Hindawi Publishing Corporation
International Journal of Statistical Mechanics
Volume 2014, Article ID 136829, 13 pages
http://dx.doi.org/10.1155/2014/136829

Research Article

The Statistical Mechanics of Random Set Packing and a Generalization of the Karp-Sipser Algorithm

C. Lucibello[1] and F. Ricci-Tersenghi[2]

[1] *Dipartimento di Fisica, Università "La Sapienza", Piazzale Aldo Moro 2, 00185 Rome, Italy*
[2] *Dipartimento di Fisica, INFN-Sezione di Roma1, CNR-IPCF UOS Roma Kerberos, Università "La Sapienza", Piazzale Aldo Moro 2, 00185 Rome, Italy*

Correspondence should be addressed to F. Ricci-Tersenghi; federico.ricci@uniroma1.it

Received 19 November 2013; Accepted 8 January 2014; Published 10 March 2014

Academic Editor: Hyunggyu Park

We analyse the asymptotic behaviour of random instances of the maximum set packing (MSP) optimization problem, also known as maximum matching or maximum strong independent set on hypergraphs. We give an analytic prediction of the MSPs size using the 1RSB cavity method from statistical mechanics of disordered systems. We also propose a heuristic algorithm, a generalization of the celebrated Karp-Sipser one, which allows us to rigorously prove that the replica symmetric cavity method prediction is exact for certain problem ensembles and breaks down when a core survives the leaf removal process. The e-phenomena threshold discovered by Karp and Sipser, marking the onset of core emergence and of replica symmetry breaking, is elegantly generalized to $C_s = e/(d-1)$ for one of the ensembles considered, where d is the size of the sets.

1. Introduction

The maximum set packing is a very much studied problem in combinatorial optimization, one of Karp's twenty-one NP-complete problems. Given a set $F = \{1, \dots, M\}$ and a collection of its subsets $\mathcal{S} = \{S_i \mid S_i \subseteq F, i \in V\}$ labeled by $V = \{1, \dots, N\}$; a set packing (SP) is a collection of the subsets S_i such that they are pairwise disjoint. The size of a SP $\mathcal{S}' \subseteq \mathcal{S}$ is $|\mathcal{S}'|$. A maximum set packing (MSP) is an SP of maximum size. The integer programming formulation of the MSP problem reads

$$\text{maximize} \quad \sum_{i \in V} n_i, \tag{1}$$

$$\text{subject to} \quad \sum_{i : r \in S_i} n_i \leq 1 \quad \forall r \in F, \tag{2}$$

$$n_i = \{0, 1\} \quad \forall i \in V. \tag{3}$$

The MSP problem, also known in the literature as the matching problem on hypergraphs or the strong independent set problem on hypergraphs, is an NP-Hard problem. This general formulation, however, can be specialized to obtain two other famous optimization problems: the restriction of the MSP problem to sets S_i of size 2 corresponds to the problem of maximum matching on ordinary graphs and can be solved in polynomial time [1]; the restriction where each element of F appears exactly 2 times in $\mathcal{S}$ is the maximum independent set and belongs to the NP-Hard class.

The formulation ((1)–(3)) of the MSP problem, therefore, encodes an ample class of packing problems and, as all packing problems, is related by duality to a covering problem, the minimum set covering problem. Another common specialization of the general MSP problem, known as k-set packing, is that in which all sets S_i have size at most k. This is one of the most studied specializations in the computer science community, the efforts concentrating on minimal degree conditions to obtain a perfect matching [2], linear relaxations [3, 4], and approximability conditions [5–7]. Motivated by this interest, we choose a k-set packing problem ensemble as the principal application of the general analytical framework developed in the following sections. The asymptotic behaviour of random sparse instances of the MSP problem has not been investigated by mathematicians

and computer scientists; only in the matching [8] and independent set [9] restrictions some work has been done. Extending some theorems of [8] (on which a part of this work is greatly inspired) to a greater class of problem ensembles is some of the main aims of the present work.

On the other hand also the statistical physics literature is lacking an accurate study of the random MSP problem. One of its specialization though, the matching problem, has been covered since the beginning of the physicists' interest in optimization problems, with the work of Parisi and Mézard on the weighted and fully connected version of the problem [10, 11]. More recently the matching problem on sparse random graphs has also been accurately studied [12, 13] using the cavity method technique. Also the independent set problem on random graphs [14] and the dual problem to set packing, the set covering problem [15], received some attention by the disordered physics community. The SP problem was investigated with the cavity method formalism in a disguised form, as a glass model on a generalized Bethe lattice, in [16, 17]. This corresponds, as we will see in the next section, to a factor graph ensemble with fixed factor and variable degrees; thus, we will not cover this case in Section 7.

The paper is organized as follows.

(i) In Section 2 we map the MSP problem ((1)–(3)) into a statistical physical model defined on a factor graph and relate the MSP size to the density ρ at infinite chemical potential.

(ii) We introduce the replica symmetric (RS) cavity method in Section 3 and give an estimate for the average MSPs size on sparse factor graph ensembles in the thermodynamic limit.

(iii) In Section 4 we establish a criterion for the validity of the RS ansatz and introduce the 1RSB formalism.

(iv) In Section 5 we propose a generalization of the Karp-Sipser heuristic algorithm [8] to the MSP problem and prove the validity of the RS ansatz for certain ensemble of problems. Moreover we find a relationship between a core emergence phenomena and the breaking of replica symmetry breaking.

(v) Section 6 describes the numerical simulations performed.

(vi) In Section 7 we apply the analytical tools developed to some problem ensembles. We compare the numerical results obtained from an exact algorithm with the analytical predictions, focusing to greater extent to one ensemble modelling the k-set packing.

2. Statistical Physics Description

In order to turn the MSP combinatorial optimization problem into a useful statistical physical model let us recast ((1)–(3)) into a graphical model using the factor graph formalism [18, 19]. We define our variable nodes set to be V and to each $i \in V$ we associate a variable n_i taking values in $\{0, 1\}$ as in (3). F will be our factor nodes set, as its elements acts as hard constrains on the variables n_i through (2). The edge set E is then naturally defined as $E = \{(i, r) \mid i \in V, r \in S_i \subseteq F\}$. We call $G = (V, F, E)$ the factor graph thus composed and can then rewrite (2) as

$$\sum_{i \in \partial r} n_i \leq 1 \quad \forall r \in F. \tag{4}$$

A SP is a configuration $\{n_i\}$ satisfying (4) and its relative size is $\rho(\{n_i\}) = (1/N) \sum_{i \in V} n_i$ that is simply the fraction of occupied sites.

It is now easy to define an appropriate Gibbs measure for the MSPs problem on G through the grand canonical partition function

$$\Xi_G(\mu) = \sum_{\{n_i\}} \prod_{i \in V} e^{\mu n_i} \prod_{r \in F} \mathbb{I}\left(\sum_{i \in \partial r} n_i \leq 1 \right). \tag{5}$$

Only SPs contribute to the partition function, and in the close packing limit, as we will call the limit $\mu \uparrow +\infty$, the measure is dominated by MSPs. Equation (5) is also a model for a particle gas with hard core repulsion and chemical potential μ located on a hypergraph and as such has been studied mainly on lattice structures and more in general on ordinary graphs. Model (5) has been studied on a generalized Bethe lattice (i.e., the ensemble $\mathbb{G}_{RR}(d, c)$ defined in Section 7, a d-uniform c-regular factor graph) in [16, 17] as a prototype of a system with finite connectivity showing a glassy behaviour. This has been the only approach, although disguised as a hard spheres model, from the statistical physics community to a general MSP problem.

The grand canonical potential is defined as

$$\omega_G(\mu) = -\frac{1}{\mu N} \log \Xi_G(\mu), \tag{6}$$

and the particle density as

$$\rho_G(\mu) = \frac{1}{N} \left\langle \sum_{i \in V} n_i \right\rangle_{G,\mu} = -\omega_G(\mu) - \mu \partial_\mu \omega_G(\mu). \tag{7}$$

Grand potential and density are related to entropy by the thermodynamic relation

$$s_G(\mu) = -\mu \left(\omega_G(\mu) + \rho_G(\mu) \right) = \mu^2 \partial_\mu \omega_G(\mu). \tag{8}$$

In the close packing limit (i.e., $\mu \uparrow +\infty$) we recover the MSP problem, since in this limit the Gibbs measure is uniformly concentrated on MSPs and ρ_G gives the MSP relative size. Since entropy remains finite in this limit, from (8) we obtain the MSP relative size

$$\rho_G \equiv \lim_{\mu \to +\infty} \rho_G(\mu) = \lim_{\mu \to +\infty} -\omega_G(\mu). \tag{9}$$

In this paper we focus on random instances of the MSP problem. As usual in statistical physics we will assume the number of variables N (the number of subsets S_i) and the number of constrains M to diverge, keeping the ratio N/M finite. We will refer to this limit as the thermodynamic limit. Instances of the MSP problem will be encoded in factor graph

ensembles which we assume to be locally tree-like in the thermodynamic limit.

