Scalable Detection of Statistically Significant Communities and Hierarchies: Message-Passing for Modularity

Pan Zhang
Cris Moore

SFI WORKING PAPER: 2014-04-009

SANTA FE INSTITUTE

Scalable detection of statistically significant communities and hierarchies: message-passing for modularity

Pan Zhang and Cristopher Moore
Santa Fe Institute, Santa Fe, New Mexico 87501, USA

Modularity is a popular measure of community structure. However, maximizing the modularity can lead to many competing partitions with almost the same modularity that are poorly correlated to each other; it can also overfit, producing illusory "communities" in random graphs where none exist. We address this problem by using the modularity as a Hamiltonian, and computing the marginals of the resulting Gibbs distribution. If we assign each node to its most-likely community under these marginals, we claim that, unlike the ground state, the resulting partition is a good measure of statistically-significant community structure.

We propose an efficient Belief Propagation (BP) algorithm to compute these marginals. In random networks with no true communities, the system has two phases as we vary the temperature: a paramagnetic phase where all marginals are equal, and a spin glass phase where BP fails to converge. In networks with real community structure, there is an additional *retrieval* phase where BP converges, and where the marginals are strongly correlated with the underlying communities. We show analytically and numerically that the proposed algorithm works all the way down to the detectability transition in networks generated by the stochastic block model. We also show that our algorithm performs well on real-world networks, revealing large communities in some networks where previous work has claimed no communities exist. Finally we show that by applying our algorithm recursively, subdividing communities until no statistically-significant subcommunities can be found, we can detect hierarchical structure in real-world networks more efficiently than previous methods. Our algorithm is highly scalable, working in time nearly linear in the number of edges: for networks with 10^5 nodes and 10^6 edges, for instance, it takes 14 seconds to find community structure.

I. INTRODUCTION

Community detection, or node clustering, is a key problem in network science, computer science, sociology, and biology. It aims to partition the nodes in a network into groups such that there are many edges connecting nodes within the same group, and comparatively few edges connecting nodes in different groups.

Many methods have been proposed for this problem. These include spectral clustering, where we cluster nodes according to eigenvectors of a linear operator such as the adjacency matrix, random walk matrix, graph Laplacian, or other linear operators [1–3]; statistical inference, where we fit the network with a generative model such as the stochastic block model [4–6]; and a wide variety of other methods, e.g. [7–9]. See [10] for a review.

We focus here on a popular measure of the quality of a partition, the modularity [7, 11–13]. We think of a partition $\{t\}$ into q groups as a function $t : V \to \{1, \dots, q\}$ where t_i is the group to which node i belongs. The modularity of a partition $\{t\}$ of a network with n nodes and m edges is defined as follows,

$$Q(\{t\}) = \frac{1}{m}\left(\sum_{\langle ij\rangle\in\mathcal{E}} \delta_{t_i t_j} - \sum_{\langle ij\rangle} \frac{d_i d_j}{2m}\delta_{t_i t_j}\right). \tag{1}$$

Here $\langle ij\rangle$ denotes an ordered pair of nodes, $\mathcal{E}$ is the set of edges, d_i is the degree of node i, and δ is the Kronecker delta function. The modularity is proportional to the number of edges connecting nodes in the same community, minus the expected number of such edges if the graph were random conditioned on its degree distribution; that is, the expectation in a null model where i and j are connected with probability proportional to $d_i d_j$.

However, maximizing over all possible partitions often gives a large modularity even in random graphs with no community structure. For instance, in a random 3-regular graph, there is typically a partition into two equal groups with only 11% of the edges crossing between the groups [14]. This is a case of overfitting, where the "optimal" partition simply reflects random noise [15–17]. Moreover, even in real-world networks, the modularity often exhibits a large amount of degeneracy, with multiple local optima that are poorly correlated with each other, and are not robust to small perturbations [18]. Thus we need some measure of robustness or statistical significance, to identify partitions that truly reflect the structure of the network.

A natural approach is to compare the modularity of a real-world network to the distribution of modularities we would see in a null model such as Erdős-Rényi graphs or the configuration model, obtaining a p-value that would allow us to reject the null hypothesis that no communities exist. However, even when there really are communities, the modularity of the corresponding partition is often no higher than that of random graphs. In Fig. 1, we show partitions

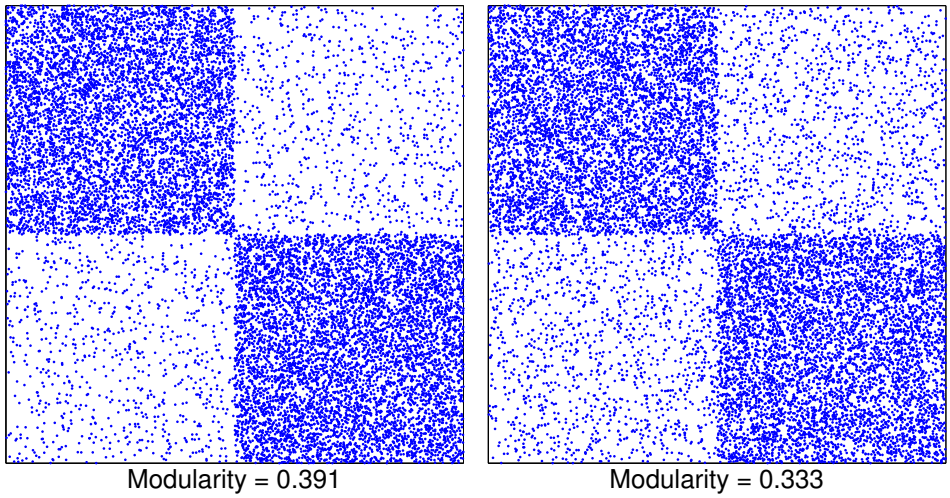

FIG. 1: The adjacency matrices of two networks, partitioned to show possible community structure. Each blue point is an edge. The network on the left is an Erdős-Rényi random graph, with no real community structure; however, a search by simulated annealing finds a partition with high modularity, 0.391. The network on the right has true communities, and is generated by the stochastic block model, however, the planted partition, i.e., the true underlying community structure, has modularity just 0.333. Thus illusory communities in random graphs can have higher modularity than true communities in structured graphs. Both networks have size $n = 5000$ and a Poisson degree distribution with mean $c = 3$; the network on the right has $c_{\text{out}}/c_{\text{in}} = 0.2$, in the easily-detectable regime of the stochastic block model.

of two networks with the same size and degree distribution: an Erdős-Rényi graph (left), and a graph generated by the stochastic block model (right), in the detectable regime where algorithms can easily find a partition correlated with the planted one [4, 5]. The (true) planted partition of the network on the right has a smaller modularity than the partition found for the random graph on the left. We can find a partition with higher modularity (and lower accuracy) on the right using e.g. simulated annealing, but then the modularities we obtain for the two networks are similar. Thus the usual approach of null distributions and p-values for hypothesis testing does not appear to work.

