Social Networks 5 (1983) 109–137
North-Holland

STOCHASTIC BLOCKMODELS: FIRST STEPS *

Paul W. HOLLAND
Educational Testing Service **

Kathryn Blackmond LASKEY and Samuel LEINHARDT
Carnegie-Mellon University †

A stochastic model is proposed for social networks in which the actors in a network are partitioned into subgroups called blocks. The model provides a stochastic generalization of the blockmodel. Estimation techniques are developed for the special case of a single relation social network, with blocks specified *a priori*. An extension of the model allows for tendencies toward reciprocation of ties beyond those explained by the partition. The extended model provides a one degree-of-freedom test of the model. A numerical example from the social network literature is used to illustrate the methods.

1. Introduction

The use of relational data in social science has increased dramatically in the last twenty years. This increase has been driven by both a general theoretical interest in the structural features of networks of relations and applied concerns focusing on the behavioral implications of various structural tendencies and patterns. Methodological tools for the analysis of these data, however, are not well advanced and, consequently, research using relational data tends to be *ad hoc*, nonreplicable and nongeneralizable. This seriously restricts the cumulation of facts about networks of relations that is necessary for the development of empirically verified substantive theory.

There are, however, two relatively new approaches to relational data analysis which are potentially very useful: blockmodels (White, Boor-

* This research was supported by HHS Grant 2 RO1 HD12506-03. The order of authorship is alphabetical.
** Educational Testing Service, Princeton, NJ 08541, U.S.A.
† Carnegie-Mellon University, Pittsburgh, PA 15213, U.S.A.

Table 1
Strengths and weaknesses of two major approaches to relational data analysis

Blockmodels	Stochastic models
Strengths	
Explicit description of global structure	Explicit models for data variability
Provide a substantive model of roles and positions	Provide a direct model for relational ties
Multiple relations easily accommodated	Parameters summarize structural features
Simple fitting algorithms	Tests of fit are standard
Weaknesses	
No explicit model for data variability	No explicit global structure
Do not model relational ties	Roles and positions not incorporated
No parameters or natural numerical summaries	Multiple relations not easily incorporated
No formal tests of fit	Complex fitting algorithms

man and Breiger 1976) and stochastic models for digraphs (Holland and Leinhardt 1981a). Table 1 outlines what we believe to be the major strengths and weaknesses of the two approaches. Because of the obvious complementarity in Table 1 a merger of the two approaches promises to overcome the limitations of each while creating a statistical methodology that is consistent, effective and broadly applicable.

In this paper we present a merger of the two approaches. It is related to the work of Fienberg and Wasserman (1981). We describe stochastic multigraphs and stochastic blockmodels in Section 2. In Section 3 we describe an extension of these models which provides formal tests of the fit of stochastic blockmodels. Section 4 contains a numerical example and the paper concludes with a discussion of the relation of stochastic blockmodels to other types of blockmodels.

2. Stochastic blockmodels

A stochastic blockmodel is a model for sociometric data obtained from a network characterized by block structure. By block structure, we mean that the nodes of the network are partitioned into subgroups called blocks, and that the distribution of the ties between nodes is dependent on the blocks to which the nodes belong. The model is intended to formalize the concepts underlying the deterministic blockmodel in a framework which allows for variability in the data.

A stochastic blockmodel is a special type of probability distribution over the space of adjacency arrays. Before defining the stochastic blockmodel, we define the more general notion of a stochastic multigraph. Let G be a set of g nodes, and let $R(1),\ldots,R(m)$ be m relations defined on the pairs of nodes. We write $iR(k)j$ to mean that node i stands in relation $R(k)$ to node j.

Definition 1. The adjacency matrix for the digraph of the single relation $R(k)$ is given by:

$$\boldsymbol{x}(k) = \big(x_{ij}(k)\big), \qquad i,j = 1,\ldots,g,$$

where

$$x_{ij}(k) = \begin{cases} 1 & \text{if } iR(k)j, \\ 0 & \text{otherwise.} \end{cases}$$

By convention, we put $x_{ii}(k) = 0$, for all nodes i.

Definition 2. The adjacency array for the multigraph of the m relations $R(1),\ldots,R(m)$ is given by the "matrix of vectors":

$$\boldsymbol{x} = (\boldsymbol{x}_{ij}),$$

where $\boldsymbol{x}_{ij}$ is the vector

$$\boldsymbol{x}_{ij} = \big(x_{ij}(1),\ldots,x_{ij}(m)\big).$$

The m-element vector $\boldsymbol{x}_{ij}$ describes the entire pattern of ties from node i to node j for all m relations. We may also represent $\boldsymbol{x}$ as the "vector of matrices", $\boldsymbol{x} = (\boldsymbol{x}(1),\ldots,\boldsymbol{x}(m))$, when $\boldsymbol{x}(k)$ is the adjacency matrix for $R(k)$.

Sociometric data are regarded as observations from a probability distribution over the space of adjacency arrays. In what follows, we adopt the standard convention of denoting random quantities by uppercase letters, and particular realizations of the random quantities by lowercase letters. If $\boldsymbol{X}$ is a random adjacency array for g nodes and m relations, then the probability distribution of $\boldsymbol{X}$ is called a *stochastic multigraph.* We will denote the probability distribution of $\boldsymbol{X}$ by $p(\boldsymbol{x}) = \Pr(\boldsymbol{X} = \boldsymbol{x})$.

A *stochastic blockmodel* is a special case of a stochastic multigraph which satisfies the following requirements.

Definition 3. Let $p(\boldsymbol{x})$ be the probability function for a stochastic multigraph, and let $\{B_1,\ldots,B_t\}$ be a partition of the nodes into mutually exclusive and exhaustive subsets called node-blocks. We say that $p(\boldsymbol{x})$ is a stochastic blockmodel with respect to the partition $\{B_1,\ldots,B_t\}$ if and only if

(1) the random vectors $\boldsymbol{X}_{ij}$ are statistically independent; and

(2) for any nodes $i \neq j$ and $i' \neq j'$, if i and i' are in the same node-block and j and j' are in the same node-block, then the random vectors $\boldsymbol{X}_{ij}$ and $\boldsymbol{X}_{i'j'}$ are identically distributed.

Requirement 2 of Definition 3 implies that nodes in the same nodeblock are stochastically equivalent in the following sense. Consider a block B_r and any node j in the network. The likelihood of any given pattern of ties with node j is the same for all nodes in the block B_r. In other words, if i and i' are two nodes (excluding j) belonging to node-block B_r, any probability statement about $\boldsymbol{X}$ can be modified by interchanging $\boldsymbol{X}_{ij}$ and $\boldsymbol{X}_{i'j}$ or by interchanging $\boldsymbol{X}_{ji}$ and $\boldsymbol{X}_{ji'}$, without changing its probability.

We formalize this as

Definition 4. Let $\boldsymbol{X}$ be a random adjacency array. We say two nodes i and i' are stochastically equivalent if and only if the probability of any event about $\boldsymbol{X}$ is unchanged by interchanging nodes i and i'.

Definition 4 generalizes the algebraic notion of structural equivalence of nodes (see Lorrain and White 1971) to that of *stochastic equivalence.* Structurally equivalent nodes are stochastically equivalent but not *vice versa.*

We say that the pair of nodes (i,j) belongs to the *pair-block* $B_r \times B_s$ if and only if i is in the node-block B_r and j is in the node-block B_s. If $p(\boldsymbol{x})$ is the probability function for a stochastic blockmodel, $\boldsymbol{X}$, then the *pair-distribution* for pair-block $B_r \times B_s$ is given by

$$p_{rs}(\boldsymbol{z}) = \Pr(\boldsymbol{X}_{ij} = \boldsymbol{z}), \qquad \text{for any } i \in B_r, j \in B_s, i \neq j \tag{1}$$

and

$$\boldsymbol{z} = (z(1),\ldots,z(m)), \qquad z(k) = 0 \text{ or } 1.$$

Thus, $p_{rs}(z)$ is the probability that z describes the pattern of relations in the observed ties from a node in B_r to a node in B_s. Note that a deterministic blockmodel is a special case of the stochastic blockmodel, in which all the $p_{rs}(z)$ are equal to 0 or 1.