The MSP relative size ρ_G is a self-averaging quantity in the thermodynamic limit and we want to compute its asymptotic value

$$\rho = \lim_{N \to +\infty} \mathbb{E}_G \left[\rho_G\right], \tag{10}$$

where we denoted with $\mathbb{E}_G[\cdot]$ the expectation over the factor graph ensemble. In the last equation the N dependence is encoded in the graph ensemble considered. Computing (10) is not an easy task and some approximation have to be taken. We will employ the cavity method from the statistical physics of disordered systems [19, 20], using both the replica symmetric (RS) and the one-step replica symmetry breaking (1RSB) ansatz. We will prove in Section 5 that the RS ansatz is exact in a certain region of the phase space, while in Section 7 we will give numerical evidence that the 1RSB approximation gives very good results outside the RS region.

3. Replica Symmetry

3.1. Bethe Approximation on a Single Instance. The RS cavity method has been known for many decades outside the statistical physics community as the Belief Propagation (BP) algorithm and only in recent years the two approaches have been bridged [18, 19]. We start with a variational approximation to the grand potential equation (6) of an instance of the problem, the Bethe free energy approximation:

$$\omega_G^{\mathrm{RS}}\left[\widehat{\boldsymbol{\nu}}\right] = \frac{1}{N}\left[\sum_{r\in\partial F}\omega_r\left[\widehat{\boldsymbol{\nu}}\right] + \sum_{i\in V}\left(1-|\partial i|\right)\omega_i\left[\widehat{\boldsymbol{\nu}}\right]\right], \tag{11}$$

with the factor and variable contributions given by

$$\begin{aligned}\omega_r\left[\widehat{\boldsymbol{\nu}}\right] &= -\frac{1}{\mu}\log\left[\sum_{n_i\in\partial r}\mathbb{I}\left(\sum_{i\in\partial r} n_i \le 1\right)\prod_{j\in\partial r}\prod_{s\in\partial j\setminus r}\widehat{\nu}_{s\to j}\left(n_j\right)\right],\\ \omega_i\left[\widehat{\boldsymbol{\nu}}\right] &= -\frac{1}{\mu}\log\left[\sum_{n_i} e^{\mu n_i}\prod_{r\in\partial i}\widehat{\nu}_{r\to i}\left(n_i\right)\right].\end{aligned} \tag{12}$$

The grand canonical potential is expressed as a function of the factor node to variable node messages $\widehat{\boldsymbol{\nu}} = \{\widehat{\nu}_{r\to i}\}$. Minimization of $\omega_G^{\mathrm{RS}}[\widehat{\boldsymbol{\nu}}]$ over the messages constrained to be normalized to one yields the fixed point BP equations for the set packing:

$$\begin{aligned}\widehat{\nu}_{r\to i}(1) &= \frac{1}{Z_{r\to i}}\prod_{j\in\partial r\setminus i}\prod_{s\in\partial j\setminus r}\widehat{\nu}_{s\to j}(0),\\ \widehat{\nu}_{r\to i}(0) &= \frac{1}{Z_{r\to i}}\left[\prod_{j\in\partial r\setminus i}\prod_{s\in\partial j\setminus r}\widehat{\nu}_{s\to j}(0)\right.\\ &\quad + e^{\mu}\sum_{j\in\partial r\setminus i}\prod_{s\in\partial j\setminus r}\widehat{\nu}_{s\to j}(1)\\ &\quad \left.\times\prod_{j'\in\partial r\setminus\{i,j\}}\prod_{s'\in\partial j'\setminus r}\widehat{\nu}_{s'\to j'}(0)\right].\end{aligned} \tag{13}$$

The coeeficients $Z_{r\to i}$ are normalization factor. Equations (13) can be simplified introducing the fields $\{t_{r\to i}\}$ defined as

$$\frac{\widehat{\nu}_{r\to i}(1)}{\widehat{\nu}_{r\to i}(0)} = e^{-\mu t_{r\to i}}, \tag{14}$$

yielding

$$t_{r\to i} = \frac{1}{\mu}\log\left[1+\sum_{j\in\partial r\setminus i} e^{\mu\left(1-\sum_{s\in\partial j\setminus r} t_{s\to j}\right)}\right]. \tag{15}$$

Since we are interested in the close packing limit to solve the problem we will straightforwardly apply the zero temperature cavity method [21]. The related BP equations which can be found as the $\mu\uparrow\infty$ limit of (15) read

$$t_{r\to i} = \max\{0\}\cup\left\{1-\sum_{s\in\partial j\setminus r} t_{s\to j}\right\}_{j\in\partial r\setminus i}. \tag{16}$$

We note that the messages $\{t_{r\to i}\}$ are bounded to take values in the interval $[0,1]$ and that if we set the initial value of each $t_{r\to i}$ in the discrete set $\{0,1\}$, at each BP iteration, all messages will take value either 0 or 1. These values can be directly interpreted as the occupational loss occurring in the subtree $r \to i$ if the subtree is connected to the occupied node i. This loss cannot be negative (thus a gain), since we put an additional constrain on the subtree demanding every j neighbour of r to be empty, and cannot be greater than 1 as well, in fact $t_{r\to i} = 1$ corresponds to the worst case scenario where an otherwise occupied node $j \in \partial r\setminus i$ has to be emptied.

The Bethe free energy for model (5) on the factor graph G can be expressed as a function of the fixed point messages $\{t_{r\to i}\}$ as

$$\begin{aligned}\omega_G^{\mathrm{RS}}(\mu) = -\frac{1}{\mu N}\left[\sum_{r\in\partial F}\log\left(1+\sum_{j\in\partial r} e^{\mu\left(1-\sum_{s\in\partial j\setminus r} t_{s\to j}\right)}\right)\right.\\ \left.+\sum_{i\in V}\left(1-|\partial i|\right)\log\left(1+e^{\mu\left(1-\sum_{r\in\partial i} t_{r\to i}\right)}\right)\right].\end{aligned} \tag{17}$$

We finally arrive to the Bethe estimation of the MSPs relative size, taking the close packing limit of (17) and using $\omega_G^{\mathrm{RS}} = -\rho_G^{\mathrm{RS}}$, which is given by

$$\rho_G^{\mathrm{RS}} = \frac{1}{N}\left[\sum_{r\in F}\max\{0\}\cup\left\{1-\sum_{s\in\partial j\setminus r} t_{s\to j}\right\}_{j\in r} + \sum_{i\in V}(1-|\partial i|)\max\left\{0,1-\sum_{r\in i} t_{r\to i}\right\}\right]. \tag{18}$$

Let us examine the various contributions to (18) since we want to convince ourselves that it exactly counts the MSP size, at least on tree factor graphs. The term $(1-|\partial i|)\max\{0,1-\sum_{r\in i} t_{r\to i}\}$ contributes with $1-|\partial i|$ to the sum only if all the incoming t messages are zero. In this case n_i is frozen to 1, that is, the variable i takes part of all the MSPs in G. Obviously all the neighbours of a variable frozen to 1 have to be frozen to 0. To all its $|\partial i|$ neighbours r, the frozen to 1 variable i sends a message $1-\sum_{s\in\partial i\setminus r} t_{s\to i}=1$, so that we have $|\partial i|$ contributions in the first sum of (18) $\max\{0\}\cup\{1-\sum_{s\in\partial j\setminus r} t_{s\to j}\}_{j\in r}=1$ and the total contribution from i correctly sums up to 1. If for a certain i we have a total field $\tau_i \equiv 1-\sum_{r\in i} t_{r\to i} < 0$ (two or more incoming messages are equal to one) the variable is frozen to 0; that is, it does not take part of any MSPs. It correctly does not contribute to ω_G^{RS} since it sends a message $1-\sum_{s\in i\setminus r} t_{s\to i} \le 0$ to each neighbour r; thus, it is not computed in $\max\{0\}\cup\{1-\sum_{s\in\partial j\setminus r} t_{s\to j}\}_{j\in r}$.

The third case is the most interesting. It concerns variables i which take part to a fraction of the MSPs. We will call them unfrozen variables. The total field on an unfrozen variable i is $\tau_i = 0$ (thus we have no contribution from the second sum in (18)) and all incoming messages are 0 except for a single $t_{r\to i}=1$. To this sole function node r, the node i sends a message 1, so that the contribution of r to the first sum is 1. Actually BP equations impose that r has to have at least another unfrozen neighbour beside i. In other terms the function node r says that whatever MSP we consider, one of my neighbours has to be occupied. The corresponding term $\max\{0\}\cup\{1-\sum_{s\in\partial j\setminus r} t_{s\to j}\}_{j\in r}=1$ in (18) accounts for that.