We propose to solve this problem with the tools of statistical physics. Like [15], we treat the modularity (or rather $-mQ$) as the Hamiltonian of a spin system. Rather than maximizing the modularity by searching for the ground state of this system, we focus on its Gibbs distribution at a finite temperature, looking for many high-modularity partitions rather than a single one. In analogy with previous work on the stochastic block model [4, 5], we define a marginalized state, or *retrieval state*, by computing the marginals of the Gibbs distribution, and assigning each node to its most-likely community. We claim that the resulting partition is a far better measure of statistically-significant community structure than the ground state, i.e., the maximum-modularity partition. That is, as in Bayesian inference of network structure (e.g. [19]) the consensus of many good solutions is better than the best single solution.

Computing marginals using Monte Carlo sampling is quite slow, since it requires many independent samples. Instead, we give an efficient Belief Propagation algorithm, which is equivalent to cavity method calculations. This algorithm is highly scalable; each iteration takes linear time on sparse networks, and it converges rapidly. Moreover, it is optimal in the sense that for synthetic graphs generated by the stochastic block model, it works all the way down to the detectability transition: that is, it detects communities whenever they can be detected. Finally, by applying this algorithm recursively, subdividing communities until no statistically significant subcommunities exist, we can uncover hierarchical structure in a way that is both efficient and mathematically principled.

We validate our approach with experiments on both real and synthetic networks. In particular, we find significant communities in some large networks where previous work claimed there were none. We also comment on connections between our method on Bayesian approaches based on generative models.

II. THE GIBBS DISTRIBUTION, BELIEF PROPAGATION, AND THE RETRIEVAL STATE

An Erdős-Rényi random graph has many partitions with high modularity. However, these partitions are nearly uncorrelated with each other. While the optimal partition might be unique, there are many others whose modularity is almost as high, but which have a large Hamming distance from the optimum and from each other. In statistical physics, we would say that the landscape of partitions is glassy, with many local optima that are far from each other.

This phenomenon has been observed in some real-world networks as well [18].

For the network on the right of Fig. 1, in contrast, we believe there is a real community not just because high-modularity partitions exist, but because there are many such partitions that are correlated with each other (and with the planted one). So to tell whether or not communities exist, we should ask not just what modularity we can obtain, but whether there are many highly-correlated partitions with high modularity. In the language of statistical physics, we want to find a set of correlated partitions with both low energy (high modularity) and high entropy. That is, we care about the free energy of the Gibbs distribution of partitions, rather than its ground state energy. We will argue that whenever the free energy is lower than that of a "paramagnetic" state where every node is equally likely to be in each community, then typical partitions drawn from the Gibbs distribution are in fact strongly correlated with each other. These partitions might have significantly lower modularity than than the "optimal" partition, i.e., the ground state, as well as different modularities from each other. Nevertheless, if we extract a partition that represents the consensus of the distribution, by assigning each node its most-likely community according to the marginals of the Gibbs distribution, it is a more accurate picture of the communities than the ground state.

The idea of using free energy to separate real community structure from random noise, and using marginals to extract a single partition from the distribution, was also explored in [4, 5]. In that work, the energy is (minus) the log-likelihood that the network is generated by the stochastic block model. The free energy was also used in [20] to perform model selection between the stochastic block model and its degree-corrected variant. However, in this paper we avoid generative models, and focus directly on the modularity. (On the other hand, as we discuss below there is a connection between the modularity and the log-likelihood of the degree-corrected stochastic block model.)

We define the energy of a partition $\{t\}$ as

$$E(\{t\}) = -mQ(\{t\}), \tag{2}$$

and introduce a Gibbs distribution as a function of inverse temperature β:

$$P(\{t\}) = \frac{1}{Z}\,\mathrm{e}^{-\beta E(\{t\})}. \tag{3}$$

The parameter β lets us tune the energy level that dominates the Gibbs measure. For example, when $\beta = 0$, every configuration has the same probability q^{-n}. When $\beta = \infty$, the only partitions with nonzero probability are those that maximize the modularity.

Our goal is to compute the free energy of the Gibbs distribution, and the marginal probability distribution that each node belongs to a given group. We could do this using a Monte Carlo Markov Chain algorithm. However, to obtain marginals we need many independent samples, and to obtain the free energy we would need to sample at many different temperatures. Thus MCMC is prohibitively slow for our purposes.

However, for sparse networks, we can use Belief Propagation [21], known in statistical physics as the cavity method [22]. BP makes a conditional independence assumption, which is exact only on trees; however, in the regimes we will consider (the detectable regime of the stochastic block model, and typical real-world graphs) its estimates of the marginals and the free energy are quite accurate.

BP works with "messages" $\psi_t^{i\to k}$: these are estimates, sent from node i to node k, of the marginal probability that $t_i = t$ based on i's interactions with nodes $j \neq k$. The update equations for these messages are as follows:

$$\psi_t^{i\to k} = \frac{1}{Z_{i\to k}} \prod_{j\in\partial i\setminus k} \sum_{s=1}^{q} \mathrm{e}^{\beta\delta_{st}} \psi_s^{j\to i} \prod_{j\neq i,k} \sum_{s=1}^{q} \mathrm{e}^{-\beta\frac{d_i d_j}{2m}\delta_{st}} \psi_s^{j\to i} \tag{4}$$

$$= \frac{1}{Z_{i\to k}} \prod_{j\in\partial i\setminus k} \left(1 + \psi_t^{j\to i}(\mathrm{e}^{\beta} - 1)\right) \prod_{j\neq i,k} \left(1 + \psi_t^{j\to i}(\mathrm{e}^{-\beta\frac{d_i d_j}{2m}} - 1)\right). \tag{5}$$

Here $Z_{i\to k}$ is simply a normalization factor, and ∂i denotes the neighborhood of node i. The BP estimate of the marginal probability that $t_i = t$ is then

$$\psi_t^{i} = \frac{1}{Z_{i}} \prod_{j\in\partial i} \sum_{s=1}^{q} \mathrm{e}^{\beta\delta_{st}} \psi_s^{j\to i} \prod_{j\neq i} \sum_{s=1}^{q} \mathrm{e}^{-\beta\frac{d_i d_j}{2m}\delta_{st}} \psi_s^{j\to i}$$

$$= \frac{1}{Z_{i}} \prod_{j\in\partial i} \left(1 + \psi_t^{j\to i}(\mathrm{e}^{\beta} - 1)\right) \prod_{j\neq i} \left(1 + \psi_t^{j\to i}(\mathrm{e}^{-\beta\frac{d_i d_j}{2m}} - 1)\right), \tag{6}$$

which is the same as (4) except that we remove the condition $j \neq k$. We can also estimate the two-point marginals, and in particular, the probability that two neighboring points belong to the same group. If $\langle ij\rangle \in \mathcal{E}$, the BP estimate

of the probability that $t_i = t$ and $t_j = s$ is

$$\psi_{st}^{ij} = \frac{1}{Z_{ij}} \mathrm{e}^{\beta\delta_{st}} \psi_s^{j\to i} \psi_t^{i\to j} \,.$$

The update equations (4) involve qn^2 messages: every node interacts with every other one, not just their neighbors. However, in the sparse case we can simplify the effect of non-neighbors, by replacing them with an external field as in [4, 5]. If $k \notin \partial i$ and $d_i, d_k \ll \sqrt{m}$, we have

$$\psi_t^i = \psi_t^{i\to k} \sum_s \mathrm{e}^{-\beta \frac{d_i d_k}{2m} \delta_{st}} \psi_s^{k\to i} \approx \psi_t^{i\to k} \left(1 - \beta \frac{d_i d_k}{2m} \psi_t^{k\to i}\right) \approx \psi_t^{i\to k} \,. \tag{7}$$

