

$$1 = \alpha_{\mathrm{p}} k(k-1)\phi_{\mathrm{p}}^{k-2} \exp[-\alpha_{\mathrm{p}} k \phi_{\mathrm{p}}^{k-1}]. \tag{86}$$

Consider first the large n regime right at the transition ($\alpha = \alpha_{\mathrm{p}}$), and assume that the decay of Q_n towards the plateau ϕ_{p} is algebraic, $Q_n \sim \phi_{\mathrm{p}} + An^{-a}$, with A a positive constant and a a positive exponent. Expanding (83) with this ansatz, we obtain

$$\widehat{Q}_n \sim \phi_{\mathrm{p}}^{k-1} + (k-1)\phi_{\mathrm{p}}^{k-2} A n^{-a} + \frac{(k-1)(k-2)}{2} \phi_{\mathrm{p}}^{k-3} A^2 n^{-2a}. \tag{87}$$

The properties of generating functions (similar to Laplace transforms) lead to algebraic singularities of R and $\widehat{R}$ around $x = 1$ [56]:

$$R(1-s) \sim 1 - \phi_{\mathrm{p}} - A\Gamma(1-a)s^a, \tag{88}$$

$$\widehat{R}(1-s) \sim 1 - \phi_{\mathrm{p}}^{k-1} - (k-1)\phi_{\mathrm{p}}^{k-2} A\Gamma(1-a)s^a - \frac{(k-1)(k-2)}{2} \phi_{\mathrm{p}}^{k-3} A^2 \Gamma(1-2a)s^{2a} \tag{89}$$

where the equivalent notation hold in the $s \to 0$ limit, and Γ is Euler's special function. Inserting these expressions in (84), one can expand in powers of s and identify the terms of order s^0, s^a and s^{2a} on both sides of the equation. The first two orders compensate because of, respectively, the relation on the order parameter (85) and its derivative (86). The order s^{2a} fixes the exponent a under the form (63), with λ given by (64).

We now consider the limit $\alpha \to \alpha_{\mathrm{p}}$ and denote $\delta = \alpha_{\mathrm{p}} - \alpha$ the (vanishing) distance to the transition. There are two scaling regimes to be distinguished; the first governs the behavior

of Q_n in the neighborhood of the plateau. Suppose this regime is reached on a scale $n_{\rm i}(\delta)$ diverging with δ and described by the following scaling function:

$$\epsilon(t) = \lim_{\delta\to 0} \delta^{-1/2}[Q_{n=tn_{\rm i}(\delta)} - \phi_{\rm p}]. \tag{90}$$

Expanding (83, 84) order by order in δ, one finds similarly (see [27] for details) that the two first orders are satisfied thanks to relations (85, 86), while the third leads to an integro-differential equation for the scaling function $\epsilon(t)$. The important feature of $\epsilon(t)$ is its behavior in the small and large t limits (entrance and exit from the plateau):

$$\epsilon(t) \overset{t\to 0}{\sim} t^{-a}, \qquad \epsilon(t) \overset{t\to\infty}{\sim} t^{b}, \tag{91}$$

where a is the same exponent as before, and b the dual one (cf. (63)). In fact the small t behavior of ϵ allows to fix the still undetermined scale $n_{\rm i}(\delta)$: for a large, yet independent of δ, value of n, the study right at $\alpha_{\rm p}$ lead to $Q_n - \phi_{\rm p} \sim n^{-a}$. For consistency we must have $n^{-a} \sim \delta^{1/2}(n/n_{\rm i}(\delta))^{-a}$, which implies $n_{\rm i}(\delta) \sim \delta^{-1/2a}$.

The second scaling regime describes the decay of Q_n from its plateau value down to zero, i.e. the distribution of the almost-frozen rearrangements whose size is diverging as α reaches $\alpha_{\rm p}$. Suppose again the existence of another scale $n_{\rm f}(\delta)$ for this to happen, and of the scaling function

$$Q(t) = \lim_{\delta\to 0} Q_{n=tn_{\rm f}(\delta)}. \tag{92}$$

Plugging this ansatz in (83, 84) one obtains another equation for $Q(t)$, which implies in particular $Q(t) - \phi_{\rm p} \sim t^b$ at small t. Matching the small t behavior of $Q(t)$ with the large t limit of the previous scaling function $\epsilon(t)$, one finds that $n_{\rm f}(\delta) \sim \delta^{-\nu}$, with $\nu = (1/2a) + (1/2b)$, as announced in the main part of the text.

SAT

The same steps, with some technical adaptations, can be followed in the case of SAT. Let us first define the integrated distributions and the generating functions for each value of the conditioning field:

$$\begin{aligned} Q_n(h) &= \sum_{n'\geq n} q_{n'}(h), \qquad & \widehat{Q}_n(u) &= \sum_{n'\geq n} \widehat{q}_{n'}(u), \\ R(h,x) &= \sum_n q_n(h) x^n, \qquad & \widehat{R}(u,x) &= \sum_n \widehat{q}_n(u) x^n. \end{aligned} \tag{93}$$

We rewrite (53, 54, 55) as

$$\widehat{Q}_n(u)\widehat{\mathcal{P}}(u) = \int \prod_{i=1}^{k-1} d\mathcal{P}(h_i)\delta(u - f(h_1,\ldots,h_{k-1})) \prod_{i=1}^{k-1} \frac{1-\tanh h_i}{2} Q_n(h_i) \quad \text{for } n \geq 1, \tag{94}$$

$$\begin{aligned} R(h,x)\mathcal{P}(h) = x \sum_{l_+,l_-=0}^{\infty} p_{l_+,l_-} \int \prod_{i=1}^{l_+} d\widehat{\mathcal{P}}(u_i^+) \prod_{i=1}^{l_-} d\widehat{\mathcal{P}}(u_i^-)\delta\left(h - \sum_{i=1}^{l_+} u_i^+ + \sum_{i=1}^{l_-} u_i^-\right) \\ \times \prod_{i=1}^{l_-} \widehat{R}(u_i^-, x), \end{aligned} \tag{95}$$

$$Q_n = \int d\mathcal{P}(h)(1-\tanh h)\, Q_n(h). \tag{96}$$

Recall that the functional order parameters $\phi(h) = \lim Q_n(h) = 1 - R(h, x = 1)$ and $\widehat{\phi}(u)$ are solutions of (59, 60); we denote $\phi_{\rm p}(h)$ and $\widehat{\phi}_{\rm p}(u)$ their values at the threshold $\alpha_{\rm p}$ for the appearance of a non-trivial solution, and $\phi_{\rm p} = \lim Q_n = \int d\mathcal{P}(h)(1 - \tanh h)\phi_{\rm p}(h)$.

