Retrieving information from a noisy ‘knowledge network’

J Barré

Université de Nice-Sophia Antipolis, Laboratoire J. A. Dieudonné UMR CNRS 6621, Parc Valrose 06108 Nice Cedex 02, France
E-mail: jbarre@unice.fr

Received 1 May 2007
Accepted 5 July 2007
Published 6 August 2007

Abstract. We address the problem of retrieving information from a noisy version of the ‘knowledge networks’ introduced by Maslov and Zhang (2001 Phys. Rev. Lett. 87 248701). We map this problem onto a disordered statistical mechanics model, which opens the door to many analytical and numerical approaches. We give the replica symmetric solution, compare it with numerical simulations, and finally discuss an application to real data from the United States Senate.

Keywords: cavity and replica method, message-passing algorithms, communication, supply and information networks

ArXiv ePrint: 0704.3983
1. Introduction

In a recent paper [1], Maslov and Zhang addressed the following problem: we are given \( N \) agents, each one represented by an \( M \)-dimensional real vector \( \vec{r}_i \); suppose we know \( K \) of the \( N(N - 1)/2 \) scalar products \( \Omega_{ij} = \vec{r}_i \cdot \vec{r}_j \) with \( i \neq j \). In this situation, can we predict the value of an unknown scalar product \( \Omega_{ij} \)? This question is relevant for instance to the problem of extracting information from the vast amount of data generated by a commercial website. The \( \vec{r}_i \) may represent in that context the interests of a person \( i \), and \( \Omega_{ij} \) the mutual appreciation of persons \( i \) and \( j \); the problem is then to predict the mutual appreciation of two persons that do not know each other. Maslov and Zhang called the network of interactions and overlaps \( \Omega_{ij} \) an ‘knowledge network’.

One of their main results is the following: there exists a critical density of known overlaps \( p_c = 2K/N(N - 1) \) above which almost all the \( a \ priori \) unknown overlaps are completely determined by the \( K \) known ones. This transition is a realization of the so-called rigidity percolation. However, their treatment leaves several important issues aside, and assumes that we have at our disposal much more information that we typically do. For instance, the size of the vectors \( M \) describing each agent is \( a \ priori \) unknown; the problem of estimating \( M \) from the data was addressed in [2]. More drastically, the data on the overlaps is necessarily noisy: if \( \vec{r}_i \) and \( \vec{r}_j \) model the interests of persons \( i \) and \( j \), their mutual appreciation \( \Omega_{ij} \) is certainly not completely determined by the overlap of their interests \( \vec{r}_i \cdot \vec{r}_j \), although it is probably biased by it. In this more realistic case of noisy information, the questions are: does the ‘phase transition’ noted by Maslov and Zhang survive? and how do we retrieve the information contained in the noisy knowledge

---

\(1\) These authors actually introduce a bipartite version of these networks.
network? We address these issues in the following by studying a simple model of this situation.

The outline of the paper is as follows: we present in section 2 the details of the model we are going to study, and the mapping onto a disordered statistical mechanics problem, which happens to be the one studied in [3] and more recently in [4]. This mapping opens the door to the use of many analytical and numerical methods. In section 3, we give the solution of this problem at the replica symmetric level, using the cavity method [5]. We then check these analytical results against numerical simulations in section 4, and real data from the United States Senate in section 5.

2. The model

We present now the noisy version of Maslov and Zhang’s ‘knowledge network’ which we are going to study; for simplicity, the variable describing each agent is discrete, and one dimensional. We consider $N$ agents; each one is characterized by an opinion $s_0^i$, with $i = 1, \ldots, N$; the $s_0^i$ may take $k$ different values, and are a priori unknown. The $s_0^i$ may be for instance political opinions, as in the example of section 5. We suppose we have some information on the $s_0^i$, given by an analog of the ‘overlaps’ of [1]: for a certain number of pairs $(i, j)$ we know a number $J_{ij}$ associated with it, constructed as follows. If $s_0^i = s_0^j$, then $J_{ij} = 1$ with probability $1 - p$, and $J_{ij} = -1$ with probability $p$; if $s_0^i \neq s_0^j$, then $J_{ij} = 1$ with probability $p$, and $J_{ij} = -1$ with probability $1 - p$. We take $p \leq 1/2$. $p$ is then a measure of the noise in the information; in the limit $p = 1/2$, the network does not convey any information on the $s_0^i$. The basic questions we ask are: how well can we reconstruct the actual opinions $s_0^i$ knowing the $J_{ij}$? and do we have an effective algorithm to do so?

