Journal of Classification 14:75-100 (1997)

Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure

Tom A. B. Snijders

University of Groningen

Krzysztof Nowicki

University of Lund

Abstract: A statistical approach to *a posteriori* blockmodeling for graphs is proposed. The model assumes that the vertices of the graph are partitioned into two unknown blocks and that the probability of an edge between two vertices depends only on the blocks to which they belong. Statistical procedures are derived for estimating the probabilities of edges and for predicting the block structure from observations of the edge pattern only. ML estimators can be computed using the EM algorithm, but this strategy is practical only for small graphs. A Bayesian estimator, based on Gibbs sampling, is proposed. This estimator is practical also for large graphs. When ML estimators are used, the block structure can be predicted based on predictive likelihood. When Gibbs sampling is used, the block structure can be predicted from posterior predictive probabilities.

A side result is that when the number of vertices tends to infinity while the probabilities remain constant, the block structure can be recovered correctly with probability tending to 1.

Keywords: Colored graph; EM algorithm; Gibbs sampling; Latent class model; Social network.

Parts of this research were carried out while the second author was a visiting scholar at the Department of Sociology of the University of Utrecht, and later when he was on leave at the University of California at Berkeley under support from the Swedish Council of the Humanities and the Social Sciences.

Authors' addresses: Tom Snijders, Department of Statistics and Measurement Theory, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands, email T.A.B.SNIJDERS@PPSW.RUG.NL. Krzysztof Nowicki, Department of Statistics, University of Lund, Box 7008, S-220 07 Lund, Sweden, email KRZYSZTOF.NOWICKI@STAT.LU.SE.

1. Introduction

Graphs and directed graphs (digraphs) are used as mathematical models for social, physical, and other phenomena where relations between units are observed. This paper applies especially to social science applications of graphs, where stochastic models and statistical inference play an essential role.

Mathematical random graph theory, which started with the classical papers of Erdös and Rényi (1959, 1960), has been mainly devoted to studies of various probabilistic properties of Bernoulli graphs and related models (uniform graph models). The monograph by Bollobás (1985) provides an excellent overview. Since these random graph models exhibit too little structure for satisfactory application to many empirical data sets, many other models have been proposed in the literature. Two important types of models for graphs and digraphs are blockmodels and stochastic models. An integration of these approaches to graph modeling was proposed by Fienberg and Wasserman (1981) and Holland, Laskey, and Leinhardt (1983).

The purpose of blockmodeling is to partition the vertex set into subsets called *blocks* in such a way that the block structure and the pattern of edges between the blocks capture the main structural features of the graph. In mathematical terminology, the block structure can be represented by the colors of a colored graph; blocks and colors can be used as equivalent concepts. Lorrain and White (1971) proposed blockmodeling based on the concept of *structural equivalence*, which states that two vertices are structurally equivalent (belong to the same block) if they relate to the other vertices in the same way. The blocks can be regarded as equivalence classes of vertices. In practice, social network researchers often use a more loose concept of approximate structural equivalence, where some deviations (exceptional lines or absences of lines, destroying the property of structural equivalence) are permitted. If the vertices of a given graph are renumbered so that blocks according to (approximate) structural equivalence correspond to sets of consecutive vertices, the adjacency matrix shows a block pattern: some blocks have (predominantly) 1 entries, other blocks have (predominantly) 0 entries.

The popularity of deterministic blockmodeling results in part from the fact that since the mid 1970's two computer packages, CONCOR (Breiger, Boorman, and Arabie 1975; Arabie, Boorman, and Levitt 1978; Schwartz 1977) and STRUCTURE (Burt 1976) have been available, both allowing us to find a permutation of the rows and columns in the adjacency matrix leading to an approximate block structure. Unfortunately, these programs are of an exploratory nature and lack a statistical framework. Discussion can be found in Wasserman and Anderson (1987), Faust (1988), and Scott (1991, pp. 134-142). These references also treat some of the many other equivalence

concepts in digraphs.

Fienberg and Wasserman (1981) and Holland, Laskey, and Leinhardt (1983) extended the concept of blockmodeling to a stochastic version. A stochastic blockmodel can be defined as a probability distribution (or family of distributions) for graphs (or digraphs) of which the vertex set is partitioned into subsets called blocks, which has the property that the probability distribution for the graph is invariant under permutations of vertices within blocks. Such a model can be described alternatively as a stochastic colored graph for which the probability distribution is invariant under permutations of similarly colored vertices. Under such a model, the probability that an edge is present between two vertices depends only on the colors of the vertices. Wasserman and Anderson (1987) defined vertices to be stochastically equivalent when they belong to the same block in a stochastic blockmodel.

For the statistical application of blockmodels, an important distinction is whether the blocks are known (e.g., through attributes of the vertices), or have to be inferred from the edge pattern. The latter situation is much more complicated, and sometimes called *a posteriori blockmodeling*. Wasserman and Anderson (1987) and Anderson, Wasserman, and Faust (1992) studied a posteriori blocking for blockmodels in the p_1 family. This is a log-linear exponential family of probability distributions for digraphs, introduced by Holland and Leinhardt (1981). This family includes vertex parameters modeling the distribution of out- and in-degrees as well as an overall parameter connected to the reciprocity of contacts between vertices. Holland and Leinhardt (1981), Fienberg, Meyer, and Wasserman (1985), and others studied various inferential aspects for log-linear models, both for single and for multiple sociometric relations. Wasserman and Faust (1994) give an overview of this model and techniques for estimation of the parameters. A version of the p_1 model for undirected graphs with n vertices can be defined by the probability function

$$P(\mathbf{y}; \alpha, \theta) = K(\alpha, \theta) \exp(\theta y_{++} + \sum_{i=1}^{n} \alpha_i y_{i+}), \tag{1}$$

where $\mathbf{y} = \{y_{ij}\}_{1 \le i \ne j \le n}$ is the adjacency matrix (see Section 2) and $y_{i+} = \Sigma_{j=1}^{n} y_{ij}$, and $y_{++} = \Sigma_{1 \le i < j \le n} y_{ij}$; where θ and $\alpha = (\alpha_1, \ldots, \alpha_n)$ are parameters with $\Sigma_{i=1}^{n} \alpha_i = 0$, while $K(\alpha, \theta)$ is a normalizing function.

A major problem in statistical inference for the p_1 model is that the number of parameters increases with the number of vertices in the observed data. This situation may lead to overfitting; in any case, the standard asymptotic properties of estimation methods such as maximum likelihood do not hold automatically. Combining blockmodels with the p_1 model is one strategy used to overcome this problem. The vertex parameters of the p_1 distribution are then defined in terms of the blocks to which the vertices belong rather

than in terms of the identity of the vertices. Wasserman and Anderson (1987) and Anderson, Wasserman, and Faust (1992) studied a posteriori blocking for blockmodels in the p_1 family. They blocked the vertices by first calculating ML estimates of the vertex parameters, and subsequently grouping the vertices on the basis of multiple comparisons of the estimated parameters. This method is reliable only if the p_1 model provides a satisfactory fit to the data; this condition is not always sufficiently stressed in the literature. Anderson, Wasserman, and Faust (1992; Section 4.1) give a review of several methods for obtaining blocks.

Wang and Wong (1987) proposed blockmodels that are obtained by adding block parameters to the basic p_1 model. Those authors proposed estimators and tests but only for the situation that the block structure is known a priori.

In this paper we study a posteriori blocking in a stochastic blockmodel for graphs. Each of the vertices of the observed graph belongs to one block; however, the block structure is not directly observed. Furthermore, the edges are independent, conditional on the block structure. Our model is more general than the analogue for undirected graphs of the model considered by Wasserman and Anderson (1987) and Anderson, Wasserman, and Faust (1992) because it is a blockmodel not subsumed in the p_1 model. This enhanced generality is purchased with a larger number of parameters: the special case of model (1), where vertex parameters depend on m blocks has m free parameters while the blockmodel considered below, for a given block structure with m blocks, has m^2 parameters (cf. Definition 1 below).

We give maximum likelihood and Bayesian estimators for the parameters of the model as well as procedures for recovering the block structure in the case of two blocks. It turns out that maximum likelihood estimators are not feasible for larger graphs (say, for more than 20 or 30 vertices). Bayesian estimators, implemented using the Gibbs sampler (see Section 5), can be used, however, as a practical possibility for larger graphs as well. Future research will be directed toward extending these procedures to directed graphs and to more than two blocks.

