Phase transitions in community detection: A solvable toy model

Greg Ver Steeg\(^1\), Cristopher Moore\(^2\), Aram Galstyan\(^1\) and Armen Allahverdyan\(^3\)

\(^1\) USC Information Sciences Institute - Marina del Rey, CA, USA
\(^2\) Santa Fe Institute - Santa Fe, NM, USA
\(^3\) Yerevan Physics Institute - Yerevan, Armenia

received 11 December 2013; accepted in final form 25 April 2014
published online 21 May 2014

PACS 89.75.Hc – Networks and genealogical trees
PACS 02.50.Tt – Inference methods
PACS 64.60.aq – Networks

Abstract – Recently, it was shown that there is a phase transition in the community detection problem. This transition was first computed using the cavity method, and has been proved rigorously in the case of \(q = 2\) groups. However, analytic calculations using the cavity method are challenging since they require us to understand probability distributions of messages. We study analogous transitions in the so-called “zero-temperature inference” model, where this distribution is supported only on the most likely messages. Furthermore, whenever several messages are equally likely, we break the tie by choosing among them with equal probability, corresponding to an infinitesimal random external field. While the resulting analysis overestimates the thresholds, it reproduces some of the qualitative features of the system. It predicts a first-order detectability transition whenever \(q > 2\) (as opposed to \(q > 4\) according to the finite-temperature cavity method). It also has a regime analogous to the “hard but detectable” phase, where the community structure can be recovered, but only when the initial messages are sufficiently accurate. Finally, we study a semisupervised setting where we are given the correct labels for a fraction \(\rho\) of the nodes. For \(q > 2\), we find a regime where the accuracy jumps discontinuously at a critical value of \(\rho\).

Introduction. – A number of recent papers have studied fundamental limits on community detection in the stochastic block model (SBM), a simple generative model of networks with tunable modularity. For networks that are dense enough, with an average degree that grows faster than \(\log n\), the communities can be recovered exactly under some circumstances [1]. However, in the sparse case where the average degree is \(O(1)\), there is a sharp transition below which the communities are undetectable [2–8]. The location of this transition was found using the cavity method [4,5], or equivalently, by analyzing the behavior of the belief propagation (BP) algorithm. It was also hypothesized that BP is an optimal inference method for community detection in SBM, so that the corresponding detectability threshold is algorithm independent [4,5]. This hypothesis was proved rigorously in the case of \(q = 2\) groups [9–11]; in the detectable regime, a polynomial-time algorithm exists that labels nodes correctly with probability bounded above \(1/2\), while in the undetectable regime, graphs generated by the SBM are indistinguishable from Erdős-Rényi random graphs, and no algorithm can label the nodes better than chance.

Below the transition, for \(q \leq 4\) belief propagation converges to a paramagnetic fixed point where every node is equally likely to belong to either community. For \(q > 4\), however, cavity method calculations [4,5] show that the situation is more complicated, including a “hard but detectable” regime where the communities can be recovered by belief propagation, but only if the algorithm is given a strong initial hint about the correct labels. However, addressing this claim analytically is difficult, given that the cavity method requires us to keep track of an entire probability distribution of messages.

Here we study the community detection problem within the zero-temperature Bethe-Peierls approximation [12–14]. Equivalently, we study a message-passing algorithm where the distribution of messages is concentrated on the most likely label of each node. Zero-temperature inference for community detection was also studied in [2]. However, we augment this algorithm with a tie-breaking mechanism: Whenever a node has several equally likely choices for its label, we break the symmetry randomly and uniformly among these choices.

Physically, tie-breaking corresponds to applying an infinitesimal random external field to each node. In terms
of inference, it corresponds to a non-uniform prior on the node labels. However, our main motivation for it is simply to gain tractability: it reduces the number of order parameters considerably, making it possible to study it analytically for any value of \( q \). The corresponding algorithm should be thought of as a “toy model” for the real community detection problem. It is far from optimal, and the thresholds we obtain are algorithm-specific, overestimating the true detectability thresholds (especially when the number of groups is large). Nevertheless, it reproduces some qualitative features of the real system, and allows insight into the interesting but analytically difficult regime where the graph contains many communities.