The presence of unfrozen variables is the reason why we cannot express the density ρ_G through the formula

$$\rho_G = \left\langle \sum_{i\in V} n_i \right\rangle. \tag{19}$$

In fact, using the infinite chemical potential formalism, we cannot compute

$$\langle n_i\rangle = \lim_{\mu\to+\infty}\frac{e^{\mu(1-\sum_{r\in i} t_{r\to i})}}{1+e^{\mu(1-\sum_{r\in i} t_{r\to i})}} \tag{20}$$

when $\lim_{\mu\to+\infty} 1-\sum_{r\in i} t_{r\to i}=0$, and we would have to use the $O(1/\mu)$ corrections to the fields $\{t_{r\to i}\}$. We bypass the problem using the grand potential ω_G^{RS} to obtain ρ_G^{RS}, also addressing a problem reported in [22] of extending an analysis suited for weighted matchings and independent sets to the unweighted case.

3.2. Ensemble Averages. To proceed further in the analysis and since one is often concerned with the average properties of a class of related factor graphs, let us consider the case where the factor graph G is sampled from a locally tree-like factor graph ensemble $\mathbb{G}(N)$. We will employ the following notation for the graph ensembles expectations: $\mathbb{E}_G[\cdot]$ for graphs averages; $\mathbb{E}_{C_0}[\cdot]$ ($\mathbb{E}_{D_0}[\cdot]$) for expectations over the factor (variable) degree distribution, which we will sometime call root degree distribution; and $\mathbb{E}_C[\cdot]$ ($\mathbb{E}_D[\cdot]$) for expectations over the excess degree distribution of factor (variable) nodes conditioned to have at least one adjacent edge, which we will sometime call residual degree distribution. The quantities c and d and the random variables C, C_0, D, and D_0, are related by

$$\mathbb{P}[C=k] = \frac{(k+1)\,\mathbb{P}[C_0=k+1]}{c}, \qquad \mathbb{P}[D=k] = \frac{(k+1)\,\mathbb{P}[D_0=k+1]}{d}. \tag{21}$$

In Section 7 we discuss some specific factor graph ensembles, where C and D are fixed to a deterministic value or Poissonian distributed.

With these definitions the distributional equation corresponding to Belief Propagation formula (16) which reads

$$T' \overset{\mathrm{d}}{=} \max\{0\}\cup\left\{1-\sum_{s=1}^{D_j} T_{sj}\right\}_{j\in\{1,\dots,C\}}, \tag{22}$$

where $\{D_j\}$ are i.i.d. random residual variable degrees, C is a random residual factor degree, and $\{T_{sj}\}$ are i.i.d. random incoming messages. Since on a given graph each message $t_{r\to i}$ takes value only on $\{0,1\}$ we can take the distribution of messages t to be of the form

$$P(t) = p\delta(t) + (1-p)\,\delta(t-1). \tag{23}$$

For this to be a fixed point of (22), the parameter p has to satisfy the self-consistent equation

$$p_* = \mathbb{E}_C\left(1-\mathbb{E}_D p_*^D\right)^C. \tag{24}$$

Using (18) and (24), we obtain the replica symmetric approximation to the asymptotic MSP relative size

$$\rho^{\mathrm{RS}} = \frac{d}{c}\left(1-\mathbb{E}_{C_0}\left(1-\mathbb{E}_D p_*^D\right)^{C_0}\right) + \mathbb{E}_{D_0}\left(1-D_0\right)p_*^{D_0}. \tag{25}$$

It turns out that the RS approximation is exact only when the ratio N/M is sufficiently low. In the SP language (25) holds true only when the number of the subsets among which we choose our MSP is not too big compared to the number of elements of which they are composed.

In the next section we will quantitatively establish the limits of validity of the RS ansatz and introduce the one-step replica symmetry breaking formalism which provides a better approximation to the exact results in the regime of large N/M ratio. In Section 5 we prove that (25) is exact for certain choices of the factor graph ensembles.

4. Replica Symmetry Breaking

4.1. RS Consistency and Bugs Propagation. Here we propose two criterions in order to check the consistency of the RS cavity method, iIf any of those fails.

The first criterion is the assumption of unicity of the fixed point of (24) and its dynamical stability under iteration. We restrict ourselves to the subspace of distributions with support on $\{0, 1\}$, although it is possible to extend the following analysis to the whole space of distributions over $[0, 1]$ with an argument based on stochastic dominance following [22]. Characterizing the distributions over $\{0, 1\}$ as in (23) with a real parameter $p \in [0, 1]$, from (22) we obtain the dynamical system

$$p' = \mathbb{E}_C\left(1 - \mathbb{E}_D p^D\right)^C \equiv f(p). \tag{26}$$

The stability criterion

$$\left|f'(p_*)\right| < 1 \tag{27}$$

suggests that the RS approximation to the MSP size (25) is exact as long as (27) is satisfied. This statement will be made rigorous in Section 5.

The second method we use to check the RS stability, called bugs proliferation, is the zero temperature analogous of spin glass susceptibility. We will compute the average number of changing t messages induced by a change in a single message $t_{r\to i}$ ($1 \to 0$ or $0 \to 1$). This is given by

$$N_{\text{ch}} = \mathbb{E}\left[\sum_{(s,j)} \sum_{\substack{a_0,a_1\\ b_0,b_1}} \mathbb{I}\left(t_{s\to j} = b_0 \longrightarrow b_1 \mid t_{r\to i} = a_0 \longrightarrow a_1\right)\right], \tag{28}$$

where a_0, a_1, b_0, and b_1 take value in $\{0, 1\}$. Since a random factor graph is locally a tree, and assuming correlations decay fast enough, last equation can be expressed as

$$N_{\text{ch}} = \sum_{s=0}^{+\infty} \left(\overline{CD}\right)^s \sum_{\substack{a_0,a_1\\ b_0,b_1}} P\left(t_s = b_0 \longrightarrow b_1 \mid t_o = a_0 \longrightarrow a_1\right), \tag{29}$$

where t_s is a message at distance s from the tree root o, and we defined the average residual degrees $\overline{C} = \mathbb{E}_C[C]$ and $\overline{D} = \mathbb{E}_D[D]$. The stability condition $N_{\text{ch}} < +\infty$ yields a constraint on the greatest eigenvalue λ_M of the transfer matrix $P(b_0 \to b_1 \mid a_0 \to a_1)$:

$$\overline{CD}\lambda_M < 1. \tag{30}$$

The two methods presented above give equivalent conditions for the RS ansatz to hold true and they simply express the independence for a finite subgraph from the tail boundary conditions.

4.2. The 1RSB Formalism. We are going to develop the 1RSB formalism for the MSP problem and then apply it in Section 7.1 to the ensemble $\mathbb{G}_{RP}$. We will not check the coherence of the 1RSB ansatz through the interstate and intrastate susceptibilities [23]; we are then not guaranteed against the need of further steps of replica symmetry breaking in order to recover the exact solution. Even in the worst case scenario though, when the 1RSB solution is trivially exact only in the RS region and a full RSB ansatz is needed otherwise, the 1RSB prediction for MSP relative size ρ should be everywhere more accurate than the RS one and possibly very close to the real value. We will refer to the textbook of Montanari and Mézard [19] for a detailed exposition of the 1RSB cavity method.

