Supporting Information

Zhao et al. 10.1073/pnas.1006642108

SI Text

1. Additional Simulation Results. Here we show results for the simulation of two pure communities with no background, comparing extraction, partition by modularity, and the block model fitted by Markov chain Monte Carlo. Specifically, we generate a network from the block model with $K = 2, P_{11} = 0.5, P_{22} = 0.4$, and $P_{12} = 0.05$ and vary the block sizes ($n_1 = 100, 200, 300, n_2 = 1{,}000 - n$). Fig. S1 shows the box plots of the adjusted Rand index for the three methods. For $n_1 = 100$, the block model does best (which is expected because the data are in fact generated by the block model), closely followed by modularity and extraction, which does a little worse because of "losing" some of the lower degree nodes in the tighter community and extracting a slightly smaller "core." For the more balanced communities ($n_1 = 200$ and 300), all three methods do perfectly.

2. A Hypothesis Test for Determining the Number of Communities. Here we give more details on the proposed hypothesis test under the block model for determining the number of communities. The null hypothesis is that the subgraph under consideration is a random realization from the Erdos–Renyi model. To test this hypothesis, we need to be able to simulate graphs from this distribution that "match" the observed subgraph in a suitable way. If the graph is unweighted, the simulation is trivial: We can simply generate N independent random Erdos–Renyi graphs with the same number of nodes n as the subgraph to be tested, and the same number of edges placed independently at random. If the graph is weighted, we first generate the same number of edges between randomly chosen pairs without weights, and then assign the weights from the original graph at random. Once the N random graphs are generated, we maximize the value of $\tilde{W}$ for each one of them. Then, if the value of $\tilde{W}$ for the proposed split on the real graph is higher than the $100(1 - \alpha)$th percentile of the N simulated values, we can reject the null hypothesis at level α and proceed with the split.

Although we have not yet been able to obtain the analytical form of this distribution, we note that the critical values depend only on n and p. As a rough guide, we provide an estimate of the 5% critical values for several values of p as a function of n in Fig. S2.

In addition, we also computed these critical values for the karate club (Table S1) and the school network data (Table 2). The karate club has a clear first split and very weak subsequent splits. For the school data, all six splits are significant (corresponding to the six grades), with the first four being particularly strong. The seventh split (not shown in the paper to facilitate comparison to the grouping into six grades) is not significant even at the 10% significance level.

3. Proofs.

Proof of Theorem 1: From assumption 3 and $R(\boldsymbol{s},\boldsymbol{c})\overset{P}{\to}R$,

$$\begin{aligned}\frac{1}{n^2}E[O_{ab}|\mathbf{c}] &= \frac{1}{n^2}\sum_{i,j}P_{c_ic_j}I(s_i = a,s_j = b)\\ &= \sum_{s,t}R_{as}(\mathbf{s},\mathbf{c})P_{st}R_{bt}(\boldsymbol{s},\boldsymbol{c}) \to (RPR^T)_{ab}. \qquad \textbf{[1]}\end{aligned}$$

From assumption 2, we have

$$\begin{aligned}\frac{1}{n^4}\text{Var}[O_{ab}|\boldsymbol{c}] &\le \frac{1}{n^4}\text{Var}\Big[\sum_{i,j}A_{ij}|\boldsymbol{c}]\\ &= \frac{1}{n^4}\sum_{i,j}\text{Var}[A_{ij}|\mathbf{c}\Big] + \frac{2}{n^4}\sum_{i=1}\sum_{j\ne i,k\ne i,j<k}\text{Cov}[A_{ij},A_{ik}|\boldsymbol{c}]\\ &\le \frac{1}{n^4}Cn\binom{n}{2} \to 0.\end{aligned}$$

Thus (a) holds by the dominated convergence theorem.

It is straightforward to show that (a) implies

$$\begin{aligned}W(S)\overset{P}{\to}f(R,P) &= \frac{1}{(r_{11}+r_{12})^2}(r_{11}^2P_{11} + 2r_{11}r_{12}P_{12} + r_{12}^2P_{22})\\ &\quad - \frac{1}{(r_{11}+r_{12})(r_{21}+r_{22})}(r_{11}r_{21}P_{11} + r_{11}r_{22}P_{12}\\ &\quad + r_{12}r_{21}P_{12} + r_{12}r_{22}P_{22}),\\ n^{-2}\tilde{W}(S)\overset{P}{\to}\tilde{f}(R,P) &= (r_{11}+r_{12})(r_{21}+r_{22})f(R,P).\end{aligned}$$

To maximize f under the constraint $\mathbf{1}^T\mathbf{R} = (\pi,1-\pi)$, we apply the transformation $t_1 = r_{11}/(r_{11}+r_{12})$, $t_2 = r_{22}/(r_{21}+r_{22})$ and obtain

$$\begin{aligned}f &= P_{22} - P_{12} + (P_{11} - 2P_{12} + P_{22})\left[t_1(t_1+t_2-1) - \frac{1}{2}(t_1+t_2)\right]\\ &\quad + \frac{1}{2}(P_{11} - P_{22})(t_1+t_2).\end{aligned}$$

