Community extraction for social networks

Yunpeng Zhao, Elizaveta Levina[1], and Ji Zhu

Department of Statistics, University of Michigan, Ann Arbor, MI 48109

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved March 15, 2011 (received for review May 17, 2010)

Analysis of networks and in particular discovering communities within networks has been a focus of recent work in several fields and has diverse applications. Most community detection methods focus on partitioning the entire network into communities, with the expectation of many ties within communities and few ties between. However, many networks contain nodes that do not fit in with any of the communities, and forcing every node into a community can distort results. Here we propose a new framework that extracts one community at a time, allowing for arbitrary structure in the remainder of the network, which can include weakly connected nodes. The main idea is that the strength of a community should depend on ties between its members and ties to the outside world, but not on ties between nonmembers. The proposed extraction criterion has a natural probabilistic interpretation in a wide class of models and performs well on simulated and real networks. For the case of the block model, we establish asymptotic consistency of estimated node labels and propose a hypothesis test for determining the number of communities.

Understanding and modeling network structures have been a focus of attention in a number of diverse fields, including physics, biology, computer science, statistics, and social sciences. Applications of network analysis include friendship and social networks, marketing and recommender systems, the World Wide Web, disease models, and food webs, among others. A fundamental problem in the study of networks is community detection (see refs. 1–3 for comprehensive recent reviews). We focus here on undirected networks $N = (V,E)$, where V is the set of nodes and E is the set of edges, possibly weighted. The community detection problem is typically formulated as finding the partition $V = V_1 \cup \ldots \cup V_K$, which gives the "best" communities in some suitable sense. The node sets $V_1,\ldots,V_K$ are usually taken to be disjoint, although there is some recent work on detecting overlapping communities (4, 5).

The extensive physics and computer science literature on networks typically thinks of communities as tightly knit groups with many connections between the group members and relatively few connections between groups. Thus detection methods focus on maximizing links within communities while minimizing links between communities. This can be achieved either implicitly through an algorithmic approach (6) or explicitly by optimizing a criterion that measures the quality of a proposed partition over all possible partitions. These criteria include ratio cuts (7), normalized cuts (8), spectral clustering (9), and modularity (10); see ref. 3 for a review. All of these are symmetric criteria, in the sense that all potential communities play the same role. There are many examples of networks where such a requirement makes sense, for example, the college football games network (11), and yet some commonly studied networks clearly do not fit this framework. One such example is when there are nodes without strong connections to any communities, such as in the high school friendship network of ref. 12 discussed later in the paper. In such cases, the partitioning methods typically split up weakly connected nodes and group them together with tighter communities. There is not much work in the networks literature focusing on such a structure, even though in traditional multivariate clustering there are methods that allow for a diffuse "background" cluster [e.g., DBSCAN (13) and DenClue (14)].

Another class of community detection methods relies on a statistical model for the network to estimate the partition, typically by maximizing some form of the likelihood directly or employing Gibbs sampling. The models used for partitioning include the stochastic block model (15–17), a mixture model (18), univariate (19) and multivariate (20) latent variable models, and latent feature models (21); for a comprehensive review of statistical models of networks, see ref. 2. In particular, the stochastic block model, described in detail later in the paper, allows for any density of connections within communities and can in principle handle any combination of sparse and tight communities. The block model assumes a uniform distribution of edges within a block, and an extension was recently proposed to accommodate arbitrary degree distributions within the blocks [the degree-corrected block model (22)].

In this paper, we propose a framework for community extraction that looks for one community at a time. Intuitively, our extraction criterion, like other partition methods, looks for a "tight" group with more links within itself than to the rest of the network; but, unlike partition methods, it allows for the rest of the network to include an arbitrary mixture of tight and weak communities or sparsely connected background nodes. Mathematically, our criterion matches the definition of community in a large class of probability models on networks, outlined below, which can be thought of as a generalization of the block model with some parameter constraints. Our goal is to extract the tightest community, which we do by focusing on the edges within the candidate community and edges connecting it to the rest of the network, and ignoring edges within the rest of the network. The key feature of the extraction criterion is that it is not symmetric in the two sets into which we are splitting the network. In practice, such extraction can be used on its own or in conjunction with graph partitioning, for example, to identify community cores.

As an illustration, consider this toy example: Out of $n = 60$ nodes, 15 belong to a community where links between members form independently with probability 0.5. The links between members and the other 45 nodes and links between the other 45 nodes all form independently with probability 0.1. A partition into two communities using modularity and our community extraction method are shown in Fig. 1. Modularity has to balance tightness of the two communities, and as a result includes a number of background nodes in the community. Extraction, on the other hand, separates out the true community perfectly.