Let us fix a factor graph G from a locally tree-like ensemble $\mathbb{G}$. We call $Q_{r\to i}(t_{r\to i})$ the distribution of messages on the directed edge $r \to i$ over the states of the system. We still expect the messages $t_{r\to i}$ to take values 0 or 1, so that $Q_{r\to i}$ can be parametrized as

$$Q_{r\to i}(t_{r\to i}) = q_{r\to i}\delta(t_{r\to i}) + (1 - q_{r\to i})\,\delta(t_{r\to i} - 1). \tag{31}$$

The 1RSB Parisi parameter $x \in [0, 1]$ has to be properly rescaled in order to correctly take the limit $\mu \uparrow \infty$. Therefore we introduce the new 1RSB parameter $y = \mu x$ which stays finite in the close packing limit and takes a value in $[0, +\infty)$. The reweighting factor $e^{-y\omega_{\text{iter}}^{r\to i}}$ is defined as

$$e^{-y\omega_{\text{iter}}^{r\to i}} = \frac{Z_{r\to i}}{\prod_{j\in\partial r\setminus i}\prod_{s\in\partial j\setminus r} Z_{s\to j}}. \tag{32}$$

Last equation combined with (13) and (14) gives $\omega_{\text{iter}}^{r\to i} = -t_{r\to i}$. Averaging over the whole ensemble we can then write the zero temperature 1RSB message passing rules (also called Survey Propagation equations):

$$q' \overset{\text{d}}{=} \frac{\prod_{j=1}^{C}\left(1 - \prod_{s=1}^{D_j} q_{sj}\right)}{e^y + (1 - e^y)\prod_{j=1}^{C}\left(1 - \prod_{s=1}^{D_j} q_{sj}\right)}. \tag{33}$$

In preceding equation $\{q_{sj}\}$ are i.i.d.r.v. on $[0, 1]$ and, as usual, C is the random variable residual degree and $\{D_j\}$ are random independent factors residual degrees. Fixed points of (33) take the form

$$P(q) = p_0\delta(q) + p_1\delta(q - 1) + p_2 P_2(q), \tag{34}$$

where $P_2(q)$ is a continuous distribution on $[0, 1]$ and $p_2 = 1 - p_0 - p_1$. Parameters p_0 and p_1 have to satisfy the closed equations

$$\begin{aligned} p_1 &= \mathbb{E}_C\left(1 - \mathbb{E}_D(1 - p_0)^D\right)^C,\\ p_0 &= 1 - \mathbb{E}_C\left(1 - \mathbb{E}_D p_1^D\right)^C. \end{aligned} \tag{35}$$

Solutions of (35) with $p_2 = 0$ correspond to replica symmetric solutions and their instability marks the onset of a spin glass phase. In this new phase the MSPs are clustered according to the general scenario displayed by constraint satisfaction problems [24].

From the stable fixed point of (33) we can calculate the 1RSB free energy functional $\phi(y)$ as

$$-y\phi(y) = \frac{d}{c}\mathbb{E}\log\left[(1-e^y)\prod_{j=1}^{C_0}\left(1-\prod_{s=1}^{D_j}q_{sj}\right)+e^y\right] + \mathbb{E}(1-D_0)\log\left[(1-e^y)\left(1-\prod_{r=1}^{D}q_r\right)+e^y\right], \tag{36}$$

and 1RSB density, $\rho_{1\mathrm{RSB}}(y) = -(\partial y\phi(y)/\partial y)$, as

$$\rho_{1\mathrm{RSB}}(y) = \frac{d}{c}\mathbb{E}\frac{e^y\left(1-\prod_{j=1}^{C_0}\left(1-\prod_{s=1}^{D_j}q_{sj}\right)\right)}{(1-e^y)\prod_{j=1}^{C_0}\left(1-\prod_{s=1}^{D_j}q_{sj}\right)+e^y} + \mathbb{E}(1-D_0)\frac{e^y\prod_{r=1}^{D_0}q_r}{(1-e^y)\left(1-\prod_{r=1}^{D_0}q_r\right)+e^y}, \tag{37}$$

with expectations intended over $\mathbb{G}$ and over fixed point messages $\{q_s\}$ and $\{q_{sj}\}$. Since the free energy functional $\phi(y)$ and the complexity $\Sigma(\rho)$ are related by the Legendre transform

$$\Sigma(\rho) = -y\rho - y\phi(y), \tag{38}$$

with $\partial\Sigma/\partial\rho = -y$, through (36) we can compute the complexity taking the inverse transform. Equilibrium states, that is, MSPs, are selected by

$$\rho_{1\mathrm{RSB}} = \arg\max_{\rho}\{\rho : \Sigma(\rho) \geq 0\} \tag{39}$$

or equivalently taking the 1RSB parameter y to be

$$y_s = \arg\max_{y\in[0,+\infty]}\phi(y). \tag{40}$$

In the static 1RSB phase we expect $\Sigma(\rho_s) = 0$ so that from (38) we have $\phi(y_s) = -\rho_s$. We will see that this is generally true except for the ensemble $\mathbb{G}_{RP}(2,c)$ of Section 7, corresponding to maximum matchings on ordinary graphs, where the equilibrium state have maximal complexity and $y_s = +\infty$. The relation

$$\rho_{1\mathrm{RSB}} = -\phi(y_s) \tag{41}$$

is always valid though, since for $y_s \uparrow \infty$ complexity stays finite.

5. A Heuristic Algorithm and Exact Results

In this section we propose a heuristic greedy algorithm to address the problem of MSP. It is a natural generalization of the algorithm that Karp and Sipser proposed to solve the maximum matching problem on Erdös-Rényi random graphs [8]; therefore, we will call it generalized Karp-Sipser (GKS). Extending their derivation concerning the leaf removal part of the algorithm we are able to prove that the RS prediction for MSP density is exact as long as the stability criterion (27) is satisfied. We will not give the proofs of the following theorems as they are lengthy but effortless extension of those given in [8]. In order to find the maximum matching on a graph Karp and Sipser noticed that as long as the graph contains a node of degree one (a leaf), its unique edge has to belong to one of the perfect matchings.

They considered the simplest randomized algorithm one can imagine: as long as there is any leaf remove it from the graph, otherwise remove a random edge; then iterate until the graph is depleted. They studied the behaviour of this leaf-removal algorithm on random graphs and were able to prove that it grants w.h.p a maximum matching (within an $o(n)$ error).

To generalize some of their results we need to extend the definition of leaf to that of pendant. We call pendant a variable node whose factor neighbours all have degree one, except for one at most. Stating the same concept in different words, all of the neighbours of a pendant have the pendant itself as their sole neighbour, except for one of them at most. See Figure 1 for a pictorial representation of a pendant (in red). The GKS algorithm is articulated in two phases: a pendant removal and a random occupation phase. We give the pseudocode for the generalized Karp-Sipser algorithm (see Algorithm 1).

At each step the algorithm prioritizes the removal of pendants over that of random variable nodes. We notice that the removal of a pendant is always an optimal choice in order to achieve a MSP; we have no guarantees though on the effect of the occupation of a random node. We call `phase 1` the execution of the algorithm up to the point where the first nonpendant variable is added to V'. It is trivial to show that `phase 1` is enough to find an MSP on a tree factor graph. The interesting thing though is that `phase 1` is also able to deplete nontree factor graphs and find an MSP as long as the factor graphs are sufficiently sparse and large enough.

We call core the subset of the variable nodes which has not been assigned to the MSP in `phase 1`. We will show how the emergence of a core is directly related to replica symmetry breaking. Let us define as usual

$$f(x) \equiv \mathbb{E}_C\left(1-\mathbb{E}_D x^D\right)^C. \tag{42}$$

The function f is continuous, nonincreasing, and satisfies the relation $1 \geq f(0) \geq f(1) = \mathbb{P}[C=0]$. It turns out that, in the large graph limit, `phase 1` of GKS is characterized by the solutions of the system of equations

$$\begin{aligned} p &= f(r), \\ r &= f(p). \end{aligned} \tag{43}$$

We notice that system (43) is equivalent to 1RSB equation (35) once the substitutions $p \to p_0$ and $r \to 1-p_1$ are made.

Lemma 1. *The system of* (43) *admits always a (unique) solution* $p_* = r_*$.

Proof. By monotony and continuity the function $g(p) = f(p) - p$ has a single zero in the interval $[0,1]$. □

Require: a factor graph $G = (V, F, E)$
Ensure: a set packing V'
  $V' = \emptyset$
  add to V' all isolated variable nodes and remove them from G
  remove from G any isolated factor node
  while V is not empty **do**
    if G has any pendant **then**
      choose a pendant i uniformly at random
      add i to V'
      remove i from G, then remove its factor neighbours and their variable neighbours
    else
      pick uniformly at random a variable node and add it to V', remove it from G, then remove its factor neighbours and their variable neighbours
    end if
    remove from G any isolated factor node
  end while
  Return V'

Algorithm 1: Algorithm 1 generalized Karp-Sipser (GKS).

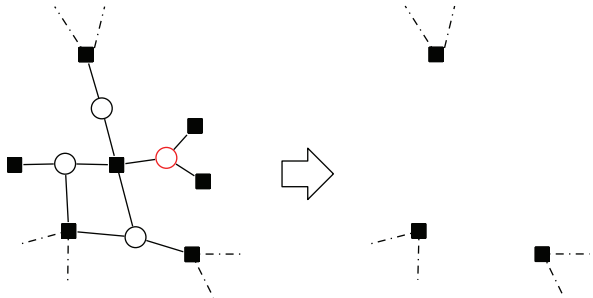

Figure 1: (left) A representation of a factor graph containing a pendant, depicted in red. (right) The factor graph after the removal of the pendant as occurs in the inner part of the while loop in Algorithm 1.

If other solutions are present the relevant one is the one with smallest p, as the following theorems certify.

