A nonparametric view of network models and Newman–Girvan and other modularities

Peter J. Bickel[a,1] and Aiyou Chen[b]

[a]University of California, Berkeley, CA 94720; and [b]Alcatel-Lucent Bell Labs, Murray Hill, NJ 07974

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved October 13, 2009 (received for review July 2, 2009)

Prompted by the increasing interest in networks in many fields, we present an attempt at unifying points of view and analyses of these objects coming from the social sciences, statistics, probability and physics communities. We apply our approach to the Newman–Girvan modularity, widely used for "community" detection, among others. Our analysis is asymptotic but we show by simulation and application to real examples that the theory is a reasonable guide to practice.

modularity | profile likelihood | ergodic model | spectral clustering

The social sciences have investigated the structure of small networks since the 1970s, and have come up with elaborate modeling strategies, both deterministic, see Doreian et al. (1) for a view, and stochastic, see Airoldi et al. (2) for a view and recent work. During the same period, starting with the work of Erdös and Rényi (3), a rich literature has developed on the probabilistic properties of stochastic models for graphs. A major contribution to this work is Bollobás et al. (4). On the whole, the goals of the analyses of ref. 4, such as emergence of the giant component, are not aimed at the statistical goals of the social science literature we have cited.

Recently, there has been a surge of interest, particularly in the physics and computer science communities in the properties of networks of many kinds, including the Internet, mobile networks, the World Wide Web, citation networks, email networks, food webs, and social and biochemical networks. Identification of "community structure" has received particular attention: the vertices in networks are often found to cluster into small communities, where vertices within a community share the same densities of connecting with vertices in the their own community as well as different ones with other communities. The ability to detect such groups can be of significant practical importance. For instance, groups within the worldwide Web may correspond to sets of web pages on related topics; groups within mobile networks may correspond to sets of friends or colleagues; groups in computer networks may correspond to users that are sharing files with peer-to-peer traffic, or collections of compromised computers controlled by remote hackers, e.g. botnets (5). A recent algorithm proposed by Newman and Girvan (6), that maximizes a so-called "Newman–Girvan" modularity function, has received particular attention because of its success in many applications in social and biological networks (7).

Our first goal is, by starting with a model somewhat less general than that of ref. 4, to construct a nonparametric statistical framework, which we will then use in the analysis, both of modularities and parametric statistical models. Our analysis is asymptotic, letting the number of vertices go to ∞. We view, as usual, asymptotics as being appropriate insofar as they are a guide to what happens for finite n. Our models can, on the one hand, be viewed as special cases of those proposed by ref. 4, and on the other, as encompassing most of the parametric and semiparametric models discussed in Airoldi et al. (2) from a statistical point of view and in Chung and Lu (8) for a probabilistic one. An advantage of our framework is the possibility of analyzing the properties of the Newman–Girvan modularity, and the reasons for its success and occasional failures. Our approach suggests an alternative modularity which is, in principle, "fail-safe" for rich enough models. Moreover, our point of view has the virtue of enabling us to think in terms of "strength of relations" between individuals not necessarily clustering them into communities beforehand.

We begin, using results of Aldous and Hoover (9), by introducing what we view as the analogues of arbitrary infinite population models on infinite unlabeled graphs which are "ergodic" and from which a subgraph with n vertices can be viewed as a piece. This development of Aldous and Hoover can be viewed as a generalization of deFinetti's famous characterization of exchangeable sequences as mixtures of i.i.d. ones. Thus, our approach can also be viewed as a first step in the generalization of the classical construction of complex statistical models out of i.i.d. ones using covariates, information about labels and relationships.

It turns out that natural classes of parametric models which approximate the nonparametric models we introduce are the "blockmodels" introduced by Holland, Laskey and Leinhardt ref. 10; see also refs. 2 and 11, which are generalizations of the Erdös–Rényi model. These can be described as follows.

In a possibly (at least conceptually) infinite population (of vertices) there are K unknown subcommunities. Unlabeled individuals (vertices) relate to each other through edges which for this paper we assume are undirected. This situation leads to the following set of probability models for undirected graphs or equivalently the corresponding adjacency matrices $\{A_{ij} : i,j \geq 1\}$, where $A_{ij} =$ 1 or 0 according as there is or is not an edge between i and j.

1. Individuals independently belong to community j with probability π_j, $1 \leq j \leq K$, $\sum_{j=1}^{K} \pi_j = 1$.
2. A symmetric $K \times K$ matrix $\{P_{kl} : 1 \leq k,l \leq K\}$ of probabilities is given such that P_{ab} is the probability that a specific individual i relates to individual j given that $i \in a, j \in b$. The membership relations between individuals are established independently. Thus $1 - \sum_{1 \leq a,b \leq K} \pi_a \pi_b P_{ab}$ is the probability that there is no edge between i and j.

The Erdös–Rényi model corresponds to $K = 1$.

We proceed to define Newman–Girvan modularity and an alternative statistically motivated modularity. We give necessary and sufficient conditions for consistency based on the parameters of the block model, properties of the modularities, and average degree of the graph. By consistency we mean that the modularities can identify the members of the block model communities perfectly. We also give examples of inconsistency when the conditions fail. We then study the validity of the asymptotics in a limited simulation and apply our approach to a classical small example, the Karate Club and a large set of Private Branch Exchange (PBX) data. We conclude with a discussion and some open problems.

Author contributions: P.J.B. and A.C. performed research and analyzed data.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: bickel@stat.berkeley.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0907096106/DCSupplemental.

