Pseudolikelihood Estimation for Social Networks

DAVID STRAUSS and MICHAEL IKEDA*

Interest in log-linear modeling for social-network data has grown steadily since Holland and Leinhardt (1981) proposed their p_1 model. That model was designed for a single binary relationship (directed graph) representing interactions between individuals. It assumed that interactions between pairs of individuals are mutually independent. Subsequent work has extended the model in various ways, including block-modeling and the case of dependence between pairs of individuals. In empirical work it would often be desirable to fit a wide variety of these models, as the differences in predictions or goodness of fit are likely to provide insights into the data. This has not been common practice, however, because estimation for some of the models has been difficult, and the maximum likelihood schemes developed for others involve different computer programs not always available in standard packages. The focus of this article is on a general estimation technique that maximizes the pseudolikelihood, the product of the probabilities of the binary variables, with each probability conditional on the rest of the data. The method is shown to be equivalent to a weighted least squares procedure and thus can be carried out with standard computer packages. In cases where true maximum likelihood estimation is available for comparison the two methods seem to work about equally well. The pseudolikelihood estimation is used in an example where the fits of a large number of different models are compared. Some of these models, such as various Markov block models, have not previously been proposed. In this example (as in others considered) it appears that the p_1-type models are overparameterized, and that much more parsimonious models give tolerable fits.

1. INTRODUCTION

During the last few years the modeling of social-network data has become increasingly popular in statistics. In its simplest form a social network is a square array, or graph, G of binary random variables y_{ij}, with the event $y_{ij} = 1$ denoting an arc, or tie, between individuals i and j. In different contexts this might indicate that i knows j, i does business with j, i reports to j, and so on. Numerous probability models for such data have been developed in sociology and other disciplines [see Strauss and Freeman (1989) for a review]. Holland and Leinhardt (1981) were the first to develop a log-linear model for network data. Their proposed p_1 model for the graph distribution of G assumes independence of the dyads (y_{ij}, y_{ji}) but includes parameters for density and reciprocity of arcs and for the individuals' tendencies to emit and attract arcs. Since then numerous generalizations of the model have been developed. These include the case of polytomous data (e.g., Wasserman and Iacobucci 1986), the treatment of multiple relationships (Fienberg, Meyer, and Wasserman 1981, 1985; Wasserman 1987), stochastic block models (Holland, Laskey, and Leinhardt 1983; Wang and Wong 1987; Wasserman and Anderson 1987), and Markov models admitting dyad dependence (Frank and Strauss 1986). The article by Wasserman and Iacobucci contains a useful summary.

Parameter estimation for the models assuming dyad independence can be performed with maximum likelihood. As noted by Holland and Leinhardt (1981), however, the relaxation of that assumption makes estimation difficult, and this has inhibited the development of dyad-dependent models. The problem arises because of an intractable normalizing function in the likelihood, which generally makes maximum likelihood estimation impossible. In this article we propose an estimation method that we call maximum pseudolikelihood estimation. The pseudolikelihood function is simply the product of the probabilities of the y_{ij}, with each probability conditional on the rest of the data. The method avoids the technical difficulty inherent in the maximum likelihood approach and can be performed with standard statistical packages.

The number of log-linear models available has become large enough that a classification scheme is useful. Here, we consider a classification based on three factors: (a) whether the model assumes dyad independence; (b) whether it is suitable for symmetric arrays (undirected graphs), with $y_{ij} \equiv y_{ji}$, or for asymmetric arrays (digraphs); (c) whether it is a block model, with parameters corresponding to an a priori grouping of individuals. This leads to the eightfold classification shown in Table 1. Models for most of the categories are already familiar, and we introduce some new models for the other cases. All of the models in Table 1 can conveniently be fitted with the pseudolikelihood method.

In most of this article we restrict attention to models for a single relationship, represented by a binary array $\{y_{ij}\}$. In Section 2 we begin by specifying the models in terms of the conditional probabilities of the y_{ij}, given the rest of the data. This conditional form proves convenient for our classification of the models; it is also the natural way to express the models when pseudolikelihood estimation is used. In Section 3 we develop the method and show its relationship to both a logistic regression procedure and maximum likelihood. It turns out that the first two are equivalent, and they are equivalent to maximum likelihood in the symmetric dyad-independence case. In Section 4 we analyze some well-known data of Sampson (1968) by fitting a variety of models. In data analysis it is often advisable to consider a wide range of possible models, and our aim here is to point out and illustrate several issues that arise in practice.

* David Strauss is Professor, Department of Statistics, University of California, Riverside, CA 92521. Michael Ikeda is Mathematical Statistician, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233. The authors thank Barry Arnold for some helpful discussions, and two referees and an associate editor for detailed comments that have substantially improved the article.