Theorem 2. *Let (p, r) be the solution with smallest p of* (43). *Then the density of factor nodes surviving* `phase 1` *in the large graphs limit is given by*

$$\psi_1 = 1 + \widetilde{p} - \widetilde{r} + cp\mathbb{E}_D\left[p^D - r^D\right] \tag{44}$$

with

$$\begin{aligned} \widetilde{r} &= \mathbb{E}_{C_0}\left(1 - \mathbb{E}_D p^D\right)^{C_0}, \\ \widetilde{p} &= \mathbb{E}_{C_0}\left(1 - \mathbb{E}_D r^D\right)^{C_0}. \end{aligned} \tag{45}$$

In particular if the smallest solution is $p_* = r_*$ the graph is depleted with high probability in `phase 1`.

As already said we will not give the proof of this and of the following theorem, as they are lengthy and can be obtained from the derivation given in Karp and Sipser's article [8] with little effort even if not in a completely trivial way. Theorem 2 affirms that as soon as (43) develops another solution a core survives `phase 1`. This phenomena coincides with the need for replica symmetry breaking in the cavity formalism. Let us now establish the exactness of the cavity prediction for ρ in the RS phase.

Theorem 3. *Let (p, r) be the solution with smallest p of* (43). *Then the density of variables assigned in* `phase 1` *in the large graphs limit is*

$$\rho_1 = \frac{d}{c}\left(1 - \widetilde{r}\right) + \mathbb{E}_{D_0}\left(1 - D_0\right) p^{D_0}, \tag{46}$$

where $\widetilde{p}$ has been defined in previous theorem. In particular if the smallest solution is $p_* = r_*$ the replica symmetric cavity method prediction (25) *is exact.*

These two theorems imply that the RS prediction for ρ, (25), holds true as long as `phase 1` manages to deplete w.h.p. the whole factor graph. A similar behaviour has been observed in other combinatorial optimization problem, for example, the random XORSAT [25]. Conversely it is easy to prove, given the equivalence between (43) and (35), that if a solution with $p \neq p_*$ exists the RS fixed point is unstable in the 1RSB distributional space. Therefore the system is in the RS phase if and only if (35) admits a unique solution.

We notice that we could have also looked for the solutions of the single equation $p = f^2(p)$ instead of the two of (43), since the value of r is uniquely determined by the value of p. Then previous theorems state that the RS results hold as long as f^2 has a single fixed point. In [22] it has been proven that the RS solution of the weighted maximum independent set and maximum weighted matching holds true as long as the corresponding squared cavity operator f^2 has a unique fixed point. Since those are special cases of the weighted MSP

problem, we conjecture that the f^2 condition holds for the general case too, as it does for the unweighted case. Probably it would be possible to further generalize the Karp-Sipser algorithm to cover analytically the weighted case as well.

The scenario arising from the analysis of the algorithm and from 1RSB cavity method is the following: in the RS phase maximum set packings form a single cluster and it is possible to connect any two of them with a path involving MSPs separated by a rearrangement of a finite number of variables [26]. In the RSB phase instead MSPs can be grouped into many connected (in the sense we mentioned before) clusters, each one of them defined by the assigning of the variables in the core. Two MSPs who differ on the core are always separated by a global rearrangement of the variables (i.e., $O(N)$). In presence of a core the GKS algorithm makes some suboptimal random choices (after `phase 1`).

We did not make an analytical study of the random variable removal part of the GKS algorithm. This has been done by Wormald for the matching problem, associating to the graph process a set of differential equations amenable to analysis [9] (something similar has been done for random XORSAT as well [25]). In that case it turns out that the algorithm still achieves optimality and yields an almost perfect matching on the core. An appropriate analysis of the second phase of the GKS algorithm and the reason of its failure for most of the SP ensembles we covered deserve further studies.

6. Numerical Simulations

While the RS cavity (24) and (25) can be easily computed, the numerical solution of the 1RSB density evolution (33) is much more involved (although simplified by the zero temperature limit) and has been obtained with a standard population dynamics algorithm [27].

The implementation of the GKS algorithm is straightforward. While it is very fast during `phase 1`, we noticed a huge slowing down in the random removal part. We were able to find set packings for factor graphs with hundred of thousands of nodes.

In order to test the cavity and GKS predictions, we also computed the exact MSP size on factor graphs of small size. First we notice that a set packing problem, coded in a factor graph structure F, is equivalent to an independent set problem on an appropriate ordinary graph G. The node set of G will be the variable node set of F and we add an edge to G between each pair of nodes having a common factor node neighbour in F. Therefore each neighborhood of a factor node in F forms a clique in G. We then solve the maximum set problem for G using an exact algorithm [28] implemented in the igraph library [29]. Since the time complexity is exponential in the size of the graph we performed our simulations on graphs containing up to only one hundred nodes.

7. Applications to Problem Ensembles

We will now apply the methods developed in the previous sections to some factor graph ensembles, each modelling a class of MSP problem instances. We consider graph ensembles containing nodes with Poissonian random degree or regular degree: $\mathbb{G}_{PR}(N,d,c)$, $\mathbb{G}_{RP}(N,d,c)$, $\mathbb{G}_{PP}(N,d,c)$, and $\mathbb{G}_{RR}(N,d,c)$. Subscript R or P indicates whether the type of nodes to which they refer (variable nodes for the first subscript, factor nodes for the second) have regular or Poissonian random degree, respectively. We parametrize these ensemble by their average variable and factor degree; that is, $d = \mathbb{E}_{D_0} D_0$ and $c = \mathbb{E}_{C_0} C_0$. They are constituted by factor graphs having N variable nodes and $M = \lfloor Nd/c \rfloor$ factor nodes but they differ both for their elements and for their probability law. We define the ensembles giving the probability of sampling one of their elements.

(i) $\mathbb{G}_{RP}(N,d,c)$: each element G has N variables and $M = \lfloor Nd/c \rfloor$ factors. Every variable node has fixed degree d. G is obtained linking each variable with d factors chosen uniformly and independently at random. The factor nodes degree distribution obtained is Poissonian of mean c with high probability. This ensemble is a model for the k-set packing and will be the main focus of our attention.

(ii) $\mathbb{G}_{PR}(N,d,c)$: each element G has N variables and $M = \lfloor Nd/c \rfloor$ factor. Every factor node has fixed degree c. G is built linking each factors with c variables chosen uniformly and independently at random. The variable nodes degree distribution obtained is Poissonian of mean d with high probability.

(iii) $\mathbb{G}_{PP}(N,d,c)$: each element G has N variables and $M = \lfloor Nd/c \rfloor$ factors. G is built adding an edge (i,r) with probability c/N independently for each choice of a variable i and a factor r. The factor graph obtained has w.h.p Poissonian variable nodes degree distribution of mean d and Poissonian factor nodes degree distribution of mean c.

(iv) $\mathbb{G}_{RR}(N,d,c)$: it is constituted of all factor graphs of N variable nodes of degree d and $M = Nd/c$ factor nodes of degree c (Nd has to be multiple of c). Every factor graph of the ensemble is equiprobable and can be sampled using a generalization of the configuration model for random regular graph [30]. This ensemble is a model for the k-set packing.

We will omit the argument N when we refer to an ensemble in the $N \uparrow \infty$ limit.

7.1. $\mathbb{G}_{RP}(d,c)$. This is the ensemble with variable node degrees fixed to d and Poissonian factor node degree that is $C \sim C_0 \sim \text{Poisson}(c)$. The case $d = 2$ corresponds to the maximum matching problem on Erdös-Rényi random graph [8, 12].

Density evolution (22) for $\mathbb{G}_{RP}$ reduces to

$$T' \stackrel{\mathrm{d}}{=} \max\{0\} \cup \left\{ 1 - \sum_{s=1}^{d-1} T_{sj} \right\}_{j \in \{1,\dots,C\}}. \tag{47}$$

Considering distributions of the form $P(t) = p\delta(t) + (1-p)\delta(1-t)$ fixed points of (47) reads

$$p_* = e^{-cp_*^{d-1}} \equiv f(p_*). \tag{48}$$

The last equation admits only one solution for each value of d and c, as it is easily seen through a monotony argument considering the left and right hand side of the equation. The values of c as a function of d satisfying $|f'(p_*)| = 1$ are the critical points $c_s(d)$ delimiting the RS phase ($c < c_s(d)$) and are given by

$$c_s(d) = \frac{e}{d-1}. \tag{49}$$

For $c > e/(d-1)$ a core survives `phase 1` of the GKS algorithm as stated by Theorem 2 and shown in Figure 2. We notice that the same threshold value $e/(d-1)$ is found in the minimum set covering problem on the dual factor graph ensemble [31], where the application of a leaf removal algorithm, that is, `phase 1` of GKS, unveils an analogous core emergence phenomena. In the RS phase, the MSP density is equal to the minimum set covering density, but in the RSB phase we cannot compare the cavity predictions for these two dual problems, since in [31] the 1RSB solution to the minimum set covering is not provided.

