Inference and Phase Transitions in the Detection of Modules in Sparse Networks

Aurelien Decelle,[1] Florent Krzakala,[2] Cristopher Moore,[3] and Lenka Zdeborová[4,*]

[1]*Université Paris-Sud and CNRS, LPTMS, UMR8626, Bâtiment 100, Université Paris-Sud 91405 Orsay, France*
[2]*CNRS and ESPCI ParisTech, 10 rue Vauquelin, UMR 7083 Gulliver, Paris 75000, France*
[3]*Computer Science Department, University of New Mexico, and the Santa Fe Institute, Santa Fe, New Mexico 87501, USA*
[4]*Institut de Physique Théorique, IPhT, CEA Saclay, and URA 2306, CNRS, 91191 Gif-sur-Yvette, France*

(Received 8 February 2011; revised manuscript received 5 July 2011; published 2 August 2011)

We present an asymptotically exact analysis of the problem of detecting communities in sparse random networks generated by stochastic block models. Using the cavity method of statistical physics and its relationship to belief propagation, we unveil a phase transition from a regime where we can infer the correct group assignments of the nodes to one where these groups are undetectable. Our approach yields an optimal inference algorithm for detecting modules, including both assortative and disassortative functional modules, assessing their significance, and learning the parameters of the underlying block model. Our algorithm is scalable and applicable to real-world networks, as long as they are well described by the block model.

DOI: 10.1103/PhysRevLett.107.065701 PACS numbers: 64.60.aq, 75.10.Hk, 89.75.Hc

In many networks, ranging from online communities to gene regulatory networks, nodes belong to modules or communities that play distinct functional roles. A fundamental problem is to detect these communities and understand what role they play in the network's structure and dynamics. Many algorithms for this problem have been suggested over the past decade. However, most of these methods suffer from two conceptual problems. First, as pointed out previously [1,2], they do not provide a measure of the significance of the division into communities, and they falsely detect communities even in purely random graphs. Second, they are limited to *assortative* community structure, where nodes are more likely to be connected to other nodes of the same type. In many real-world networks, such as food webs, metabolic, and word adjacency networks, nodes belong to *functional* communities—rather than connecting to each other, they connect to the rest of the network in similar ways.

Here we present an approach that resolves both these problems and is asymptotically exact for networks generated by a widely used stochastic block model. It is applicable to real-world networks that are reasonably well described by this model and can also be generalized to other local generative models such as Ref. [3]. We combine a Bayesian approach with the cavity method developed in statistical physics [4,5], leading to a message-passing algorithm, known in computer science and information theory as belief propagation (BP) [6], for detecting functional modules and learning the model parameters. Our approach provides a natural measure of the significance of the modules in the network, as it outputs the marginal probability that a given node belongs to a given group. If the network does not contain any modules, it correctly infers this fact by making these marginals uniform. Both these aspects are missing in the vast majority of current approaches to community detection.

We also unveil striking theoretical aspects of the block model. As a function of its parameters, we discover several sharp phase transitions. We distinguish between a *detectable* phase, where it is possible to learn the model's parameters and the group memberships of the nodes, and a nonintuitive *undetectable* phase, where learning is impossible because the network's topology does not retain enough information about the true group memberships. The existence of a phase where particular algorithms are unable to detect communities was previously predicted [1,7,8], but our results about undetectability are much stronger: Assuming that BP determines the correct marginals, which it does modulo standard assumptions, our results are algorithm independent, showing that *no* algorithm can determine the groups. We also find a transition from a ''hard detectable'' phase, where the network has enough information to determine the groups but where, we believe, no polynomial algorithm can find them, to an ''easy'' phase where polynomial-time algorithms do exist.

Stochastic block model.—We consider networks of N nodes. Each node i has a hidden label $t_i \in \{1, \ldots, q\}$, specifying which of q groups it is a member of. These labels are chosen independently, where n_a is the probability that a given node has label $a \in \{1, \ldots, q\}$ (normalized so that $\sum_{a=1}^{q} n_a = 1$). If N_a is the number of nodes in each group, we have $n_a = \lim_{N\to\infty} N_a/N$.