Definition 3 formalizes the concept of "internal homogeneity" (Breiger 1981) within the pair-blocks. According to our interpretation of the term, any pair-block $B_r \times B_s$ is *internally homogeneous;* that is, the ties are distributed without any apparent pattern among the pairs of individuals in $B_r \times B_s$. In other words, the rectangular submatrix of $\boldsymbol{X}$ consisting of the $X_{ij}(k)$ for which $i \in B_r$ and $j \in B_s$ shows no regular pattern. This property of the pair-blocks is implied by Definition 3 which requires that the distribution of relationships between any pair of nodes in a given pair-block is the same as that of any other pair of nodes in the same pair-block, and is independent of ties between any other pairs of nodes. This is the formal sense in which we use the term homogeneity.

The pair-distributions give the distributions of the entire vector $\boldsymbol{X}_{ij}$. For any $k \leqslant m$, and any two distinct nodes $i \in B_r$ and $j \in B_s$, the marginal distribution of $X_{ij}(k)$ is given by:

$$p_{rs(k)}(z) = \Pr\big(X_{ij}(k) = z\big) = \sum_{z(k)=z} p_{rs}(z(1),\ldots,z(k),\ldots,z(m))$$

$$\text{for } z = 0 \text{ or } 1. \tag{2}$$

We call the $\{p_{rs(k)}(z)\}$ the *marginal pair-distributions* of $\boldsymbol{X}(k)$.

We now present a theorem which is useful when working with the marginal distribution of $\boldsymbol{X}(k)$, or certain other stochastic multigraphs obtained from $\boldsymbol{X}$.

Theorem 1. Suppose $\boldsymbol{X} = (X_{ij}(k))$ is a stochastic blockmodel with node partition $(B_1,\ldots,B_t)$. Suppose further that $\boldsymbol{Y} = (Y_{ij}(k))$ is a multigraph obtained from $\boldsymbol{X}$ by $\boldsymbol{Y}_{ij} = f(\boldsymbol{X}_{ij})$, for $i \neq j$, for some function f. Then $\boldsymbol{Y}$ is a stochastic blockmodel with the same node partition $(B_1,\ldots,B_t)$.

Proof. If $\boldsymbol{Y}_{ij} = f(\boldsymbol{X}_{ij})$ and the $\boldsymbol{X}_{ij}$ are independent, then so are the $\boldsymbol{Y}_{ij}$. Further, if (i,j) and (i',j') belong to $B_r \times B_s$, then X_{ij} and $X_{i'j'}$ have the same distribution. This implies that $\boldsymbol{Y}_{ij} = f(\boldsymbol{X}_{ij})$ and $Y_{i'j'} = f(X_{i'j'})$ are also identically distributed.

Several corollaries follow immediately from Theorem 1.

Corollary 1. Suppose $\boldsymbol{X}$ is a stochastic blockmodel with blocks $B_1,\ldots,$ B_t. Let $\{R(k_1),\ldots,R(k_r)\}$ be a subset of $\{R(1),\ldots,R(m)\}$. Then the submatrix $\boldsymbol{X}^* = (X(k_1),\ldots,X(k_r))$ consisting of the multigraph of these r relations is also a stochastic blockmodel with the same node partition $B_1,\ldots,B_t$.

In particular, the adjacency matrix $\boldsymbol{X}(k)$ for the single relation $R(k)$ is a stochastic blockmodel if $\boldsymbol{X}$ is. The next corollary applies to aggregating sociometric relations.

Corollary 2. Suppose $\boldsymbol{X}$ is a stochastic blockmodel with blocks $B_1,\ldots,$ B_t. Let the adjacency matrix $\boldsymbol{Y}$ be given by

$$Y_{ij} = \begin{cases} 1 & \text{if } X_{ij}(k) = 1 \text{ for some } k, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Then $\boldsymbol{Y}$ is also a stochastic blockmodel with the node partition $B_1,\ldots,B_t$.

Corollary 2 implies, for example, that aggregating a series of measures of affect into a single affect relation preserves a stochastic blockmodel.

Estimation: A single relation, a priori *blocks*

For the remainder of this section, we consider the special case of a single sociometric relation with blocks specified *a priori.* We discuss this case in detail because it arises frequently in substantively interesting problems, and maximum likelihood estimation is completely trivial for this case.

In many cases of interest, we have reasonable hypotheses for the makeup of the blocks. Many variables (for example, sex and race) are natural blocking criteria in many common applications. Sometimes data are available only for a single relation. When data exist for multiple relations, it is often the case that the adjacency matrices are so highly correlated that not much information is lost by aggregation. Recall that Theorem 1 implies that aggregating relational data generated by a stochastic blockmodel produces another stochastic blockmodel.

Let $\boldsymbol{X} = \boldsymbol{X}(1)$ be the adjacency matrix for the stochastic blockmodel with a single relation. Let $p_{rs}(z) = p_{rs(1)}(z)$ denote the pair-distributions. In this case each pair-distribution is characterized by a *block* density π_{rs}, given by

$$\pi_{rs} = p_{rs(1)}(1) = \Pr(X_{ij} = 1) \text{ for } i \in B_r, j \in B_s, i \neq j. \tag{4}$$

The block density is the probability of a tie between any two distinct nodes in pair-block $B_r \times B_s$.

The implied distribution of the matrix $\boldsymbol{X}$ is

$$p(\boldsymbol{x}) = \Pr(\boldsymbol{X} = \boldsymbol{x}) = \prod_{r,s} \prod_{i \in B_r} \prod_{\substack{j \in B_s \\ j \neq i}} \pi_{rs}^{x_{ij}} (1 - \pi_{rs})^{(1 - x_{ij})}. \tag{5}$$

Let b_r be the number of nodes in node-block B_r. Then the number of pairs in pair-block $B_r \times B_s$ is given by

$$b_{rs} = \begin{cases} b_r(b_r - 1) & r = s \\ b_r b_s & r \neq s \end{cases}. \tag{6}$$

Let $X_{++}(r, s)$ be the number of 1s in the $b_r \times b_s$ submatrix of $\boldsymbol{X}$ corresponding to pair-block $B_r \times B_s$. That is,

$$X_{++}(r, s) = \sum_{i \in B_r} \sum_{\substack{j \in B_s \\ j \neq i}} X_{ij}. \tag{7}$$

Then (5) can be rewritten as

$$p(\boldsymbol{x}) = \Pr(\boldsymbol{X} = \boldsymbol{x}) = \prod_{r,s} \pi_{rs}^{x_{++}(r,s)} (1 - \pi_{rs})^{(b_{rs} - x_{++}(r,s))}. \tag{8}$$

From Definition 3 it follows that the X_{ij} are independent Bernoulli random variables, where the Bernoulli parameter is π_{rs} if $(i, j) \in B_r \times B_s$. Therefore, the $X_{++}(r, s)$ are independent binomial (b_{rs}, π_{rs}) random variables, and the marginal distribution of $X_{++}(r, s)$ is given by

$$\Pr(X_{++}(r, s) = k) = \binom{b_{rs}}{k} (\pi_{rs})^k (1 - \pi_{rs})^{(b_{rs} - k)}. \tag{9}$$

Because the $X_{++}(r, s)$ are independent binomial random variables, maximum likelihood estimation is particularly tractable for a stochastic blockmodel. Let $\boldsymbol{X}$ be the observed adjacency matrix arising from the stochastic blockmodel (5), where the blocks are given but $\boldsymbol{\pi} = (\pi_{rs})$ is unknown. Then the likelihood function for the data is

$$L(\boldsymbol{\pi}) = \prod_{r,s} \pi_{rs}^{x_{++}(r,s)} (1 - \pi_{rs})^{(b_{rs} - x_{++}(r,s))}. \tag{10}$$

Because the $X_{++}(r, s)$ are independent and there are no constraints on the π_{rs}, the maximum likelihood estimate for $\boldsymbol{\pi}$ is obtained by maximizing each of the marginal likelihood functions

$$L_{rs}(\pi_{rs}) = \pi_{rs}^{x_{++}(r,s)} (1 - \pi_{rs})^{(b_{rs} - x_{++}(r,s))}, \tag{11}$$

with respect to π_{rs}. The maximum occurs at the sample proportions

$$\hat{\pi}_{rs} = X_{++}(r, s)/b_{rs}. \tag{12}$$

Thus, the maximum likelihood estimate $\hat{\boldsymbol{\pi}}$ of the matrix of block densities $\boldsymbol{\pi}$ is given by the matrix of observed block densities.