In the matching case, that is, $d = 2$ equation (49) expresses the notorious e-phenomena discovered by Karp and Sipser, while for higher values of d it provides an extension of the critical threshold.

We can recover the same critical condition equation (49) through the bug propagation method, as the transfer matrix $P(b_0 \to b_1 \mid a_0 \to a_1)$ non-zero elements are only:

$$\begin{aligned} P(0 \longrightarrow 1 \mid 1 \longrightarrow 0) &= p^{d-1}, \\ P(1 \longrightarrow 0 \mid 0 \longrightarrow 1) &= p^{d-1}, \end{aligned} \tag{50}$$

which give $\lambda_M = p^{d-1}$. The average branching factor is $\overline{CD} = c(d-1)$ so that (30) and (48) yield (49). The analytical value for the relative size of MSPs, that is, the particle density ρ, is

$$\rho(d,c) = \frac{d}{c}(1-p_*) + (1-d)\,p_*^d \quad \text{for } c < c_s(d). \tag{51}$$

In Figure 3 we compared ρ from (51) as a function of c for some values d with an exact algorithm applied to finite factor graphs (as explained in Section 6), both above and below c_s. Clearly for $c > c_s$ the RS approximation is increasingly more inaccurate.

We continue our analysis of $\mathbb{G}_{RP}$ above the critical value c_s through the 1RSB cavity method as outlined in Section 4.2. Fixed point messages of (33) are distributed as

$$P(q) = p_0\delta(q) + p_1\delta(q-1) + p_2 P_2(q), \tag{52}$$

where

$$\begin{aligned} p_1 &= e^{-c(1-p_0)^{d-1}}, \\ p_0 &= 1 - e^{-cp_1^d}, \\ p_2 &= 1 - p_0 - p_1, \end{aligned} \tag{53}$$

and $P_2(q)$ has to be determined through (33). Equation (53) admits always an RS solution $p_1 = 1 - p_0 = p_*$ which is stable up to c_s, as already noticed. Above c_s a new stable fixed point, with $p_2 > 0$, continuously arises and we study it numerically with a population dynamics algorithm.

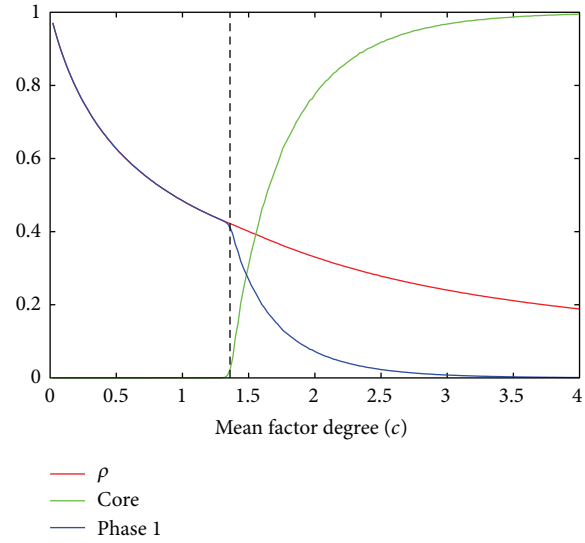

Figure 2: Results from GKS algorithm applied to $\mathbb{G}_{RP}(3,c)$. The set packing size after phase 1 (blue line) and after the algorithm stops (red line) are reported. The vertical line is $c_s = e/2$ and is drawn as a visual aid to determine the point where the RS assumptions stop to hold true. Above c_s a nonempty core emerges continuously (green line).

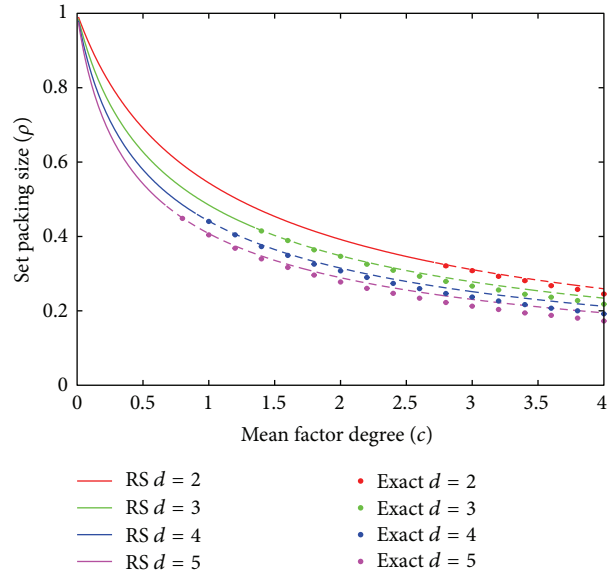

Figure 3: RS cavity method analytical prediction equation (51) for MSP size in $\mathbb{G}_{RP}(d,c)$ compared with the average size of MSPs obtained from 1000 samples of factor graphs with 100 variable nodes (*dots*). Dashed parts of the lines are the RS estimations in the RSB phase, that is, for $c > e/(d-1)$, where it is no longer exact.

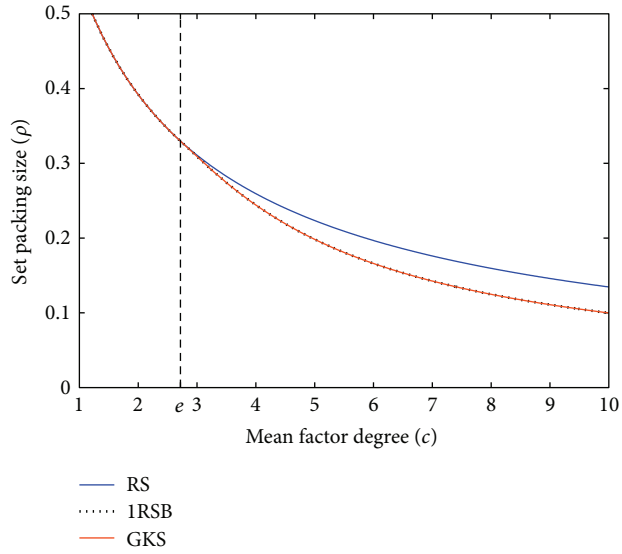

FIGURE 4: Maximum set packing size for $\mathbb{G}_{RP}(d, c)$ as a function of c for $d = 2$. Blue line corresponds to the RS analytical value (51), red line is from the Karp-Sipser algorithm and dotted black line is the 1RSB solution (54) with $y \gg 1$. Karp-Sipser and 1RSB cavity method yield the same and exact result.

The 1RSB free energy, as a function of the Parisi parameter y, takes the form

$$\phi(y) = -\frac{d}{yc}\mathbb{E}\log\left[(1-e^{y})\prod_{j=1}^{C}\left(1-\prod_{s=1}^{d-1}q_{sj}\right)+e^{y}\right] + \frac{d-1}{y}\mathbb{E}\log\left[(1-e^{y})\left(1-\prod_{r=1}^{d}q_{r}\right)+e^{y}\right]. \quad (54)$$

As prescribed by the cavity method, the value y_s which maximizes $\phi(y)$ over $[0, +\infty]$ yields the correct free energy; therefore, we have $\phi(y_s) = -\rho_{1\mathrm{RSB}}$.

Unsurprisingly, as they belong to different computational classes, the cases $d = 2$ and $d \geq 3$ show qualitatively different pictures. In the case of maximum matching on the Poissonian graph ensemble, numerical estimates suggest that complexity is an increasing function of y on the whole real positive axis. Correct choice for parameter y is then $y_s = +\infty$, as already conjectured in [12], and we find that maximum matching size prediction from 1RSB cavity method fully agrees with rigorous results from Karp and Sipser [8] and with the size of the matchings given by their algorithm (see Figure 4). The 1RSB ansatz is therefore exact for $d = 2$.