Finally, we briefly mention other related work. The core-periphery partition methods (23, 24) use a different criterion to separate a tight "core" from a sparse "periphery," whereas our criterion is designed to extract the tightest community regardless of whether the rest of the network is sparse or contains other communities. Local community detection (25, 26) looks for the tightest community around a given node, but not the globally tightest community. Finally, the hierarchical network model

Author contributions: Y.Z., E.L., and J.Z. designed research; Y.Z., E.L., and J.Z. performed research; Y.Z. analyzed data; and Y.Z., E.L., and J.Z. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: elevina@umich.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1006642108/-/DCSupplemental.

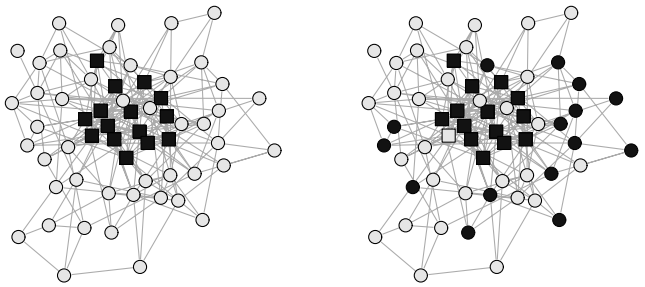

Fig. 1. Toy example: shapes represent the truth and colors represent results using our extraction method (*Left*) and modularity (*Right*).

based on ensembles of trees proposed in ref. 27 also has the feature that the strength of a community does not depend on unrelated nodes.

The Community Extraction Framework

First we introduce basic notation. A network $N = (V,E)$ with $|V| = n$ nodes can be represented by an $n \times n$ adjacency matrix $\mathbf{A} = [A_{ij}]$, where $A_{ij} > 0$ if there is an edge between nodes i and j and $A_{ij} = 0$ otherwise. If the edges have weights, the positive A_{ij}'s are the weights; if not, they are set to 1. For undirected networks, $\mathbf{A}$ is symmetric.

For simplicity, we start with partitioning into two sets V_1, V_2, where $V_1 \cap V_2 = \varnothing$ and $V = V_1 \cup V_2$. A naive way to partition a network is to minimize the total weight R of edges connecting V_1 and V_2 (the min-cut method), $R = \sum_{i \in V_1, j \in V_2} A_{ij}$. However, minimizing R yields a trivial solution of $V_1 = V$, so various adjustments have been proposed. For example, the ratio cut (7) avoids the trivial solution by minimizing $R/(|V_1| \cdot |V_2|)$, where $|V_1|$ and $|V_2|$ are the sizes of the two groups. The important point for us here is that in all these criteria the sets V_1 and V_2 can be interchanged.

The criterion we propose extracts one community at a time by looking for a set with a large number of links within itself and a small number of links to the rest of the network. The links within the complement of this set do not affect the value of the criterion. To emphasize the lack of symmetry, we denote the community to be extracted by S and its complement by S^c (rather than V_1 and V_2). Then we maximize the following extraction criterion over all possible S:

$$W(S) = \frac{O(S)}{|S|^2} - \frac{B(S)}{|S||S^c|}, \qquad [1]$$

where

$$O(S) = \sum_{i,j \in S} A_{ij}, \qquad B(S) = \sum_{i \in S, j \in S^c} A_{ij}.$$

The term $O(S)$ is twice the weight of the edges within S, and $B(S)$ represents connections between S and the rest of the network. Each term is normalized by the total number of possible edges in each case, which gives these quantities a natural interpretation as probability estimates, discussed further below. Note that we normalize the first term by $|S|^2$ rather than $|S|(|S| - 1)$, thus not explicitly excluding self-loops in order to be consistent with the probability models discussed below, but in practice this makes little difference. Subject to this small difference, our criterion can be described as the intracluster density minus the intercluster density; it is also related to conductance (3).

Criterion **1**, like the graph min-cut, does not explicitly guard against splitting off small communities. The trivial solution does not maximize W, but in a large sparse network a very small community can give a high value of W, because the second term will be made negligible by the large $|S^c|$. To avoid this situation, we make an adjustment in the spirit of the ratio cut and maximize the following criterion instead:

$$\tilde{W}(S) = |S||S^c| \left[\frac{O(S)}{|S|^2} - \frac{B(S)}{|S||S^c|} \right]. \qquad [2]$$

Because $|S||S^c|$ is maximized at $|S| = n/2$, this factor penalizes very small and very large communities and produces more balanced solutions. Empirically, we found the adjustment helps in sparse networks, but plays no role in dense networks. Later we show that asymptotically both criteria are consistent under an appropriate probability model.

Probabilistic Interpretation. The criterion (Eq. **1**) is motivated by the intuitive definition of community as a tightly knit group rather than by fitting a probability model to the network. However, it has a clear interpretation in the context of probability models on graphs. Consider a probability distribution P on symmetric adjacency matrices A that satisfies the following assumptions:

1. Each node i is associated with latent variables c_i and θ_i, where c_i is the community label and θ_i can contain any other node information. The labels c_i are independent and have a multinomial distribution with parameter $\pi = (\pi_1,\ldots,\pi_K)^T$.
2. For any pair $(i\,j)$ and $(i'\,j')$ that have no nodes in common, A_{ij} and $A_{i'j'}$ are independent given the labels $\mathbf{c}$.
3. $E(A_{ij}|\mathbf{c}) = P_{c_i c_j}$ for all i, j.

