Statistical mechanics of reputation systems in autonomous networks

Andre Manoel[1] and Renato Vicente[2]

[1] Departamento de Física Geral, Instituto de Física, Universidade de São Paulo, 05314-090, São Paulo-SP, Brazil
[2] Department of Applied Mathematics, Instituto de Matemática e Estatística, Universidade de São Paulo, 05508-090, São Paulo-SP, Brazil
E-mail: amanoel@if.usp.br and rvicente@ime.usp.br

Abstract. Reputation systems seek to infer which members of a community can be trusted based on ratings they issue about each other. We construct a Bayesian inference model and simulate approximate estimates using belief propagation (BP). The model is then mapped onto computing equilibrium properties of a spin glass in a random field and analyzed by employing the replica symmetric cavity approach. Having the fraction of positive ratings and the environment noise level as control parameters, we evaluate in different scenarios the robustness of the BP approximation and its theoretical performance in terms of estimation error. Regions of degraded performance are then explained by the convergence properties of the BP algorithm and by the emergence of a glassy phase.

J. Stat. Mech. (2013) P08002

Contents

1. Introduction

Ad hoc [1] and wireless sensor networks [2] work in the absence of a central authority and are increasingly pervasive in modern computer systems. The secure operation of these autonomous networks depends on the capability of establishing trust among network entities. Usually it is reasonable to assume that reputation and trust are positively correlated quantities and then employ a mutual scoring system as a source of data that can be used to estimate reputations [3]–[5].

Here we are concerned with the part of a reputation system [6] that identifies ill-intentioned individuals or malfunctioning devices by estimating reputations. This task would be trivial if scores were a reliable representation of an entity reputation. Instead, evaluation mistakes may happen or misleading ratings may be issued on purpose [7].

Reputation systems are particularly prone to attacks by malicious entities who can corrupt the recommendation process [8]–[10]. This happens for instance when multiple entities conspire to emit negative ratings about well-intentioned agents while emitting positive ratings about co-conspirators. In another form of attack, known as a *Sybil attack*, a single entity could impersonate others and trick the reputation mechanism.

The simplest algorithms employed by online communities use average ratings to determine reputations. Despite having the advantage of being easy to understand, these algorithms do not take into account the possibility of entities committing mistakes or acting deceitfully, which often leads to inferior results. More sophisticated algorithms

employ Bayesian inference [11] or fuzzy logic [12]. Recently, iterative formulas over looped or arbitrarily long chains—the so called *flow models*—have also been proposed (e.g. the PageRank algorithm [13]). For a more thorough exposition of the range of techniques employed we suggest recent reviews, such as [3, 10].

In this paper, we employ statistical mechanics techniques to study the performance of a belief propagation algorithm to approximately estimate reputations. This analysis provides insights into the general structure of the inference problem and suggests improvement directions.

The material is organized as follows. In section 2 we introduce the inference model, an algorithm for estimating reputations and a performance measure. In section 3 we simulate the algorithm and discuss the results. A theoretical analysis is presented in section 4 and phase diagrams are calculated. Section 5 discusses the dynamical properties of the approximate inference that impact performance. In section 6 we discuss the algorithm robustness by analyzing attacks, parameter mismatches and different topologies. Finally, conclusions are provided in section 8.

2. Estimating reputations

We model a reputation system along the lines of [14, 15]. An entity $i = 1, \ldots, n$ has a reputation r_i, and issues ratings J_{ij} about other entities $j = 1, \ldots, n$ with $j \neq i$. We define a set Ω of ordered pairs that contains (i, j) if the rating that i issues about j is present. We assume that ratings and reputations are related by a given function

$$\boldsymbol{J} = f(\boldsymbol{r}, \{\xi\}), \tag{1}$$

where $\{\xi\}$ is a set of random variables representing externalities such as, for example, uncertainties affecting opinion formation or transmission. A model is defined by specifying the domains of r_i and J_{ij}, the distribution of $\{\xi\}$ and the function f.

A good model should be able to describe realistic scenarios while remaining amenable to analytical treatment. Ratings should represent true reputations, namely, $J_{ij} \propto r_j$. The model should also take into account emitter reputations, as we have to consider that an unreliable entity may emit ratings defaming well-intentioned individuals or groups, and that a sufficient number of such ratings can misguide the reputation system—the collusion phenomena depicted in figure 1. A simple choice is

$$J_{ij} = \xi_{ij} r_i r_j, \tag{2}$$

with ξ_{ij} representing noise in the communication channel. We start by choosing $r_i, J_{ij}, \xi_{ij} \in \{-1, +1\}$, with ξ_{ij} being a random variable such that $\xi_{ij} = +1$ with probability p (*signal level*).

This choice of $f(\boldsymbol{r}, \{\xi\})$ mitigates the collusion phenomena, as $J_{ij} = -1$ represents, in the noiseless case, either an ill-intentioned positive recommendation about an unreliable entity or a negative recommendation about a reliable entity issued by an unreliable node. By introducing ξ_{ij} we also build into the model misjudgments and transmission failures. We assume a prior distribution for $\boldsymbol{r}$ supposing independence and a fraction q of reliable agents (*reputation bias*).

Alternatively, the dependence structure of the inference model may be represented by a directed graph $\mathcal{G} = (V, E)$, with each vertex $v \in V$ standing for an entity, and an arc (i, j) being connected if and only if $(i, j) \in \Omega$.

The methods of equilibrium statistical mechanics require a symmetric $\boldsymbol{J}$, which can be achieved by grouping terms as

$$\sum_{(i,j) \in \Omega} J_{ij} r_i r_j = \tfrac{1}{2} \sum_{ij} (J_{ij} + J_{ji}) r_i r_j,$$

where on the rhs we assume that $J_{ij} = 0$ if $(i, j) \notin \Omega$. Here we consider only undirected graphs, to say, both arcs (i, j) and (j, i) are in $\mathcal{G}$ if $(i, j) \in \Omega$. By replacing $\tfrac{1}{2}(J_{ij} + J_{ji})$ for J_{ij}, we work with $J_{ij} \in \{-1, 0, +1\}$, discarding dissonant opinions that, fortunately, are few in many cases of interest [10]. The case where a single arc may be present can be modeled by a simple extension that considers $J_{ij} \in \{-1, -\tfrac{1}{2}, 0, +\tfrac{1}{2}, +1\}$.

Given a sample of ratings $\{J_{ij}\}$, our goal is to find an estimate $\hat{\boldsymbol{r}}$ for the reputations $\boldsymbol{r}$ while keeping the error

$$\varepsilon(\hat{\boldsymbol{r}}, \boldsymbol{r}) = \frac{1}{2} \left(1 - \frac{\hat{\boldsymbol{r}} \cdot \boldsymbol{r}}{n} \right), \tag{7}$$

as small as possible.

A naive solution calculates reputations according to the majority of recommendations (a majority rule)

$$\hat{r}_i = \mathrm{sgn} \left(\sum_{k \in \partial i} J_{ki} \right), \tag{8}$$

where ∂i stands for the neighborhood of the vertex i in $\mathcal{G}$. Assuming that ratings are produced according to equation (2) we can write

$$\hat{r}_i = \mathrm{sgn} \left[\left(\sum_{k \in \partial i} \xi_{ki} r_k \right) r_i \right]. \tag{9}$$