**Our contributions.** – We derive a fixed-point equation for the proposed message-passing algorithm, and use it to analytically explore the community detection problem. In this model, we find that the detectability transition is continuous for \( q = 2 \), and discontinuous for \( q > 2 \); in contrast, the finite-temperature cavity method [4,5] shows that the transition is continuous for \( q \leq 4 \) and discontinuous for \( q > 4 \) (in the assortative case). Analogous to the “hard but detectable” regime [4,5], we also find a regime in which there are two fixed points; a paramagnetic one where all labels are equally likely, and a second one which has high accuracy. In this regime, the algorithm is able to recover the underlying community structure only if the initial messages are sufficiently close to the true labels; otherwise, it converges to the paramagnetic solution.

We also analyze the SBM reconstruction problem in “semisupervised” settings where one is given the true group labels for a fraction \( \rho \) of nodes. For \( q = 2 \), even a tiny amount of prior information suppresses the detectability transition in the zero-temperature model [3]. Here we show that for \( q > 2 \), the behavior of the inference problem is much richer. Namely, while the prior information always improves the accuracy, there is a line of discontinuities where the accuracy jumps discontinuously at a critical value of \( \rho \), again in qualitative agreement with the cavity method for \( q \) sufficiently large [15].

**Stochastic block models.** – Consider a network of \( N \) nodes, where each node belongs to one of the \( q \) groups or communities. Let \( s_i \in \{1, \ldots, q\} \) be the community label of node \( i \), and let \( s = (s_i)_{i=1}^N \) be the probability of a link between two nodes \( i,j \) in groups \( k = s_i \) and \( \ell = s_j \) is given by a \( q \times q \) matrix \( p_{k\ell} \). We focus on the case in which \( p \) depends only on whether the nodes are in the same or different groups: that is, \( p_{kk} = p_{ii} \) if \( k = \ell \) and \( p_{k\ell} = p_{\text{out}} \) if \( k \neq \ell \). We assume that the network is assortative, i.e., that \( p_{\text{in}} > p_{\text{out}} \). Finally, we assume that it is sparse, i.e., that \( p_{\text{in}}, p_{\text{out}} = O(1/N) \).

Let \( A \) be the adjacency matrix of a graph generated by the above block model. The generative model is fully described by the following joint probability:

\[
p(A, s) = \pi(s)p(A|s) = \pi(s) \prod_{i<j} p_{s_is_j}^{A_{ij}} (1 - p_{s_is_j})^{1 - A_{ij}},
\]

where \( \pi(s) \) encodes prior information about the community assignment one might have.

Given the observed network \( A \), we are interested in reconstructing the unknown state \( s \). Toward this goal, we define the posterior probability of \( s \) given \( A \), which, by the use of Bayes theorem, can be written as follows:

\[
p(s|A) = \frac{\pi(s)p(A|s)}{p(A)},
\]

If the prior \( p(A) \) is constant, this gives a Gibbs distribution \( p(s|A) \propto \pi(s)p(A|s) \) at unit temperature.

There are several approaches for deciding the community assignments from the Gibbs distribution, and different approaches are optimal for different loss functions [16]. For instance, the fraction of correctly inferred labels is maximized by computing marginal probabilities for each node, \( p(s_i|A) \), and choosing the most likely label for each of them. Here we focus on a different approach known as maximum a posteriori estimation that tries to find the state \( s \) that jointly maximizes \( p(s|A) \). This is the ground state of a generalized Potts Hamiltonian,

\[
\mathcal{H} = -\sum_{i,j} A_{ij} \delta(s_i, s_j) + H_\pi(s),
\]

where the second term represents prior knowledge about the community assignments.

Since exact minimization is computationally intractable for large graphs, one has to resort to approximate methods. A popular family of such methods are message-passing algorithms such as belief propagation (BP). BP is exact on trees; although there are no convergence guarantees for general graphs with loops, for locally tree-like graphs, such as those generated by the SBM, we expect BP to be asymptotically correct in the thermodynamic limit, at least in some regimes [4,5].