Once the group assignments are chosen, the model generates a graph G as follows. For each pair of nodes (i, j) we put an edge between i and j independently with probability p_{t_i,t_j}, leaving them unconnected with probability $1 - p_{t_i,t_j}$. We call p_{ab} the *affinity* matrix. Since we are interested in the sparse case where $p_{ab} = O(1/N)$, we will use the

rescaled affinity matrix $c_{ab} = Np_{ab}$ and assume that $c_{ab} = O(1)$ in the limit $N \to \infty$.

In our setting, the adjacency matrix A_{ij} of the graph is the only information available to us. Our goal is to learn the parameters q, $\{n_a\}$, and $\{c_{ab}\}$ of the block model, as well as the group assignments $\{t_i\}$. Special cases of this model have often been considered in the literature. Planted partitioning, when $n_a = 1/q$, $c_{ab} = c_{\text{out}}$ for $a \neq b$, and $c_{aa} = c_{\text{in}}$ with $c_{\text{in}} > c_{\text{out}}$, is a classical problem in computer science and has been used as a benchmark for community detection [2,7,9–11]. Planted coloring, where $n_a = 1/q$, $c_{aa} = 0$, and $c_{ab} = cq/(q-1)$, is a fundamental problem in constraint optimization [5] and was studied by using the cavity method in Ref. [12].

Bayesian inference for block models.—Bayesian inference has been applied to community detection before. However, except for some very specific generative models (e.g., [13,14]), the likelihood function for sparse networks must be computed approximately, either through Monte Carlo sampling (e.g., [15]) or variational methods [11]. We note that the authors of Ref. [16] studied Bayesian inference for dense networks, $c_{ab} = O(N)$, and their results can be recovered from our work in the limit $c \to \infty$.

The crucial contribution of our work is that the quantities that follow from Bayesian inference can be analyzed *exactly* in the thermodynamic limit by using the cavity method [4,5]. The probability that the model parameters take a given set of values $\{\theta\} = (q, \{n_a\}, \{c_{ab}\})$, conditioned on the topology of the network G, is

$$P(\{\theta\} \mid G) = \frac{P(\{\theta\})}{P(G)} \sum_{\{t_i\}} P(G, \{t_i\} \mid \{\theta\}). \tag{1}$$

The sum is over all possible group assignments $\{t_i\}$, where $t_i \in \{1, \ldots, q\}$ for each node i. The prior $P(\{\theta\})$ includes all graph-independent information about the values of the parameters. We will assume there is no such information available and hence this prior is uniform. In that case, maximizing $P(\{\theta\} \mid G)$ over $\{\theta\}$ is equivalent to maximizing the sum $\sum_{\{t_i\}} P(G, \{t_i\} \mid \{\theta\})$.

The function $P(G, \{t_i\} \mid \{\theta\})$ is called the *likelihood*. It is the probability that the model would produce the group assignment $\{t_i\}$ and the network G, assuming that its parameters are $\{\theta\}$. We can write the likelihood exactly for many different generative models; for the stochastic block model defined above, it is

$$P(G, \{t_i\} \mid \{\theta\}) = \prod_i n_{t_i} \prod_{i<j} [p_{t_i,t_j}^{A_{ij}} (1 - p_{t_i,t_j})^{1-A_{ij}}].$$

Thus $P(\{\theta\} \mid G)$ is proportional to the partition sum $Z(\{\theta\})$ of a generalized Potts model, with Hamiltonian

$$\mathcal{H}(\{t_i\}) = -\sum_i \log n_{t_i} - \sum_{i<j} \begin{cases} \log c_{t_i,t_j} & A_{ij} = 1, \\ \log(1 - \frac{c_{t_i,t_j}}{N}) & A_{ij} = 0. \end{cases} \tag{2}$$

There is a strong $O(1)$ interaction between connected nodes and a weak $O(1/N)$ one between unconnected nodes. The $\log n_{t_i}$ play the role of local fields, enforcing the prior distribution $\{n_a\}$ on group assignments.