Thus we can calculate the probability of agreement between $\hat{r}_i$ and the real reputation r_i by considering that r_k and ξ_{ki} are random variables sampled from $P(r)$ and $P(\xi)$ (while keeping $\xi_{ik} = \xi_{ki}$), and by introducing a random variable λ sampled from the degree distribution of graph $\mathcal{G}$:

$$P(\hat{r}_i = r_i) = \left\langle\!\!\left\langle P \left(\sum_{k=1}^{\lambda} \xi_{ik} r_k > 0 \right) \right\rangle\!\!\right\rangle_\lambda$$

$$= \left\langle\!\!\left\langle \sum_{n=0}^{\lfloor (\lambda-1)/2 \rfloor} \binom{\lambda}{n} \varpi^{(\lambda-n)} (1 - \varpi)^n \right\rangle\!\!\right\rangle_\lambda, \tag{10}$$

with $\varpi = pq + (1 - p)(1 - q)$, and $\langle\!\langle \cdot \rangle\!\rangle_\lambda$ denoting the average with respect to λ. Here, $\xi_{ik} r_k$ is a ± 1 random variable with parameter ϖ (i.e., it is $+1$ with probability ϖ). The sum of a set of these variables is binomially distributed, and the expression on the rhs represents the cumulative distribution.
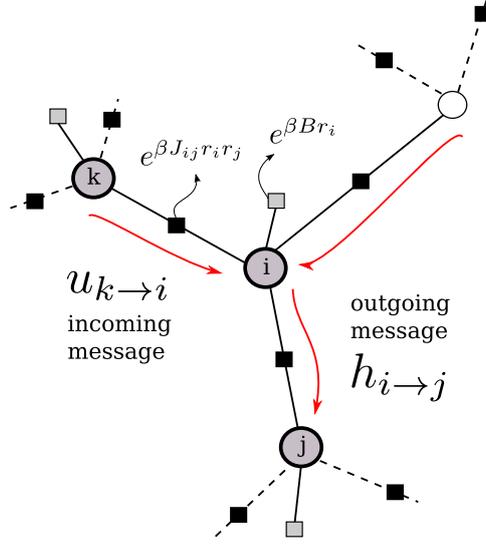
Figure 2. Factor graph and BP message-passing. Snapshot of a factor graph representing the posterior (11). Variable nodes are represented by circles, factor nodes are represented by small squares. At an iteration outgoing messages $h_{i \to j}^{(t+1)}$ are computed after combining incoming messages $u_{k \to i}^{(t)}$. When convergence is attained we compute effective fields $\hat{h}_i$ and the approximation for marginal posteriors $P(r_i)$.

Until section 6 we assume a random regular graph $\mathcal{G}$ with degree c. Also, we deal with a scenario such that the signal prevails, $p > 0.5$, and most of the nodes are reliable, $q > 0.5$.

For an entity i the inference task consists of maximizing the posterior for $\boldsymbol{r}$ marginalized over every component except i. The posterior is factorizable and can be put in the following form

$$P(\boldsymbol{r}|\boldsymbol{J}) = \frac{1}{\mathcal{Z}} \prod_{(i,j) \in \mathcal{G}} \exp(\beta J_{ij} r_i r_j) \prod_i \exp(\beta B r_i), \tag{11}$$

where β and B are parameters which are optimally set at Nishimori's condition [16] $\beta = \alpha_q$ and $B = \alpha_q / \alpha_p$. Factorizable distributions such as this are well represented by factor graphs [17]–[19], with variable nodes associated to each r_i and function (or factor) nodes representing the functions linking them, in this case the exponentials. Figure 2 zooms in a factor graph representation of the posterior (11).

Given this posterior distribution we calculate marginal distributions $P(r_i) = \sum_{r_j \neq r_i} P(\boldsymbol{r}|\boldsymbol{J})$ efficiently by employing the message-passing scheme of belief propagation (BP) on the factor graph associated with the posterior equation (11) [19]. In our case the outgoing BP messages (see figure 2 for illustration) are

$$h_{i \to j}^{(t+1)} = B + \sum_{k \in \partial i / j} u_{k \to i}^{(t)}(J_{ki}, h_{k \to i}^{(t)}), \tag{12}$$

while $u_{k \to i}^{(t)} = (1/\beta)\tanh^{-1}\left[\tanh(\beta J_{ki}) \tanh(\beta h_{k \to i}^{(t)})\right]$ are incoming messages.
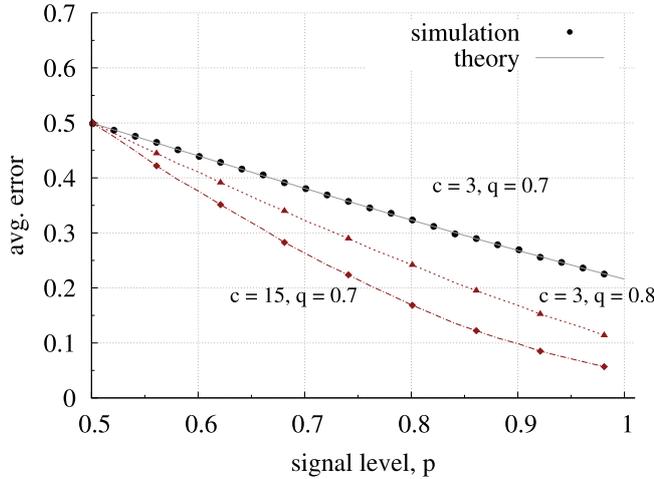
Figure 3. Estimation error for the majority rule. Average errors in simulations (circles) are compared with theoretical errors calculated by averaging equation (7) over (10) (lines). Averages are over 3000 runs, $q = P(r_i = 1) = 0.7$ or 0.8. Graphs $\mathcal{G}$ are random and regular with degrees $c = 3$ or 15, and $n = 100$. Error bars are smaller than the symbols in all three curves.

Iterating this set of equations until convergence we obtain $\{h^*_{i \to j}, u^*_{i \to j}\}$, which yields an approximation for marginals $P(r_i) \propto \exp(\beta \hat{h}_i r_i)$, with effective fields given by

$$\hat{h}_i = B + \sum_{k \in \partial i} u^*_{k \to i}. \tag{13}$$

This algorithm is exact on trees, but can be used in graphs of any topology, leading to good approximations provided that the average cycle length is large [19].

After convergence of the BP scheme, reputations r_i are estimated by marginal posterior maximization (MPM)

$$\hat{r}_i = \arg\max_{r_i} P(r_i) = \mathrm{sgn}(\hat{h}_i). \tag{14}$$

3. Simulations

As a basis of comparison we run the majority rule algorithm defined by equation (8) for 3000 scenarios with reputations $\boldsymbol{r}$ chosen randomly with reputation bias q and symmetric ξ_{ik} with signal level p. In figure 3 we compare the average error in simulations of the majority rule with the theoretical error calculated by averaging equation (7) over the distribution (10). The majority rule is not very far from what is used in common reputation systems on e-commerce websites. Note, however, that the error of this very simple scheme can be larger than $1 - q$ if the signal level p is low enough. The message is therefore clear: in noisy environments, assigning good reputations by default may actually be more effective than using the majority recommendation.

In figure 4 we compare the majority rule with the average over 3000 runs of the MPM estimate computed with the BP algorithm. For convenience, the algorithm is presented as
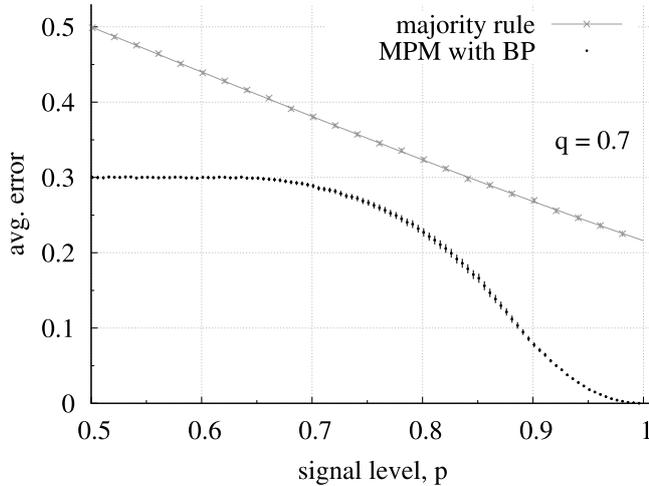
Figure 4. MPM estimate versus the majority rule. Comparison between the average error of the majority rule (equation (8)) and of the MPM estimate (equation (14)), for $q \equiv P(r_i = 1) = 0.7$, a regular random graph with degree $c = 3$ and $n = 100$. For the majority rule we show theory and simulations, with symbols larger than the error bars. For the MPM estimate we show the average over 3000 simulation runs, with error bars representing one standard deviation.

a pseudocode in appendix A.[3] The gains in performance when the collusion phenomenon is built into the inference model are considerable even in very noisy environments.