Journal of the American Statistical Association
March 1990, Vol. 85, No. 409, Theory and Methods

Table 1. Classification of Models

Classification	*Nonblock models*	*Block models*
Dyad-independent		
Directed graph	p_1, (2.3)	Wang and Wong (1987), (2.6)
Undirected graph	Symmetric p_1, (2.5)	(2.8)
Dyad-dependent (Markov)		
Directed	(p_1, σ, τ), (2.11)	See Section 4.
Undirected	(ρ, σ, τ), (2.9)	(ρ, σ, τ), block model, (2.12)

NOTE: The display numbers correspond to models defined in Section 2. Some additional examples are given in Section 4.

2. MODELS

Let G be a realization of a $g \times g$ random binary array (graph); that is, $G = \{y_{ij}: i \neq j; 1 \leq i, j \leq g\}$. The dyad D_{ij} is the ordered pair (y_{ij}, y_{ji}). We write G_{ij}^{+} for the realization G with y_{ij} set to 1, G_{ij}^{-} for the realization with y_{ij} set to 0, and C_{ij} (complement) for a specification of $\{y_{rs}: (r, s) \neq (i, j)\}$. When no confusion arises the subscripts on C will be suppressed. We are concerned with log-linear models of the form

$$\Pr(G) = \{1/Z(\theta)\}\exp\{\theta' x(G)\}, \tag{2.1}$$

where θ is a vector of parameters and $x(G)$ a corresponding vector of graph statistics. For a symmetric graph the components of x might be the number of lines $\sum y_{ij}$, the number of triads $\sum\sum\sum y_{ij}y_{jk}y_{ki}$, and so on. The normalizing function $Z(\theta)$ is the sum of $\exp\{\theta' x(G)\}$ over all $2^{g(g-1)}$ possible graphs. For graphs with $g \leq 6$, say, it is feasible to compute Z explicitly, but for large graphs in the dyad-dependent case the Z function is intractable. As a result, maximum likelihood estimation for the dyad-dependent models is not available. From (2.1), however, it follows that $\Pr(y_{ij} = 1 \mid C) = \Pr(G^{-})/\{\Pr(G^{+}) + \Pr(G^{-})\}$, a form that does not involve Z. With the notation $\text{logit}(t) = \log\{t/(1-t)\}$ we have, more compactly,

$$\begin{aligned}\text{logit}\,\Pr(y_{ij} = 1 \mid C) &= \theta'\{x(G^{+}) - x(G^{-})\} \\ &= \theta'\,\Delta x_{ij},\end{aligned} \tag{2.2}$$

where $\Delta x_{ij} = x(G^{+}) - x(G^{-})$ is the vector of changes in $x(G)$ when y_{ij} changes from 1 to 0. We refer to a valid specification (2.2) as a *logit model.*

The conditional probabilities (2.2) are not necessarily compatible in the sense of being consistent with some joint probability distribution $\Pr(G)$. Arnold and Press (1989) gave sufficient conditions for compatibility. In the present context, however, one would normally know $\Pr(G)$ in advance or be able to deduce it from inspection of (2.2). It is worth noting that the logit models are identifiable in the sense that the distribution $\Pr(G)$ corresponding to (2.2) is unique. This too follows from a result of Arnold and Press (1989). Their method of proof, in our context, is to construct a Markov chain of graphs G, with the conditional probabilities from (2.2) as transition probabilities. The state space is the set of all possible graphs, and the transitions are the possible changes of a single y_{ij}, either to y_{ij} or $1 - y_{ij}$. The pairs ij are taken cyclically in some fixed order. The chain is evidently aperiodic and irreducible, and thus has a long-run distribution that must be the unique distribution consistent with the logit model.

One might ask why the models are expressed in terms of conditional probabilities of the variables y_{ij} rather than the dyads D_{ij}, since the latter have been the traditional modeling unit. The reason is that if we define a pseudolikelihood as the product of the conditional likelihoods of dyads rather than y_{ij}'s, the maximization would no longer be equivalent to a regression procedure that can be performed with standard computing packages. Much of the advantage of the pseudolikelihood approach would thus be lost. The choice between the two conditional specifications is in any event only a matter of notation: As we have seen, there is a one-to-one correspondence between graph models (2.1) and compatible logit models of form (2.2), and a similar result holds for compatible specifications of dyad-conditional probabilities. Thus each logit model is a dyad model, and vice versa.

We now consider the model classification of Table 1 and express various cases in logit form. Holland and Leinhardt (1981) defined their p_1 model by

$$p_1(G) \propto \exp\left[\psi m + \rho y_{++} + \sum_i \alpha_i y_{i+} + \sum_j \beta_j y_{+j}\right]. \tag{2.3}$$

Here $m = \frac{1}{2}\sum_i \sum_j y_{ij}y_{ji}$, the number of mutual arcs, y_{++} is the total number of arcs, and so on. The parameters ψ, ρ, $\{\alpha_i\}$, and $\{\beta_j\}$ relate to overall mutuality, density, expansiveness, and attractiveness, respectively. (Holland and Leinhardt used the symbols ρ and θ in place of our ψ and ρ.) Side conditions $\sum \alpha_i = \sum \beta_j = 0$ ensure that the model is identifiable. Expressed in logit form, (2.3) is simply

$$\text{logit}\, p_1(y_{ij} = 1 \mid C) = \rho + \psi y_{ji} + \alpha_i + \beta_j. \tag{2.4}$$

To obtain a symmetric form, appropriate for an undirected graph, note that in this case $m = y_{++}$ and $y_{i+} = y_{+i}$. This suggests that in (2.3) $\psi + \rho$ be replaced by a new parameter ρ, and $\alpha_i + \beta_i$ be replaced by α_i. The resulting logit model is

$$\text{logit}\, p_1(y_{ij} = 1 \mid C) = \rho + \alpha_i + \alpha_j. \tag{2.5}$$

A side condition for the α_i is required for identifiability; the natural choice is $\sum \alpha_i = 0$.