In that case, we can identify the messages $\psi_t^{i\to k}$ that i sends to its non-neighbors k with its marginal ψ_t^i. Then (4) simplifies to

$$\begin{aligned} \psi_t^{i\to k} &= \frac{1}{Z_{i\to k}} \prod_{j\in\partial i\setminus k} \left(1 + \psi_t^{j\to i}(\mathrm{e}^\beta - 1)\right) \prod_{j\neq i,k} \left(1 + \psi_t^j (\mathrm{e}^{-\beta \frac{d_i d_j}{2m}} - 1)\right) \\ &\approx \frac{1}{Z_{i\to k}} \exp\left(-\frac{\beta d_i}{2m}\theta_t + \sum_{j\in\partial i\setminus k} \log\left(1 + \psi_t^{j\to i}(\mathrm{e}^\beta - 1)\right)\right) \,, \end{aligned} \tag{8}$$

where

$$\theta_t = \sum_{j=1}^n d_j \psi_t^j \tag{9}$$

denotes an external field acting on nodes in group t, which we update after each BP iteration. Iterating (8) now has computational complexity qm, which is linear in the number of edges when q is fixed. Moreover, these updates can be easily parallelized. The number of iterations required to converge appears to depend very weakly on the network size (although in some cases it must grow at least logarithmically).

Observe that the *factorized* solution, $\psi_t^{j\to i} = 1/q$, where each node is equally likely to be in each possible group, is always a fixed point of the BP equations (8). If BP converges to this solution, we cannot label the nodes better than chance, and the retrieval modularity is zero. There are two other possibilities: BP fails to converge, or it converges to a non-factorized fixed point, which we call the *retrieval state*. In the latter case, we can compute the marginals and define a partition $\hat{t}$ that assigns each node to its most-likely community,

$$\hat{t}_i = \operatorname*{argmax}_{t_i} \psi_{t_i}^i, \,. \tag{10}$$

This partition represents the consensus of the Gibbs distribution: there are many high-modularity partitions that are correlated with it. We call it the *retrieval partition*, and call its modularity $Q(\{\hat{t}\})$ the *retrieval modularity*. In the spin glass phase, we are free to define a retrieval modularity from the current marginals, but it fluctuates as the algorithm jumps from one partition to another.

The *Bethe free energy* of a BP fixed point is a function of the messages:

$$f_{\mathrm{Bethe}} = -\frac{1}{n\beta} \left(\sum_i \log Z_i - \sum_{\langle ij\rangle\in\mathcal{E}} \log Z_{ij} + \frac{\beta}{4m} \sum_t \theta_t^2 \right) \,, \tag{11}$$

where Z_i and Z_{ij} are the normalization constants for the one- and two-point marginals. BP fixed points are stationary points of the Bethe free energy. Assuming it does not get stuck in a local minimum, BP converges to a retrieval state whenever its Bethe free energy is less than that of the factorized state. If the network has average degree c, this is simply

$$f_{\mathrm{Bethe}}^{\mathrm{fact}} \;=\; -\frac{1}{\beta} \left(\log q + \frac{c}{2} \log\left(1 - \frac{1}{q} + \frac{\mathrm{e}^\beta}{q}\right) - \frac{c\beta}{2q} \right) \,. \tag{12}$$

On the other hand, if BP does not converge, this means that neither the factorized solution nor any other fixed point is locally stable; the spin glass susceptibility [23] diverges, and replica symmetry is broken. In other words, the

space of partitions breaks into an exponential number of clusters, and BP jumps from one to another. The retrieval partition obtained using the current marginals will change to a very different partition if we run BP a bit longer, or if we perturb the initial BP messages slightly.

The linear stability of the factorized solution can be characterized by computing the derivatives of messages with respect to each other at the factorized fixed point. Using (8), we define a $q \times q$ matrix

$$T_{st} = \left. \frac{\partial \psi_t^{i \to k}}{\partial \psi_s^{j \to i}} \right|_{\frac{1}{q}} = \frac{\mathrm{e}^\beta - 1}{\mathrm{e}^\beta - 1 + q} \left(\delta_{st} - \frac{1}{q} \right) . \tag{13}$$

Its largest eigenvalue (in magnitude) is

$$\lambda = \frac{\mathrm{e}^\beta - 1}{\mathrm{e}^\beta - 1 + q} . \tag{14}$$

On locally tree-like graphs with Poisson degree distributions and average degree c, the factorized fixed point is then unstable with respect to random noise whenever $c\lambda^2 > 1$. This is also known as the de Almeida-Thouless local stability condition [24], the Kesten-Stigum bound [25, 26], or the threshold for census or robust reconstruction [27, 28]. In our case, it shows that the temperature must be below a certain threshold, or that β must exceed a critical β^*:

$$\beta > \beta^*(q,c) = \log \left(\frac{q}{\sqrt{c}-1} + 1 \right) . \tag{15}$$

If the degree distribution is not Poisson but the network is random in the configuration model, this expression holds where c is the average excess degree, i.e., the expected number of additional neighbors of the endpoint of a random edge. If a_d is the fraction of nodes with degree d, this is

$$c = \frac{\sum_d d(d-1)a_d}{\sum_d d a_d} = \frac{\langle d^2 \rangle}{\langle d \rangle} - 1 .$$

This suggests two scenarios. If there is no statistically significant community structure, then BP has just two phases, the paramagnetic one and the spin glass: for $\beta < \beta^*$ it converges to the factorized fixed point, and for $\beta > \beta^*$ it doesn't converge at all. On the other hand, if there are statistically significant communities, there is an additional *retrieval phase*, in the range $\beta_{\mathrm{R}} < \beta < \beta_{\mathrm{SG}}$, where BP converges to a retrieval state. Typically $\beta_{\mathrm{R}} < \beta^*$ and β^* is in the retrieval phase, since even if the factorized fixed point is locally stable, BP can still converge to a retrieval state if its free energy is lower than $f_{\mathrm{Bethe}}^{\mathrm{fact}}$. Thus we can test for statistically significant communities by running BP at $\beta = \beta^*$. (Note that our calculation of β^* in (15) assumes that the network is random conditioned on its degree distribution; in principle β^* could fall outside the retrieval phase in real-world networks, although we have not encountered this. One can always scan values of β in the vicinity of β^*.)

Our experiments in the next section, on both real and synthetic networks, bear this out. Moreover, as we will see below, we can choose the number of groups either by maximizing the retrieval modularity, or by recursively subdividing the network until no statistically significant subcommunities can be found.

There is one caveat to this picture. Namely, the system could be a "hard but detectable" phase, where there are statistically significant communities, but where they are hard to find. Specifically, the retrieval state has lower free energy than the factorized state, but it has an exponentially small basin of attraction. In this case, BP with random initial messages either converges to the factorized fixed point if $\beta < \beta^*$, or fails to converge. In the stochastic block model, this phase exists for $q \geq 5$; happily, in the assortative case where nodes are more likely to be connected to others in the same group (i.e., where the modularity is positive) the corresponding range of parameters is quite narrow, and we have found no evidence of this phase in our experiments on real-world networks.

