Supporting Information

Zhang and Moore 10.1073/pnas.1409770111

SI Text

Belief Propagation Equation and Bethe Free Energy

In this section we derive the BP update equations appearing in the main text. 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 $j \neq k$. If the Hamiltonian is $-mQ$, the update equations for these messages are as follows:

$$\begin{aligned}\psi_t^{i\to k} &= \frac{1}{Z_{i\to k}} \prod_{j\in\partial i\setminus k} \sum_{s=1}^{q} \mathrm{e}^{\beta\delta_{st}} \psi_s^{j\to i} \prod_{j\neq i,k} \sum_{s=1}^{q} \mathrm{e}^{-\beta\left(d_i d_j/2m\right)\delta_{st}} \psi_s^{j\to i} \\ &= \frac{1}{Z_{i\to k}} \prod_{j\in\partial i\setminus k} \left(1+\psi_t^{j\to i}\left(\mathrm{e}^{\beta}-1\right)\right) \prod_{j\neq i,k} \left(1+\psi_t^{j\to i}\left(\mathrm{e}^{-\beta\left(d_i d_j/2m\right)}-1\right)\right).\end{aligned} \quad \textbf{[S1]}$$

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 $\psi_t^i = \Pr[t_i = t]$ is then

$$\begin{aligned}\psi_t^{i} &= \frac{1}{Z_{i}} \prod_{j\in\partial i} \sum_{s=1}^{q} \mathrm{e}^{\beta\delta_{st}} \psi_s^{j\to i} \prod_{j\neq i} \sum_{s=1}^{q} \mathrm{e}^{-\beta\left(d_i d_j/2m\right)\delta_{st}} \psi_s^{j\to i} \\ &= \frac{1}{Z_{i}} \prod_{j\in\partial i} \left(1+\psi_t^{j\to i}\left(\mathrm{e}^{\beta}-1\right)\right) \prod_{j\neq i} \left(1+\psi_t^{j\to i}\left(\mathrm{e}^{-\beta\left(d_i d_j/2m\right)}-1\right)\right),\end{aligned} \quad \textbf{[S2]}$$

which is the same as [**S1**] 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

$$\psi_{st}^{ij} = \frac{1}{Z_{ij}} \mathrm{e}^{\beta\delta_{st}} \psi_s^{j\to i} \psi_t^{i\to j} \quad . \quad \textbf{[S3]}$$

The update Eq. **S1** involves qn^2 messages: Every node interacts with every other one, not just its neighbors. However, in the sparse case we can simplify the effect of nonneighbors, by replacing them with an external field as in refs. 1 and 2. If $k \notin \partial i$ and $d_i, d_k \ll \sqrt{m}$, we have

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

In that case, we can identify the messages $\psi_t^{i\to k}$ that i sends to its nonneighbors k with its marginal ψ_t^i. Then [**S1**] simplifies to

$$\begin{aligned}\psi_t^{i\to k} &= \frac{1}{Z_{i\to k}} \prod_{j\in\partial i\setminus k} \left(1+\psi_t^{j\to i}\left(\mathrm{e}^{\beta}-1\right)\right) \prod_{j\neq i,k} \left(1+\psi_t^{j}\left(\mathrm{e}^{-\beta\left(d_i d_j/2m\right)}-1\right)\right) \\ &\approx \frac{1}{Z_{i\to k}} \exp\left(-\frac{\beta d_i}{2m}\theta_t + \sum_{j\in\partial i\setminus k} \log\left(1+\psi_t^{j\to i}\left(\mathrm{e}^{\beta}-1\right)\right)\right),\end{aligned} \quad \textbf{[S4]}$$

where

$$\theta_t = \sum_{j=1}^{n} d_j \psi_t^j \quad \textbf{[S5]}$$

denotes an external field acting on nodes in group t, which we update after each BP iteration. Iterating [**S4**] now has computational complexity qm, which is linear in the number of edges when q is fixed.

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

$$f_{\mathrm{Bethe}} = -\frac{1}{n\beta}\left(\sum_i \log Z_i - \sum_{\langle ij\rangle\in\mathcal{E}} \log Z_{ij} + \frac{\beta}{4m}\sum_t \theta_t^2\right), \quad \textbf{[S6]}$$

where Z_i and Z_{ij} are the normalization constants for the one- and two-point marginals appearing in [**S2**] and [**S3**]. BP fixed points are also stationary points of the Bethe free energy (3).

