The Annals of Statistics
2011, Vol. 39, No. 5, 2280–2301
DOI: 10.1214/11-AOS904

THE METHOD OF MOMENTS AND DEGREE DISTRIBUTIONS FOR NETWORK MODELS

BY PETER J. BICKEL[1], AIYOU CHEN[2] AND ELIZAVETA LEVINA[3]

University of California, Berkeley, Google Inc. and University of Michigan

This research is dedicated to Erich L. Lehmann, the thesis advisor of one of us and "grand thesis advisor" of the others. It is a work in which we try to develop nonparametric methods for doing inference in a setting, unlabeled networks, that he never considered. However, his influence shows in our attempt to formulate and develop a nonparametric model in this context. We also intend to study to what extent a potentially "optimal" method such as maximum likelihood can be analyzed and used in this context. In this respect, this is the first step on a road he always felt was the main one to stick to.

Probability models on graphs are becoming increasingly important in many applications, but statistical tools for fitting such models are not yet well developed. Here we propose a general method of moments approach that can be used to fit a large class of probability models through empirical counts of certain patterns in a graph. We establish some general asymptotic properties of empirical graph moments and prove consistency of the estimates as the graph size grows for all ranges of the average degree including $\Omega(1)$. Additional results are obtained for the important special case of degree distributions.

1. Introduction. The analysis of network data has become an important component of doing research in many fields; examples include social and friendship networks, food webs, protein interaction and regulatory networks in genomics, the World Wide web and computer networks. On the algorithmic side, many algorithms for identifying important network structures such as communities have been proposed, mainly by computer scientists and physicists; on the mathematical side, various probability models for random graphs have been studied. However, there has only been a limited amount of research on statistical inference for networks, and on learning the network features by fitting models to data; to a large extent, this is due to the gap between the relatively simple models that are analytically

Received February 2011; revised June 2011.

[1]Supported in part by NSF Grant DMS-09-06808.

[2]Part of this work was done while A. Chen was at Alcatel-Lucent Bell Labs.

[3]Supported in part by NSF Grants DMS-08-05798 and DMS-11-06772.

MSC2010 subject classifications. Primary 62E10; secondary 62G05.

Key words and phrases. Social networks, block model, community detection.

tractable and the complex features of real networks not easily reproduced by these models.

Probability models on infinite graphs have a nice general representation based on results [Aldous (1981), Hoover (1979), Kallenberg (2005), Diaconis and Janson (2008)], analogous to de Finetti's theorem, for exchangeable matrices. Here, we give a brief summary closely following the notation of Bickel and Chen (2009). Graphs can be represented through their adjacency matrix A, where $A_{ij} = 1$ if there is an edge from node i to j and 0 otherwise. We assume $A_{ii} = 0$, that is, there are no self-loops. A_{ij}'s can also represent edge weights if the graph is weighted, and for undirected graphs, which is our focus here, $A_{ij} = A_{ji}$. For an unlabeled random graph, it is natural to require its probability distribution P on the set of all matrices $\{[A_{ij}], i, j \geq 1\}$ to satisfy $[A_{\sigma_i \sigma_j}] \sim P$, where σ is an arbitrary permutation of node indices. In that case, using the characterizations above one can write

$$A_{ij} = g(\alpha, \xi_i, \xi_j, \lambda_{ij}), \tag{1.1}$$

where α, ξ_i and λ_{ij} are i.i.d. random variables distributed uniformly on $(0, 1)$, $\lambda_{ij} = \lambda_{ji}$ and g is a function symmetric in its second and third arguments. α as in de Finetti's theorem corresponds to the mixing distribution and is not identifiable. The equivalent of the i.i.d. sequences in de Finetti's theorem here are distributions of the form $A_{ij} = g(\xi_i, \xi_j, \lambda_{ij})$. This representation is not unique, and g is not identifiable. These distributions can be parametrized through the function

$$h(u, v) = \mathbb{P}[A_{ij} = 1 | \xi_i = u, \xi_j = v]. \tag{1.2}$$

The function h is still not unique, but it can be shown that if two functions h_1 and h_2 define the same distribution P, they can be related through a measure-preserving transformation, and a unique canonical h can be defined, with the property that $\int_0^1 h_{\mathrm{can}}(u, v)\, dv$ is monotone nondecreasing in u; see Bickel and Chen (2009) for details. From now on, h will refer to the canonical h_{can}. We use the following parametrization of h: let

$$\rho = \int_0^1 \int_0^1 h(u, v)\, du\, dv \tag{1.3}$$

be the probability of an edge in the network. Then the density of (ξ_i, ξ_j) conditional on $A_{ij} = 1$ is given by

$$w(u, v) = \rho^{-1} h(u, v). \tag{1.4}$$

With this parametrization, it is natural to let $\rho = \rho_n$, make w independent of n and control the rate of the expected degree $\lambda_n = (n - 1)\rho_n$ as $n \to \infty$. The case most studied in probability on random graphs is $\lambda_n = \Omega(1)$ [where $a_n = \Omega(b_n)$ means $a_n = O(b_n)$ and $b_n = O(a_n)$]. The case of $\lambda_n = 1$ corresponds to the so-called phase transition, with the giant connected component emerging for $\lambda_n > 1$.

Many previously studied probability models for networks fall into this class. It includes the block model [Holland, Laskey and Leinhardt (1983), Snijders and Nowicki (1997), Nowicki and Snijders (2001)], the configuration model [Chung and Lu (2002)] and many latent variable models, including the univariate [Hoff, Raftery and Handcock (2002)] and multivariate [Handcock, Raftery and Tantrum (2007)] latent variable models, and latent feature models [Hoff (2007)]. In fact, dynamically defined models such as the "preferential attachment" model [which seems to have been first mentioned by Yule in the 1920s, formally described by de Solla Price (1965) and given its modern name by Barabási and Albert (1999)] can also be thought of in this way if the dynamical construction process continues forever producing an infinite graph; see Section 16 of Bollobás, Janson and Riordan (2007).

Bickel and Chen (2009) pointed out that the block model provides a natural parametric approximation to the nonparametric model (1.2), and the block model is the main parametric model we consider in this paper; see more details in Section 3. The block model can be defined as follows: each node $i = 1, \dots, n$ is assigned to one of K blocks independently of the other nodes, with $\mathbb{P}(c_i = a) = \pi_a$, $1 \le a \le K$, $\sum_{a=1}^{K} \pi_a = 1$, where K is known, and $c = (c_1, \dots, c_n)$ is the $n \times 1$ vector of labels representing node assignments to blocks. Then, conditional on c, edges are generated independently with probabilities $\mathbb{P}[A_{ij} = 1 | c_i = a, c_j = b] = F_{ab}$. The vector of probabilities $\pi = \{\pi_1, \dots, \pi_K\}$ and the $K \times K$ symmetric matrix $F = [F_{ab}]_{1 \le a,b \le K}$ together specify a block model. The block model is typically fitted either in the Bayesian framework through some type of Gibbs sampling [Snijders and Nowicki (1997)] or by maximizing the profile likelihood using a stochastic search over the node labels [Bickel and Chen (2009)]. Bickel and Chen (2009) also established conditions on modularity-type criteria such as the Newman–Girvan modularity [see Newman (2006) and references therein] give consistent estimates of the node labels in the block model, under the condition of the graph degree growing faster than $\log n$, where n is the number of nodes. They showed that the profile likelihood criterion satisfies these conditions.

The block model is very attractive from the analytical point of view and useful in a number of applications, but the class (1.2) is much richer than the block model itself. Moreover, the block model cannot deal with nonuniform edge distributions within blocks, such as the commonly encountered "hubs," although a modification of the block model introducing extra node-specific parameters has been recently proposed by Karrer and Newman (2011) to address this shortcoming. It may also be difficult to obtain accurate results from fitting the block model by maximum likelihood when the graph is sparse.

In this paper, we develop an alternative approach to fitting models of type (1.2), via the classical tool of the method of moments. By moments, we mean empirical or theoretical frequencies of occurrences of particular patterns in a graph, such as commonly used triangles and stars, although the theory is for general patterns.

While specific parametric models like the block model can be fitted by other methods, the method of moments applies much more generally, and leads to some general theoretical results on graph moments along the way. We note that related work on the method of moments was carried out for some specific parametric models in Picard et al. (2008).