We are interested in the probability of any set of opinions $\{s_i\}_{i=1,\ldots,N}$, given the $J_{ij}$ representing our knowledge; from Bayes formula, we can write:

$$P(\{s_i\}|\{J_{ij}\}) = \frac{P(\{s_i\})}{P(\{J_{ij}\})} P(\{J_{ij}\}|\{s_i\}).$$

(1)

The factor $P(\{s_i\})$ is the prior probability on the $s_i$; we suppose from now on that it is flat, so that this term is independent of the $s_i$. It would be possible however to consider another prior probability. The factor $P(\{J_{ij}\})$ is difficult to compute, as the $J_{ij}$ are correlated in an intricate way; however, it is in any case independent of the $s_i$, so it acts as a normalization factor for the distribution (1). Finally, the $P(\{J_{ij}\}|\{s_i\})$ is easy to compute, since once the $s_i$ are given, the $J_{ij}$ are independent. Let us consider two agents 1 and 2 with opinions $s_1$ and $s_2$; then from simple algebra one checks that

$$P(J_{12}|s_1, s_2) = \sqrt{\frac{1-p}{p}} \left( \sqrt{\frac{1-p}{p}} \right)^{1/2} J_{12} \delta_{s_1,s_2}.$$

(2)

Since the $J_{ij}$ are independent once the $s_i$ are given, equation (1) may be rewritten as

$$P(\{s_i\}|\{J_{ij}\}) \propto \prod_{(i,j)} \left( \sqrt{\frac{1-p}{p}} \right)^{(1/2)J_{ij}\delta_{s_i,s_j}}.$$

(3)
Retrieving information from a noisy ‘knowledge network’

where the index \( \langle i, j \rangle \) means that the product runs over the pairs \( (i, j) \) that are connected by a known \( J_{ij} \). Taking the logarithm, we have:

\[
H[\{s_i\}] = -\log P(\{s_i\}|\{J_{ij}\}) = \text{Cste} - B \sum_{(i,j)} J_{ij} (2\delta_{s_i s_j} - 1),
\]

(4)

with

\[
B = \frac{1}{2} \log \left( \frac{1-p}{p} \right).
\]

Equation (4) can be seen as the Hamiltonian of a disordered Potts model, which opens the door to the use of many analytical and numerical tools to study it. From now on, we will concentrate for simplicity on the Ising case, where each agent may have only two opinions, \( s_i = +1 \) or \( s_i = -1 \). In this Ising case, the Hamiltonian reads:

\[
H[\{s_i\}] = -\log P(\{s_i\}|\{J_{ij}\}) = \text{Cste} - B \sum_{(i,j)} J_{ij} s_i s_j.
\]

(5)

The sets \( \{s_i\} \) with maximum probability are the minimizers of equation (4); the minimizer is not necessarily unique. The question of how well we can reconstruct the real opinions knowing the \( J_{ij} \) is then rephrased as: given a minimizer \( \{s_i^*\} \) of equation (4), how far is it from the real opinions \( \{s_i^0\} \)? We answer this question in the next section. We note that this rephrasing of the problem bears some resemblance with the community detection, or clustering problem as stated in [6]; in this work however, the probabilistic analysis yields a Potts-like model without disorder.

3. Cavity solution

3.1. Gauge transformation

Hamiltonian (5) is not as well-suited for analytical treatment as it seems to be. It is a disordered Ising model, but the probability distribution of the couplings \( J_{ij} \) is not known, and actually very complicated: the relevant information we want to extract is precisely hidden in the correlations between the \( J_{ij} \). The following gauge transformation, somewhat miraculously, yields a tractable problem.

We define \( \tilde{s}_i = s_i^0 s_i \) and \( \tilde{J}_{ij} = s_i^0 s_j^0 J_{ij} \) (the \( s_i^0 \) are the true opinions of the agents); then

\[
H[\{s_i\}] = -B \sum_{(i,j)} \tilde{J}_{ij} \tilde{s}_i \tilde{s}_j.
\]

(6)

The distribution of the \( \tilde{J}_{ij} \) does not depend any more on the \( s_i^0 \): \( \tilde{J}_{ij} = 1 \) with probability \( 1-p \) and \( \tilde{J}_{ij} = -1 \) with probability \( p \): all correlations in the couplings have disappeared. Furthermore, given a set \( \{\tilde{s}_i\} \), it is easy to know how far the corresponding set \( \{s_i\} \) is from the original \( \{s_i^0\} \): it is enough to compute the number of \( \tilde{s}_i \) equal to \(-1\). Thus we are left with the study of Hamiltonian (6), which is that of a ferromagnetically biased Ising spin glass. We would like to compute the magnetization of the ground state of such a Hamiltonian. From now on, we remove the \( \sim \) on the \( J \) and the \( s \). Let us note that the ground state does not depend on \( B \), so we may take \( B = 1 \) for simplicity (as long
as $B > 0$, that is $p < 1/2$). All explicit dependence on $p$ is then removed, which is very convenient for practical purposes, as $p$ is a priori unknown: the knowledge of the $J_{ij}$ is sufficient to determine the minimizers of (6). We need however to keep $p$ as a parameter in the theoretical analysis, and will turn later to the issue of estimating it.