Assumption 1 is similar to the block model, except it allows for additional latent variables θ_i. An assumption related to assumption 2 was proposed in ref. 28, stipulating independence conditional on other edges. In assumption 3, the expectation integrates out θ, so the left-hand side is always a function of $\mathbf{c}$; the assumption stipulates that it can depend only on the two labels c_i and c_j, as in the block model.

Let $\mathbf{s}$ be an arbitrary label assignment, let $O_{ab}(\mathbf{s}) = \sum_{ij} A_{ij} I(s_i = a, s_j = b)$, and $R_{ab}(\mathbf{s},\mathbf{c}) = n^{-1} \sum_{i=1}^{n} I(s_i = a, c_i = b)$. Suppose $R(\mathbf{s},\mathbf{c}) \xrightarrow{P} R$ as $n \to \infty$. Note that R satisfies $\mathbf{1}^T R = \pi$. Then the following holds:

Theorem 1. *(a) Under assumptions 1–3, as* $n \to \infty$,

$$\frac{O_{ab}(\mathbf{s})}{n^2} \xrightarrow{P} (RPR^T)_{ab}.$$

(b) Assume that $K = 2$, *and, without loss of generality,* $P_{11} \geq P_{22}$. *Then* $W(S) \xrightarrow{P} f(R,P)$, $n^{-2}\tilde{W}(S) \xrightarrow{P} \tilde{f}(R,P)$. *Further, if* $P_{11} + P_{22} > 2P_{12}$, *then both* f *and* $\tilde{f}$ *are maximized by* $R = \mathrm{diag}(\pi)$, *under the constraint* $\mathbf{1}^T R = \pi$. The proof and the expressions for functions f and $\tilde{f}$ can be found in the *SI Text*. The theorem establishes that our extraction criterion is a natural data-based approximation to a population criterion that is maximized by the correct label assignment. Note that even though in part (b) we have $K = 2$ because it applies to a single extraction step, this does not mean that extraction fails if $K > 2$ (see more on this in Example 4). Next, we give several examples of models that satisfy our conditions.

Example 1: The stochastic block model corresponds to $\theta_i = \text{const}$ and independent edges. Under the block model, each node is first assigned to one of the K blocks independently of other nodes. Then, conditional on $\mathbf{c}$, edges are generated independently with probabilities $P[A_{ij} = 1 | c_i = a, c_j = b] = P_{ab}$, which evidently satisfies assumptions 2 and 3. The block model is fully specified by its

parameters π and the $K \times K$ symmetric matrix P. Thus part (a) of the theorem applies. Part (b) applies if we additionally assume $K = 2$ and $P_{11} + P_{22} > 2P_{12}$.

Example 2: The degree-corrected block model. This generalization of the block model has been recently introduced to allow for different expected node degrees within blocks (22). There the labels $\boldsymbol{c}$ are treated as fixed and fixed node-specific parameters θ are introduced to reflect the nodes' individual tendencies to form ties. We use exactly the same model but treat $\boldsymbol{c}$ and θ as random, with $\boldsymbol{c}$ satisfying our assumption 1. Conditional on $\boldsymbol{c}$ and θ, A_{ij} are independent Poisson random variables with mean $\theta_i\theta_j P_{c_i c_j}$. The relaxation of the Bernoulli distribution to the Poisson is primarily for ease of technical derivations in ref. 22 and has few practical consequences. In either distributional form, the model satisfies assumptions 1–3 and part (a) holds. The model can be further constrained to satisfy conditions for (b).

Example 3: General exchangeable distributions. It is well known that any probability distribution invariant to node permutations on the matrix A can be written in the form $A_{ij} = h(\mu,\xi_i,\xi_j,\lambda_{ij})$, where μ, ξ_i's, and λ_{ij} are independent and identically distributed (i.i.d.), and h is symmetric in its second and third arguments (for details, see, e.g., ref. 29). The equivalent of i.i.d. sequences in this class has the form $A_{ij} = h(\xi_i,\xi_j,\lambda_{ij})$. Let $\xi_i = (c_i,\theta_i)$, where c_i is the class label, and θ_i is other node information that can be correlated with c_i. If the distribution of i.i.d. pairs (c_i,θ_i) is such that the marginal distribution of $\boldsymbol{c}$ is multinomial, this model satisfies conditions 1–3. One concrete example of such a model is the latent position cluster model of ref. 20. Again, the general model can be further constrained to satisfy (b).