A well-studied class of random graph models where moments play a big role is the exponential random graph models (ERGMs). ERGMs are an exponential family of probability distributions on graphs of fixed size that use network moments such as number of edges, p-stars and triangles as sufficient statistics. ERGMs were first proposed by Holland and Leinhardt (1981) and Frank and Strauss (1986) and have then been generalized in various ways by including nodal covariates or forcing particular constraints on the parameter space; see Robins et al. (2007) and references therein. While the ERGMs are relatively tractable, fitting them is difficult since the partition function can be notoriously hard to estimate. Moreover, they often fail to provide a good fit to data. Recent research has shown that a wide range of ERGMs are asymptotically either too simplistic, that is, they become equivalent to Erdös–Renyi graphs, or nearly degenerate, that is, have no edges or are complete; see Handcock (2003) for empirical studies and Chatterjee and Diaconis (2011) and Shalizi and Rinaldo (2011) for theoretical analysis.

The rest of the paper is organized as follows. In Section 2, we set up the notation and problem formulation and study the distribution of empirical moments, proving a central limit theorem for acyclic patterns. We also work out examples for several specific patterns. In Section 3 we show how to use the method of moments to fit the block model, as well as identify a general nonparametric model of type (1.2). In Section 4, we focus on degree distributions, which characterize (asymptotically) the model (1.2). Section 5 discusses the relationship between normalized degrees and more complicated pattern counts that can be used to simplify computation of empirical moments. Section 6 concludes with a discussion. Proofs and additional lemmas are given in the Appendix.

2. The asymptotic distribution of moments.

2.1. *Notation and theory.* We start by setting up notation. Let G_n be a random graph on vertices $1, \dots, n$, generated by

$$\mathbb{P}(A_{ij} = 1 | \xi_i = u, \xi_j = v) = h_n(u, v) = \rho_n w(u, v) I(w \le \rho_n^{-1}), \tag{2.1}$$

where $w(u, v) \ge 0$, symmetric, $0 \le u, v \le 1$, $\rho_n \to 0$. We cannot, unfortunately, treat ρ_n and w as two completely free parameters, as we need to ensure that $h \le 1$. We can either assume that the sequence ρ_n is such that $\rho_n w \le 1$ for all n, or restrict our attention to classes where $w_n(u, v) = w(u, v) I(w(u, v) \le \rho_n^{-1}) \stackrel{L_2}{\to} w(u, v)$. In either case, we can ignore the weak dependence of w_n on ρ_n and effectively replace w_n with w.

Let $T : \mathcal{L}_2(0,1) \to \mathcal{L}_2(0,1)$ be the operator defined by

$$[Tf](u) \equiv \int_0^1 h(u,v) f(v)\,dv.$$

We drop the subscript n on h, T when convenient. Similarly, let $T_w : \mathcal{L}_2(0,1) \to \mathcal{L}_2(0,1)$ be defined by w. Let

$$D_i = \sum_j A_{ij}, \qquad \bar{D} = \frac{1}{n}\sum_{i=1}^n D_i = \frac{2L}{n}.$$

Thus D_i is the degree of node i, $\bar{D}$ is the average degree and L is the total number of edges in G_n.

Let R be a subset of $\{(i,j) : 1 \le i < j \le n\}$. We identify R with the vertex set $V(R) = \{i : (i,j) \text{ or } (j,i) \in R \text{ for some } j\}$ and the edge set $E(R) = R$. Let $G_n(R)$ be the subgraph of G_n induced by $V(R)$. Recall that two graphs R_1 and R_2 are called isomorphic ($R_1 \sim R_2$) if there exists a one-to-one map σ of $V(R_1)$ to $V(R_2)$ such that the map $(i,j) \to (\sigma_i, \sigma_j)$ is one-to-one from $E(R_1)$ to $E(R_2)$.

Throughout the paper, we will be using two key quantities defined next:

$$Q(R) = \mathbb{P}\big(A_{ij} = 1, \text{ all } (i,j) \in R\big),$$
$$P(R) = \mathbb{P}\big(E(G_n(R)) = R\big).$$

Next, we give a proposition summarizing some simple relationships between P and Q. The proof, which is elementary, is given in the Appendix. Similar results are implicit in Diaconis and Janson (2008).

PROPOSITION 1. *If G_n is a random graph, and R a subset of $\{(i,j) : 1 \le i < j \le n\}$, then*

$$P(R) = \mathbb{E}\bigg\{\prod_{(i,j)\in R} h(\xi_i,\xi_j) \prod_{(i,j)\in\bar{R}} \big(1 - h(\xi_i,\xi_j)\big)\bigg\}$$
$$= Q(R) - \sum\big\{Q\big(R \cup (i,j)\big) : (i,j) \in \bar{R}\big\} \tag{2.2}$$
$$+ \sum\big\{Q\big(R \cup \{(i,j),(k,l)\}\big) : (i,j),(k,l) \in \bar{R}\big\} - \cdots,$$

where $\bar{R} = \{(i,j) \notin R, i \in V(R), j \in V(R)\}$. Further,

$$Q(R) = \sum\{P(S) : S \supset R, V(S) = V(R)\}. \tag{2.3}$$

Here $R \subset S$ refers to $S \subset \{(i,j) : i,j \in V(R)\}$.

The quantities $P(R)$ and $Q(R)$ are unknown population quantities which we can estimate from data, that is, from the graph G_n. Define, for $R \subset \{(i,j) : 1 \le i < j \le n\}$ with $|V(R)| = p$,

$$\hat{P}(R) = \frac{1}{\binom{n}{p} N(R)} \sum\{1(G \sim R) : G \subset G_n\},$$

where $N(R)$ is the number of graphs isomorphic to R on vertices $1, \dots, p$. For instance, if R is a 2-star consisting of two edges $(1, 2)$, $(1, 3)$, then $N(R) = 3$. Further, let

$$\hat{Q}(R) = \sum\{\hat{P}(S) : S \supset R, V(S) = V(R)\}.$$

Here we use R and S to denote both a subset and a subgraph. Evidently,

$$\mathbb{E}\hat{P}(R) = P(R), \qquad \mathbb{E}\hat{Q}(R) = Q(R).$$

The scaling here is controlled by the parameter ρ_n, the natural assumption for which is $\rho_n \to 0$. In that case, $P(R) \to 0$ for any fixed R with a fixed number of vertices p. Therefore we consider the following rescaling of $P(R)$ and $Q(R)$: writing $|R|$ for $|E(R)|$, let

$$\tilde{P}(R) = \rho_n^{-|R|} P(R), \qquad \tilde{Q}(R) = \rho_n^{-|R|} Q(R).$$

Then we have

$$\tilde{P}(R) = \mathbb{E} \prod_{(i,j) \in R} w_n(\xi_i, \xi_j) + O\left(\frac{\lambda_n}{n}\right) \tag{2.4}$$

since

$$\rho_n^{-|R|} \mathbb{E} \prod_{(i,j) \in R} h_n(\xi_i, \xi_j) \left[\prod_{(i,j) \in \bar{R}} (1 - h_n(\xi_i, \xi_j)) - 1 \right] = O(\rho_n) = O\left(\frac{\lambda_n}{n}\right),$$

if $\int w^{2(|R|+1)}(u, v)\, du\, dv < \infty$.

Next, we define the natural sample estimates of the population quantities $\tilde{P}$ and $\tilde{Q}$ by

$$\check{P}(R) = \hat{\rho}_n^{-|R|} \hat{P}(R), \qquad \check{Q}(R) = \hat{\rho}_n^{-|R|} \hat{Q}(R),$$

where $\hat{\rho}_n = \frac{\bar{D}}{n-1} = \frac{2L}{n(n-1)}$ is the estimated probability of an edge. For these rescaled versions of P and Q, we have the following theorem.

THEOREM 1. *Suppose* $\int_0^1 \int_0^1 w^2(u, v)\, dv\, du < \infty$.

(a) *If* $\lambda_n \to \infty$, *then*

$$\frac{\hat{\rho}_n}{\rho_n} \to_P 1, \tag{2.5}$$

$$\sqrt{n}\left(\frac{\hat{\rho}_n}{\rho_n} - 1\right) \Rightarrow \mathcal{N}(0, \sigma^2) \tag{2.6}$$

for some $\sigma^2 > 0$. *Suppose further* R *is fixed, acyclic with* $|V(R)| = p$ *and* $\int w^{2|R|}(u, v)\, du\, dv < \infty$. *Then,*

$$\begin{gathered} \check{P}(R) \to_P \tilde{P}(R), \\ \sqrt{n}(\check{P}(R) - \tilde{P}(R)) \Rightarrow \mathcal{N}(0, \sigma^2(R)). \end{gathered} \tag{2.7}$$

More generally, for any fixed $\{R_1, \dots, R_k\}$ *as above with* $|V(R_j)| \le p$,

$$\sqrt{n}\big((\check{P}(R_1), \dots, \check{P}(R_k)) - (\tilde{P}(R_1), \dots, \tilde{P}(R_k))\big) \Rightarrow \mathcal{N}(\mathbf{0}, \Sigma(\mathbf{R})). \tag{2.8}$$

(b) *Suppose* $\lambda_n \to \lambda < \infty$. *Conclusions* (2.5)–(2.8) *continue to hold save that* $\sigma^2(R)$, $\Sigma(R)$ *depend on* λ *as well as* R.