A detailed view of the error surface for the approximate MPM estimates in terms of reputation bias q and signal level p is depicted in figure 5(a). Two regions can be discerned, with large error for low signal level (high noise) and low reputation bias. Other average quantities can also be evaluated in the simulations in order to access the algorithm's performance. Figure 5(b), for instance, shows the average number of iterations to convergence. A distinctive region is observed with degraded convergence time.

The true values p and q of the signal level and reputation bias are usually unknown, and the inference algorithm has to assume (or estimate) values $\hat{p}$ and $\hat{q}$. However, both p and q are related to the observable fraction of positive ratings, $w = P(J_{ij} = +1)$, in the following way

$$(2w - 1) = (2p - 1)(2q - 1)^2 \tag{15}$$

and actually only one of the control parameters has to be estimated.

The dashed white lines in figures 5(a) and (b) represent fixed values of w in the p–q plane; due to (15), for a given w we have $w \leq p \leq 1$ and $\frac{1}{2}(1 + \sqrt{2w - 1}) \leq q \leq 1$. As can be seen following the white lines, errors are small for p or q large, when ratings are more informative, while increasing for intermediate values.

Ideally, the parameters have to be set to the same values used to generate the data. However, the environment can change without warning and we would also like to know how the inference scheme would perform in such circumstances. By simulation we can generate the vector $\boldsymbol{r}$ and the symmetric matrix ξ as random variables with probability

---

[3] Source code is also available at https://github.com/amanoel/repsys.
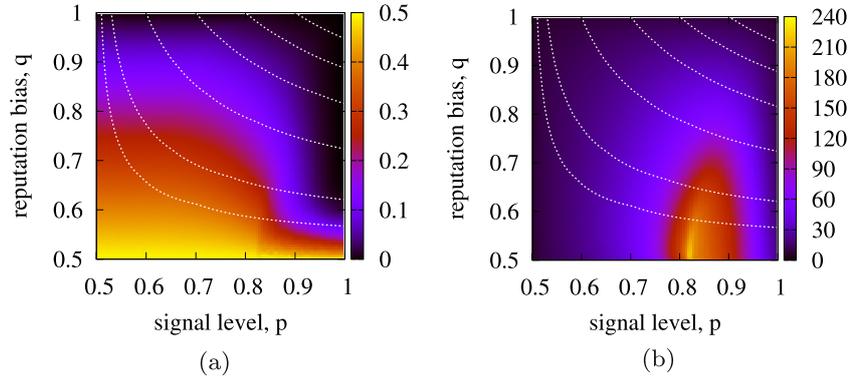
Figure 5. Average error and number of iterations to convergence: 3000 runs of the BP algorithm in a grid of values for signal level p and reputation bias q, with $n = 100$ and $c = 3$. In each run $\mathcal{G}$, $\{r_i\}$ and $\{\xi_{ij}\}$ are sampled and ratings $\{J_{ij} = \xi_{ij} r_i r_j\}$ are calculated. Panel (a) depicts the average error as in equation (7); panel (b) represents the average number of iterations needed for convergence of the BP algorithm. The dashed lines follow fixed values of w, from bottom to top: $w = 0.51, 0.53, 0.6, 0.7, 0.8, 0.9$.
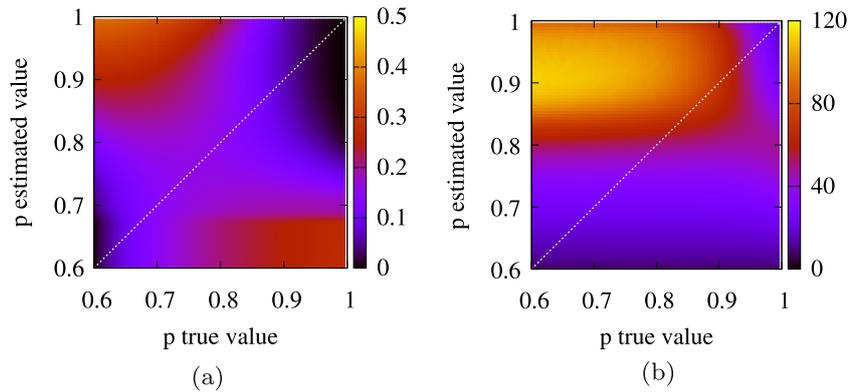
Figure 6. Empirical analysis for mismatched signal level p with the fraction of positive ratings fixed to $w = 0.6$. Panel (a) empirical error and panel (b) average number of iterations to convergence.

of being $+1$ set to q and p, respectively, and then run the algorithm assuming $\alpha_{\hat{p}}$ and $\alpha_{\hat{q}}$. As w can be determined from the data, we assume it to be known, so once p and $\hat{p}$ are chosen, q and $\hat{q}$ may be calculated using equation (15).

Figure 6 depicts results of this simulations in the plane p–$\hat{p}$ for $w = 0.6$. Small mismatches between $\hat{p}$ and p are in general well absorbed by the inference scheme. Performance, however, deteriorates when estimates, in this case for the signal level, are too optimistic or too pessimistic. Figure 8(b) also reveals a region with degraded convergence times for optimistic estimates.

In section 4 we use equilibrium statistical mechanics to calculate the phase diagram as a function of the control parameters p–q and p–$\hat{p}$, and to explain regions of degraded performance.

4. Theoretical analysis

The posterior distribution in equation (6) suggests the description of the problem of inferring reputations in terms of the equilibrium properties of a spin glass in an external field

$$
\mathcal{H}(\boldsymbol{s}|\boldsymbol{J}, B) = - \sum_{(i,j) \in \Omega} J_{ij} s_i s_j - B \sum_i s_i,
\tag{16}
$$

where we designate the dynamic variable as s_i, and the *target* variables r_i are fixed (*quenched*). The MPM estimates described in section 3 correspond to $\hat{r}_i = \mathrm{sgn}(m_i)$, with $m_i = \tanh(\beta \hat{h}_i)$ representing local equilibrium magnetizations.

At Nishimori's condition [16], the temperature and field are chosen as $\beta_N = \alpha_p$ and $B_N = \alpha_q/\alpha_p$, and the microstates of this physical system are distributed according to a Gibbs measure

$$
P(\boldsymbol{s}|\boldsymbol{J}) = \frac{1}{\mathcal{Z}} \mathrm{e}^{-\beta_N \mathcal{H}(\boldsymbol{s}|\boldsymbol{J}, B_N)}.
\tag{17}
$$

Other values of β and B, corresponding to misspecified p and q, can also be studied along the same lines.

We calculate the equilibrium average error in equation (7), which corresponds to the magnetization of the Hamiltonian in equation (16) (gauge) transformed with $r_i s_i \to s_i$. As this Hamiltonian is not gauge invariant we now have to deal with a spin glass in a random field

$$
\mathcal{H}(\boldsymbol{s}|\boldsymbol{\xi}, \boldsymbol{r}, B) = - \sum_{(i,j) \in \Omega} \xi_{ij} s_i s_j - B \sum_i r_i s_i.
\tag{18}
$$

Here $\{\xi_{ij}\}$ and $\{r_i\}$ are quenched variables with $P(\xi_{ij} = \xi_{ji} = +1) = p$ and $P(r_i = +1) = q$. The error $\varepsilon(\hat{\boldsymbol{r}}, \boldsymbol{r})$ in the gauge transformed variables can be written as $\bar{\varepsilon} = \frac{1}{2}(1 - \langle \mathrm{sgn}(m) \rangle)$, where $m = (1/n)\sum_i s_i$ is the gauge transformed equilibrium magnetization.

In our analysis we employ the replica symmetric cavity method along the lines of [19, 20]. We start this section by calculating the thermodynamic properties at Nishimori's condition. We first write an equation for the distribution of cavity fields for the gauge transformed variables as calculated by the BP procedure in equation (12)

$$
P(h) = \int \prod_{i=1}^{c-1} \mathrm{d}h_i \, P(h_i) \left\langle\!\!\!\left\langle \delta\left(h - Br - \sum_{i=1}^{c-1} u_i(\xi_i, h_i)\right) \right\rangle\!\!\!\right\rangle_{r, \{\xi_i\}},
\tag{19}
$$

or, more concisely

$$
h \overset{d}{=} \left\langle\!\!\!\left\langle Br + \sum_{i=1}^{c-1} u_i(\xi_i, h_i) \right\rangle\!\!\!\right\rangle_{r, \{\xi_i\}},
\tag{20}
$$

with $\overset{d}{=}$ indicating equality in distribution. In this context, h is the cavity field, and

$$
u_i(\xi_i, h_i) = \frac{1}{\beta} \tanh^{-1}\left[\tanh(\beta \xi_i) \tanh(\beta h_i)\right]
$$

are cavity biases [19, 21]. Note that we work here under the assumption that it makes sense to describe fixed points of the BP equations (12) in terms of a unique density $P(h)$. That is the replica symmetry (RS) assumption.