2. Some Notation for Graphs and Blocks

An *undirected graph* G consists of a pair $(V(G),E(G))$ of a set $V(G)$ of elements called *vertices*, and a subset $E(G)$ of the collection of unordered pairs from $V(G)$. The elements of $E(G)$ are called *edges*. A graph G on the finite set $V(G) = \{1, \ldots, n\}$ of vertices can be represented by its adjacency matrix $\mathbf{y} = \{y_{ij}\}_{1 \le i \ne j \le n}$, where

$$y_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } i \text{ and } j \\ 0 & \text{otherwise,} \end{cases}$$

where $y_{ii} = 0$ for all i, since loops between a vertex and itself are excluded. Clearly, $y_{ij} = y_{ji}$ for each pair $1 \le i \ne j \le n$.

Consider an undirected graph in which the vertices belong to m different categories. Let those categories, referred to as *blocks*, be represented by a vertex variable; designate outcomes of this variable as *colors*. We introduce the vector $\mathbf{x} = (x_i)_{i=1}^n$, where

$x_i = k$ if vertex i has color k,

for $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$.

Thus, the *colored graph* C can be represented by the array $(\mathbf{y},\mathbf{x})$; $\mathbf{x}$ is also called the *block structure* of the colored graph. In this paper we shall consider random graphs to be defined by probability distributions over the set of undirected graphs G with a fixed vertex set $V(G) = \{1, \ldots, n\}$ and an arbitrary edge set. Random variables will be denoted by capital letters. For a random colored graph, the number n of vertices will be assumed to be fixed, but the adjacency matrix $\mathbf{Y}$ and the color vector $\mathbf{X}$ will be random.

3. Stochastic Blockmodels for Graphs

The definitions in Holland et al. (1983) for directed graphs can be extended as follows to define *stochastic blockmodels* with independent edges for undirected graphs G. The vertex set is $\{1, \ldots, n\}$, and the random adjacency matrix is (Y_{ij}).

1. The random variables Y_{ij} for $i < j$ are statistically independent; furthermore, $Y_{ij} \equiv Y_{ji}$, and $Y_{ii} \equiv 0$.
2. There exists a partition of the vertex set $\{1,\ldots,n\}$ into blocks such that for any vertices i, j, h with $i \ne j \ne h$, if i and h belong to the same block, then Y_{ij} and Y_{hj} are identically distributed.

Condition 2 can be rephrased by stating that $P(Y_{ij} = 1)$ depends on the vertices i and j through their colors. In this paper we consider a probabilistic blockmodel for graphs with random vertex colors.

Definition 1. *A* random blockmodel *is a family of probability distributions for a colored graph C with vertex set* $\{1,\ldots,n\}$ *and color set* $\{1,\ldots,m\}$, *defined as follows.*

1. *The parameters are the vector* $\theta = (\theta_1, \ldots, \theta_m)$ *of color probabilities*

and the matrix $\eta = (\eta_{kl})_{1 \le k \le l \le m}$ *of color-dependent edge probabilities.*

2. *The vector of vertex colors consists of i.i.d. rv's* $(X_i)_{i=1}^n$, *where* $P(X_i = k) = \theta_k$ *for* $k = 1, \ldots, m$.
3. *Conditional on the vertex colors* X_i, *the edges* Y_{ij} *are independent, with* $Y_{ij} \sim$ *Bernoulli* (η_{X_i, X_j}).

If $(\mathbf{y}, \mathbf{x})$ represents such a colored graph C, the probability function is given by

$$P(\mathbf{y}, \mathbf{x}; \theta, \eta) = \theta_1^{n_1} \cdots \theta_m^{n_m} \prod_{1 \le k \le l \le m} \eta_{kl}^{e_{kl}} (1 - \eta_{kl})^{n_{kl} - e_{kl}}, \tag{2}$$

where $n_k = \sum_{i=1}^n I(x_i = k)$ denotes the number of vertices in C having color k,

$$e_{kl} = \frac{1}{1 + \delta_{kl}} \sum_{1 \le i \ne j \le n} y_{ij} I(x_i = k) I(x_j = l) \tag{3}$$

denotes the number of edges in C having one vertex of color k and the other of color l, and

$$n_{kl} = \begin{cases} n_k n_l & \text{if } k \ne l \\ \binom{n_k}{2} & \text{if } k = l, \end{cases}$$

while $\delta_{kl} = 1$ for $k = l$ and $\delta_{kl} = 0$ for $k \ne l$. Furthermore, we denote

$$s = \sum_{1 \le i < j \le n} y_{ij}, \tag{4}$$

the total number of edges.

The conditional distribution of the colored graph, given the vector of colors $(x_i)_{i=1}^n$, is a stochastic blockmodel with independent edges in which the colors $x_1, \ldots, x_n$ have the role of parameters. In the latter model the number of parameters tends to infinity with n, whereas in our model with random blocks the number of parameters is fixed at $m(m + 3)/2$, and the colors x_i occur as random variables. The situation where the number of parameters increases with n (called a situation with incidental parameters) is undesirable from a statistical point of view, because it often leads to inconsistent estimation. Therefore we prefer to work with a model where the blocks are random and the statistical parameters correspond to probabilities of colors rather than to realized colors.

In applications of blockmodels, the edges often refer to a relation which is more frequent within blocks than between blocks; e.g., friendship relations between persons where blocks correspond to groups with similar attitudes. In such cases, η_{kl} for $k < l$ will tend to be smaller than η_{kk} and η_{ll}. However, it is also possible to think of applications where the diagonal values

η_{kk} typically are smaller than off-diagonal η_{kl} $(k < l)$, e.g., when the relation is mutual sexual attraction and the blocks are defined by gender in a mainly heterosexual population. Both these orderings are incompatible with the blockmodel version of the p_1 model defined by (1). In such a model we have that $\eta_{kl} = 1 - (1 + \exp(\theta + \alpha_k + \alpha_l))^{-1}$ so that the values of off-diagonal probabilities η_{kl} necessarily are between the values of the corresponding diagonal probabilities. Wasserman and Galaskiewicz (1984) discuss a stochastic blockmodel which is an extension of the p_1 model not subject to these order restrictions, but they do not treat posterior blockmodeling.

Various stochastic properties of the random blockmodel have been studied in the literature. In addition to the references given in Section 1, we mention the following. Frank and Harary (1982), motivated by entropy calculations, discussed statistical inference for $s_2 = \Sigma_{i=1}^m \theta_i^2$ and $s_3 = \Sigma_{i=1}^m \theta_i^3$ under the assumption that $\eta_{kl} = (1 - \alpha)\,\delta_{kl} + \beta(1 - \delta_{kl})$. They proposed several moment-based estimators for s_2 and s_3 under various restrictions on $\theta_1, \ldots, \theta_m$, α, and β. Frank (1988) obtained the expectation and the variance for the number of edges and for the vector of triad counts. Frank (1988) and Wellman, Frank, Espinoza, Lundquist, and Wilson (1991) consider statistical inference for certain models for randomly colored graphs assuming that the edges as well as the colors are observed. Janson and Nowicki (1991) studied the asymptotic distributions of the vector of suitable normalized subgraph counts and obtained, depending on the topology of the subgraph, convergence to either the normal or the X^2 distribution.

4. Maximum Likelihood Estimation

We consider the colored graph model as given by (2) for $m = 2$ colors, assuming that only the edge structure $\mathbf{Y}$ can be observed, i.e., the color vector is unobserved (latent). Thus, the probability of observing edge pattern $\mathbf{y}$ is

$$P(\mathbf{y};\theta,\eta) = \sum_{\mathbf{x}\in\{1,2\}^n} P(\mathbf{y},\mathbf{x};\theta,\eta)\,, \tag{5}$$

where

$$P(\mathbf{y},\mathbf{x};\theta,\eta) = (1-\theta)^n \left[\frac{\theta}{1-\theta}\right]^{n_2} \times \tilde{\eta}_{11}^{\binom{n_1}{2}}\, \tilde{\eta}_{12}^{n_1 n_2}\, \tilde{\eta}_{22}^{\binom{n_2}{2}} \left[\frac{\eta_{11}}{\tilde{\eta}_{11}}\right]^{e_{11}} \left[\frac{\eta_{12}}{\tilde{\eta}_{12}}\right]^{e_{12}} \left[\frac{\eta_{22}}{\tilde{\eta}_{22}}\right]^{e_{22}}, \tag{6}$$

where $n_k = \Sigma_{i=1}^n I(x_i = k)$, e_{kl} is as defined in (3), $\tilde{\eta}_{kl} = 1 - \eta_{kl}$ for $k,l = 1,2$, and θ now denotes $P(X_i = 2) = 1 - P(X_i = 1)$. Note that $n_1 + n_2 = n$ and $e_{11} + e_{12} + e_{22} = s$, defined in (4).