If we want to find the ground state rather than the marginals, however, it makes sense to consider a zero-temperature version of BP, where the messages are concentrated on the most likely labels. We describe this algorithm, and our simplification of it, in the next section.

**Zero-temperature message passing.** – In the zero-temperature form of belief propagation, also known as the max-product algorithm, the messages have a particular simple form. Namely, they are binary vectors \( u = (u_1, u_2, \ldots, u_q) \), where \( u_k \in \{0, 1\} \) for all \( 1 \leq k \leq q \) and at least one of the \( u_k \) is positive. The message \( u_{i \rightarrow j} \) (also referred to as cavity bias) from node \( i \) to node \( j \) describes the preferred state of node \( i \) in the absence of node \( j \). To calculate this message, node \( i \) sums the messages from all its neighboring nodes except \( j \), obtaining the cavity field \( h_{i \rightarrow j} = \sum_{k \neq j} A_{ik} u_k \). It then constructs a new message \( u_{i \rightarrow j}^{\text{new}} = \hat{u}(h_{i \rightarrow j}) \), where the function \( \hat{u}(h) \) picks the maximum component of its argument and sets it to 1, while setting all the other components to zero: \( \hat{u}_k(h_{i \rightarrow j}) = \delta(h_k, \max_k h_k) \). Thus, \( u_k(h_{i \rightarrow j}) = 1 \) if \( k \) is one of
Phase transitions in community detection: A solvable toy model

the most likely groups for $i$ to belong to, given the most likely group memberships of its neighbors other than $j$.

There are $2^q - 1$ possible messages. Furthermore, the probability of a particular message depends on the true label of the node it originates from, giving $q(2^q - 1)$ probabilities. However, due to symmetry, one can show that there are only $2q - 1$ relevant order parameters [2]. Namely, what matters is (a) whether $u_k = 1$, where $k$ is the correct label $s_i$, and (b) the number of other non-zero entries of $u$. Thus, the cavity field distributions can be parameterized as $n_{c,w}$, where $\ell = u_{s_i} \in \{0,1\}$, and $w = ||u||^2 = \ell \in \{0,\ldots,q-1\}$.

The fixed point of the message-passing procedure can be found by solving the so-called cavity equation. For the SBM, this equation seems to have a closed-form solution only in simple cases, such as $q = 2$ [2,3,6]. In general, one has to resort to numerical methods such as population dynamics. Here one considers a pool of messages that are dynamically updated according to the rules specified above, while choosing the number of neighbors a node has in each group from the appropriate degree distribution. In essence, this simulates the message-passing algorithm within the annealed approximation, where the network is redrawn at each iteration.

Here we modify the message-passing scheme by only allowing messages where exactly one of $u$’s components is 1. In our update, if the procedure above gives a message $u$ with more than one non-zero component, we break the tie by choosing one of those components with equal probability. In that case, by symmetry and normalization, the only relevant order parameter is $m = \eta_{i,0} - \eta_{i,1}$, which we can think of as a magnetization.

Analysis for $q = 2$. Below we use $\alpha = \nu_\alpha N/q$ and $\gamma = \nu_{\text{out}} N/q$ to denote the average number of neighbors a node has in its own group and in each other group respectively. The total connectivity, or average degree, is then $c = \alpha + (q-1)\gamma$. We write $\delta = \alpha - \gamma$ as a measure of the strength of the community structure.

We start by analyzing zero-temperature message passing in the case $q = 2$. There are three order parameters, which we denote $\eta_+ = \eta_{i,0} - \eta_{i,1}$, and $\eta_0 = \eta_{i,1}$, corresponding to correct, incorrect, and non-informative messages, respectively. The update rule is a majority vote, with $\eta_0$ corresponding to a tie. Specifically, let $k_1$ be the number of correct messages $i$ receives from neighbors in its own group, plus the number of incorrect messages it receives from the other group; and let $k_2$ be the number of incorrect messages it receives from its own group, plus the number of correct messages from the other group. If $k = k_1 - k_2$, then $i$’s message is correct, incorrect, or uninformative if $k > 0$, $k < 0$, or $k = 0$, respectively.