Given $\hat{\boldsymbol{\pi}}$, we can obtain the maximum likelihood estimate of the expected value of any function of $\boldsymbol{X}$. For instance, the number of mutual, or reciprocated, ties in pair-block $B_r \times B_s$, denoted by $M(r, s)$, is a binomial random variable, since

$$M(r, s) = \begin{cases} \displaystyle\sum_{i \in B_r} \sum_{\substack{j \in B_s \\ i<j}} X_{ij} X_{ji} & r = s, \\ \displaystyle\sum_{i \in B_r} \sum_{j \in B_s} X_{ij} X_{ji} & r \neq s, \end{cases} \tag{13}$$

and the summands $X_{ij} X_{ji}$ are independent Bernoulli variables with parameter $\pi_{rs}\pi_{sr}$. Therefore, the maximum likelihood estimates of the mean and variance of $M(r, s)$ are given by

$$\overline{M}(r, s) = E_{\hat{\boldsymbol{\pi}}}[M(r, s)] = \begin{cases} \frac{1}{2} b_{rr} \hat{\pi}_{rr}^2 & r = s, \\ b_{rs} \hat{\pi}_{rs} \hat{\pi}_{sr} & r \neq s, \end{cases} \tag{14}$$

$$\overline{V}_M(r, s) = \mathrm{Var}_{\hat{\boldsymbol{\pi}}}[M(r, s)] = \begin{cases} \frac{1}{2} b_{rr} \hat{\pi}_{rr}^2 (1 - \hat{\pi}_{rr}^2) & r = s, \\ b_{rs} \hat{\pi}_{rs} \hat{\pi}_{sr} (1 - \hat{\pi}_{rs} \hat{\pi}_{sr}) & r \neq s. \end{cases} \tag{15}$$

A standardized measure of tendency toward reciprocity can be obtained from (14) and (15). Let

$$\overline{M} = \sum_{r \le s} \overline{M}(r, s)$$

and

$$\overline{V} = \sum_{r \le s} \overline{V}(r, s).$$

and set

$$\delta_M = \frac{M - \overline{M}}{\sqrt{\overline{V}}}. \tag{16}$$

The statistic δ_M is a measure of the reciprocation in excess of that predicted by the stochastic blockmodel. The asymptotic distribution of δ_M is standard normal when the number of blocks is small and the number of nodes is large.

3. An extension: Pair-level structure

It is plausible to expect a tendency toward reciprocation of choices in some types of social relationships (e.g. friendship), and a tendency away from reciprocation in others (e.g. power). Indeed, Moreno (1934) observed that the number of mutual ties usually exceeded the number

i	X									X_{i+}	
1	0	0	0	1	1	1	1	1	1	6	
2	0	0	0	1	1	1	1	1	1	6	
3	0	0	0	1	1	1	1	1	1	6	$M = 18$
4	1	1	1	0	0	0	0	0	0	3	
5	1	1	1	0	0	0	0	0	0	3	
6	1	1	1	0	0	0	0	0	0	3	$E(M \mid X_{i+}) = 8.86$
7	1	1	1	0	0	0	0	0	0	3	
8	1	1	1	0	0	0	0	0	0	3	
9	1	1	1	0	0	0	0	0	0	3	

Fig. 1. Reciprocity explained by blockmodel.

expected "by chance." The number of mutual ties to be expected "by chance", however, depends on the null distribution from which the chance expectation is calculated. The null model used by Moreno was one in which all adjacency matrices with row sums (number of choices made, or *out-degrees*) equal to the given adjacency matrix were considered equally likely.

The following example demonstrates that an apparent tendency toward mutuality can be explained by a blockmodel. Consider the hypothetical blockmodel of Fig. 1, which consists of two subgroups, with members of each subgroup choosing only members of the other subgroup. If, for example, $\boldsymbol{X}$ were the graph of the antagonism relation, the blockmodel could be interpreted as describing two mutually antagonistic cliques.

Let M be the number of (unordered) pairs $\{i, j\}$ such that $X_{ij} = X_{ji} = 1$. In the example of Figure 1, $M = 18$. Because members of each group choose and are chosen only by members of the other group, all 18 mutual ties are explained by the blockmodel. However, the null model used by Moreno predicts only about 9 mutual ties. (The formula for calculating the expected number of mutual ties using Moreno's null distribution is given in Holland and Leinhardt (1981a).) In this example, an apparent tendency toward reciprocity is explained by the blockmodel.

If $\boldsymbol{X}$ is a stochastic blockmodel, then the observed ties are due only to the the pair-distributions $\{p_{rs}\}$ and random noise. There are no tendencies toward reciprocation of ties, transitivity, or any other type of structure, except those which can be explained by the block memberships. In the remainder of this section we develop a model which allows for a correlation between ties from node i to node j and ties from node j to node i, and simultaneously takes into account the block structure. The *pair-dependent stochastic blockmodel,* which we define below, is a generalization of the stochastic blockmodel developed in the last section. It accounts for pair-level structure such as tendencies toward reciprocation of ties or toward exchange of one type of tie for another. In the single relation case, a specialization of the pair-dependent stochastic blockmodel provides a one degree-of-freedom test of the stochastic blockmodel for tendencies toward reciprocity.

The model proposed in this section is based on the $2m$-element random vectors $\boldsymbol{D}_{ij} = (\boldsymbol{X}_{ij}, \boldsymbol{X}_{ji})$, which we call *dyad vectors.* We now define the pair-dependent stochastic blockmodel (abbreviated PSB).

Definition 5. Let $p(\boldsymbol{x})$ be the probability function for a stochastic multigraph, and let $\{B_1,\ldots,B_t\}$ be a partition of the nodes of the stochastic multigraph into mutually exclusive and exhaustive node-blocks. We say that $p(\boldsymbol{x})$ satisfies a pair-dependent stochastic blockmodel with respect to the partition $\{B_1,\ldots,B_L\}$ if and only if:

(1) the random vectors $\boldsymbol{D}_{ij}$ are statistically independent, and

(2) for any nodes $i \neq j$ and $i' \neq j'$, if (i,j) and (i',j') belong to the same pair-block, then the random vectors $\boldsymbol{D}_{ij}$ and $\boldsymbol{D}_{i'j'}$ are identically distributed.

The stochastic blockmodel is a special case of the PSB in which the dyad vectors $\boldsymbol{D}_{ij}$ are made up of two statistically independent random vectors $\boldsymbol{X}_{ij}$ and $\boldsymbol{X}_{ji}$.

The pair-distributions for the PSB are defined over the dyad vectors. The pair-distribution for pair-block $B_r \times B_s$ is given by:

$$P_{rs}(z_1, z_2) = \Pr\big(\boldsymbol{D}_{ij} = (z_1, z_2)\big), \qquad \text{for all } i \in B_r, j \in B_s, i \neq j \tag{17}$$

$$z_i = (z_i(1),\ldots,z_i(m)), \qquad z_i(k) = 0 \text{ or } 1,\ i = 1, 2.$$

The marginal distribution of $D_{ij}(k) = (X_{ij}(k), X_{ji}(k))$ is given by:

$$P_{rs(k)}\big(D_{ij}(k)\big) = \Pr\big(D_{ij}(k) = (z_1, z_2)\big) = \sum_{z_i(k) = z_i} p_{rs}(z_1, z_2)$$

$$\text{for } z_i = 0 \textit{ or } 1. \tag{18}$$

Theorem 2 is an analogue of Theorem 1 for the pair-dependent stochastic blockmodel.

Theorem 2. Suppose $\boldsymbol{X} = (X_{ij}(k))$ is a pair-dependent stochastic blockmodel with node partition $(B_1,\ldots,B_t)$. Suppose that $\boldsymbol{Y} = (Y_{ij}(k))$ is another stochastic multigraph. Suppose further that there exists a function f such that $(\boldsymbol{Y}_{ij}, \boldsymbol{Y}_{ji}) = f(\boldsymbol{X}_{ij}, \boldsymbol{X}_{ji})$, for all $i \neq j$. Then $\boldsymbol{Y}$ is a pair-dependent stochastic blockmodel with the same node partition $(B_1,\ldots,B_t)$.

The proof of Theorem 2 is analogous to the proof of Theorem 1. We now state several obvious corollaries to Theorem 2. The first two are direct analogues to Corollaries 1 and 2.