(c) *Even if* R *is not necessarily acyclic, the same conclusions apply to* $\check{Q}$ *and* $\tilde{Q}$ *if* λ_n *is of order* $n^{1-2/p}$ *or higher, and to* $\check{P}$ *and* $\tilde{P}$ *under the same condition on* λ_n.

The proof is given in the Appendix.

Remarks. (1) Note that part (b) yields consistency and asymptotic normality of acyclic graph moment estimates across the phase transition to a giant component, that is, for $\lambda < 1$ as well as $\lambda \ge 1$.

(2) Note that we are, throughout, estimating features of the canonical w. Unnormalized P and Q are trivially 0 if λ_n is not of order n.

(3) In view of (2.4), we can use $\check{P}(R)$ as an estimate of $\tilde{Q}(R)$ if R is acyclic and $\lambda_n = o(n^{1/2})$, since in this case the bias of $\check{P}$ is of order $o(n^{-1/2})$. The reason for not using $\check{Q}(R)$ directly even if R is acyclic is that by (2.3), there may exist $S \supset R$ which are not acyclic, and we can therefore not conclude that the theorem also applies to $\check{Q}$ unless we are in case (c).

(4) Part (c) of the theorem shows that for graphs with $\lambda_n = \Omega(n)$, $\check{Q}$ always gives $\sqrt{n}$-consistent estimates of any pattern while $\check{P}$ is not consistent unless we assume acyclic graphs, since the bias is of order $O(\lambda_n/n) = O(1)$. In the range $\lambda_n = o(n^{1/2})$ to $\Omega(n)$, what is possible depends on the pattern. For instance, if $\Delta = \{(1,2), (2,3), (3,1)\}$, a triangle, $\check{P}(\Delta) = \check{Q}(\Delta)$ (because there is no other graph on three nodes containing Δ), and $\check{P}$ is $\sqrt{n}$-consistent if $\lambda_n \ge \varepsilon n^{1/3}$ by part (c) but otherwise only consistent if $\lambda_n \to \infty$.

2.2. *Examples of specific patterns*. Next we give explicit formulas for several specific R. Our main focus is on wheels (defined next), which, as we shall see, in principle can determine the canonical w.

DEFINITION 1 (Wheels). A (k,l)-wheel is a graph with $kl + 1$ vertices and kl edges isomorphic to the graph with edges $\{(1,2), \dots, (k, k+1); (1, k+2), \dots, (2k, 2k+1); \dots, (1, (l-1)k+2), \dots, (lk, lk+1)\}$.

In other words, a wheel consists of node 1 at the center and l "spokes" connected to the center, and each spoke is a chain of k edges. We consider only $k \ge 2$. The number of isomorphic (k,l)-wheels on vertices $1, \dots, p$ is $N(R) = (kl+1)!/l!$.

If the graph R is a (k,l)-wheel, the theoretical moments have a simple form and can be expressed in terms of the operator T as follows:

$$Q(R) = \mathbb{E}(T^k(1)(\xi_1))^l. \tag{2.9}$$

This follows from

$$\begin{aligned}Q(R) &= \mathbb{E}\Big(\mathbb{E}\Big(\prod\{h(\xi_i,\xi_j):(i,j)\in E(R)\}|\xi_1\Big)\Big)\\ &= \left(\int_0^1\cdots\int_0^1 h(\xi_1,\xi_2)\cdots h(\xi_k,\xi_{k+1})\,d\xi_2\cdots d\xi_{k+1}\right)^l\\ &= \mathbb{E}(T^k(1)(\xi_1))^l,\end{aligned}$$

where the first equality holds by the definition of Q and the second by the structure of a (k,l)-wheel.

For a (k,l)-wheel R, from our general considerations, $\mathbb{E}\check{P}(R) = \tilde{P}(R) = \tilde{Q}(R) + o(1)$ if $\lambda_n = o(n)$ and in view of (2.8), $\check{P}(R)$ always consistently estimates $\tilde{Q}(R)$. However, $\sqrt{n}$-consistency of $\check{P}$ (converging to $\tilde{Q}$) holds in general only if $\lambda_n = o(n^{1/2})$. By part (c) $\check{Q}$ is $\sqrt{n}$ consistent for $\tilde{Q}$ only if λ_n is of order larger than $n^{1-2/(kl+1)}$. In the λ_n range between $O(n^{1/2})$ and $O(n^{1-2/(kl+1)})$, we do not exhibit a $\sqrt{n}$-consistent estimate though we conjecture that by appropriate de-biasing of $\check{P}$ such an estimate may be constructed. However, $\lambda_n = o(n^{1/2})$ seems a reasonable assumption for most graphs in practice, and then we can use the more easily computed $\check{P}$.

DEFINITION 2 (Generalized wheels). A $(\mathbf{k},\mathbf{l})$-wheel, where $\mathbf{k} = (k_1,\dots,k_t)$, $\mathbf{l} = (l_1,\dots,l_t)$ are vectors and the k_j's are distinct integers, is the union $R_1\cup\cdots\cup R_t$, where R_j is a (k_j,l_j)-wheel, $j = 1,\dots,t$, and the wheels $R_1,\dots,R_t$ share a common hub but all their spokes are disjoint.

A $(\mathbf{k},\mathbf{l})$-wheel has a total of $p = \sum_j l_jk_j + 1$ vertices and $\sum_j l_jk_j$ edges. For example, a graph defined by $E = \{(1,2); (1,3), (3,4); (1,5), (5,6); (1,7), (7,8), (8,9)\}$ is a $(\mathbf{k},\mathbf{l})$-wheel with $\mathbf{k} = (1,2,3)$ and $\mathbf{l} = (1,2,1)$. The number of distinct isomorphic $(\mathbf{k},\mathbf{l})$-wheels on p vertices is $N(R) = p!(\prod_j l_j!)^{-1}$.

We can compute, defining $A(R) = \prod\{A_{ij}:(i,j)\in R\}$,

$$\begin{aligned}Q(R) &= \mathbb{P}\left(\bigcap_{j=1}^t [A(R_j)=1]\right)\\ &= \mathbb{E}\left\{\prod_{j=1}^t \mathbb{P}\big(A(R_j)=1|\mathrm{Hub}\big)\right\}\qquad(2.10)\\ &= \mathbb{E}\prod_{j=1}^t [T^{k_j}(\xi)]^{l_j}.\end{aligned}$$

Thus $(\mathbf{k},\mathbf{l})$-wheels give us all cross moments of $T^m(\xi)$, $m\ge 1$. Note that all $(\mathbf{k},\mathbf{l})$-wheels are acyclic.

We are not aware of other patterns for which the moment formulas are as simple as those for wheels. For example, if R is a triangle, then

$$Q(R) = \int_0^1 \int_0^1 \int_0^1 h(u, v)h(v, w)h(w, u)\, du\, dv\, dw$$
$$= \int_0^1 \int_0^1 h^{(2)}(u, w)h(w, u)\, du\, dw,$$

where $h^{(2)}(u, w) = \int_0^1 h(u, v)h(v, w)\, dv$ corresponds to $T^2 f \equiv \int_0^1 h^{(2)}(u, v) \times f(v)\, dv$.

In general, unions of $(\mathbf{k}, \mathbf{l})$-wheels are also more complicated. If R_1, R_2 are $(\mathbf{k}_1, \mathbf{l}_1)$, $(\mathbf{k}_2, \mathbf{l}_2)$-wheels which share a single node $[V(R_1) \cap V(R_2) = \{a\}]$, we can compute $P(R_1 \cup R_2) = \mathbb{E}P(R_1|\xi_a)P(R_2|\xi_a)$. If a is the hub of both wheels, then evidently $R_1 \cup R_2$ is itself a generalized wheel, and (2.10) applies. Otherwise, the formula, as for triangles, is more complex. However, such unions of $(\mathbf{k}, \mathbf{l})$-wheels are acyclic.

3. Moments and model identifiability. We establish two results in this section: identifiability of block models with known K using $\{\check{P}(R) : R \text{ a } (k, l)\text{-wheel}, 1 \le l \le 2K - 1, 2 \le k \le K\}$, and the general identifiability of the function w from $\{\check{P}(R)\}$ using all $(\mathbf{k}, \mathbf{l})$-wheels R.