Example 4: Consider a block model with true $K = 3$ where two blocks have been merged under one label. Let c_i (merged labels) take values 1 and 2 with probabilities π_c and $1 - \pi_c$, and let θ_i (true labels) have the distribution $P(\theta_i = 1|c_i = 1) = 1$, $P(\theta_i = 2|c_i = 2) = \pi_\theta$, and $P(\theta_i = 3|c_i = 2) = 1 - \pi_\theta$. If Q is the 3×3 matrix corresponding to the true block model, and Π is the 2×3 matrix giving the joint probability distribution of the pair (c,θ), let $P = \Pi Q \Pi^T$. Then the population version of our criterion is maximized by extracting the first community as long as $P_{11} > P_{12}$, $P_{11} > P_{22}$, $P_{11} + P_{22} > 2P_{12}$, which ensures the condition of part (b) holds. This is true, for example, if $Q_{11} > Q_{ij}$ for all $(ij) \neq (1,1)$ and $Q_{ij} \geq Q_{1k}$ for all $1 < i,j,k \leq 3$. In other words, if the ties within the first community are stronger than its ties to the mixture of second and third, the criterion will extract it correctly. Note that this is exactly the situation in the counterexample to consistency of modularity given by ref. 29. In that example, both profile likelihood and extraction (in two stages) are consistent, but modularity is not.

Maximizing the Extraction Criterion. To maximize the extraction criteria, we use a local optimization technique based on label switching known as tabu search (30, 31). The key idea of tabu search is that once a node label has been switched, it cannot be switched again for the next T iterations (the node has "tabu" status) This guards against being trapped in a local maximum. The algorithm starts from an initial value and examines all current nontabu nodes in order. If the current value of the global maximum can be improved, the node label is switched, its status changed to tabu, and the algorithm returns to node 1. If no node can be switched to improve the global maximum, the node that gives the largest increase in the current criterion value is switched, and if no increase is possible, the node that gives the smallest decrease is switched. The algorithm is run for a prescribed number of iterations, and the best solution seen in the course of these iterations is taken to be the final solution. Note that the value of $\tilde{W}$ can be updated efficiently in $O(n)$ operations for a single label switch. To help guard against local maxima, we run the algorithm for a number of random starting values and random orderings of nodes. Each run will converge to a local maximum, and although the algorithm is not guaranteed to find the global maximum, we have not encountered any problems with local maxima in either simulations or real data examples.

The Stochastic Block Model Case

If we focus on the special case of the stochastic block model, we can obtain additional results on the properties of the extraction criterion. First, we show that the estimated node labels are asymptotically consistent, using the recent results of ref. 29. Second, we describe a hypothesis test that can be performed at every sequential split to determine whether the remainder after extraction contains any more communities.

Asymptotic Consistency. We consider asymptotic consistency of label assignments by the extraction method as the number of nodes $n \to \infty$. If P does not change with n, the network becomes very dense as n grows, so we allow P_n to depend on n and write $P_n = \rho_n P$, where $\rho_n = P[A_{ij} = 1] \to 0$ is the probability of an edge between arbitrary nodes i and j. The expected node degree $\lambda_n = n\rho_n$ becomes the natural parameter to control as $n \to \infty$.

Bickel and Chen (29) developed a general framework for checking whether a community-finding criterion can recover the true node labels as $n \to \infty$, under the assumptions of the block model and $\lambda_n / \log n \to \infty$. The latter may not be universally applicable, but in many examples the degree does grow with n, and faster than logarithmic growth is a very mild requirement. Further details are given in *SI Text*; briefly, the main condition is that the proposed criterion is maximized by the true label assignment when all the sample quantities in the criterion are replaced by their population equivalents. This can be viewed as a special case of the theory of minimum contrast estimation (32).

We focus on checking one-step consistency for the case $K = 2$ (one extracted community plus the rest of the network). The matrix P is 2×2 with three unique parameters P_{11}, P_{22}, P_{12}, and the vector of class probabilities $\{\pi, 1 - \pi\}$ is determined by the single parameter π. Let $\hat{\boldsymbol{c}}^{(n)}$ be the maximizer of criterion **1**, and $\tilde{\boldsymbol{c}}^{(n)}$ of criterion **2**. It turns out that the adjustment factor of $|S||S^c|$ has no effect in the limit, and both criteria are asymptotically consistent, as shown in the following theorem:

Theorem 2. *Suppose* $\lambda_n / \log n \to \infty$, $P_{11} > P_{12}$, $P_{11} > P_{22}$ *and* $P_{11} + P_{22} > 2P_{12}$, *and* $\boldsymbol{c}$ *are the true labels. Then*

$$P[\hat{\boldsymbol{c}}^{(n)} = \boldsymbol{c}] \to 1 \quad \textit{and} \quad P[\tilde{\boldsymbol{c}}^{(n)} = \boldsymbol{c}] \to 1.$$

Note that the simplest case of our toy example (one community with other weakly connected nodes, $P_{12} = P_{22} = p$) is covered by the theorem as long as $P_{11} > p$. The proof is given in *SI Text*.