Corollary 3. Suppose $\boldsymbol{X}$ is a random adjacency matrix for the m relations $R(1),\ldots,R(m)$, and suppose $\boldsymbol{X}$ is a PSB with blocks $B_1,\ldots,B_t$. Let $\{R(k_1),\ldots,R(k_r)\}$ be a subset of $\{R(1),\ldots,R(m)\}$. Then the subarray $\boldsymbol{X}^* = (X(k_1),\ldots,X(k_r))$ consisting of the multigraph of these r relations is also a PSB with the same node partition $B_1,\ldots,B_t$.

Corollary 4. Suppose $\boldsymbol{X}$ is a random adjacency matrix for the m relations $R(1),\ldots,R(m)$, and suppose $\boldsymbol{X}$ is generated by a PSB with blocks $B_1,\ldots,B_t$. Let the random adjacency matrix $\boldsymbol{Y}$ be given by

$$Y_{ij} = \begin{cases} 1 & \text{if } X_{ij}(k) = 1 \text{ for some } k, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Then $\boldsymbol{Y}$ is also a PSB with the node partition $B_1,\ldots,B_t$.

Therefore, a PSB is preserved by taking subsets of the set of m relations and by aggregating relations. Furthermore, breaking a PSB up into *mutual, asymmetric* and *null* multigraphs also produces a PSB, as stated in the next corollary.

Corollary 5. Suppose $\boldsymbol{X}$ is a random adjacency matrix for a single relation and suppose $\boldsymbol{X}$ is a PSB with blocks $B_1,\ldots,B_t$. Let the random adjacency matrices $\boldsymbol{Y}^{(m)}$, $\boldsymbol{Y}^{(a)}$ and $\boldsymbol{Y}^{(n)}$ be given by

$$Y_{ij}^{(m)} = X_{ij}X_{ji}, \tag{20}$$

$$Y_{ij}^{(a)} = X_{ij}(1 - X_{ji}), \tag{21}$$

and

$$Y_{ij}^{(n)} = (1 - X_{ij})(1 - X_{ji}). \tag{22}$$

Then $\boldsymbol{Y} = (\boldsymbol{Y}^{(m)}, \boldsymbol{Y}^{(a)}, \boldsymbol{Y}^{(n)})$ is a PSB with the node partition $B_1,\ldots,B_t$.

3.1. A special case: Stochastic blockmodel with reciprocity

In this section, we develop a special case of the PSB, which we call the *stochastic blockmodel with reciprocity,* or SBR. Let $\boldsymbol{X}$ be a PSB with one relation and denote its probability function by $p(\boldsymbol{x})$. Using the

notation of Holland and Leinhardt (1981a), the pair-distributions $p_{rs}(z_1, z_2)$ and $p_{sr}(z_1, z_2)$ are characterized by specifying the four quantities:

$$m_{rs} = p_{rs}(1, 1) = p_{sr}(1, 1), \tag{23}$$

$$a_{rs} = p_{rs}(1, 0) = p_{sr}(0, 1), \tag{24}$$

$$a_{sr} = p_{rs}(0, 1) = p_{sr}(1, 0), \tag{25}$$

and

$$n_{rs} = p_{rs}(0, 0) = p_{sr}(0, 0), \tag{26}$$

such that

$$m_{rs} + a_{rs} + a_{sr} + n_{rs} = 1. \tag{27}$$

For a dyad vector in the pair-block $B_r \times B_s$, m_{rs} is the probability of a mutual pair, a_{rs} and a_{sr} are the probabilities of the two types of asymmetric pair, and n_{rs} is the probability of a null pair. To avoid degenerate cases we assume that m_{rs}, a_{rs}, a_{sr} and n_{rs} are positive for all r and s. The distribution of the matrix X is given by

$$p(\boldsymbol{x}) = \prod_{r,s} \prod_{i \in B_r} \prod_{\substack{j \in B_s \\ i<j}} m_{rs}^{x_{ij}x_{ji}} a_{rs}^{x_{ij}(1-x_{ji})} a_{sr}^{(1-x_{ij})x_{ji}} n_{rs}^{(1-x_{ij})(1-x_{ji})}. \tag{28}$$

The single-relation, pair-dependent, stochastic blockmodel is a special case of the general exponential family model for random directed graphs with pair-level correlations described in Holland and Leinhardt (1981a: 36). Fienberg and Wasserman (1980) also discuss the distribution (28) for grouped relational data.

Equation (28) can be re-expressed in terms of the natural parameters of the exponential family model. This re-expression facilitates comparison with work appearing elsewhere (Holland and Leinhardt 1981a; Fienberg and Wasserman 1980, 1981). Further, the stochastic blockmodel with reciprocity, a special case of (28) with one parameter more than the stochastic blockmodel, is most conveniently defined in terms of the natural parameters of (28).

Define

$$\theta_{rs} = \log\left(\frac{a_{rs}}{n_{rs}}\right) \qquad \text{for all } r, s \tag{29}$$

$$\rho_{rs} = \log\left(\frac{m_{rs}n_{rs}}{a_{rs}a_{sr}}\right) \qquad \text{for all } r, s, \tag{30}$$

and let $M(r, s)$ be the number of mutual dyads in pair-block $B_r \times B_s$ as defined in (13). With some algebraic manipulation, (28) can be written as follows.

$$p(\boldsymbol{x}) = \exp\left\{\sum_{r \leqslant s} \rho_{rs} M(r, s) + \sum_{r,s} \theta_{rs} X_{++}(r, s)\right\} K(\{\rho_{rs}, \theta_{rs}\}). \tag{31}$$

It can also be shown that the stochastic blockmodel of Section 2 with block densities π_{rs} is a special case of (31) with $\rho_{rs} = 0$ for all r, s and

$$\theta_{rs} = \log\left(\frac{\pi_{rs}}{1 - \pi_{rs}}\right). \tag{32}$$

Thus, if $p^*(\boldsymbol{x})$ is the probability distribution of a stochastic blockmodel, then p^* satisfies (8), which can be written in the form

$$p^*(\boldsymbol{x}) = \exp\left\{\sum_{r,s} \theta_{rs} X_{++}(r, s)\right\} K(\{\theta_{rs}\}). \tag{33}$$

The parameter ρ_{rs} measures the tendency of ties between pair-block B_r and pair-block B_s to be reciprocated. Holland and Leinhardt (1981a: 36) show that when ρ_{rs} is positive, if $X_{ij} = 1$ for $(i, j) \in B_r \times B_s$ then the likelihood of observing $X_{ji} = 1$ is increased. When $\rho_{rs} = 0$, the parameter θ_{rs} governs the overall density of pair-block $B_r \times B_s$.

We are now ready to define the stochastic blockmodel with reciprocity. This submodel of the single relation pair-dependent stochastic blockmodel is obtained by restricting the parameters ρ_{rs} so that

$$\rho_{rs} = \rho \qquad \text{for all } r, s. \tag{34}$$

Definition 6. Let $p(\boldsymbol{x})$ be the probability function for a single-relation pair-dependent stochastic blockmodel with blocks $\{B_1, \ldots, B_t\}$. We say

$p(\boldsymbol{x})$ is a stochastic blockmodel with reciprocity if there exist parameters ρ and $\{\theta_{rs}\}$ such that

$$p(\boldsymbol{x}) = \exp\left\{ \rho M + \sum_{r,s} \theta_{rs} X_{++}(r, s) \right\} K(\{\rho, \theta_{rs}\}), \tag{35}$$

where M is the total number of mutual pairs in the graph, given by

$$M = \sum_{r \leqslant s} M(r, s). \tag{36}$$

Fienberg and Wasserman (1980) suggest restricting the $\{\theta_{rs}\}$ so that they satisfy the additive model

$$\theta_{rs} = \theta + \alpha_r + \beta_s \qquad \text{for all } r, s. \tag{37}$$

With the restriction (37), the stochastic blockmodel with reciprocity is a special case of the p_1 distribution proposed by Holland and Leinhardt (1981a). Holland and Leinhardt interpret the parameters $\{\theta, \alpha_r, \beta_s\}$ as follows. The parameter θ is a measurement of overall choice density, the parameter α_r is a measurement of the "productivity" of nodes in node-block r, and β_s is a measurement of the "attractiveness" of nodes in node-block s. In our view, this additive decomposition is unacceptable for modeling grouped data in many social contexts. Cliquing is one social phenomenon which cannot be explained unless an interaction term is allowed in Eq. (37). Cliques are characterized by high density diagonal pair-blocks and low density off-diagonal pair-blocks. When cliques are present, θ_{rs} is larger than would be predicted by an additive model. We prefer leaving the $\{\theta_{rs}\}$ unconstrained.