Random Graph Models

Consider any probability distribution $\mathbb{P}$ on an infinite undirected graph, or equivalently a probability distribution on the set of all matrices $||A_{ij} : i,j \geq 1||$ where $A_{ij} = 1$ or 0, $A_{ij} = A_{ji}$ for all i,j pairs, and $A_{ii} = 0$ for all i, thus excluding self relation. If the graph is unlabeled, it is natural to restrict attention to $\mathbb{P}$ such that $||A_{\sigma_i\sigma_j}|| \sim \mathbb{P}$ for any permutation σ of $\{1,2,3,\ldots\}$. Hoover (see ref. 9) has shown that all such probability distributions can be represented as,

$$A_{ij} = g(\alpha, \xi_i, \xi_j, \lambda_{ij})$$

where $\alpha, \{\xi_i\}$ and $\{\lambda_{ij}\}$ are i.i.d. $U(0,1)$ variables and

$$g(u,v,w,z) = g(u,w,v,z)$$

for all u,v,w,z. The variables ξ correspond to latent variables, λ being completely individual specific, ξ generating relations between individuals and α a mixture variable which is unidentifiable even for an infinite graph. Note that g is unidentifiable and the ξ and λ could be put on another scale, e.g. Gaussian. We note that, this point of departure was also recently proposed by Hoff (12) but was followed to a different end.

It is clear that the distributions representable as,

$$A_{ij} = g(\xi_i, \xi_j, \lambda_{ij}) \quad \textbf{[1]}$$

where $\lambda_{ij} = \lambda_{ji}$, are the extreme points of this set and play the same role as sequences of i.i.d. variables play in de Finetti's theorem. Since given ξ_i and ξ_j, the λ_{ij} are i.i.d., these distributions are naturally parametrized by the function

$$h(u,v) \equiv \mathbb{P}[A_{ij} = 1 | \xi_i = u, \xi_j = v].$$

As Diaconis and Janson (13) point out $h(.,.)$ does not uniquely determine $\mathbb{P}$ but if h_1 and h_2 define the same $\mathbb{P}$, then there exists $\varphi : [0,1] \to [0,1]$ which is: measure preserving, i.e. such that $\varphi(\xi_1)$ has a $U(0,1)$ distribution; and $h_1(u,v) = h_2(\varphi(u), \varphi(v))$.

Given any h corresponding to $\mathbb{P}$, let

$$\mathbb{P}[X_{ij} = 1 | \xi_i = u] = g(u) = \int_0^1 h(u,v)dv.$$

It is well known (see section 10 of ref. 14) that there exists a measure preserving φ_g such that, $g(\varphi_g(v))$ is monotone non decreasing.

Define

$$h_{CAN}(u,v) = h(\varphi_g(u), \varphi_g(v)).$$

We claim that

$$g_{CAN}(u) \equiv \int_0^1 h_{CAN}(u,v)dv = F^{-1}(u)$$

where F is the cdf of $g_{CAN}(\xi_i)$, and h_{CAN} is unique up to sets of measure 0. To see this note that if h corresponds to $\mathbb{P}$ and $g(u) \equiv \int_0^1 h(u,v)dv$ is non decreasing, then since F is determined by $\mathbb{P}$ only, $g(u) = F^{-1}(u)$. But $g(\varphi_g(u)) = g_{CAN}(u)$ and $\varphi_g(u) = g^{-1}g_{CAN}(u) = u$.

There is a reparametrization of h_{CAN} (we drop the *CAN* subscript in the future) which enables us to think of our model in terms more familiar to statisticians.

Let

$$\rho = \mathbb{P}(\text{Edge}) = \int_0^1 \int_0^1 h(u,v)dudv.$$

Then the conditional density of (ξ_i, ξ_j) given that there is an edge between i and j is $w(u,v) = \rho^{-1}h(u,v)$. This parametrization also permits us to decouple $\rho \propto \mathbf{E}(\text{Degree})$ of the graph from the inhomogeneity structure. It is natural finally to let ρ depend on n but $w(\cdot,\cdot)$ to be fixed. If $\lambda_n \equiv \mathbb{E}(\text{Degree}) \to \infty$, we have what we may call the "dense graph" limit. If $\lambda_n = \Omega(1)$, we are in the case most studied in probability theory where, for instance, $\lambda_n = 1$ is the threshold at which the so called "giant component" appears. This is the situation Bollobas et al. focus on.

As we have noted, block models are of this type. Here we can think of the reparametrization as being ρ,π, $||S_{ab}|| = ||\rho^{-1}P_{ab}||$, or $||W_{ab}|| \equiv ||\rho^{-1}P_{ab}\pi_a\pi_b||$. The models studied by Chung and others (8) given by,

$$h(u,v) \propto a(u)a(v) \quad \textbf{[2]}$$

also fall under our description. The mixture model of Newman and Leicht (15) where, given communities $1,\cdots,K$,

$$\mathbb{P}[X_{ij} = 1 | i \in s, j \in r] = \theta_{ri}\theta_{sj}$$

is not of this type, since it is not invariant under permutations. It can be made invariant by summing over all permutations of $\{1,\cdots,n\}$, but is then generally not ergodic. Such models can be developed from our framework by permitting covariates Z_i depending on vertex identity or Z_{ij} depending on edge identity. Newman and Leicht's example where the communities are WEB pages falls under this observation. From a statistical point of view, these models bear the same relation to our models as regression models do to single population models.