For our purposes it will be sufficient to work with the simplified versions of (94, 95) obtained by integration over the fields:

$$\int d\widehat{\mathcal{P}}(u)\widehat{Q}_n(u) = \left(\int d\mathcal{P}(h)\frac{1-\tanh h}{2}Q_n(h)\right)^{k-1} = \frac{1}{2^{k-1}}Q_n^{k-1}, \tag{97}$$

$$\int d\mathcal{P}(h)R(h,x) = x\exp\left[-\frac{\alpha k}{2} + \frac{\alpha k}{2}\int d\widehat{\mathcal{P}}(u)\widehat{R}(u,x)\right]. \tag{98}$$

Consider now the behavior of these quantities right at the transition $\alpha_{\rm p}$. The simplest hypothesis is to assume the existence of a single exponent a describing the decay of the integrated distributions $Q_n(h)$, $\widehat{Q}_n(u)$, towards their limit (as $n \to \infty$) $\phi(h)$, $\widehat{\phi}(u)$, independently of h, u. This hypothesis is customary in the formally analog mode coupling theory of liquids [32], where the role of the conditioning field is held by a wave vector. We thus make the ansatz $Q_n(h) \sim \phi(h) + A(h)n^{-a}$ with $A(h)$ a positive function. Expanding (97), this leads to

$$\int d\widehat{\mathcal{P}}(u)\widehat{Q}_n(u) \sim \left(\frac{\phi_{\rm p}}{2}\right)^{k-1} + \frac{k-1}{2^{k-1}}\phi_{\rm p}^{k-2}\left(\int d\mathcal{P}(h)(1-\tanh h)A(h)\right)n^{-a} \tag{99}$$

$$+ \frac{(k-1)(k-2)}{2^k}\phi_{\rm p}^{k-3}\left(\int d\mathcal{P}(h)(1-\tanh h)A(h)\right)^2 n^{-2a}. \tag{100}$$

These algebraic decays at large n translate into singularities in the generating function around $x = 1$,

$$\int d\mathcal{P}(h)R(h,1-s) \sim 1 - \int d\mathcal{P}(h)\phi_{\rm p}(h) - \left(\int d\mathcal{P}(h)A(h)\right)\Gamma(1-a)s^a, \tag{101}$$

$$\int d\widehat{\mathcal{P}}(u)\widehat{R}(u,1-s)$$

$$\sim 1 - \left(\frac{\phi_{\rm p}}{2}\right)^{k-1} - \frac{k-1}{2^{k-1}}\phi_{\rm p}^{k-2}\left(\int d\mathcal{P}(h)(1-\tanh h)A(h)\right)\Gamma(1-a)s^a \tag{102}$$

$$- \frac{(k-1)(k-2)}{2^k}\phi_{\rm p}^{k-3}\left(\int d\mathcal{P}(h)(1-\tanh h)A(h)\right)^2\Gamma(1-2a)s^{2a}. \tag{103}$$

Finally these expansions are inserted in (98); collecting the terms of order s^0, s^a, s^{2a} yields the following three equations:

$$\int d\mathcal{P}(h)\phi_{\rm p}(h) = 1 - \exp\left[-\frac{\alpha_{\rm p}k}{2^k}\phi_{\rm p}^{k-1}\right], \tag{104}$$

$$\int d\mathcal{P}(h)A(h) = \frac{\alpha_{\rm p}k(k-1)}{2^k}\phi_{\rm p}^{k-2}\exp\left[-\frac{\alpha_{\rm p}k}{2^k}\phi_{\rm p}^{k-1}\right]\left(\int d\mathcal{P}(h)(1-\tanh h)A(h)\right), \tag{105}$$

$$\frac{\Gamma(1-a)^2}{\Gamma(1-2a)} = \lambda = \frac{2^k(k-2)}{\alpha_{\rm p}k(k-1)\phi_{\rm p}^{k-1}}. \tag{106}$$

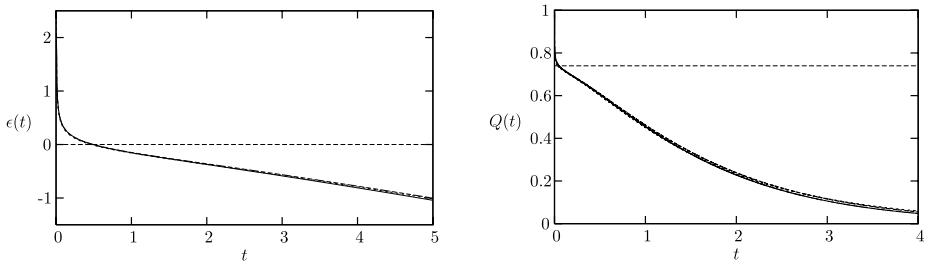

Fig. 5 The scaling functions of the average m.s.r.d. for the random tree ensemble of 3-SAT. The almost superimposed curves correspond to $\alpha = 4.39, 4.392, 4.396$. *Left*: intermediate scale $t = n(\alpha_\mathrm{p} - \alpha)^{1/2a}$, see (90). *Right*: final scale $t = n(\alpha_\mathrm{p} - \alpha)^{\nu}$, cf. (92), the *dashed horizontal line* indicates the order parameter ϕ_p. Numerical values of the exponents can be found in Table 1

The first is a direct consequence of (59, 60) on the order parameter, and can also be derived from (97, 98), setting $x = 1$ in the latter.