In networks generated by the stochastic block model, the number of neighbors a node has in its own group or in the other group are Poisson-distributed with mean $\alpha$ and $\gamma$, respectively (note that we can generalize this to other degree distributions). As in the cavity method, we assume that the messages sent by these neighbors are independent. Thus, $k_1$ and $k_2$ are Poisson-distributed with mean $\lambda_1$ and $\lambda_2$, respectively, where

$$\lambda_1 = \alpha \eta_+ + \gamma \eta_- , \quad \lambda_2 = \alpha \eta_- + \gamma \eta_+ .$$

Their difference $k = k_1 - k_2$ is then distributed according to the Skellam distribution:

$$P(k) = e^{-(\lambda_1 + \lambda_2)} \left( \frac{\lambda_1}{\lambda_2} \right)^{\frac{k}{2}} I_{|k|} \left( 2\sqrt{\lambda_1 \lambda_2} \right) ,$$

where $I_k(z)$ is the modified Bessel function of the first kind. Without tie-breaking, the fixed-point equations of population dynamics are thus

$$\eta_+ = \sum_{k>0} P(k), \quad \eta_- = \sum_{k<0} P(k).$$

If we define

$$m = \eta_+ - \eta_- , \quad \tilde{q} = \eta_+ + \eta_- \equiv 1 - \eta_0 ,$$

then $m$ and $\tilde{q}$ are the magnetization and the Edwards-Anderson parameter, respectively. The fixed-point equations can be rewritten in terms of these parameters,

$$1 - \tilde{q} = e^{-c\tilde{q}} I_0(x),$$

$$m = e^{-c\tilde{q}} \sum_{k=1}^\infty \left[ \left( \frac{\lambda_1}{\lambda_2} \right)^\frac{k}{2} - \left( \frac{\lambda_1}{\lambda_2} \right)^{-\frac{k}{2}} \right] I_k(x)$$

$$= 2e^{-c\tilde{q}} \sum_{k=1}^\infty I_k(x) \sinh ky ,$$

where we have defined

$$x \equiv 2\sqrt{\lambda_1 \lambda_2} = \sqrt{\epsilon^2 \tilde{q}^2 - \delta^2 m^2} ,$$

$$y \equiv \tanh \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} = \tanh \frac{\delta m}{c\tilde{q}} .$$

These equations were obtained via zero-temperature cavity methods [3], but this derivation is considerably simpler. In the vicinity of the second-order phase transition we linearize (9) around $m = 0$ to obtain

$$m = 2e^{-c\tilde{q}} \delta m \sum_{k=1}^\infty \frac{k I_k(c\tilde{q})}{c\tilde{q}} .$$

Using the identity $2kI_k(z)/z = I_{k-1}(z) - I_{k+1}(z)$, the sum in (12) telescopes, giving an equation for the detection threshold,

$$\delta_c = \frac{e^{c\tilde{q}}}{I_0(c\tilde{q}) + I_1(c\tilde{q})} ,$$

where $\tilde{q}$ is given by taking $m \to 0$ in (8),

$$1 - \tilde{q} = e^{-c\tilde{q}} I_0(c\tilde{q}) .$$

We now consider the case with tie-breaking, flipping a coin whenever $k = 0$. In that case (6) becomes,

$$\eta_+ = \sum_{k>0} P(k) + \frac{1}{2} P(0) .$$

48004-p3
and $\eta_+ + \eta_- = 1$. After some manipulation, we obtain

$$m = 2e^{-c} \sum_{k=1}^{\infty} I_k(x) \sinh ky$$

(16)

but where now $x$, $y$ are defined by

$$x \equiv \sqrt{c^2 - \delta^2 m^2}, \quad y \equiv \frac{\delta m}{c}.$$  

(17)

Reasoning as before, we obtain for the threshold

$$\delta_c = \frac{e^c}{I_0(c) + I_1(c)}.$$  

(18)

Comparing (13) and (18) we see that tie-breaking is equivalent to setting the Edwards-Anderson parameter $\tilde{q}$ to 1. In fig. 1 we show both thresholds as a function of $c$. The threshold with tie-breaking is higher, showing that it can be helpful to report ties rather than break them; this behavior is similar to distributed algorithms for approximate majority [17]. In fact, the tie-breaking algorithm fails to find communities even when $\gamma = 0$, i.e., where all links are within groups, if $c < 1.849$, since at this point $\delta_c = c$. In contrast, without tie-breaking $\delta_c < c$ for all $c > 1$.