Block models or models where

$$h(u,v,\theta) \propto \sum_{k=1}^{K_n} \theta_k a_k(u)a_k(v)$$

for known functions $\{a_k\}$ can be used to approximate general h. The latent eigenvalue model of Hoff (12) is of this type, but with a_k which are extremely rough and unidentifiable since the $a_j(\xi)$ are independent, and for which no unique choice exists. We can think of the canonical version of the block model as corresponding to a labeling $1,\cdots,K$ of the communities in the order $W_1 \leq \cdots \leq W_K$ where $W_j = \sum_k W_{jk}$, which is proportional to the expected degree of a member of community j. The function $h(\cdot,\cdot)$ then takes value P_{ab} on the (a,b) block of the product partition in which each axis is divided into consecutive intervals, of lengths $\pi_1,\cdots,\pi_K$. Each corresponding vertical slice exhibits the relation pattern for that community with the diagonal block identifying the members of the community. The nonparametric $h(\cdot,\cdot)$ gives the same intuitive picture on an arbitrarily fine scale. We note that, as in nonparametric statistics, to estimate h or w, regularization is needed. That is, we need to consider $K_n \to \infty$ at rates which depend on n and λ_n to obtain good estimates of h or w by using estimates of θ above or of block model parameters. We will discuss this further later.

Newman–Girvan and Likelihood Modularities

The task of determining K communities corresponds to finding a good assignment for the vertices $\mathbf{e} \equiv \{e_1,\cdots,e_n\}$ where $e_j \in \{1,\cdots,K\}$. There are K^n such assignments. Suppose that the distribution of A follows a K block model with parameters $\pi = (\pi_1,\cdots,\pi_K)$ and $P = ||P_{ab}||_{K\times K}$. The observed A is a consequence of a realization $\mathbf{c} = (c_1,\cdots,c_n)$ of n independent *Multinomial*$(1,\pi)$ variables. Evidently we can measure the adequacy of an assignment through the matrix

$$R(\mathbf{c},\mathbf{e}) = ||\boldsymbol{R}_{ab}||_{K\times K},$$

where $\boldsymbol{R}_{ab} = n^{-1}\sum_{i=1}^n I(c_i = b, e_i = a)$, the fraction of b members classified as a members if we use $\mathbf{e}$. It is natural to ask for consistency of an assignment $\mathbf{e}$, that is, $\mathbf{e} = \mathbf{c}$, i.e.

$$R(\mathbf{c},\mathbf{e}) = diag(\mathbf{f}(\mathbf{c}))$$

where $\mathbf{f}_a(\mathbf{e}) = n^{-1}n_a(\mathbf{e})$ and $n_a(\mathbf{e}) = \sum_{i=1}^n I(e_i = a)$. The Newman–Girvan modularity $Q_{NG}(\mathbf{e},A)$ is defined as follows. Let $\{i : e_i = k\}$ denote e-community k, i.e. as estimated by $\mathbf{e}$. Define

$$O_{kl}(\mathbf{e},A) = \sum_{1\le i,j\le n} A_{ij} I(e_i = k, e_j = l)$$

to be the block sum of A. Obviously, O_{kk} is twice the number of edges among nodes in the k-th e-community and for $k \neq l$, O_{kl} is the number of edges between nodes in the k-th e-community and nodes in the l-th e-community. Let $D_k(\mathbf{e},A) = \sum_{l=1}^K O_{kl}(\mathbf{e},A)$. It is easy to verify that D_k is the sum of degrees for nodes in the k-th e-community. Let $L = \sum_{k=1}^K D_k$ be twice the number of edges among all nodes. Then the Newman–Girvan modularity is defined by

$$Q_{NG}(\mathbf{e},A) = \sum_{k=1}^K \frac{O_{kk}}{L} - \left(\frac{D_k}{L}\right)^2.$$

The Newman–Girvan algorithm then searches for the membership assignment vector $\mathbf{e}$ that maximizes Q_{NG}. Notice that if edges are randomly generated uniformly among all pairs of nodes with given node degrees, then the number of edges between the kth e-community and lth e-community is expected to be $L^{-1}D_kD_l$. Therefore, the Newman–Girvan modularity measures the fraction of the edges on the graph that connect vertices of the same type (i.e. within-community edges) minus the expected value of the same quantity on a graph with the same community divisions but random connections between the vertices (6). Newman (16) contrasts and compares his modularity with spectral clustering, another common "community identification" method which we will also compare to the likelihood modularity below. We seek conditions under which the official N-G assignment

$$\hat{\mathbf{c}} = \arg\max Q_{NG}(\mathbf{e},A)$$

is consistent with probability tending to 1. Before doing so we consider alternative modularities.

For fixed $\mathbf{e}$, the conditional, given $\{n_k(\mathbf{e})\}_1^K$, log-likelihood of A is $\frac{1}{2}\sum_{1\le a,b\le K}(O_{ab}\log(P_{ab}) + (n_{ab} - O_{ab})\log(1 - P_{ab}))$, where $n_{ab} = n_an_b$ if $a \neq b$, $n_{aa} = n_a(n_a - 1)$. If we maximize over P, we obtain by letting $\tau(x) = x\log x + (1-x)\log(1-x)$,

$$Q_{LM}(\mathbf{e},A) = \frac{1}{2}\sum_{a,b} n_{ab}\tau\left(\frac{O_{ab}}{n_{ab}}\right)$$

which we call the likelihood modularity. This is not a true likelihood but a profile likelihood where we treat $\mathbf{e}$ as an unknown parameter. We will argue below that the profile likelihood is optimal in the usual parametric sense if $\frac{\lambda_n}{\log n} \to \infty$. But so are all other consistent modularities as defined below. However, we expect that if $\frac{\lambda_n}{\log n}$ is bounded and certainly if $\lambda_n = \Omega(1)$, the most important case, this is false. We are deriving optimal and computationally implementable procedures for this case.