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 Eq. **S4**. 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

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

In Fig. S1 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 $f_{\mathrm{Bethe}}^{\mathrm{fact}}$, because they are all in the paramagnetic phase. For each value of ε, there is a critical β_{R} at which the free energy splits off from the others, where it makes a transition to a retrieval state with $f_{\mathrm{Bethe}} < f_{\mathrm{Bethe}}^{\mathrm{fact}}$. 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 $\beta^* = \beta_{\mathrm{SG}}$ 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.

Relation with the Degree-Corrected Stochastic Block Model

The degree-corrected stochastic block model (DCSBM) was introduced in ref. 4 to overcome the fact that the SBM typically places low-degree and high-degree vertices into different groups, because 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

$$\begin{aligned}L(\{t\}) &= \log P\left(G\middle|\{\omega_{ab}\},\{t\}\right) \\ &= \log\left(\prod_{\langle ij\rangle\in\mathcal{E}} d_i d_j \omega_{t_i t_j} \prod_{\langle ij\rangle} \mathrm{e}^{-d_i d_j \omega_{t_i t_j}}\right).\end{aligned} \quad \textbf{[S7]}$$

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

$$L = \sum_{\langle ij \rangle} \left(\log(d_i d_j \omega_{\text{out}}) - d_i d_j \omega_{\text{out}} \right) + \left(\log \frac{\omega_{\text{in}}}{\omega_{\text{out}}} \right) \left[\sum_{\langle ij \rangle \in \mathcal{E}} \delta_{t_i t_j} - \frac{\omega_{\text{in}} - \omega_{\text{out}}}{\log(\omega_{\text{in}}/\omega_{\text{out}})} \sum_{\langle ij \rangle} d_i d_j \delta_{t_i t_j} \right]. \quad \textbf{[S8]}$$

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

$$\beta = \log \frac{\omega_{\text{in}}}{\omega_{\text{out}}} \quad \text{and} \quad 2m = \frac{\log(\omega_{\text{in}}/\omega_{\text{out}})}{\omega_{\text{in}} - \omega_{\text{out}}}, \quad \textbf{[S9]}$$

then the second term in **[S8]** is $\beta m Q(\{t\})$. Because the first term in **[S8]** does not depend on $\{t\}$, we have

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

and the Gibbs distribution is exactly the Gibbs distribution of partitions in the DCSBM.

Thus, for any fixed β, there are parameters $\omega_{\text{in}}, \omega_{\text{out}}$ 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 ref. 5, and one can optimize the parameters $\omega_{\text{in}}, \omega_{\text{out}}$ through an expectation-maximization algorithm analogous to that in refs. 1 and 2. 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 in the main text; at worst, we just have to a scan a small region.
- For real-world networks the retrieval modularity appears to be a good guide to the number of groups q^*, whereas 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 LFR benchmark (6) and the network of political blogs, where the DCSBM does much better (4). In particular, we have no need to do model selection between SBM and DCSBM, as was done using the Bethe free energy in ref. 5.

The Nishimori Line and the Optimal Temperature

When data are produced by an underlying generative model, inference of the latent parameters can be done optimally along the Nishimori line (7, 8), where the Gibbs distribution is exactly the posterior distribution of the latent parameters (in this case the group labels or partitions). If the network is generated by the DCSBM, then Eq. **S9** gives a $\beta_{\text{Nishimori}}$ that corresponds to the correct parameters at the Nishimori line. Determining the parameters, and therefore $\beta_{\text{Nishimori}}$, could be done with an EM algorithm as in refs. 1 and 2, but our goal is to avoid this additional learning step. Moreover, if the network is not actually generated by the DCSBM, there is a priori no value of β that corresponds to the Nishimori line and no way to determine the optimal β without access to the ground truth.