It is easy to verify that the function $g(t_1,t_2) = t_1(t_1+t_2-1)-(t_1+t_2)/2$ has two maximizers, $t_1 = 1,t_2 = 1$ and $t_1 = 0,t_2 = 0$. Thus under the condition $P_{11} - 2P_{12} + P_{22} > 0,P_{11} > P_{22}$, the unique maximizer of f is $t_1 = 1,t_2 = 1$, or equivalently, $\boldsymbol{R} = \text{diag}(\pi,1-\pi)$.

For $\tilde{f}$, applying the same transformation, we obtain

$$\begin{aligned}\tilde{f} &= \frac{(t_1-\pi)(t_2-(1-\pi))}{(t_1+t_2-1)^2}\Big\{P_{22} - P_{12}\\ &\quad + (P_{11} - 2P_{12} + P_{22})\left[t_1(t_1+t_2-1) - \frac{1}{2}(t_1+t_2)\right]\\ &\quad + \frac{1}{2}(P_{11} - P_{22})(t_1+t_2)\Big\},\end{aligned}$$

where $(t_1,t_2) \in [0,\pi]\times[0,1-\pi]\cup[\pi,1]\times[1-\pi,1]$. The only interior point t^* that potentially satisfies $\nabla f(t^*) = 0$ is

$$t_1^* = \frac{P_{22} - P_{12}}{P_{11} + P_{22} - 2P_{12}} \qquad t_2^* = \frac{P_{11} - P_{12}}{P_{11} + P_{22} - 2P_{12}}.$$

However, because $t_1^* + t_2^* = 1$, the only intersection with the feasible region is at $t_1^* = \pi$, $t_2^* = 1 - \pi$, and thus $\tilde{f}$ can be maximized only on the boundary of the feasible region. Because all functions involved are monotone and convex, it is easy to check the boundary values; comparing all possible solutions shows that the unique maximizer of $\tilde{f}$ is $t_1 = 1,t_2 = 1$, or equivalently, $r_{11} = \pi,r_{22} = 1 - \pi$. This completes the proof of (b).

To prove Theorem 2, we first state a simpler version of the main theorem of Bickel and Chen from ref. 1. The theorem holds for a general K but to simplify notation we state it only for $K = 2$. This theorem allows $\rho_n = P[A_{ij} = 1] \to 0$. Letting $\mu_n = n^2\rho_n$, we can write W and $\tilde{W}$ (up to a multiplicative factor) in the form

$$Q(\boldsymbol{s},\boldsymbol{A}) = F\left(\frac{O(\mathbf{s},\mathbf{A})}{\mu_n},g(\boldsymbol{s})\right).$$

Further, following the proof of Theorem 1, it is easy to verify that

$$\frac{\mathbb{E}(\mathbf{O}(\boldsymbol{s},\boldsymbol{A})|\boldsymbol{c})}{\mu_n} = \boldsymbol{R}(\boldsymbol{s},\boldsymbol{c})\boldsymbol{P}\boldsymbol{R}^T(\boldsymbol{s},\boldsymbol{c}),$$

where $P = P_n/\rho_n$. Thus the population version of Q is $F(\boldsymbol{RPR}^T,\boldsymbol{R}\mathbf{1})$.

Theorem A1. *Suppose F, $\boldsymbol{P}$, and π satisfy the following conditions:*

- (C1) $F(\boldsymbol{RPR}^T,\boldsymbol{R}\mathbf{1})$ is uniquely maximized over $\mathscr{R} = \{\boldsymbol{R}:\boldsymbol{R} \geq 0, \boldsymbol{R}^T\mathbf{1} = (\pi,1-\pi)'\}$ by $\boldsymbol{R} = \mathbf{D}(\pi) \equiv diag(\pi,1-\pi)$, for all $(\pi,\boldsymbol{P})$ in an open set Θ.
- (C2) $\boldsymbol{P}$ has no identical columns.
- (C3) (a) F is Lipschitz in its arguments; (b) let $\boldsymbol{W} = \boldsymbol{D}(\pi)\boldsymbol{PD}(\pi)$. The directional derivatives $\frac{\partial^2 F}{\partial\epsilon^2}(\boldsymbol{M}_0 + \epsilon(\boldsymbol{M}_1 - \boldsymbol{M}_0),\boldsymbol{t}_1 + \epsilon(\boldsymbol{t}_1 - \boldsymbol{t}_0))|_{\epsilon=0+}$ are continuous in $(\boldsymbol{M}_1,\boldsymbol{t}_1)$ for all $(\boldsymbol{M}_0,\boldsymbol{t}0)$ in a neighborhood of $(\boldsymbol{W},\boldsymbol{C}(\pi))$, where $\boldsymbol{C}(\pi) = (\pi,1-\pi)^T$; (c) let $G(\boldsymbol{R},\boldsymbol{P}) = F(\boldsymbol{RPR}^T,\boldsymbol{R}1)$. Then on $\mathscr{R}$, $\frac{\partial G((1-\epsilon)\boldsymbol{D}(\pi)+\epsilon\boldsymbol{R},\boldsymbol{P})}{\partial\epsilon}|_{\epsilon=0+} < -C < 0$ for all $(\pi,\boldsymbol{P}) \in \Theta$.