The parameters in this statistical model are not identifiable, which means that several distinct parameter vectors can be associated with the same probability distribution for **Y**. Indeed, replacing θ by $1-\theta$ and interchanging η_{11} and η_{22} will yield the same distribution; furthermore, if $\eta_{11} = \eta_{12} = \eta_{22}$, then we are back at the Bernoulli graph model, and the value of θ is irrelevant. This lack of identifiability can be remedied formally by requiring $\eta_{11} < \eta_{22}$. We shall not institute this requirement, but will instead take the non-identifiability into account when interpreting results; moreover, algorithms and statistical evaluations will have to take into account the possibility of nonuniqueness of maximum likelihood and other estimators.

We now turn to the problem of estimation of the parameters in the latent blockmodel. Unfortunately, because of the intractable form of the likelihood function, explicit formulae for the maximum like- lihood estimators cannot be obtained. Instead we must use numerical methods for maximizing the likelihood function. Two numerical estimation methods are studied here: (a) the direct numerical maximization of the likelihood function, and (b) the EM algorithm (Dempster, Laird, and Rubin 1977).

4.1 The Direct Maximization

The expression given in (5)-(6) for $P(\mathbf{y};\theta,\eta)$ is not convenient for numerical calculations because the number of terms in the sum, 2^n, is unpleasantly large. By combining terms it is possible to obtain an expression containing a sum where the number of terms increases only polynomially with n. Toward this end, write (6) as

$$P(\mathbf{y},\mathbf{x};\theta,\eta) = a(\theta,\eta_{11})\, b^s(\eta_{12})\, \beta(e_{11},e_{22},n_2;\theta,\eta)\,, \tag{7}$$

where

$$a(\theta,\eta_{11}) = (1-\theta)^n\, \tilde{\eta}_{11}^{\binom{n}{2}}\,,$$

$$b(\eta_{12}) = \frac{\eta_{12}}{\tilde{\eta}_{12}}\,,$$

$$\beta(e_{11},e_{22},n_2;\theta,\eta) =$$

$$c^{e_{11}}(\eta_{11},\eta_{12})d^{e_{22}}(\eta_{12},\eta_{22})e^{n_2}(\theta,\eta_{11},\eta_{12},\eta_{22})f^{n_2^2}(\eta_{11},\eta_{12},\eta_{22})\,,$$

and where

$$c(\eta_{11},\eta_{12}) = \frac{\eta_{11}}{\tilde{\eta}_{11}}\,\frac{\tilde{\eta}_{12}}{\eta_{12}}\,,$$

$$d(\eta_{12},\eta_{22}) = \frac{\eta_{22}}{\tilde{\eta}_{22}} \frac{\tilde{\eta}_{12}}{\eta_{12}},$$

$$e(\theta,\eta_{11},\eta_{12},\eta_{22}) = \frac{\theta\tilde{\eta}_{12}^{n}}{(1-\theta)\tilde{\eta}_{11}^{n-1/2}\tilde{\eta}_{22}^{1/2}},$$

and

$$f(\eta_{11},\eta_{12},\eta_{22}) = \frac{\tilde{\eta}_{11}^{1/2}\tilde{\eta}_{22}^{1/2}}{\tilde{\eta}_{12}}.$$

To simplify (5) we introduce, for a given graph G,

$$F_k(l,m) = \#\{(x_1,\ldots,x_n): n_2 = k,\ e_{11} = l,\ e_{22} = m\},$$

which is the number of partitions of $V(G) = \{1,\ldots,n\}$ into two sets with $n-k$ and k vertices, respectively, such that subgraphs of G induced by these sets have l and m edges, respectively. We then have that $F_k(l,m) = F_{n-k}(m,l)$, so we need to calculate $F_k(l,m)$ only for $k \le [n/2]$, where $[r]$ denotes the largest integer not exceeding r. The probability function (5) now can be rewritten as

$$\begin{aligned} P(\mathbf{y};\theta,\eta) = {} & a(\theta,\eta_{11})b^{s}(\eta_{12}) \\ & \Sigma_{k=0}^{[n/2]}(1+\delta_{k,(n-k)})^{-1}\Sigma_{l,m}(F_k(l,m)\beta(l,m,k;\theta,\eta) \\ & + F_k(m,l)\beta(m,l,n-k;\theta,\eta)). \end{aligned} \tag{8}$$

Maximum likelihood estimators for (θ,η) can be calculated by applying a standard numerical maximization routine to (8). A considerable part of the necessary computer time is taken by the calculation of the numbers $F_k(l,m)$, even though this computation is done only once, before the maximization. The number of steps in calculating $F_k(l,m)$ is an exponential function of n, which restricts the applicability of the direct numerical maximization of the likelihood function to small values of n.

4.2 The EM Algorithm

Dempster, Laird, and Rubin (1977) introduced the EM (expectation-maximization) algorithm for the calculation of maximum likelihood estimates in statistical problems with missing data. In our model, the vertex color vector $\mathbf{x}$ can be regarded as missing data, and the EM algorithm offers an alternative to the straight numerical maximization of (5) as a function of (θ,η). Application of Dempster et al. (1977) leads to the following algorithm. First, for given data $\mathbf{y}$, define

$$Q(\theta',\eta' \mid \theta,\eta) = E(\log p(\mathbf{y},\mathbf{X};\theta',\eta' \mid \mathbf{y},\theta,\eta)) .$$

The EM iteration $(\theta^{(p)},\eta^{(p)}) \to (\theta^{(p+1)},\eta^{(p+1)})$ now proceeds as follows:

E: Compute $Q(\theta,\eta \mid \theta^{(p)},\eta^{(p)})$;
M: Choose $(\theta^{(p+1)},\eta^{(p+1)})$ to be a value of (θ,η) that maximizes $Q(\theta,\eta \mid \theta^{(p)},\eta^{(p)})$.

It follows from Dempster et al. (1977) that this iteration scheme converges to the maximum likelihood estimate.

Since

$$\log p(\mathbf{y},\mathbf{x};\theta,\eta) = \log a(\theta,\eta_{11}) + s \log b(\eta_{12}) + e_{11} \log c(\eta_{11},\eta_{12})$$

$$+ e_{22} \log d(\eta_{12},\eta_{22}) + n_2 \log e(\theta,\eta_{11},\eta_{12},\eta_{22}) + n_2^2 \log f(\eta_{11},\eta_{12},\eta_{22}) ,$$

where s is a function of $\mathbf{y}$ only, whereas e_{11} and e_{22} also depend on $\mathbf{x}$, the E-step consists of the calculation of the conditional expectations of the vector of canonical sufficient statistics $(N_2,N_2^2,E_{11},E_{22})$. Denote these conditional expectations by $(\bar{n}_2,\bar{m}_2,\bar{e}_{11},\bar{e}_{22})$. Thus, e.g., $\bar{e}_{ii}$ for $i = 1,2$, is given by

$$\bar{e}_{ii} = E(E_{ii} \mid \mathbf{y},\theta,\eta) =$$

$$(a(\theta,\eta_{11})b^s(\eta_{12})/P(\mathbf{y};\theta,\eta)) \sum_{k=0}^{[n/2]} (1 + \delta_{k,(n-k)})^{-1}$$

$$\sum_{e_{11},e_{22}} (e_{ii}F_k(e_{11},e_{22})\beta(e_{11},e_{22},k;\theta,\eta) +$$

$$e_{(3-i)(3-i)}F_k(e_{22},e_{11})\beta(e_{22},e_{11},n-k;\theta,\eta)) .$$

For the specification of the M-step it may be noted that the family of complete data distributions $P(\mathbf{y},\mathbf{x};\theta,\eta)$ forms a curved exponential family, so that the remarks in Dempster *et al.* (1977, pp. 5-6) apply. It follows that the M-step corresponds to maximum likelihood estimation of (θ,η) under the assumption that the outcome of the vector of canonical sufficient statistics $(N_2,N_2^2,E_{11},E_{22})$ for the complete data problem is given by $(\bar{n}_2,\bar{m}_2,\bar{e}_{11},\bar{e}_{22})$. The result is

$$\hat{\theta} = \frac{\bar{n}_2}{n} ,$$

$$\hat{\eta}_{11} = \frac{\bar{e}_{11}}{\binom{n}{2} - \left[n - \frac{1}{2}\right]\bar{n}_2 + \frac{1}{2}\bar{m}_2} ,$$

$$\hat{\eta}_{22} = \frac{2\bar{e}_{22}}{\bar{m}_2 - \bar{n}_2},$$

and

$$\hat{\eta}_{12} = \frac{s - \bar{e}_{11} - \bar{e}_{22}}{n\bar{n}_2 - \bar{m}_2}.$$

The EM algorithm is, in our experience, considerably faster than the direct maximization of (8) by standard multiparameter maximization methods. Its use is restricted, however, by the fact that the number of operations required to calculate the numbers $F_k(l,m)$ is an exponential function of n.