However, for synthetic networks generated by the SBM, we can construct an approximate Nishimori line by omitting the difference between the SBM and the DCSBM, by assuming that the expected degrees are actually the same. This gives

$$\beta_{\text{Nishimori}} = \log \frac{c_{\text{in}}}{c_{\text{out}}} = -\log \epsilon.$$

In Fig. S2 we show the phase diagram from the main text with this approximate Nishimori line added. It passes through the critical point (ϵ^*, β^*) (one can check analytically that $\beta^* = -\log \epsilon^*$) and then it avoids the spin-glass phase, passing directly from the paramagnetic phase to the retrieval phase. This recovers the fact that replica symmetry breaking cannot occur on the Nishimori line (9).

Choosing the Number of Groups

Choosing the number q of groups in a network is a classic 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, whereas 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 (1, 2) 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 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 figure 8 of ref. 2. Thus, for networks not generated by the SBM, it is not clear that this method works.

Here we propose to use the retrieval modularity $Q(\{\hat{t}\})$ as a criterion for choosing q. Namely, we claim that $Q(\{\hat{t}\})$ increases with q until we reach the correct value q^*. For $q > q^*$, either $Q(\{\hat{t}\})$ stays the same or the retrieval phase disappears and we enter the spin-glass phase. In Fig. S3 we plot $Q(\{\hat{t}\})$ and BP convergence time for the karate club network with different values of q. With $q = 2$, i.e., the ground-truth number of groups, the retrieval phase is very large. For larger q, the retrieval phase becomes narrower, and $Q(\{\hat{t}\})$ does not increase. Note the similarity with Fig. 2, *Right* in the main text.

In Fig. S4, we plot $Q(\{\hat{t}\})$ 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 $q^* = 2$ (10), and a network of political books with $q^* = 3$ (11). In each case, $Q(\{\hat{t}\})$ stops growing at $q = q^*$ and is nearly independent of β throughout the retrieval phase. (To deal with fluctuations, in practice we don't increase q unless the retrieval modularity increases by at least some threshold value.) Thus, our method gives the correct number of communities, rather than overfitting.

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

Additional Comparisons with Louvain and OSLOM

In Fig. S5 we show comparisons between our BP algorithm, Louvain (12), and OSLOM (13) on networks with power-law degree distributions. In Fig. S5, *Left*, the graphs are generated by the LFR benchmark process (6). We show the normalized mutual information (14) as a function of the mixing parameter μ. As for the SBM graphs shown in the main text, there is a parameter range where BP achieves a higher NMI than the other algorithms. In Fig. S5, *Right*, we show results for a network

with no community structure, where the degree distribution follows a power law with exponent −2. Whereas BP correctly chooses $q^* = 1$ as the number of groups, the other algorithms overfit, finding a number of communities that grow with the network size. These results are similar to those shown in Fig. 5 of the main text.

The Resolution Limit

In this section we describe results of our algorithm on the ring-of-cliques network, which is the standard example of the resolution limit (15). This network has size $n = ab$; it consists of a cliques, each of which is composed of b nodes, which are connected to the neighboring cliques by a single link. Thus, the intuitively correct partition of the network puts each clique into one group. However, when b is sufficiently small compared with a, maximizing the modularity forces us to combine multiple cliques (15–20). For example, if $a = 24$ and $b = 5$, the correct partition with 24 groups has modularity 0.8674, whereas the division with 12 groups of two cliques each has modularity 0.8712. As a consequence, maximizing the modularity fails to divide the network correctly into the cliques.

In Fig. S6 we plot the dendrogram obtained by our hierarchical clustering algorithm starting from three different initial conditions (from *Top* to *Bottom*). All three dendrograms have two levels below the root. The first split creates groups consisting of multiple cliques, but the second split correctly assigns each clique to its own group. At that point the algorithm concludes that the cliques have no internal structure, and it stops subdividing. This suggests that our hierarchical clustering algorithm may be able to avoid the resolution limit.