We note that the paramagnetic, retrieval, and spin glass states were also studied in [29], using a generalized Potts model and a heat bath MCMC algorithm. However, their Hamiltonian depends on a tunable cut-size parameter, rather than on a general measure of community structure such as the modularity.

III. RESULTS ON THE STOCHASTIC BLOCK MODEL

In this section we apply our algorithm to synthetic networks generated by the stochastic block model (SBM). We confirm the existence of the retrieval phase, and argue that our algorithm is optimal in the sense that it succeeds all the way down to the Kesten-Stigum transition—that is, that it detects the communities whenever they can be

efficiently detected. In the process, we compute $\beta_{\rm R}$ analytically, locating the transition from the paramagnetic phase to the retrieval phase.

Also called the planted partition model, the SBM is a widely-used model for generating benchmark networks with community structure. There are q groups of nodes, and each node i has a group label $t_i^* \in \{1,\dots,q\}$; thus $\{t^*\}$ is the planted partition. Edges are generated independently according to a $q \times q$ matrix p, by connecting each pair of nodes $\langle ij \rangle$ with probability $p_{t_i^*,t_j^*}$. Here for simplicity we discuss the commonly studied case where the q groups have equal size and where p has only two distinct entries, $p_{rs} = c_{\rm in}/n$ if $r = s$ and $c_{\rm out}/n$ if $r \neq s$. We use $\epsilon = c_{\rm out}/c_{\rm in}$ to denote the ratio between these two entries. In the assortative case, $c_{\rm in} > c_{\rm out}$ and $\epsilon < 1$. When ϵ is small, the community structure is quite strong, and the modularity is large; when $\epsilon = 1$, the network becomes an Erdős-Rényi graph where ever pair of nodes is equally likely to be connected, and the modularity is zero in the limit $n \to \infty$.

For a given average degree $c = (c_{\rm in}+(q-1)c_{\rm out})/q$, there is a phase transition, called the detectability transition [4, 5], at a critical value

$$\epsilon^* = \frac{\sqrt{c}-1}{\sqrt{c}-1+q}\,. \tag{16}$$

For $\epsilon < \epsilon^*$, BP can label the nodes with high accuracy; for $\epsilon > \epsilon^*$, neither BP nor any other algorithm can label the nodes better than chance, and indeed no algorithm can distinguish the network from an Erdős-Rényi graph with high probability. This transition was recently established rigorously in the case $q = 2$ [30–32].

For larger number of groups, the situation is more complicated. For $q \leq 4$, in the assortative case, this detectability transition coincides with the Kesten-Stigum bound discussed in the previous section. For $q \geq 5$ the Kesten-Stigum bound marks a conjectured transition to a "hard but detectable" phase where community detection is still possible but takes exponential time, while the detectability transition is at a larger value of ϵ; that is, the thresholds for reconstruction and robust reconstruction become different. Our claim is that our algorithm succeeds down to the Kesten-Stigum bound, i.e., throughout the detectable regime for $q \leq 4$ and the easily detectable regime for $q \geq 5$.

In Fig. 2 we compare the behavior our algorithm on Erdős-Rényi graphs and a network generated by the SBM in the detectable regime. Both graphs have the same size and average degree $c = 3$. The ER graph is shown on the left. As described above, there are just two phases, separated by a transition at $\beta^* = 1.317$: the paramagnetic phase where BP converges to the factorized fixed point (with a convergence time that diverges as we approach the transition from below), and the spin glass phase where BP fails to converge. The retrieval modularity is zero or undefined, and there is no statistically significant community structure.

The SBM network shown on the right of Fig. 2 has a strong community structure with $\epsilon = 0.2$, well inside the detectable regime. In addition to the paramagnetic and spin glass phases, there is now a retrieval phase in a range of β. The retrieval modularity jumps sharply at $\beta_{\rm R} = 1.072$ when we first enter this phase, and then increases gently to 0.393 as β increases; for comparison, the modularity of the planted partition is $M_{\rm hidden}(\epsilon) = 1/(1+\epsilon) - 1/2 = 0.33$. When we enter the spin glass phase at $\beta_{\rm SG} = 2.27$, the retrieval modularity starts to fluctuate and then decreases. The BP convergence time diverges at both transitions.

In Fig. 3 we compare the free energy, convergence time, and retrieval modularity for networks generated by the stochastic block model at three different values of ϵ, alongside an Erdős-Rényi graph of the same average degree $c = 3$. For small enough β, their free energies are all equal to $f_{\rm Bethe}^{\rm fact}$, since they are all in the paramagnetic phase. For each value of ϵ, there is a critical $\beta_{\rm R}$ at which the free energy splits off from the others, where makes a transition to a retrieval state with $f_{\rm Bethe} < f_{\rm Bethe}^{\rm fact}$. The retrieval modularity jumps to a nonzero value, indicating community structure, and the convergence time diverges at the transition. For the Erdős-Rényi graph, the apparent modularity also jumps, but at $\beta^* = \beta_{\rm SG}$ it enters the spin glass phase rather than the retrieval phase: BP fails to converge and the retrieval modularity fluctuates, indicating partitions that are uncorrelated with each other.

To determine $\beta_{\rm R}$, and to argue that our algorithm succeeds all the way down to the Kesten-Stigum transition, we again consider the linear stability of BP around the factorized fixed point; but now we consider arbitrary perturbations, as opposed to random noise. Let T be the $q \times q$ matrix defined in (13). The matrix of derivatives of all $2qm$ messages with respect to each other is a tensor product $T \otimes B$, where B is the non-backtracking matrix [3]. The adaptive external field in the BP equations suppresses eigenvectors where every node is in the same community. As a result, the relevant eigenvalue is $\lambda\mu$ where λ is the largest eigenvalue of T, and μ is the second-largest eigenvalue of B, and the factorized fixed point is unstable whenever $\lambda\mu > 1$.

For networks generated by the SBM, we have [3]

$$\mu = \frac{c(1-\epsilon)}{1+(q-1)\epsilon}\,. \tag{17}$$

Combining this with (14) and setting $\lambda\mu = 1$ gives

$$\beta_{\rm R}(\epsilon) = \log\left(\frac{q(1+(q-1)\epsilon)}{c(1-\epsilon)-(1+(q-1)\epsilon)}+1\right)\,. \tag{18}$$

