Collegio Carlo Alberto

# The stochastic matching problem

Fabrizio Altarelli
Alfredo Braunstein
Abolfazl Ramezanpour
Riccardo Zecchina

**No. 182**
**December 2010**

Carlo Alberto Notebooks

**www.carloalberto.org/working_papers**

© 2010 by Fabrizio Altarelli, Alfredo Braunstein, Abolfazl Ramezanpour and Riccardo Zecchina. Any opinions expressed here are those of the authors and not those of the Collegio Carlo Alberto.

# The stochastic matching problem

F. Altarelli,[1,2] A. Braunstein,[3] A. Ramezanpour,[1] and R. Zecchina[1,2]

[1] *Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*
[2] *Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy*
[3] *Human Genetic Foundation, Torino, via Nizza 52, and Centre for Computational Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

(Dated:)

Stochastic matching is the stochastic version of the well understood and famous matching problem. It consists in finding the minimum final cost of a matching given just some partial information about the structure of the problem. Both the deterministic and the stochastic versions of the matching problem play a basic role in combinatorial optimization and in statistical mechanics. However, while the deterministic case can be solved in polynomial time, its stochastic variants are wort case intractable. We propose a novel method which combines some features of the survey propagation equations and of the cavity method which allows us to solve stochastic matching problems efficiently. We apply our approach to stochastic matching over random bipartite graphs for which analyze the phase diagram and compare the results with exact bounds. In the replica symmetry broken region where smart heuristics fail, our approach is shown numerically to be highly efficient. Eventually we discuss how the method can be generalized to other problems of optimization under uncertainty.

One important aspect of the statistical physics approaches to disordered systems is the broad range of their interdisciplinary applications. Systems with frustration, structural disorder and uncertainties are in fact ubiquitous in many fields of science and their study has greatly benefited from the algorithms which have emerged at the interface between statistical physics of disordered systems and computer science.

One of the key problems has been the so called matching problem, which was among the firsts to be solved by statistical physics methods and later by rigorous mathematical techniques.

Matching is a constituent part of many problems in different fields, ranging from physics (dimers models), to computers science (vision), economics (auctions) and computational biology (pattern matching ). Among its virtues there is simplicity of formulation: given a weighted graph $G$, find a matching of nodes on $G$ which maximized the total weight. From the computer science point of view, it belongs to the class of polynomial complexity.

The stochastic version of matching is a basic example of *optimization under uncertainty*, namely the problem of finding the minimum of a cost function which depends on some stochastic parameters, given just some partial information about their value (e.g. their probability distribution). Problems of this kind pervade all areas of natural and applied sciences, but they are extremely difficult to solve. Stochastic matching problems are in fact known to belong to higher computational complexity classes ranging from NP-hard to PSPACE depending on how stochasticity is introduced.

Most real-world optimization problems involve uncertainty: the precise value of some of the parameters is often unknown, either because they are measured with insufficient accuracy, or because they are stochastic in nature and revealed only *after* some decisions have been taken. The objective of the optimization process is thus to find solutions which are optimal in some probabilistic sense, a fact which introduces fundamental conceptual and computational challenges [1].

Here we propose a new approach to stochastic matching which builds on the formalism of Survey Propagation (SP) [17–19] and of the cavity method. The method is partly analytic and allows to optimize the expectation of a stochastic cost function by estimating the statistics of its minima, without resorting to explicit (and costly) sampling techniques. We focus on the so called two-stage stochastic matching, in which some of the variables have to be assigned before the stochastic parameters are specified (first stage), and the remaining variables are assigned after (second stage). This version of the problem is known to be NP-hard.

In what follows we illustrate our new approach by studying the stochastic maximum size Matching problem with uncertainties the second stage vertices. We derive the phase diagram of the problem in the case of random bipartite graphs and show how our new technique is effective also in the replica symmetry broken region where usual heuristics fail. Eventually we discuss how the method can be applied to other stochastic optimization problems, giving few examples.