If $\hat{\boldsymbol{c}}^{(n)}$ is the maximizer of $Q(\boldsymbol{s},\boldsymbol{A})$ and $\frac{\lambda_n}{\log n} \to \infty$, then, for all $(\pi,\boldsymbol{P}) \in \Theta$,

$$\limsup_{n\to\infty}\frac{P(\hat{\boldsymbol{c}}^{(n)} \neq \boldsymbol{c})}{\lambda_n} \leq -s_Q(\pi,\boldsymbol{P}) < 0.$$

Proof of Theorem 2. *Condition (C1) has already been checked in the proof of Theorem 1(b), condition (C2) holds trivially, and it is entirely straightforward to check condition (C3). Thus Theorem 2 follows from Theorem A1.*

1. . Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. *J Am Stat Assoc* 97:1090–1098.

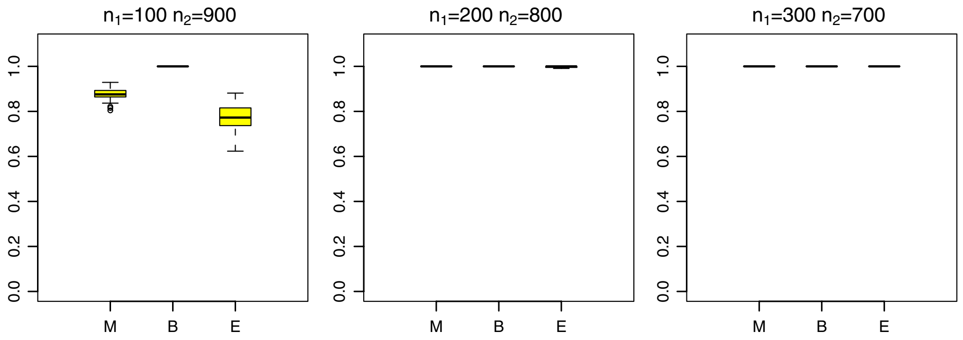

Fig. S1. Results for two communities with no background.

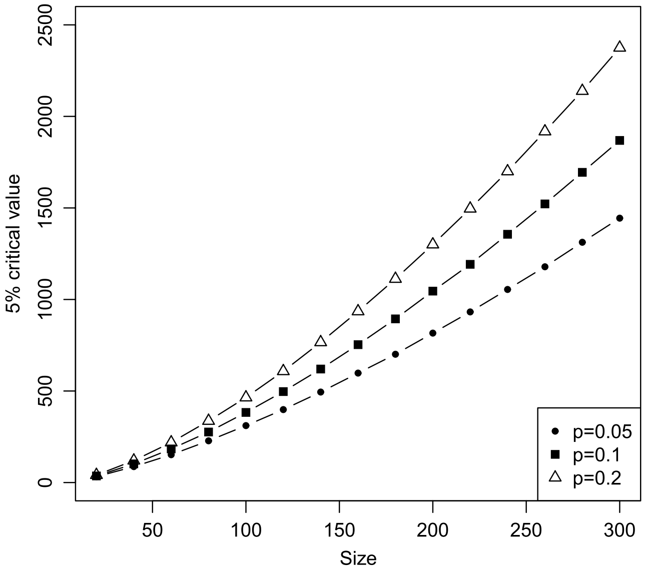

Fig. S2. Critical values for testing the hypothesis of an Erdos–Renyi graph.

Table S1. Critical values for the karate club

Split	1	2	3	4
Test statistic	108.7	60.7	40.8	26.0
1%	94.7	**66.7**	38.8	28.0
5%	88.0	**62.4**	37.0	23.0
10%	84.0	59.4	35.0	19.5

The first nonsignificant test at each level is shown in boldface.

Table S2. Critical values for the school friendship network

Split	1	2	3	4	5	6	7
Test statistic	651.9	423.8	318.4	196.3	136.8	93.9	48.0
1%	368.7	305.0	236.6	170.8	**137.0**	98.6	66.0
5%	352.8	290.5	220.0	157.8	127.5	92.0	**56.7**
10%	343.9	280.3	216.2	151.8	122.0	89.0	**54.0**

The first nonsignificant test at each level is shown in boldface.