We write general modularities in the form,

$$Q(\mathbf{e},A) = F_n\left(\frac{O}{\mu_n}, \frac{L}{\mu_n}, \mathbf{f}(\mathbf{e})\right)$$

where $\mu_n = E(L) = n(n-1)\rho_n$ and the matrix $O(\mathbf{e},A)$ is defined by its elements $O_{ab}(\mathbf{e},A)$, $\mathbf{f}(\mathbf{e}) = (\frac{n_1(\mathbf{e})}{n}, \cdots, \frac{n_K(\mathbf{e})}{n})^T$, and $F_n : \mathcal{M}\times\mathcal{R}^+\times\mathcal{G} \to \mathcal{R}$ where $\mathcal{M}$ is the set of all nonnegative $K\times K$ symmetric matrices and $\mathcal{G}$ is the K simplex. Note that both Q_{NG} and Q_{LM} can be written as such up to a proportionality constant. It is easy to see that if the K block model holds,

$$\frac{\mathbb{E}(O(\mathbf{e},A)|\mathbf{c})}{E(L)} = R(\mathbf{c},\mathbf{e})SR^T(\mathbf{c},\mathbf{e}) \qquad [3]$$

where, by definition, $R^T\mathbf{1} = \mathbf{f}(\mathbf{c})$, $R\mathbf{1} = \mathbf{f}(\mathbf{e})$, $\mathbf{1} = (1,\cdots,1)^T$, and $S_{ab} = \rho_n^{-1}\mathbb{P}(A_{12} = 1|c_1 = a, c_2 = b)$. Note that $W \equiv \mathcal{D}(\pi)S\mathcal{D}(\pi)$, where $\mathcal{D}(v) \equiv diag(v)$ for v $K\times 1$.

We define asymptotic consistency of a sequence of assignments $\hat{\mathbf{c}}$ by

$$\mathbb{P}[\hat{\mathbf{c}} = \mathbf{c}] \to 1 \qquad [4]$$

as $n \to \infty$.

We will assume that there exists a function $F : \mathcal{M}\times\mathbb{R}^+\times\mathcal{G} \to \mathbb{R}$ such that F_n is approximated by F evaluated at the conditional expectation given $\mathbf{c}$ of the argument of F_n. Suppose first $F_n \equiv F$. It is intuitively clear from [3] that if $\hat{\mathbf{c}} \to \mathbf{c}$, then $R(\mathbf{c},\hat{\mathbf{c}}) \to \mathcal{D}(\pi)$. Then, since $\mathbf{f}(\mathbf{c}) \to \pi$, the following condition is natural.

I. $F(RSR^T, 1, R\mathbf{1})$ is uniquely maximized over $\mathcal{R} = \{R : R \ge 0, R^T\mathbf{1} = \pi\}$ by $R = \mathcal{D}(\pi)$, for all (π,S) in an open set Θ.

This means $\mathbf{c}$ with $\mathbf{f}(\mathbf{c}) = \pi$ is the right assignment for the limiting problem. Note that since F is not concave in R, this is a strong condition.

For (π,S) to be identifiable uniquely, we clearly also need that:

II. S does not have two identical columns (Two communities cannot have identical probabilities of being related to other communities and within themselves) and π has all entries positive (Each community has some members).

We also need a few more technical conditions of a standard type.

III. a) F is Lipschitz in its arguments. b) The directional derivatives $\frac{\partial^2 F}{\partial\epsilon^2}(M_0+\epsilon(M_1-M_0), r_0+\epsilon(r_1-r_0), \mathbf{t}_0+\epsilon(\mathbf{t}-\mathbf{t}_0))|_{\epsilon=0+}$ are continuous in $(M_1,r_1,\mathbf{t})$ for all $(M_0,r_0,\mathbf{t}_0)$ in a neighborhood of $(W,1,\pi)$. c) Let $G(R,S) = F(RSR^T,1,R\mathbf{1})$. Assume that on $\mathcal{R}$, $\frac{\partial G((1-\epsilon)\mathcal{D}(\pi)+\epsilon R,S}{\partial\epsilon}|_{\epsilon=0+} < -C < 0$ for all $(\pi,S) \in \Theta$.

Theorem 1. *Suppose F, S and π satisfy I–III and $\hat{\mathbf{c}}$ is the maximizer of $Q(\mathbf{e},A)$. Suppose $\frac{\lambda_n}{\log n} \to \infty$. Then, for all $(\pi,S) \in \Theta$,*

$$\overline{limit}_{n\to\infty}\frac{\log\mathbb{P}(\hat{\mathbf{c}} \neq \mathbf{c})}{\lambda_n} \le -s_Q(\pi,S) < 0.$$

The proof is given in *SI Appendix*.

We note that Snijders and Nowicki (11) established a related result, exponential convergence to 0 of the mis-classification probability for $\lambda_n = \Omega(n)$ using node degree K-means clustering for $K = 2$.