*Solving two-stage stochastic optimization problems.* The problem we study is defined by an energy function $\mathcal{E}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{t})$ depending on two sets of variables $\mathbf{x}_1, \mathbf{x}_2$ and one set of independent stochastic parameters $\mathbf{t}$ with distribution $P(\mathbf{t})$ and consists in computing

$$\mathbf{x}_1^* = \arg\min_{\mathbf{x}_1} \mathbb{E}_{\mathbf{t}} \min_{\mathbf{x}_2} \mathcal{E}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{t})\,. \tag{1}$$

The method we propose consists in computing the

chain of operations in (1) by performing the minimizations with MS and the expectation with BP. Let us first briefly remind the BP and MS equations, for a system with unknows $s_{ij}$ associated to edges on a graph $G = (V, E)$ and an energy $H(\mathbf{s}) = \sum_i H_i(\mathbf{s}_i)$ where $\mathbf{s}_i = \{s_{ij} : (ij) \in E\}$. The method can be easily generalized to any graphical model. The BP Equations consist on a method to compute statistical properties of a Gibbs distribution $P(\mathbf{s}) = \frac{1}{Z} \exp(-\beta H(\mathbf{s}))$. Specifically, these are equations for certain single variable probabilities $m_{ij}(s_{ij})$ (called "messages") indexed by *directed* edges $(ij) \in E$ on the graph:

$$m_{ij}(s_{ij}) \propto \sum_{\{s_{ki}:k\in\partial i\setminus j\}} \prod_{k\in\partial i\setminus j} m_{ki}(s_{ki}) e^{-\beta H_i(\mathbf{s}_i)} \qquad (2)$$

where we used the notation $\partial i = \{k : (ik) \in E\}$. From a solution of these equations (normally obtained by an itererative scheme), one can compute several useful quantities, such as the average of an observable (like the energy of the system or the marginal probabilities $P_{ij}(s_{ij}) \propto m_{ij}(s_{ij}) m_{ji}(s_{ij})$) or the entropy of the system. On the $\beta \to \infty$ limit, the Gibbs measure concentrates on the minimums of $H$ and the MS equations can be derived as an explicit $\beta = \infty$ limit of (2) after the change of variables $h_{ij} = \frac{1}{\beta} \log m_{ij}$:

$$h_{ij}(s_{ij}) = \max_{\{s_{ki}:k\in\partial i\setminus j\}} \left\{ \sum_{k\in\partial i\setminus j} h_{ki}(s_{ki}) - H_i(\mathbf{s}_i) \right\} \qquad (3)$$

where the equation holds up to an additive constant independent from $s_{ij}$. Naturally, the $\beta \to \infty$ expression for the BP average energy gives an expression for the minimum energy as a function of MS messages.

We will now describe a general method for solving TSSOs. Proceed as follows. **(i)** Call $\chi_{ij}$ the MS equation (3) for the inner minimum in (1), and build the following energy function of $\mathbf{t}$ and the MS messages: $-\log\left(P(\mathbf{t}) \prod_{ij} \mathbb{1}[\chi_{ij}]\right)$, which equals $+\infty$ if $\chi_{ij}$ is not satisfied for some $ij$. **(ii)** Obtain BP equations (2) for this energy function. These equations can be considered SP equations [19]. **(iii)** Treat the expression for the minimum energy from the MS of the first step as an observable and compute an expression $\mathcal{E}^*(\mathbf{x}_1)$ for its average as a function of BP messages. Up to here, variables $\mathbf{x}_1$ have been considered constant. **(iv)** Finally, employ MS again to find the minimum of $\mathcal{E}^*(\mathbf{x}_1)$ over both messages and $\mathbf{x}_1$, where messages are contrained by the BP equations.

*Application to stochastic matching.* We shall now illustrate the general method by applying it to a stochastic bipartite matching problem, otherwise known as stochastic assignment problem, introduced in [21–23] and defined as follows (refer to Figure 1 for notations). An instance of the problem is specified by: 1) a bipartite graph $G = (L, R, E)$ in which the set $L$ is partitioned in 2 subsets, $L = L_1 \cup L_2$; and 2) a set of independent probabilities $\{p_l \in ]0, 1] : l \in L_2\}$. In the first stage, vertices in $L_1$ must be matched to vertices in $R$. Then, a subset of vertices $\mathcal{L}_2 \subset L_2$ is extracted such that each $l \in L_2$ is present in $\mathcal{L}_2$ with probability $p_l$, and the vertices in $\mathcal{L}_2$ must be matched to the remaining vertices in $R$ (or are left unmatched). The objective is to minimize the number of vertices that remain unmatched in $L_1 \cup \mathcal{L}_2 \cup R$.