3.2. Estimation: Stochastic blockmodel with reciprocity

Let $p(\boldsymbol{x})$ be a stochastic blockmodel with reciprocity. Since $p(\boldsymbol{x})$ is an exponential family probability function, the maximum likelihood estimates (MLEs) of the parameters ρ and $\{\theta_{rs}\}$ are the values for which the sufficient statistics of the distribution are equal to their expected values. The sufficient statistics of $p(\boldsymbol{x})$ are the number of mutuals M and the observed number of edges in each pair block $X_{++}(r, s)$. The likelihood equations which define the maximum likelihood estimates are given by equating the sufficient statistics to their expected values.

The expected value of M is given by

$$E_{(\rho,\theta)}(M) = \tfrac{1}{2}\sum_{r,s}\sum_{i\in B_r}\sum_{j\in B_s}\Pr(D_{ij} = (1,1)) = \tfrac{1}{2}\sum_{r,s} b_{rs}m_{rs}. \tag{38}$$

Recall that m_{rs} is the probability of a mutual tie for any pair in pair-block $B_r \times B_s$ and b_{rs} is the number of pairs in pair-block $B_r \times B_s$. The factor $\frac{1}{2}$ enters because each dyad is counted twice in (38). The expected value of $X_{++}(r, s)$ is given by

$$E_{(\rho,\theta)}(X_{++}(r,s)) = \sum_{i\in B_r}\sum_{j\in B_s}\Pr(X_{ij} = 1) = b_{rs}(m_{rs} + a_{rs}). \tag{39}$$

Note that $m_{rs} + a_{rs}$ is the probability that $X_{ij} = 1$ for $(i, j) \in B_r \times B_s$. Equations (38) and (39) imply that the likelihood equations for the data are given by

$$M = \tfrac{1}{2}\sum_{r,s} b_{rs}m_{rs}, \tag{40}$$

$$X_{++}(r,s) = b_{rs}(m_{rs} + a_{rs}) \qquad r, s = 1,\ldots,t. \tag{41}$$

The MLEs of m_{rs}, a_{rs}, a_{sr} and n_{rs} are the values $\hat{m}_{rs}$, $\hat{a}_{rs}$, $\hat{a}_{sr}$ and $\hat{n}_{rs}$ which solve Equations (40) and (41). These equations cannot be solved in closed form. Fienberg and Wasserman (1980, 1981) have shown that the maximum likelihood estimates for models of this type can be found using standard iterative proportional fitting routines (*e.g.*, BMDP3F) for fitting log-linear models. Fienberg and Wasserman's method depends on representing the directed graph $\boldsymbol{X}$ as a $t \times t \times 2 \times 2$ contingency table $\boldsymbol{Y}$, where

$$Y(r,s,0,0) = \sum_{i\in B_r}\sum_{j\in B_s}(1 - X_{rs})(1 - X_{sr}), \tag{42}$$

$$Y(r,s,0,1) = \sum_{i\in B_r}\sum_{j\in B_s}(1 - X_{rs})X_{sr}, \tag{43}$$

$$Y(r,s,1,0) = \sum_{i\in B_r}\sum_{j\in B_s}X_{rs}(1 - X_{sr}), \tag{44}$$

and

$$Y(r, s, 1, 1) = \sum_{i \in B_r} \sum_{j \in B_s} X_{rs} X_{sr}. \tag{45}$$

Certain marginal sums of $\boldsymbol{Y}$ correspond to sufficient statistics of the stochastic blockmodel with reciprocity. For example, by summing over the first two dimensions of the contingency table, one can calculate the number of mutual dyads in the observed adjacency matrix.

$$2M = Y(+, +, 1, 1) = \sum_{r,s} Y(r, s, 1, 1) = \sum_{r,s} \sum_{i \in B_r} \sum_{j \in B_s} X_{ij} X_{ji}. \tag{46}$$

The observed block densities can be calculated in two ways, by summing over dimension 3, or by summing over dimension 4.

$$\begin{aligned} X_{++}(r, s) &= Y(r, s, 1, +) = \sum_{k} Y(r, s, 1, k) \\ &= \sum_{i \in B_r} \sum_{j \in B_s} X_{ij} X_{ji} + \sum_{i \in B_r} \sum_{j \in B_s} X_{ij}(1 - X_{ji}), \end{aligned} \tag{47}$$

$$\begin{aligned} X_{++}(s, r) &= Y(r, s, +, 1) = \sum_{k} Y(r, s, k, 1) \\ &= \sum_{i \in B_r} \sum_{j \in B_s} X_{ij} X_{ji} + \sum_{i \in B_r} \sum_{j \in B_s} (1 - X_{ij}) X_{ji}. \end{aligned} \tag{48}$$

The maximum likelihood estimates $\hat{m}_{rs}$, $\hat{a}_{rs}$, $\hat{a}_{sr}$ and $\hat{n}_{rs}$ can be found by fitting the log-linear model which satisfies (46), (47) and (48). Using the notation of Fienberg (1980), Eq. (46) corresponds to fitting the [34] margin of Y, that is, fitting the marginal totals corresponding to fixing dimensions 3 and 4, and summing over the other dimensions. Similarly, Eq. (47) corresponds to fitting the [123] margin, and Eq. (48) corresponds to fitting the [124] margin.

3.3. Testing the fit of a stochastic blockmodel

The MLEs of the parameters of a stochastic blockmodel are obtained by applying the constraint $\rho = 0$ to the corresponding SBR and calculating the maximum likelihood estimates for the parameters of this

submodel. A likelihood statistic can be constructed which corresponds to testing the hypothesis

$$H_0\colon \rho = 0, \boldsymbol{\theta} \text{ unconstrained}, \tag{49}$$

against the alternative hypothesis

$$H_A\colon \rho, \boldsymbol{\theta} \text{ unconstrained}. \tag{50}$$

Let $\hat{m}_{rs}$, $\hat{a}_{rs}$, $\hat{a}_{sr}$ and $\hat{n}_{rs}$ denote the MLEs of the parameters of the SBR, and let $\hat{\pi}_{rs}$ denote the MLEs of the parameters of the stochastic blockmodel. The usual log-likelihood statistic for the test of H_0 against H_A is given by

$$LLR = L_m + L_a + L_n. \tag{51}$$

where

$$L_m = \sum_{r,s} Y(r, s, 1, 1) \log\left(\frac{\hat{m}_{rs}}{\hat{\pi}_{rs}\hat{\pi}_{sr}} \right), \tag{52}$$

$$L_a = 2\sum_{r,s} Y(r, s, 1, 0) \log\left(\frac{\hat{a}_{rs}}{\hat{\pi}_{rs}(1 - \hat{\pi}_{sr})} \right), \tag{53}$$

$$L_n = \sum_{r,s} Y(r, s, 0, 0) \log\left(\frac{\hat{n}_{rs}}{(1 - \hat{\pi}_{rs})(1 - \hat{\pi}_{sr})} \right). \tag{54}$$

The duplication of entries in $Y(r, s, 1, 1)$ and $Y(r, s, 0, 0)$ means that no factor of 2 is needed in Equations (52) and (54). The LLR statistic is asymptotically distributed as chi-square with one degree of freedom.

Our use of this test is to ascertain how well the blockmodel fits the data. If H_0 is accepted, the conclusion is that the blockmodel does explain the observed degree of reciprocity in the data. If H_0 is rejected we may conclude that there is reciprocity in the data that is not explained by the blockmodel. Section 4 illustrates this point.