1. Decelle A, Krzakala F, Moore C, Zdeborová L (2011) Inference and phase transitions in the detection of modules in sparse networks. *Phys Rev Lett* 107(6):065701.
2. Decelle A, Krzakala F, Moore C, Zdeborová L (2011) Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys Rev E Stat Nonlin Soft Matter Phys* 84(6 Pt 2):066106.
3. Yedidia J, Freeman W, Weiss Y (2001) Understanding belief propagation and its generalizations. *Proceedings of the International Joint Conference on Artificial Intelligence* (Morgan Kaufmann Publishers Inc., San Francisco).
4. Karrer B, Newman MEJ (2011) Stochastic block models and community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 83(1 Pt 2):016107.
5. Yan X, et al. (2014) Model selection for degree-corrected block models. *J Stat Mech* 2014:P05007.
6. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. *Phys Rev E Stat Nonlin Soft Matter Phys* 78(4 Pt 2):046110.
7. Iba Y (1999) The Nishimori line and Bayesian statistics. *J Phys Math Gen* 32:3875.
8. Nishimori H (2012) *Statistical Physics of Spin Glasses and Information Processing* (Oxford Univ Press, Oxford).
9. Montanari A (2008) Estimating random variables from random sparse observations. *Eur Trans Telecomm* 19:385.
10. Zachary WW (1977) An information flow model for conflict and fission in small groups. *J Anthropol Res* 33(4):452–473.
11. Krebs V, Social Network Analysis software & services for organizations, communities, and their consultants. Available at www.orgnet.com/. Accessed October 26, 2014.
12. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech* 2008:P10008.
13. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. *PLoS ONE* 6(4):e18961.
14. Danon L, Diaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. *J Stat Mech* 2005:P09008.
15. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. *Proc Natl Acad Sci USA* 104(1):36–41.
16. Krzakała F, Montanari A, Ricci-Tersenghi F, Semerjian G, Zdeborová L (2007) Gibbs states and the set of solutions of random constraint satisfaction problems. *Proc Natl Acad Sci USA* 104(25):10318–10323.
17. Adamic LA, Glance N (2005) The political blogosphere and the 2004 US election: Divided they blog. *Proceedings of the Third International Workshop on Link Discovery* (ACM, New York), pp 36–43.
18. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2009) Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Math* 6:29.
19. Lusseau D, et al. (2003) The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav Ecol Sociobiol* 54:396.
20. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E Stat Nonlin Soft Matter Phys* 74(3 Pt 2):036104.

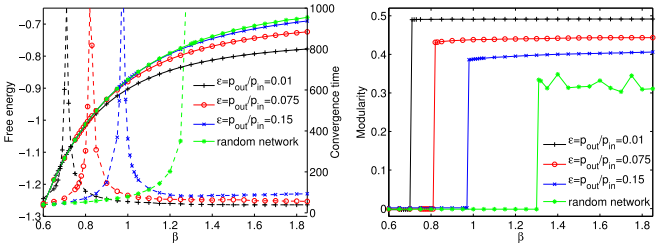

Fig. S1. (*Left*) Free energy (solid lines) and convergence time (dashed lines) as a function of β for networks generated by the stochastic block model for three different values of $\epsilon = c_{out}/c_{in}$, also compared with an Erdős–Rényi graph. (*Right*) Retrieval modularity for these networks. 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.

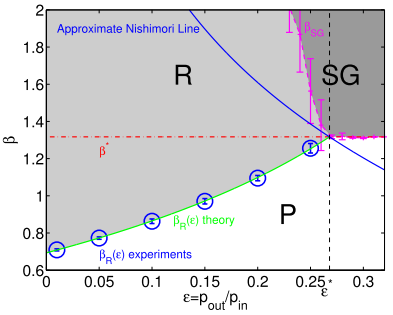

Fig. S2. The phase diagram from the main text for networks generated by the stochastic block model, with the approximate Nishimori line $\beta_{\text{Nishimori}} = -\log \epsilon$ added (blue line). Replica symmetry breaking cannot occur on the Nishimori line, and indeed it avoids the spin-glass phase. Inference at $\beta_{\text{Nishimori}}$ would be optimal, but it would require us to learn, or infer, the correct value of the parameter ε.