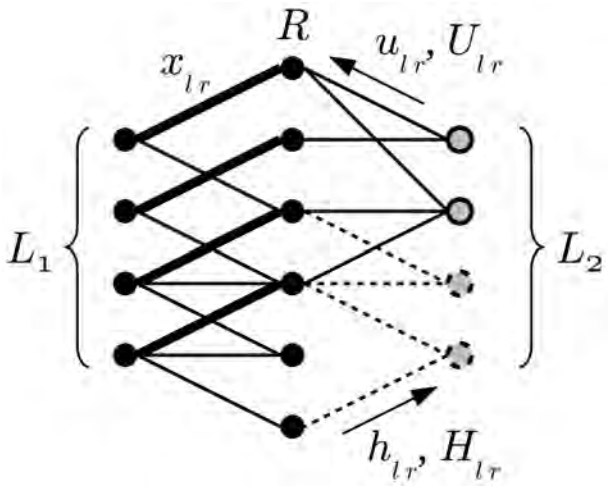


FIG. 1. Example graph for stochastic bipartite matching. The full graph is known in advance, but $L_2$ nodes may or may not be present, with known probabilities. Nodes in $L_1$ must be matched with nodes in $R$ on the first stage, and then nodes in $L_2$ are extracted (the dotted nodes and links turn out to be absent in the realization given). A greedy decision on the first stage (bold links) leads to a suboptimal final configuration in this case. Assignment variables $x_{lr}$ and messages traveling on links ($h_{lr}$,$H_{lr}$,$u_{lr}$,$U_{lr}$) are also shown.

A configuration of the problem can be described associating a binary variable $x_{lr} \in \{0, 1\}$ to each edge $(lr) \in E$ where $x_{lr} = 1$ means that $l$ is matched to $r$. A configuration $\mathbf{x} \equiv \{x_{lr} : (lr) \in E\}$ is a matching iff the following constraints are satisfied: $\sum_r x_{lr} \leq 1 \ \forall l$ and $\sum_l x_{lr} \leq 1 \ \forall r$. We also introduce the stochastic parameters $\mathbf{t} = \{t_l : l \in L_2\}$ with $t_l \in \{0, 1\}$ where $t_l = 1$ means that $l$ is present in $\mathcal{L}_2$, and the associated constraints $x_{lr} \leq t_l$ for each $(lr) \in E$ with $l \in L_2$. The "energy" of a configuration $\mathbf{x}$ for some given $\mathbf{t}$ is the number of unmatched vertices, given by $\mathcal{E}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{t}) = \sum_{l\in L_1} (1 - \sum_r x_{lr}) + \sum_{l\in L_2} (t_l - \sum_r x_{lr}) + \sum_{r\in R} (1 - \sum_l x_{lr})$.

For a given partial configuration $(\mathbf{x}_1, \mathbf{t})$ the optimum value of $\mathbf{x}_2$ can be found with MS by associating to each variable $x_{lr}$ two messages $u_{lr}$ (going from left vertices to right vertices) and $h_{lr}$ (going from right to left) satisfying the equations $u_{lr} = -\max[-1, \max_{r'\neq r} h_{lr'}]$ and $h_{lr} = -\max[-1, \max_{l'\neq l} u_{l'r}]$, as discussed in [24]. These equations admit several solutions and (at least for average connectivity $c < e$) the correct one has support on $\{-1, 1\}$. For $c > e$ the simpler solution with $u_{lr}, h_{lr} \in \{-1, 1\}$ is not necesarily correct and the 0 value

should aditionally be allowed.

To proceed with the computation of the average in (1), we compute BP equations for variables $\mathbf{h}, \mathbf{u}$ and $\mathbf{t}$ and an energy function containing both the MS constrains in the previous paragraph and the prior probability for $\mathbf{t}$. Normally, we would need to consider the couple $(u_{lr}, h_{lr})$ as a single joint variable, and construct messages representing the cavity marginals for these joint variables, going from left to right as well as from right to left. However, as discussed in [20], the message passing procedure over variables that are themselves messages of an underlying message passing algorithm can be simplified: it is only needed to introduce $U_{lr} = \mathbb{P}[u_{lr} = 1]$ propagating from left to right, and $H_{lr} = \mathbb{P}[h_{lr} = 1]$ propagating from right to left. The BP equations for $U_{lr}$ and $H_{lr}$ are:

$$U_{lr} = p_l \prod_{r' \neq r} (1 - H_{lr'}) \ , \ H_{lr} = \prod_{l' \neq l} (1 - U_{l'r}) \quad (4)$$

Equations (4) can be derived by observing that $U_{lr} = \mathbb{P}[t_l = 1]\, \mathbb{P}[-\max(-1, \max_{r' \neq r} h_{lr}) = 1 | \, t_l = 1]$, and similarly for $H_{lr}$.