3.1. *The block model.* Let w correspond to a K-block model defined by parameters $\theta \equiv (\pi, \rho_n, S)$, where π_a is the probability of a node being assigned to block a as before, and

$$F_{ab} \equiv \mathbb{P}(A_{ij} = 1|i \in a, j \in b) = \rho_n S_{ab}, \qquad 1 \le a, b \le K.$$

Recall that the function h in (1.2) is not unique, but a canonical h can be defined. For the block model, we use the canonical h given by Bickel and Chen (2009). Let $H_{ab} = S_{ab}\pi_a\pi_b$. Let the labeling of the communities $1, \dots, K$ satisfy $H_1 \le \cdots \le H_K$, where $H_a = \sum_b H_{ab}$ is proportional to the expected degree for a member of block a. The canonical function h then takes the value F_{ab} on the (a, b) block of the product partition where each axis is divided into intervals of lengths $\pi_1, \dots, \pi_K$. Let $F \equiv \|F_{ab}\|$.

In view of (2.6), we will treat ρ_n as known. Let $\{W_{kl} : 1 \le l \le 2K - 1, 2 \le k \le K\}$ be the specified set of (k, l)-wheels, and let

$$\tau_{kl} = \rho^{-kl} P(W_{kl}) = \tilde{P}(W_{kl}), \qquad \check{\tau}_{kl} = \check{P}(W_{kl}).$$

Let $f : \Theta \to \mathbb{R}^{(2K-1)(K-1)}$ be the map carrying the parameters of the block model $\theta \equiv (\pi, S)$ to $\tau \equiv \|\tau_{kl}\|$. Θ here is the appropriate open subset of $\mathbb{R}^{K(K+3)/2-2}$. Note that the number of free parameters in the block model is $K - 1$ for π and $K(K + 1)/2$ for F, but S only has $K(K + 1)/2 - 1$ free parameters, to account for ρ.

THEOREM 2. *Suppose* $\theta = (\pi, S)$ *defines a block model with known* K, *and the vectors* $\pi, F\pi, \dots, F^{K-1}\pi$ *are linearly independent. Suppose* $\varepsilon \le \lambda_n = o(n^{1/2})$. *Then*:

(a) $\{\tau_{kl} : l = 1, \dots, 2K-1, k = 2, \dots, K\}$ *identify the* $K(K+3)/2 - 2$ *parameters of the block model other than* ρ (*i.e., the map* f *is one to one*).

(b) *If* f *has a gradient which is of rank* $\frac{K(K+3)}{2} - 2$ *at the true* (π_0, S_0), *then* $f^{-1}(P(\check{\tau}))$ *is a* $\sqrt{n}$-*consistent estimate of* (π_0, S_0), *where* $\check{\tau} = \|\check{\tau}_{kl}\|$ *and* $P(\check{\tau})$ *is the closest point in the range of* f *to* $\check{\tau}$.

Note that the linear independence condition rules out all matrices F that have 1 as an eigenvector. In particular, it rules out the case of F_{aa} equal for all a, F_{ab} equal for all $a \ne b$, which was studied in detail by Decelle et al. (2011). Using physics arguments, they showed that in that particular case, when $\lambda = O(1)$, there are regions of the parameter space where neither the parameters nor the block assignments can be estimated by any method.

Part (b) shows $\sqrt{n}$-consistency of nonlinear least squares estimation of (π, S) using $\check{\tau}$ to estimate $\tilde{\tau}(\theta, S)$. The variance of $\check{\tau}_{kl}$ is proportional asymptotically to that of $\mathbb{E}\{\prod_{(i,j)\in S} w(\xi_i, \xi_j)|\xi_1\}$, where ξ_1 corresponds to the hub, which we expect increases exponentially in $p = kl + 1$. If we knew these variances, we could use weighted nonlinear least squares. In Section 5, we suggest a bootstrap method by which such variances can be estimated, but we do not pursue this further in this paper.

3.2. *The nonparametric model.* In the general case, we express everything in terms of the operator $T_w \equiv T/\rho_n$ induced by the canonical w. We require that:

(A) the joint distribution of $\{T_w^l(1)(\xi) : l \ge 1\}$ is determined by the cross moments of $(T_w^{l_1}(\xi), \dots, T_w^{l_k}(\xi))$, for $l_1, \dots, l_k$ arbitrary.

A simple sufficient condition for (A) is $|w| \le M < \infty$. A more elaborate one is the following:

(A′)

$$\mathbb{E}e^{sw^k(\xi_1,\xi_2)} < \infty, \qquad 0 \le |s| \le \varepsilon \text{ all } k \text{ some } \varepsilon > 0.$$

PROPOSITION 2. *Condition* (A′) *implies* (A).

The proof is given in the Appendix.

Let w characterize T_w, where $\int_0^1 w^2(u, v)\, du\, dv < \infty$. By Mercer's theorem,

$$w(u, v) = \sum_j \lambda_j \phi_j(u)\phi_j(v), \tag{3.1}$$

where the ϕ_j are orthonormal eigenfunctions and the λ_j eigenvalues, $\sum \lambda_j^2 < \infty$.

THEOREM 3. *Suppose* $\int_0^1 \int_0^1 w^2(u,v)\,du\,dv < \infty$. *Assume the eigenvalues* $\lambda_1 > \lambda_2 > \cdots$ *of* T_w *are each of multiplicity* 1 *with corresponding eigenfunction* ϕ_j, *and* $\int_0^1 \phi_j(u)\,du \neq 0$ *for all* j. *The joint distribution of* $(T_w(1)(\xi), \ldots, T_w^m(1)(\xi), \ldots)$ *then determines*, *and is determined by*, $w(\cdot,\cdot)$.

Note again that interesting cases are ruled out by the condition that all eigenfunctions of T are not orthogonal to 1. The general analogue to the block model case is that $P(A_{ij} = 1|\xi_i)$ cannot be constant for all i and j. Constancy can be interpreted as saying that A_{ij} and the latent variable ξ_i associated with vertex i are independent. The proof of Theorem 3 is given in the Appendix. The almost immediate application to wheels is stated next.

THEOREM 4. *Suppose assumption* (A) *and the conditions of Theorem* 3 *hold. Let* $\tau_{\mathbf{kl}} = \tilde{P}(S_{\mathbf{kl}})$ *where* $S_{\mathbf{kl}}$ *is a* $(\mathbf{k},\mathbf{l})$-*wheel. Then* $\mathcal{S} \equiv \{\tau_{\mathbf{kl}}$: *all* $\mathbf{k},\mathbf{l}\}$ *determines* T. *If* $\check{\tau}_{\mathbf{kl}} \equiv \check{P}(S_{\mathbf{kl}})$, $\check{\tau}_{\mathbf{kl}}$ *are* $\sqrt{n}$-*consistent estimates of* $\tau_{\mathbf{kl}}$, *provided that* $\lambda_n = o(n^{1/2})$.

PROOF. Since $\mathbf{T}_l \equiv (T(1)(\xi), \ldots, T^l(\xi))$ has a moment generating function converging on $0 < |s| \le \varepsilon_l$, the moments (including cross moments) determine the distribution of the vector. By (2.10), the $\tau_{\mathbf{kl}}$ give all moments of the vector $\mathbf{T}_l$ for all l. By Theorem 1, the $\check{\tau}_{\mathbf{kl}}$ are $\sqrt{n}$-consistent. $\square$

4. Degree distributions. The average degree $\bar{D}$ is, as we have seen in Theorem 1, a natural data dependent normalizer for moment statistics which eliminates the need to "know" ρ_n. In fact, as we show in this section, the joint empirical distribution of degrees and what we shall call m degrees below can be used in estimating asymptotic approximations to $w(\cdot,\cdot)$ in a somewhat more direct way than moment statistics. They can also be used to approximate moment estimates based on $(\mathbf{k},\mathbf{l})$-wheels in a way that potentially simplifies computation.

We define the m-degree of i, $D_i^{(m)}$, as the total number of loopless paths of length m between i and other vertices. Note that the $D_i^{(m)}$ can be interpreted as the "volume" of the radius m geodesic sphere around i. As for regular degrees, we normalize and consider $D_i^{(m)}/\bar{D}^m$, $i = 1, \ldots, n$, and the empirical joint distribution of vectors $\mathbf{D}_i^{(m)} \equiv (\frac{D_i}{\bar{D}}, \frac{D_i^{(2)}}{\bar{D}^2}, \ldots, \frac{D_i^{(m)}}{\bar{D}^m})$, $i = 1, \ldots, n$. The generalized degrees can be computed as follows: for all entries of A^m, eliminate all terms in the sum defining each entry in which an index appears more than once to obtain a modified matrix $\tilde{A}^{(m)} = [\tilde{A}_{ij}^{(m)}]$; then the $D_i^{(m)}$ are given by row sums of $\tilde{A}^{(m)}$. In other words, letting $A_{E(R)} = \prod_{(i,j)\in E(R)} A_{ij}$ we can write

$$\tilde{A}_{ij}^{(m)} = \sum \{A_{E(R)} : R = \{(i,i_1),(i_1,i_2),\ldots,(i_{m-1},j)\},$$
$$i, i_1, \ldots, i_{m-1}, j \text{ distinct}\}.$$

The complexity of this computation is $O((n+m)\lambda_n^m)$ (first term is for computing the row sums of A^m and the second for eliminating the loops).