Let $F(\mathbf{c},\mathbf{e}) \equiv F(R(\mathbf{c},\mathbf{e})SR^T(\mathbf{c},\mathbf{e}), \mathbf{f}^T(\mathbf{c})S\mathbf{f}(\mathbf{c}), \mathbf{f}(\mathbf{e}))$. In the general case it suffices to show,

$$\overline{limit}_{n\to\infty}\frac{\mathbb{P}\left(\sup\{|Q(\mathbf{e},A) - F(\mathbf{c},\mathbf{e})| : \mathbf{e}\} \le \delta\Delta_n\right)}{\lambda_n} \le -\gamma \qquad [5]$$

for all $\delta > 0$, some $\gamma > 0$, where

$$\Delta_n = \inf\{|F(\mathbf{c},\mathbf{e}) - F(\mathbf{c},\mathbf{c})| : |\mathbf{e} - \mathbf{c}| \ge 1\}$$

and $|\mathbf{e} - \mathbf{c}| = \sum_{i=1}^n 1(e_i \neq c_i)$. We show in *SI Appendix* that Q_{NG} and Q_{LM} satisfy I and III and Eq. **5** for selected (π,S) for Q_{NG} and all (π,S) for Q_{LM}.

An immediate consequence of the theorem is:

Corollary 1: *If the conditions of Theorem 1 hold and,*

$$\hat{W} \equiv L^{-1}O(\hat{\mathbf{c}},A),$$
$$\hat{\pi} = \left\{\frac{1}{n}\sum_{i=1}^{n} I(\hat{\mathbf{c}}_i = a) : a = 1,\cdots,K\right\},$$

then,

$$\sqrt{n}(\hat{\pi} - \pi) \Rightarrow \mathcal{N}(0, \mathcal{D}(\pi) - \pi\pi^T),$$
$$\sqrt{n}(\hat{W} - W) \Rightarrow S \cdot (\pi\eta^T + \eta\pi^T) - 2(\pi^T S\eta)W,$$
$$\eta = \mathcal{N}(0, \mathcal{D}(\pi) - \pi\pi^T),$$

with $A \cdot B$ denoting point-wise product. The limiting variances are what we would get for maximum likelihood estimates if $\hat{\mathbf{c}} = \mathbf{c}$, i.e. we knew the assignment to begin with. So consistent modularities lead to efficient estimates of the parameters.

This follows since with probability tending to 1, $\hat{\mathbf{c}} = \mathbf{c}$.

To estimate $w(\cdot,\cdot)$ in the nonparametric case we need $K \to \infty$, and $w(\cdot,\cdot)$ and $\pi(\cdot)$ smooth. We approximate by $W_K \sim K^{-2}||w(aK^{-1}, bK^{-1})||$, $\pi_K(a) \sim K^{-1}\pi(aK^{-1})$, where $w(\cdot,\cdot)$, W_K are canonical and the modularity defining F_K, $F_K(\cdot,\cdot,\cdot)$ is of order K^{-2}. We have preliminary results in that direction but their formulation is complicated and we do not treat them further here.

Consistency of N-G, L-M

We show in *SI Appendix* using the appropriate F_{NG}, F_{LM} that the likelihood modularity is always consistent while the Newman–Girvan is not. This is perhaps not surprising since N-G focuses on the diagonal of O. In fact, we would hope that N-G is consistent under the submodel $\{(\rho,\pi,W) : W_{aa} > \sum_{b\neq a} W_{ab}$ for all $a\}$, which corresponds to Newman and Girvan's motivation. We have shown this for $K = 2$ but it surprisingly fails for $K > 2$. Here is a counterexample. Let $K = 3$, $\pi = (1/3, 1/3, 1/3)^T$ and

$$P = \begin{bmatrix} .06 & .04 & 0 \\ .04 & .12 & .04 \\ 0 & .04 & .66 \end{bmatrix}.$$

As $n \to \infty$, with true labeling, Q_{NG} approaches 0.033. However, the maximum Q_{NG}, about 0.038, is achieved by merging the first two communities. That is, two sparser communities are merged. This is consistent with an observation of Fortunato and Barthelemy (17).

If for the profile likelihood we maximize only over $\mathbf{e}$ such that $\hat{W}_{aa}(\mathbf{e}) > \sum_{b\neq a}\hat{W}_{ab}(\mathbf{e})$ for all a, we obtain $\hat{\mathbf{c}}$ which is consistent under the submodel above, and in the Karate Club example performs like N-G.

Computational Issues

Computation of optimal assignments using modularities is, in principle, NP hard. However, although the surface is multimodal, in the examples we have considered and generally when the signal is strong, optimization from several starting points using a label switching algorithm (19) works well.

Simulation

We generate random matrices A and maximize Q_{NG}, Q_{LM} to obtain node labels respectively, where Q_{LM} is maximized using a label switching algorithm. To make a fair comparison, the initial labeling for Q_{NG} and Q_{LM} is to randomly choose 50% of the nodes with correct labels and the other 50% with random labels. For spectral clustering, we adopt the algorithm of (18) by using the first K eigenvectors of $\mathcal{D}(\mathbf{d})^{-1/2}A\mathcal{D}(\mathbf{d})^{-1/2}$, where $\mathbf{d} = (d_1,\cdots,d_n)^T$ and d_i is the degree of the i-th node. We generate the P matrix randomly by forcing symmetry and then add a constant to diagonal entries such that I holds. The π is generated randomly from the simplex. To be precise, the values for Fig. 1 are $\pi = (.203, .286, .511)^T$ and

$$P = bn^{-1}\log n \cdot \begin{bmatrix} .43 & .06 & .13 \\ .06 & .34 & .17 \\ .13 & .17 & .40 \end{bmatrix},$$

where n varies from 200 to 1,500 and b varies from 10 to 100. Obviously, Fig. 1 says that the likelihood method exhibits much less incorrect labeling than Newman–Girvan and spectral clustering. This is consistent with theoretical comparison.