3.2. Replica symmetric solution

It turns out that the ferromagnetically biased Ising spin glass given by Hamiltonian (6) has been studied recently by Castellani et al in [4] for fixed connectivity graphs. In the present context, it is more natural to consider random graphs of Erdős–Rényi type, with a Poissonian distribution of connectivity. However, such a change from fixed to Poissonian connectivity usually does not induce any qualitative change in the phase diagram.

Castellani et al use the cavity method [5] to compute, among other quantities, the one

$\pi(p) = e^{-\gamma k}/k!$

where $s\gamma n$ is the sign function, taken to be zero when the argument is zero;

It turns out that the ferromagnetically biased Ising spin glass given by Hamiltonian (6) is then described by a single probability distribution:

$P(u) = q_+ \delta(u - 1) + q_0 \delta(u) + q_- \delta(u + 1)$.

We write a recursion relation for the probability distribution $P$ as follows:

$P(u) = \sum_{k=0}^{\infty} e^{-\gamma k} E_J \prod_{i=1}^{k} dP(u_i) \delta \left( u - \text{sgn} \left( J \sum_{i=1}^{k} u_i \right) \right)$,

(8)

where $\text{sgn}$ is the sign function, taken to be zero when the argument is zero; $E_J$ means ‘expectation’ over the coupling $J$. Equation (8) straightforwardly translates into three fixed point equations for $q_0, q_+$ and $q_-$ (see figure 1 for an explanation of $k_+, k_-$ and $k_0$):

\begin{align*}
q_0 &= \sum_{k=0}^{\infty} e^{-\gamma k} \sum_{k_0=0}^{k} \sum_{k_0+k_+=k} \frac{q_+ q_- q_0}{k_0! k_+! k_-! k_0!}, \\
q_+ &= \sum_{k=0}^{\infty} e^{-\gamma k} \left[ (1 - p) \sum_{k_0=0}^{k} \sum_{k_0+k_+<k} \frac{q_+ q_- q_0}{k_0! k_-! k_0!} + p \sum_{k_0=0}^{k} \sum_{k_0+k_+>k} \frac{q_+ q_- q_0}{k_0! k_+! k_0!} \right], \\
q_- &= 1 - q_+ - q_0.
\end{align*}

(9)
Once \( q_0, q_+ \) and \( q_- \) are known, the ground state magnetization is given by the expression:

\[
m = \sum_{k=0}^{\infty} e^{-\gamma \gamma^k} \left[ \sum_{k_0=0}^{k} \sum_{k_0+k_+=k_-=k} \frac{k_+q_-q_0}{k_+!k_--!k_0!} - \sum_{k_0=0}^{k} \sum_{k_0+k_+=k} \frac{k_-q_+q_0}{k_-!k_+!k_0!} \right].
\]  

(10)

To compute the ground state energy, one computes the energy shifts \( \Delta E_s \) due to the addition of a site, and \( \Delta E_l \) due to the addition of a link. One gets after straightforward calculations:

\[
\Delta E_s = \sum_{k=0}^{\infty} e^{-\gamma \gamma^k} \sum_{k_0+k_+=k} \frac{k_+q_-q_0}{k_+!k_--!k_0!} (-k_0 - |k_+ - k_-|)
\]  

(11)

\[
\Delta E_l = -\frac{(q_+ - q_-)^2}{1 - 2p} - 2q_0^2 + q_0^2.
\]  

(12)

The ground state energy \( e_{gs} \) is then given by

\[
e_{gs} = \Delta E_s - \frac{\gamma}{2} \Delta E_l.
\]  

(13)

The qualitative picture emerging from this replica symmetric analysis is the following: for each mean connectivity \( \gamma > 1 \), there is a critical value \( p_c^{RS}(\gamma) \) such that for \( p < p_c^{RS}(\gamma) \), it is possible to extract information from the knowledge network. The error rate \( \varepsilon \) in the \( N \to \infty \) limit is directly related to the ground state magnetization \( m \):

\[
\varepsilon(p, \gamma) = \frac{1 - m(p, \gamma)}{2}.
\]

For \( p > p_c^{RS}(\gamma) \), it is not possible any more to extract meaningful information from the data in the limit \( N \to \infty \): the error rate tends to 1/2.