Neither version of zero-temperature inference performs as well as belief propagation. In particular, the detectability thresholds for both methods are noticeably higher than the algorithm-independent threshold predicted by [4,5] and established rigorously in [9–11]. Note, however, that those thresholds do scale correctly for large $c$: both (13) and (18) approach $\delta_c = \sqrt{\pi/2}c$ as $c \to \infty$, while the true threshold [4,5] is $\delta_c = \sqrt{c}$.

**Analysis for arbitrary $q$.** – We now consider the case with tie-breaking for arbitrary $q$. By symmetry, we again have just two types of messages: correct ones with density $\eta_+$, and incorrect ones with density $\eta_- = 1 - \eta_+$. Incorrect messages are spread uniformly over the $q - 1$ incorrect groups.

Let $k_0$ denote the number of messages a node $i$ receives carrying its own group label. These are either correct messages from neighbors in its group, or a $1/(q - 1)$ fraction of the incorrect messages from other groups. The expected total number of neighbors $i$ has in other groups is $(q - 1)\gamma$, so $k_0$ is Poisson-distributed with mean

$$\lambda_1 = \alpha \eta_+ + \gamma \eta_- = \gamma + \delta \eta_+.$$  

(19)

For each of the other groups, we label $\ell \in \{1, \ldots, q - 1\}$, let $k_\ell$ be the number of messages $i$ receives with label $\ell$. Then $k_\ell$ is Poisson-distributed with mean

$$\lambda_2 = \gamma \eta_+ + \frac{(\alpha + (q - 2)\gamma)\eta_-}{q - 1} = \frac{(c - \gamma) - \delta \eta_+}{q - 1}.$$  

(20)

The population dynamics then works as follows. Let $k = \max\{k_0, k_1, \ldots, k_{q-1}\}$ and let $n = \sum_{\ell=1}^{q-1} \delta(k_\ell, k)$ be the number of incorrect colors that achieve this maximum. Then $i$ emits a correct message with probability $1/(n + 1)$ if $k_0 = k$, and an incorrect message otherwise.

The joint probability that $n$ incorrect colors have $k_\ell = k$ and that the other $q - 1 - n$ have $k_\ell < k$ is

$$\tilde{p}(k, n) = \binom{q - 1}{n} P_{\lambda_1}(k)^n Q_{\lambda_1}(k)^{q-1-n},$$  

(21)

where $P_{\lambda}(k) = e^{-\lambda} \lambda^k / k!$ is the Poisson distribution with mean $\lambda$, and $Q_{\lambda}(k) = \sum_{j<k} P_{\lambda}(j)$ is the regularized Gamma function. The fixed-point equation is then

$$\eta_+ = g(\eta_+),$$

where

$$g(\eta) = \sum_{k=0}^{\infty} P_{\lambda_1}(k) \sum_{n=0}^{q-1} \tilde{p}(k, n) n + 1.$$  

(22)

We illustrate the fixed-point equation in fig. 2 for $q = 10$ and $c = 10$. A close inspection reveals that there are several different phases separated by two phase transitions, which we denote $\delta_1^{(1)}$ and $\delta_1^{(2)}$. For $\delta < \delta_1^{(1)}$, there is a single fixed point $\eta_+ = 1/q$ corresponding to the paramagnetic solution. At $\delta_1^{(1)}$ a second solution $\eta_2 > 1/q$ emerges, giving an accurate labeling of the nodes. This occurs when $g(\eta_2) = 1/q$ and $\frac{dg}{d\eta_2} |_{\eta_2} = 1$. Similarly, $\delta_1^{(2)}$ is defined by $\frac{dg}{d\eta_2} |_{\eta_2} = 1$. For $q > 2$ this transition is first order; that is, $\eta_2$ is bounded above $1/q$. In fact, the detectability transition (in the assortative case) is continuous for $q < 4$ and first order for $q > 4$ [4,5], but the zero-temperature model does give some intuition about why it becomes discontinuous at larger $q$.