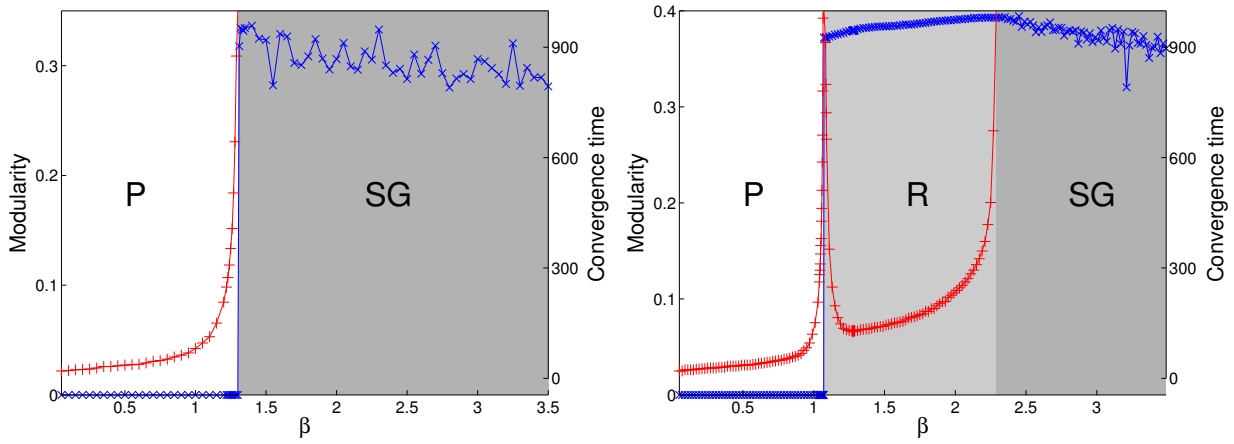

FIG. 2: Retrieval modularity (blue ×) and BP convergence time (red +) of an Erdős-Rényi random graph (left) and a network generated by the stochastic block model in the detectable regime. Both networks have $n = 1000$ and average degree $c = 3$, and the network on the right has $\epsilon = 0.2$. In both cases we ran BP with $q = 2$ groups. In the ER network, which has no community structure, there are two phases, paramagnetic (P) and spin glass (SG), with a transition at $\beta^* = 1.317$. In the SBM network with community structure, there is an additional retrieval phase (R) between $\beta_{\rm R} = 1.072$ and $\beta_{\rm SG} = 2.27$, where BP finds a partition $\{\hat{t}\}$ with high retrieval modularity.

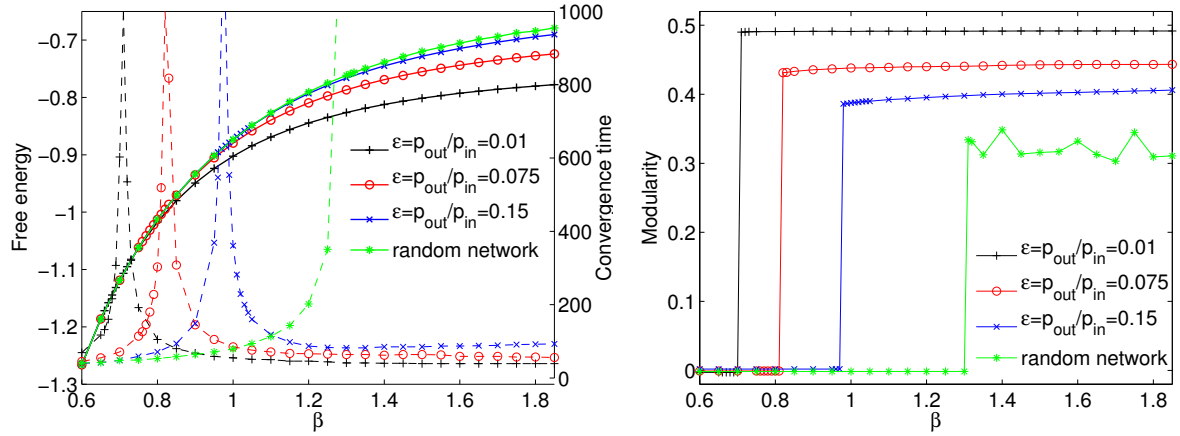

FIG. 3: Left: Free energy (solid) and convergence time (dashed) as a function of β for networks generated by the stochastic block model for three different values of $\epsilon = c_{\rm out}/c_{\rm in}$, also compared with an Erdős-Rényi graph. Right: retrieval modularity for these networks. All networks have size $n = 10^4$ and average degree $c = 3$. The networks generated by the SBM have $q = 2$ groups of equal size.

However, this assumes that the corresponding eigenvector of B is correlated with the community structure, so that perturbing BP away from the factorized fixed point will lead to the retrieval state. This is true as long as μ is outside the bulk of B's eigenvalues, which are confined to a disk of radius $\sqrt{c}$ in the complex plane [3]; if it is inside the bulk, then the community structure is washed out by localized eigenvectors and becomes hard to find. Thus the communities are detectable as long as $\mu > \sqrt{c}$: this is equivalent to $\beta_{\rm R} < \beta^*$, or equivalently $\epsilon < \epsilon^*$. Thus the retrieval state exists all the way down to the Kesten-Stigum transition where $\epsilon = \epsilon^*$, $\mu = \sqrt{c}$, $\beta_{\rm R} = \beta^*$. At that point, the relevant eigenvalue crosses into the bulk, and the retrieval phase disappears.

On the left in Fig. 4, we show the phase diagram of our algorithm on the SBM, including the paramagnetic, retrieval, and spin glass phases as a function of ϵ, with $q = 2$ and $c = 3$. The boundary $\beta_{\rm R}(\epsilon)$ between the paramagnetic and retrieval phases is in excellent agreement with to our expression (18). For $\epsilon < \epsilon^* \approx 0.267$, our algorithm finds a retrieval state for β between $\beta_{\rm R}(\epsilon)$ and β_{SG}. On the right, we show the accuracy of the retrieval partition $\{\hat{t}\}$, defined as its overlap with the planted partition:

$$O(\{t\},\{t^*\}) = \frac{1}{n}\sum_i \delta_{t_i^*,t_i}\,, \tag{19}$$

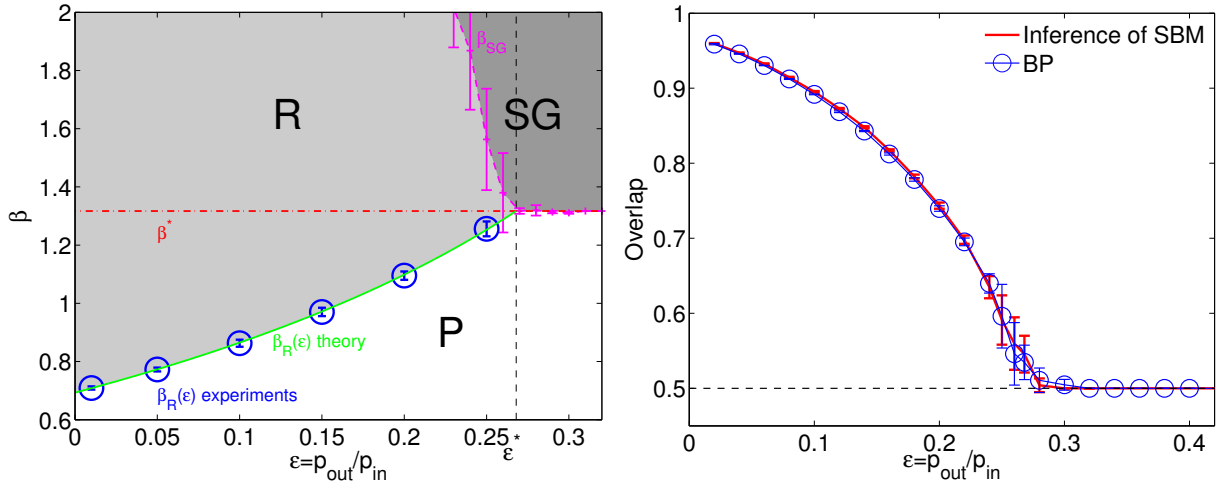

FIG. 4: Left: phase diagram for networks generated by the stochastic block model. Blue circles with error bars denote experimental estimates of $\beta_{\rm R}(\epsilon)$, the boundary between the paramagnetic and retrieval phases, and the solid green line shows our theoretical expression (18). The spin glass instability occurs for $\beta > \beta^*(2,3)$ (red dash-dotted line) and ϵ^* is the detectability transition (black dashed line). The white area is the paramagnetic phase, where the factorized fixed point is stable; the lightly shaded area is the retrieval phase, and the heavily shaded area is the spin glass phase. The magenta dashed line with error bars is the boundary between the retrieval and spin glass phases. Right: The overlap of the retrieval partition at $\beta = 1.315 \approx \beta^*(2,3)$ (blue circles) and the partition obtained with the algorithm of [4], which infers the parameters of the SBM. Experiments were carried out on networks with $n = 10^5$, $q = 2$ equal groups, and average degree $c = 3$, averaged over 10 random instances.

We also show the overlap of the partition obtained with the algorithm of [4, 5], which infers the parameters of the block model. Our algorithm provides the same overlap (and the same modularity) but without the need to learn the parameters through an expectation-maximization algorithm. Experiments with $q > 2$ gave similar behavior.

IV. RESULTS ON REAL-WORLD NETWORKS

In this section we consider real-world networks with ground-truth partitions such as the karate club network [33], a network of political blogs [34], and others.

In table I we list the retrieval modularity, and the overlap between the retrieval partition and the ground-truth partition, for several real-world networks. In each case, we set $\beta = \beta^*(q^*, c)$ as in (15), where q^* is the ground-truth number of groups and c is the average degree. From the table we see that our algorithm finds a retrieval state in all these networks, and that the retrieval modularity and overlap are high.

While we have focused here on assortative strucutre, the last network in the table, the adjacency network of common adjectives and nouns in the novel *David Copperfield* [2], is a disassortative network, since nouns are more likely to be adjacent to adjectives than other nouns and vice versa. In this case, we found a retrieval state with negative modularity, and high overlap with the ground truth, by setting β to $-\beta^*(q^*, c)$.

In addition, for the Gnutella and Epinions networks, our algorithm found 7 groups with modularity 0.519, and 4 groups with modularity 0.431 respectively. This suggests that these networks possess statistically significant large-scale communities, in contrast to the claim made in [35].

V. CHOOSING THE NUMBER OF GROUPS BY MAXIMIZING RETRIEVAL MODULARITY

Choosing the number q of groups in a network is a classical model selection problem. Setting q by maximizing the modularity is a widely-used heuristic in the network literature; however, as we have already seen, it is prone to overfitting. For example, the maximum modularity for an Erdős-Rényi graph is an increasing function of q, while the correct model has $q = 1$. Similarly, in the stochastic block model the likelihood increases, or the ground state energy decreases, until every node is assigned to its own group.

One approach [4, 5] is to use the free energy rather than the ground state energy. In essence, the entropic term penalizes overfitting, and gives us the total likelihood of the model summed over all partitions, as opposed to the likelihood of the best partition. This approach works well on synthetic graphs: the free energy decreases until we

Network	q^*	n	c	retrieval modularity	overlap
Zachary's karate club	2	34	4.588	0.371	1
Dolphin social network	2	62	5.129	0.395	0.887
Books about US politics	3	105	8.4	0.521	0.829
Political blogs	2	1222	27.36	0.426	0.948
Word adjacencies	2	112	7.589	-0.275	0.848
Gnutella	7	62586	4.726	0.519	
Epinions	4	75888	10.693	0.431	
Web-Google	5	916428	9.432	0.753	

TABLE I: Retrieval modularity and the overlap between the retrieval partition and the ground-truth partition for some real-world networks [2, 33, 34, 36, 37]. We also show results for the Gnutella and Epinions networks; here no ground truth is known, but based on our results we claim, contrary to [35], that these networks have statistically significant large-scale communities. Here q^* denotes the ground truth number of groups, except for Gnutella and Epinions, where we determined q^* using the procedure in Section V.

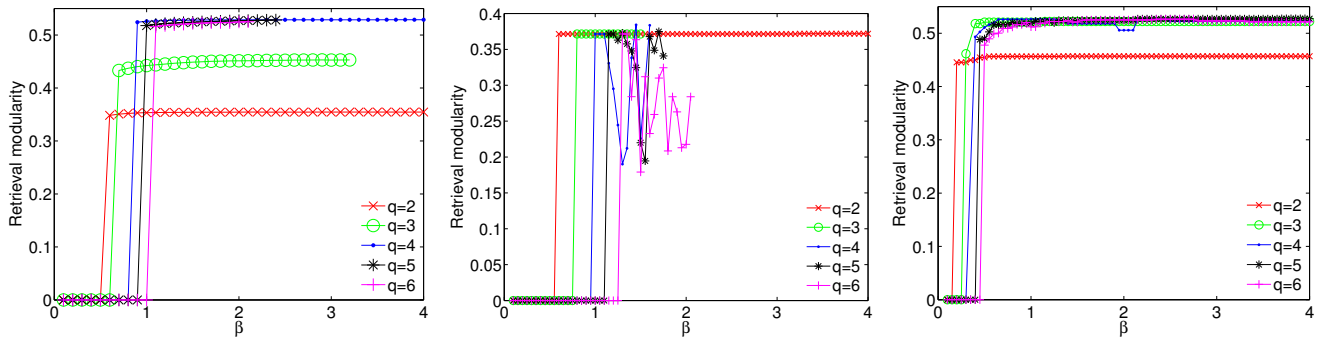

FIG. 5: Retrieval modularity when BP converges as a function of number of groups q used for a network generated by stochastic block model with $q^* = 4$ groups, 10^4 nodes and $\epsilon = 0.1$ (left), karate club network with $q^* = 2$ (middle) and network of political books with $q^* = 3$ (right).

reach the correct number of groups, after which it stays roughly constant. However, on real-world networks the free energy continues to decrease with q, for example as shown in Fig. 8 of [4]. Thus, for networks not generated by the SBM, it is not clear that this method works.

Here we propose to use retrieval modularity as a criterion for choosing q. Namely, we claim that the retrieval modularity increases with q until we reach the correct value q^*. For $q > q^*$, either the retrieval modularity stays the same, or the retrieval state disappears and we enter the spin glass state.

In Fig. 5, we plot the retrieval modularity $Q(\{\hat{t}\})$ for different values of q as a function of β for three networks with known community structure: a synthetic network generated by the SBM with $q^* = 4$, the Karate Club with $q^* = 2$, and a network of political books with $q^* = 3$. In each case, the retrieval modularity is largest for $q = q^*$, and is nearly independent of β throughout the retrieval state. Increasing q beyond q^* does not increase $Q(\{\hat{t}\})$; it stays constant or becomes unstable. Thus, as we argue above, our method gives the correct number of statistically-significant communities, rather than overfitting.