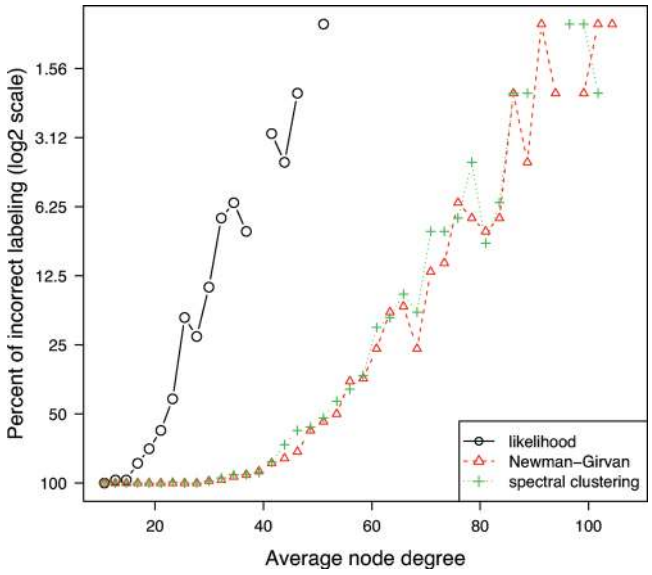

Fig. 1. Empirical comparison of Newman–Girvan, likelihood modularities and spectral clustering (18), where $K = 3$, the number of nodes n varies from 200 to 1500, and the percent of correct labeling is computed from 100 replicates of each simulation case. Here π, P are given in the text.

Data Examples

We compare the L-M and N-G modularity algorithms below with applications to two real data sets. To deal with the issue of non-convex optimization, we simply use many restarting points.

Zachary's "Karate Club" Network. We first compare L-M and N-G with the famous "Karate Club" network of ref. 20, from the social science literature, which has become something of a standard test for community detection algorithms. The network shows the patterns of friendship between the members of a karate club at a US university in the 1970s. The example is of particular interest because shortly after the observation and construction of the network, the club in question split into two components separated by the dashed line as shown in Figs. 2 and 3 as a result of an internal dispute. Fig. 2 *Left* shows two communities identified by maximizing the likelihood modularity where the shapes of the vertices denote the membership of the corresponding individuals, and similarly the right panel shows communities identified by N-G. Obviously,the N-G communities match the two sub-divisions identified by the split save for one mis-classified individual. The L-M communities are quite different, and obviously one community consists of five individuals with central importance that connect with many other nodes while the other community consists of the remaining individuals. Although not reflecting the split this corresponds to other plausible distinguishing characteristics of the individuals. However, if we force the constraint that within-community density is no less than the density of relationship to all other communities, the submodel we discussed, then we obtain two L-M communities that match the split perfectly. The same partitions as ours with and without constraint have also been reported

STATISTICS

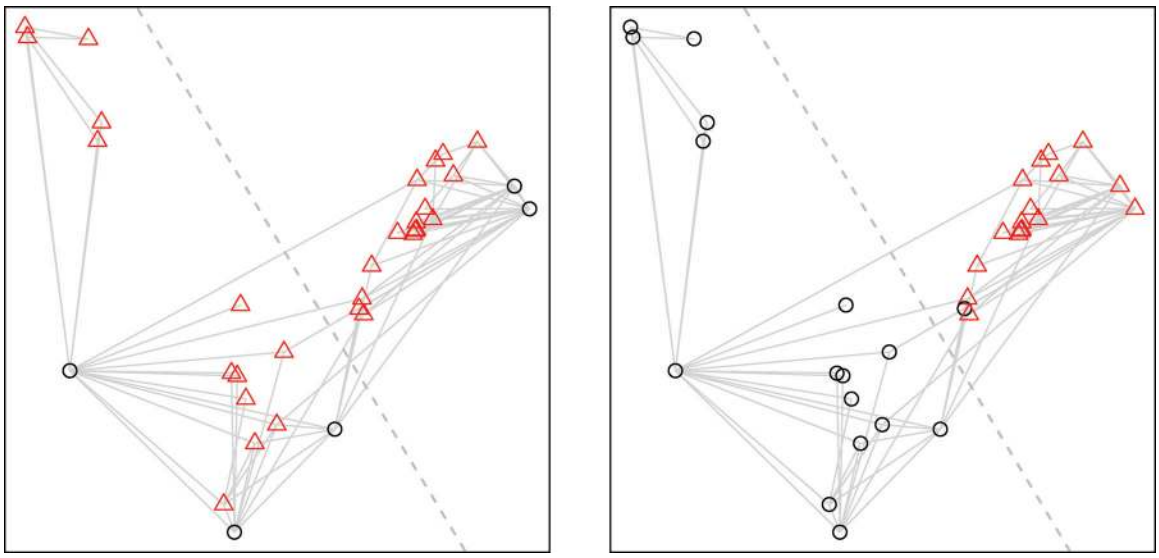

Fig. 2. Zachary's karate club network. Communities were identified by maximizing the likelihood modularity (*Left*) and by maximizing the Newman–Girvan modularity with $K = 2$ (*Right*), where the shapes of vertices indicate the membership of the corresponding individuals. The dashed line cuts the nodes into two groups which are the "known" communities that the club was split into.

by Rosvall and Bergstrom using a data compression criterion (21), which is closely related to L-M. We note that, as is usual in clustering, there is no ground truth, only features which can be validated ex post fact. It is interesting to note that, if instead of $K = 2$, we put $K = 4$, as in Fig. 3, it is evident for both modularities that merging the communities on either side of the eigenvector split, gives the "correct" Karate Club split. This suggests that the standard policy mentioned by Newman (16) of increasing the number of communities by splitting is not necessarily ideal since in this case the "misclassified" individual of Fig. 2 would never be "correctly" classified.