The minimum energy can be easily computed from MS messages $\mathbf{h}, \mathbf{u}$ as the $\beta \to \infty$ limit of (21) in [24]:

$$\begin{aligned} \mathcal{E}^*(\mathbf{x}_1, \mathbf{t}) &= \min_{\mathbf{x}_2} \mathcal{E}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{t}) \\ &= \epsilon(\mathbf{x}_1) - \sum_{l \in \mathcal{L}_2(\mathbf{t})} \max_r h_{lr} - \sum_r \max_{l \in \mathcal{L}_2(\mathbf{t})} u_{lr} + \\ &\quad + 2 \sum_{(lr): l \in \mathcal{L}_2(\mathbf{t})} \mathbb{1}[h_{lr} + u_{lr} = 2] \end{aligned} \quad (5)$$

where $\epsilon(\mathbf{x}_1)$ represents the energy contribution of the vertices in $L_1$ and is constant with respect to $\mathbf{x}_2$, and therefore to $u_{lr}$ and $h_{lr}$. This expression is a sum of local terms, and its average over $\mathbf{t}$ can be computed from a solution of the BP equations (4). Considering (5) as an observable, the average minimum energy is then computed by averaging over $u_{lr}$ and $h_{lr}$ using $U_{lr} = \mathbb{P}[u_{lr} = 1]$, and $H_{lr} = \mathbb{P}[h_{lr} = 1]$:

$$\begin{aligned} \mathcal{E}^*(\mathbf{x}_1) \equiv \mathbb{E}_{\mathbf{t}} \mathcal{E}^*(\mathbf{x}_1, \mathbf{t}) = \sum_l p_l \left[ 2 \prod_r (1 - H_{lr}) - 1 \right] + \\ + \sum_r \left[ 2 \prod_l (1 - U_{lr}) - 1 \right] + 2 \sum_{(lr)} H_{lr} U_{lr} \end{aligned} \quad (6)$$

where the term $\epsilon(\mathbf{x}_1)$ is included and represented with the convention $U_{lr} \equiv H_{lr} \equiv x_{lr} \ \forall l \in L_1$. For example, the contribution from a vertex $l \in L_2$ is $+1$ if the vertex is present and if all the incoming values of $h_{lr}$ are $-1$, which happens with probability $p_l \prod_r (1 - H_{lr})$; the same contribution will be $-1$ if the vertex is present and if there is at least one incoming value of $h_{lr}$ equal to $+1$, which happens with probability $p_l [1 - \prod_r (1 - H_{lr})]$. The average of the contribution is then $p_l [2 \prod_r (1 - H_{lr}) - 1]$. The average of all remaining terms is computed similarly.

We can then proceed to minimize this energy with MS by considering the messages $U_{lr}$ and $H_{lr}$ as variables of a new problem, introducing the cavity messages $\mathcal{U}_{lr}(U) = \log \mathbb{P}[U_{lr} = U]$ propagating from left to right and $\mathcal{H}_{lr}(H) = \log \mathbb{P}[H_{lr} = H]$ from right to left. Notice that if $l \in L_1$, we will have $U_{lr} = x_{lr} \in \{0, 1\}$ satisfying the matching constraints $\sum_r U_{lr} \leq 1$, while $U_{lr} \in [0, 1]$ satisfying the SP update equations (4) if $l \in L_2$, and similarly for $H_{lr}$. The continuous distributions over messages associated to $\mathbf{x}_2$ variables can be discretized for numerical purposes. The update equations for the messages $\mathcal{U}_{lr}$ and $\mathcal{H}_{lr}$ are obtained as usual for MS, i.e. $\mathcal{U}_{lr}(U_{lr}) = \max \left[ -E_{lr}(U_{lr}, H_{lr'}) + \sum_{r' \neq r} \mathcal{H}_{lr'}(H_{lr'}) \right]$, where $E_{lr}(U_{lr}, H_{lr'})$ is the sum of the terms in (6) containing $U_{lr}$, and where the maximisation is over the values of the incoming messages $\{H_{lr'} : r' \neq r\}$ subject to the appropriate constraints; the update of $\mathcal{H}_{lr}(H_{lr})$ is obtained similarly. We don't report these equations for brevity. All these maximizations can be performed efficiently by exploiting their associativity.