Determining the Number of Communities. The full extraction procedure consists of sequentially applying criterion **2**: We extract a community and apply the extraction again to its complement. Ideally, the user would have information on the true or desired number of communities K to be extracted. In the absence of such prior information, determining the number of communities in a network is an open problem, and a rigorous solution would require fully specifying a statistical model. Here we propose a hypothesis test that can be used under the assumption of the block model. In this case, we need to test the null hypothesis H_0: $K = 1$ against the alternative H_a: $K \geq 2$ for the subgraph induced by the nodes in the remainder. H_0 means that the remainder is an Erdos–Renyi (ER) graph, where all edges form independently with the same probability. This approach has an

analogue in clustering (33). Deriving the null distribution of our criterion analytically is difficult (currently work in progress); however, the test can always be conducted by simulation, in the manner of a permutation test. Specifically, we simulate N ER graphs with the probability of an edge matching that of the graph being tested, and maximize the criterion $\tilde{W}$ for each graph. The empirical distribution of these N values of $\tilde{W}$ approximates the distribution of $\tilde{W}$ under the null hypothesis. If the value of $\tilde{W}$ for the proposed split on the real graph is larger than $100(1-\alpha)$th percentile of these N values for some standard choice of α, we reject the null hypothesis and apply extraction to the remainder; if not, we stop. Note that this procedure is an analogue of backward elimination in classical linear regression. As with backward elimination, it is more important to look for a gap in p values than at the exact levels of significance; however, if those are of interest, the overall family error rate can be controlled with a multiple testing correction. For example, the Bonferroni correction, which is conservative, guarantees the overall error rate of no more than α_0 if each test is conducted at the level $\alpha_0/T_{\max}$, where $T_{\max}$ is the maximum number of tests allowed. Note that this approach generalizes easily to the degree-corrected block model from Example 2, and specific applications may suggest other appropriate null models. *SI Text* gives further details.

Numerical Evaluation

We compare our extraction criterion $\tilde{W}$ to modularity, fitting the block model via Markov chain Monte Carlo (MCMC) using the BLOCKS program of ref. 17, and the latent position cluster model of ref. 20 using the R package latentnet. In this section, we compare performance of extraction, modularity, and MCMC for several simulated scenarios. The latentnet package cannot handle networks of the size used in simulations, but we do include it in comparisons on real data in the next section. The modularity partition into two communities is computed from the approximate eigenvector solution described in ref. 34. For partitioning into more than two communities, we follow the sequential splitting approach in ref. 1. Following the code kindly shared with us by Mark Newman (University of Michigan, Ann Arbor, MI), we use only the eigendecomposition-based approximation for subsequent splits as an initial value for a search, which substantially improves performance. For MCMC, we run 10^5 iterations for burn-in and another 10^5 until convergence; for real data, which has fewer nodes, the number of iterations is reduced by half. BLOCKS does not assign a class to a node if its label varies too much across MCMC runs; those nodes were excluded from analysis. We compare the methods using the adjusted Rand index (35), a widely used measure for comparing partitions. The adjustment is done so that the expected value of the index is 0 for comparing two random partitions, and higher values indicate better agreement. The results in this section are presented via boxplots of the adjusted Rand index from 50 replications of each simulated scenario. All simulations are performed with $n = 1{,}000$ nodes.

First, we check the standard case of a network consisting of two tight communities. Briefly, all methods do well here, except when community sizes are very unbalanced, in which case the block model fitted by MCMC does best, closely followed by modularity and extraction (see *SI Text*).

Next, we consider a network consisting of one community and weakly connected background, generated from the block model with $P_{12} = P_{22} = 0.05$. We consider community sizes $n_1 = 100$ and 200, and $P_{11} = 0.15$ and 0.2. Fig. 2 shows that in this scenario extraction does best in all cases, but as P_{11} and n_1 increase, modularity and the MCMC fit to the block model improve.

A more interesting situation is two tight communities with similar densities and a weakly connected background. We generate this from a block model with $K = 3$, and $P_{31} = P_{32} = P_{33} = 0.05$. We consider balanced ($n_1 = n_2 = 200$) and unbalanced ($n_1 = 100, n_2 = 200$) communities, and two levels of community strength ($P_{11} = 0.15$, $P_{22} = 0.12$, and $P_{11} = 0.20$, $P_{22} = 0.16$). For a fair comparison, we extract two communities by our method, but partition into three parts by modularity to allow for background, and fit the block model with $K = 3$. The results are presented in Fig. 3, *Top*. Extraction performs the best, with MCMC catching up as the signal gets stronger. Modularity improves slightly for denser communities, but even though it usually separates the two communities correctly, it tends to add the background nodes to them, resulting in poor overall Rand index.