5. Bayesian Estimation

The algorithms for maximum likelihood estimation of θ and η, discussed in the preceding section, are not practical unless n is rather small, say, not more than 20. An excellent alternative, practical also for larger graphs, is offered by a Bayesian approach. Gibbs sampling, proposed by Geman and Geman (1983) and explained by Gelfand and Smith (1990) and Casella and George (1992), is a simulation method that can be used also for larger values of n to calculate Bayes estimates of the parameters. For those who have such strong objectivistic feelings that they would object to a Bayesian approach, it may be a valid counter-argument that already for intermediate values of n, the information contained in the data $\mathbf{y}$ about the parameters η (note that the number of Bernoulli variables relevant for estimation of these parameters is a quadratic function of n) and the vertex colors $\mathbf{x}$ is so large that the influence of the prior is quite small. Bayes estimators and the posterior standard deviations of the parameters can approximate for $n \to \infty$ the maximum likelihood estimators for θ and η_{hk} and their standard errors; see Lehmann (1983, Section 6.7) or Press (1989, Theorem 3.2.1).

The Gibbs sampler is an iterative simulation scheme that operates by repeatedly drawing in turn each of a set of unknown random variables or vectors, each conditionally on the values of all the other random variables. For a more extensive explanation we refer to the literature cited earlier. We apply this scheme to $(\theta,\eta),X_1,\ldots,X_n$, treating (θ,η) as a single random vector. This approach leads to the following procedure. Let the prior density of the parameter vector (θ,η) be by $f(\theta,\eta)$. Given current values $\mathbf{X}^{(p)}$, $\theta^{(p)},\eta^{(p)}$, the next values $\mathbf{X}^{(p+1)}$, $\theta^{(p+1)},\eta^{(p+1)}$ are determined as follows:

1. $\theta^{(p+1)}$, $\eta^{(p+1)}$ is drawn from the posterior distribution of (θ,η), given the complete data $(\mathbf{X}^{(p)},\mathbf{y})$;

2. $X_1^{(p+1)}$ is drawn from the conditional distribution of X_1 given the values $\theta^{(p+1)}, \eta^{(p+1)}, \mathbf{y}, X_2^{(p)}, \ldots, X_n^{(p)}$;

   For each value $i = 2, \ldots, n-1$ in turn, $X_i^{(p+1)}$ is drawn from the conditional distribution of X_i given the values $\theta^{(p+1)}, \eta^{(p+1)}, \mathbf{y}, X_h^{(p+1)}$ for $h = 1, \ldots, i-1$, and $X_h^{(p)}$ for $h = i+1, \ldots, n$;

   $X_n^{(p+1)}$ is drawn from the conditional distribution of X_n given the values $\theta^{(p+1)}, \eta^{(p+1)}, \mathbf{y}, X_1^{(p+1)}, \ldots, X_{n-1}^{(p+1)}$.

It follows from the convergence theorem in Geman and Geman (1983) that for this iteration scheme, irrespective of the starting values, the distribution of $(\theta^{(p)}, \eta^{(p)})$ converges to the posterior distribution given the observed data $\mathbf{y}$, and the distribution of $\mathbf{X}^{(p)}$ converges to the Bayesian posterior predictive distribution with probability function

$$f(\mathbf{x} \mid \mathbf{y}) \propto \int P(\mathbf{y},\mathbf{x} \mid \theta,\eta) f(\theta,\eta) d\theta d\eta \,.$$

The Gibbs sampler is so convenient because the conditional distributions from which $\mathbf{X}^{(p+1)}$ and $(\theta^{(p+1)}, \eta^{(p+1)})$ are drawn are quite simple. The conditional distribution of X_i given θ, η, $\mathbf{y}$, and $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, can be derived from (6). For a given i, and for $k = 1,2$, define

$$n_k' = \sum_{1 \le j \le n,\, j \ne i} I(X_j = k)$$

and

$$e_k' = \sum_{1 \le j \le n,\, j \ne i} y_{ij} I(X_j = k) \,.$$

Scrutiny of the influence of X_i on probability (6) yields

$$\frac{P(X_i = 2 \mid \mathbf{y}, X_j \text{ for } j \ne i)}{P(X_i = 1 \mid \mathbf{y}, X_j \text{ for } j \ne i)} = \frac{\theta}{1-\theta} \tilde{\eta}_{11}^{-n_1'} \tilde{\eta}_{12}^{n_1' - n_2'} \tilde{\eta}_{22}^{n_2'} \left[\frac{\eta_{11}}{\tilde{\eta}_{11}}\right]^{-e_1'} \left[\frac{\eta_{12}}{\tilde{\eta}_{12}}\right]^{e_1' - e_2'} \left[\frac{\eta_{22}}{\tilde{\eta}_{22}}\right]^{e_2'} . \tag{9}$$

This ratio determines the probability distribution from which $X_i^{(p+1)}$ is to be drawn in Step (2) of the Gibbs sampler.

The posterior distribution of (θ,η), given the complete data, has density proportional to $f(\theta,\eta)P(\mathbf{y},\mathbf{x} \mid \theta,\eta)$ with $P(\mathbf{y},\mathbf{x} \mid \theta,\eta)$ given in (6). One way of dealing with the identifiability problem mentioned above for the parameters, is to use a prior distribution with support $\{\eta_{11} \le \eta_{22}\}$. If a flat prior is required that indicates absence of special prior information, the uniform distribution on

$$\{(\theta,\eta_{11},\eta_{12},\eta_{22}) \in [0,1]^4 \mid \eta_{11} \leq \eta_{22}\}$$

is a natural choice. Suppose that this is indeed the prior distribution. Since the uniform distribution is identical to the Beta(1,1) distribution, well-known results on the Bayesian analysis of binomially distributed data with Beta prior distributions can be used to derive the posterior (see, e.g., Press 1989, p. 53). The posterior distribution of $(\theta,\eta_{11},\eta_{12},\eta_{22})$ given the complete data $(\mathbf{y},\mathbf{x})$ is given by Beta distributions with parameters, respectively,

$$(n_1 + 1, n_2 + 1), \left[\binom{n_1}{2} - e_{11} + 1, e_{11} + 1\right],$$

$$(n_1 n_2 - e_{12} + 1, e_{12} + 1), \left[\binom{n_2}{2} - e_{22} + 1, e_{22} + 1\right],$$

conditional on $\eta_{11} \leq \eta_{22}$ but with otherwise independent elements.

Drawing a random vector from this posterior is awkward because of the restriction $\{\eta_{11} \leq \eta_{22}\}$. A simple way to obtain the posterior with this restriction from the Gibbs sampler is the following. First note that if the uniform distribution of (θ,η) (without the restriction $\{\eta_{11} \leq \eta_{22}\}$) is used as a prior, the joint posterior distribution of $(\mathbf{x},\theta,\eta)$ is invariant under the transformation that replaces θ by $1-\theta$, interchanges η_{11} and η_{22}, and replaces x_i by $3-x_i$ (all i). Also note that $\eta_{11} \leq \eta_{22}$ holds for exactly one of the resulting pair of outcomes of $(\mathbf{x},\theta,\eta)$, unless $\eta_{11} = \eta_{22}$, which has probability 0. The procedure now is that in Step 1 of the Gibbs sampler, the restriction $\{\eta_{11} \leq \eta_{22}\}$ is not made and, at the end of Step 1, the following is added:

1^+. If $\eta_{11}^{(p+1)} > \eta_{22}^{(p+1)}$, then interchange $\eta_{11}^{(p+1)}$ and $\eta_{22}^{(p+1)}$, and replace $\theta^{(p+1)}$ by $1-\theta^{(p+1)}$.

The average of $(\theta^{(p)},\eta^{(p)})$ over a large number of runs after convergence is a good Monte Carlo estimate of the Bayes estimate (the posterior mean); the standard deviation of $(\theta^{(p)},\eta^{(p)})$ is an estimate of the posterior standard deviation and hence an approximation of the standard error of estimation. We shall see below that the empirical distributions of the $X_i^{(p)}$ can be used for the prediction of the vertex colors.

The iteration steps of the Gibbs sampler are quite simple. Detecting convergence, however, is not straightforward (as also observed by Casella and George 1992, Section 5.1), because the process converges not to a single value but to a stationary probability *distribution*. Gelman and Rubin (1992) and their discussants debate this question; Gelman and Rubin propose to use multiple starting points. We do not wish to go deeply into this discussion, but we are convinced that it is very sensible to use multiple starting points, and we have implemented the following procedure.

1. Determine several (e.g., $g = 10$) "local" modes of the joint probability function $P(\mathbf{y},\mathbf{x} \mid \theta,\eta)$ as a function of $(\mathbf{x},\theta,\eta)$, using the method of Section 7.1 below;
2. Apply the Gibbs sampler g times with each of these "local" modes as starting points; iterate the algorithm 10,000 times to achieve convergence and then another 10,000 times to estimate the posterior probability distribution of (θ,η) and of $(X_1, \ldots, X_n)$;
3. If these g estimated probability distributions are quite close to each other, then assume that the algorithm has converged; if not, rerun the Gibbs sampler with more than 10,000 iterations.