Define the empirical distribution of the vector of normalized degrees

$$\hat{F}_m(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} 1\big(\mathbf{D}_i^{(m)} \le \mathbf{x}\big).$$

Further, recall the Mallows 2-distance between two distributions P and Q, defined by $M_2(P, Q) = \min_F\{(\mathbb{E}\|X - Y\|^2)^{1/2} : (X, Y) \sim F, X \sim P, Y \sim Q\}$. A sequence of distribution functions F_n converges to F in M_2 ($F_n \overset{M_2}{\to} F$) if and only if $F_n \Rightarrow F$ in distribution, and F_n, F have second moments such that $\int |\mathbf{x}|^2\, dF_n(\mathbf{x}) \to \int |\mathbf{x}|^2\, dF(\mathbf{x})$.

THEOREM 5. *Suppose $\lambda_n \to \infty$ and $|w_{2m}| < \infty$. Then $\hat{F}_m \overset{M_2}{\to} F_m$ as $n \to \infty$, where F_m is the distribution of $\boldsymbol{\theta}_m(\xi) = (\tau_w(\xi), \dots, T_w^{m-1}(\tau_w)(\xi))$, and $\tau_w(\xi) = \int_0^1 w(\xi, v)\, dv$ is monotone increasing. Moreover, if $\hat{G}_m(\mathbf{x}, \mathbf{y})$ is the empirical distribution of $(\mathbf{D}_i^{(m)}, \boldsymbol{\theta}_m(\xi_i))$, then*

$$\int |\mathbf{x} - \mathbf{y}|^2\, d\hat{G}_m(\mathbf{x}, \mathbf{y}) \overset{P}{\to} 0. \tag{4.1}$$

The proof is given in the Appendix.

There is an attractive interpretation of the last statement of Theorem 5. If $\lambda_n \to \infty$, $\lambda_n = o(n^{1/(m-1)})$, $m \ge 2$, then D_i/λ_n can be identified with $\tau(\xi_i)$ in the following sense: While ξ_i is unobserved but $D_i/\bar{D}$ is, on average, $\tau(\xi_i)$ and $D_i/\bar{D}$ are close. Since τ is monotone increasing in ξ, that is, is a measure of ξ on another scale, we can treat D_i/λ_n as the latent affinity of i to form relationships.

Bollobás, Janson and Riordan (2007) show that if $m = 1$, $\lambda_n = O(1)$, then the limit of the empirical distribution of the degrees can be described as follows: given $\xi \sim \mathcal{U}(0, 1)$, the limit distribution is Poisson with mean $\tau_w(\xi)$. The limit of the joint degree distribution in this case can be determined but does not seem to give much insight.

Remark. Theorem 5 shows that the normalized degree distributions can be used for estimation of parameters only if $\lambda_n \to \infty$. If that is the case we can proceed as follows:

(1) Let $\hat{\tau}_1, \dots, \hat{\tau}_n$ be the empirical quantiles of the normalized 1-degree distribution, and let $\hat{T}^m(\hat{\tau}_k)$ be the m-degree of the vertex with normalized degree $\hat{\tau}_k$.

(2) Fit smooth curves to $(\hat{\tau}_k, \hat{T}^m(\hat{\tau}_k))$ viewed as observations of functions at $\hat{\tau}_k$, $k = 1, \dots, n$, for each m, and call these $\hat{T}^m(\cdot)$ (on R). By Theorem 5, $\hat{T}^m(t) \to T^{m-1}(\tau)(\tau^{-1}(t))$ for all t. If $T^{m-1}(\tau^{-1}(\cdot))$ are smooth, the convergence can be made uniform on compacts.

(3) From the fitted functions $\hat{T}^m(\cdot)$, we can estimate the parameters of block models of any order consistently by replacing $\mathbf{v}_m$ in the proof of identifiability of block models by fitting the $\hat{T}^m(t)$ by $T^m(t)$ of the type specified by block models and then using the corresponding $\hat{\mathbf{v}}_m$. We only need the conditions of Theorem 5.

5. Computation of moment estimates and estimation of their variances. General acyclic graph moment estimates including those corresponding to patterns arising from $(\mathbf{k},\mathbf{l})$-wheels are computationally difficult. For (k,l)-wheels with small k and l, we can use brute force counting, but unfortunately, the complexity of moment computation even for (k,l)-wheels appears to be $O(n\lambda_n^k)$. Note that we need to count the sets of loopless paths of length k, $S_{i\mathbf{a}}$, for each i, where $S_{i\mathbf{a}}$ is the set of all paths of length k originating at node i which intersect another such path at $a_1 < \cdots < a_m$, $1 \le m \le k$, and S_{i0} is the set of all paths of length k from i which do not intersect. The number of (k,l)-wheels with hub i is then the number of l-tuples of such paths selected so that elements from $S_{i\mathbf{a}}$ appear at most once, with the remaining paths coming from S_{i0}. This is computationally nontrivial.

For very sparse graphs, however, intersecting paths can be ignored up to a certain order, and the wheel counts can be related to normalized m-degrees via a following approximation. If the conditions of Theorem 5 hold and $\lambda_n = o(n^\alpha)$ for all $\alpha > 0$, then

$$\hat{\tau}_{kl} = \frac{1}{n}\sum_{i=1}^{n} \frac{(D_i^{(k)})_l}{\bar{D}^{kl}} + o_P(n^{-1/2}). \tag{5.1}$$

A similar formula holds for $\hat{\tau}_{\mathbf{kl}}$.

The heuristic argument for (5.1) is that the expected number of paths of lengths k from i is $O(\lambda_n^k)$. The expected number of pairs of such paths which intersect at least once is

$$\begin{aligned} &O(\lambda_n^{2k})\mathbb{P}[\text{two specified paths intersect at least once}] \\ &\quad = O\big(\lambda_n^{2k}(1-(1-\lambda_n/n)^k)\big) = O\left(\frac{k\lambda_n^{2k+1}}{n}\right) = o(1), \end{aligned}$$

if $\lambda_n = o(n^\alpha)$ for all $\alpha > 0$. Note that for K-block models this condition is not necessary for all α, since we only need to count a finite number of (k,l)-wheels.

Estimation of variances of moment estimates even for $(\mathbf{k},\mathbf{l})$-wheels involve the counting of more complicated patterns. However, we propose the following bootstrap method:

(i) Associate with each vertex i the counts of $(\mathbf{k},\mathbf{l})$-wheels for which it is a hub, $S_i = \{n_{i\mathbf{kl}}\colon \text{all } \mathbf{k},\mathbf{l}\}$, $i = 1,\dots,n$.

(ii) Sample without replacement m vertices $\{i_1,\dots,i_m\}$, and let

$$\bar{D}^* = \frac{1}{m}\sum_{j=1}^{m} D_{i_j}.$$

For R a $(\mathbf{k},\mathbf{l})$-wheel, define

$$\hat{P}^*(R) = \frac{(n/m)\sum_{j=1}^{m} n_{i_j\mathbf{kl}}}{\binom{n}{p}N(R)},$$

$$\check{P}^*(R) = \hat{P}^*(R)\left(\frac{\bar{D}^*}{m}\right)^{-|R|}.$$

(iii) Repeat this B times to obtain $\check{P}_1^*, \dots, \check{P}_B^*$, and let

$$\hat{\sigma}^2 = \frac{m}{n}\frac{1}{B}\sum_{b=1}^{B}(\check{P}_b^* - \check{P}_\cdot^*)^2.$$

Then $\hat{\sigma}^2$ is an estimate of the variance of $\check{P}(R)$ if $\frac{m}{n} \to 0, m \to \infty$.

This scheme works if $\lambda_n \to \infty$ since, given that the first term of $\check{P}(R) - \tilde{P}(R)$ is of lower order given $\xi_1, \dots, \xi_n$, each $\tilde{P}^*(R)$ corresponds to a sample without replacement from the set of possible $\{\xi_i\}$. We conjecture that this bootstrap still works if $\lambda_n = O(1)$. A similar device can be applied to approximation (5.1).

