Biometrika (2012), **99**, 2, *pp.* 273–284
doi: 10.1093/biomet/asr053
Advance Access publication 17 April 2012

Stochastic blockmodels with a growing number of classes

By D. S. CHOI

School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, U.S.A.
dchoi@seas.harvard.edu

P. J. WOLFE

Department of Statistical Science, University College London, London WC1E 6BT, U.K.
patrick@stats.ucl.ac.uk

and E. M. AIROLDI

Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.
airoldi@fas.harvard.edu

Summary

We present asymptotic and finite-sample results on the use of stochastic blockmodels for the analysis of network data. We show that the fraction of misclassified network nodes converges in probability to zero under maximum likelihood fitting when the number of classes is allowed to grow as the root of the network size and the average network degree grows at least polylogarithmically in this size. We also establish finite-sample confidence bounds on maximum-likelihood blockmodel parameter estimates from data comprising independent Bernoulli random variates; these results hold uniformly over class assignment. We provide simulations verifying the conditions sufficient for our results, and conclude by fitting a logit parameterization of a stochastic blockmodel with covariates to a network data example comprising self-reported school friendships, resulting in block estimates that reveal residual structure.

Some key words: Likelihood-based inference; Social network analysis; Sparse random graph; Stochastic blockmodel.

1. Introduction

The global structure of social, biological, and information networks is sometimes envisioned as the aggregate of many local interactions whose effects propagate in ways that are not yet well understood. There is increasing opportunity to collect data on an appropriate scale for such systems, but their analysis remains challenging (Goldenberg et al., 2009). Here we analyse a statistical model for network data known as the single-membership stochastic blockmodel. Its salient feature is that it partitions the N nodes of a network into K distinct classes whose members all interact similarly with the network. Blockmodels were first associated with the deterministic concept of structural equivalence in social network analysis (Lorrain & White, 1971), where two nodes were considered interchangeable if their connections were equivalent in a formal sense. This concept was adapted to stochastic settings and gave rise to the stochastic blockmodel in the work by Holland et al. (1983) and Fienberg et al. (1985). The model and extensions thereof have since been applied in a variety of disciplines

(Airoldi et al., 2008; Hoff, 2008; Nowicki & Snijders, 2001; Girvan & Newman, 2002; Handcock et al., 2007; Copic et al., 2009; Mariadassou et al., 2010; Karrer & Newman, 2011).

In this work we provide a finite-sample confidence bound that can be used when estimating network structure from data modelled by independent Bernoulli random variates, and also show that under maximum likelihood fitting of a correctly specified K-class blockmodel, the fraction of misclassified network nodes converges in probability to zero even when the number of classes K grows with N. As noted by Rohe et al. (2011) this is advantageous if we expect class sizes to remain relatively constant even as N increases. Related results for fixed K have been shown by Snijders & Nowicki (1997) for networks with a linearly increasing degree, and in a stronger sense for sparse graphs with poly-logarithmically increasing degrees by Bickel & Chen (2009).

Our results can be related to those of Rohe et al. (2011), who use spectral methods to bound the number of misclassified nodes in the stochastic blockmodel with increasing K, although with the more restrictive requirement of nearly linearly increasing degree. As noted by those authors, this assumption may not hold in many practical settings. Our manner of proof requires only poly-logarithmically increasing degree, and is more closely related to the fixed-K proof of Bickel & Chen (2009), although we note that spectral clustering as suggested by Rohe et al. (2011) provides a computationally appealing alternative to maximum likelihood fitting in practice.

As discussed by Bickel & Chen (2009), one may assume exchangeability in lieu of a generative K-class blockmodel: an analogue to de Finetti's theorem for exchangeable sequences states that the probability distribution of an infinite exchangeable random graph is expressible as a mixture of distributions whose components can be approximated by blockmodels (Kallenberg, 2005; Bickel & Chen, 2009). An observed network can then be viewed as a sample drawn from this infinite conceptual population, and so in this case the fitted blockmodel describes one mixture component thereof.

2. Statement of results

2·1. *Problem formulation and definitions*

We consider likelihood-based inference for independent Bernoulli data $\{A_{ij}\}$ $(i = 1, \ldots, N;\ j = i+1, \ldots, N)$, both when no structure linking the success probabilities $\{P_{ij}\}$ is assumed, as well as the special case when a stochastic blockmodel of known order K is assumed to apply. To this end, let $A \in \{0, 1\}^{N \times N}$ denote the symmetric adjacency matrix of a simple, undirected graph on N nodes whose entries $\{A_{ij}\}$ for $i < j$ are assumed independent $\mathrm{Ber}(P_{ij})$ random variates, and whose main diagonal $\{A_{ii}\}_{i=1}^{N}$ is fixed to zero. The average degree of this graph is $2M/N$, where $M = \sum_{i<j} P_{ij}$ is its expected number of edges. Under a K-class stochastic blockmodel, these edge probabilities are further restricted to satisfy

$$P_{ij} = \theta_{z_i z_j} \quad (i = 1, \ldots, N;\ j = i+1, \ldots, N) \tag{1}$$

for some symmetric matrix $\theta \in [0, 1]^{K \times K}$ and membership vector $z \in \{1, \ldots, K\}^N$. Thus the probability of an edge between two nodes is assumed to depend only on the class of each node.

Let $L(A; z, \theta)$ denote the loglikelihood of observing data matrix A under a K-class blockmodel with parameters (z, θ), and $\bar{L}_P(z, \theta)$ its expectation:

$$L(A; z, \theta) = \sum_{i<j} \{A_{ij} \log \theta_{z_i z_j} + (1 - A_{ij}) \log(1 - \theta_{z_i z_j})\},$$

$$\bar{L}_P(z, \theta) = \sum_{i<j} \{P_{ij} \log \theta_{z_i z_j} + (1 - P_{ij}) \log(1 - \theta_{z_i z_j})\}.$$