Inferring the parameters $\{\theta\}$ is equivalent to minimizing the free energy $f(\{\theta\}) = -\log Z(\{\theta\})/N$ associated with (2). If $f(\{\theta\})$ has a nondegenerate minimum, then, from the saddle point method, $\{\theta\}$ is with high probability exactly the set of parameters used in the generation of the network. In that case, inferring the parameters of the underlying model is possible.

Assuming that we know, or have learned, the correct parameters $\{\theta\}$, how should we determine the group assignment of the nodes? The most likely assignment $\{t_i\}$ is the ground state of the Hamiltonian (2). However, if we want to find an assignment $\{t_i\}$ that maximizes the number of correctly labeled nodes, we need to compute the marginal distribution $\nu_i(t_i) = \sum_{\{t_j\}_{j\neq i}} \mu(\{t_j\}_{j\neq i}, t_i)$ of the label of each node i, where μ is the Boltzmann distribution of (2). Note that a configuration chosen according to the Boltzmann distribution has, asymptotically, the correct group sizes and the correct number of edges between each pair of groups, while for the ground state this is not true; finding the minimum bisection, for instance, creates the illusion of two groups even in a completely random graph [17]. The marginal $\nu_i(t_i)$ is the probability that node i belongs to group t_i, and the most probable group assignment is $t_i^* = \text{argmax}_{t_i} \nu_i(t_i)$. The expected fraction of correctly labeled nodes when $N \to \infty$ is $A = \sum_i \nu_i(t_i^*)/N = \max_{\pi \in S_q} \sum_i \delta_{t_i^*, \pi(t_i)}/N$, where t_i is the true group assignment and S_q is the permutation group of q elements. We define the overlap as $Q = (A - \max_a n_a)/(1 - \max_a n_a)$. Note that this is zero, for instance, if all nodes are assigned to the largest group.

Belief propagation.—We could estimate the free energy by using Monte Carlo (MC) sampling, and we do this below and in Ref. [18] for comparison. But a faster algorithm is BP, known in physics as the cavity method [4,5]. It is exact in the thermodynamic limit as long as the network is locally treelike and as long as correlations decay rapidly as a function of topological distance. BP was proposed for community detection in Ref. [10] but without the crucial ability to learn the parameters of the underlying model.

To derive the BP equations [5,6], one introduces cavity marginals, or "messages," $\psi_{t_i}^{i\to j}$ and $\psi_{t_j}^{j\to i}$ that are sent from one node to another along each edge (i, j). For instance, $\psi_{t_i}^{i\to j}$ is the probability that i would be in group t_i if j were absent from the network. Assuming conditional independence between the neighbors of each node and neglecting lower-order terms, the messages must be a fixed point of a self-consistency equation:

$$\psi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}} n_{t_i} e^{-h_{t_i}} \prod_{k \in \partial i \setminus j} \left[\sum_{t_k=1}^{q} c_{t_k t_i} \psi_{t_k}^{k\to i} \right] \tag{3}$$

for each edge (i, j). Here ∂i is the set of i's neighbors, $Z^{i\to j}$ is a normalizing factor ensuring that $\sum_{t_i} \psi_{t_i}^{i\to j} = 1$, and the external field $h_{t_i} = \frac{1}{N}\sum_{k=1}^{N}\sum_{t_k=1}^{q} c_{t_k t_i}\nu_k(t_k)$ summarizes the influence of the nonedges via $\nu_i(t_i) = (1/Z^i)n_{t_i}e^{-h_{t_i}}\prod_{j\in\partial i}[\sum_{t_j} c_{t_j t_i}\psi_{t_j}^{j\to i}]$. For more details about implementing BP, see [18].

We start with random messages and iterate (3) until we reach a fixed point. The marginals corresponding to the fixed point are $\nu_i(t_i)$ [18], and the free energy is

$$f_{\mathrm{BP}}(\{\theta\}) = -\frac{1}{N}\sum_i \log Z^i + \frac{1}{N}\sum_{(i,j)\in E}\log Z^{ij} - \frac{c}{2},$$

where $Z^{ij} = \sum_{a>b} c_{ab}(\psi_a^{i\to j}\psi_b^{j\to i} + \psi_b^{i\to j}\psi_a^{j\to i}) + \sum_a c_{aa}\psi_a^{i\to j}\psi_a^{j\to i}$. Requiring that $f_{\mathrm{BP}}(\{\theta\})$ is stationary, we update the parameters to their most likely values:

$$c'_{ab} = \sum_{(i,j)\in E} c_{ab}(\psi_a^{i\to j}\psi_b^{j\to i} + \psi_b^{i\to j}\psi_a^{j\to i})/(Z^{ij}n_a n_b N),$$

and $n'_a = \sum_i \nu_i(a)/N$. Starting with a suitable initial value $\{\theta_0\}$, we compute $\{\theta'\}$ and iterate until convergence (see Fig. 1 and Ref. [18]) as in the expectation-maximization algorithm [19]. Each iteration takes time proportional to the number of edges, and a constant number of iterations is needed for convergence (see Fig. 2). Thus like the expectation-maximization algorithms of Refs. [13,14], the total running time of our algorithm is linear in N.

In order to learn the number of groups, we note that the free energy $f_{\mathrm{BP}}(q)$ decays with q and then stays constant (in the limit $N \to \infty$) for $q \geq q_{\mathrm{actual}}$. Then the correct number of groups q_{actual} is learned by running the algorithm for several values of q until f_{BP} stops decreasing.

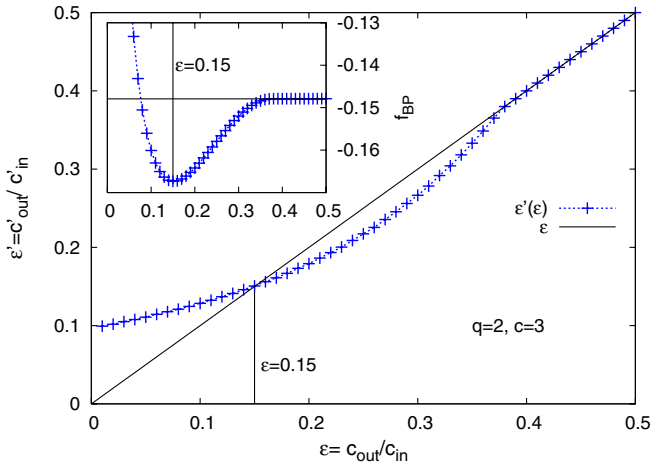

FIG. 1 (color online). Learning for $q = 2$ groups with $n_a = 1/2$, average degree $c = 3$, and $\epsilon = c_{\mathrm{out}}/c_{\mathrm{in}} = 0.15$. If we initialize our algorithm in the ordered region, i.e., with $\epsilon_0 < 0.37$, it infers the correct value of ϵ. Inset: The free energy as a function of ϵ. Note the minimum at $\epsilon = 0.15$ and the paramagnetic region for $\epsilon > 0.37$.

Phase diagrams.—For illustration, we use the case of planted partitions and colorings: $n_a = 1/q$, $c_{ab} = c_{\mathrm{out}}$ for $a \neq b$, and $c_{aa} = c_{\mathrm{in}}$. We observe three different cases governing the free energy landscape $f_{\mathrm{BP}}\{\theta\}$. In the "paramagnetic" phase, the free energy is constant in the vicinity of the true parameters $\{\theta\}$. Learning is impossible, and the marginals are $\nu_i(t_i) = 1/q$ for all nodes. In this case the overlap between the true assignment and the one resulting from BP marginalization is zero, and the true assignment is undetectable. Generalizing [12,20], one can show there is essentially no difference between a graph produced by the block model and a completely random graph of the same average degree.

In the *ordered* phase, f_{BP} has an attractive global minimum at the true parameters $\{\theta\}$, and BP rapidly infers them. This is illustrated in Fig. 2. As $\epsilon = c_{\mathrm{out}}/c_{\mathrm{in}}$ varies from 0 (q separate groups) to 1 (a purely random graph), we observe a continuous phase transition from an ordered phase with positive overlap to a paramagnetic phase with zero overlap. Thus there is a second-order transition from a detectable to an undetectable phase.