6. Discussion.

6.1. *Estimation of canonical w generally.* Our Theorem 4 suggests that we might be able to construct consistent nonparametric estimates of w_{CAN}. That is, $\boldsymbol{\tau}_M = \{\tau_{\mathbf{kl}} : |\mathbf{k}| \leq M, |\mathbf{l}| \leq M\}$ can be estimated at rate $n^{-1/2}$ for all $M < \infty$. But $\{\boldsymbol{\tau}_M, M \geq 1\}$ determines T_w, and thus in principle we can estimate T_w arbitrarily closely using $\{\hat{\tau}_{\mathbf{kl}}\}$. This appears difficult both theoretically and practically. Theoretically, one difficulty seems to be that we would need to analyze the expectation of moments or degree distributions when the block model does not hold, which is doable. What is worse is that the passage to w from moments is very ill-conditioned, involving first inversion via solution of the moment problem, and then estimation of eigenvectors and eigenvalues from a sequence of iterates $T_w(1), T_w^2(1)$, etc. If we assume $\lambda_n \to \infty$ so that we can use consistency of the degree distributions, we bypass the moment problem, but the eigenfunction estimation problem remains. A step in this direction is a result of Rohe, Chatterjee and Yu (2011) which shows that spectral clustering can be used to estimate the parameters of k block models if $\lambda \to \infty$ sufficiently, even if $k \to \infty$ slowly. Unfortunately this does not deal with the problem we have just discussed, how to pick a block model which is a good approximation to the nonparametric model. For reasons which will appear in a future paper, smoothness assumptions on w have to be treated with caution.

While $\lambda_n \to \infty$ has not occurred in practice in the past, networks with high average degrees are now appearing routinely. In particular, university Facebook networks have λ of 15 or more with n in the low thousands. In any case $\lambda_n \to \infty$ can still be useful as an asymptotic regime that can help us understand some general patterns, in the same way that the sample size going to infinity does in ordinary statistics. Note that most of the time we do not specify the rate of growth of λ_n, which can be very slow.

6.2. *Adding covariates and directed graphs.* In principle, adding covariates X_i at each vertex or X_{ij} at each edge simply converts our latent variable model, $w(\cdot,\cdot)$ into a mixed model

$$\mathbb{P}_\theta(A_{ij} = 1 | X_i, X_j, X_{ij}, \xi_i, \xi_j) = w_\theta(\xi_i, \xi_j, X_i, X_j, X_{ij}),$$

which can be turned into a logistic mixed model. Special cases of such models have been considered in the literature; see Hoff (2007) and references therein. We do not pursue this here. The extension of this model to directed graphs is also straightforward.

6.3. *Dynamic models.* Many models in the literature have been specified dynamically; see Newman (2010). For instance, the "preferential attachment" model constructs an n graph by adding 1 vertex at a time, with edges of that vertex to previous vertices formed with probabilities which are functions of the degree of the candidate "old" vertex. If we let $n \to \infty$, we obtain models of the type we have considered whose w function can be based on an integral equation for $\tau(\xi)$, our proxy for the degree of the vertex with latent variable ξ. We shall pursue this elsewhere also.

APPENDIX: ADDITIONAL LEMMAS AND PROOFS

PROOF OF PROPOSITION 1. The first line of (2.2) is immediate, conditioning on $\{\xi_1, \dots, \xi_n\}$. The second line in (2.2) follows by expanding the second product. Finally, (2.2) follows directly from the definitions of P and Q. $\square$

The following standard result is used in the proof of Theorem 1.

LEMMA 1. *Suppose* (U_n, V_n) *are random elements such that*,

$$\mathcal{L}(U_n) \longrightarrow \mathcal{L}(U),$$
$$\mathcal{L}(V_n | U_n) \longrightarrow \mathcal{L}(V)$$

in probability. Then U_n, V_n *are asymptotically independent*,

$$\mathcal{L}(V_n) \longrightarrow \mathcal{L}(V).$$

PROOF OF THEOREM 1. By definition, $\mathbb{E}(\frac{L}{n\lambda_n}) = \frac{1}{2}$. Moreover,

$$\begin{aligned}\mathrm{Var}\biggl(\frac{1}{n\lambda_n}\sum\{A_{ij}\colon \text{all } 1 \le i < j \le n\}\biggr) &= (n\lambda_n)^{-2}\mathbb{E}\biggl(\mathrm{Var}\biggl(\sum_{i<j} A_{ij} | \boldsymbol{\xi}\biggr)\biggr) \\ &\quad + \rho_n^2 (n\lambda_n)^{-2}\,\mathrm{Var}\biggl(\sum_{i<j} w(\xi_i, \xi_j)\biggr) \\ &\equiv \mathrm{Var}(T_1) + \mathrm{Var}(T_2),\end{aligned}$$

where

$$T_1 = (n\lambda_n)^{-1} \sum_{i<j} (A_{ij} - \rho_n w(\xi_i, \xi_j)),$$

$$T_2 = \rho_n (n\lambda_n)^{-1} \sum_{i<j} w(\xi_i, \xi_j) - \frac{1}{2}.$$

Since $\lambda_n = (n-1)\rho_n$, the first term is

$$\begin{aligned}(n\lambda_n)^{-2}\mathbb{E}\sum\{h(\xi_i, \xi_j)(1 - h(\xi_i, \xi_j)) \text{ all } i, j\} \\ \le \frac{\rho_n n^2}{2n^2\lambda_n^2} = O((n^2\rho_n)^{-1}) = O((n\lambda_n)^{-1}).\end{aligned}$$

The second term is a U-statistic of order 2, which is well known to be $O(n^{-1})$. Thus, (2.5) follows in case (a).

To establish (2.6) and (b), we note that the conditional distribution of $\sqrt{n\lambda_n}T_1$ given $\boldsymbol{\xi}$ is that of a sum of independent random variables with conditional variance

$$\frac{1}{n\lambda_n} \sum_{i<j} \rho_n w(\xi_i, \xi_j)(1 - \rho_n w_n(\xi_i, \xi_j)) = \frac{1}{n^2} \sum_{i<j} w(\xi_i, \xi_j)(1 + o_P(1)) \xrightarrow{P} \frac{1}{2}.$$

This sum is approximated by a U-statistic of order 2. Note that $\mathbb{E}w(\xi_i, \xi_j) = 1$. Since the max of the summands in $\sqrt{n\lambda_n}T_1$ is $\frac{1}{\sqrt{n\lambda_n}} \to 0$, by the Lindeberg–Feller theorem, the conditional distribution tends to $\mathcal{N}(0, \frac{1}{2})$ in probability. We can similarly apply the limit theorem for U-statistics [see Serfling (1980)] to conclude that

$$\sqrt{n}T_2 \Rightarrow \mathcal{N}(0, \mathrm{Var}(\tau(\xi))).$$

Applying Lemma 1, we see that if $\lambda_n = O(1)$, (b) follows. On the other hand, if $\lambda_n \to \infty$, $\sqrt{n}T_1$ is negligible, and the Gaussian limit is determined by T_2.

The proof of (2.7) and (2.8) is similar. We shall decompose $\check{P}(R)$ as $U_1 + U_2$ as we did $\frac{L}{n\lambda_n}$. If $\lambda_n \to \infty$, it is enough to prove that

$$\sqrt{n}(\check{P}(R) - \tilde{P}(R)) \Rightarrow \mathcal{N}(0, \sigma^2(R))$$

since replacing $\bar{D}$ by $n\rho_n = \lambda_n$ gives a perturbation of order $(n\lambda_n)^{-1/2} = o(n^{-1/2})$.

In case (b), it is enough to show that the joint distribution of $\sqrt{n}((\hat{P}(R) - P(R))\rho_n^{-|R|}, T_1, T_2)$ is Gaussian in the limit, since in view of (2.5) and (2.6) we can apply the delta method to $\check{P}(R)$. Let $p \equiv |V(R)|$, $q \equiv |R|$. Each term in $\check{P}(R)$ is of the form

$$T(S) \equiv \frac{1}{\binom{n}{p}N(R)} \prod\{A_{i_l j_l} : (i_l, j_l) \in E(S), S \sim R\}.$$

Condition on $\boldsymbol{\xi} = \{\xi_1, \dots, \xi_n\}$. Then terms $T(S)$, as above, yield

$$\mathbb{E}(\hat{P}(R)|\boldsymbol{\xi}) = \frac{1}{\binom{n}{p} N(R)} \sum_{S \sim R} \biggl(\prod_{(i,j) \in E(S)} [w(\xi_i, \xi_j)] \biggr) + O(n^{-1}\lambda_n). \tag{A.1}$$

Thus,

$$U_2 = \mathbb{E}(\hat{P}(R)|\boldsymbol{\xi})\rho_n^{-q} - P(R),$$
$$U_1 = \rho_n^{-q} \sum \{T(S) - E(T(S)|\boldsymbol{\xi}) : S \sim R\}.$$

We begin by considering $\mathrm{Var}(U_1|\xi)$ which we can write as

$$\sum \mathrm{cov}(T(S_1), T(S_2)|\boldsymbol{\xi})\rho_n^{-2q},$$

where the sum ranges over all $S_1 \sim R$, $S_2 \sim R$.