The $d \geq 3$ case analysis does not yield such a definite result. The complexity of states Σ is no more a strictly increasing function of y. It reaches its maximum in y_d, the choice of y that selects the most numerous states, which could be those where local search greedy algorithms are more likely to be trapped. Then it decreases up to the finite value y_s where complexity changes sign and takes negative values. Therefore y_s is the correct choice for the Parisi parameter which maximizes $\phi(y)$. Plotted as a function of ρ, complexity

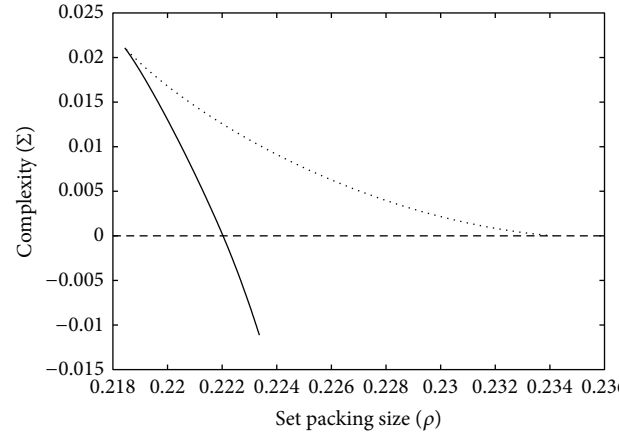

FIGURE 5: Complexity in $\mathbb{G}_{RP}(d, c)$, with $d = 3$ and $c = 4.0$, as a function of the relative MSP size ρ.

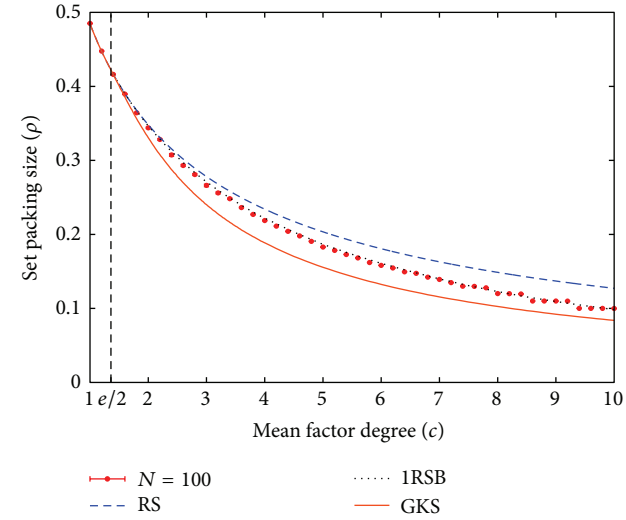

FIGURE 6: MSP size in $\mathbb{G}_{RP}(3, c)$ as a function of c. RS and 1RSB cavity method is confronted with the GKS algorithm and with an exact algorithm on factor graphs of 100 variable nodes.

Σ has a convex nonphysical part, with extrema the RS solution (on the right) and the point corresponding to the dynamic 1RSB solution (on the left) and a concave physically relevant for $y \in [y_d, y_s]$ (see Figure 5). The 1RSB seems to be in very good agreement with the exact algorithm and we are inclined to believe that no further steps of replica symmetry breaking are needed in this ensemble. The GKS algorithm instead falls short of the exact value (see Figure 6); therefore, it constitutes a lower bound which is not strict but at least it could probably be made rigorous carrying on the analysis of the GKS algorithm beyond `phase 1`.

7.2. $\mathbb{G}_{PR}(d,c)$. The ensemble $\mathbb{G}_{PR}(d, c)$ is constitute factor graphs containing factor nodes of degree fixed to c and variable nodes of Poissonian random degree of mean d. It has statistical properties with respect to the MSP problem quite different from those encountered in $\mathbb{G}_{RP}$, as we will readily show. The MSP problem on $\mathbb{G}_{PR}(d, c)$

with $d = 2$ is equivalent to the well known problem of maximum independent sets on Poissonian graphs [14, 32]. The real parameter p characterizing discrete support distributions of messages has to satisfy the fixed point (24) that reads

$$p_* = \left(1 - e^{d(p_*-1)}\right)^{c-1}. \tag{55}$$

At fixed value of c the RS ansatz holds up to the critical value $d_s(c)$, which is implicitly given by first derivative condition (27):

$$(c-1)\, p_*^{(c-2)/(c-1)} e^{(p_*-1)d_s} d_s = 1. \tag{56}$$

Although critical condition equation (56) is not as elegant as the one we obtained for the ensemble $\mathbb{G}_{RP}$, it can be easily solved numerically for d_s as a function of c. For $c = 2$ the threshold value is exactly $d_s(2) = e$. For $c > 2$ instead d_s is an increasing function of the factor degree c. Thanks to (25) we readily compute the MSP size in the RS phase:

$$\rho = \frac{d}{c}\left(1 - p_*^{c/(c-1)}\right) + \left(1 - p_* d\right) e^{d(p_*-1)} \quad \text{for } c < c_s(d). \tag{57}$$

We can see in Figure 7 that the MSP size $\rho(d, c)$ is a decreasing function in both arguments as expected. Equation (57) can be taken as the RS estimate for MSP size for $d > d_s(c)$. The RS estimate is strictly greater than the average size of SPs given by the GKS algorithm at all values of $d > d_s(c)$ (see Figure 7).

7.3. $\mathbb{G}_{PP}(d,c)$. We will now briefly examine our MSP model on the ensemble $\mathbb{G}_{PP}(d, c)$ where both factor nodes and variable nodes have Poissonian random degrees of mean c and d, respectively. From (23) and (24) we obtain the fixed point condition for the probability distribution of messages:

$$p_* = e^{-ce^{d(p_*-1)}}. \tag{58}$$

As usual p is the parameter characterizing the distribution of messages, $P(t) = p\delta(t) + (1-p)\delta(1-p)$. Equation (58) admits one and only one fixed point solution p_* for each value of d and c. In fact f is continuous, strictly decreasing, and $f(0) > 0$, $f(1) < 1$. The first derivative condition $|f'(p)| = 1$ defines the critical line $d_s(c)$ through

$$d_s(c) = -\frac{1}{p_* \log(p_*)}. \tag{59}$$

The curve $d_s(c)$ separates the RS phase from the RSB phase in the $c - d$ parametric space (see Figure 8). The unbounded RS region shares some resemblance with the corresponding (although d-discretized) region of $\mathbb{G}_{PR}(d, c)$ and is at variance with the compact area of the RS phase in $\mathbb{G}_{RP}$.

We can compute the MSP relative size ρ through (25) and obtain

$$\rho = \frac{d}{c}(1 - p_*) + (1 - dp_*) e^{d(p-1)} \quad \text{for } d < d_s(c), \tag{60}$$

which holds only as an approximation in the RSB phase. We can see from Figure 8 that ρ is decreasing both in c and d as was observed in the other ensembles as well.

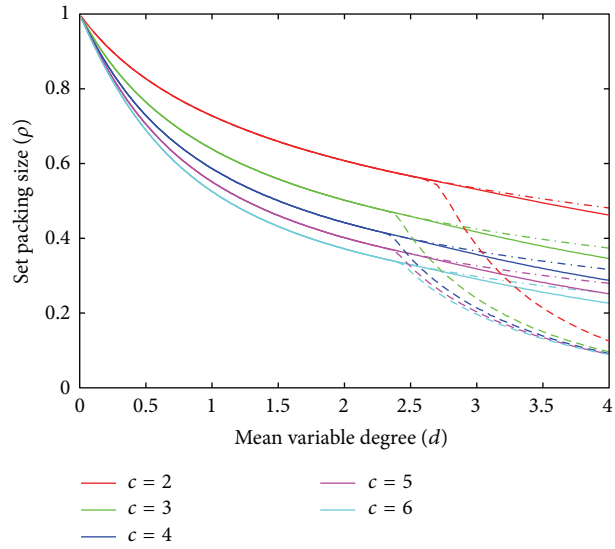

FIGURE 7: RS cavity method analytical prediction (57) for MSP size in $\mathbb{G}_{PR}(d, c)$ (dot-dashed) compared with the set packing size from `phase 1` of the GKS algorithm (dashed lines) and with a complete run of the algorithm (continuous lines).

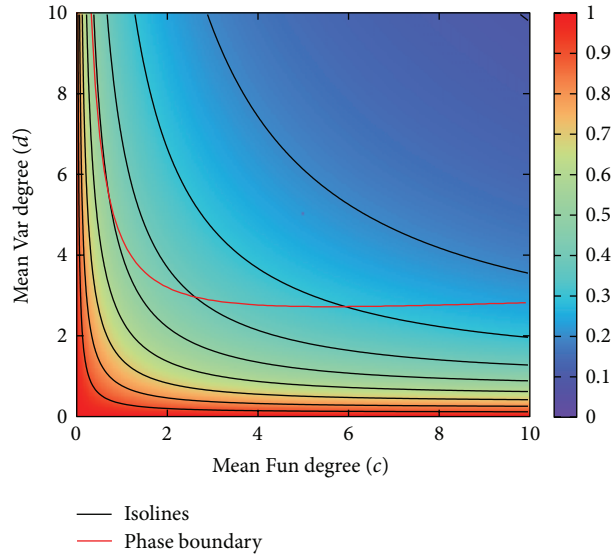

FIGURE 8: MSP relative size for the ensemble $\mathbb{G}_{PP}(d, c)$ as given by (60). Above the boundary separating RS and RSB phase values are only approximation to the exact value of ρ.