The population dynamics can be described, in a suitable time scale, as $d\eta_+ / dt = g(\eta_+) - \eta_+$. Therefore, for $\delta_1^{(1)} < \delta < \delta_1^{(2)}$, both $1/q$ and $\eta_2$ are locally stable, with an unstable fixed point between them. At $\delta_1^{(2)}$, the paramagnetic solution $1/q$ becomes unstable.

These results fit qualitatively with the results from the cavity method in [4,5], albeit with overestimated values.
accurate enough to converge to the effect of initial messages that are random (solid line) and instability of the paramagnetic solution, respectively. We show and different values of \( \eta \) and \( \rho \) to the appearance of the second solution as a function of \( q \).

\[ \text{Fig. 2: (Color online) Fixed-point equation for } q=10, c=10, \text{ and different values of } \delta. \text{ From bottom to top, } \delta < \delta_c^{(1)}, \delta = \delta_c^{(1)}, \delta_c^{(1)} < \delta < \delta_c^{(2)}, \text{ and } \delta = \delta_c^{(2)}. \text{ Observe that at } \delta_c^{(1)}, \text{ the second solution emerges discontinuously, indicating a first-order transition.} \]

of \( \delta_c^{(1)} \) and \( \delta_c^{(2)} \). If \( \delta < \delta_c^{(1)} \), the communities are undetectable, since the algorithm converges to the paramagnetic fixed point. If \( \delta > \delta_c^{(2)} \), the communities are easy to detect, since a small perturbation away from the paramagnetic fixed point will lead to \( \eta_2 \). Finally, \( \delta_c^{(1)} < \delta < \delta_c^{(2)} \) is the “hard but detectable” regime: we can converge to \( \eta_2 \) and label the nodes accurately, but only if the initial messages are accurate enough.

In fig. 3 we plot the accuracy \( \eta_+ \) as a function of \( \delta \). The two curves correspond to different ways to initialize the messages; randomly (solid line) and accurately enough to converge to \( \eta_2 \) (dashed line). Compare with fig. 2(c) in [4].

\[ \text{Fig. 3: (Color online) The accuracy } \eta_+ \text{ as a function of } \delta \text{ for } q=10 \text{ and } c=20. \text{ The thresholds } \delta_c^{(1)} \text{ and } \delta_c^{(2)} \text{ correspond to the appearance of the second solution } \eta_2 > 1/q \text{ and the instability of the paramagnetic solution, respectively. We show the effect of initial messages that are random (solid line) and accurate enough to converge to } \eta_2 \text{ (dashed line). Compare with fig. 2(c) in [4].} \]

\[ \text{Fig. 4: (Color online) Zero-temperature thresholds } \delta_c^{(1)} \text{ (blue) and } \delta_c^{(2)} \text{ (green) as a function of the number of groups } q \text{ for } c=30. \text{ The red line shows the true hard/easy threshold } \sqrt{c}. \text{ It appears that } \delta_c^{(2)} \text{ increases with } q. \text{ This suggests that, when starting from random messages, zero-temperature inference with tie-breaking performs poorly when } q \text{ is large.} \]

Semisupervised inference. – So far, we have assumed that the only information available to us is the graph structure. We now focus on “semisupervised” inference, where we are also given some prior information about the true group assignment.

One can distinguish two possible scenarios. In the first, we have noisy information about every node, biasing us toward its correct label. Physically, this introduces a local field at each node, giving us a helpful hint. We can represent this by giving the correct label some weight \( \beta > 1 \) in the tie-breaking rule, so that the probability \( 1/(n+1) \) of a correct message in (22) becomes \( \beta/(n+1) \).

In the second scenario, our information is perfectly accurate, but limited: we know the true labels of a fraction \( \rho \) of the nodes [3]. We found that these two scenarios produce qualitatively similar results, and we focus on the latter.

For \( q=2 \), even a tiny amount of prior information suppresses the detectability transition [3]. The intuition behind this qualitative change is that local fields act as stabilizers during the inference process. We expect that prior information should facilitate community detection for \( q > 2 \) as well. However, the situation is more complex.