Note that here q^* refers to the top level of organization in the network. In Section VII below, we discuss using our approach to recursively divide communities into subcommunities. In that case, we use our procedure here to determine the number q^* of subcommunities we should split into at each stage, and stop splitting when we reach communities with $q^* = 1$.

VI. RELATION WITH DEGREE-CORRECTED STOCHASTIC BLOCK MODEL

In this section we pause to relate the modularity to the log-likelihood of a popular generative model, and to compare the corresponding message-passing algorithms. The degree-corrected stochastic block model (DCSBM) was introduced in [6] to overcome the fact that the SBM typically places low-degree and high-degree vertices into different groups, since it expects the degree distribution within each group to be Poisson.

The DCSBM's parameters are the expected node degrees $\{d_i\}$ and a $q \times q$ matrix of parameters ω_{rs}. Given a partition $\{t\}$, the number of edges A_{ij} between each pair $\langle ij \rangle$ is Poisson-distributed with mean $d_i d_j \omega_{t_i,t_j}$. In the simple graph case where $A_{ij} = 1$ if $\langle ij \rangle \in \mathcal{E}$ and $A_{ij} = 0$ otherwise, the log-likelihood of the network is then

$$L(\{t\}) = \log P(G|\{\omega_{ab}\},\{t\}) = \log \left(\prod_{\langle ij \rangle \in \mathcal{E}} d_i d_j \omega_{t_i t_j} \prod_{\langle ij \rangle} \mathrm{e}^{-d_i d_j \omega_{t_i t_j}} \right) . \tag{20}$$

If $\omega_{rs} = \omega_{\mathrm{in}}$ for $r = s$ and ω_{out} for $r \neq s$, the likelihood can be written as

$$L = \sum_{\langle ij \rangle} \big(\log(d_i d_j \omega_{\mathrm{out}}) - d_i d_j \omega_{\mathrm{out}} \big) + \left(\log \frac{\omega_{\mathrm{in}}}{\omega_{\mathrm{out}}} \right) \left[\sum_{\langle ij \rangle \in \mathcal{E}} \delta_{t_i t_j} - \frac{\omega_{\mathrm{in}} - \omega_{\mathrm{out}}}{\log(\omega_{\mathrm{in}}/\omega_{\mathrm{out}})} \sum_{\langle ij \rangle} d_i d_j \delta_{t_i t_j} \right] . \tag{21}$$

Comparing with the definition (1) of modularity, if we set ω_{in} and ω_{out} such that

$$\beta = \log \frac{\omega_{\mathrm{in}}}{\omega_{\mathrm{out}}} \quad \text{and} \quad 2m = \frac{\log(\omega_{\mathrm{in}}/\omega_{\mathrm{out}})}{\omega_{\mathrm{in}} - \omega_{\mathrm{out}}} , \tag{22}$$

then the second term in (21) is $\beta m Q(\{t\})$. Since the first term in (21) does not depend on $\{t\}$, we have

$$\mathrm{e}^{L(\{t\})} \propto \mathrm{e}^{\beta m Q(\{t\})} ,$$

and the Gibbs distribution (3) is exactly the Gibbs distribution of partitions in the DCSBM.

Thus, for any fixed β, there are parameters $\omega_{\mathrm{in}}, \omega_{\mathrm{out}}$ of the DCSBM such that these distributions have the same free energy and the same ground state. Belief propagation on the DCSBM was described in [20], and one can optimize the parameters $\omega_{\mathrm{in}}, \omega_{\mathrm{out}}$ through an expectation-maximization algorithm analogous to [4, 5]. However, our approach is different in several ways.

- We define community structure directly in terms of a classic measure, the modularity, as opposed to the log-likelihood of a generative model.
- Rather than having to fit the parameters of the DCSBM with an EM algorithm, we have a single temperature parameter β. We can usually detect communities by setting $\beta = \beta^*$ as (15); at worst, we just have to a scan a small region.
- As stated in the previous section, for real-world networks the retrieval modularity appears to be a good guide to the number of groups q^*, while the free energy of the (DC)SBM continues to decrease for $q > q^*$.
- Our approach appears to work equally well for networks with Poisson degree distributions (generated by the SBM) and those with heavy-tailed degree distributions (such as the network of political blogs, where the DCSBM does much better [6]). In particular, we have no need to do model selection between SBM and DCSBM, as was done using the Bethe free energy in [20].

VII. HIERARCHICAL CLUSTERING

Many networks appear to have hierarchical structure, and a number of algorithms have been proposed to find it (e.g. [2, 7, 19, 38, 39]). One approach is to work from the top down, dividing the network into communities, asking whether there are substructures in the subgraphs consisting of the nodes and internal edges in each one. We can do this by applying the same criteria to each subgraph that we did to the entire graph: namely, whether it has statistically significant communities (and if so, how many) as opposed to being essentially a random graph. We then apply our technique recursively, checking for a retrieval state in each subgraph, finding the optimal number q^* of subgroups, dividing it into smaller subgraphs, and so on. We conclude when we reach subgraphs with no retrieval state, indicating that they have no internal structure.

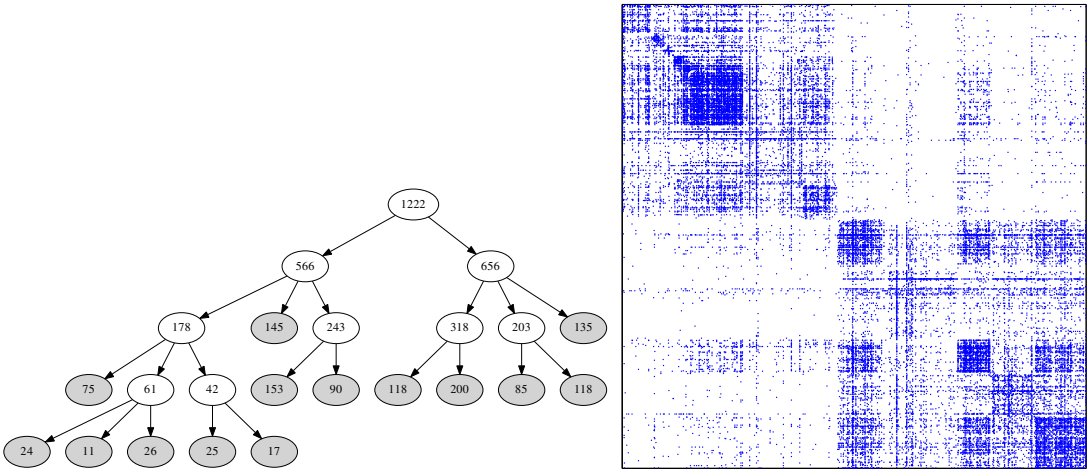

FIG. 6: Left, a hierarchical division of the political blog network [34] where we apply our technique recursively, looking for a retrieval state and optimizing the number of groups in which to split the community at each stage. The number in each circle indicates the size of the group. We stop when no retrieval state is detected, indicating that the remaining groups (shaded) have no statistically significant subcommunities. Right, the adjacency matrix of the network ordered according to this partition.

For networks generated by the SBM, the communities are Erdős-Rényi graphs, and indeed our algorithm finds no retrieval state in the subgraphs. We find the same situation in some small real-world networks, e.g. Zachary's karate club, where our algorithm concludes that the communities have no internal structure. In some larger real-world networks, on the other hand, our algorithm repeatedly finds a retrieval state in the subgraphs, suggesting a deep hierarchical structure.