The second is a functional analog of (86) and deserves a short explanation. The order parameter $\phi(h)$ is defined as the solution of a fixed-point functional equation of the type $\phi_\alpha = V[\phi_\alpha, \alpha]$, where we keep implicit the functional character of ϕ but emphasize the dependence on the control parameter α. The relevant non-trivial solution of this equation which exists for $\alpha \geq \alpha_\mathrm{p}$ disappears at α_p: this is a bifurcation point in the vocabulary of discrete dynamical systems. A powerful tool in this context is the implicit function theorem: if for some value α_0 there is a solution ϕ_{α_0} and if the differential of V with respect to ϕ in $(\phi_{\alpha_0}, \alpha_0)$ has no eigenvector of eigenvalue 1, then the solution ϕ_α can be continuously followed in a neighborhood of α_0. At the bifurcation point α_p the hypothesis of the theorem must be violated. Linearizing (59, 60), the reader will easily verify that an eigenvector of eigenvalue 1 of the differential satisfies (105). We can thus assume $A(h)$ to be in this eigenspace for the second condition to be verified.[8] Note that for a real order parameter equation $\phi = V(\phi, \alpha)$, this condition is nothing but the equality of the derivatives $1 = \partial_\phi V$ at a transition, as used for instance in (86).

The third equation fixes the exponent a and gives the value of the exponent λ, as was claimed in the main part of the text (cf. (63) and (66)).

The study of the intermediate and final scaling regimes can be done following the lines sketched above on the XORSAT example; for instance the behavior around the plateau is described, for all values of the cavity fields, by a single scaling function, generalizing (90) to

$$Q_{n=tn_\mathrm{i}(\delta)}(h) \sim \phi_\mathrm{p}(h) + \delta^{1/2} A(h) \epsilon(t). \tag{107}$$

Provided $A(h)$ is chosen in such a way that (105) is verified, $\epsilon(t)$ obeys the same kind of integro-differential equation as the scaling function of the XORSAT problem, and in particular its behavior at small and large t is identical (see (91)). We thus reach exactly the same conclusions on the behavior of $n_\mathrm{i}(\delta)$ and $n_\mathrm{f}(\delta)$. This is confirmed in Fig. 5, which shows, in the two regimes, a good collapse of numerically determined distributions Q_n for three values of α approaching α_p.

[8] This explanation is of course heuristic; the functional character of the fixed point equation makes the invocation of the implicit function theorem rather fuzzy.

Asymptotics at Large k, q

We justify now the statements made in the main part of the text on the large k, q behavior of the freezing thresholds. This analysis is simple in the case of XORSAT: from (85, 86) one obtains a closed equation on the order parameter at the transition,

$$\frac{1}{k-1} = -\frac{(1-\phi_{\rm p}(k))\ln(1-\phi_{\rm p}(k))}{\phi_{\rm p}(k)}, \tag{108}$$

which can be inverted to obtain an asymptotic expansion of $\phi_{\rm p}(k)$. Reinserting it in (86) yields

$$\alpha_{\rm p}(k) = \frac{1}{k}\left(\ln k + \ln\ln k + 1 + O\left(\frac{\ln\ln k}{\ln k}\right)\right). \tag{109}$$

The formal correspondence with the COL problem (see (58)) leads immediately to the left-hand side of (78).

The distributions of fields $\mathcal{P}(h)$, $\widehat{P}(u)$ for random SAT formulas can be shown from (52) to concentrate in the large k limit around, respectively, 0 and 2^{-k}. Equations (53, 54) on $q_n(h)$, $\widehat{q}_n(u)$ can thus be simplified at the leading order in k by retaining only these deterministic values of the conditioning fields. A simple transformation then shows that the distribution $q_n(h=0)$ collapses onto the solution of the XORSAT equations (31, 32), provided the connectivity α is divided by a factor of 2^k. This leads to the asymptotic behavior of the freezing threshold stated in (77), and to the equivalence at large k of the exponents a, b, ν in the SAT and XORSAT problems. A systematic expansion in powers of 2^{-k} of the deviations between the two models could be set up from this starting point.

Appendix 2: Minimal Size Rearrangements at the 1RSB Level

General Case

We consider in this appendix the computation proposed in Sect. 6.3, namely the determination of the m.s.r.d. for a finite tree factor graph whose initial configuration is drawn according to the law $\mu^{(1)}$ (see (72)). To characterize it we introduce on each directed edge of the factor graph a distribution of cavity fields, denoted $P_{i\to a}(\eta)$ and $\widehat{P}_{a\to i}(\nu)$. They obey the following set of equations,

$$\widehat{P}_{a\to i}(\nu) = \frac{1}{Z(\{P_{j\to a}\})}\int \prod_{j\in\partial a\setminus i} dP_{j\to a}(\eta_{j\to a})\delta(\nu - f(\{\eta_{j\to a}\}))z(\{\eta_{j\to a}\})^m, \tag{110}$$

$$P_{i\to a}(\eta) = \frac{1}{Z(\{\widehat{P}_{b\to i}\})}\int \prod_{b\in\partial i\setminus a} d\widehat{P}_{b\to i}(\nu_{b\to i})\delta(\eta - g(\{\nu_{b\to i}\}))z(\{\nu_{b\to i}\})^m, \tag{111}$$

where the functions f, g and z are the ones defined in (12, 13) for the corresponding edges. The boundary condition is given by $P_{i\to a}(\eta) = P_{{\rm ext},i}(\eta)$ if $i \in B$, otherwise $P_{i\to a}(\eta) = \delta(\eta - \overline{\eta})$. The marginals of $\mu^{(1)}$ can be obtained from these distributions, for instance for a single variable one obtains

$$\mu^{(1)}(\sigma_i) = \int d P_i(\eta)\eta(\sigma_i),$$
$$P_i(\eta) = \frac{1}{Z(\{\widehat{P}_{a\to i}\})}\int \prod_{a\in\partial i} d\widehat{P}_{a\to i}(\nu_{a\to i})\delta(\eta - g(\{\nu_{a\to i}\}))z(\{\nu_{a\to i}\})^m. \tag{112}$$