In our experience, the 10 Gibbs sequences usually produced quite similar distributions.

6. Asymptotic Recovery of Colors

One of the main goals in posterior blockmodeling is to recover (or predict) the colors x_i from the observation of the edge pattern $\mathbf{y}$. It turns out that asymptotically for $n \to \infty$, it is possible, under certain weak conditions, to recover the colors x_i correctly with probability tending to 1. This property will be called *the asymptotically correct distinction of vertex colors.* In our experience, depending on the values of the parameters (θ,η), a quite good recovery of the colors for a latent two-blockmodel is possible for values $n = 30$ and higher. This finding is important because it implies that, if the assumption contained in Definition 1 is valid and one uses a good procedure for recovery of colors, statistical inference in a posteriori blockmodeling can be almost as good as inference in blockmodeling with a priori given blocks.

More formally, this statement can be expressed as follows. This section shows that there exists a function $F(\mathbf{Y})$ such that $P(\mathbf{X} = F(\mathbf{Y}) \mid \theta,\eta) \to 1$ for all θ,η, as $n \to \infty$. Therefore, for any statistical procedure (test, estimator, or whatever) $T(\mathbf{X},\mathbf{Y})$ that can be derived for the model in which $\mathbf{X}$ as well as $\mathbf{Y}$ are observed, there corresponds a procedure $T(F(\mathbf{Y}),\mathbf{Y})$ which is a function only of $\mathbf{Y}$ and not of $\mathbf{X}$, and which has asymptotically the same properties as $T(\mathbf{X},\mathbf{Y})$ in the sense that for all θ,η,

$$\lim_{n\to\infty} P(T(F(\mathbf{Y}),\mathbf{Y}) = T(\mathbf{X},\mathbf{Y}) \mid \theta,\eta) = 1 .$$

A procedure to accomplish the asymptotically correct distinction of vertex colors in a two-blockmodel can be based on the degrees, as indicated in the following theorem. This procedure was also proposed in Frank and Nowicki (1993, Section 4). Grusho (1984) proved a similar result under more restrictive model conditions.

Theorem 1. *Denote by* $y_{(i)+}$ *the ordered degrees:* $y_{(1)+} \le y_{(2)+} \le \ldots \le y_{(n)+}$. *Let I be the index i with* $1 \le i \le n-1$ *for which* $y_{(i+1)+} - y_{(i)+}$ *is maximal (if there are several such indices, let I be the smallest among them) and denote* $D = y_{(I)+}$. *Define*

$$F_i(\mathbf{y}) = \begin{cases} 1 & \text{if } y_{i+} \le D; \\ 2 & \text{if } y_{i+} > D. \end{cases}$$

Let $n \to \infty$ *while* θ *and* η *are fixed, and assume that*

$$n_2/n \to \theta \in (0,1)\,,$$

$$\theta\eta_{12} + (1-\theta)\eta_{11} < \theta\eta_{22} + (1-\theta)\eta_{12}\,. \tag{10}$$

Then

$$P(X_i = F_i(\mathbf{Y}) \text{ for } i = 1, \ldots, n \mid \mathbf{X} = \mathbf{x}) \to 1\,.$$

Remarks.

1. The condition $n_2/n \to \theta$ holds with probability 1 if the two-blockmodel is valid.
2. If the reverse inequality sign > holds in (10), then the same conclusion holds when the colors 1 and 2 in the definition of F_i are interchanged. Therefore, condition (10) effectively excludes only a 1-dimensional subset in the 4-dimensional parameter space.

The proof of this theorem is given in the Appendix.

The procedure can be expressed in the following words: order the degrees and find the greatest gap between the degrees; vertices with a degree higher than this gap are designated white, the remaining vertices are designated black.

Condition (10) is necessary and sufficient for the vertex colors to be distinguishable correctly with probability tending to 1 on the basis of only the degrees. It is, however, not a necessary condition for asymptotically correct distinction *per se*. We now describe a slightly more complicated procedure for the asymptotically correct distinction of vertex colors that works whenever the three probabilities η_{11}, η_{12}, and η_{22} are not all the same (if $\eta_{11} = \eta_{12} = \eta_{22}$, then the graph really is a Bernoulli graph and one single block rather than two is sufficient). This procedure makes use of the inner products $C_{ij} = \Sigma_{h \ne i,j} Y_{hi} Y_{hj}$. Note that C_{ij} is the number of other vertices with which both i and j are connected, which can also be expressed as the cardinality of the intersection of the neighbourhoods of i and j. (There is a remote similarity to the CONCOV procedure (see, e.g., Schwartz 1977), because CONCOV employs covariances between rows of the adjacency matrix while

our procedure utilizes inner products; however, the remainder of CONCOV and our procedure are completely different, among other reasons because CONCOV works with iterated covariances.) According to some of our Monte Carlo simulations, this second procedure works better in practice than the procedure of Theorem 1. The procedure is a kind of clustering algorithm, carried out by first ordering the vertices and then splitting the ordered set of vertices into two blocks.

The sequential ordering of the vertices produces the vector $\mathbf{s}$ of the ordered vertex numbers and an ordering vector $\mathbf{d}$ and works as follows. First, we choose the pair of vertices i and j, say, with the maximum value of C_{ij} and set $s_1 = i$, $s_2 = j$ and $d_1 = d_2 = C_{ij}$. Next, suppose that we obtained $s_1, \ldots, s_i$ and $d_1, \ldots, d_i$. To determine s_{i+1} calculate, for each remaining vertex $j \notin \{s_1, \ldots, s_i\}$,

$$C^*_{ij} = \min(C_{hj} \mid h \in \{s_1, \ldots, s_i\}). \tag{11}$$

Choose then vertex, r, say, such that C^*_{ir} is maximal, and set $s_{i+1} = r$ and $d_{i+1} = C^*_{ir}$.

The splitting step proceeds as follows: let $d_{t+1} - d_t$ be the greatest gap between two consecutive values in $\mathbf{d}$. We then assign vertices $\{s_1, \ldots, s_t\}$ to one block and all the remaining vertices to the other.

This approach produces the desired block structure. It is proven in the Appendix that this procedure also produces an asymptotically correct recovery of vertex colors.

7. Color Prediction for Finite n

The procedures of the previous section yield asymptotically for $n \to \infty$ the correct coloration, but are not necessarily satisfactory for small and intermediate values of n. In this section, we consider the problem of predicting the colors $X_1, \ldots, X_n$ of the vertices $1, \ldots, n$ based on the observed graph $\mathbf{y}$ in the setting of a fixed value for n. This problem of a posteriori stochastic blockmodeling was studied by Wasserman and Anderson (1987) and Anderson, Wasserman, and Faust (1992) in the context of the log-linear p_1 model for digraphs. Those authors proposed a classification of the vertices based on similarity of vertex parameters estimated under the assumption that the p_1 model holds. We do not make this assumption.

In this section we present three predictive approaches. The first is based on a profile predictive likelihood and the second on a conditional predictive likelihood. The third approach is Bayesian and is a direct result of the Gibbs sampling procedure of Section 5.

7.1 The Profile Predictive Likelihood

Prediction of $\mathbf{X}$ can be based on Mathiasen's (1979) likelihood-based function given by

$$L_p(\mathbf{x} \mid \mathbf{y}) = \sup_{\theta,\eta} P(\mathbf{y},\mathbf{x};\theta,\eta) .$$

This function is motivated by replacing the unknown parameters $\theta,\eta_{11},\eta_{12},\eta_{22}$ with their most likely values, given the complete data $(\mathbf{y},\mathbf{x})$. Because of the correspondence to the profile likelihood in parametric inference, L_p has been called the profile predictive likelihood; see Bjørnstad (1990) for more details.

First, we observe that ML-estimators given the complete data are given by $\hat{\theta} = n_2/n$, $\hat{\eta}_{11} = e_{11} / \binom{n_1}{2}$, $\hat{\eta}_{12} = e_{12}/n_1 n_2$ and $\hat{\eta}_{22} = e_{22} / \binom{n_2}{2}$. Thus we obtain that

$$L_p(\mathbf{x} \mid \mathbf{y}) =$$

$$\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2} \left(1 - \frac{e_{11}}{\binom{n_1}{2}}\right)^{\binom{n_1}{2}-e_{11}} \left(1 - \frac{e_{12}}{n_1 n_2}\right)^{n_1 n_2 - e_{12}} \times$$

$$\left(1 - \frac{e_{22}}{\binom{n_2}{2}}\right)^{\binom{n_2}{2}-e_{22}} \left(\frac{e_{11}}{\binom{n_1}{2}}\right)^{e_{11}} \left(\frac{e_{12}}{n_1 n_2}\right)^{e_{12}} \left(\frac{e_{22}}{\binom{n_2}{2}}\right)^{e_{22}} .$$

If n is not too large, $L_p(\mathbf{x} \mid \mathbf{y})$ can be calculated for all $\mathbf{x}$, and the values for $\mathbf{x}$ with the highest value for $L_p(\mathbf{x} \mid \mathbf{y})$ can be chosen as likely partitions of the vertex set into blocks. The disadvantage of this approach is that it assumes the unknown parameters to be equal to their maximum likelihood estimates, which may imply a misleading impression of precision, especially for low values of n.