An example is the network of political blogs with $n = 1222$ [34]. Our algorithm first finds two large communities of sizes 566 and 656 corresponding to liberals and conservatives, and agreeing with the ground-truth labels on 95% of the nodes. But as shown in Fig. 6, our algorithm finds subcommunities at multiple scales. It divides the first community into 3 subgroups with 178, 243 and 145 nodes. The subgroup with 145 nodes has no retrieval state, while the others are subdivided further. Similarly, the second community is divided into 3 subgroups with 318, 203 and 135 nodes. Eventually, the algorithm terminates with a total of 14 subgroups (the shaded leaves of the tree in Fig. 6) in which no retrieval state can be found. We show the adjacency matrix with nodes ordered by this final partition on the right of Fig. 6, and the hierarchical structure is clearly visible.

We note that using a nested SBM to explore hierarchical structure was studied in [39], where the blog network was also reported to have hierarchical structure. Our results are slightly different, giving 14 rather than 17 subgroups, but the first 3 levels of subdivision are similar.

VIII. CONCLUSION AND DISCUSSION

We have presented a scalable message-passing algorithm for finding statistically significant community structures in networks, and determining the number of groups. By applying our algorithm recursively, we can find hierarchical structure, breaking communities into subcommunities until the remaining groups have no internal structure. Our approach returns a quantity we call the retrieval modularity, namely the modularity of a typical partition from the Gibbs distribution.

We claim that this is a much more robust approach than the traditional one of maximizing the modularity, which is both computationally difficult and prone to overfitting. While our algorithm is related to the degree-corrected block model, we do not have to learn the parameters of the block model with an expectation-maximization algorithm, or perform model selection between the stochastic block model and its degree-corrected variant.

Another recent proposal for distinguishing statistically significant communities, and determining the number of groups, is to use the number of real eigenvalues of the non-backtracking matrix, outside the bulk of the spectrum [3]. For some networks, such as the network of political blogs, this gives a larger number than the q^* we found here; it may be that, in some sense, this method detects not just top-level communities, but subcommunities deeper in the hierarchy. It would be interesting to perform a detailed comparison of the two methods.

We note that our BP equations can easily be extended to generalizations of the modularity, where the graph is weighted, or where a parameter γ represents the relative importance of the expected number of internal edges [15].

Finally, it would be interesting to apply BP to other objective functions, such as normalized cut or conductance, devising Hamiltonians from them and considering the resulting Gibbs distributions. We leave this for future work.

Acknowledgments

We are grateful to Lenka Zdeborová and Mark Newman for helpful discussions. This work was supported by AFOSR and DARPA under grant FA9550-12-1-0432.

[1] U. V. Luxburg, M. Belkin, O. Bousquet, and Pertinence, Stat. Comput (2007).
[2] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006), URL http://link.aps.org/doi/10.1103/PhysRevE.74.036104.
[3] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, Proceedings of the National Academy of Sciences **110**, 20935 (2013), http://www.pnas.org/content/110/52/20935.full.pdf+html, URL http://www.pnas.org/content/110/52/20935.abstract.
[4] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. E **84**, 066106 (2011), URL http://link.aps.org/doi/10.1103/PhysRevE.84.066106.
[5] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. Lett. **107**, 065701 (2011), URL http://link.aps.org/doi/10.1103/PhysRevLett.107.065701.
[6] B. Karrer and M. E. J. Newman, Phys. Rev. E **83**, 016107 (2011), URL http://link.aps.org/doi/10.1103/PhysRevE.83.016107.
[7] A. Clauset, M. E. Newman, and C. Moore, Physical Review E **70**, 066111 (2004).
[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment **2008**, P10008 (2008).
[9] M. Rosvall and C. T. Bergstrom, Proceedings of the National Academy of Sciences **105**, 1118 (2008).
[10] S. Fortunato, Physics Reports **486**, 75 (2010), ISSN 0370-1573, URL http://www.sciencedirect.com/science/article/pii/S0370157309002841.
[11] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004), URL http://link.aps.org/doi/10.1103/PhysRevE.69.026113.
[12] M. E. Newman, Physical Review E **69**, 066133 (2004).
[13] J. Duch and A. Arenas, Physical Review E **72**, 027104 (2005).
[14] L. Zdeborová and S. Boettcher, Journal of Statistical Mechanics: Theory and Experiment **2010**, P02020 (2010).
[15] J. Reichardt and S. Bornholdt, Physical Review E **74**, 016110 (2006).
[16] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Physical Review E **70**, 025101 (2004).
[17] P. Šulc and L. Zdeborová, Journal of Physics A Mathematical General **43**, B5003 (2010), 0912.3563.
[18] B. H. Good, Y.-A. de Montjoye, and A. Clauset, Physical Review E **81**, 046106 (2010).
[19] A. Clauset, C. Moore, and M. E. Newman, Nature **453**, 98 (2008).
[20] X. Yan, J. E. Jensen, F. Krzakala, C. Moore, C. R. Shalizi, L. Zdeborova, P. Zhang, and Y. Zhu, arXiv preprint arXiv:1207.3994 (2012).
[21] J. Yedidia, W. Freeman, and Y. Weiss, in *International Joint Conference on Artificial Intelligence (IJCAI)* (2001).
[22] M. Mézard and G. Parisi, Eur. Phys. J. B **20**, 217 (2001).
[23] O. Rivoire, G. Biroli, O. C. Martin, and M. Mézard, The European Physical Journal B-Condensed Matter and Complex Systems **37**, 55 (2004).
[24] J. De Almeida and D. Thouless, Journal of Physics A: Mathematical and General **11**, 983 (1978).
[25] H. Kesten and B. P. Stigum, The Annals of Mathematical Statistics **37**, 1211 (1966).
[26] H. Kesten and B. P. Stigum, The Annals of Mathematical Statistics **37**, 1463 (1966).
[27] M. Mézard and A. Montanari, J. Stat. Phys. **124**, 1317 (2006).
[28] S. Janson and E. Mossel, Annals of probability pp. 2630–2649 (2004).
[29] D. Hu, P. Ronhovde, and Z. Nussinov, Philosophical Magazine **92**, 406 (2012).
[30] E. Mossel, J. Neeman, and A. Sly, arXiv preprint arXiv:1202.1499 (2012).
[31] L. Massoulie, arXiv preprint arXiv:1311.3085 (2013).
[32] E. Mossel, J. Neeman, and A. Sly, arXiv preprint arXiv:1311.4115 (2013).
[33] W. W. Zachary, Journal of anthropological research pp. 452–473 (1977).
[34] L. A. Adamic and N. Glance, in *Proceedings of the 3rd international workshop on Link discovery* (ACM, 2005), pp. 36–43.
[35] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Mathematics **6**, 29 (2009).
[36] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, Behavioral Ecology and Sociobiology **54**, 396 (2003).
[37] V. Krebs, social Network Analysis software & services for organizations, communities, and their consultants. Available at www.orgnet.com/.
[38] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, Proceedings of the National Academy of Sciences **104**, 15224 (2007).

[39] T. P. Peixoto, arXiv preprint arXiv:1310.4377 (2013).