Private Branch Exchange. Our second example is of a telephone communication network where connections are made among the internal telephones of a private business organization, a so called PBX. PBXs are differentiated from "key systems" in that users of key systems usually select their own outgoing lines, while PBXs select the outgoing line automatically. Our data contains 621 individuals. Fig. 4 *Left* shows the results of community detection by L-M, where the adjacency matrix is plotted but the nodes are sorted according to the membership of the corresponding individuals identified by maximizing the likelihood modularity. Similarly, the right panel of Fig. 4 shows the communities identified by N-G, where the maximum Newman–Girvan modularity is 0.4217. Note that the identified communities by L-M have sizes 323, 81, 78, 97, 41, and 1, respectively. The communities are ordered simply by their average node degrees, essentially the order for h_{CAN}. Interestingly, the last L-M community has only one node that communicates with almost everyone else, nodes in the second community only communicate with internal nodes, nodes in the fourth community and the sixth community, but not with others; Similarly, the third community only communicates with the fifth and sixth communities, and so on. In other words, communication between communities is sparse. However, the communities identified by N-G are quite different with only the fifth community heavily overlapping with a community identified by L-M. This

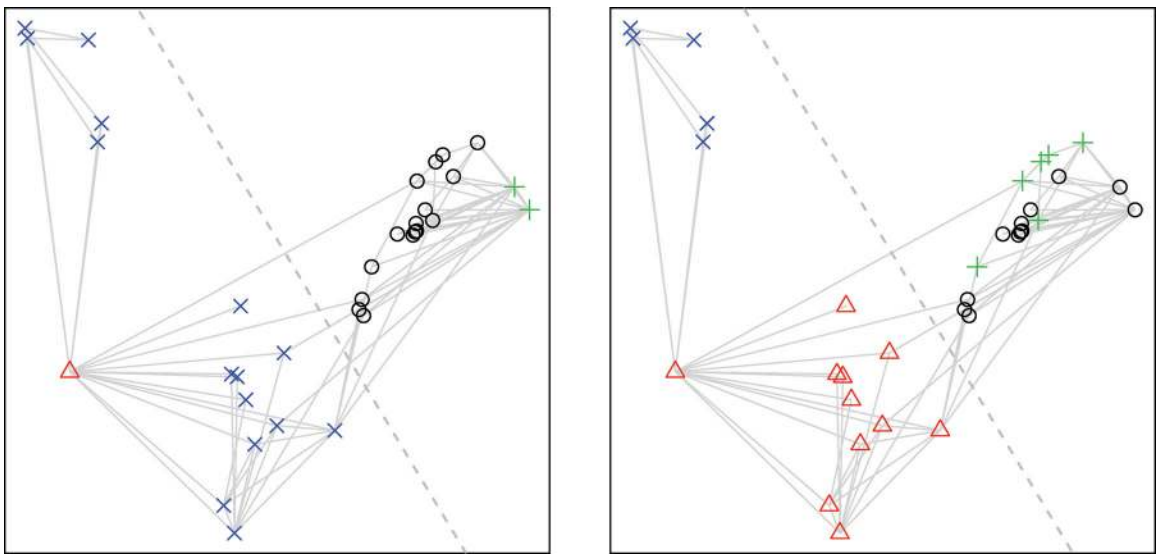

Fig. 3. Zachary's karate club network. Communities were identified by maximizing the likelihood modularity (*Left*) and by maximizing the Newman–Girvan modularity with $K = 4$ (*Right*), where the shapes of vertices indicate the membership of the corresponding individuals. The dashed line cuts the nodes into two groups which are the known communities that the club was split into.