Wasserman and Galaskiewicz (1984), Wang and Wong (1987), and others proposed a directed-graph model that retains the dyad independence but introduces parameters

associated with a block structure. In logit form, the model may be written as

$$\text{logit Pr}(y_{ij} = 1 \mid C) = \rho + \psi y_{ji} + \alpha_i + \beta_j + \sum_k \sum_l \lambda_{kl}\delta_{ij,kl}, \quad (2.6)$$

where the λ_{kl} are the block parameters and

$$\delta_{ij,kl} = 1 \quad \text{if } (i, j) \text{ belongs to block } k, l$$
$$= 0 \quad \text{otherwise.} \quad (2.7)$$

A block-diagonal form of (2.6) arises when $\lambda_{kl} = 0$ for $k \neq l$. Wang and Wong considered a single block-parameter model, where $\lambda_{kl} = \lambda$ if $k = l$ and 0 otherwise. A version of (2.6) for undirected graphs is easily constructed as before. It is

$$\text{logit Pr}(y_{ij} = 1 \mid C) = \rho + \alpha_i + \alpha_j + \sum \sum \lambda_{kl}\delta_{ij,kl}. \quad (2.8)$$

We now turn to models that do not assume dyad independence. The general class is too complicated to be useful for data analysis, and we restrict ourselves to Markov models (Frank and Strauss 1986). For such models $\Pr(y_{ij} = 1 \mid C)$ depends solely on those y_{rs} such that at least one of r and s is equal to i or j. One of the simplest, called the (ρ, σ, τ) model by Frank and Strauss, is the undirected-graph model

$$\Pr(G) \propto \exp(\rho R + \sigma S + \tau T). \quad (2.9)$$

Here R is the number of arcs (cases where $y_{ij} = 1$), S is the number of two-stars (i.e., distinct subscripts i, j, k such that $y_{ij}y_{ik} = 1$), and T is the number of triads (i, j, k) such that $y_{ij}y_{jk}y_{ki} = 1$. As before, ρ is a parameter related to overall density; σ and τ correspond to clustering and to transitivity of arcs. The logit form is

$$\text{logit Pr}(y_{ij} = 1 \mid C) = \rho + \sigma\Delta S + \tau\Delta T, \quad (2.10)$$

where ΔS is the change in the number of two-stars when y_{ij} changes from 1 to 0, for example. This form conveniently displays how the odds of an arc between i and j depend on the neighbors.

Frank and Strauss considered directed Markov graphs as well. The range of possible models then becomes very broad; for example, two-stars may be "outstars" of form $\{ij, ik\}$ (both arcs originating from the same vertex i), "instars" of form $\{ji, ki\}$, or of mixed form $\{ij, jk\}$. The number of each type could be taken as a graph statistic, each with its own parameter. Similarly, there are several triad-count statistics (Frank and Strauss 1986; Holland and Leinhardt 1976) that may be included. In practice one would usually require a model with rather few parameters. We do not attempt to list the possible cases, but instead give some examples in Section 4.

Of course, it is possible to create a model combining the Markov property with the expansiveness and attractiveness parameters of the p_1 model. One simple version that encompasses both (2.4) and (2.10) is

$$\text{logit Pr}(y_{ij} = 1 \mid C) = \rho + \psi y_{ji} + \alpha_i + \beta_j + \sigma\Delta S + \tau\Delta T, \quad (2.11)$$

where we might take S to be the number of two-stars of form $\{ij, ik\}$ and T to be the number of $\{ij, jk, ki\}$ triads, for example. This may be called a (p_1, σ, τ) model. A natural question to consider in practice is whether (2.11) gives a substantially better fit than (2.4) or (2.10).

Finally, we propose a class of *Markov block models*. These may be defined in a way that parallels the block models (2.6). For example, consider a symmetric array $\{y_{ij}\}$ partitioned into blocks B_{kl}. One simple Markov block-diagonal model is

$$\Pr(G) \propto \exp\left\{\rho R + \sum_k \sigma_k S^{(k)} + \sum_k \tau_k T^{(k)}\right\}, \quad (2.12)$$

where

$$S^{(k)} = \sum_{h,i,j \in \text{block } k} y_{hi}y_{hj}, \quad (2.13)$$

the number of two-stars within block k, and

$$T^{(k)} = \sum_{h,i,j} y_{hi}y_{ij}y_{hj}, \quad (2.14)$$

the number of transitive triads within block k. Of course, one could make the parameters σ_k (or the parameters τ_k) equal, include parameters for off-diagonal blocks, and include p_1-type parameters α_i, β_j, and so on. Equation (2.12) specifies one example of a Markov block model for undirected graphs. It is straightforward to define corresponding Markov block models for directed graphs as well. We see some of these models in the example of Section 4.