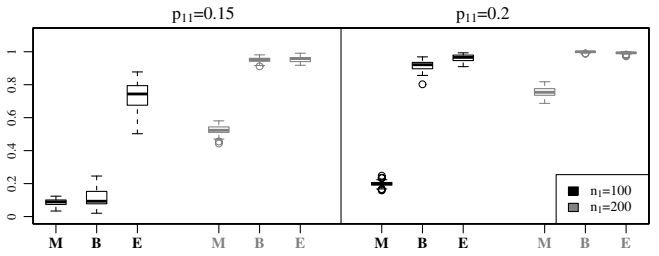

Fig. 2. Adjusted Rand index for one community with background. M: Modularity; B: Block model fitted via MCMC; E: Extraction.

Our final simulation is designed to test the methods in the presence of "hubs," a phenomenon present in many datasets, for example, in the karate club dataset described in the next section. For ease of comparison, we use the same setup as in the previous simulation, except in each community we double the degrees of the 10 highest-degree nodes by adding edges to randomly selected nodes within the community and delete the same number of edges at random from all other edges *within* the community, thus keeping the total number of edges within the community constant. Comparing the results for this case (Fig. 3, *Bottom*) with those for the block model (*Top*), we see that results for modularity are virtually unchanged and for extraction just slightly worse, whereas the MCMC-fitted block model deteriorates substantially. This is because in many cases (though not all, which explains high variance) it groups all high-degree nodes together into a single block. This agrees with other empirical observations on the block model in the presence of hubs (22, 29). We conjecture that the additional constraint $P_{11} > P_{12}$ that extraction enforces explains its relative robustness in this case.

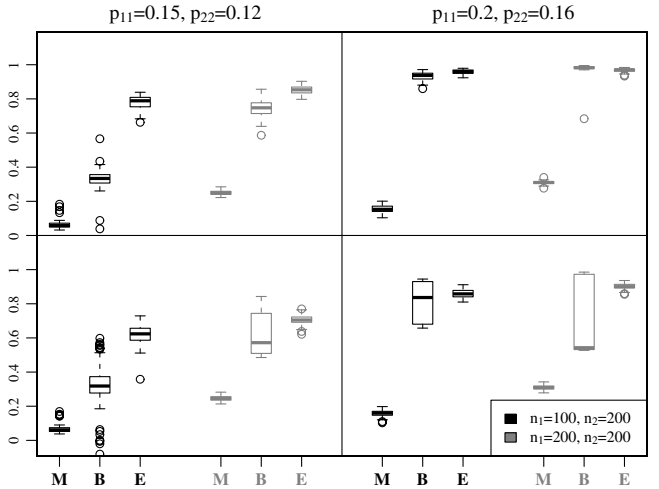

Fig. 3. Box plots of adjusted Rand index for two communities with background. M: Modularity; B: Block model fitted via MCMC; E: Extraction. (*Top*) Data generated from the block model; (*Bottom*) doubling the degree for the highest-degree nodes.

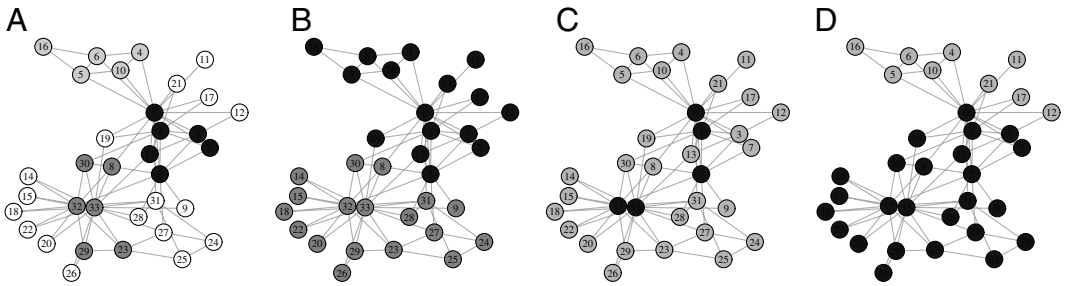

Fig. 4. Results for the karate club network: (*A*) extraction, (*B*) modularity, (*C*) the block model fitted via MCMC, and (*D*) the latent position cluster model.

Examples

The Karate Club Network. Our first example is a well-known network representing friendships between 34 members of a karate club (36). This club later split into two parts following a disagreement between an instructor (node 0) and an administrator (node 33), and these two groups are used as the "ground truth" communities in benchmark studies. Modularity partitions this network into exactly the true factions (1). The extraction approach can be used to supplement this division with more information by identifying the "cores" of each faction. Fig. 4 shows the first three extracted communities—the cores of two factions, and a small tight subcommunity. Note that the members of the two factions are never grouped together. The hypothesis test under the block model suggests that the first extraction split is strongly significant, and the second is marginal (see *SI Text*). The block model fitted with MCMC with $K = 2$ puts four nodes with the highest degrees into one block, and everything else into another block, a result independently obtained in ref. 29 using profile likelihood. The latent position cluster model splits one of the communities into what might be described as core and periphery, and groups the core together with the other community.