For fixed class assignment z, let N_a denote the number of nodes assigned to class a, and let n_{ab} denote the maximum number of possible edges between classes a and b; i.e., $n_{ab} = N_a N_b$ if $a \neq b$ and $n_{aa} = N_a!/\{(N_a - 2)!2!\}$. Further, let $\hat{\theta}^{(z)}$ and $\bar{\theta}^{(z)}$ be symmetric matrices in $[0, 1]^{K \times K}$, with

$$\hat{\theta}^{(z)}_{ab} = \frac{1}{n_{ab}} \sum_{i<j} A_{ij} 1(z_i = a, z_j = b),$$

$$\bar{\theta}^{(z)}_{ab} = \frac{1}{n_{ab}} \sum_{i<j} P_{ij} 1(z_i = a, z_j = b) \quad (a = 1, \ldots, K; b = a, \ldots, K)$$

defined whenever $n_{ab} \neq 0$. Observe that $\hat{\theta}^{(z)}$ comprises sample proportion estimators as a function of z, whereas $\bar{\theta}^{(z)}$ is its expectation under the independent $\{\mathrm{Ber}(P_{ij})\}$ model. Taken over all class assignments $z \in \{1, \ldots, K\}^N$, the sets $\{\hat{\theta}^{(z)}\}$ comprise a sufficient statistic for the family of K-class stochastic blockmodels, and for each z, $\hat{\theta}^{(z)}$ maximizes $L(A; z, \cdot)$. Analogously, the sets $\{\bar{\theta}^{(z)}\}$ are functions of the model parameters $\{P_{ij}\}_{i<j}$, and maximize $\bar{L}_P(z, \cdot)$. We write $\hat{\theta}$ and $\bar{\theta}$ when the choice of z is understood, and $L(A; z)$ and $\bar{L}_P(z)$ to abbreviate $\sup_\theta L(A; z, \theta)$ and $\sup_\theta \bar{L}_P(z, \theta)$, respectively.

Finally, observe that when a blockmodel with parameters $(\bar{z}, \bar{\theta})$ is in force, then $P_{ij} = \bar{\theta}_{\bar{z}_i \bar{z}_j}$ in accordance with (1), and consequently $\bar{L}_P$ is maximized by the true parameter values $(\bar{z}, \bar{\theta})$:

$$\bar{L}_P(\bar{z}, \bar{\theta}) - \bar{L}_P(z, \theta) = \sum_{i<j} D(P_{ij} \,||\, \theta_{z_i z_j}) \geqslant \sum_{i<j} 2(P_{ij} - \theta_{z_i z_j})^2 \geqslant 0,$$

where $D(p \,||\, p')$ denotes the Kullback–Leibler divergence of a $\mathrm{Ber}(p')$ distribution from a $\mathrm{Ber}(p)$ one.

2·2. *Fitting a K-class stochastic blockmodel to independent Bernoulli trials*

Fitting a K-class stochastic blockmodel to independent $\mathrm{Ber}(P_{ij})$ trials yields estimates $\hat{\theta}^{(z)}$ of averages $\bar{\theta}^{(z)}$ of subsets of the parameter set $\{P_{ij}\}$, with each class assignment z inducing a partition of that set. We begin with a basic lemma that expresses the difference $L(A; z) - \bar{L}_P(z)$ in terms of $\hat{\theta}^{(z)}$ and $\bar{\theta}^{(z)}$, and follows directly from their respective maximizing properties.

LEMMA 1. *Let* $\{A_{ij}\}_{i<j}$ *comprise independent* $\mathrm{Ber}(P_{ij})$ *trials. Then the difference* $\sup_\theta L(A; z, \theta) - \sup_\theta \bar{L}_P(z, \theta)$ *can be expressed for* $X = \sum_{i<j} A_{ij} \log\{\bar{\theta}_{z_i z_j}/(1 - \bar{\theta}_{z_i z_j})\}$ *as*

$$L(A; z) - \bar{L}_P(z) = \textstyle\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab}) + X - E(X).$$

We first bound the former quantity in this expression, which provides a measure of the distance between $\hat{\theta}$ and its estimand $\bar{\theta}$ under the setting of Lemma 1. The bound is used in subsequent asymptotic results, and also yields a kind of confidence measure on $\hat{\theta}$ in the finite-sample regime.

THEOREM 1. *Suppose that a* K*-class stochastic blockmodel is fitted to data* $\{A_{ij}\}_{i<j}$ *comprising* $N!/\{(N-2)!2!\}$ *independent* $\mathrm{Ber}(P_{ij})$ *trials, where, for any class assignment* z*, estimate* $\hat{\theta}$ *maximizes the blockmodel loglikelihood* $L(A; z, \cdot)$*. Then with probability at least* $1 - \delta$*,*

$$\max_z \left\{ \sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab}) \right\} < N \log K + (K^2 + K) \log\left(\frac{N}{K} + 1\right) - \log \delta. \qquad (2)$$

Theorem 1 is proved in the Appendix via the method of types: for fixed z, the probability of any realization of $\hat{\theta}$ is first bounded by $\exp\{-\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab})\}$. A counting argument then yields a deviation result in terms of $(N/K+1)^{K^2+K}$, and finally a union bound is applied so that the result holds uniformly over all K^N possible choices of assignment vector z.

Our second result is asymptotic, and combines Theorem 1 with a Bernstein inequality for bounded random variables, applied to the latter terms $X - E(X)$ in Lemma 1. To ensure boundedness we assume minimal restrictions on each P_{ij}; this Bernstein inequality, coupled with a union bound to ensure that the result holds uniformly over all z, dictates growth restrictions on K and M.

THEOREM 2. *Assume the setting of Theorem 1, whereby a K-class blockmodel is fitted to $N!/\{(N-2)!2!\}$ independent* $\mathrm{Ber}(P_{ij})$ *random variates $\{A_{ij}\}_{i<j}$, and further assume that $1/N^2 \leqslant P_{ij} \leqslant 1 - 1/N^2$ for all N and $i < j$. Then if $K = \mathcal{O}(N^{1/2})$ and $M = \omega(N(\log N)^{3+\delta})$ for some $\delta > 0$,*

$$\max_z |L(A; z) - \bar{L}_P(z)| = o_P(M). \tag{3}$$

Thus whenever each P_{ij} is bounded away from 0 and 1 in the manner above, the maximized loglikelihood function $L(A; z) = \sup_\theta L(A; z, \theta)$ is asymptotically well behaved in network size N as long as the network's average degree $2M/N$ grows faster than $(\log N)^{3+\delta}$ and the number K of classes fitted to it grows no faster than $N^{1/2}$.