4. An empirical example: Sampson's monastery data

As an example of the application of stochastic blockmodeling methods we present results of an analysis of data which have been widely

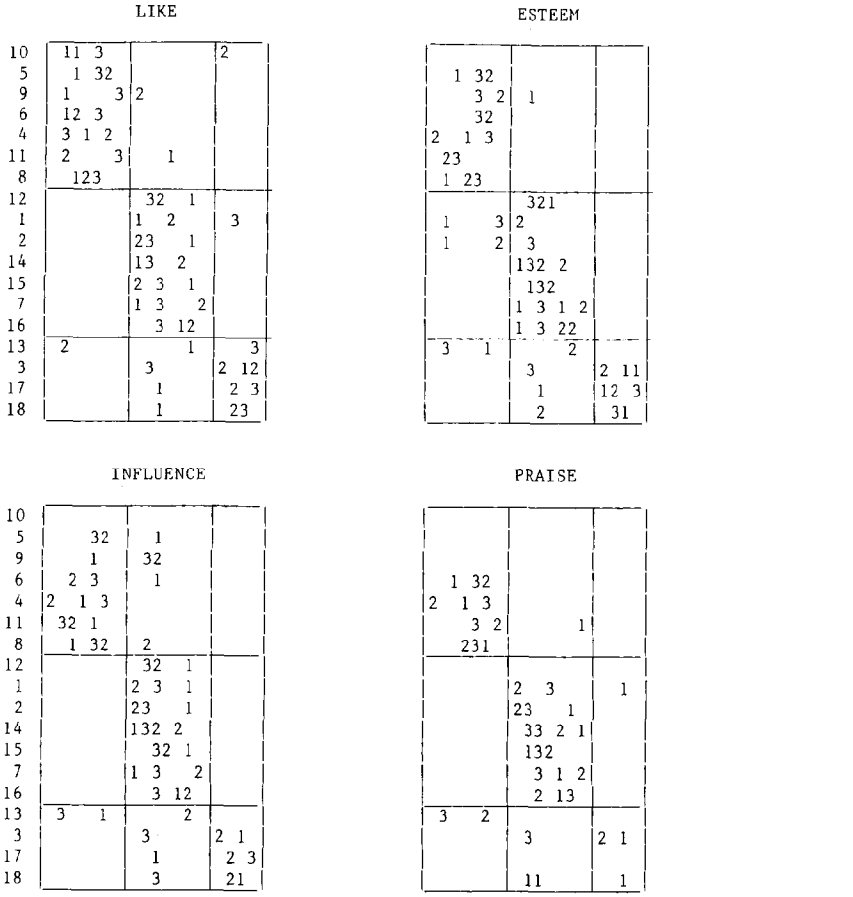

Fig. 2. Blockmodel for Sampson data: Positive choice data matrices. Taken from White *et al.* (1976: 750). Strong ties are coded as 3, weak ties as 1. Blanks indicate no ties. (Source: Sampson 1969).

used as an example of the application of blockmodel analysis to sociometric data. The data were originally collected by Sampson (1969), who spent a 12-month period observing and collecting data on social relations between the monks of an American monastery. A major social conflict erupted near the end of the study, resulting in the expulsion or resignation of many of the monastery's members. The first reported analysis of these data using blockmodeling methods appeared in White,

ANTAGONISM

10			
5			321
9		1 2	32
6		231	
4		13	2
11		31	2
8			
12	1	2 3	
1	3 2		1
2	3 2		1
14	1		32
15	23 1		
7	3		12
16	2	1 3	
13	23	2 1	
3	31 2		
17	2 1 3		
18			

DISESTEEM

			132
			1 32
	1	3	22
	231		
	13 1		2
	1		23
	2 3		1
3			12
22			3 11
31			211
32 2			21
3			221
3	2		1
23	1		
32 1			
1 22 3			

NEGINFL

10			
5			132
9	1		32
6		3	21
4	231		
11		3	12
8	1	3	2
12			
1	2	1	3
2	3		2
14	2		3 11
15	3		21
7	32 2		21
16	32		1
13	2	1 3	
3	23	1	
17	32 1		
18	132		

BLAME

		3	1 21
	32 1		
	1		23
			123
	2		13
2			3 1
			3
32			1
1			32 1
12			3
3			1222
	1		23
3	21		

Fig. 3. Blockmodel for Sampson data: Negative choice data matrices. Taken from White *et al.* (1976: 750). Strong ties are coded as 3, weak ties as 1. Blanks indicate no ties. (Source: Sampson 1969).

Boorman and Breiger (1976). The data were partially reanalyzed by Holland and Leinhardt (1981a) using the p_1 model. (See also the discussion in Breiger (1981) and in Holland and Leinhardt (1981b)).

Sampson's data include responses to four sociometric criteria: affect, esteem, influence, and sanction. Both positive and negative interpersonal choices on each of the four criteria were obtained. The sociometric data are shown in Figures 2 and 3 as they were presented in White,

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	X_{i+}
1		1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	4
2	0		*M*	0	*M*	*M*	0	0	0	0	0	0	0	0	0	0	0	0	3
3	0	*M*		0	0	0	*M*	1	0	0	0	0	0	0	0	0	0	0	3
4	0	1	1		*M*	0	0	0	0	0	0	0	0	0	0	0	0	0	3
5	0	*M*	0	*M*		1	0	0	0	0	0	0	0	0	0	0	0	0	3
6	0	*M*	0	0	0		1	0	0	0	1	0	0	0	0	0	0	0	3
7	0	0	*M*	1	1	0		0	0	0	0	0	0	0	0	0	0	0	3
8	0	0	0	0	0	0	0		*M*	*M*	0	0	*M*	0	0	0	0	0	3
9	0	0	0	0	0	0	0	*M*		0	*M*	0	0	0	0	*M*	0	0	3
10	0	0	0	0	0	0	0	*M*	1		0	0	*M*	0	0	0	0	0	3
11	0	0	0	0	0	0	0	1	*M*	0		1	0	0	0	0	0	0	3
12	0	0	0	0	0	0	0	1	0	1	0		1	0	0	0	0	0	3
13	0	0	0	0	0	0	0	*M*	0	*M*	0	0		*M*	0	0	0	0	3
14	0	0	0	0	0	0	0	0	0	1	0	1	*M*		0	0	0	0	3
15	0	1	0	0	0	0	0	0	0	0	0	0	1	0		0	0	1	3
16	0	0	0	0	0	0	0	0	*M*	0	0	0	0	0	1		*M*	*M*	4
17	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	*M*		*M*	3
18	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	*M*	*M*		3
X_{+j}	0	6	4	2	4	2	2	6	4	6	2	2	5	1	2	3	2	3	$56 = X_{++}$

Figure 4. Adjacency matrix from Sampson (1969) as presented in White, Boorman and Breiger (1976). Dashed lines indicate high and low tie-density blocks found by White *et al.* *M* denotes a reciprocated tie while 1 denotes a tie that is not reciprocated.

Boorman and Breiger (1976). The data are disaggregated into positive (Figure 2) and negative (Figure 3) responses on each of the four criteria. The sociomatrices are arranged in the blocks which White, Boorman and Breiger deduced using the CONCOR algorithm (White *et al.* 1976; Schwartz 1977).

For the purposes of this example we focused solely on the data generated by the monks' positive responses to the "like" criterion, the matrix in the upper left hand corner of Figure 2. We modified these data by dropping the strength distinction so that an adjacency matrix was created in which a one was entered if there was either a one, two or three in the original data matrix; otherwise the entry was zero. The resulting data matrix appears in Figure 4 where we have, for purposes of clarity in the display, placed an M in the i, j entry if the i, j and j, i entries were both one. Figure 4 also shows marginal sums.