The logit form of (2.12) is

$$\text{logit Pr}(y_{ij} = 1 \mid C) = \rho + \sigma_k\Delta S_{ij}^{(k)} + \tau\Delta T_{ij}^{(k)}, \quad (2.15)$$

where $\Delta S_{ij}^{(k)}$ is the increase in the number of two-stars within block k resulting from a change of y_{ij} from 0 to 1, for example. The increase will be 0 unless i and j are both in block k. In that case $\Delta S_{ij}^{(k)}$ is just $\Sigma\ (y_{ih} + y_{jh})$, with the sum over h in block k.

For some parameter values the Markov models without blocks are unsuitable for large data sets because of the possibility of degeneracy (Strauss 1986). To see this, consider for simplicity the (ρ, σ, τ) model (2.9) with $\sigma = 0$. It can be shown that if τ is positive, then as the number of vertices g tends to infinity the probability that an arbitrarily large proportion of lines are present tends to 1. Loosely speaking, a sufficiently large graph is sure to be almost complete, whatever the value of ρ, if τ is positive. An analogous situation obtains for the case $\sigma > 0$, $\tau = 0$. This means that for large graphs the model is unlikely to generate data sets displaying more than one cluster of arcs. Nevertheless, if the data display clustering (or cliquing) and if it seems legitimate to take the corresponding block structure as given, then the Markov block models may be a plausible choice. As we shall see, they are no harder to

fit than the dyad-independence block models, and which are preferable for a given data set is largely an empirical question.

3. ESTIMATION

We have noted that for dyad-dependence models the awkward normalizing constant $Z(\theta)$ in the likelihood (2.1) generally makes maximum likelihood estimation impossible. The dyad-dependence models are in many respects analogous to interactive models on the rectangular lattice, such as the Ising model; there too an intractable normalizing constant rules out the possibility of maximum likelihood estimation (Strauss 1986). Our proposed estimation method is related to procedures suggested by Besag (1974, 1975) for the lattice case. We define the pseudolikelihood function to be

$$\mathrm{PL}(\theta) = \prod_{i \neq j} \Pr(y_{ij} \mid C_{ij}), \tag{3.1}$$

and a maximum pseudolikelihood estimator (MPE) to be a value of θ that maximizes (3.1). Since no conditional probability in (3.1) involves $Z(\theta)$, $\mathrm{PL}(\theta)$ should be much easier to maximize than the true likelihood (2.1). The MPE generally differs from the maximum likelihood estimator (MLE) except when the conditional probabilities in (3.1) are independent of C. This occurs in the symmetric block model (2.8) and its special case, the symmetric p_1 model (2.5).

We now show that maximization of (3.1) can be performed by a logistic regression. More formally, we have the following theorem.

Theorem. For a given logit model of form (2.2), maximization of (3.1) is equivalent to a maximum likelihood fit of logistic regression to the model (2.2) for independent observations y_{ij}. It can be implemented as an iteratively reweighted Gauss–Newton least squares procedure.

Proof. It is convenient to replace the indexes (i, j) with a single index r. Let P_r denote the conditional probability of y_r, given C_r [as defined by (2.2)], and set $Q_r = 1 - P_r$. Denote the components of θ by θ_k $(k = 1, 2, \ldots)$. The pseudolikelihood (3.1) can be written as $\mathrm{PL} = \Pi_r P_r^{y_r} Q_r^{1-y_r}$. The pseudolikelihood estimators are thus solutions of the equations

$$\frac{\partial}{\partial \theta_k} \log \mathrm{PL} = \sum_r \left\{ \frac{y_r}{P_r} \frac{\partial P_r}{\partial \theta_k} + \frac{1 - y_r}{Q_r} \left(-\frac{\partial P_r}{\partial \theta_k} \right) \right\} = 0, \qquad k = 1, 2, \ldots, \tag{3.2}$$

which simplify to

$$\sum_r \frac{1}{P_r Q_r} (y_r - P_r) \frac{\partial P_r}{\partial \theta_k} = 0, \qquad k = 1, 2, \ldots. \tag{3.3}$$

This can be written as

$$\frac{\partial}{\partial \theta_k} \sum_r w_r (y_r - P_r)^2 = 0, \qquad k = 1, 2, \ldots,$$

where $w_r = 1/(P_r Q_r)$, provided that the w_r are treated as constants in the differentiation. But this is iteratively reweighted least squares, with the weights w_r recomputed at each step from the current value of θ.