We calculate numerical solutions to equation (19) by a population dynamics algorithm (see appendix B for details) and then calculate thermodynamic quantities. The magnetization is given by $m = \langle\tanh(\beta\hat{h})\rangle_{\hat{h}}$, where

$$\hat{h} \stackrel{d}{=} \left\langle\!\!\!\left\langle Br + \sum_{i=1}^{c} u_i(\xi_i, h_i) \right\rangle\!\!\!\right\rangle_{r,\{\xi_i\}} \tag{21}$$

is the effective field. Note that the sum here ranges from 1 to c.

The onset of a spin glass phase can be detected by finding divergences in the spin glass susceptibility. Provided that the disorder is spatially homogeneous, the spin glass susceptibility averaged over this disorder can be written as:

$$\chi_{\text{sg}} = \sum_{\ell=0}^{\infty} \mathcal{N}(\ell) \left\langle\!\!\!\left\langle [\langle s_0 s_\ell\rangle - \langle s_0\rangle\langle s_\ell\rangle]^2 \right\rangle\!\!\!\right\rangle_{r,\{\xi_i\},\mathcal{G}}, \tag{22}$$

where s_0 is a variable at an arbitrary central site, s_ℓ is an arbitrary variable at a site separated from 0 by a chemical distance ℓ and $\mathcal{N}(\ell)$ is the number of sites at a distance ℓ from 0.

The fluctuation-dissipation theorem, the symmetry introduced by averaging and the BP equations (12) yield

$$\chi_{\text{sg}} = \sum_{\ell=0}^{\infty} \mathcal{N}(\ell) \left\langle\!\!\!\left\langle \left[\frac{\partial m_0}{\partial\hat{h}_0}\frac{\partial\hat{h}_0}{\partial u_{\cdot\rightarrow 0}}\frac{\partial u_{\cdot\rightarrow 0}}{\partial u_{\cdot\rightarrow\ell}}\frac{\partial u_{\cdot\rightarrow\ell}}{\partial\hat{h}_\ell}\right]^2 \right\rangle\!\!\!\right\rangle_{r,\{\xi_i\},\mathcal{G}}, \tag{23}$$

where $u_{\cdot\rightarrow 0}$ and $u_{\cdot\rightarrow\ell}$ represent incoming messages in a path connecting 0 to ℓ. A sufficient condition for χ_{sg} to diverge is, therefore, that

$$\lim_{\ell\rightarrow\infty} \left[\frac{\partial u_{\cdot\rightarrow 0}}{\partial u_{\cdot\rightarrow\ell}}\right]^2 > 0. \tag{24}$$

This quantity measures the sensibility of the incoming message at a central site 0 to a perturbation in a message forming at a far outside distance ℓ. In terms of cavity field distributions we can write

$$\rho = \left\langle\!\!\!\left\langle \left\langle \lim_{\ell\rightarrow\infty} \left[\frac{\partial u_0[u_\ell(h)]}{\partial u_\ell}\right]^2 \right\rangle_h \right\rangle\!\!\!\right\rangle_{r,\{\xi_i\}}. \tag{25}$$

A number of numerical methods can be used to evaluate ρ [21, 20]. Using population dynamics (described in appendix B), we introduce two slightly different initial states such that $\boldsymbol{u}_0'[i] - \boldsymbol{u}_0[i] = \delta$ with $i = 1, \ldots, N$ and $\delta = 10^{-4}$. After a large number τ of iterations of the population dynamics algorithm we calculate

$$\rho \approx \frac{1}{N}\sum_i (\boldsymbol{u}_\tau'[i] - \boldsymbol{u}_\tau[i])^2. \tag{26}$$
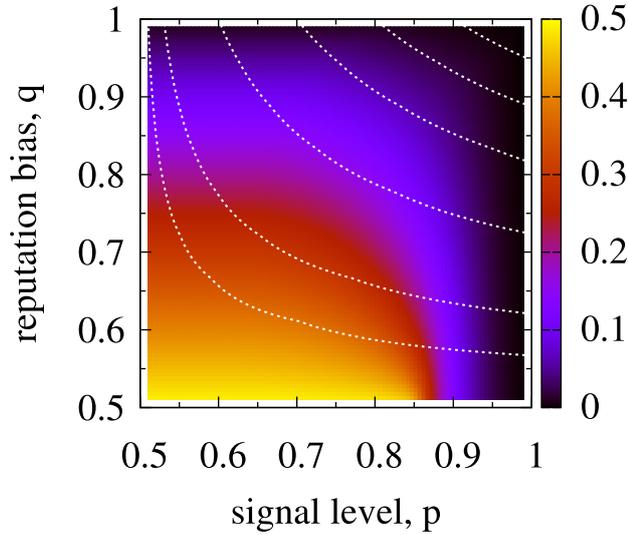
Figure 7. Theoretical error at Nishimori's condition: theoretical error $\bar{\varepsilon} = \frac{1}{2}(1 - \langle \text{sgn}(\hat{h}) \rangle_{\hat{h}})$ for different values of p and q. For $q = 0.5$, the model is gauge invariant and a transition from a paramagnetic phase (to the left) to a ferromagnetic phase is observed at $p_c \approx 0.853$.

The order parameters m and ρ allow the identification of four different thermodynamic phases: paramagnetic if $\rho = 0$, $m = 0$, ferromagnetic if $\rho = 0$, $m > 0$, glassy if $\rho \neq 0$, $m = 0$, and mixed or ferromagnetic ordered spin glass for $\rho \neq 0$, $m > 0$.

To measure the theoretical error, we also calculate $\langle \text{sgn}(\hat{h}) \rangle_{\hat{h}}$. The theoretical results depicted in figure 7 are corroborated by simulations depicted in figure 5(a). Under Nishimori's condition, $\rho = 0$ for all values of p and q, and thus no glassy or mixed phase is present. Likewise, for all values of $p, q > 0.5$, $m > 0$, and the ferromagnetic phase covers the whole diagram.

The same theoretical analysis can be repeated for the mismatched parameters case introduced in the previous section, by setting $\beta = \alpha_{\hat{p}}$ and $B = \alpha_{\hat{q}}/\alpha_{\hat{p}}$ while considering r_i and ξ_i as quenched ± 1 random variables with parameters q and p. Here again, the theoretical error depicted in figure 8(a) reproduces the empirical error in figure 6(a). In the p–$\hat{p}$ plane, however, a mixed phase also emerges as $\rho > 0$ (see figure 8(c)) for a region of parameters depicted in figure 8(b). In this region, the free energy landscape becomes rugged, and BP or any inference procedure will not find a globally optimal solution in practical time. Also the RS cavity analysis does not necessarily provide asymptotically correct results, so the thermodynamic quantities computed in this region may not reflect the actual behavior of the system at equilibrium.