White, Boorman and Breiger (1976) have proposed a block structure for Sampson's data. We use techniques developed in Section 2 to evaluate the fit of their model. The relevant computations are simple

	1	2	3
1	$\pi_{11} = 0.452,\ \frac{1}{2}b_{11} = 21$ $\pi_{11}^2 = 0.2043$ $1 - \pi_{11}^2 = 0.7957$ $\frac{1}{2}b_{11}\pi_{11}^2 = 4.2903$ $\frac{1}{2}b_{11}\pi_{11}^2(1 - \pi_{11}^2) = 3.4138$	$\pi_{12} = 0.041,\ b_{12} = 49$ $\pi_{12}\pi_{21} = 0$ $1 - \pi_{12}\pi_{21} = 1.000$ $b_{12}\pi_{12}\pi_{21} = 0$ $b_{12}\pi_{12}\pi_{21}(1 - \pi_{12}\pi_{21}) = 0$	$\pi_{13} = 0.036,\ b_{13} = 28$ $\pi_{13}\pi_{31} = 0.0013$ $1 - \pi_{13}\pi_{31} = 0.9987$ $b_{13}\pi_{13}\pi_{31} = 0.0364$ $b_{13}\pi_{13}\pi_{31}(1 - \pi_{13}\pi_{31}) = 0.0364$
2	$\pi_{21} = 0.000,\ b_{21} = 49$ $\pi_{21}\pi_{12} = 0.000$ $1 - \pi_{21}\pi_{12} = 1.000$ $b_{21}\pi_{21}\pi_{12} = 0.000$ $b_{21}\pi_{21}\pi_{12}(1 - \pi_{21}\pi_{12}) = 0$	$\pi_{22} = 0.476,\ \frac{1}{2}b_{22} = 21$ $\pi_{22}^2 = 0.2266$ $1 - \pi_{22}^2 = 0.7734$ $\frac{1}{2}b_{22}\pi_{22}^2 = 4.7586$ $\frac{1}{2}b_{22}\pi_{22}^2(1 - \pi_{22}^2) = 3.6803$	$\pi_{23} = 0.036,\ b_{23} = 28$ $\pi_{23}\pi_{32} = 0.0051$ $1 - \pi_{23}\pi_{32} = 0.9949$ $b_{23}\pi_{23}\pi_{32} = 0.1428$ $b_{23}\pi_{23}\pi_{32}(1 - \pi_{23}\pi_{32}) = 0.1421$
3	$\pi_{31} = 0.036,\ b_{31} = 28$ $\pi_{31}\pi_{13} = 0.0013$ $1 - \pi_{31}\pi_{13} = 0.9987$ $b_{31}\pi_{31}\pi_{13} = 0.0364$ $b_{31}\pi_{31}\pi_{13}(1 - \pi_{31}\pi_{13}) = 0.0364$	$\pi_{32} = 0.143,\ b_{32} = 28$ $\pi_{32}\pi_{23} = 0.0051$ $1 - \pi_{32}\pi_{23} = 0.9949$ $b_{32}\pi_{32}\pi_{23} = 0.1428$ $b_{32}\pi_{32}\pi_{23}(1 - \pi_{32}\pi_{23}) = 0.1421$	$\pi_{33} = 0.667\ \frac{1}{2}b_{33} = 6$ $\pi_{33}^2 = 0.4449$ $1 - \pi_{33}^2 = 0.5551$ $\frac{1}{2}b_{33}\pi_{33}^2 = 2.6694$ $\frac{1}{2}b_{33}\pi_{33}^2(1 - \pi_{33}^2) = 1.4818$

Figure 5. Worksheet of calculations to obtain δ_M for Sampson's liking data. $\overline{M} = 11.8975$, $\overline{V} = 8.7544$, $\sqrt{\overline{V}} = 2.959$, $\delta_M = (15 - 11.8975)/2.959 = 1.048$.

and straightforward and are specified in Equations (14), (15) and (16). We use the maximum likelihood estimates of the block densities, $\hat{\pi}_{rs}$, and the number of pairs in each block to obtain maximum likelihood estimates of the mean and variance of the number of mutuals given the blocking. These are used to calculate a statistic, δ_M (Equation (16)) which, when referred to a table of the standard normal, allows us to test the blockmodel's ability to explain the reciprocity in the data. A sufficiently large value of δ_M would lead us to reject the proposed blockmodel.

Figure 5 consists of a worksheet of the computations necessary to obtain δ_M for Sampson's liking data given the blockmodel proposed in White, Boorman and Breiger (1976). The obtained value of δ_M is 1.048 from which we infer that rejection of the blockmodel is not warranted.

A comparison of this analysis with the one presented in Holland and Leinhardt (1981a) points up the ability of the blockmodel to explain an apparent tendency toward reciprocation. Holland and Leinhardt analyzed these data using the p_1 distribution. The p_1 model accounts for reciprocation of ties but not for block structure. The reciprocation parameters for the p_1 model and the SBR (both denoted by ρ) are log-odds ratios and can be directly compared.

When the p_1 model is fit to the Sampson data, an estimate $\hat{\rho} = 3.14$ is obtained. [1] The substantive interpretation of this result is that the odds of observing $X_{ij} = 1$ are about 23 times higher [2] when $X_{ji} = 1$ than when $X_{ji} = 0$. The likelihood ratio statistic for testing $\rho \neq 0$ is 30.41, a value much larger than the conventional significance levels of the chi-square distribution on 1 degree of freedom. In contrast, a fit of the SBR to these data yields the estimate $\hat{\rho} = 1.01$. This corresponds to less than a three-fold increase in the odds that $X_{ij} = 1$ due to $X_{ji} = 1$. Further, the likelihood ratio statistic for testing $\rho \neq 0$ is only 3.13. This value is less than the usual significance levels for the chi-square distribution. Thus, the observed degree of reciprocation is consistent with the hypothesis that a stochastic blockmodel generated the data.

These results indicate that the blockmodel, in effect, "explains" the apparent mutuality of choice in the data. Except for one pair, the mutual choices occur within the diagonal blocks where choice density is

[1] The value of $\hat{\rho} = 3.10$ given on page 45 of Holland and Leinhardt (1981a) is in error. The value given in Table 7 on page 47 is correct.

[2] The parameter ρ is the log of the increase in odds. The odds increase is given by $e^{3.14} = 23$.

expected to be high and so reciprocation is expected. The one mutual pair (9, 16) is an important exception to the fit as is the column of zeros at $j = 1$. Further investigation of these exceptions would seem warranted in light of our results.

It is important to note that we have chosen this example for its pedagogic value. Sampson's monastery data are by now well known in the research literature on social networks. Reader familiarity with these data plus the moderate size and clear cut blocking of the data facilitate following the analysis. Of course, with other data, one might obtain a value of δ_M which is significant, leading to rejection of the stochastic blockmodel. In such a case the analysis would indicate that there is more reciprocation than the stochastic blockmodel can explain. There are several explanations for such an outcome. First, the blockmodel may contain too few blocks. Second, a block structure may be inappropriate to describe the data. Third, a more complicated generalization of the blockmodel, such as the pair-dependent stochastic blockmodels discussed in Section 3.2, may be needed. If the third explanation seems most appropriate then the equations appearing in Section 3 can be used to compute an estimate for the ρ parameter, yielding a quantitative measure of the tendency toward reciprocity above and beyond that explained by the blocking. The log-likelihood test described in Section 3.3 then provides a test of this more complicated model.

5. Discussion

There are two important topics that we will comment on briefly in this section. They are both concerned with the current uses of nonstochastic blockmodels.

5.1. The problem of closure

In the algebraic theories that are used to give a rationale for blockmodeling techniques (*i.e.*, Lorrain and White 1971) an important role is played by the semi-group of relations generated by the m relations $R_{(1)}, \ldots, R_{(m)}$. It is natural to ask what role this semi-group might play in our theory of stochastic blockmodels. Unfortunately, there is no simple or direct connection between stochastic blockmodels and the semi-group of relations. One symptom of the problems that

occur in trying to make this connection is the fact that a stochastic blockmodel is not closed under the operation of forming the binary product of two adjacency matrices. Theorems 1 and 2 show that stochastic blockmodels are closed under various natural operations. The binary product of $X(1)$ and $X(2)$ is defined as

$$(X(1) * X(2))_{ij} = \begin{cases} 1 & \text{if } X_{ik}(1) * X_{kj}(2) = 1 \text{ for some } k, \\ 0 & \text{otherwise.} \end{cases}$$

This product is used to define indirect relationships between nodes. Indirect relationships from node i to node j go through one or more intermediaries, . However, it is easy to see that even if $X(1)$ and $X(2)$ have the joint distribution of a stochastic blockmodel (with or without reciprocation) the product $X(1) * X(2)$ can fail to be a stochastic blockmodel. It fails because of correlation that is introduced between the entries $(X(1) * X(2))_{ij}$ and $(X(1) * X(2))_{ij'}$ for $j \neq j'$. A formal proof is left to the reader, but a heuristic explanation is given below. The condition that elements of $X(1) * X(2)$ in the same pair-block have the same distribution is not violated by the binary product. Thus the essential conflict between binary products and stochastic blockmodels is focused on the notions of indirect ties and of randomness within pair-blocks. The product relation describes indirect ties between individuals. A link between i and j occurs in the product relation if there is a third individual k such that i is related to k in the first relation and k is related to j in the second relation. Both $(X(1) * X(2))_{ij}$ and $(X(1) * X(2))_{ij'}$ are more likely to be 1 when $X_{ik}(1) = 1$. Thus, the two entries cannot be independent. Since randomness within pair-blocks is the vehicle that we have used to formalize the intuitive idea that a blockmodel "explains" a set of data we believe that this conflict is a basic one that requires further work.