2·3. *Fitting a correctly specified K-class stochastic blockmodel*

The above results apply to the general case of independent Bernoulli data $\{A_{ij}\}$, with no additional structure assumed amongst the set of success probabilities $\{P_{ij}\}$; if we further assume the data to be generated by a K-class stochastic blockmodel whose parameters $(\bar{z}, \bar{\theta})$ are subject to suitable identifiability conditions, it is possible to characterize the behaviour of the class assignment estimator $\hat{z}$ under maximum likelihood fitting of a correctly specified K-class blockmodel.

THEOREM 3. *If (3) holds, and data are generated according to a K-class blockmodel with membership vector $\bar{z}$, then*

$$\bar{L}_P(\bar{z}) - \bar{L}_P(\hat{z}) = o_P(M), \tag{4}$$

with respect to the maximum-likelihood K-class blockmodel class assignment estimator $\hat{z}$.

Let $N_e(\hat{z})$ be the number of incorrect class assignments under $\hat{z}$, counted for every node whose true class under $\bar{z}$ is not in the majority within its estimated class under $\hat{z}$. If furthermore the following identifiability conditions hold with respect to the model sequence:

(i) *for all blockmodel classes $a = 1, \ldots, K$, class size N_a grows as $\min_a(N_a) = \Omega(N/K)$;*
(ii) *the following holds over all distinct class pairs (a, b) and all classes c:*

$$\min_{(a,b)} \max_c \left\{ D\left(\bar{\theta}_{ac} \,||\, \frac{\bar{\theta}_{ac} + \bar{\theta}_{bc}}{2}\right) + D\left(\bar{\theta}_{bc} \,||\, \frac{\bar{\theta}_{ac} + \bar{\theta}_{bc}}{2}\right) \right\} = \Omega\left(\frac{MK}{N^2}\right),$$

then it follows from (4) that $N_e(\hat{z}) = o_P(N)$.

Thus the conclusion of Theorem 3 is that under suitable conditions the fraction N_e/N of misclassified nodes goes to zero in N, yielding a convergence result for stochastic blockmodels with a growing number of classes. Condition (i) stipulates that all class sizes grow at a rate that is eventually bounded below by a single constant times N/K, while condition (ii) ensures that any two rows of θ differ in at least one entry by an amount that is eventually bounded by a single constant

time MK/N^2. Observe that if eventually $K = N^{1/2}$ and $M = N(\log N)^4$ so that conditions on K and M sufficient for Theorem 2 are met, then since $(\log N)^4 = o(N^{1/2})$, it follows that MK/N^2 goes to zero in N.

3. Numerical results

We now present results of a small simulation study undertaken to investigate the assumptions and conditions of Theorems 1–3, in which K-class blockmodels were fitted to various networks generated at random from models corresponding to each of the three theorems. Because exact maximization in z of the blockmodel loglikelihood $L(A; z, \theta)$ is computationally intractable even for moderate N, we instead employed Gibbs sampling to explore the function $\max_\theta L(A; z, \theta)$ and recorded the best value of z visited by the sampler. As the results of Theorems 1 and 2 hold uniformly in z, however, we expect $\bar{\theta}$ and $\bar{L}_P(z)$ to be close to their empirical estimates whenever N is sufficiently large, regardless of the approach employed to select z. This fact also suggests that a single-class blockmodel may come closest to achieving equality in Theorems 1 and 2, as many class assignments are equally likely a priori to have high likelihood. By similar reasoning, a weakly identifiable model should come closest to achieving the error bound in Theorem 3, such as one with nearly identical within- and between-class edge probabilities. We describe each of these cases empirically in the remainder of this section.

First, the tightness of the confidence bound of (2) from Theorem 1 was investigated by fitting K-class blockmodels to Erdös–Rényi networks comprising $N!/\{(N-2)!2!\}$ independent $\mathrm{Ber}(p)$ trials, with $N = 500$ nodes, $p = 0{\cdot}075$ chosen to match the data analysis example in the sequel, and $K \in \{5, 10, 20, 30, 40, 50\}$. For each K, the error terms $\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab})$ and $\{\sum_{a \leqslant b} n_{ab}(\hat{\theta}_{ab} - \bar{\theta}_{ab})^2\}^{1/2}$ were recorded for each of 100 trials and compared with the respective 95% confidence bounds, $\delta = 0{\cdot}05$, derived from Theorem 1. The bounds overestimated the respective errors by a factor of 3–7 on average, with small standard deviation. In this worst-case scenario, the bound is loose, but not unusable; the errors never exceeded the 95% confidence bounds in any of the trials.

To test whether the assumptions of Theorem 2 are necessary as well as sufficient to obtain convergence of $L(A; z)/M$ to $\bar{L}_P(z)/M$, blockmodels were next fitted to Erdös–Rényi networks of increasing size, for N in the range 50–1050. The corresponding normalized loglikelihood error $|L(A; z) - \bar{L}_P(z)|/M$ for different rates of growth in the expected number of edges M and the number of fitted classes K is shown in Fig. 1. Observe from Fig. 1(a) that when $M = N(\log N)^4$ and $K = N^{1/2}$, as prescribed by the theorem, this error decreases in N. If the edge density is reduced to $M/N = (\log N)^2$, we observe in Fig. 1(b) convergence when $K = N^{1/2}$ and divergence when $K = N^{3/5}$. This suggests that the error as a function of K follows Theorem 2 closely, but that the network can be somewhat more sparse than it requires.

To test the conditions of Theorem 3, blockmodels with parameters $(\bar{z}, \bar{\theta})$ and increasing class size K were used to generate data, and corresponding node misclassification error rates $N_e(z)/N$ were recorded as a function of correctly specified K-class blockmodel fitting. Model parameter $\bar{z}$ was chosen to yield equally sized blocks, so as to meet identifiability condition (i) of Theorem 3. Parameter $\bar{\theta} = \alpha I + \beta 1 1^{\mathrm{T}}$ was chosen to yield within- and between-class success probabilities with the property that for any class pair (a, b), the condition $D(\theta_{aa} \,||\, (\theta_{aa} + \theta_{ab})/2) = MK^\gamma/(20N^2)$ was satisfied, with $\gamma \in \{4/5, 9/10, 1\}$; identifiability condition (ii) was thus met only in the $\gamma = 1$ case. Figure 1(c) shows the fraction $N_e(z)/N$ of misclassified nodes when $M = N(\log N)^2$ and $K = N^{1/2}$, corresponding to the setting in which convergence of $L(A; z)/M$ to $\bar{L}_P(z)/M$ was observed above; this fraction is seen to decay when $\gamma = 1$ or $9/10$, but to increase when $\gamma = 4/5$. This behaviour conforms with Theorem 3 and suggests that its identifiability conditions are close to being necessary as well as sufficient.