A third situation arises if $f_{\mathrm{BP}}\{\theta\}$ has both a paramagnetic fixed point *and* the ordered fixed point at the true $\{\theta\}$. In this case, the two phases coexist and the detectability transition is first-order; see Fig. 2 on the right. The phase transition is located by comparing the free energies of the two phases. However, even if the ordered fixed point has a lower free energy, it is not easy to find it unless the initial messages are close to the true group assignment. All but an exponentially small set of initial messages will lead to the paramagnetic fixed point. This situation is typical of mean-field first-order phase transitions. In fact, recent results about random optimization problems show that finding the lower-free-energy phase in this case is an extremely hard problem [12].

Only when the paramagnetic phase is no longer locally stable does inference become easy. We can compute the location of the transition to this easily detectable phase analytically by analyzing how a small random perturbation to the paramagnetic fixed point propagates as the BP equations are iterated [12,21]. It follows that for

$$|c_{\mathrm{in}} - c_{\mathrm{out}}| > q\sqrt{c}, \tag{4}$$

the true group assignment is dynamically attractive. In that case, many algorithms, e.g., MC or BP, will converge to it in linear time. But in the easy phase it may still be hard to compute the ground state of (2), even though we can compute the marginals, and therefore the optimal estimate of the group assignment, asymptotically exactly.

On the other hand, if (4) is not satisfied, then community detection is either impossible or, at best, as hard as solving the hardest known optimization problems. When $c_{\mathrm{out}} < c_{\mathrm{in}}$, the phase transition is of first order for $q > 4$, as can be retrieved from data presented in Ref. [21]. However, the detectable but hard region is so narrow that it is quite unlikely to appear in realistic situations.

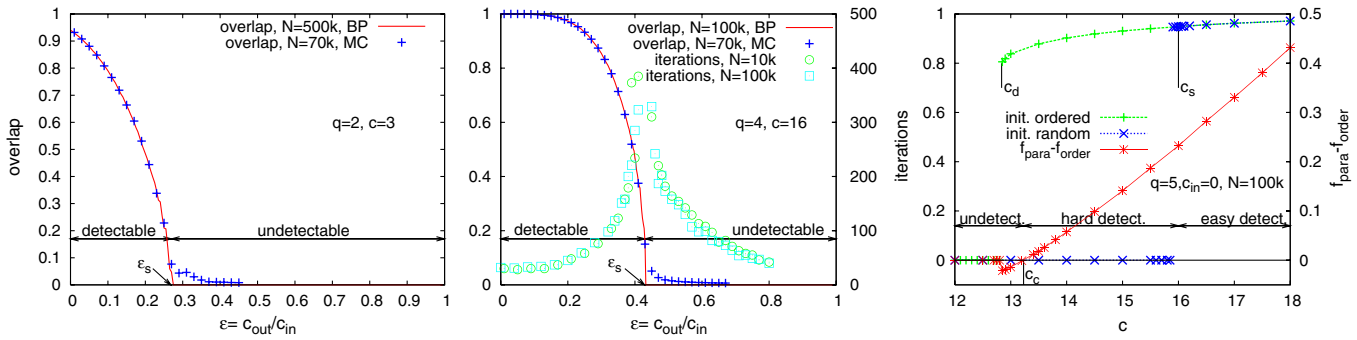

FIG. 2 (color online). The best possible overlap between the inferred and true group assignments. Left: Community detection with $q = 2$, $c = 3$, and different values of $\epsilon = c_{\text{out}}/c_{\text{in}}$. A continuous phase transition between a detectable and a nondetectable phase arises at the critical point ϵ_s given by (4). Middle: The 4-group community detection benchmark of Ref. [9] with $c = 16$, with the same phenomenology. The number of BP iterations needed for convergence is a constant with respect to N and diverges at the critical point. The results agree well with MC simulations, except very close to the critical point where finite-size effects are stronger. Right: A planted coloring problem with $q = 5$ and $c_{\text{in}} = 0$, $c = c_{\text{out}}(1 - 1/q)$. Both the ordered fixed point (green +'s, obtained by initializing in the actual group assignment) and the paramagnetic one (blue ×'s, obtained by initializing the algorithm in a random configuration) exist between c_d and c_s. The difference Δf (red) between the paramagnetic and ordered free energies shows that modules are, in principle, detectable as soon as $c > c_c$ when $\Delta f > 0$. In practice, it is exponentially hard to find the corresponding fixed point, and detection becomes feasible only after the phase transition point c_s given by (4).