7.4. $\mathbb{G}_{RR}(d,c)$. The MSP problem on the ensemble $\mathbb{G}_{RR}(d, c)$, the straightforward generalization to factor graphs of the random regular graphs ensemble, poses some simplification to the cavity formalism thanks to his homogeneity. It has already been the object of preliminary studies by Weigt and Hartmann [16] and then a much more deep work of Hansen-Goos and Weigt [17] who disguised it as a hard spheres model on a generalized Bethe lattice. The authors studied through the cavity method this hard spheres model on $\mathbb{G}_{RR}(d, c)$ both at finite chemical potential μ and in the close packing limit and found out that the 1RSB solution is unstable in the close

packing limit, therefore suggesting the need of a full RSB treatment of the problem.

8. Conclusions

We studied the average asymptotic behaviour of random instances of the maximum set packing problem, both from a mathematical and a physical viewpoint. We contributed to the known list of models where the replica symmetric cavity method can be proven to give exact results, thanks to the generalization of an algorithm (and of its analysis) first proposed by Karp and Sipser [8]. Moreover, our analysis address a problem reported in recent work on weighted maximum matchings and independent sets on random graphs [22], where the authors could not extend their results to the unweighted cases. We achieve here the desired result making use the grand canonical potential instead of the direct computation of single variable expectations. We also extend their condition for the system to be in what physicists call a replica symmetric phase, namely, the uniqueness of the fixed point of the square of a certain operator (which is the analogue of the one defined in (42)), to the more general setting of maximum set packing (although without weights). On some problem ensembles, where the assumptions of Theorems 2 and 3 no longer hold and the RS cavity method fails, we used the 1RSB cavity method machinery to obtain an analytical estimation of the MSP size. Numerical simulations show very good agreement of the 1RSB estimation with the exact values, although comparisons have been done only with small random problems due to the exact algorithm being of exponential time complexity. The GKS algorithm instead generally fails to find any MSPs except for some special instances of the problem.

Some questions remain open for further investigation. To validate the 1RSB approach the stability of the 1RSB solution has to be checked against more steps of replica symmetry breaking. Moreover a thorough analysis of the second phase of the GKS algorithm could shed some light on the mechanism of replica symmetry breaking and give a rigorous lower bound to the average maximum packing size.

Regarding the problem of looking for optimal solutions in single samples, we believe that an efficient heuristic algorithm, able to obtain near-to-optimal configurations also in the RSB phase, could be constructed following the ideas of [33].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research has received financial support from the European Research Council (ERC) through Grant agreement no. 247328 and from the Italian Research Minister through the FIRB project no. RBFR086NN1.

References

[1] S. Micali and V. V. Vazirani, "An O(|*v*||v||c|E||) algoithm for finding maximum matching in general graphs," in *Proceedings of the 21st Annual IEEE Symposium on Foundations of Computer Science*, pp. 17–27, 1980.

[2] N. Alon, J.-H. Kim, and J. Spencer, "Nearly perfect matchings in regular simple hypergraphs," *Israel Journal of Mathematics*, vol. 100, pp. 171–187, 1997.

[3] Z. Füredi, J. Kahn, and P. D. Seymu, "On the fractional matching polytope of a hypergraph," *Combinatorica*, vol. 13, no. 2, pp. 167–180, 1993.

[4] Y. H. Chan and L. C. Lau, "On linear and semidefinite programming relaxations for hypergraph matching," *Mathematical Programming*, vol. 135, pp. 123–148, 2011.

[5] V. Rödl and L. Thoma, "Asymptotic packing and the random greedy algorithm," *Random Structures & Algorithms*, vol. 8, no. 3, pp. 161–177, 1996.

[6] M. M. Halldórsson, "Approximations of weighted independent set and hereditary subset problems," *Journal of Graph Algorithms and Applications*, vol. 4, no. 1, pp. 1–6, 2000.

[7] E. Hazan, S. Safra, and O. Schwartz, "On the complexity of approximating *K*-set packing," *Computational Complexity*, vol. 15, no. 1, pp. 20–39, 2006.

[8] R. M. Karp and M. Sipser, "Maximum matchings in sparse random graphs," in *Proceedings of the 22nd Annual Symposium on Foundations of Computer Science*, vol. 12, pp. 364–375, IEEE Computer Society, 1981.

[9] N. C. Wormald, "The differential equation method for random graph processes and greedy algorithms," in *Lectures on Approximation and Randomized Algorithms*, pp. 73–155, 1999.

[10] G. Parisi and M. Mézard, "Replicas and optimization," *Journal de Physique Lettres*, vol. 46, no. 17, pp. 771–778, 1985.

[11] G. Parisi and M. Mézard, "A replica analysis of the travelling salesman problem," *Journal de Physique*, vol. 47, pp. 1285–1296, 1986.

[12] H. Zhou and Z.-C. Ou-Yang, "Maximum matching on random-graphs," In press, http://arxiv.org/abs/cond-mat/0309348.

[13] L. Zdeborová and M. Mézard, "The number of matchings in random graphs," *Journal of Statistical Mechanics*, vol. 2006, Article ID P05003, 2006.

[14] L. Dallasta, A. Ramezanpour, and P. Pin, "Statistical mechanics of maximal independent sets," *Physical Review E*, vol. 80, Article ID 061136, 2009.

[15] M. Mézard and M. Tarzia, "Statistical mechanics of the hitting set problem," *Physical Review E*, vol. 76, no. 4, Article ID 041124, 10 pages, 2007.

[16] M. Weigt and A. K. Hartmann, "Glassy behavior induced by geometrical frustration in a hard-core lattice-gas model," *Europhysics Letters (EPL)*, vol. 62, Article ID 021005, p. 533, 2003.

[17] H. Hansen-Goos and M. Weigt, "A hard-sphere model on generalized bethe lattices: statics," *Journal of Statistical Mechanics*, vol. 2005, Article ID P04006, 2005.

[18] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[19] A. Montanari and M. Mézard, *Information, Physics and Computation*, Oxford University Press, 2009.

[20] G. Parisi, M. Mézard, and M. A. Virasoro, *Spin Glass Theory and Beyond*, World Scientific, Singapore, 1987.

[21] M. Mézard and G. Parisi, "The cavity method at zero temperature," *Journal of Statistical Physics*, vol. 22, no. 1-2, pp. 1–34, 2003.

[22] D. Gamarnik, T. Nowicki, and G. Swirszcz, "Maximum weight independent sets and matchings in sparse random graphs. Exact results using the local weak convergence method," *Random Structures and Algorithms*, vol. 28, no. 1, pp. 76–106, 2006.

[23] A. Montanari, G. Parisi, and F. Ricci-Tersenghi, "Instability of one-step replica-symmetry-broken phase in satisfiability problems," *Journal of Physics A*, vol. 37, no. 6, p. 2073, 2004.

[24] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová, "Gibbs states and the set of solutions of random constraint satisfaction problems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 25, pp. 10318–10323, 2007.

[25] M. Mézard, F. Ricci-Tersenghi, and R. Zecchina, "Two solutions to diluted *p*-spin models and XORSAT problems," *Journal of Statistical Physics*, vol. 111, no. 3-4, pp. 505–533, 2002.

[26] G. Semerjian, "On the freezing of variables in random constraint satisfaction problems," *Journal of Statistical Physics*, vol. 130, no. 2, pp. 251–293, 2008.

[27] M. Mézard and G. Parisi, "The bethe lattice spin glass revisited," *The European Physical Journal B*, vol. 20, no. 2, pp. 217–233, 2001.

[28] S. Tsukiyama, M. Ide, H. Ariyoshi, and I. Shirakawa, "A new algorithm for generating all the maximal independent sets," *SIAM Journal on Computing*, vol. 6, no. 3, pp. 505–517, 1977.

[29] G. Csárdi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, Complex Systems*, p. 1695, 2006.

[30] N. C. Wormald, "Models of random regular graphs," in *Surveys in Combinatorics*, London Mathematical Society Lecture Note, pp. 239–298, 1999.

[31] S. Takabe and K. Hukushima, "Minimum vertex cover problems on random hypergraphs: replica symmetric solutionand a leaf removal algorithm," In press, http://arxiv.org/abs/1301.5769.

[32] S. Sanghavi, D. Shah, and A. S. Willsky, "Message passing for maximum weight independent set," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 4822–4834, 2009.

[33] M. Weigt and H. Zhou, "Message passing for vertex covers," *Physical Review E*, vol. 74, Article ID 046110, 19 pages, 2006.