3.3. Discussion

We compare these replica symmetric analytic results to numerical simulations in the next section. We can make however some \textit{a priori} remarks on the validity of the calculation. First, we expect the calculations to be exact at small enough \( p \); we then expect a replica...
symmetry breaking transition at some $p_{RSB}(\gamma) < p_c^{RS}(\gamma)$. For $p > p_{RSB}(\gamma)$, the replica symmetric results are not reliable any more. We expect that the phase transition described above towards a non-magnetized ground state is shifted to some $p_c^{RSB} \neq p_c^{RS}$. However, the qualitative result of a transition between one phase which contains some information and another one which does not should still hold true.

Another word of caution is in order: the authors of [4] note strong finite size effects for a fixed connectivity network; this is likely to be the case also for a Poissonian network, and it may smear out somewhat the transition for finite $N$.

3.4. Estimating $p$

As already noted above, equation (6) only depends on $p$ through the parameter $B$, so an a priori knowledge of $p$ is not necessary to carry out the minimization. This is an interesting practical advantage. However, the amount of errors contained in the minimizer strongly depends on $p$, as explained above. So it would be useful to have some information about the value of $p$, to get an estimate of the amount of errors contained in the ground state. It is indeed in some cases possible to estimate $p$ from the only available data, the $J_{ij}$. Suppose we are given a network. It is possible to compute for this network $e_{GS}(p)$, the ground state energy as a function of $p$, by randomly choosing the $J_{ij}$ with probability $p$; this can be done analytically in some cases with the cavity method, or numerically. Then one computes the ground state of the network with the real $J_{ij}$ from the data; comparing with the $e_{GS}(p)$ curve is not flat.

4. Numerical simulations

We now compare the analytical prediction of the previous section to data generated randomly: we randomly assign a value $S_{i}^{(0)} = 1$ or $S_{i}^{(0)} = -1$ to $N$ spins; we randomly draw a network connecting these spins, and randomly assign a value 1 or −1 to each link $J_{ij}$ connecting spins $i$ and $j$, following the rule:

$$J_{ij} = S_{i}^{(0)} S_{j}^{(0)} \quad \text{with probability } 1 - p,$$

$$J_{ij} = -S_{i}^{(0)} S_{j}^{(0)} \quad \text{with probability } p.$$

We then numerically minimize the corresponding Hamiltonian. For this purpose, we may use simulated annealing. It is simple to program, but not very fast, and does not perform well in the replica symmetry broken phase. However, the structure of the problem may suggest to use another class of algorithm, intensively studied in different contexts recently (see for instance [8] for a pedagogical introduction in the context of error correcting codes): Belief Propagation (BP). BP is not expected to perform better than simulated annealing in the replica symmetry broken phase, and it may sometimes fail to converge. However, it performs overall very well, and is much faster than simulated annealing, which allows us to reach higher $N$: this is crucial to deal with large data sets.

In figure 2, one sees that the agreement between simulations using BP and replica symmetric calculations is very good for low $p$. For larger $p$, there are important discrepancies, that may have two origins. First, one expects a replica symmetry breaking, as in [4]; this means that the replica symmetric calculation is not exact any more, and that BP is not expected to perform well. Second, as already noticed in [4], finite size effects are
Retrieving information from a noisy ‘knowledge network’

Figure 2. Energy (left) and error rate (right) as a function of $p$, for a $\gamma = 3$ Poissonian random graph. Symbols are from numerical simulations using the BP algorithm, with $N = 8000$; the solid and dashed lines are the replica symmetric analytical results.

strong. However, the numerical results seem compatible with the main analytical finding: the presence of a transition between a low $p$ phase which contains information, and a high $p$ one that does not. We also note that the error rate obtained with BP is always smaller than the theoretical one estimated from the replica symmetric analysis. Finally, it is interesting to compare quantitatively these results with those of [4] for regular graphs: both theory and numerics predict a significantly higher threshold between the informative and non-informative phases for a Poissonian network, for a given mean connectivity.

BP does have another big advantage over simulated annealing: its outcome is a magnetization for each site, so we also have an indication on which sites are most likely to be wrongly guessed (those with magnetization close to zero). As a final remark, it could be possible to improve performance in the replica symmetry broken phase by using a survey propagation algorithm [7].