We also have to introduce distributions of the size messages, $q_{\vec{n}}^{(i\to a,\sigma_i)}(\eta)$ and $\widehat{q}_{\vec{n}}^{(a\to i,\sigma_i)}(\nu)$, which corresponds to the weighted averages of the distributions in a single $\mu^{(0)}$. From (18, 19) one obtains

$$\widehat{P}_{a\to i}(\nu)\widehat{q}_{\vec{n}}^{(a\to i,\sigma_i)}(\nu) = \frac{1}{Z(\{P_{j\to a}\})}\int \prod_{j\in\partial a\setminus i} d P_{j\to a}(\eta_{j\to a})\delta(\nu - f(\{\eta_{j\to a}\}))z(\{\eta_{j\to a}\})^m$$
$$\times \sum_{\underline{\sigma}_{a\setminus i}} \mu(\underline{\sigma}_{a\setminus i}|\sigma_i, \{\eta_{j\to a}\}) \prod_{j\in\partial a\setminus i}\sum_{\vec{n}_{j\to a}} q_{\vec{n}_{j\to a}}^{(j\to a,\sigma_j)}(\eta_{j\to a})\delta_{\vec{n},\widetilde{f}(\{\vec{n}_{j\to a}\})} \tag{113}$$

and

$$P_{i\to a}(\eta)q_{\vec{n}}^{(i\to a,\sigma_i)}(\eta) = \frac{1}{Z(\{\widehat{P}_{b\to i}\})}\int \prod_{b\in\partial i\setminus a} d\widehat{P}_{b\to i}(\nu_{b\to i})\delta(\eta - g(\{\nu_{b\to i}\}))z(\{\nu_{b\to i}\})^m$$
$$\times \prod_{b\in\partial i\setminus a}\sum_{\vec{n}_{b\to i}} \widehat{q}_{\vec{n}_{b\to i}}^{(b\to i,\sigma_i)}(\nu_{b\to i})\delta_{\vec{n},\widetilde{g}_{\sigma_i}(\{\vec{n}_{b\to i}\})}, \tag{114}$$

with the boundary condition at the leaves $q_{\vec{n}}^{(i\to a,\sigma)}(\eta) = \delta_{\vec{n},\vec{o}(\sigma)}$. Finally the m.s.r.d. with respect to $\mu^{(1)}$ for a variable i reads

$$q_n^{(i)} = \int d P_i(\eta)\sum_{\sigma_i}\eta(\sigma_i)\sum_{\vec{n}} q_{\vec{n}}^{(i,\sigma_i)}(\eta)\delta_{n,\ \min_{\tau_i\neq\sigma_i}[\vec{n}]_{\tau_i}}, \tag{115}$$

with

$$P_i(\eta)q_{\vec{n}}^{(i,\sigma_i)}(\eta) = \frac{1}{Z(\{\widehat{P}_{a\to i}\})}\int \prod_{a\in\partial i} d\widehat{P}_{a\to i}(\nu_{a\to i})\delta(\eta - g(\{\nu_{a\to i}\}))z(\{\nu_{a\to i}\})^m$$
$$\times \prod_{a\in\partial i}\sum_{\vec{n}_{a\to i}} \widehat{q}_{\vec{n}_{a\to i}}^{(a\to i,\sigma_i)}(\nu_{a\to i})\delta_{\vec{n},\widetilde{g}_{\sigma_i}(\{\vec{n}_{a\to i}\})}. \tag{116}$$

Note that this computation reduces to the one of Sect. 3.1.2 either when the distribution of cavity fields are concentrated on a single value or when $m = 1$, defining in the latter case

$$\eta_{i\to a} = \int d P_{i\to a}(\eta)\eta, \qquad q_{\vec{n}}^{(i\to a,\sigma_i)} = \frac{\int d P_{i\to a}(\eta)\eta(\sigma_i)q_{\vec{n}}^{(i\to a,\sigma_i)}(\eta)}{\eta_{i\to a}(\sigma_i)}, \tag{117}$$

and similarly $\nu_{a\to i}$ and $\widehat{q}_{\vec{n}}^{(a\to i,\sigma_i)}$. For a generic value of m one proceeds with the computation of the average m.s.r.d. for a random tree; the only modification with respects to Sect. 3.1.3 is a replacement of the distribution of external fields $\mathcal{P}(\eta)$ by a distribution of distribution of fields, $\mathcal{P}(P)$. One has thus to define $q_{\eta,\vec{n}}^{(\sigma,L)}(P)$, the average of the joint law $P_i(\eta)q_{\vec{n}}^{(i,\sigma_i)}(\eta)$ on η and $\vec{n}$ for the root of $\mathbb{T}_L$, conditioned on the event $P = P_i$, and similarly $\widehat{q}_{\nu,\vec{n}}^{(\sigma,L)}(\widehat{P})$ for $\widehat{\mathbb{T}}_L$. These quantities can be obtained by recursions on L through equations formally similar to (22, 23), which could in principle be solved numerically using a population of population of elements $(\eta, \vec{n}^{(1)}, \ldots, \vec{n}^{(q)})$. We shall give the explicit form of these equations in the two particular cases of SAT and COL in the following two subsections.