If $E(S_1) \cap E(S_2) = \phi$ the covariance is 0. In general, suppose the graph $S_1 \cap S_2$ has c vertices and d edges. Since R is acyclic any subgraph is acyclic. By Corollary 3.2 of Chartrand, Lesniak and Behzad (1986) for every acyclic graph, $|V(S)| \geq |E(S)| + 1$. Now,

$$\rho_n^{-2q} \,\mathrm{cov}(T(S_1), T(S_2)|\boldsymbol{\xi}) \leq n^{-2p} \rho_n^{-d} \prod_{(i,j) \in S_1 \cup S_2} w_n(\xi_i, \xi_j) \tag{A.2}$$

since, if $d \geq 1$,

$$\begin{aligned} &\mathbb{E}\Bigl[\prod \{A_{ij} : (i,j) \in \overline{S_1 \cap S_2}\} \prod \{A_{ij}^2 : (i,j) \in S_1 \cap S_2\} |\boldsymbol{\xi}\Bigr] \\ &\quad = \rho_n^{2q-d} \prod \{w_n(\xi_i, \xi_j) : (i,j) \in S_1 \cup S_2\}. \end{aligned} \tag{A.3}$$

There are $O(n^{2p-c})$ terms in (A.1) which have c vertices in common. Therefore by (A.2) the total contribution of all such terms to $\mathrm{Var}(U_1)$ is

$$O\biggl(n^{-c} \rho_n^{-d} \int w^{2q}(u,v) \, du \, dv\biggr),$$

after using Hölder's inequality on $\mathbb{E} \prod \{w(\xi_i, \xi_j) : (i,j) \in S_1 \cup S_2\}$. From (A.3) and our assumptions we conclude that

$$\mathrm{Var}(U_1) = O(n^{-1}\lambda_n^{-d}) = o(n^{-1}),$$

if $\lambda_n \to \infty$. On the other hand

$$U_2 = \frac{1}{\binom{n}{p} N(R)} \sum_{S \sim R} \biggl\{ \prod_{(i,j) \in S} w(\xi_i, \xi_j) \prod_{(i,j) \in \bar{S}} \bigl(1 - h_n(\xi_i, \xi_j)\bigr) - \tilde{P}(S) \biggr\}$$

is a U-statistic. Its kernel

$$\prod_S w(\xi_i, \xi_j) \prod_{\bar{S}} \bigl(1 - h_n(\xi_i, \xi_j)\bigr) - \tilde{P}(S) \quad \stackrel{L_2}{\to} \quad \prod_S w(\xi_i, \xi_j) - \mathbb{E} \prod_S w(\xi_i, \xi_j).$$

Thus, $\sqrt{n}(U_1, U_2)$ are jointly asymptotically Gaussian; see, for instance, Serfling (1980).

Since if $\lambda_n \to \infty$, $T_1, U_1 = o_P(n^{-1/2})$, the result follows if $\lambda_n \to \infty$. If $\lambda_n = O(1)$, we note that $\sqrt{n}(T_1, U_1)$ are sums of q dependent random variables in the sense of Bulinski [see Doukhan (1994)] and hence, given $\boldsymbol{\xi}$, are jointly asymptotically Gaussian. It is not hard to see that the limiting conditional covariance matrix is independent of ξ, as it was for T_1 marginally. By Lemma 1 again (T_1, U_1) and (T_2, U_2) are asymptotically independent and (a) and (b) follow.

Finally we prove (c). To have $n^{-1/2}$ consistency for $\check{P}(R)$, $\tilde{P}(R)$ and hence for $\check{Q}(R)$, $\tilde{Q}(R)$ by (2.3) we need to argue that if $S \subset R$, $c \equiv |S| \le p$ $|E(S)| = d$, then for a universal M,

$$n^{-c}\rho^{-d} \le Mn^{-1}.$$

Since $\rho = \frac{\lambda_n}{n}$ we obtain

$$n^c\left(\frac{\lambda_n}{n}\right)^d \ge n, \qquad \lambda_n \ge n^{1-(c-1)/d}.$$

For fixed $c \ge 1$ this is maximized by $d = \frac{c(c-1)}{2}$ and $n^{1-2/c}$ is maximized for $c \le p$ by $c = p$. $\square$

PROOF OF THEOREM 2. Since T corresponds to the canonical h,

$$T(1)(\xi) = v_{(1)}, \qquad 0 \le \xi \le \pi_1,$$

$$T(1)(\xi) = v_{(j)}, \qquad \sum_{k=1}^{j-1}\pi_k \le \xi \le \sum_{k=1}^{j}\pi_k, \qquad 1 \le j \le K,$$

where $v_{(1)} < \cdots < v_{(k)}$ are the ordered $\{v_j\}$, $v_j = \sum_{i=1}^K \pi_i F_{ij}$. By a theorem of Hausdorff and Hamburger [Feller (1971)], the distribution of the random variable $T(1)(\xi_1)$ which takes on only K distinct values above is completely determined and uniquely so by its first $2K-1$ moments $\mathbb{E}(T(1)(\xi_1))^l$, $l = 1, \dots, 2K-1$. Therefore for our model $\pi_1, \dots, \pi_K$ are completely determined since $T(1)(\xi_1)$ takes values v_j with probability π_j, $j = 1, \dots, K$.

Let $v^{(1)} = (v_{(1)}, \dots, v_{(K)})^T = F\pi$. Note that $\mathbb{E}(T^2(1)(\xi_1))^l$, $l = 1, \dots, 2K-1$, similarly determines the distribution of $T^2(1)(\xi_1)$. Hence,

$$v^{(2)} = Fv^{(1)}.$$

Continuing we see that the $(K-1)(2K-1)$ moments $\{\tau_{kl} : 2 \le k \le K, 1 \le l \le 2K-1\}$ yield

$$v^{(j)} = Fv^{(j-1)} \tag{A.4}$$

for $j = 1, \dots, K$ where $v^{(0)} \equiv \pi$.

Given $\pi, v^{(1)}, \dots, v^{(K)}$ linearly independent, we can compute F since by (A.4), we can write

$$F_{K\times K} V^{(1)}_{K\times K} = V^{(2)}_{K\times K},$$

where $V^{(1)} = (v^{(0)}, \dots, v^{(K-1)})^T$ and $V^{(2)} = (v^{(1)}, \dots, v^{(K)})^T$ and hence

$$F = V^{(2)}\big[V^{(1)}\big]^{-1}.$$

Consistency and $\sqrt{n}$-consistency follow from Theorem 1 and the delta method. $\square$

PROOF OF PROPOSITION 2. Note that

$$\begin{aligned}\mathbb{E}\exp sT^l(1)(\xi) &= \mathbb{E}\exp s\mathbb{E}\big(w(\xi,\xi_1)\cdots w(\xi_{l-1},\xi_l)|\xi\big)\\ &\le \mathbb{E}\exp s\big(w(\xi,\xi_1)\cdots w(\xi_{l-1},\xi_l)\big).\end{aligned}\tag{A.5}$$

Taking $\xi = \xi_0$,

$$\text{(A.5)} \le \mathbb{E}\exp|s|\left(\frac{1}{l}\sum_{j=0}^{l} w^l(\xi_j,\xi_{j+1})\right)\tag{A.6}$$

by the arithmetic/geometric mean and Minkowski inequalities. By Hölder's inequality (A.6) is bounded by

$$\prod_{j=0}^{l}[\mathbb{E}\exp|s|w^l(\xi_j,\xi_{j+1})]^{1/l}.$$

It is easy to show that (A$'$) implies that $\mathbb{E}\exp\{\sum_{j=1}^m s_j T^j(1)(\xi)\}$ converges for $0 < |s| < \varepsilon$ for some ε depending on m and hence by a classical result that (A$'$) implies (A). $\square$