For larger values of n, maximizing $L_p(\mathbf{x} \mid \mathbf{y})$ by enumerating all $\mathbf{x}$ is not practical. A feasible alternative numerical method for seeking the value of $\mathbf{x}$ that maximizes $L_p(\mathbf{x} \mid \mathbf{y})$ is based on the observation that such an $\mathbf{x}$ is also the $\mathbf{x}$-coordinate of the maximum of $P(\mathbf{y},\mathbf{x};\theta,\eta)$ as a function of $\mathbf{x},\theta,\eta$ for fixed $\mathbf{y}$. We define a "local" mode of the likelihood $P(\mathbf{y},\mathbf{x};\theta,\eta)$ as a value of $(\theta,\eta,\mathbf{x})$, where a change in (θ,η) or in any of the x_i separately does not lead to an increase in the likelihood. Such "local" modes can found by alternating the two following steps:

1. Maximize $P(\mathbf{y},\mathbf{x};\theta,\eta)$ over (θ,η); this approach is the same as computing the maximum-likelihood estimate for observed data $\mathbf{y},\mathbf{x}$;
2. For each $i = 1,...,n$, maximize $P(\mathbf{y},\mathbf{x};\theta,\eta)$ as a function of $x_i \in \{1,2\}$.

These steps can be taken from random starting points, or starting from the values for $\mathbf{x}$ obtained by the procedures proposed in Section 6; the procedure converges when the second step does not lead to any changes in the current value of $\mathbf{x}$. In our experience, this procedure is quite satisfactory when applied to a reasonably large number (e.g., 100) of random starting points, but it cannot be proved that this approach yields a global maximum of the function $L_p(\mathbf{x} \mid \mathbf{y})$.

7.2 The Conditional Predictive Likelihood

Butler (1986) proposed the conditional predictive likelihood based on the minimal sufficient statistic $t(\mathbf{y},\mathbf{x})$ for $(\mathbf{y},\mathbf{x})$, defined as

$$L_c(\mathbf{x} \mid \mathbf{y}) = \frac{P(\mathbf{y},\mathbf{x};\theta,\eta)}{P(t(\mathbf{y},\mathbf{x});\theta,\eta)}.$$

We refer to Butler (1986) for the motivation for this expression. The minimal sufficient statistic for $(\mathbf{y},\mathbf{x})$ is given by $t(\mathbf{y},\mathbf{x}) = (N_2,E_{11},E_{22},E_{12})$.

To calculate the probability function of $t(\mathbf{y},\mathbf{x})$ we note that, conditional on the fact that there are $n_1 = n - n_2$ vertices of color 1, E_{11} has the binomial distribution with parameters $\binom{n_1}{2}$ and η_{11}. A similar reasoning can be applied to E_{22} and E_{12}. Moreover, E_{11}, E_{12} and E_{22} are independent, conditional on n_2. Hence the conditional probability function for E_{11},E_{22},E_{12} given n_2 is

$$\begin{aligned}
&P(e_{11},e_{22},e_{12} \mid n_2) \\
&= P(e_{11} \mid n_2)P(e_{22} \mid n_2)P(e_{12} \mid n_2) \\
&= \binom{\binom{n_1}{2}}{e_{11}} \binom{n_1 n_2}{e_{12}} \binom{\binom{n_2}{2}}{e_{22}} \\
&\times \tilde{\eta}_{11}^{\binom{n_1}{2}} \tilde{\eta}_{12}^{n_1 n_2} \tilde{\eta}_{22}^{\binom{n_2}{2}} \left(\frac{\eta_{11}}{\tilde{\eta}_{11}}\right)^{e_{11}} \left(\frac{\eta_{12}}{\tilde{\eta}_{12}}\right)^{e_{12}} \left(\frac{\eta_{22}}{\tilde{\eta}_{22}}\right)^{e_{22}}.
\end{aligned}$$

Further, recall that N_2 is binomially distributed with parameters n and θ. Hence, the conditional predictive likelihood is given by

$$L_c(\mathbf{x} \mid \mathbf{y}) = \left[\binom{n}{n_2} \binom{n_1}{2} \binom{n_1 n_2}{e_{11}} \binom{n_1 n_2}{e_{12}} \binom{n_2}{2} \binom{n_2}{e_{22}} \right]^{-1} .$$

Butler (1986) suggests that this function is more reliable than $L_p(\mathbf{x} \mid \mathbf{y})$ in predicting the partitioning of the vertex set into blocks. To find the maximum of $L_c(\mathbf{x} \mid \mathbf{y})$, procedures similar to those of Section 7.1 can be followed: maximization by complete enumeration (feasible only for small n), or determination of "local" maxima by maximizing successively over each of the coordinates x_i.

7.3 Bayesian Prediction

The Gibbs sampler of Section 5 can be used not only for parameter estimation but also for color prediction. The limiting distribution of $\mathbf{X}^{(p)}$, when the algorithm of Section 5 is followed, is the Bayesian posterior predictive distribution with probability function

$$f(\mathbf{x} \mid \mathbf{y}) \propto \int P(\mathbf{y},\mathbf{x} \mid \theta,\eta) f(\theta,\eta) d\theta d\eta .$$

This expression is just the conditional distribution of the vertex colors $\mathbf{X}$ given the observed edges $\mathbf{Y}$, when the parameters (θ,η) have prior density function $f(\theta,\eta)$.

Hence the relative frequencies of $\{X_i^{(p)} = k\}$ for $k = 1,2$ are Monte Carlo estimates of the Bayesian posterior predictive probabilities for $X_i = k$ and can be used for predicting the vertex colors. A satisfactory two-block structure may be assumed to have been found if these relative frequencies are close to 0 or 1 for all vertices. It can be concluded that the Gibbs sampler automatically produces a color prediction.

In many examples that we tried, the second procedure of Section 6 (based on the matrix $\mathbf{C}$), followed by the iteration steps of Section 7.1 to obtain a "local" mode of $P(\mathbf{y},\mathbf{x};\theta,\eta)$, yielded a block structure $\mathbf{x}$ that was also the mode of the posterior predictive distribution of $\mathbf{X}$ as estimated by the Gibbs sampler.

8. Example: Hansell's Student Data

Here we will discuss Hansell's (1984) classroom data previously studied in the context of stochastic blockmodels by Wang and Wong (1987). Our example consists of the sociomatrix of friendship among 27 classmates: 13 male and 14 female sixth-graders in an inner-city Baltimore elementary school; see Hansell (1984) or Wang and Wong (1987) for more details. Since the original data are asymmetric, we have chosen to symmetrize the data in our analysis by assuming the presence of friendship between two students whenever at least one of the two students expressed liking for the other.

Table 1: Adjacency matrix of Hansell's friendship data.

	1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
1	- 1 1 1 1 0 0 1 1 1 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0
2	1 - 1 1 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0
3	1 1 - 1 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 1 0 0 1 0 0 0
4	1 1 1 - 1 1 1 1 1 1 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 0 0
5	1 1 0 1 - 1 1 1 1 1 1 0 1 0 1 1 1 1 1 0 1 0 1 1 1 1 0 0
6	0 1 0 1 1 - 1 1 0 1 0 1 1 1 0 1 1 1 0 1 0 1 1 0 1 0 0
7	0 0 0 1 1 1 - 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
8	1 0 0 1 1 1 0 - 1 1 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0
9	1 0 1 1 1 0 1 1 - 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10	1 0 0 1 1 1 0 1 0 - 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0
11	0 0 0 1 0 0 0 1 0 0 - 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0
12	0 0 0 0 1 1 0 0 0 1 0 - 0 0 0 0 0 0 1 1 0 0 0 0 0 0
13	0 0 1 1 0 1 1 1 1 0 0 0 - 0 0 0 1 0 0 0 1 0 0 0 0 0 0
14	0 0 1 0 1 1 0 0 0 0 0 0 0 - 1 1 1 1 1 0 1 0 0 1 1 0 0
15	1 1 0 0 1 0 0 0 0 0 0 0 0 1 - 1 0 1 1 1 1 1 1 1 1 0 0
16	1 1 0 1 1 1 0 1 0 1 1 0 0 1 1 - 1 1 1 0 1 1 1 1 1 0 1
17	0 0 0 1 1 1 0 0 0 0 1 0 1 1 0 1 - 1 1 1 0 0 1 0 1 0 0
18	0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 1 - 1 1 1 1 0 0 1 0 0
19	0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 - 0 1 1 0 1 1 0 0
20	1 0 1 0 1 1 1 0 0 0 1 1 0 0 1 0 1 1 0 - 0 0 0 1 0 0 0
21	1 0 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 - 0 1 0 1 1 0
22	0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 - 1 1 1 0 1
23	0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 1 1 - 1 1 0 0
24	0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 1 0 0 1 1 0 1 1 - 1 0 0
25	0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 1 1 1 - 0 0
26	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 - 0
27	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 -

The adjacency matrix given in Table 1 presents data blocked according to sex; individuals labeled 1-13 are male students, while those labeled 14-27 are female students. Among same-sex pairs, the fraction of friendship ties is 0.53 for male students and 0.56 for female students. Among opposite-sex pairs, the fraction of friendship ties is 0.23.