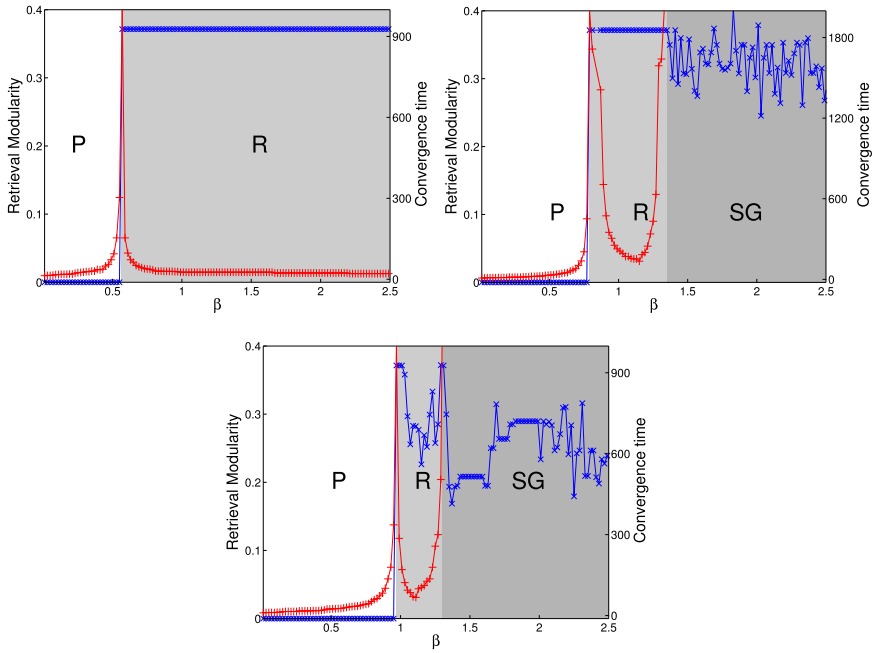

Fig. S3. Retrieval modularity (blue ×) and BP convergence time (red +) of the karate club network with two groups (*Top Left*), three groups (*Top Right*), and four groups (*Bottom*). With $q = 2$, which is the ground-truth value, the system has a very strong community structure, represented by a large retrieval phase starting at $\beta_R = 0.565$. With $q = 3$, the retrieval phase exists between $\beta_R = 0.79$ and $\beta_{SG} = 1.35$; compare Fig. 2, *Right* in the main text. With $q = 4$ groups, the retrieval phase becomes even narrower, between $\beta_R = 0.97$ and $\beta_{SG} = 1.3$.

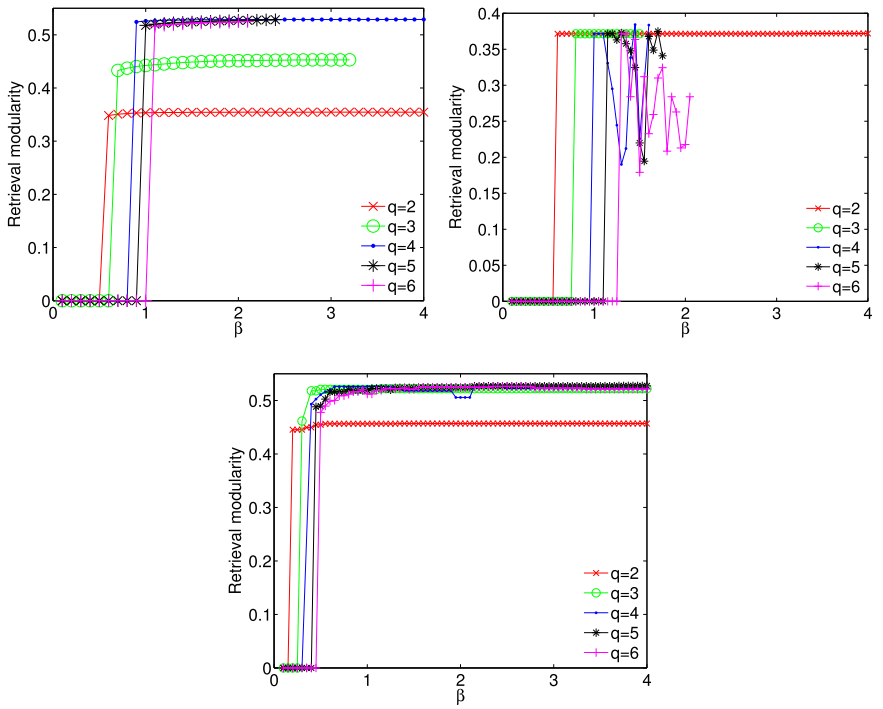

Fig. S4. Retrieval modularity as a function of q for three networks where the number of groups is known: a network generated by the stochastic block model with $q^* = 4$, $n = 10^4$, and $\epsilon = 0.1$ (*Top Left*); the karate club with $q^* = 2$ (*Top Right*); and the network of political books with $q^* = 3$ (*Bottom*). In each case, for $q > q^*$ the retrieval modularity stops growing until the spin-glass phase appears.