PROOF OF THEOREM 3. Clearly w determines the joint distribution of moments. We can take $\tau_w(\xi) = T_w(1)(\xi)$ monotone, corresponding to the canonical w, to be the quantile function of the marginal distribution of $T_w(1)(\xi)$. Now the joint distribution of $(T_w(1)(\xi), T_w^2(1)(\xi))$ determines $\tau_w(\cdot)$, $T_w\tau_w(\cdot)$, except on a set of measure 0. Continuing this argument, we can determine the entire sequence of functions $\tau_w, T_w\tau_w, T_w^2\tau_w, \dots$. Since T_w is bounded self-adjoint, these functions are all in L_2. Let $g_k^{(1)}(\cdot) = T_w(\frac{g_{k-1}^{(1)}}{|g_{k-1}^{(1)}|})$, $g_0^{(1)}(\cdot) = 1$, where $|f|$ and (f,g) are, respectively, the norm and the inner product in L_2. Then $g_k \to_{L_2} \lambda_1\phi_1$ where λ_1 is the first eigenvalue, ϕ_1 the first eigenfunction and $\frac{g_k}{|g_k|} \to \phi_1$. This is just the "powering up" method applied to the function 1 with convergence guaranteed since λ_1 is unique, and 1 is not orthogonal to ϕ_1 or any other eigenfunction. So λ_1 and ϕ_1 are also determined. Thus we can compute $g_0^{(2)} \equiv 1 - (1, \phi_1)\phi_1$. Further,

$$g_1^{(2)} = T_w\left(\frac{g_0^{(2)}}{|g_0^{(2)}|}\right) = \frac{T_w 1(\cdot) - \lambda_1(1, \phi_1)\phi_1}{|1 - (1, \phi_1)\phi_1|}$$

is computable since we know $T_w 1(\cdot)$ and the eigenfunction ϕ_1 and eigenvalue λ_1. More generally, $T_w^k g_1^{(2)}$, $|g_{k-1}^{(2)}|$ can be similarly determined. Then, by the same argument as before, using 1 not orthogonal to ϕ_2, we obtain $g_k^{(1)} \to_{L_2} \lambda_2\phi_2$ and $g_k^{(1)}/|g_k^{(1)}| \to_{L_2} \phi_2$. Now form $g_0^{(3)} \equiv 1 - \lambda_1(1, \phi_1)\phi_1 - \lambda_2(1, \phi_2)\phi_2$ and proceed as before, and continue to determine λ_k, ϕ_k for all k. This and (3.1) complete the proof. $\square$

PROOF OF THEOREM 5. Note first that (4.1) implies that the M_2 distance between $\hat{F}_m$ and the empirical distribution of $\{\boldsymbol{\theta}_m(\xi_i)\}$ tends to 0. The first conclusion of the theorem now follows by the Glivenko–Cantelli theorem and the Law of Large Numbers.

To show (4.1), note that

$$\frac{1}{n}\sum_{i=1}^{n}|\tilde{D}_i^{(m)} - \theta_m(\xi_i)|^2 \xrightarrow{P} 0, \tag{A.7}$$

where $\tilde{D}_i^{(m)} \equiv (\frac{D_i}{\bar{D}}, \ldots, \frac{D_i^{(m)}}{\bar{D}^m})^T$. By Theorem 1, we can replace $\bar{D}$ by λ_n if $\lambda_n \geq \varepsilon$. Then (A.7) is implied by

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left|\sum_{j=1}^{n}\frac{\tilde{A}_{ij}^{(m)}}{\lambda_n^m} - \theta_m(\xi_i)\right|^2 \to 0. \tag{A.8}$$

Now,

$$\begin{aligned}\sum_{j=1}^{n}\mathbb{E}\left(\frac{\tilde{A}_{ij}^{(m)}}{\lambda_n^m}\middle|\boldsymbol{\xi}\right) &= \frac{1}{n^m}\sum\{w_{E(R)} : R = \{(i, i_1), \ldots, (i_{m-1}, j)\},\\ &\qquad\qquad \text{all vertices distinct}\},\end{aligned} \tag{A.9}$$

where $w_{E(R)} = \prod_{(a,b)\in E(R)} w(\xi_a, \xi_b)$. Further, (A.9) is a U-statistic of order m under $|w_{2m}| < \infty$ and

$$\mathbb{E}\left|\sum_{j=1}^{n}\mathbb{E}\left(\frac{\tilde{A}_{ij}^{(m)}}{\lambda_n^m}\middle|\boldsymbol{\xi}\right) - \mathbb{E}(w_{E(R)}|\xi_i)\right|^2 \leq \frac{C|w_{2m}|}{n}$$

by standard theory [Serfling (1980)].

Since $\mathbb{E}(w_{E(R)}|\xi_i) = \theta_m(\xi_i)$, we can consider

$$\begin{aligned} &\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left|\sum_{j=1}^{n}\frac{\tilde{A}_{ij}^{(m)} - \mathbb{E}(\tilde{A}_{ij}^{(m)}|\boldsymbol{\xi})}{\lambda_n^m}\right|^2\right) \\ &\quad \le \max_i \frac{\mathbb{E}|\sum_{j=1}^{n}(\tilde{A}_{ij}^{(m)} - \mathbb{E}(\tilde{A}_{ij}^{(m)}|\boldsymbol{\xi}))|^2}{\lambda_n^{2m}}. \end{aligned} \tag{A.10}$$

Note that $R = \{(i, i_1), (i_1, i_2), \ldots, (i_{m-1}, j)\}$ is acyclic if all vertices are distinct. As in the proof of Theorem 1, all nonzero covariance terms in (A.10) are of order $\rho^{2m-d}n^{2m-c}$ where $c \ge d$ since the intersection graphs all have i in common but are otherwise acyclic. The largest order term corresponds to $c = d = m$, so that

$$\mathbb{E}\left|\sum_{j=1}^{n}(\lambda_n^{-m}\tilde{A}_{ij}^{(m)} - \theta_m(\xi_i))\right|^2 \le C\lambda_n^{-m},$$

where C depends on $|w_{2m}|$ only. Thus (A.8) holds if $\lambda_n \to \infty$. $\square$

Acknowledgment. Thanks to Allan Sly for a helpful discussion.

REFERENCES

ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937

BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BOLLOBÁS, B., JANSON, S. and RIORDAN, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures Algorithms* **31** 3–122. MR2337396

CHARTRAND, G., LESNIAK, L. and BEHZAD, M. (1986). *Graphs and Digraphs*, 2nd ed. Wadsworth and Brooks, Monterey, CA.

CHATTERJEE, S. and DIACONIS, P. (2011). Estimating and understanding exponential random graph models. Unpublished manuscript. Available at arXiv:1102.2650.

CHUNG, F. and LU, L. (2002). Connected components in random graphs with given expected degree sequences. *Ann. Comb.* **6** 125–145. MR1955514

DE SOLLA PRICE, D. J. (1965). Networks of scientific papers. *Science* **149** 510–515.

DECELLE, A., KRZAKALA, F., MOORE, C. and ZDEBOROVÁ, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Available at arXiv:1109.3041.

DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439

DOUKHAN, P. (1994). *Mixing: Properties and Examples. Lecture Notes in Statistics* **85**. Springer, New York. MR1312160

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications. Vol. II*, 2nd ed. Wiley, New York. MR0270403

FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. MR0860518

HANDCOCK, M. (2003). Assessing degeneracy in statistical models of social networks. Working Paper 39, Center for Statistics and the Social Sciences.

HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300

HOFF, P. D. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems* **19**. MIT Press, Cambridge, MA.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088

HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. MR0608176

HOOVER, D. (1979). Relations on probability spaces and arrays of random variables. Technical report, Institute for Advanced Study, Princeton, NJ.

KALLENBERG, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer, New York. MR2161313

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107. MR2788206

NEWMAN, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* (3) **74** 036104. MR2282139

NEWMAN, M. E. J. (2010). *Networks*: *An Introduction*. Oxford Univ. Press, Oxford. MR2676073

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255

PICARD, F., DAUDIN, J. J., KOSKAS, M., SCHBATH, S. and ROBIN, S. (2008). Assessing the exceptionality of network motifs. *J. Comput. Biol.* **15** 1–20. MR2383618

ROBINS, G., SNIJDERS, T., WANG, P., HANDCOCK, M. and PATTISON, P. (2007). Recent developments in exponential random graphs models (p^*) for social networks. *Social Networks* **29** 192–215.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic block model. *Ann. Statist.* To appear.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

SHALIZI, C. R. and RINALDO, A. (2011). Projective structure and parametric inference in exponential families. Carnegie Mellon Univ. Unpublished manuscript.

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742

P. J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
367 EVANS HALL
BERKELEY, CALIFORNIA 94720-3860
USA
E-MAIL: bickel@stat.berkeley.edu

A. CHEN
GOOGLE INC.
1600 AMPHITHEATRE PKWY
MOUNTAIN VIEW, CALIFORNIA 94043
USA
E-MAIL: aiyouchen@google.com

E. LEVINA
DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
439 WEST HALL
1085 S. UNIVERSITY AVE.
ANN ARBOR, MICHIGAN 48109-1107
USA
E-MAIL: elevina@umich.edu