It is worth mentioning that the numerical evaluation of ρ yields slightly positive values still inside the ferromagnetic phase in the p–$\hat{p}$ plane (see the two top panels of figure 8(d)). We have verified, however, that these values can be arbitrarily reduced as the precision of the floating-point representation employed in the calculation is increased, thus suggesting the result follows from an underflow and that ρ is actually equal to zero in this region. Interestingly, the region where this numerical phenomena happens is correlated to the region of degraded performance in figure 6.
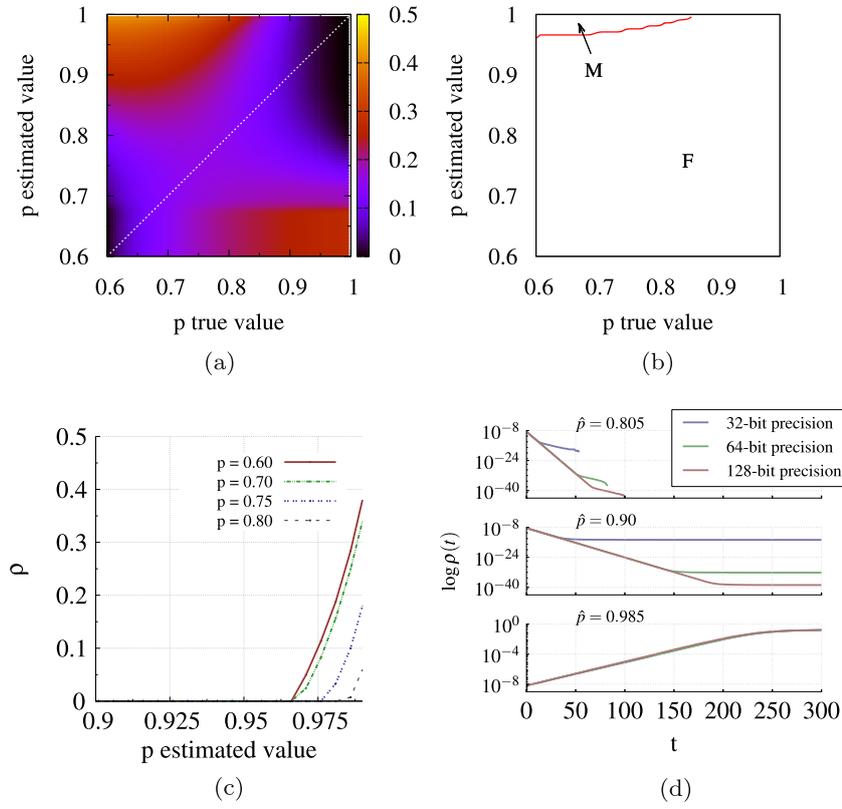
Figure 8. Theoretical analysis for mismatched signal level p with the fraction of positive ratings fixed to $w = 0.6$. Panel (a) shows the theoretical error, in good agreement with simulations. Panel (b) gives the phase diagram, in which a mixed phase emerges, partially explaining the degraded performance region. Panel (c) $\rho > 0$ is the signature for the onset of the mixed phase as $\hat{p}$ increases. Panel (d) shows the evolution of ρ throughout the population dynamics for $p = 0.7$ and different values of $\hat{p}$; as the mixed phase approaches, ρ converges to values which can be made arbitrarily small by increasing the floating-point precision employed in the calculations.

We observe that at least three rigorous results on phase diagrams of similar models are available in the literature: (A) if the Hamiltonian is gauge invariant the Nishimori line does not cross a spin glass phase [22]; (B) there is no spin glass phase in a random field Ising model [23]; and (C) provided that the parameters employed in the inference task are identical to those used to generate data (namely, we are at Nishimori's condition), in a random graph with bounded maximum degree the BP scheme converges to the correct marginals in the thermodynamic limit [24].

In order to check our results we observe that choosing $q = 1/2$ ($\alpha_q = 0$) yields a gauge invariant model. Consistently with result (A), $\rho = 0$ over the line $q = 0.5$ and only paramagnetic and ferromagnetic phases are observed, with a transition around $p_c \approx 0.853$ for our example (see figure 7). Result (B) is only relevant if we can choose parameters such that $\xi_{ij} \geq 0$ (or $p = 1$) for any $(i, j) \in \Omega$, any random field r_{ij} and any B in the Hamiltonian of equation (18). At Nishimori's condition, however, $p = 1$ implies that $\beta_N = \infty$ and $B_N = 0$. Thus the model is a trivial ferromagnet and result (B) is irrelevant.
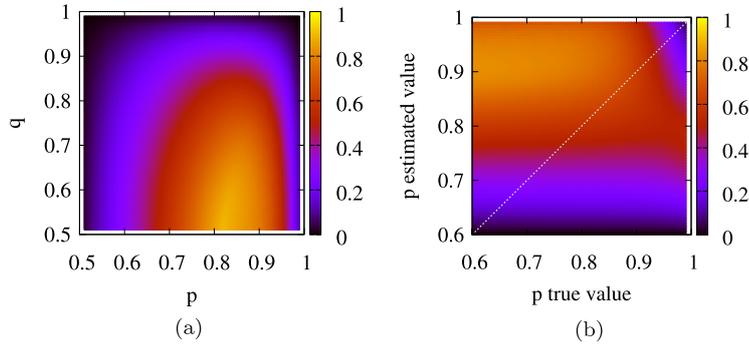
Figure 9. Average spectral radius of the Jacobian of the BP dynamical system. For each value of p and q (p and $\hat{p}$), the Jacobian's spectral radius was computed for 2500 runs of the algorithm and then averaged. Larger values of R correlate with slower convergence times. Panel (a), at Nishimori's condition, is to be compared with figure 5(b). Panel (b), with a mismatch between p and $\hat{p}$, is to be compared with figure 6(b).

Yet if the inference model assumes $\hat{p} < 1$ while ξ_{ij} are generated with $p = 1$, rigorous result (B) forbids either a spin glass or a mixed phase to show up. Accordingly, figure 8(b) exemplifies a phase diagram for $q > 0.5$ with no mixed phase for $p = 1$ (true value) and $\hat{p} < 1$ (estimated value). Finally, result (C) implies the absence of a mixed phase anywhere at Nishimori's condition, which is indeed the case, as $\rho = 0$.

The onset of a mixed phase explains in part the degradation in algorithm's performance for the upper left corner of the p–$\hat{p}$ plane. However, empirical observation shows that convergence rates also worsen outside this region of parameters, as well as on the p–q plane, even at Nishimori's condition, where no glassy phase is to be found. In the next section, we investigate this issue further.

5. Dynamical properties

In order to understand further the deterioration in the algorithm's performance, we now study the BP dynamical system. From equation (12), we obtain

$$\frac{\partial h_{i \to j}^{(t+1)}}{\partial h_{k \to i}^{(t)}} = \frac{\tanh(\beta) J_{ki}}{\cosh^2(\beta h_{k \to i}^{(t)}) - \tanh^2(\beta)\sinh^2(\beta h_{k \to i}^{(t)})} \mathbb{I}(k \in \partial i/j), \qquad (27)$$

where $\mathbb{I}(A) = 1$ if A is true and $\mathbb{I}(A) = 0$ otherwise.

The dynamical system in question has nc equations, one for each direction of each edge on the graph. In order to study the linear stability of the BP dynamical system, we have calculated the spectral radius of the Jacobian matrix evaluated at a fixed point, namely, $R = \max |\lambda(\mathbb{J})|$, where $\mathbb{J}$ is a $nc \times nc$ matrix with entries $\mathbb{J}_{ij,ki} = \partial h_{i \to j}^* / \partial h_{k \to i}^*$.

The spectral radius gives us a measure of the *convergence rate* of the algorithm. In fact, as figure 9 shows, regions where the value of R is larger coincide with those where the algorithm on average converges more slowly.