These equations can be solved by iteration starting with uniform initial conditions. It is important to note that these are the *only* message passing equations that need to be solved numerically At the fixed point, the values of $\mathcal{U}_{lr}$ and $\mathcal{H}_{lr}$ provide the optimal values of $\mathbf{x}_1$ by setting $x_{lr} = 1$ iif $[\mathcal{U}_{lr}(1) - \mathcal{U}_{lr}(0)] + [\mathcal{H}_{lr}(1) - \mathcal{H}_{lr}(0)] + 2 > 0$. Once $\mathbf{x}_1$ has been assigned and $\mathbf{t}$ has been extracted it is easy to perform the minimization over $\mathbf{x}_2$.

Figure 2 shows some results obtained with this algorithm, both for average connectivity $c < e$ and for $c > e$.

*Approximations for multi-stage problems.* The efficiency of the algorithm described above for $k = 2$ allows to use it to obtain an approximate solution to the (generally intractable) $k$-stage problem with $k > 2$. With obvious notations, the problem at stage $i$ s to find the partial configuration $\mathbf{x}_i^*$ which minimizes the expected value of the final cost function $\mathcal{E}(\mathbf{x}, \mathbf{t})$, given the previously assigned variables $\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}$ and the previously set parameters $\mathbf{t}_1, \ldots, \mathbf{t}_{i-1}$:

$$\mathbf{x}_i^* = \arg\min_{\mathbf{x}_i} \mathbb{E}_{\mathbf{t}_i} \min_{\mathbf{x}_{i+1}} \mathbb{E}_{\mathbf{t}_{i+1}} \cdots \min_{\mathbf{x}_k} \mathbb{E}_{\mathbf{t}_k} \mathcal{E}(\mathbf{x}, \mathbf{t}). \quad (7)$$

Instead, a simple heuristics consists in finding $\tilde{\mathbf{x}}_i^* = \arg\min_{\mathbf{x}_i} \mathbb{E}_{\mathbf{t}_{i+1} \cdots \mathbf{t}_k} \min_{\mathbf{x}_{i+1} \cdots \mathbf{x}_k} \mathcal{E}(\mathbf{x}, \mathbf{t})$ by repeatedly applying the algorithm for $k = 2$, producing lower bounds for the expected value of the energy at each stage.

As $k$ increases, the effect of more and more of these approximations cumulates. Yet, extensive numerical experiments for the $k$-stage stochastic matching problem with large $k$ show that the final energy can be interestingly good and the convergence time of the algorithm is extremely fast (see data of Figure 2).

*Possible extensions and future work.* Let us stress that, in principle, the general procedure we described for $k = 2$ can be extended to $k > 2$ to obtain an *exact* (within the limits of the cavity method) solution, rather

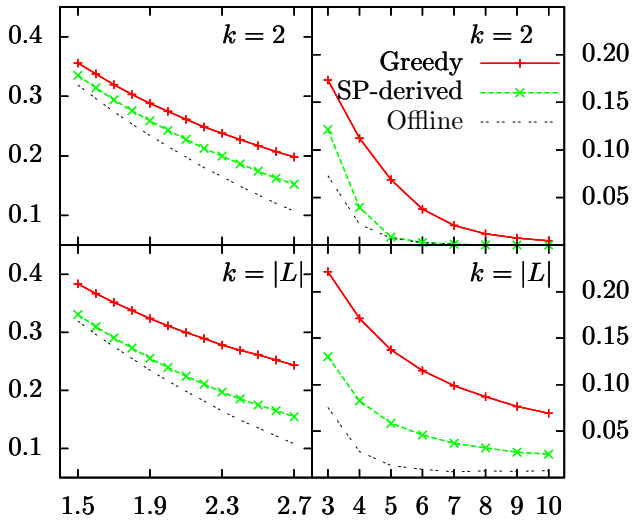


FIG. 2. Average fraction of unmatched vertices (energy) vs. average connectivity. The three lines correspond to a greedy algorithm (assign $\mathbf{x}_1$ as if $\mathbf{x}_2$ did not exist for $k = 2$, RANKING [25] for $k = |L|$), to the SP-derived algorithm, and to the offline optimum (with full knowledge of $\mathbf{t}$) lower bound. Each point is an average of at least 130 to 350 instances, with error bars smaller than the point sizes; for $k = 2$ the energy of each instance is averaged over the realizations of $\mathbf{t}$. The top panels have $|L_1| = 1\,000$ and $|L_2| = |R| = 2\,000$; the bottom ones have $|L| = 4\,000$ and $|R| = 4\,000$. In all cases $p_l$ is distributed uniformly in $]0, 1]$, and the average number of left vertices is equal to the number of right vertices. Typical instances with $k = 2$ and $c < e$ are solved in a few seconds. For $c > e$ convergence is improved by reinforcement and friction terms [26, 27].