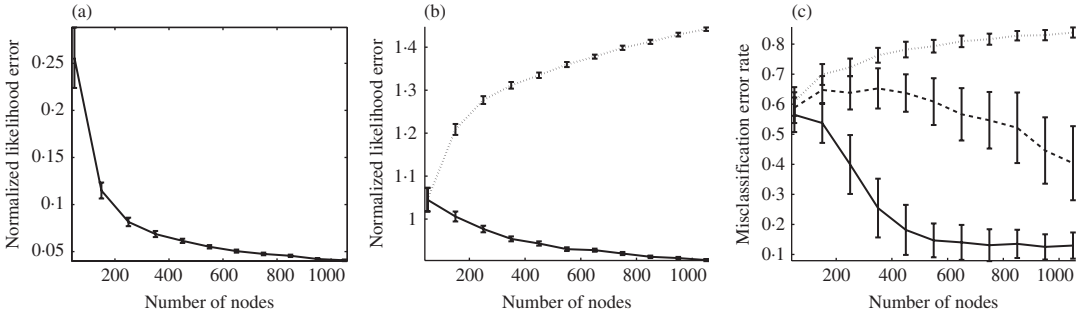

Fig. 1. Simulation study results illustrating Theorems 1–3. (a) Likelihood error $|L(A; z) - \bar{L}_P(z)|/M$ as a function of network size N, shown for $M = N(\log N)^4$ with $K = N^{1/2}$. (b) Same quantity for $M = N(\log N)^2$ with $K = N^{3/5}$ (dotted) and $K = N^{1/2}$ (solid). (c) Error rate $N_e(\hat{z})/N$ for $M = N(\log N)^2$ with $K = N^{1/2}$ and $\gamma = 4/5$ (dotted), $\gamma = 9/10$ (dashed), $\gamma = 1$ (solid).

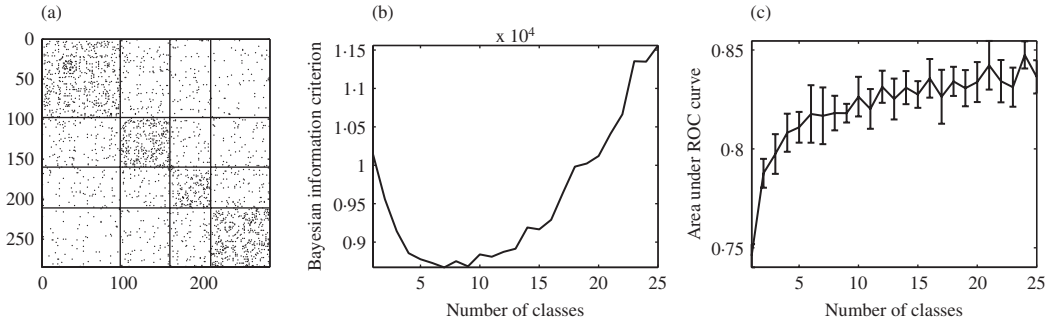

Fig. 2. Social network dataset and its fitting statistics for a varying number of blockmodel classes K. (a) Adjacency data matrix with students ordered by school year. (b) Model order statistic for fitted logit blockmodels as a function of K. (c) Out-of-sample prediction error using five-fold crossvalidation, as a function of K. Error bars indicate standard deviation.

4. NETWORK DATA EXAMPLE

4·1. *Adolescent health social network dataset*

To illustrate the use of our results in the fitting of K-class stochastic blockmodels to network data, we employed the Comm18 friendship network from the National Longitudinal Survey of Adolescent Health, in which $N = 284$ students at a school in the United States were asked to list up to five friends of each gender, yielding a network with 1189 edges signifying that one or both of the students had listed the other as a friend. The students also supplied additional information including their gender, school year and race. Further details of the study can be found in, e.g., Goodreau et al. (2009).

Of the three covariates, shared school year is reported by Goodreau et al. (2009) to be the best predictor of community structure. This finding is borne out in Fig. 2(a), which shows the adjacency structure under an ordering of students by school year and reveals strong community divisions between years.

4·2. *Logit blockmodel parameterization and fitting procedure*

Here we build on the observation of school year clustering by taking covariate information explicitly into account when fitting the dataset described above. Specifically, by assuming only that links are independent Bernoulli variates and then employing confidence bounds to assess fitted blocks by way of parameter $\bar{\theta}^{(z)}$, we examine these data for residual community structure beyond that well explained by the covariates themselves.

Since the results of Theorems 1 and 2 hold uniformly over all choices of blockmodel membership vector z, we may select z in any manner, including those that depend on covariates. For this example, we determined an approximate maximum likelihood estimate $\hat{z}$ under a logit blockmodel that allows the direct incorporation of covariates. The model is parameterized such that the log-odds ratio of an edge occurrence between nodes i and j is given by

$$\log \frac{P_{ij}}{1-P_{ij}} = \tilde{\theta}_{z_i z_j} + x(i,j)^{\mathrm{T}}\beta \quad (i=1,\ldots,N;\ j=i+1,\ldots,N), \tag{5}$$

where $x(i,j)$ a vector of covariates indicating shared group membership, and model parameters $(\tilde{\theta}, \beta, z)$ are estimated from the data. Three covariates were used, indicating shared gender, difference in school years, and a six-category covariate indicating the range of the observed degree of each node; see Karrer & Newman (2011) for related discussion on this point. The matrix $\tilde{\theta}$ is analogous to blockmodel parameter θ, the vector z specifies the blockmodel class assignment and the vector β was implemented here with sum-to-zero identifiability constraints.