Interestingly, this quantity may be also studied within the RS cavity scheme. The population dynamics algorithm, which we have used to obtain samples of $P(h)$, may be
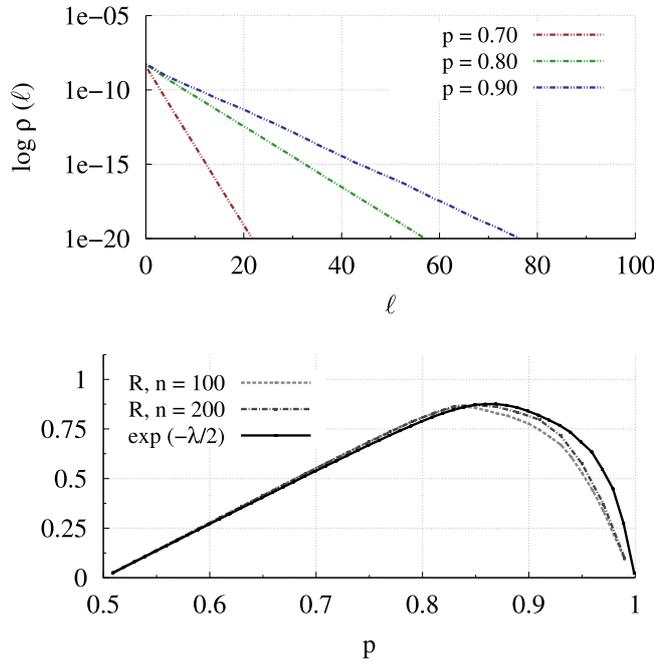
Figure 10. Decay of $\rho(\ell)$. Top panel: using population dynamics we find the rate λ of decay as a function of p for $q = 0.6$. Bottom panel: average spectral radius R for finite size ($n = 100, 200$). For comparison we also plot $e^{-\lambda/2}$. Finite size effects are larger in the large p region.

also treated as a dynamical system (the so called, *density evolution* equations):

$$
\begin{aligned}
u_i^{(\ell)} &= \frac{1}{\beta}\tanh^{-1}\left[\tanh(\beta\xi)\tanh(\beta h_i^{(\ell)})\right], \\
h_i^{(\ell+1)} &= Br + \sum_{j=1}^{c-1} u_{\gamma(j)}^{(\ell)},
\end{aligned}
\tag{28}
$$

with r and ξ independently sampled for each i, and $\gamma(\cdot)$ representing random indices, that is $\gamma(\cdot) \sim \text{Uniform}([N])$.

In this context, the evolution of the order parameter ρ, $\rho(\ell) = (1/N)\sum_{i=1}^{N}(u_i'^{(\ell)} - u_i^{(\ell)})^2$, can be studied. As can be seen in figure 10, $\rho(\ell)$ decays exponentially at a constant rate λ, that is, $\rho \propto e^{-\lambda\ell}$. In the neighborhood of a fixed point the decay rate of $\sqrt{\rho}$ is given by the dominant eigenvalue of the Jacobian matrix. We thus expect the relation $R \propto e^{-\lambda/2}$ to hold, where R is the spectral radius of the Jacobian matrix. This relationship is clearly depicted in figure 10—thus by computing the decay rate λ for ρ, we also learn about the algorithm convergence rate.

We have observed two mechanisms leading to performance degradation: the onset of a glassy phase and the decreased stability of the BP fixed point. The former is a limitation intrinsic to the inference problem, the latter an issue that probably could be addressed by modifying the approximate inference algorithm. We, however, observe that the stability of the fixed point seems to decrease as the parameters approach a mixed phase, as $p = 0.8536, q = 0.5$ defines a multicritical point in a model with $c = 3$ [20], thus suggesting this can also be an intrinsic limitation of the problem.
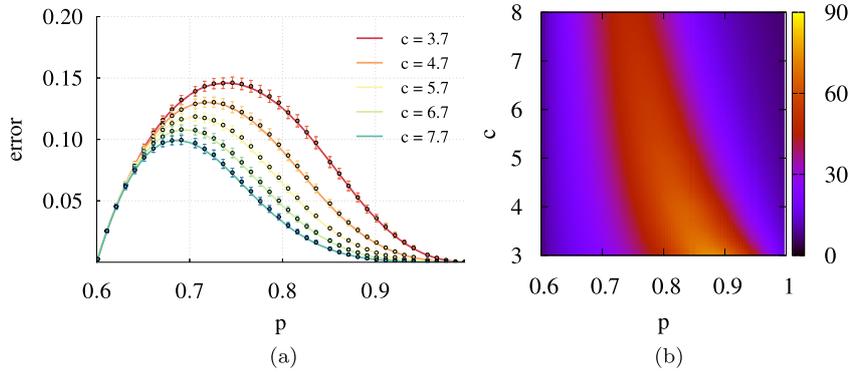
Figure 11. Empirical and theoretical analysis at Nishimori's condition for a random graph with degree distribution given by equation (29) and $w = 0.6$. The parameter c is the average node degree while p is the signal level. Panel (a) compares empirical (bars) and theoretical (lines) errors as a function of p for (from top to bottom) $c = 3.7$ to $c = 7.7$ in steps of size 1 . Panel (b) depicts the average number of iterations for the BP algorithm to converge.

6. Robustness

Up to this point our analysis has only considered ratings $\{J_{ij}\}$ distributed over a regular random graph of fixed degree $c = 3$ and issued exactly as assumed by the inference model. In this section, we relax these assumptions to access both the performance of the algorithm and the validity of the theoretical analysis under more general conditions.

6.1. Graph topologies

The set Ω of ratings issued by network entities define a graph $\mathcal{G}$ with each vertex representing an entity and an edge (i, j) being connected if and only if $(i, j) \in \Omega$. The theoretical analysis based on the replica symmetric cavity equation (19) relies on specifying an ensemble of graphs represented by a particular degree distribution (or profile). In the previous discussion we have used an ensemble of regular random graphs with a degree profile given by $\Lambda_c(\gamma) = \delta(\gamma - c)$. A natural extension to that is allowing non-integer values of c, for which we introduce:

$$\Lambda_c(\gamma) = \begin{cases} 1 - (c - \lfloor c \rfloor) & \text{for } \gamma = \lfloor c \rfloor, \\ c - \lfloor c \rfloor & \text{for } \gamma = \lfloor c \rfloor + 1, \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

In this way for $c = 2.3$ we would have 30% of the nodes with degree 3 and the remaining with degree 2.

The empirical analysis is done by simulating instances sampled from this ensemble of graphs. Figure 11(a) compares empirical and theoretical errors at Nishimori's condition as a function of p for several c values, $w = 0.6$. Figure 11(b) depicts the average number of iterations to convergence as a function of p and c, which is again explained by the stability of the unique BP fixed point.

16
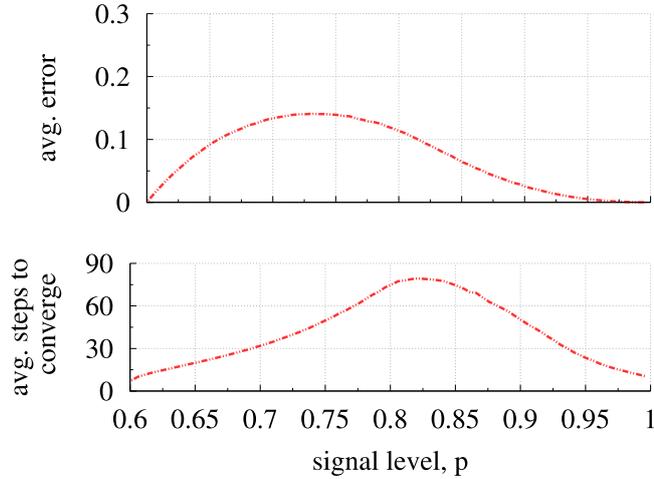
Figure 12. Simulations in a square lattice, for $w = 0.6$ and different values of p. Despite the short cycle lengths, the algorithm exhibits a good performance.

The BP algorithm calculates exact marginals and allows for optimal Bayesian inference when the subjacent graph is a tree. The performance of the BP algorithm, however, can be studied by simulation on any topology. Figure 12 shows the resulting performance measures as a function of p for $w = 0.6$ and for $\mathcal{G}$ chosen to be a square lattice in two dimensions. The average error is always smaller than $1 - q$ and vanishes as $p \to 1$, showing that even in this case the BP algorithm may yield good results.

7. Attacks

Malicious entities may issue ill-intentioned ratings to trick the reputation system and malfunctioning devices may issue erroneous ratings. These scenarios of targeted attacks can also be considered by the inference model and studied within the same theoretical framework and simulation techniques.

We suppose that a fraction f of the ratings are issued by noisy entities while the inference process remains unchanged. Let us call the subset of noisy ratings $\Omega_f \subset \Omega$. To simulate this scenario, we uniformly sample and fix a fraction $f/2$ of the entities to be noisy, issuing ratings as $J_{ij} = \eta_{ij}$, where $\eta_{ij} = \eta_{ji}$ is a ± 1 random variable with parameter $z = \frac{1}{2}$ and $(i, j) \in \Omega_f$—since we require $J_{ji} = J_{ij}$, the total fractions of noisy ratings will be f. We then run the BP algorithm and study how the performance changes with f.