This result is (almost) a special case of a more general result on maximum likelihood for the generalized linear model (e.g., see McCullagh and Nelder 1983, sec. 2.5). It shows that maximization of (3.1) can be performed by a standard logistic-regression computer routine, even though the variables y_{ij} in (2.2) are not conditionally independent. We have used the version in the BMDP package throughout our work. To implement it one constructs a set of $g(g - 1)$ observations, each consisting of a binary "dependent" variable y_{ij} and a vector of "independent" variables Δx_{ij}. The latter are given by $x(G^+) - x(G^-)$, as in (2.1), and are easily constructed from the data matrix G. We note that logistic regression had previously been suggested as an ad hoc procedure by Frank and Strauss (1986) in connection with the (ρ, σ, τ) model (2.10).

Maximum likelihood estimation is feasible for all of the dyad-independent models discussed here; the methods of Darroch and Ratcliff (1972) and Feinberg et al. (1985), among others, have been used to implement it. Nevertheless, if one wishes to fit both dyad-dependent and dyad-independent models to the same data set, it seems desirable to do so with a single computer package (as can be done with the MPE). In addition, much of the supplementary information automatically supplied by the program will be useful, even though the true model is not a standard logistic regression. For example, the number of y_{ij} correctly classified by the regression function with various cut points will usually be of interest. The information on which explanatory variables in the logistic regression contribute most of the predictions of the y_{ij} is sometimes useful as well. One caution should be noted, however: The quoted standard errors of the estimated parameters do not apply, because the $g(g - 1)$ observations in the regression are certainly not independent.

Since it is difficult to compare the MLE and MPE analytically, we offer some experimental comparisons. For Markovian models MLE is only really feasible when there is a single parameter, in which case a graphical method is available (Strauss 1986). Using a Metropolis-type simulation method given in that article, we generated realizations of the (ρ, σ, τ) model with $\rho = \tau = 0$ and various values of σ. Results are shown in Table 2. For each value of σ there were five replicates, with the mean and root mean squared error computed for the MPE and the MLE. The two methods appear to give estimators that are about equally good. A similar conclusion applies to results in Table 3, which corresponds to $\rho = \sigma = 0$ and various values of τ. In addition, we performed some experiments with smaller graphs, of size $g = 15$ and 20. As expected, both methods gave more variable estimates than those for the larger graphs, but the root mean squared errors for the two methods were again quite close in all cases.

For the multiparameter Markov models it becomes difficult to perform systematic comparisons because the MLE is not available. In the case of the dyad-independence

Table 2. Comparison of Estimators for Data From a One-Parameter Markov Model: (ρ, σ, τ) Model With $\rho = \tau = 0$; $g = 30$

True σ	Mean of MLE's	Mean of MPE's	Root mean squared error: MLE's	Root mean squared error: MPE's
−.10	−.097	−.096	.010	.010
−.05	−.054	−.054	.009	.008
0	−.007	−.006	.009	.008
.075	.076	.077	.011	.012

models, however, the MPE and MLE may be readily compared, and they have consistently given similar results. A typical case is the Sampson data of Section 4. We estimated the parameters of the p_1 model with both methods and derived the two sets of unconditional fitted probabilities. For this 18×18 data set there were $18 \times 17 = 306$ pairs. In 254 cases the absolute difference of the fitted probabilities was less than .05, and in 299 cases less than .10. A similar procedure with the p_1 block-diagonal model (2.6) gave even closer agreement, with corresponding counts of 292 and 304. The correlation between the two sets of predictors was .935 in the first case and .997 in the second. We found similar agreement between the two sets of parameter estimates (see Sec. 4).

The consistency of the MPE and MLE deserves some comment. Consistency is not a meaningful term unless the parameter set remains fixed as the number of vertices g tends to infinity. This condition is not satisfied by the general p_1 model, for example. The (ρ, σ, τ) model (2.9) does satisfy it, but as we have seen, that model is sometimes degenerate. In certain simple cases matters can be resolved satisfactorily. For example, consider, the p_1 model where all the parameters α_i are constrained to be equal and the parameters β_j are constrained to be equal. Then, the common values may both be taken to be 0, and the counts of mutual, asymmetric and null dyads follow a trinomial distribution. It is easy to show that the maximum likelihood estimates of ψ and ρ in (2.3) are then almost surely consistent. It can be verified directly that the MPE estimates are identical to the MLE's for this case, even though the likelihood and pseudolikelihood functions are not equal. Similar conclusions hold for the block-model extension (2.6) of this simple case.

If we have a random sample of n realizations from a $g \times g$ graph model, we can compute an MPE by maximizing the product of the n pseudolikelihoods. The behavior of this estimator as n tends to infinity follows from general results of Arnold and Strauss (1988): The estimator is consistent and asymptotically normal, with asymptotic efficiency given by certain information quantities. Although these results are encouraging for the use of the MPE, note that the case of one realization from a "large" graph is more common than a large random sample of n graphs of a fixed order g. We noted earlier that interactive lattice models share many features of the dyad-dependence models. Thus it is worth remarking that Geman and Graffigne (1987) proved the consistency of the MPE as the lattice becomes large.