Because exact maximization of the loglikelihood function $L(A; \tilde{\theta}, \beta, z)$ corresponding to (5) is computationally intractable, we instead employed an approach that alternated between Markov chain Monte Carlo exploration of z while holding $(\tilde{\theta}, \beta)$ constant, and optimization of $\tilde{\theta}$ and β while holding z constant. We tested different initialization methods and observed that highest likelihoods were consistently produced by first fitting the class assignment vector z. This fitting procedure provides a means of estimating averages $\bar{\theta}^{(z)}$ over subsets of the set $\{P_{ij}\}_{i<j}$, under the assumption that the network data comprise independent $\mathrm{Ber}(P_{ij})$ trials.

4·3. *Data analysis*

We fitted the logit blockmodel of (5) for values of K ranging from 1 to 25 using the stochastic maximization procedure described in the preceding paragraph, and gauged model order by the Bayesian information criterion and out-of-sample prediction shown, respectively, in Figs. 2(b) and (c). The minimum of the Bayesian information criterion corresponds in location with the knee of the out-of-sample prediction curve, suggesting a model order between 4 and 7. The corresponding 95% confidence bounds on the divergence of $\hat{\theta}^{(z)}$ from $\bar{\theta}^{(z)}$ provided by Theorem 1 also yield small values for K in this range also: for example, when $K=4$, the normalized sum of Kullback–Leibler divergences $N!/\{N-2)!2!\}\sum_{a\leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab})$ is bounded by 0·0120.

The top two rows of Fig. 3 depict approximate maximum likelihood estimates of z for K in the range 4–7. Larger values of K also reveal block structure, but exhibit correspondingly larger confidence bound evaluations; for example, when $K=10$, the Kullback–Leibler divergence bound of 0·026 no longer excludes an Erdos–Renyi random graph whose density matches the observed network. Adjacency structures permuted to show block divisions under $\hat{z}$ within each school year are shown in the top row, with the corresponding values of $\hat{\mu}$ shown in the bottom row. We note that the total number of visible communities shown in the top row appears to exceed K, due to the interaction of school year and latent class effects.

As K is increased, the groups do not become isolated but rather continue to exhibit cross-group friendships, suggesting fewer than four tightly demarcated communities per school year. Within each school year, the K groups can be separated into two meta-groups whose membership remained roughly constant, with 234 students whose meta-group membership did not change at all as K ranged from 4 to 7. The two meta-groups have similar school year and nodal degree distributions, with a two-sample Kolmogorov–Smirnov test returning p-values of 0·63 and 0·08 for school year and degree, respectively. The bottom row of Fig. 3 shows differing racial

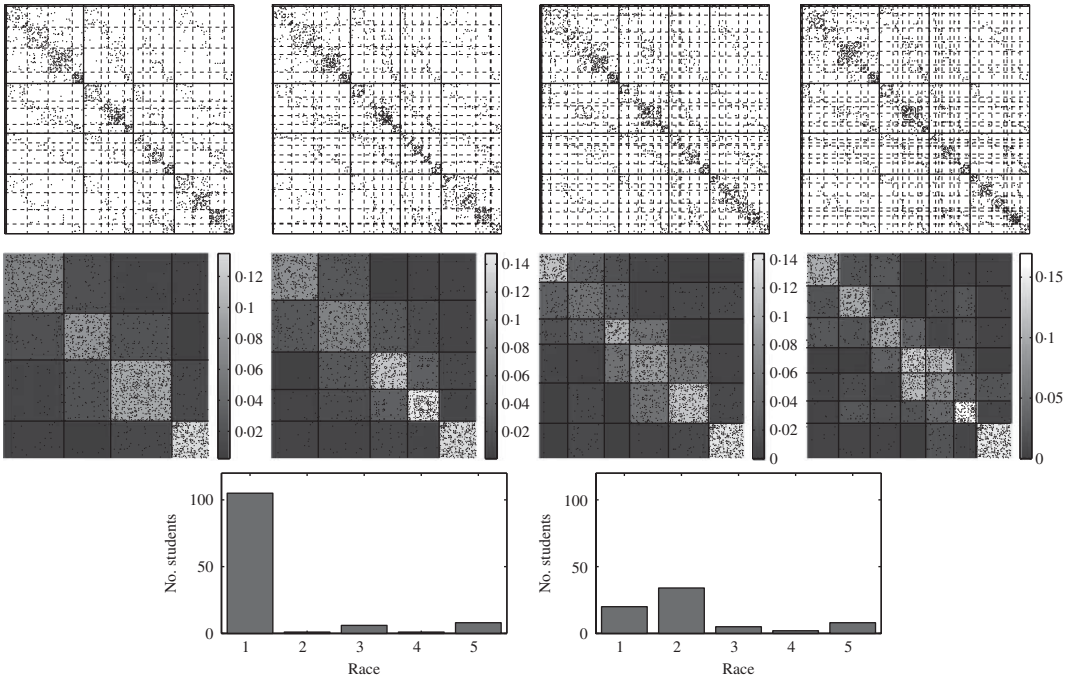

Fig. 3. Results of logit blockmodel fitting to the data of Fig. 2 for each of $K \in \{4, 5, 6, 7\}$ classes. Top row: Adjacency structure of the data, permuted to show class year and block assignments for $K \in \{4, 5, 6, 7\}$. Second row: Corresponding estimates $\hat{\theta}$, with Kullback–Leibler divergence bounds 0·0057, 0·0067, 0·0077, and 0·0086. Bottom row: Racial identity of students whose grouping remained constant over these four values of K.

compositions for the meta-groups, with race 2 concentrated almost exclusively in meta-group 2. However, membership was not determined solely by race; we note that race 1 students in the second meta-group had a higher density of friendships with race 2 than did the race 1 students in the first meta-group by a factor of ten, justifying their inclusion in the second meta-group.

ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation, National Institutes of Health, Army Research Office and the Office of Naval Research, U.S.A. Additional funding was provided by the Harvard Medical School's Milton Fund.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the dataset from § 4·1.