SAT

For random SAT instances the stationarity conditions for the distribution of distribution of fields $\mathcal{P}(P)$, $\widehat{\mathcal{P}}(\widehat{P})$ can be written in their distributional form as

$$\widehat{P} \stackrel{\mathrm{d}}{=} F(P_1, \ldots, P_{k-1}), \qquad P \stackrel{\mathrm{d}}{=} G(\widehat{P}_1^+, \ldots, \widehat{P}_{l_+}^+, \widehat{P}_1^-, \ldots, \widehat{P}_{l_-}^-). \tag{118}$$

The functionals F and G are defined by

$$\widehat{P}(u) = \frac{1}{Z(\{P_i\})} \int \prod_{i=1}^{k-1} dP_i(h_i)\delta(u - f(h_1, \ldots, h_{k-1}))z(h_1, \ldots, h_{k-1})^m, \tag{119}$$

$$\begin{aligned} P(h) = {} & \frac{1}{Z(\{\widehat{P}_i^{\pm}\})} \int \prod_{i=1}^{l_+} d\widehat{P}_i^+(u_i^+) \prod_{i=1}^{l_-} d\widehat{P}_i^-(u_i^-)\delta\left(h - \sum_{i=1}^{l_+} u_i^+ + \sum_{i=1}^{l_-} u_i^-\right) \\ & \times z(u_1^+, \ldots, u_{l_+}^+, u_1^-, \ldots, u_{l_-}^-)^m, \end{aligned} \tag{120}$$

where

$$z(h_1, \ldots, h_{k-1}) = 2 - \prod_{i=1}^{k-1} \frac{1 - \tanh h_i}{2}, \tag{121}$$

$$\begin{aligned} z(u_1^+, \ldots, u_{l_+}^+, u_1^-, \ldots, u_{l_-}^-) = {} & \prod_{i=1}^{l_+} \frac{1 + \tanh u_i^+}{2} \prod_{i=1}^{l_-} \frac{1 - \tanh u_i^-}{2} \\ & + \prod_{i=1}^{l_+} \frac{1 - \tanh u_i^+}{2} \prod_{i=1}^{l_-} \frac{1 + \tanh u_i^-}{2}. \end{aligned} \tag{122}$$

The conditional average of the joint law of cavity field and sizes obey the two following equations:

$$\begin{aligned} \widehat{q}_{u,n}^{(L)}(\widehat{P})\widehat{\mathcal{P}}(\widehat{P}) = {} & \int \prod_{i=1}^{k-1} d\mathcal{P}(P_i)\delta(P - F(\{P_i\})) \frac{1}{Z(\{P_i\})} \\ & \times \int \prod_{i=1}^{k-1} dh_i \delta(u - f(h_1, \ldots, h_{k-1}))z(h_1, \ldots, h_{k-1})^m \sum_{n_1, \ldots, n_{k-1}} \prod_{i=1}^{k-1} q_{h_i, n_i}^{(L)}(P_i) \\ & \times \left[\left(1 - \prod_{i=1}^{k-1} \frac{1 - \tanh h_i}{2}\right) \delta_{n,0} + \left(\prod_{i=1}^{k-1} \frac{1 - \tanh h_i}{2}\right) \delta_{n, \min[n_1, \ldots, n_{k-1}]} \right], \end{aligned} \tag{123}$$

$$\begin{aligned} q_{h,n}^{(L+1)}(P)\mathcal{P}(P) = {} & \sum_{l_+, l_-} p_{l_+, l_-} \int \prod_{i=1}^{l_+} d\widehat{\mathcal{P}}(\widehat{P}_i^+) \prod_{i=1}^{l_-} d\widehat{\mathcal{P}}(\widehat{P}_i^-)\delta(P - G(\{\widehat{P}_i^{\pm}\})) \frac{1}{Z(\{\widehat{P}_i^{\pm}\})} \\ & \times \int \prod_{i=1}^{l_+} du_i^+ \prod_{i=1}^{l_-} du_i^- \delta\left(h - \sum_{i=1}^{l_+} u_i^+ + \sum_{i=1}^{l_-} u_i^-\right) z(\{u_i^{\pm}\})^m \prod_{i=1}^{l_+} \widehat{P}_i^+(u_i^+) \\ & \times \sum_{n_1, \ldots, n_{l_-}} \prod_{i=1}^{l_-} \widehat{q}_{u_i^-, n_i}^{(L)}(\widehat{P}_i^-)\delta_{n, 1 + n_1 + \cdots + n_{l_-}}. \end{aligned} \tag{124}$$

These equations conserve the conditions $\sum_n q_{h,n}^{(L)}(P) = P(h)$ which follow from the definition of $q_{h,n}^{(L)}(P)$. Finally the average m.s.r.d. for the root of $\mathbb{T}_L$ reads

$$q_n^{(L)} = \int d\mathcal{P}(P) \int dh (1 - \tanh h) q_{h,n}^{(L)}(P). \tag{125}$$

As a consistency check one can reduce these equations to the ones developed in the main part of the text (cf. (53–55)) when $m = 1$, using the identity (117).

Let us come back on the 1RSB equations (118–120). It is possible for the distributions $\widehat{P}(u)$ in the support of $\widehat{\mathcal{P}}$ to acquire a peak on the hard field value $u = +\infty$, of intensity denoted $\widehat{\phi}(\widehat{P})$. This corresponds to a field forcing the variable node to satisfy the constraint node emitting the message. Similarly we call $\phi(P)$ the intensity of the peak in $h = -\infty$, signaling a clause that the emitting variable is forced to unsatisfy it. These intensities are found from (118–120) to obey

$$\widehat{\phi}(\widehat{P})\widehat{\mathcal{P}}(\widehat{P}) = \int \prod_{i=1}^{k-1} d\mathcal{P}(P_i) \delta(P - F(\{P_i\})) \frac{1}{Z(\{P_i\})} \prod_{i=1}^{k-1} \phi(P_i), \tag{126}$$

$$\begin{aligned}\phi(P)\mathcal{P}(P) = \sum_{l_+,l_-} p_{l_+,l_-} \int \prod_{i=1}^{l_+} d\widehat{\mathcal{P}}(\widehat{P}_i^+) \prod_{i=1}^{l_-} d\widehat{\mathcal{P}}(\widehat{P}_i^-) \delta(P - G(\{\widehat{P}_i^\pm\})) \frac{1}{Z(\{\widehat{P}_i^\pm\})} \\ \times \prod_{i=1}^{l_+} \int d\widehat{P}_i^+(u) \left(\frac{1 - \tanh u}{2}\right)^m \prod_{i=1}^{l_-} \int d\widehat{P}_i^-(u) \left(\frac{1 + \tanh u}{2}\right)^m \\ \times \left[1 - \prod_{i=1}^{l_-} \left(1 - \frac{\widehat{\phi}(\widehat{P}_i^-)}{\int d\widehat{P}_i^-(u) \left(\frac{1+\tanh u}{2}\right)^m}\right)\right].\end{aligned} \tag{127}$$