For the theoretical analysis we calculate performance measures averaged over every disorder component: a regular random graph (with $c = 3$ in our example), and symmetric communication noise $\boldsymbol{\xi}$ and random ratings $\boldsymbol{\eta}$. As we are interested in checking the algorithm robustness, we also assume that the inference scheme has no knowledge that the reputation system is under attack.

We first rewrite the posterior by taking into account noisy ratings:

$$P(\boldsymbol{r}|\boldsymbol{\xi}, \boldsymbol{\eta}) \propto \prod_{(i,j) \in \Omega \backslash \Omega_f} P(\xi_{ij}) \prod_{(i,j) \in \Omega_f} P(\eta_{ij}) \prod_i P(r_i). \tag{30}$$

17
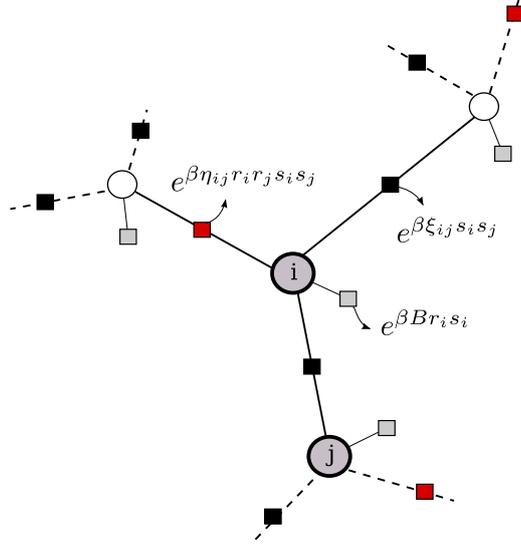
Figure 13. Factor graph for the posterior (31). A fraction f of the function nodes represent noisy ratings (red squares).

Following the previous steps yields:

$$P(\boldsymbol{r}|J) \propto \exp\left[\alpha_p \sum_{(i,j)\in\Omega\backslash\Omega_f} J_{ij}r_ir_j + \alpha_z \sum_{(i,j)\in\Omega_f} J_{ij} + \alpha_q \sum_i r_i\right]. \tag{31}$$

Since the algorithm considers the ratings Ω_f as subject to the same communication noise as regular ratings, we have $\alpha_z = \alpha_p$. The gauge transformed Hamiltonian for an equilibrium statistical mechanics description is

$$H(\boldsymbol{s}) = -\sum_{(i,j)\in\Omega\backslash\Omega_f} \xi_{ij}s_is_j - \sum_{(i,j)\in\Omega_f} \eta_{ij}r_ir_js_is_j - B\sum_i r_is_i. \tag{32}$$

There is a new term $-\sum_{(i,j)\in\Omega_f}\eta_{ij}r_ir_js_is_j$ in the Hamiltonian, which can be treated by the inclusion of a new type of function node to the factor graph representation for the posterior of equation (31). Figure 13 provides a snapshot of this factor graph. The replica symmetric cavity description can then be written as

$$h \stackrel{d}{=} \left\langle\!\!\left\langle Br + \sum_{i=1}^{c-1} u_i(h)\right\rangle\!\!\right\rangle_{r,\xi,\eta}, \tag{33}$$

where $u_i(h) = (1/\beta)\tanh^{-1}\left[\tanh(\beta\xi_i)\tanh(\beta h_i)\right]$ with probability $1 - f$, and with probability f

$$u_i(h) = \frac{1}{\beta}\tanh^{-1}\left[\tanh(\beta\eta_i rr_i)\tanh(\beta h_i)\right],$$

where η_i is a ± 1 random variable with $P(\eta_i = 1) = \frac{1}{2}$, r is the same used in equation (33) and the $\{r_i\}$ are independently sampled from r.

Figure 14 depicts simulation results for the error in panel (a) and for the number of iterations to convergence in panel (b). Panels (c) and (d) show the theoretical error
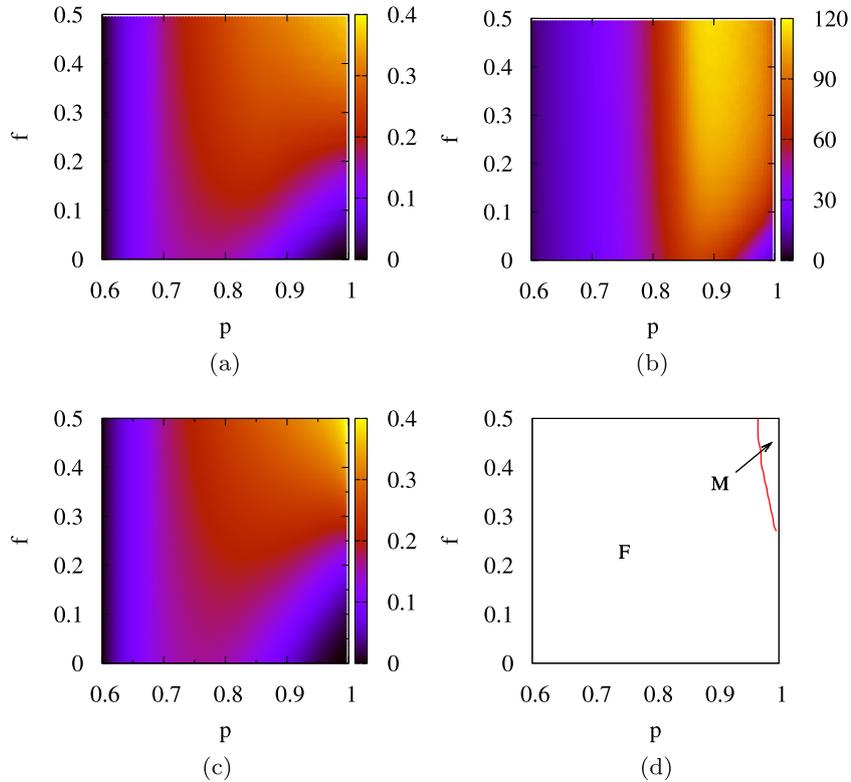
Figure 14. Fraction f of nodes broadcasting random ratings with the fraction of positive ratings fixed to $w = 0.6$. Panels (a) and (b) depict simulation results for the error and for the average number of iterations to convergence. Panels (c) and (d) show the theoretical error and phase diagram using the replica symmetric cavity approach. As can be seen, the error for intermediate values of p increases with f.

and the phase diagram using the replica symmetric cavity approach. The increased time to convergence inside the ferromagnetic phase is explained by the decreased stability of the still unique BP fixed point. Inside the mixed phase new fixed points emerge and the inference process is fundamentally faulty, as can be seen in the top right corner of the error surface.

8. Conclusions

The use of a belief propagation (BP) algorithm for approximate inference in reputation systems has been introduced in [25, 26]. Here we extend previous work by calculating performance measures using the replica symmetric cavity approach after expressing the inference problem in terms of equilibrium statistical mechanics. We also study the convergence times of the BP approach by looking at the algorithm as a dynamical system and apply the framework to a basic scenario and to three simple variations.

The framework is very general and allows a study of the algorithm performance when subjected to several scenarios of practical interest, such as the presence of collusions, parameter mismatches and targeted attacks. Other questions of practical interest remain.

Algorithms based on BP approximate inference seem to represent an interesting alternative for reputation systems such as wireless sensor networks; however, implementation details that have been purposefully ignored in our analysis certainly deserve a more thorough analysis. For instance, sensors often operate with very limited resources, so that sampling of ratings and running of the algorithm should be scheduled taking these limitations into account. Also, faulty elements would behave differently with lower signal to noise rates. In another direction, in a distributed scheme it would be interesting to study the role of different prescriptions for the matrix $\boldsymbol{J}$.

From a theoretical point of view reinforced belief propagation or survey propagation techniques promise better results in the glassy phase. Also, expectation maximization-belief propagation [27] could allow the algorithm to run without the need to supply the signal level p as an input. For scenarios involving targeted attacks more information could be built into the rating mechanism, introducing a duality relationship between the attack and the inference algorithm.