5.2. Generating the blocks: A Bayesian problem

In the development that we have given here the blocks $B_1, \ldots, B_t$ have been assumed to be specified *a priori*. This need not be the case and, in fact, the most exciting uses of blockmodel techniques have come from the discovery of blocks of nodes from the analysis of relationship data (White, Boorman and Breiger 1976). We have concentrated on the case of *a priori* specified blocks because it is a natural first step. One

approach to the problem of discovering the blocks *a posteriori* is to formulate it in Bayesian terms. In the remainder of this section we outline an approach to this. For simplicity we shall describe it in terms of a single adjacency matrix, $\boldsymbol{X} = \boldsymbol{X}(1)$.

To begin, we shall assume that we know that there are t node-blocks. In an actual analysis, the value of t would be varied to see how it changes the results. The block sizes can be controlled in the following way. Let

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_t)$$

be a probability vector, $\lambda_i > 0$, $\Sigma_i \lambda_i = 1$. In the simplest model we let the *a priori* probability that a node is in block j be λ_j. The block sizes are then multinomially distributed with parameters g and $\boldsymbol{\lambda}$. For example, for a 4-block model where we wish to have *approximately* the same number of nodes in each block we let $t = 4$ and $\lambda_i = 1/4$ $i = 1, 2, 3, 4$. Such assumptions do not force the node-blocks to be equal but do keep their sizes similar. Again, a sensitivity analysis varying $\boldsymbol{\lambda}$ over a reasonable range will reveal how much the final results depend on this value.

Let $\boldsymbol{C}$ be a g by t block-membership indicator-matrix, *i.e.*,

$$C_{ij} = \begin{cases} 1 & \text{if } i \in B_j \\ 0 & \text{otherwise.} \end{cases}$$

Then the rows of $\boldsymbol{C}$ are independent and identically distributed given $\boldsymbol{\lambda}$ and t with

$$P(C_{ij} = 1 | \boldsymbol{\lambda}, t) = \lambda_j.$$

The next prior assumption concerns the pattern of pair-block density parameters that are allowed by the model. We divide the density parameters π_{kl} $k = 1, \ldots, t$ into two types – *low* (type-L) and *unspecified* (type-U). The prior assumption that we make on the (π_{kl}) assumes that the division of pair-blocks into those of type-L and those of type-U is completely specified. For example, in a 4-block "dominance" model this specification might look like:

	1	2	3	4
1	U	L	L	L
2	U	U	L	L
3	U	U	U	L
4	U	U	U	U

This is analogous to the image matrix used in BLOCKER (Heil and White 1976). We emphasize that this specification is a prior assumption. Once the specification has been made we then assume that the (π_{kl}) for type-L pair-blocks are independent and identically distributed from a beta prior of the form

$$P_{\mathrm{L}}(\pi) = K(1-\pi)^{b-1},$$

where K is a normalizing constant. This type of prior is concentrated on low values of π. The prior parameter, b, controls the degree of this concentration. The parameter b plays the role of the tolerance parameters used in some blockmodel analyses to accommodate the difficulty of finding pure zero-blocks. If b is high, the block density parameters for L-type pair-blocks is assumed *a priori* to be very small.

The prior for the $\{\pi_{kl}\}$ for U-type pair-blocks is less obvious. We wish to avoid the convenient fiction of "one-blocks" (White, Boorman and Breiger 1976) since they never exist in real data. Rather we view the prior for a π_{kl} U-type pair block as being "flat" in some sense, *i.e.*, unspecified. A simple choice is to assume that it is uniform, *i.e.*,

$$P_{\mathrm{U}}(\pi) = 1.$$

In summary, our Bayesian model for this simple case of a stochastic blockmodel is characterized by:

(a) the number of blocks, t,
(b) the block size distribution vector, λ,
(c) the pattern of L and U type pair blocks,
(d) the distributions $P_{\mathrm{L}}(\pi)$ and $P_{\mathrm{U}}(\pi)$ of the block densities π_{kl}.

Given these quantities, a sociomatrix $\boldsymbol{X}$ can be generated from the Bayesian stochastic blockmodel as follows.

(a) Assign nodes independently to the t node-blocks according to the distribution given by λ.
(b) Generate the block densities π_{kl} from the distribution $P_L(\pi)$ for the L-type pair blocks, and from the distribution $P_U(\pi)$ for the U-type pair-blocks.
(c) Generate X from the π_{kl} as in an ordinary SB with *a priori* blocks.

Given a matrix $\boldsymbol{X}$ generated from this Bayesian model, we can use Bayes' theorem to compute a posterior probability distribution for the block membership of each node. This has the form

$$P(C_{ij} = 1|\boldsymbol{X}) \qquad i = 1,\ldots,g, j = 1,\ldots,t.$$

This $g \times t$ array of values gives the probability that each node belongs to each block. If the blockmodel "works" then these probabilities should be near zero or one. If the blockmodel does not work then the posterior probabilities will be near their *a priori* values, *i.e.*, λ_1, through λ_t. If the posterior probabilities are not near 0 or 1 then the blockmodel is not informative and one cannot tell which nodes go into which blocks. For example, suppose $t = 2$ and $g = 10$. The posterior block membership probabilities might look like this:

Node	Block	
	1	2
1	0.90	0.10
2	0.85	0.15
3	0.95	0.05
4	0.70	0.30
5	0.50	0.50
6	0.45	0.55
7	0.05	0.95
8	0.10	0.90
9	0.15	0.85
10	0.05	0.95

The interpretation of the array is that block 1 contains nodes 1, 2, 3, and 4 with high probability; block 2 contains nodes 7, 8, 9 and 10, with high probability, and the data do not give strong evidence for the block membership of either nodes 5 or 6.

We believe that the Bayesian model outlined above has great potential for formalizing the current techniques of blockmodeling and in-

tegrating them as components of a consistent, systematic approach to the study of relational data.

References

Breiger, R.L.
1981 "Comment on An Exponential Family of Probability Distributions for Directed Graphs". *Journal of the American Statistical Association 76*: 373, 51–53.

Fienberg, S.E.
1980 *The Analysis of Cross-Classified Categorical Data* (2nd edition) Cambridge, MA: MIT Press.

Fienberg, S.E. and S.S. Wasserman
1981 "Categorial Data Analysis of Single Sociometric Relations". In S. Leinhardt (ed.) *Sociological Methodology 1981*. San Francisco: Jossey-Bass, 156–192.
1980 "Methods for the Analysis of Data for Multivariate Directed Graphs". In *Proceedings of the Conference on Recent Developments in Statistical Methods and Applications*. Taipei, Taiwan: Institute of Mathematics, Academia Sinica.

Heil, G.H. and H.C. White
1976 "An Algorithm for Finding Simultaneous Homomorphic Correspondences Between Graphs and their Image Graphs". *Behavioral Science 21*: 26–35.

Holland, P.W., and S. Leinhardt
1981a "An Exponential Family of Probability Distributions for Directed Graphs". *Journal of the American Statistical Association 76*: 373, 33–50.
1981b "Reply to Comments on An Exponential Family of Probability Distributions for Directed Graphs". *Journal of the American Statistical Association 76*: 373, 62–65.

Lorrain, F. and H.C. White
1971 "Structural Equivalence of Individuals in Social Networks". *Journal of Mathematical Sociology 1*: 49–80.

Moreno, J.L.
1934 *Who Shall Survive?* Washington, D.C.: Nervous and Mental Disease Publishing Company.

Sampson, S.F.
1969 "Crisis in a Cloister", Ph.D. thesis, Cornell University: University Microfilms, No. 69–5775, Ann Arbor, Michigan.

Schwartz, Joseph E.
1977 "An Examination of CONCOR and Related Methods for Blocking Sociometric Data". In D. Heise (ed.), *Sociological Methodology 1977*, San Francisco: Jossey-Bass, 255–282.

White, H.C., S.A. Boorman and R.L. Breiger
1976 "Social Structure from Multiple Networks I: Blockmodels of Roles and Positions". *American Journal of Sociology 81*: 4: 730–780.