A randomly chosen variable will receive a forcing hard field in a randomly chosen pure state with probability

$$\phi = 2 \int d\mathcal{P}(P) \phi(P), \tag{128}$$

where the factor 2 comes from the symmetry between positive and negative literals; ϕ is also the order parameter of the freezing transition. Equations (123, 124), in the $L \to \infty$ limit, admit a solution where $\phi(P)$ (resp. $\widehat{\phi}(\widehat{P})$) is the intensity of a Dirac peak on $(h, n) = (-\infty, \infty)$ (resp. $(u, n) = (+\infty, \infty)$). The fraction of diverging rearrangements in (125) is then seen to be equal to ϕ.

In order to discuss the critical behavior of the m.s.r.d. it is convenient to derive an integrated version of (123, 124),

$$\begin{aligned}&\int d\widehat{\mathcal{P}}(\widehat{P}) \frac{\int du \widehat{Q}_{u,n}(\widehat{P})(1 + \tanh u)^m}{\int d\widehat{P}(u)(1 + \tanh u)^m} \\ &\quad = \left(\int d\mathcal{P}(P) \int dh \frac{1 - \tanh h}{2} Q_{h,n}(P)\right)^{k-1} = \frac{1}{2^{k-1}} Q_n^{k-1},\end{aligned} \tag{129}$$

$$\int d\mathcal{P}(P)\frac{\int dh R_{h,x}(P)(1-\tanh h)^m}{\int dP(h)(1-\tanh h)^m}$$
$$= x\exp\left[-\frac{\alpha k}{2}+\frac{\alpha k}{2}\int d\widehat{\mathcal{P}}(\widehat{P})\frac{\int du \widehat{R}_{u,x}(\widehat{P})(1+\tanh u)^m}{\int d\widehat{P}(u)(1+\tanh u)^m}\right], \tag{130}$$

where the former is valid for $n \geq 1$ and following our conventions we defined

$$Q_{h,n}(P)=\sum_{n'\geq n} q_{h,n'}(P), \qquad R_{h,x}(P)=\sum_n x^n q_{h,n}(P). \tag{131}$$

Let us call $\alpha_{\mathrm{p}}^{(1)}$ the threshold value for the appearance of a non-trivial solution to (126, 127), and $\phi_{\mathrm{p}}^{(1)}$ the corresponding order parameter. We want to determine the critical behavior of q_n in the neighborhood of this threshold, expecting to recover the phenomenology obtained in the $m=1$ case. For simplicity we shall consider only the first critical regime at $\alpha=\alpha_{\mathrm{p}}^{(1)}$, supposing an algebraic decay of Q_n with an exponent a to its asymptotic value $\phi_{\mathrm{p}}^{(1)}$. More precisely we make the ansatz $Q_{h,n}(P)=\delta(h+\infty)(\phi(P)+A(P)n^{-a})+o(n^{-a})$, with $A(P)$ a positive function. The computation proceeds as in Sect. 1: one inserts this ansatz in (129) and expands to order n^{-2a}. The algebraic decays translate into singularities around $x=1$ in the generating functions of (130), matching the three leading orders one obtains

$$\int d\mathcal{P}(P)\frac{\phi(P)}{\int dP(h)(\frac{1-\tanh h}{2})^m}=1-\exp\left[-\frac{\alpha_{\mathrm{p}}^{(1)}k}{2^k}(\phi_{\mathrm{p}}^{(1)})^{k-1}\right], \tag{132}$$

$$\int d\mathcal{P}(P)\frac{A(P)}{\int dP(h)(\frac{1-\tanh h}{2})^m}$$
$$=\frac{\alpha_{\mathrm{p}}^{(1)}k(k-1)}{2^k}(\phi_{\mathrm{p}}^{(1)})^{k-2}\exp\left[-\frac{\alpha_{\mathrm{p}}^{(1)}k}{2^k}(\phi_{\mathrm{p}}^{(1)})^{k-1}\right]\int d\mathcal{P}(P)A(P), \tag{133}$$

$$\frac{\Gamma(1-a)^2}{\Gamma(1-2a)}=\lambda^{(1)}=\frac{2^k(k-2)}{\alpha_{\mathrm{p}}^{(1)}k(k-1)(\phi_{\mathrm{p}}^{(1)})^{k-1}}. \tag{134}$$

The first equality is a direct consequence of (126, 127), the second is fulfilled by taking $A(P)$ in the eigenspace of eigenvalue 1 of the differential of (126, 127), while the third fixes the exponent a. The computation of the parameter λ at the RSB level thus leads to the expression found in the RS approach (cf. (66)), apart from the replacement of the critical connectivity and order parameter with their corresponding RSB values.