APPENDIX

Proof of Theorem 1. To begin, observe that for any fixed class assignment z, every $\hat{\theta}_{ab}$ is a sum of n_{ab} independent Bernoulli random variables, with corresponding mean $\bar{\theta}_{ab}$. A Chernoff bound (Dubhashi & Panconesi, 2009) shows

$$\text{pr}(\hat{\theta}_{ab} \geqslant \bar{\theta}_{ab} + t) \leqslant e^{-n_{ab} D(\bar{\theta}_{ab}+t \| \bar{\theta}_{ab})}, \quad 0 < t \leqslant 1 - \bar{\theta}_{ab},$$

$$\text{pr}(\hat{\theta}_{ab} \leqslant \bar{\theta}_{ab} - t) \leqslant e^{-n_{ab} D(\bar{\theta}_{ab}-t \| \bar{\theta}_{ab})}, \quad 0 < t \leqslant \bar{\theta}_{ab}.$$

Since these bounds also hold, respectively, for $\mathrm{pr}(\hat{\theta}_{ab} = \bar{\theta}_{ab} \pm t)$, we may bound the probability of any given realization $\vartheta \in \{0, 1/n_{ab}, \ldots, 1\}$ of $\hat{\theta}_{ab}$ in terms of the Kullback–Leibler divergence of $\bar{\theta}_{ab}$ from ϑ:

$$\mathrm{pr}(\hat{\theta}_{ab} = \vartheta) \leqslant e^{-n_{ab} D(\vartheta || \bar{\theta}_{ab})}.$$

By independence of the $\{A_{ij}\}_{i<j}$, this implies a corresponding bound on the probability of any $\hat{\theta}$:

$$\mathrm{pr}(\hat{\theta}) \leqslant \exp\left\{ -\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab}) \right\}. \tag{A1}$$

Now, let $\hat{\Theta}$ denote the range of $\hat{\theta}$ for fixed z, and observe that since each of the $(K+1)!/\{(K-1)!2!\}$ lower-diagonal entries $\{\hat{\theta}_{ab}\}_{a \leqslant b}$ of $\hat{\theta}$ can independently take on $n_{ab}+1$ distinct values, we have that $|\hat{\Theta}| = \prod_{a \leqslant b}(n_{ab}+1)$. Subject to the constraint that $\sum_{a \leqslant b} n_{ab} = N!/\{(N-2)!2!\}$, we see that this quantity is maximized when $n_{ab} = N!(K-1)!/\{(N-2)!(K-1)!\}$ for all $a \leqslant b$, and hence

$$|\hat{\Theta}| \leqslant \left\{ \binom{N}{2} \Big/ \binom{K+1}{2} + 1 \right\}^{\binom{K+1}{2}} < (N^2/K^2 + 1)^{(K^2+K)/2} < (N/K+1)^{K^2+K}. \tag{A2}$$

Now consider the event that $\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab})$ is at least as large as some $\epsilon > 0$; the probability of this event is given by $\mathrm{pr}(\hat{\Theta}_\epsilon)$ for

$$\hat{\Theta}_\epsilon = \left\{ \hat{\theta} \in \hat{\Theta} : \sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab}) \geqslant \epsilon \right\}. \tag{A3}$$

Since $\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab}) \geqslant \epsilon$ for all $\hat{\theta} \in \hat{\Theta}_\epsilon$, we have from (A1) and (A3) that

$$\mathrm{pr}(\hat{\Theta}_\epsilon) = \sum_{\hat{\theta} \in \hat{\Theta}_\epsilon} \mathrm{pr}(\hat{\theta}) \leqslant \sum_{\hat{\theta} \in \hat{\Theta}_\epsilon} e^{-\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} || \bar{\theta}_{ab})} \leqslant \sum_{\hat{\theta} \in \hat{\Theta}_\epsilon} e^{-\epsilon} = |\hat{\Theta}_\epsilon| e^{-\epsilon},$$

and since $|\hat{\Theta}_\epsilon| \leqslant |\hat{\Theta}|$, we may use (A2) to obtain, for fixed class assignment z,

$$\mathrm{pr}\left\{ \sum_{a \leqslant b} n_{ab} D(\hat{\theta} \,||\, \bar{\theta}) \geqslant \epsilon \right\} < (N/K+1)^{K^2+K} e^{-\epsilon}. \tag{A4}$$

Appealing to a union bound over all K^N possible class assignments and setting $\epsilon = \log\{K^N (N/K+1)^{K^2+K}/\delta\}$ then yields the claimed result. □

Proof of Theorem 2. By Lemma 1, the difference $L(A; z) - \bar{L}_P(z)$ can be expressed for any fixed class assignment z as $\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab}) + X - E(X)$, where the first term satisfies the deviation bound of (A4), and $X = \sum_{i<j} A_{ij} \log\{\bar{\theta}_{z_i z_j}/(1-\bar{\theta}_{z_i z_j})\}$ comprises a weighted sum of independent $\mathrm{Ber}(P_{ij})$ random variables.

To bound the quantity $|X - E(X)|$, observe that since by assumption $N^{-2} \leqslant P_{ij} \leqslant 1 - N^{-2}$, the same is true for each corresponding average $\bar{\theta}_{z_i z_j}$. As a result, the random variables $X_{ij} = A_{ij} \log\{\bar{\theta}_{z_i z_j}/(1-\bar{\theta}_{z_i z_j})\}$ comprising X are each bounded in magnitude by $C = 2 \log N$. This allows us to apply a Bernstein inequality for sums of bounded independent random variables due to Chung & Lu (2006, Theorems 2.8 and 2.9, p. 27), which states that for any $\epsilon > 0$,

$$\mathrm{pr}\{|X - E(X)| \geqslant \epsilon\} \leqslant 2 \exp\left\{ -\frac{\epsilon^2}{2\sum_{i<j} E(X_{ij}^2) + (2/3)\epsilon C} \right\}. \tag{A5}$$

Finally, observe that since the event $|L(A;z) - \bar{L}_P(z)| > 2\epsilon M$ implies either the event $\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab}) \geqslant \epsilon M$ or the event $|X - E(X)| \geqslant \epsilon M$, we have for fixed assignment z that

$$\mathrm{pr}\{|L(A;z) - \bar{L}_P(z) \geqslant 2\epsilon M\} \leqslant \mathrm{pr}\left[\left\{\sum_{a \leqslant b} n_{ab} D(\hat{\theta}_{ab} \,||\, \bar{\theta}_{ab}) \geqslant \epsilon M\right\} \cup \left\{|X - E(X)| \geqslant \epsilon M\right\}\right].$$