Table 3. Comparison of Estimators for Data From a One-Parameter Markov Model: (ρ, σ, τ) Model With $\rho = \sigma = 0$; $g = 30$

True τ	Mean of MLE's	Mean of MPE's	Root mean squared error: MLE's	Root mean squared error: MPE's
−.2	−.182	−.186	.028	.026
−.1	−.109	−.113	.042	.046
0	.010	.008	.014	.014
.1	.106	.105	.014	.014
.15	.147	.147	.017	.016

Parameter identifiability can be an issue in block modeling. For example, suppose that we have a partition of integers $1, \ldots, g$ into disjoint subsets, and let B be the set of pairs (ij) such that i and j belong to the same subset. Given a symmetric graph G, let

$$T = \sum\sum\sum_{i<j<k} y_{ij}y_{jk}y_{ik} \tag{3.4}$$

be the number of transitive triads, and let T_B be the number of such triads with ij, jk, and ik belonging to B. Consider the Markov block model

$$\Pr(G) \propto \exp\{\rho R + \tau_B T_T + \tau T\}, \tag{3.5}$$

which is equivalent to

$$\text{logit } \Pr(y_{ij} = 1 \mid C) = \rho + \tau_B \Delta T_B + \tau \Delta T. \tag{3.6}$$

The parameters in (3.5) are certainly identifiable. Nevertheless, suppose that we observe a graph with $y_{ij} = 0$ whenever (ij) does not belong to B. Then, the columns ΔT_B and ΔT in (3.6) are identical, and the logistic regression is ill-conditioned. A similar issue would arise in maximum likelihood estimation of (3.5). In applications there should be no problem, because one would not contemplate a model such as (3.5) unless the block structure were already known, and in that case the redundancy of the term τT should be apparent by inspection of the data.

4. AN EXAMPLE

To illustrate the methods and some special points that arise in applications, we consider the well-known monastery data of Sampson (1968). In Table 4 the 1s indicate some degree of liking between two monks. The partition lines correspond to a by-now traditional block structure for the 18 monks; for example, see Wasserman and Anderson (1987).

Table 5 summarizes the fitting of the directed-graph models of Section 2; some explanation of the symbols for the models is given in the footnote. The MPE has been used in all cases. The table shows various summary statistics associated with the models. The maximized pseudolikelihood is one criterion for model comparisons. Even for the Markovian models this can be interpreted as a true likelihood for the logistic regression model (3.2), so [following Holland and Leinhardt (1981) and Wang and Wong (1987)] we might refer twice the difference of the log-pseudolikelihood to the χ^2 distribution as an informal test

Table 4. Sampson's (1968) Data

	Monk																	
Monk	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
2	0		1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1		0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
4	0	1	1		1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1	0	1		1	0	0	0	0	0	0	0	0	0	0	0	0
6	0	1	0	0	0		1	0	0	0	1	0	0	0	0	0	0	0
7	0	0	1	1	1	0		0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0		1	1	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	1		0	1	0	0	0	0	1	0	0
10	0	0	0	0	0	0	0	1	1		0	0	1	0	0	0	0	0
11	0	0	0	0	0	0	0	1	1	0		1	0	0	0	0	0	0
12	0	0	0	0	0	0	0	1	0	1	0		1	0	0	0	0	0
13	0	0	0	0	0	0	0	1	0	1	0	0		1	0	0	0	0
14	0	0	0	0	0	0	0	0	0	1	0	1	1		0	0	0	0
15	0	1	0	0	0	0	0	0	0	0	0	0	1	0		0	0	1
16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1		1	1
17	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1		1
18	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	

statistic. The sum of absolute residuals $\Sigma\,|y_{ij} - \tilde{p}_{ij}|$ is a useful measure of fit as well. The weighted sum of squared residuals, which is the quantity minimized by MPE, is not shown because it seems to provide little additional information. The final column shows the maximum number of correct predictions of the value of y_{ij}, based on the logistic regression with known C_{ij}'s and an optimal cut point. This is a standard feature in discriminant analysis, and it is an option with logistic regression in most computer packages. The total number of predictions is $18 \times 17 = 306$. Predicted probabilities derived from MPE are generally conditional on the C_{ij}'s and thus may not agree closely with the unconditional predictions from MLE (except, of course, when the methods are equivalent). Note that the comparisons discussed in the following are based on conditional prediction.

Model 1, the one-parameter Bernoulli model, is included to provide a baseline for comparisons; its fit is poor. Model 2 is a p_1 model with all α_i and β_j set to 0. This leaves just the mutuality and density parameters ψ and ρ in (2.3). Model 3 is the p_1, and according to the informal pseudolikelihood ratio test its improvement in fit over Model 2 is insufficient to justify the additional 34 parameters. (The same conclusion could be reached by comparison of the true log-likelihoods, -133.9 and -118.5.) The inclusion of an additional parameter τ_C, corresponding to a count of cyclic triads with $y_{ij}y_{jk}y_{ki} = 1$, gives Model 4, which fits only slightly better.