The Political Books Network. The nodes in this network (34) are 105 recent political books with links representing pairs of books reported by Amazon as "frequently bought together." Following ref. 34, we show the modularity solution with node colors representing the components of the leading principal vector of the modularity matrix in Fig. 5*A*. These values result from relaxing the labels from ± 1 to real-valued, and the modularity partition is computed from their signs. The node colors can be interpreted to represent a book's position on the political spectrum, with blue being the most liberal and red the most conservative (34). Fig. 5 shows that in addition to a few clear "red" and "blue" books, many nodes are in fact "purple" and may not clearly belong to either the Left or the Right. From the colors alone (i.e., component magnitudes), it is not clear how to separate out the blue and the red from the more centrist purple. Fig. 5 shows the first two extracted communities, which clearly correspond to the cores of the Left and the Right. The block model and the latent position cluster model were fitted with $K = 3$ (to allow for purple) and do reasonably well in this case, although the partitions are not nearly as clean as the one obtained by extraction; also, the block model left some nodes unclassified (shown in light blue) due to lack of agreement between MCMC runs.

The School Friendship Network. This dataset is a school friendship network compiled from the National Longitudinal Study of Adolescent Health (see ref. 12 for more information). The directed links in this dataset connect students to their self-reported friends, with weights representing friendship strength. Here we analyze the network of school 1 from this dataset, converting the data to an undirected network by averaging the weights on the two edges connecting each pair of nodes. The resulting network with 71 nodes is shown in Fig. 6, with colors representing grades, along with results of extraction, modularity, the block model, and the latent position cluster model, all with six groups to match the number of grades. Modularity and extraction agree fairly well with each other and with the grades, but extraction has the ability to leave out sparsely connected nodes, including the two isolated nodes. The block model and the latent position cluster model partly agree with the grades, but also produce some communities that consist of disjoint parts.

Summary and Discussion

We have proposed a previously undescribed framework for analysis of social networks, which extracts tight communities out of the network one at a time, allowing for complex structures in the remainder of the network. In the examples we considered, it offers an additional insight into the network structure and can be used as either an alternative or a complement to network partitioning. The probabilistic interpretation of extraction shows that it is applicable to a wide range of models. Although we have obtained good results with the tabu search, we would like to be able to formulate the extraction criterion as an eigenvalue problem, which is work in progress. More analytical work is needed on the stopping criterion outside of the block model case, although this is a general question common to all community detection methods. A related issue is assessing the quality of a proposed extraction/partition; one solution based on robustness to perturbations was proposed in ref. 37, which can be applied to extraction as well. Extensions to networks with more complex community structures, such as overlapping communities or disassortative networks, are also a topic for future work.

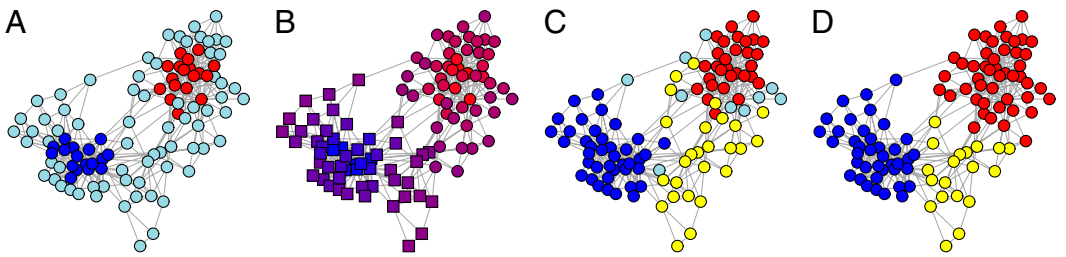

Fig. 5. Results for the political books network: (*A*) extraction, (*B*) modularity, (*C*) the block model fitted via MCMC, and (*D*) the latent position cluster model. For modularity, node colors represent the components of the principal eigenvector of the modularity matrix, and node shapes represent the partition.

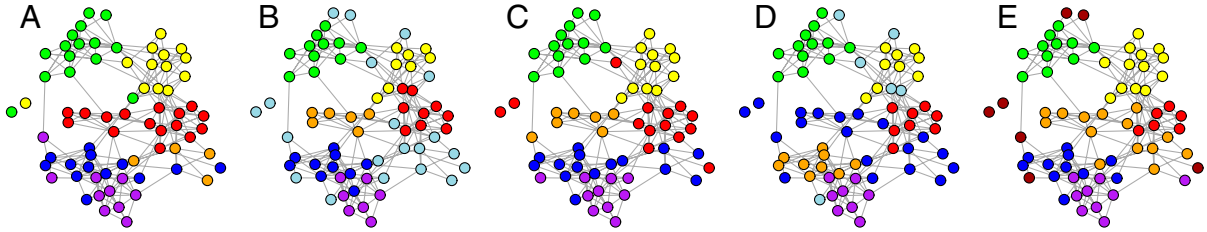

Fig. 6. The school friendship network. (*A*) grades, (*B*) extraction, (*C*) modularity, (*D*) the block model fitted via MCMC, and (*E*) the latent position cluster model.