COL

The random q-COL model is described at the 1RSB level by a distribution $\mathcal{P}(P)$ over (invariant under the color permutations) distributions $P(\eta)$ of fields (laws on $\mathcal{X}=\{1,\ldots,q\}$). $\mathcal{P}$ is solution of the distributional equation $P\overset{\mathrm{d}}{=}F(P_1,\ldots,P_l)$, where l is a Poisson random variable of mean c and F is defined by

$$P(\eta)=\frac{1}{Z(P_1,\ldots,P_l)}\int dP_i(\eta_i)\delta(\eta-f(\eta_1,\ldots,\eta_l))z(\eta_1,\ldots,\eta_l)^m,$$
$$f(\{\eta_i\})(\sigma)=\frac{1}{z(\{\eta_i\})}\prod_{i=1}^{l}(1-\eta_i(\sigma)). \tag{135}$$

One can distinguish the hard fields which constrain a variable to take a definite color and define

$$P(\eta) = \phi(P)\frac{1}{q}\sum_{\sigma=1}^{q}\delta(\eta - d_\sigma) + (1-\phi(P))\widetilde{P}(\eta), \quad d_\sigma(\tau) = \delta_{\sigma,\tau}, \tag{136}$$

where $\widetilde{P}$ is a normalized distribution with no intensity on the hard fields d_σ. The order parameter $\phi(P)$ is found from (135) to obey:

$$\begin{aligned}\phi(P)\mathcal{P}(P) = &\sum_{l=0}^{\infty} p_l \int \prod_{i=1}^{l} d\mathcal{P}(P_i)\delta(P - F(P_1,\ldots,P_l))\frac{1}{Z(P_1,\ldots,P_l)} \\ &\times \sum_{p=0}^{q-1} q\binom{q-1}{p}(-1)^p \prod_{i=1}^{l}\left(\int dP_i(\eta)(1-\eta(\sigma))^m - \frac{p}{q}\phi(P_i)\right).\end{aligned} \tag{137}$$

The average m.s.r.d. on random trees where the initial configurations are drawn from the 1RSB measure $\mu^{(1)}$ reads

$$q_n^{(L)} = \int d\mathcal{P}(P)\int d\eta \sum_{\sigma}\eta(\sigma)q_{\eta,n}^{(\sigma,L)}(P), \tag{138}$$

where $q_{\eta,n}^{(\sigma,L)}(P)$ is the conditional average of the joint law of size and fields messages. Note that all values of σ contribute in the same way above, by the symmetry between colors. The equation governing $q_{\eta,n}^{(\sigma,L)}(P)$ is

$$\begin{aligned}q_{\eta,n}^{(\sigma,L+1)}(P)\mathcal{P}(P) = &\sum_{l=0}^{\infty} p_l \int \prod_{i=1}^{l} d\mathcal{P}(P_i)\delta(P - F(P_1,\ldots,P_l))\frac{1}{Z(P_1,\ldots,P_l)} \\ &\times \int \prod_{i=1}^{l} d\eta_i \delta(\eta - f(\eta_1,\ldots,\eta_l))z(\eta_1,\ldots,\eta_l)^m \\ &\times \sum_{\substack{\sigma_1,\ldots,\sigma_l \\ n_1,\ldots,n_l}} \prod_{i=1}^{l}\mu(\sigma_i|\sigma;\eta_i)q_{\eta_i,n_i}^{(\sigma_i,L)}(P_i)\mathbb{I}\left(n = 1 + \min_{\tau\neq\sigma}\sum_{i=1}^{l}\delta_{\tau,\sigma_i}n_i\right),\end{aligned} \tag{139}$$

with

$$\mu(\sigma_i|\sigma;\eta_i) = \frac{\eta_i(\sigma_i)}{1-\eta_i(\sigma)}\mathbb{I}(\sigma_i \neq \sigma). \tag{140}$$

The order parameter $\phi = \int d\mathcal{P}(P)\phi(P)$ is again the height of the plateau in the $L\to\infty$ limit of the integrated average m.s.r.d. Q_n. One can indeed check that $q_{\eta,n}^{(\sigma)}(P)$ has a Dirac peak of intensity $\phi(P)/q$ in $(\eta,n) = (d_\sigma,\infty)$.

The study of the critical behavior at the transition $c_{\mathrm{p}}^{(1)}$ corresponding to the appearance of hard fields in the 1RSB distributions is similar to the SAT case. We first write an integrated version of (139),

$$\int d\mathcal{P}(P)\frac{\int d\eta\,\eta(\sigma)^m q^{(\sigma)}_{\eta,n}(P)}{\int dP(\eta)\eta(\sigma)^m}$$
$$=\sum_{l=0}^{\infty}\frac{e^{-c}c^l}{l!}\frac{1}{(q-1)^l}\sum_{\sigma_1,\ldots,\sigma_l=2}^{q}\sum_{n_1,\ldots,n_l}\mathbb{I}\left(n=1+\min_{\tau=2,\ldots,q}\left[\sum_{i=1}^{l}\delta_{\tau,\sigma_i}n_i\right]\right)$$
$$\times\prod_{i=1}^{l}\left[(q-1)\int d\mathcal{P}(P)\frac{\int d\eta(1-\eta(1))^{m-1}\eta(\sigma_i)q^{(\sigma_i)}_{\eta,n_i}(P)}{\int dP(\eta)(1-\eta(\sigma))^m}\right], \tag{141}$$

which is independent on the value of σ. The ansatz $Q^{(\sigma)}_{\eta,n}(P)=\delta(\eta-d_\sigma)(\phi(P)+A(P)n^{-a})+o(n^{-a})$ is then inserted in this equation. The first two orders in an asymptotic expansion at large n in powers of n^{-a} are satisfied thanks to (137) and by choosing $A(P)$ in the eigenspace of eigenvalue 1 of its differential. The third order fixes the value of the exponent a through

$$\frac{\Gamma(1-a)^2}{\Gamma(1-2a)}=(q-2)\frac{1-\widetilde{\phi}^{1/(q-1)}}{\widetilde{\phi}^{1/(q-1)}},\quad \widetilde{\phi}=\frac{1}{q}\int d\mathcal{P}(P)\frac{\phi(P)}{\int dP(\eta)\eta(\sigma)^m}, \tag{142}$$

which corresponds for $m=1$ to the expression found in (65).