We now turn to the block models, taking the block structure of Table 4 as given a priori. Model 5 (denoted by p_1, λ) is a one-parameter extension to p_1 [see (2.6)]. It was previously fitted to this data by Wang and Wong (1987). The relatively high log-pseudolikelihood of -79.9 is close to the true log-likelihood of -82.1. For comparison, the values of $\{\alpha_i\}$ obtained by MLE and MPE are plotted in Figure 1a; the sets of estimators are in good agreement.

Table 5. Models for Sampson's Data

Model	Number of parameters	Maximized log-pseudolikelihood	Sum of absolute value of residuals	Number of residuels less than .1	Number of residuels less than .3	Number of residuels less than .7	Maximum number of correct predictions
1. Bernoulli	1	−145.6	91.5	0	0	306	250
2. ψ, ρ	2	−122.2	74.5	0	224	280	254
3. p_1	36	−102.6	62.7	149	238	281	268
4. p_1, τ_C	37	−98.0	60.5	157	238	282	268
Block models							
5. p_1, λ	37	−79.9	49.1	191	245	287	276
6. p_1, σ'_M	37	−78.9	49.0	192	245	287	275
7. ψ, ρ, λ	3	−100.3	62.5	194	201	297	259
8. ψ, ρ, σ'_I	3	−104.5	63.6	197	238	285	263
9. ψ, ρ, σ'_O	3	−118.0	73.4	194	224	280	259
10. ψ, ρ, σ'_M	3	−109.3	68.1	194	223	276	256
11. ψ, ρ, σ'_{Ik}	5	−103.3	62.6	197	239	285	264
12. ψ, ρ, τ'_C	3	−119.6	73.4	201	222	280	256
13. ψ, ρ, τ'_{TR}	3	−116.2	72.2	196	224	273	253

NOTE: ψ and ρ are the mutuality and density parameters, as in the p_1 model; λ is the block parameter in the model of Wang and Wong (2.6); subscripts *I*, *O*, and *M* on σ indicate indegree, outdegree, and mixed, respectively; subscripts *C* and *TR* on τ denote cyclic or transitive triads, respectively; a prime on σ or τ indicates a within-block count; and a subscript *k* on a parameter indicates a separate value for each block.

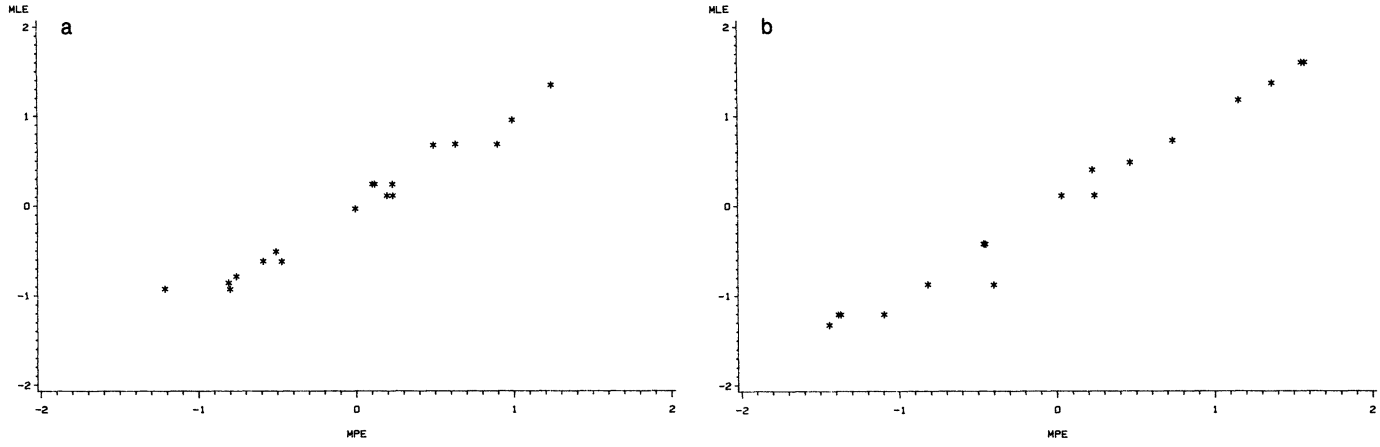

Figure 1. Parameter Estimates for the (p_1, λ) Model (2.6) by Maximum Likelihood and Maximum Pseudolikelihood: (a) Estimates of $\{\alpha_i\}$ ($1 \le i \le 18$); (b) Estimates of $\{\beta_j\}$ ($2 \le j \le 18$).

The same is true for the estimates of $\{\beta_j\}$, shown in Figure 1b.

Model 5 fits the data substantially better than the others considered so far. An almost identical fit is provided by the Markov block model, Model 6. [See (2.12); the prime on the clustering parameter σ indicates a count within blocks only, and the suffix M denotes mixed pairs of form $y_{ij}y_{jk}$.] The (ψ, ρ, λ) model, 7, has 34 fewer parameters than Model 5, yet its fit is not much worse. As before, the informal χ^2 test indicates that inclusion of the $\{\alpha_i\}$ and $\{\beta_j\}$ is unnecessary.