than an heuristics like the one described above. The only practical limitation to this is that the messages become nested "distributions of distributions". However we consider it to be an interesting subject for further work.

Moreover, it is clear that the procedure described to solve (1) can be easily adapted to solve other interesting problems, such as $\mathbf{x}_1^* = \arg\min_{\mathbf{x}_1} \max_{\mathbf{t}} \min_{\mathbf{x}_2} \mathcal{E}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{t})$, corresponding to a two player game; or, replacing the $\max_{\mathbf{t}}$ with a "finite temperature" version, corresponding to robust optimization (i.e. against adverse events that can happen with some finite probability).

A more detailed discussion of the multi-stage problem, together with other applications to some relevant network optimization problems (e.g. stochastic Minimum Steiner Tree problem) for which we have promising preliminary results, will be given in a forthcoming work.


[1] C. Papadimitriou, *Computational complexity* , J. Wiley (2003)
[2] A. Prekopa, *Stochastic programming*, Springer (1995)
[3] J. Birge and F. Louveaux, *Introduction to stochastic programming*, Springer Verlag (1997)
[4] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on stochastic programming: modeling and theory*, SIAM, Philadelphia, PA (2009)
[5] R. Motwani and P. Raghavan, *Randomized algorithms*, Cambridge University Press (1995)
[6] J. Schneider and S. Kirkpatrick, *Stochastic optimization*, Springer Verlag (2006)
[7] Y. Kabashima, J. Phys. A: Math Gen. **36**, 11111 (2003)
[8] A. Braunstein, R. Mulet, A. Pagnani, M. Weigt, and R. Zecchina, Phys. Rev. E **68**, 36702 (2003)
[9] J. Neirotti and D. Saad, EPL (Europhysics Letters) **71**, 866 (2005)
[10] A. Montanari and T. Rizzo, J. Stat. Mech. (2005), P10011
[11] M. Chertkov and V. Chernyak, J. Stat. Mech. (2006), P06009
[12] L. Zdeborova and F. Krzakala Phys. Rev. E **76**, 31131 (2007)
[13] B. Frey and D. Dueck, Science **315**, 972 (2007)
[14] A. Braunstein, R. Mulet, and A. Pagnani, BMC Bioinformatics **9**, 240 (2008)
[15] M. Bayati, C. Borgs, A. Braunstein, J. Chayes, A. Ramezanpour, and R. Zecchina, Phys. Rev. Lett. **101**, 37208 (2008)
[16] F. Ricci-Tersenghi and G. Semerjian, J. Stat. Mech. (2009), P09001
[17] M. Mezard and R. Zecchina, Phys. Rev. E **66**, 56126 (2002)
[18] A. Braunstein and R. Zecchina, J. Stat. Mech. (2004), P06007
[19] A. Braunstein, M. Mezard, and R. Zecchina, Random Structures and Algorithms **27**, 201 (2005)
[20] M. Mezard and A. Montanari, *Information, physics, and computation*, Oxford University Press (2009)
[21] N. Kong and A. Schaefer, European Journal of Operational Research **172**, 740 (2006)
[22] I. Katriel, C. Kenyon-Mathieu, and E. Upfal, Theoretical Computer Science **408**, 213 (2008)
[23] B. Escoffier, L. Gourves, J. Monnot, and O. Spanjaard, European Journal of Operational Research **205**, 19 (2009)
[24] L. Zdeborova and M. Mezard, J. Stat. Mech. (2006), P05003
[25] R. Karp, U. Vazirani, and V. Vazirani, in Proceedings of the twenty-second annual ACM symposium on Theory of computing (ACM, 1990), pp. 352358
[26] A. Braunstein and R. Zecchina, Phys. Rev. Lett. **96**, 30201 (2006)
[27] F. Altarelli, A. Braunstein, J. Realpe Gomez, and R. Zecchina, J. Stat. Mech. (2009), P07002