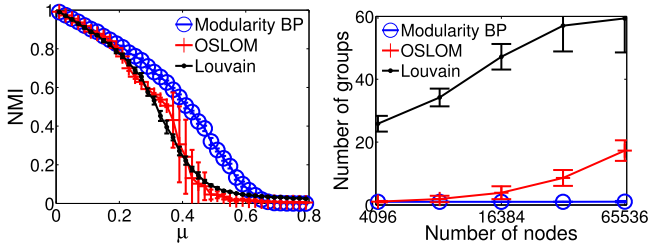

Fig. S5. Comparison of BP, the Louvain method, and OSLOM on benchmark networks with power-law degree distributions. (*Left*) Networks are LFR benchmarks with $n = 10^4$ and $c = 4$. The distribution of community sizes follows a power law with exponent -1, ranging from 200 to 400. The degree distribution is a power law with exponent -2, and the maximum degree is 30. We show the normalized mutual information (NMI) as a function of the mixing parameter μ, and there is a range of μ where BP achieves a higher NMI than the other algorithms. (*Right*) We show results on a random graph with no community structure, with a power-law degree distribution with exponent -2 and mean $c = 6$. Here BP correctly chooses $q^* = 1$ for the number of groups, whereas the other algorithms overfit, selecting a number of groups that grow with n. For both graphs, each data point is averaged over 20 instances. Compare Fig. 5 in the main text.

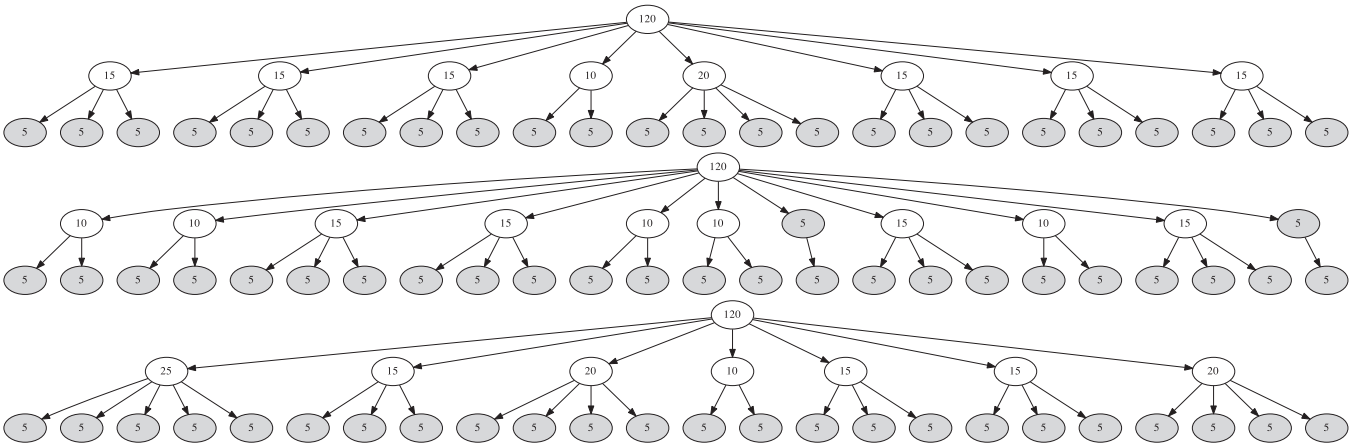

Fig. S6. Three dendrograms obtained by our hierarchical clustering algorithm on the ring of cliques, generated by independent runs with different initial conditions. Here there are $a = 24$ cliques of size $b = 5$ each. The number inside each node indicates the number of nodes in it. In all three runs, the first level of splitting merges multiple cliques together, but the second level correctly divides the network into individual cliques. This offers some evidence that our hierarchical algorithm can overcome the resolution limit, as opposed to algorithms that maximize the modularity.