In our analysis we wish to divide the students into two groups based on their friendship pattern, and investigate whether the partitioning of the class into two subgroups based on friendship ties differs from the partitioning based exclusively on gender. (Wang and Wong (1987) used another block-related approach by trying to improve the fit provided by the p_1 model through making use of the extra nodal information about the gender of the students.)

A preliminary block division of students can be carried out by applying the second procedure from Section 6, based on the matrix **C**. We obtain the block structure

$$\mathbf{x}^{(1)} = (111122111111122222212212211).$$

Using multiple starting points of the Gibbs sampler (block structure $\mathbf{x}^{(1)}$ and local modes as defined in Section 7.1) led to the conclusion that the block structure found does not depend on the starting point. The use of $\mathbf{x}^{(1)}$ as the starting point led to the posterior means:

$$\hat{\theta} = 0.479, \quad \hat{\eta}_{11} = 0.326,$$

$$\hat{\eta}_{12} = 0.248, \quad \hat{\eta}_{22} = 0.763,$$

with the estimated posterior covariance matrix:

$$\hat{\Sigma}_{\theta,\eta} = \begin{bmatrix} 0.0148 & -0.0013 & -0.0036 & -0.0033 \\ -0.0013 & 0.0038 & -0.0005 & 0.0015 \\ -0.0036 & -0.0005 & 0.0054 & 0.0006 \\ -0.0033 & 0.0015 & 0.0006 & 0.0064 \end{bmatrix}.$$

The square roots of the posterior variances may be treated as standard errors of the estimated parameters:

$$S.E.(\hat{\theta}) = 0.122, \quad S.E.(\hat{\eta}_{11}) = 0.062,$$
$$S.E.(\hat{\eta}_{12}) = 0.073, \quad S.E.(\hat{\eta}_{22}) = 0.080.$$

For individuals labeled 5-6, 14-19, and 21-25, the posterior predictive probability of belonging to block 2 was .89 or more with the exception of individual 21 (0.81). For the others, the posterior predictive probability of belonging to block 1 was .93 or more with the exception of individual 20 (0.80). These results imply a quite clear-cut posterior block structure. Block 2 has a high within-block density; the densities within Block 1 and between the blocks are much lower.

This finding can now be compared with blocking induced by gender. The posterior blocking does not entirely follow gender lines, because male students labeled 5 and 6 belong now to the "female" group and female students labeled 20 and 26-27 belong now to the "male" group. Table 1 shows that these individuals indeed have different friendship patterns from the others: 5 and 6 have relatively many female friends, 20 has more male than female friends, while 26 and 27 have few friends of either gender.

9. Discussion

In this paper we have presented a statistical approach to probabilistic posterior blockmodeling. We restricted attention to the simple case of a two-blockmodel for an undirected graph. This allows us to uncover what is, and what is not, possible in such an approach in the simplest situation. Implementation possibilities depend strongly on the number n of vertices. A maximum likelihood solution appeared to be practically feasible only for small values of n (up to about 20). It turned out that a Bayesian solution, using the Gibbs sampling algorithm, can be satisfactorily applied also for large n. For intermediate and large values of n, the Bayesian estimator (i.e., the posterior mean) with a high probability comes very close to the maximum likelihood estimator. The posterior predictive distribution can be used to estimate the block structure.

We also obtained the result that if a probabilistic two-blockmodel holds, the block structure (represented by the vector $\mathbf{x}$) can be correctly recovered with probability tending to 1, as n increases. We found indeed, when applying the Gibbs sampler to artificial data generated according to a probabilistic two-blockmodel, that for $n \geq 30$ and θ not too close to 0 or 1, the posterior predictive probability distribution for $\mathbf{X}$ tended to be concentrated closely around the true value.

From this work on the simple situation of a two-blockmodel for an undirected graph, we conclude that maximum likelihood estimation of the parameters is not worth the trouble of being developed for multiple-blockmodels for undirected or directed graphs. Our future research will be directed toward using the Gibbs sampling method for blockmodels for undirected and directed graphs with an arbitrary number of blocks.

Appendix

Proof of Theorem 1.

Suppose that $x_i = 1$; then the conditional distribution of Y_{i+} given the colors $x_1,\ldots,x_n$ is the convolution of $\mathrm{Bin}(n_1 - 1,\eta_{11})$ and $\mathrm{Bin}(n_2,\eta_{12})$. Well-known large deviations theorems imply that if S has the binomial $\mathrm{Bin}(m,p)$ distribution, then

$$P(\mid S/m - p \mid > \varepsilon) \leq 2\exp(-2m\varepsilon^2)$$

(see, e.g., Feller 1968, section VIII.4).

Define $p_i(\mathbf{x}) = E(Y_{i+}/(n-1) \mid \mathbf{X} = x)$; then

$$p_i(\mathbf{x}) = \begin{cases} \eta_{11}(n_1 - 1)/(n-1) + \eta_{12}n_2/(n-1) & \text{for } x_i = 1; \\ \eta_{12}n_1/(n-1) + \eta_{22}(n_2 - 1)/(n-1) & \text{for } x_i = 2. \end{cases}$$

Now let ε be an arbitrary positive number. Applying the large deviations theorem separately to $Y_{i+}^{(s)}$ and to $Y_{i+}^{(d)}$, where $Y_{i+}^{(s)}$ denotes the number of edges from i to all other vertices of the same color as i and $Y_{i+}^{(d)}$ the number of edges to differently colored vertices, and using Boole's inequality, yields

$$\begin{aligned} P(\mid Y_{i+}/(n-1) - p_i(\mathbf{x}) \mid > \varepsilon \mid \mathbf{X} = x) \\ &\leq 2\exp(-\varepsilon^2(n-1)^2/(2n_1)) \\ &\quad + 2\exp(-\varepsilon^2(n-1)^2/(2n_2)) \\ &\leq 4\exp(-\varepsilon^2(n-1)/2). \end{aligned}$$

Using Boole's inequality again, it follows that

$$\begin{aligned} P(\mid Y_{i+}/(n-1) - p_i(\mathbf{x}) \mid > \varepsilon \text{ for at least one } i \mid \mathbf{X} = x) \\ &\leq 4n\exp(-(n-1)\varepsilon^2/2) \to 0. \end{aligned} \tag{12}$$

Now let

$$\alpha_1 = \theta\eta_{12} + (1-\theta)\eta_{11}, \quad \alpha_2 = \theta\eta_{22} + (1-\theta)\eta_{12},$$

and

$$\varepsilon = \mid \alpha_1 - \alpha_2 \mid /4.$$

The assumptions imply that $\varepsilon > 0$. It follows from (12) that for n sufficiently large,

$$P(\mid Y_{i+}/(n-1) - \alpha_{x_i} \mid > \varepsilon \text{ for at least one } i \mid \mathbf{X} = x) \to 0. \tag{13}$$

In words, the probability is almost 1 that all normalized degrees $Y_{i+}/(n-1)$ of vertices with color k ($k = 1,2$) are closely clustered around the value α_k. Hence, with probability tending to 1, the distances between normalized degrees of vertices with the same color are all less than 2ε, while all distances between normalized degrees of differently colored vertices are greater than 2ε. This result implies that, with probability tending to 1, the greatest gap between the degrees will separate the vertices of color 1 from those of color 2. In other words, (13) implies that, with probability tending to 1, $X_i = F_i(\mathbf{Y})$ for $i = 1, \ldots, n$. ■

Convergence proof for the second procedure of Section 6.