To proceed further, we define \( \eta_+ \) as the accuracy we achieve on the unknown nodes. In that case, we can modify our previous analysis by assuming that a fraction \( \rho \) of the incoming messages is from known nodes, and is automatically correct. Thus, we still solve the fixed-point equation \( \eta_+ = g(\eta_+) \), where \( g(\eta_+) \) is defined as before, but we now replace \( \lambda_1 \) and \( \lambda_2 \) in (21) and (22) with

\[ \lambda_1(\rho) = \rho \alpha + (1-\rho) \lambda_1, \quad \lambda_2(\rho) = \rho \gamma + (1-\rho) \lambda_2. \quad (23) \]

Figure 5(a) shows the accuracy \( \eta_+ \) as a function of \( \delta \) for different amounts of prior information, \( \rho = 0, 0.005, 0.05 \). We see that even a small value of \( \rho \) lets us jump to an accurate fixed point analogous to \( \eta_2 \) at some \( \delta < \delta_c^{(2)} \), letting us label the nodes even when we are some distance...
inside the “hard but detectable” regime. Note that this is in stark contrast with the behavior reported in ref. [3], where the detection threshold disappeared for any \( \rho > 0 \).

As shown in fig. 5(b), there is a range of \( \delta \) where the accuracy jumps discontinuously at a critical value of \( \rho \). These discontinuities disappear at a particular value of \( \delta \), corresponding to a tricritical point. Below this \( \delta \) the accuracy increases steeply, but continuously, as a function of \( \rho \). This qualitatively reproduces the picture of cavity method calculations for large \( q \) [15].

Discussion. – We analyzed community detection in the stochastic block model, based on a zero-temperature message-passing algorithm. By breaking ties randomly, we reduced the number of order parameters to one, giving us an analytically tractable model for any number of groups. This corresponds to applying an infinitesimal random external field at each node, or a random non-uniform prior on the labels.

The randomized message-passing algorithm considered here is not optimal, and the thresholds reported here can only be viewed as bounds on the true (algorithm-independent) detection thresholds. Nevertheless, it lets us analytically reproduce some qualitative aspects of the system. For \( q > 2 \) it predicts a first-order detectability transition, and a “hard but detectable” regime. We note that the finite-temperature cavity method shows that this regime appears when \( q > 4 \) [4,5].

We also analyzed a “semisupervised” setting where one is given the true labels of a fraction \( \rho \) of nodes. In contrast to \( q = 2 \) [3], for \( q > 2 \) even a small value of \( \rho \) significantly moves the boundary between the hard and easy regimes, and there is a range of \( \delta \) where the accuracy jumps discontinuously as a function of \( \rho \). This is again in qualitative agreement with the cavity method [15].

We limited our analysis to the case where the groups are of equal size, and where the connectivity between nodes depends only on whether they are in the same group or not. Our approach can be extended to more general cases, although the analysis will be more complicated.

Fig. 5: (Color online) (a) Accuracy \( \eta_+ \) vs. \( \delta \) for \( q = 10 \) and \( c = 20 \), and for varying amounts of prior information, \( \rho = 0.0, 0.005, 0.05 \). At \( \rho = 0 \), we do no better than chance until \( \delta = \delta^{(q)} \), as in fig. 3. However, even a small \( \rho \) moves the boundary between the easy and hard regimes downward, letting us jump to an accurate fixed point. (b) Accuracy vs. \( \rho \), for the same parameters as in (a), and for three different values of \( \delta \). There is a range of \( \delta \) where the accuracy jumps discontinuously at a critical value of \( \rho \). At a critical value of \( \delta \), this discontinuity disappears.

AG and GVS thank the Santa Fe Institute for their hospitality. AG and GVS were supported in part by the US AFOSR MURI grant FA9550-10-1-0569, and US DTRA grant HDTRA1-10-1-0086. CM is supported by the AFOSR and DARPA under grant No. FA9550-12-1-0432. The authors thank Lenka Zdeborová, Pan Zhang, Elchanan Mossel, and Allan Sly for helpful conversations.