Summing the right-hand sides of (A4) and (A5), and then over all K^N possible assignments, yields

$$\mathrm{pr}\left\{\max_z |L(A;z) - \bar{L}_P(z)| \geqslant 2\epsilon M\right\} \leqslant \exp\{K \log N + (K^2 + K)\log(N/K + 1) - \epsilon M\} + 2\exp\left\{K \log N - \frac{\epsilon^2 M}{8\log^2 N + (4/3)\epsilon \log N}\right\},$$

where we have used the fact that $\sum_{i<j} E(X_{ij}^2) \leqslant 4M \log^2(N)$ in (A5). It follows directly that if $K = \mathcal{O}(N^{1/2})$ and $M = \omega(N(\log N)^{3+\delta})$, then $\lim_{N\to\infty} \mathrm{pr}\{\max_z |L(A;z) - \bar{L}_P(z)|/M \geqslant \epsilon\} = 0$ for every fixed $\epsilon > 0$ as claimed. □

Proof of Theorem 3. To begin, note that Theorem 2 holds uniformly in z, and thus implies that

$$|\bar{L}_P(\bar{z}) - L(A;\bar{z})| + |\bar{L}_P(\hat{z}) - L(A;\hat{z})| = o_P(M).$$

Since $\hat{z}$ is the maximum-likelihood estimate of class assignment $\bar{z}$, we know that $L(A;\hat{z}) \geqslant L(A;\bar{z})$, implying that $L(A;\hat{z}) = L(A;\bar{z}) + \delta$ for some $\delta \geqslant 0$. Thus, by the triangle inequality,

$$|\bar{L}_P(\bar{z}) - \bar{L}_P(\hat{z}) + \delta| \leqslant |\bar{L}_P(\bar{z}) - L(A;\bar{z})| + |\bar{L}_P(\hat{z}) - (L(A;\bar{z}) + \delta)| = o_P(M),$$

and since $\bar{L}_P(\bar{z}) \geqslant \bar{L}_P(\hat{z})$ under any blockmodel with parameter $\bar{z}$, we have $\bar{L}_P(\bar{z}) - \bar{L}_P(\hat{z}) = o_P(M)$.

Under conditions (i) and (ii) of Theorem 3, we will now show that also

$$\bar{L}_P(\bar{z}) - \bar{L}_P(\hat{z}) = \frac{N_e(\hat{z})}{N}\Omega(M), \tag{A6}$$

holds for every realization of $\hat{z}$, thus implying that $N_e(\hat{z}) = o_P(N)$ and proving the theorem.

To show (A6), first observe that any blockmodel class assignment vector z induces a corresponding partition of the set $\{P_{ij}\}_{i<j}$ according to $(i,j) \mapsto (z_i, z_j)$. Formally, z partitions $\{P_{ij}\}_{i<j}$ into L subsets $(S_1, \ldots, S_L)$ via the mapping

$$\zeta_{ij} : (i = 1, \ldots, N;\ j = i+1, \ldots, N) \to (l = 1, \ldots, L).$$

This partition is separable in the sense that there exists a bijection between $\{1, \ldots, L\}$ and the upper triangular portion of blockmodel parameter θ, such that we write $\theta_{\zeta_{ij}} = \theta_{z_i z_j}$ for membership vector z. More generally, for any partition Π of $\{P_{ij}\}_{i<j}$, we may define $\bar{\theta}_l = |S_l|^{-1} \sum_{i<j} P_{ij}\, 1\{P_{ij} \in S_l\}$ as the arithmetic average over all P_{ij} in the subset S_l indexed by $\zeta_{ij} = l$. Thus we may also define

$$\bar{L}_P^*(\Pi) = \sum_{i<j} \{P_{ij} \log \bar{\theta}_{\zeta_{ij}} + (1 - P_{ij}) \log(1 - \bar{\theta}_{\zeta_{ij}})\},$$

so that $\bar{L}_P^*$ and $\bar{L}_P$ coincide on partitions corresponding to admissible blockmodel assignments z.

The establishment of (A6) proceeds in three steps: first, we construct and analyse a refinement of the partition Π^z induced by any blockmodel assignment vector z in terms of its error $N_e(z)$; then, we show that refinements increase $\bar{L}_P^*(\cdot)$; finally, we apply these results to the maximum-likelihood estimate $\hat{z}$.

LEMMA A1. *Consider a K-class stochastic blockmodel with membership vector $\bar{z}$, and let Π^z denote the partition of its associated $\{P_{ij}\}_{1\leqslant i<j\leqslant N}$ induced by any $z \in \{1, \ldots, K\}^N$. For every Π^z, there exists a*

partition Π^* *that refines* Π^z *and with the property that, if conditions* (i) *and* (ii) *of Theorem 3 hold,*

$$\bar{L}_P(\bar{z}) - \bar{L}_P^*(\Pi^*) = \frac{N_e(\hat{z})}{N}\Omega(M), \tag{A7}$$

where $N_e(z)$ *counts the number of nodes whose true class assignments under* $\bar{z}$ *are not in the majority within their respective class assignments under* z.

LEMMA A2. *Let* Π' *be a refinement of any partition* Π *of the set* $\{P_{ij}\}_{i<j}$*; then* $\bar{L}_P^*(\Pi') \geqslant \bar{L}_P^*(\Pi)$.

Since Lemma A1 applies to any admissible blockmodel assignment vector z, it also applies to the maximum-likelihood estimate $\hat{z}$ for any realization of the data; each $\hat{z}$ in turn induces a partition $\Pi^{\wedge}$ of blockmodel edge probabilities $\{P_{ij}\}_{i<j}$, and (A7) holds with respect to its refinement Π^*. By Lemma A2, $\bar{L}_P^*(\Pi^{\wedge}) \leqslant \bar{L}_P^*(\Pi^*)$. Finally, observe that $\bar{L}_P(\hat{z}) = \bar{L}_P^*(\Pi^{\wedge})$ by the definition of $\bar{L}_P^*$, and so $\bar{L}_P(\bar{z}) - \bar{L}_P(\hat{z}) \geqslant \bar{L}_P(\bar{z}) - \bar{L}_P^*(\Pi^*)$, thereby establishing (A6). □

Proof of Lemma A1. The construction of Π^* will take several steps. For a given membership class under z, partition the corresponding set of nodes into subclasses according to the true class assignment $\bar{z}$ of each node. Then remove one node from each of the two largest subclasses so obtained, and group them together as a pair; continue this pairing process until no more than one nonempty subclass remains, then terminate. Observe that if we denote pairs by their node indices as (i, j), then by construction $z_i = z_j$ but $\bar{z}_i \neq \bar{z}_j$.