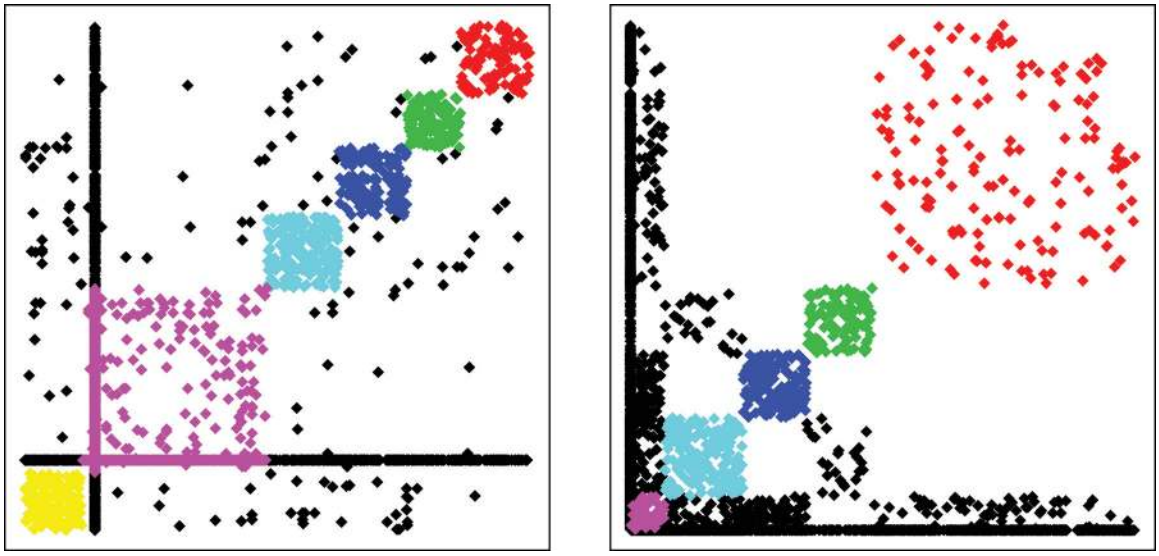

Fig. 4. Private branch exchange data. (*Left*) The adjacency matrix where the nodes are sorted according to the membership of the corresponding individuals identified by maximizing L-M. (*Right*) Same as *Left*, but with individuals identified by maximizing N-G with $K = 6$. The colors on the within-community edges are used to differentiate the communities for both L-M and N-G.

difference appears to be caused by the nodes in the 5th and 6th L-M communities which have many more between-community connections than within-community connections while N-G more or less maximizes within-community connections. We have verified that the group communicating with all others is a service group.

Discussion

1. As we noted, under our conditions the usual statistical goal of estimating the parameters π and P is trivial, since, once we have assigned individuals to the K communities consistently, the natural estimates, $\hat{W}$ and $\hat{\pi}$, are not just consistent but efficient. However, in the more realistic case where $\lambda_n = \Omega(1)$, or even just $\lambda_n = \Omega(\log n)$, this is no longer true. Elsewhere, we shall show that, indeed, estimation of parameters by maximum likelihood and Bayes classification of individuals (no longer perfect) is optimal.
2. A difficulty faced by all these methods, modularities or likelihoods, is that if K is large, searching over the space of classifications becomes prohibitively expensive. In subsequent work we intend to show that this difficulty may be partly overcome by using the method of moments to first estimate π and P, and then study the likelihood in a neighborhood of the estimated values.

Open Problems

1. A fundamental difficulty not considered in the literature is the choice of K. From our nonparametric point of view, this can equally well be seen as, how to balance bias and variance in the estimation of $w(\cdot,\cdot)$. We would like to argue that, as in nonparametric statistics, estimating $w(\cdot,\cdot)$ without prior prejudices on its structure is as important an exploratory step in this context as, using histograms in ordinary statistics.
2. The linking of this framework to covariates depending on vertice or edge identity is crucial, permitting relationship strength to be assessed as a function of vector variables.
3. The links of our approach to spectral graph clustering and more generally clustering on the basis of similarities seem intriguing.

ACKNOWLEDGMENTS. We thank Tin K Ho for help in obtaining the PBX data and for helpful discussions. We also thank the referees, whose references and comments improved this article immeasurably.

1. Doreian P, Batagelj V, Ferligoj A (2005) *Generalized Blockmodeling* (Cambridge Univ Press, Cambridge, UK).
2. Airoldi EM, Blei DM, Fienberg SE, Xing XP (2008) Mixed-membership stochastic blockmodels. *J Machine Learning Res* 9:1981–2014.
3. Erdös P, Rényi A (1960) On the evolution of random graphs. *Publ Math Inst Hungar Acad Sci* 5:17–61.
4. Bollobas B, Janson S, Riordan O (2007) The phase transition in inhomogeneous random graphs. *Random Struct Algorithms* 31:3–122.
5. Zhao Y, et al. (2009) Botgraph: Large scale spamming botnet detection. *Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation* (USENIX, Berkeley, CA), pp 321–334.
6. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69:026113.
7. Guimera R, Amaral LAN (2005) Functional cartography of complex metabolic networks. *Nature* 433:895–900.
8. Chung FRK, Lu L (2006) *Complex Graphs and Networks*. CBMS Regional Conference Series in Mathematics (Am Math Soc, Providence, RI).
9. Kallenberg O (2005) Probabilistic symmetries and invariance principles. *Probability and Its Application* (Springer, New York).
10. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. *Soc Networks* 5:109–137.
11. Snijders T, Nowicki K. (1997) Estimation and prediction for stochastic block-structures for graphs with latent block structure. *J Classification* 14:75–100.
12. Hoff PD (2008) Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in Neural Information Processing Systems*, eds Platt J, Koller D, Roweis S (MIT Press, Cambridge, MA) Vol 20, pp 657–664.
13. Diaconis P, Janson S (2008) Graph limits and exchangeable random graphs. *Rendiconti di Matematica* 28:33–61.
14. Hardy G, Littlewood J, Polya G (1988) *Inequalities*. (Cambridge Univ Press, Cambridge, UK), 2nd Ed.
15. Newman MEJ, Leicht EA (2007) Mixture models and exploratory analysis in networks. *Proc Natl Acad Sci USA* 104:9564–9569.
16. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E* 74:036104.
17. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. *Proc Natl Acad Sci USA* 104:36–41.
18. Ng AY, Jordan MI, Weiss Y (2001) On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA) Vol 14, pp 849–856.
19. Stephens M (2000) Dealing with label-switching in mixture models. *J R Stat Soc B* 62:795–809.
20. Zachary W (1977) An information flow model for conflict and fission in small groups. *J Anthropol Res* 33:452–473.
21. Rosvall M, Bergstrom C (2007) An information-theoretic framework for resolving community structure in complex network. *Proc Natl Acad Sci USA* 104:7327–7331.