Acknowledgments

Appendix A. Computation of marginals using belief propagation

The algorithm takes as input a distribution $P(h)$ from which the messages are initially sampled, a maximum to the number of iterations $t_{\max}$, a precision ϵ for convergence and estimated values of p and q, $\{\hat{p}, \hat{q}\}$. In what follows, we have used $P^{(0)}(h) = \delta(h)$, $t_{\max} \sim 100$, $\epsilon \sim 10^{-7}$ and $\{\hat{p}, \hat{q}\} = \{p, q\}$—this last condition is later relaxed. The complete pseudocode is as follows:

Input: $P^{(0)}(h)$, $t_{\max}$, ϵ, $\{\hat{p}, \hat{q}\}$; $\{J_{ij}\}$, $\mathcal{G}$
Output: $\{\hat{r}_i\}$
1: initialize $\{h_{i \to j}\}$ sampling from $P^{(0)}(h)$
2: $\beta \leftarrow \alpha_{\hat{p}}$, $B \leftarrow \alpha_{\hat{q}}/\alpha_{\hat{p}}$
3: **while** $\Delta \geq \epsilon$ **and** $t < t_{\max}$ **do**
4:     **for** $i = 1 \to n$, $j \in \partial i$ **do**
5:         $h'_{i \to j} = B + \sum_{k \in \partial i/j} u_{k \to i}(J_{ki}, h_{k \to i})$
6:     **end for**
7:     $\Delta \leftarrow \max |h'_{i \to j} - h_{i \to j}|$
8:     $t \leftarrow t + 1$, $\{h_{i \to j}\} \leftarrow \{h'_{i \to j}\}$
9: **end while**
10: **if** $\Delta < \epsilon$ **then**
11:     **for** $i = 1$ **to** n **do**
12:         $\hat{h}_i = B + \sum_{k \in \partial i} u_{k \to i}(J_{ki}, h_{k \to i})$
13:         $\hat{r}_i = \mathrm{sgn}(\hat{h}_i)$
14:     **end for**
15: **end if**

Appendix B. Population dynamics

The population dynamics algorithm provides an approximate solution to (20) by iterating

$$
\begin{aligned}
u_i^{(\ell)} &= \frac{1}{\beta}\tanh^{-1}\left[\tanh(\beta\xi)\tanh(\beta h_i^{(\ell)})\right], \\
h_i^{(\ell+1)} &= Br + \sum_{j=1}^{c-1} u_j^{(\ell)}.
\end{aligned}
\tag{B.1}
$$

We introduce two arrays of length $N = 10^4$: $\boldsymbol{h}$ and $\boldsymbol{u}$. At the first step, the elements of $\boldsymbol{u}$ are initialized. We have considered two possible ways of initializing $\boldsymbol{u}$: by uniformly sampling from $[-\varepsilon, \varepsilon]$, $\varepsilon = 10^{-2}$, or by assigning $u_i^{(0)} = \xi_i$ (i.e., $\tanh(\beta h_i^{(0)}) = 1$); the results obtained in our analysis were very similar for both. For discussions regarding the use of different initial conditions, the reader may refer to [21, 20].

Next, the elements of $\boldsymbol{h}$ are updated according to the rule, with $\{u_j\}$ uniformly sampled from $\boldsymbol{u}$ and r sampled from $P(r)$; and the elements of $\boldsymbol{u}$ are calculated from the respective element in $\boldsymbol{h}$ and ξ sampled from $P(\xi)$. The process is repeated $\tau = 5000$ times. After this large number of iterations, the array $\boldsymbol{h}$ should be approximately distributed as the real distribution $P(h)$, and we are then able to calculate the desired averages.

In order to calculate $m = \left\langle \tanh(\beta\hat{h}) \right\rangle_{\hat{h}}$, we may introduce an array $\hat{\boldsymbol{h}}$ with the same length N, and since $\hat{h}$ is given by a sum with an extra term, the array elements are computed by simply summing the elements of $\boldsymbol{h}$ with some uniformly sampled element of $\boldsymbol{u}$. We then have $m \approx (1/N)\sum_i \tanh(\hat{\boldsymbol{h}}[i])$.

References

[1] Barbeau M and Kranakis E, 2007 *Principles of Ad Hoc Networking* (New York: Wiley)
[2] Dargie W and Poellabauer C, 2010 *Fundamentals of Wireless Sensor Networks: Theory and Practice* (New York: Wiley)
[3] Sabater J and Sierra C, *Review on computational trust and reputation models*, 2005 *Artif. Intell. Rev.* **24** 33
[4] Mui L, *Computational models of trust and reputation: agents, evolutionary games, and social networks*, 2002 *PhD Thesis* Massachusetts Institute of Technology
[5] Buchmann E, *Trust mechanisms and reputation systems*, 2007 *Algorithms for Sensor and Ad Hoc Networks* ed D Wagner and R Wattenhofer (Berlin: Springer) pp 325–36
[6] Marmol F G and Pérez G M, *Towards pre-standardization of trust and reputation models for distributed and heterogeneous systems*, 2010 *Comput. Standards Interfaces* **32** 185
[7] Josang A and Golbeck J, *Challenges for robust of trust and reputation systems*, 2009 *Proc. 5th Int. Workshop on Security and Trust Management (Sept.)*
[8] Marmol F G and Pérez G M, *Security threats scenarios in trust and reputation models for distributed systems*, 2009 *Comput. Security* **28** 545
[9] Hoffman K, Zage D and Nita-Rotaru C, *A survey of attack and defense techniques for reputation systems*, 2009 *ACM Comput. Surv.* **42** 1
[10] Jøsang A, Ismail R and Boyd C, *A survey of trust and reputation systems for online service provision*, 2007 *Decision Support Syst.* **43** 618
[11] Mui L, Halberstadt A and Mohtashemi M, *Notions of reputation in multi-agents*, 2002 *Proc. AAMAS '02* (New York: ACM Press) pp 280–7
[12] Sabater J and Sierra C, *REGRET: reputation in gregarious societies*, 2001 *Proc. AGENTS '01* (New York: ACM Press) pp 194–5
[13] Page L, Brin S, Motwani R and Winograd T, *The PageRank citation ranking: bringing order to the web*, 1998 *Technical Report* Stanford University

[14] Ermon S, Schenato L and Zampieri S, *Trust estimation in autonomic networks: a statistical mechanics approach*, 2009 *CDC: Proc. 48th IEEE Conf. on Decision and Control* pp 4790–5

[15] Jiang T and Baras J S, *Trust evaluation in anarchy: a case study on autonomous networks*, 2006 *Proc. IEEE Infocom*

[16] Iba Y, *The Nishimori line and Bayesian statistics*, 1999 *J. Phys. A: Math. Gen.* **32** 3875

[17] Kschischang F R, Frey B J and Loeliger H-A, *Factor graphs and the sum–product algorithm*, 2001 *IEEE Trans. Inform. Theory* **47** 498

[18] Loeliger H-A, 2004 *IEEE Signal Process. Mag.* **21** 28

[19] Mezard M and Montanari A, 2009 *Information, Physics, and Computation* (Oxford: Oxford University Press)

[20] Matsuda Y, Nishimori H, Zdeborová L and Krzakala F, *Random-field p-spin-glass model on regular random graphs*, 2011 *J. Phys. A: Math. Theor.* **44** 185002

[21] Zdeborová L, *Statistical physics of hard optimization problems*, 2009 *Acta Phys. Slovaca* **59** 169

[22] Nishimori H, 2001 *Statistical Physics of Spin Glasses and Information Processing* (Oxford: Oxford University Press)

[23] Krzakala F, Ricci-Tersenghi F and Zdeborová L, *Elusive spin-glass phase in the random field Ising model*, 2010 *Phys. Rev. Lett.* **104** 207208

[24] Montanari A, *Estimating random variables from random sparse observations*, 2008 *Eur. Trans. Telecommun.* **19** 385

[25] Ermon S, *Trust estimation in autonomic networks: a message passing approach*, 2009 *Proc. NECSYS (Venice, 2009)*

[26] Ayday E, *Application of belief propagation to trust and reputation management*, 2011 *ISIT: Proc. IEEE Int. Symp. on Information Theory* pp 2173–7

[27] Decelle A, Krzakala F, Moore C and Zdeborová L, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, 2011 *Phys. Rev. E* **84** 066106