Repeat the above procedure for each class under z, and let C_1 denote the total number of pairs thus formed. For each of the C_1 pairs (i, j), find all other distinct indices k for which the following holds:

$$D\left(P_{ik} \,||\, \frac{P_{ik}+P_{jk}}{2}\right) + D\left(P_{jk} \,||\, \frac{P_{ik}+P_{jk}}{2}\right) \geqslant C\frac{MK}{N^2}, \tag{A8}$$

where C is the constant from condition (ii) of Theorem 3, and indices ik and jk in (A8) are to be interpreted, respectively, as ki whenever $k < i$, and kj whenever $k < j$. Let C_2 denote the total number of distinct triples that can be formed in this manner.

We are now ready to construct the partition Π^* of the probabilities $\{P_{ij}\}_{1\leqslant i<j\leqslant N}$ as follows: For each of the C_2 triples (i, j, k), remove P_{ik} (or P_{ki} if $k < i$) and P_{jk} (or P_{kj}) from their previous subset assignment under Π^z, and place them both in a new, distinct two-element subset. We observe the following:

(i) the partition Π^* is a refinement of the partition Π^z induced by z: Since nodes i and j have the same class label under z in that $z_i = z_j$, it follows that for any k, P_{ik} and P_{jk} are in the same subset under Π^z;

(ii) since for each class at most one nonempty subclass remains after the pairing process, the number of pairs is at least half the number of misclassifications in that class. Therefore, we conclude $C_1 \geqslant N_e(z)/2$;

(iii) condition (ii) of Theorem 3 implies that for every pair of classes (a, b), there exists at least one class c for which (A8) holds eventually. Thus eventually, for any of the C_1 pairs (i, j), we obtain a number of triples at least as large as the cardinality of class c. Condition (i) in turn implies that the cardinality of the smallest class grows as $\Omega(N/K)$, and thus we may write $C_2 = C_1\Omega(N/K)$.

We can now express the difference $\bar{L}_P(\bar{z}) - \bar{L}_P^*(\Pi^*)$ as a sum of nonnegative divergences $D(P_{ij} \,||\, \bar{\theta}_{\zeta_{ij}^*})$, where ζ_{ij}^* is the assignment mapping associated to Π^*, and use (A8) to bound this difference below:

$$\bar{L}_P(\bar{z}) - \bar{L}_P^*(\Pi^*) = \sum_{i<j} D(P_{ij} \,||\, \bar{\theta}_{\zeta_{ij}^*}) = C_2\Omega\left(\frac{MK}{N^2}\right) = \frac{N_e(z)}{2}\Omega\left(\frac{M}{N}\right).$$

□

Proof of Lemma A2. Let Π' be a refinement of any partition Π of the set $\{P_{ij}\}_{i<j}$, and given $a \in \{1, \ldots, L'\}$ indexing S'_a, let $F(a)$ denote its index under Π. We show that $\bar{L}^*_P(\Pi') \geqslant \bar{L}^*_P(\Pi)$ as follows:

$$\begin{aligned}
\bar{L}^*_P(\Pi') &= \sum_{a=1}^{L'} |S'_a| \{\bar{\theta}'_a \log \bar{\theta}'_a + (1 - \bar{\theta}'_a) \log(1 - \bar{\theta}'_a)\} \\
&\geqslant \sum_{a=1}^{L'} |S'_a| \{\bar{\theta}'_a \log \bar{\theta}_{F(a)} + (1 - \bar{\theta}'_a) \log(1 - \bar{\theta}_{F(a)})\} \\
&= \sum_{b=1}^{L} |S_b| \{\bar{\theta}_b \log \bar{\theta}_b + (1 - \bar{\theta}_b) \log(1 - \bar{\theta}_b)\} = \bar{L}^*_P(\Pi),
\end{aligned}$$

where the first inequality holds by nonnegativity of Kullback–Leibler divergence, and the second equality follows from the fact that Π' is a refinement of Π. □

REFERENCES

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. & XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9**, 1981–2014.

BICKEL, P. J. & CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Nat. Acad. Sci. U.S.A.* **106**, 21068–73.

CHUNG, F. R. K. & LU, L. (2006). *Complex Graphs and Networks*. Providence, Rhode Island: American Mathematical Society.

COPIC, J., JACKSON, M. O. & KIRMAN, A. (2009). Identifying community structures from network data via maximum likelihood methods. *Berk. Electron. J. Theor. Economet.* **9**. RePEc:bpj:bejtec:v:9:y:2009:i:1:n:30.

DUBHASHI, D. P. & PANCONESI, A. (2009). *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge, U.K.: Cambridge University Press.

FIENBERG, S. E., MEYER, M. M. & WASSERMAN, S. S. (1985). Statistical analysis of multiple sociometric relations. *J. Am. Statist. Assoc.* **80**, 51–67.

GIRVAN, M. & NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Nat. Acad. Sci. U.S.A.* **99**, 7821–6.

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. & AIROLDI, E. M. (2009). A survey of statistical network models. *Foundat. Trend Mach. Learn.* **2**, 129–233.

GOODREAU, S., KITTS, J. & MORRIS, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* **46**, 103–25.

HANDCOCK, M. S., RAFTERY, A. E. & TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. R. Statist. Soc. A* **170**, 301–54.

HOFF, P. D. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, Ed. J. C. Platt, D. Koller, Y. Singer & S. Roweis, pp. 657–64, vol. 20. Cambridge, Massachusetts: MIT Press.

HOLLAND, P., LASKEY, K. B. & LEINHARDT, S. (1983). Stochastic blockmodels: some first steps. *Social Networks* **5**, 109–37.

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

KARRER, B. & NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107–1–10.

LORRAIN, F. & WHITE, H. C. (1971). Structural equivalence of individuals in social networks. *J. Math. Sociol.* **1**, 49–80.

MARIADASSOU, M., ROBIN, S. & VACHER, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Statist.* **4**, 715–42.

NOWICKI, K. & SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Am. Statist. Assoc.* **96**, 1077–87.

ROHE, K., CHATTERJEE, S. & YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39**, 1878–915.

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

[*Received November* 2010. *Revised July* 2011]