5. The US Senate example

The analytical results of section 3 are strengthened by the numerical simulations of section 4; however, unlike the numerical data, any real data set does not follow exactly the probabilistic model underlying our study. It is thus important to assess how robust the results are with respect to some uncertainty in the model. In this section, we will analyze data from the United States senate votes, and show that the strategy of minimizing Hamiltonian (6) does allow us to retrieve some information from the data; the amount of information retrieved is in reasonable quantitative agreement with the predictions of section 3.2

We consider here as agents the 100 US Senators serving in 2001. The party of each senator plays the role of the unknown opinion $s_i^0$; say $s_i^0 = -1$ if senator $i$ is

2 We certainly do not claim that the present method is the best possible to extract information from the US Senate data; we only try to test the robustness of our results on a real data set.
Figure 3. The parameter $p$ is fixed, $p = 0.2$. For each value of $\gamma$, the symbols correspond to 100 realizations. The solid line is the analytical replica symmetric result.

Varying the random network and the random pick of the roll call votes for each link, we can generate many different instances of the ‘knowledge network’ for each $\gamma$.

As senators from the same (respectively, different) party tend to cast the same (respectively, different) vote, they tend to be linked by edges with positive (respectively, negative) $J$. The fact that senators do not always vote like the majority of their colleagues from the same party plays the role of a noise. We crudely model this situation as in section 2, assuming that $J_{ij} = s_i^0 s_j^0$ with probability $1 - p$, and $J_{ij} = -s_i^0 s_j^0$ with probability $p$, $p$ being unknown, smaller than $1/2$. We now want to retrieve some information about the $s_i^0$ (i.e. the party of each senator), using the method described in this paper.

Based only on the set of the $J_{ij}$, we run the BP algorithm for each instance of the ‘knowledge network’, without using any a priori knowledge on the parameter $p$; we then split the senate into Republicans and Democrats, according to the BP results. We can check how many errors we have, and compare this with the theory of section 3. Note that we can choose the connectivity of the random network $\gamma$. We have no control however on the parameter $p$.

\footnote{In practice, we have collected the data from 50 roll call votes in early 2001.}
Retrieving information from a noisy 'knowledge network'

The results are presented in figure 3, and compared to the replica symmetric analytical calculations. They seem to be consistent with the main qualitative analytical result: the existence of a threshold separating a phase containing almost no information (low $\gamma$) and a phase which contains some (high $\gamma$). We also see in figure 3, that there is a strong sample to sample variability; for small error rates however (large values of the mean connectivity $\gamma$), the agreement is rather good; for smaller $\gamma$, the agreement is poor. There are two explanations for that, besides the fact that the votes are not random: replica symmetry is probably broken, and, more important for such small systems ($N = 100$), finite size effects create large bias. We note however that the practical error rate is usually smaller than the analytical one.

6. Conclusion

We have extended the 'knowledge network' formalism of [1] to the more realistic case of noisy data. We have shown that there is a phase transition between an information-rich phase, and a phase that essentially contains no information. In the former situation, the information may be efficiently retrieved through a Belief Propagation algorithm.

There are several possible extensions to this work. The most direct ones are the study of non-binary opinions (Potts-like models), or multidimensional opinions. With the applications to commercial websites in mind presented in [1, 2], it would also be interesting to consider bipartite networks. For all of these cases, it seems that the disordered statistical mechanics point of view used in this paper may be fruitful, by suggesting the use of some powerful analytical as well as numerical techniques.

Acknowledgment

The author thanks Christophe Giraud and Florent Krzakala for discussions about this work.

References

[1] Maslov S and Zhang Y-C, Extracting hidden information from knowledge networks, 2001 Phys. Rev. Lett. 87 248701
[2] Bagnoli F, Berrones A and Franci F, De gustibus disputandum (forecasting opinions by knowledge networks), 2004 Physica A 332 509
[3] Kwon C and Thouless D, Ising spin glass at zero temperature on the Bethe lattice, 1988 Phys. Rev. B 37 7649
[4] Castellani T, Krzakala F and Ricci-Tersenghi F, Spin glass models with ferromagnetically biased couplings on the Bethe lattice: analytic solutions and numerical simulations, 2005 Europhys. J. B 47 1434
[5] Mézard M and Parisi G, The cavity method at zero temperature, 2003 J. Stat. Phys. 111 1
[6] Hastings M B, Community detection as an inference problem, 2006 Phys. Rev. E 74 035102
[7] Mézard M and Zecchina R, Random K-satisfiability problem: From an analytic solution to an efficient algorithm, 2002 Phys. Rev. E 66 056126
[8] Montanari A, Two lectures on iterative coding and statistical mechanics, 2005 Les Houches [cond-mat/0512296]