References

1. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman, New York (1983)
2. Schaefer, T.: In: Proceedings of the 10th Annual ACM Symposium on the Theory of Computing, San Diego, p. 216 (1978)
3. Janson, S., Luczak, T., Rucinski, A.: Random Graphs. Wiley, New York (2000)
4. Mitchell, D., Selman, B., Levesque, H.: In: Proc. of the 10th National Conference on Artificial Intelligence, p. 459 (1992)
5. Mézard, M., Ricci-Tersenghi, F., Zecchina, R.: J. Stat. Phys. **111**, 505 (2003)
6. Cocco, S., Dubois, O., Mandler, J., Monasson, R.: Phys. Rev. Lett. **90**, 047205 (2003)
7. Goerdt, A.: J. Comput. Sci. **53**, 469 (1996)
8. Friedgut, E.: J. Am. Math. Soc. **12**, 1017 (1999)
9. Franco, J.: Theor. Comput. Sci. **265**, 147 (2001)
10. Dubois, O.: Theor. Comput. Sci. **265**, 187 (2001)
11. Mézard, M., Parisi, G., Virasoro, M.A.: Spin Glass Theory and Beyond. World Scientific, Singapore (1987)
12. Monasson, R., Zecchina, R.: Phys. Rev. Lett. **76**, 3881 (1996)
13. van Mourik, J., Saad, D.: Phys. Rev. E **66**, 056120 (2002)
14. Biroli, G., Monasson, R., Weigt, M.: Eur. Phys. J. B **14**, 551 (2000)
15. Mézard, M., Parisi, G., Zecchina, R.: Science **297**, 812 (2002)
16. Mézard, M., Zecchina, R.: Phys. Rev. E **66**, 056126 (2002)
17. Maneva, E.N., Mossel, E., Wainwright, M.J.: Proceedings of the Symposium on Discrete Algorithms, Vancouver, January 2005
18. Mézard, M., Mora, T., Zecchina, R.: Phys. Rev. Lett. **94**, 197205 (2005)
19. Achlioptas, D., Ricci-Tersenghi, F.: arXiv:cs/0611052. In: Proceedings of the Symposium on the Theory of Computing (2006)
20. Gopalan, P., Kolaitis, P., Maneva, E., Papadimitriou, C.H.: In: Proceedings of ICALP 2006, p. 346 (2006)
21. Krzakala, F., Montanari, A., Ricci-Tersenghi, F., Semerjian, G., Zdeborová, L.: Proc. Nat. Acad. Sci. **104**, 10318 (2007)
22. Krzakala, F., Zdeborová, L.: arXiv:0704.1269 (2007)
23. Montanari, A., Ricci-Tersenghi, F., Semerjian, G.: In preparation
24. Mertens, S., Mézard, M., Zecchina, R.: Random Struct. Alg. **28**, 340 (2006)
25. Krzakala, F., Pagnani, A., Weigt, M.: Phys. Rev. E **70**, 046705 (2004)

26. Mézard, M., Palassini, M., Rivoire, O.: Phys. Rev. Lett. **95**, 200202 (2005)
27. Montanari, A., Semerjian, G.: J. Stat. Phys. **124**, 103 (2006)
28. Zhou, H.: New J. Phys. **7**, 123 (2005)
29. Kschischang, F.R., Frey, B.J., Loeliger, H.-A.: IEEE Trans. Inform. Theory **47**, 498 (2001)
30. Mézard, M., Parisi, G.: Eur. Phys. J. B **20**, 217 (2001)
31. Lindvall, T.: Lectures on the Coupling Method. Dover, New York (2002)
32. Götze, W., Sjögren, L.: Rep. Prog. Phys. **55**, 241 (1992)
33. Mossel, E.: In: Nestril, J., Winkler, P. (eds.) Graphs, Morphisms and Statistical Physics. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, Providence (2004)
34. Georgii, H.-O.: Gibbs Measures and Phase Transitions. De Gruyter, Berlin (1988)
35. Mézard, M., Montanari, A.: J. Stat. Phys. **124**, 1317 (2006)
36. Talagrand, M.: Spin Glasses: a Challenge for Mathematicians. Springer, Berlin (2003)
37. Gerschenfeld, A., Montanari, A.: arXiv:0704.3293 (2007)
38. Guerra, F.: Comm. Math. Phys. **233**, 1 (2003)
39. Franz, S., Leone, M.: J. Stat. Phys. **111**, 535 (2003)
40. Panchenko, D., Talagrand, M.: Probab. Theory Relat. Fields **130**, 319 (2004)
41. Talagrand, M.: Ann. Math. **163**, 221 (2006)
42. Montanari, A., Parisi, G., Ricci-Tersenghi, F.: J. Phys. A **37**, 2073 (2004)
43. Aldous, D., Steele, J.M.: In: Kesten, H. (ed.) Probability on Discrete Structures, p. 1. Springer, Berlin (2003)
44. Aldous, D., Bandyopadhyay, A.: Ann. Appl. Probab. **15**, 1047 (2005)
45. Bandyopadhyay, A., Gamarnik, D.: In: Random Structures and Algorithms. math.PR/0510471 (2005, to appear) (preliminary version in Proceedings of SODA 2006)
46. Montanari, A., Shah, D.: In: Proceedings of SODA 2007. cs.DM/0607073 (2006)
47. Bayati, M., Nair, C.: In: Annual Allerton Conference on Communication, Control and Computing. cond-mat/0607290 (2006)
48. Montanari, A., Ricci-Tersenghi, F.: Eur. Phys. J. B **33**, 339 (2003)
49. Mora, T., Mézard, M.: J. Stat. Mech., P10007 (2006)
50. Duxbury, P.M., Jacobs, D.J., Thorpe, M.F., Moukarzel, C.: Phys. Rev. E **59**, 2084 (1999)
51. Schwarz, J.M., Liu, A.J., Chayes, L.Q.: Europhys. Lett. **73**, 560 (2006)
52. Sellitto, M., Biroli, G., Toninelli, C.: Europhys. Lett. **69**, 496 (2005)
53. Montanari, A., Semerjian, G.: Unpublished (2005)
54. Krzakala, F., Kurchan, J.: cond-mat/0702546 (2007)
55. Castellani, T., Napolano, V., Ricci-Tersenghi, F., Zecchina, R.: J. Phys. A **36**, 11037 (2003)
56. Flajolet, P., Odlyzko, A.: SIAM J. Discrete Math. **3**, 216 (1990)