Acknowledgments

We thank the editor and two referees for many helpful comments and Mark Newman for constructive feedback and sharing his code. This work is partially supported by National Science Foundation Grants DMS-0805798 (to E.L.) and DMS-0748389 (to J.Z.).

1. Newman MEJ (2006) Modularity and community structure in networks. *Proc Natl Acad Sci USA* 103:8577–8582.
2. Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM (2009) A survey of statistical network models. *Found Trends Mach Learn* 2:129–233.
3. Fortunato S (2010) Community detection in graphs. *Phys Rep* 486:75–174.
4. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435:814–818.
5. Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008) Mixed membership stochastic blockmodels. *J Mach Learn Res* 9:1981–2014.
6. Newman MEJ (2004) Detecting community structure in networks. *Eur Phys J B* 38:321–330.
7. Wei Y-C, Cheng C-K (1989) Toward efficient hierarchical designs by ratio cut partitioning. *Proceedings of the IEEE International Conference on Computer Aided Design* (IEEE, New York), pp 298–301.
8. Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE T Pattern Anal* 22:888–905.
9. Ng A, Jordan M, Weiss Y (2001) On spectral clustering: Analysis and an algorithm. *Neural Information Processing Systems 14*, eds T Dietterich, S Becker, and Z Ghahramani (MIT Press, Cambridge, MA), pp 849–856.
10. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69:026113.
11. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99:7821–7826.
12. Hunter DR, Goodreau SM, Handcock MS (2008) Goodness of fit of social network models. *J Am Stat Assoc* 103:248–258.
13. Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)* (AAAI Press, Menlo Park, CA), pp 226–231.
14. Hinneburg A, Keim DA (1998) An efficient approach to clustering in large multimedia databases with noise. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)* (AAAI Press, Menlo Park, CA), pp 58–65.
15. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. *Soc Networks* 5:109–137.
16. Snijders T, Nowicki K (1997) Estimation and prediction for stochastic block-structures for graphs with latent block structure. *J Classif* 14:75–100.
17. Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. *J Am Stat Assoc* 96:1077–1087.
18. Newman MEJ, Leicht EA (2007) Mixture models and exploratory analysis in networks. *Proc Natl Acad Sci USA* 104:9564–9569.
19. Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. *J Am Stat Assoc* 97:1090–1098.
20. Handcock MD, Raftery AE, Tantrum JM (2007) Model-based clustering for social networks. *J R Stat Soc A-G* 170:301–354.
21. Hoff PD (2007) Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in Neural Information Processing Systems*, (MIT Press, Cambridge, MA), Vol 19.
22. Karrer B, Newman MEJ (2011) Stochastic blockmodels and community structure in networks. *Phys Rev E* 83:016107.
23. Borgatti SP, Everett MG (2000) Models of core/periphery structures. *Soc Networks* 21:375–395.
24. Everett MG, Borgatti SP (2000) Peripheries of cohesive subsets. *Soc Networks* 21:397–407.
25. Flake GM, Lawrence S, Giles CL, Coetzee FM (2002) Self-organization and identification of web communities. *IEEE Computer* 35:66–71.
26. Clauset A (2005) Finding local community structure in networks. *Phys Rev E* 72:026132.
27. Clauset A, Moore C, Newman MEJ (2008) Hierarchical structure and the prediction of missing links in networks. *Nature* 453:98–101.
28. Frank O, Strauss D (1986) Markov graphs. *J Am Stat Assoc* 81:832–842.
29. Bickel PJ, Chen A (2009) A nonparametric view of network models and Newman-Girvan and other modularities. *Proc Natl Acad Sci USA* 106:21068–21073.
30. Beasley JE (1998) Heuristic algorithms for the unconstrained binary quadratic programming problem. Technical report, Management School, Imperial College, London, UK.
31. Glover FW, Lagunas M (1997) *Tabu Search* (Kluwer Academic, Boston).
32. Bickel PJ, Doksum KA (2007) *Mathematical Statistics: Basic Ideas and Selected Topics*, (Pearson Prentice Hall, Upper Saddle River, NJ), 2nd Ed, Vol 1.
33. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a dataset via the gap statistic. *J R Stat Soc A* 63:411–423.
34. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E* 74:036104.
35. Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2(1):193–218.
36. Zachary WW (1977) An information flow model for conflict and fission in small groups. *J Anthropol Res* 33:452–473.
37. Karrer B, Levina E, Newman MEJ (2008) Robustness of community structure in networks. *Phys Rev E* 77:046119.