We sketch a proof of the result that the second procedure of Section 6 yields an asymptotically correct recovery of vertex colors. Some details that are similar to parts of the preceding proof are not repeated. It is assumed that not all three of η_{11}, η_{12}, and η_{22} are identical. As in Theorem 1, it is assumed that $n_2/n \to \theta \in (0,1)$.

As a first remark, note that under the blockmodel conditional on $\mathbf{x}$, the distributions of the C_{ij} are convolutions of two binomial distributions with expected values

$$E(C_{ij} \mid \mathbf{X} = x) = \begin{cases} (n_1 - 2)\eta_{11}^2 + n_2\eta_{12}^2 & \text{if } x_i = x_j = 1; \\ (n_1 - 1)\eta_{11}\eta_{12} + (n_2 - 1)\eta_{12}\eta_{22} & \text{if } x_i \neq x_j; \\ n_1\eta_{12}^2 + (n_2 - 2)\eta_{22}^2 & \text{if } x_i = x_j = 2. \end{cases}$$

Denote the limits of these expected values, divided by n, by

$$\mu_{11} = (1-\theta)\eta_{11}^2 + \theta\eta_{12}^2 ,$$

$$\mu_{12} = (1-\theta)\eta_{11}\eta_{12} + \theta\eta_{12}\eta_{22} ,$$

and

$$\mu_{22} = (1-\theta)\eta_{12}^2 + \theta\eta_{22}^2 .$$

Then

$$\mu_{11} + \mu_{22} - 2\mu_{12} = (1-\theta)(\eta_{11} - \eta_{12})^2 + \theta(\eta_{12} - \eta_{22})^2 > 0 .$$

Just as in the proof of Theorem 1 it can be shown that, for $n \to \infty$, the values C_{ij}/n tend with probability 1 to μ_{kl} for $k = x_i$, $l = x_j$, and that this convergence in probability is uniform in $1 \leq i \neq j \leq n$.

The second procedure of Section 6 is based on the property that

$$\mu_{12} < \max(\mu_{11}, \mu_{22}) ,$$

which follows from inequality (9). Assume, arbitrarily, that $\mu_{11} \geq \mu_{22}$. Then $\mu_{12} < \mu_{11}$.

The reordering that is the first step of the procedure yields an order of the vertices for which the matrix with elements (C_{ij}/n) has, apart from asymptotically infinitesimal deviations, the block structure

$$\begin{bmatrix} \mu_{11}E_{tt} & \mu_{12}E_{t(n-t)} \\ \mu_{12}E_{(n-t)t} & \mu_{22}E_{(n-t)(n-t)} \end{bmatrix}, \tag{14}$$

where E_{ts} denotes the $t \times s$ matrix with all elements 1. For vertices j with color 1,

$$d_j/n \approx \mu_{11} ,$$

while for vertices j with color 2,

$$d_j/n \approx \min(\mu_{12},\mu_{22}) < \mu_{11} .$$

This result implies that, with probability 1, for large n, the largest gap between the d_j must occur precisely between the blocks in the block structure (14); this block structure coincides with the vertex coloring for the reordered vertices. ■

References

ANDERSON, C. J., WASSERMAN, S., and FAUST, K. (1992), "Building Stochastic Blockmodels," *Social Networks, 14*, 137-161.

ARABIE P., BOORMAN, S. A., and LEVITT, P.R. (1978), "Constructing Blockmodels: How and Why," *Journal of Mathematical Psychology, 17*, 21-63.

BJØRNSTAD, J. F. (1990), "Predictive Likelihood: A Review" (with discussion), *Statistical Science, 1*, 242-265.

BOLLOBÁS, B. (1985), *Random Graphs*, New York: Academic Press.

BORGATTI, S., EVERETT, M. G., and FREEMAN, L. C. (1992), *UCINET IV Version 1.0 Reference Manual*, Columbia, SC: Analytic Technologies.

BREIGER, R. L., BOORMAN, S. A., and ARABIE, P. (1975), "An Algorithm for Clustering Relational Data, with Applications to Social Network Analysis and Comparison with Multidimensional Scaling," *Journal of Mathematical Psychology, 12*, 328-383.

BURT, R. S. (1976), "Positions in Networks," *Social Forces, 55*, 93-122.

BUTLER, R. (1986), "Predictive Likelihood Inference with Applications" (with discussion), *Journal of the Royal Statistical Society, Series B, 48}*, 1-38.

CASELLA, G., and GEORGE, E. I. (1992), "Explaining the Gibbs Sampler," *American Statistician, 46*, 167-174.

DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

ERDÖS, P., and RÉNYI, A. (1959), "On Random Graphs 1," *Publications in Mathematics, Debrecen, 6*, 290-297.

ERDÖS, P., and RÉNYI, A. (1960), "On the Evolution of Random Graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5*, 17-61.

FAUST, K. (1988), "Comparison of Methods for Positional Analysis: Structural and General Equivalences," *Social Networks, 10*, 313-341.

FAUST, K., and WASSERMAN, S. (1992), "Blockmodels: Interpretation and Evaluation," *Social Networks, 14*, 5-61.

FELLER, W. (1968), *An Introduction to Probability Theory and its Applications, Vol. I*, New York: Wiley.

FIENBERG, S. E., MEYER, M. M., and WASSERMAN, S. (1985), "Statistical Analysis of Multiple Sociometric Relations," *Journal of the American Statistical Association, 80*, 51-67.

FIENBERG, S. E., and WASSERMAN, S. (1981), "Categorical Data Analysis of Single Sociometric Relations," in *Sociological Methodology - 1981*, Ed., S. Leinhardt, San Francisco, CA: Jossey-Bass, 156-192.

FRANK, O. (1988), "Triad Count Statistics," *Discrete Mathematics, 72*, 141-149.

FRANK, O. (1988), "Random Sampling and Social Networks: A Survey of Various Approaches," *Mathématiques, Informatique et Sciences Humaines, 26:104*, 19-33.

FRANK, O., and HARARY, F. (1982), "Cluster Inference by Using Transitivity Indices in Empirical Graphs," *Journal of the American Statistical Association, 81*, 835-840.

FRANK, O., and NOWICKI, K. (1993), "Exploratory Statistical Analysis of Networks," *Annals of Discrete Mathematics, 55*, 349-366.

GELFAND, A. E., and SMITH, A. F. M. (1990), "Sampling-based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association, 85*, 398-409.

GELMAN, A., and RUBIN, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science, 7*, 457-511.

GEMAN, S., and GEMAN, D. (1983), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI 6*, 721-741.

GRUSHO, A. A. (1982), "Certain Statistic Problems on Graphs," *Matematicheskie Zametki, 36*, 269-277.

HANSELL, S. (1984), "Cooperative Groups, Weak Ties, and the Integration of Peer Friendships," *Social Psychology Quarterly, 76*, 316-328.

HOLLAND, P., LASKEY, K. B., and LEINHARDT, S. (1983), "Stochastic Blockmodels: Some First Steps," *Social Networks, 5*, 109-137.

HOLLAND, P., and LEINHARDT, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association, 76*, 33-50.

JANSON, S., and NOWICKI, K. (1991), "The Asymptotic Distributions of Generalized U-Statistics with Applications to Random Graphs," *Probability Theory and Related Fields, 90*, 341-375.

LEHMANN, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.

LORRAIN, F., and WHITE, H. C. (1971), "Structural Equivalence of Individuals in Social Networks," *Journal of Mathematical Sociology, 1*, 49-80.

MATHIASEN, P. E. (1979), "Prediction Functions," *Scandinavian Journal of Statistics, 6*, 1-21.

PRESS, S. J. (1989), *Bayesian Statistics*, New York: Wiley.

SCHWARTZ, J. E. (1977), "An Examination of CONCOR and Related Methods for Blocking Sociometric Data," in *Sociological Methodology - 1977*, Ed., D. Heise, San Francisco, CA: Jossey-Bass, 255-282.

SCOTT, J. (1991), *Social Network Analysis: A Handbook*, Newbury Park, CA: Sage Publications.

WANG, Y. J., and WONG, G. Y. (1987), "Stochastic Blockmodels for Directed Graphs," *Journal of the American Statistical Association, 82*, 8-19.

WASSERMAN, S., and ANDERSON, C. (1987), "Stochastic a posteriori Blockmodels: Construction and Assessment," *Social Networks, 9*, 1-36.

WASSERMAN, S., and FAUST, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge and New York: Cambridge University Press.

WASSERMAN, S., and GALASKIEWICZ, J. (1984), "Some Generalizations of p_1: External Constraints, Interactions and Non-binary Relations," *Social Networks, 6*, 177-192.

WELLMAN, B., FRANK, O., ESPINOZA, V., LUNDQUIST, S., and WILSON, C. (1991), "Integrating Individual, Relational and Structural Analysis," *Social Networks, 13*, 223-249.