Real-world networks.—Our algorithm is not restricted to large random networks; it is applicable to real-world networks as well, including those which have small loops rather than being locally treelike. As a proof of concept, we illustrate this on the "karate club" network [22], a common benchmark for community detection. For $q = 2$, BP leads to two different fixed points. One corresponds to the actual known division into two groups. The other has a smaller free energy (thus a larger likelihood) and splits the network into high-degree nodes and low-degree nodes as found in Ref. [3]. These two fixed points correspond to two local minima of f_{BP} for $q = 2$, and depending on the initial value $\{\theta_0\}$ BP converges to one or the other.

On a network of this size, MC quickly reaches equilibrium; we found that MC gives results almost identical to those of BP for the parameters and marginals and identical in terms of the estimated group assignments.

Conclusion.—We have presented an asymptotically exact analysis of the detection of communities in networks generated by the stochastic block model. We found that there is a strict limit on detectability due to a sharp phase transition. In some cases the communities are detectable, but the problem is exponentially hard because the attractive region around the correct fixed point is exponentially small. We have also presented a learning algorithm, which for large sparse networks generated from the model is able to infer the number of groups, their exact sizes, and the affinity matrix c_{ab}. The resulting BP algorithm is also applicable to real-world networks, and it is not restricted to assortative modular structures.

We note that, for many real networks, the stochastic block model is not a good fit to the network's structure. However, our BP algorithm can be generalized to other generative models where the likelihood is a product of local terms, e.g., the degree-corrected block model of Ref. [3].

We are grateful to Mark Newman for useful discussions on BP and expectation-maximization algorithms. C. M. is funded by the McDonnell Foundation.

*Corresponding author.
lenka.zdeborova@cea.fr

[1] B. H. Good, Y.-A. de Montjoye, and A. Clauset, Phys. Rev. E **81**, 046106 (2010).
[2] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[3] B. Karrer and M. E. J. Newman, Phys. Rev. E **83**, 016107 (2011).
[4] M. Mézard and G. Parisi, Eur. Phys. J. B **20**, 217 (2001).
[5] M. Mézard and A. Montanari, *Physics, Information, Computation* (Oxford University, New York, 2009).
[6] J. Yedidia, W. Freeman, and Y. Weiss, *Exploring Artificial Intelligence in the New Millennium* (Morgan Kaufmann, San Francisco, CA, 2003), pp. 239–236.
[7] J. Reichardt and M. Leone, Phys. Rev. Lett. **101**, 078701 (2008).
[8] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[9] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[10] M. B. Hastings, Phys. Rev. E **74**, 035102 (2006).
[11] J. M. Hofman and C. H. Wiggins, Phys. Rev. Lett. **100**, 258701 (2008).
[12] F. Krzakala and L. Zdeborová, Phys. Rev. Lett. **102**, 238701 (2009).
[13] M. E. J. Newman and E. A. Leicht, Proc. Natl. Acad. Sci. U.S.A. **104**, 9564 (2007).

[14] B. Ball, B. Karrer, and M. E. J. Newman, arXiv:1104.3590.
[15] C. Moore *et al.*, in *Proceedings of the Seventeenth ACM Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, 2011).
[16] P. J. Bickel and A. Chen, Proc. Natl. Acad. Sci. U.S.A. **106**, 21 068 (2009).
[17] D. Coppersmith, D. Gamarnik, M. T. Hajiaghayi, and G. B. Sorkin, Random Struct. Algorithms **24**, 502 (2004).
[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.107.065701 for more detailed derivations.
[19] A. Dempster, N. Laird, and D. Rubin, J. R. Stat. Soc. **39**, 138 (1977).
[20] D. Achlioptas and A. Coja-Oghlan, in *Proceedings of the 49th Annual Symposium on Foundations of Computer Science* (IEEE, New York, 2008).
[21] M. Mézard and A. Montanari, J. Stat. Phys. **124**, 1317 (2006).
[22] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).