The remaining cases in the table are Markov block models. The number of potential models is very large, and we have reported only a few of them. Model 8 is the best-

Table 6. Predicted Conditional Probabilities, Multiplied by 100, for Sampson's (1968) Data

(ψ, ρ, λ) block model (7) with ψ = 1.03, ρ = −1.86, and λ = 1.31																		
0	37	37	37	37	37	37	4	4	4	4	4	4	4	4	4	4	4	
62	0	62	62	62	62	37	4	4	4	4	4	4	4	11	4	4	4	
62	62	0	62	37	37	62	4	4	4	4	4	4	4	4	4	4	4	
37	37	37	0	62	37	62	4	4	4	4	4	4	4	4	4	4	4	
62	62	37	62	0	37	62	4	4	4	4	4	4	4	4	4	4	4	
37	62	37	37	62	0	37	4	4	4	4	4	4	4	4	4	4	4	
37	37	62	37	37	62	0	4	4	4	4	4	4	4	4	4	4	4	
4	4	11	4	4	4	4	0	62	62	62	62	62	37	4	4	4	4	
4	4	4	4	4	4	4	62	0	62	62	37	37	37	4	11	4	4	
4	4	4	4	4	4	4	62	37	0	37	62	62	62	4	4	11	11	
4	4	4	4	4	11	4	37	62	37	0	37	37	37	4	4	4	4	
4	4	4	4	4	4	4	37	37	37	62	0	37	62	4	4	4	4	
4	4	4	4	4	4	4	62	37	62	37	62	0	62	11	4	4	4	
4	4	4	4	4	4	4	37	37	37	37	37	62	0	4	4	4	4	
11	4	4	4	4	4	4	4	4	4	4	4	4	4	0	62	37	37	
4	4	4	4	4	4	4	4	11	4	4	4	4	4	37	0	62	62	
4	4	4	4	4	4	4	4	4	4	4	4	4	4	37	62	0	62	
4	4	4	4	4	4	4	4	4	4	4	4	4	4	62	62	62	0	
(ψ, ρ, σ_1') Markov model (8), with ρ = −2.75, ψ = 1.72, and σ_1' = .71																		
0	52	35	21	35	21	21	6	6	6	6	6	6	6	6	6	6	6	
26	0	75	59	75	42	21	6	6	6	6	6	6	6	26	6	6	6	
26	86	0	59	52	21	42	6	6	6	6	6	6	6	6	6	6	6	
6	52	35	0	75	21	59	6	6	6	6	6	6	6	6	6	6	6	
26	86	52	42	0	12	59	6	6	6	6	6	6	6	6	6	6	6	
6	86	52	21	86	0	12	6	6	6	6	6	6	6	6	6	6	6	
6	69	75	12	35	59	0	6	6	6	6	6	6	6	6	6	6	6	
6	6	26	6	6	6	6	0	59	75	42	59	75	12	6	6	6	6	
6	6	6	6	6	6	6	86	0	86	26	21	52	12	6	26	6	6	
6	6	6	6	6	6	6	86	21	0	12	59	75	42	6	6	26	26	
6	6	6	6	6	26	6	52	59	52	0	12	52	12	6	6	6	6	
6	6	6	6	6	6	6	52	35	35	42	0	35	42	6	6	6	6	
6	6	6	6	6	6	6	86	35	75	12	59	0	26	26	6	6	6	
6	6	6	6	6	6	6	69	35	35	12	12	75	0	6	6	6	6	
26	6	6	6	6	6	6	6	6	6	6	6	6	6	0	59	21	21	
6	6	6	6	6	6	6	6	26	6	6	6	6	6	6	0	42	59	
6	6	6	6	6	6	6	6	6	6	6	6	6	6	12	42	0	59	
6	6	6	6	6	6	6	6	6	6	6	6	6	6	42	42	42	0	

fitting three-parameter Markov block model. Its fit is very similar to (if marginally worse than) that of Model 7. The subscript 0 on σ' in Model 9 indicates outstars, that is, the sum within blocks of $y_{ij}y_{ik}$. The poor performance of that model undoubtedly reflects Sampson's sampling scheme, whereby most of the outdegrees were constrained to be three. Note that Model 11, which allows the clustering parameter in Model 8 to differ among blocks, offers very little improvement. This parallels Wang and Wong's (1987) finding that different block parameters were unnecessary in their model. The three-parameter Markov models 12 and 13 are included only for illustration. They differ in which types of triads are counted: cyclic, as in *hi*, *ij*, and *jh*, or transitive, as in *hi*, *ij*, and *hj*. Frank and Strauss (1986) categorize the complete set of sufficient statistics needed for the most general homogeneous Markov graph.

Taking account of parsimony, the best models in Table 5 appear to be the three-parameter models 7 and 8. Each has a mutuality and a density parameter, and one additional parameter to account for the higher density of arcs within blocks. The interpretations of the block parameters λ and σ_1' are quite different, and there are no obvious grounds for preferring one over the other. Table 6 shows the fitted conditional probabilities (of $y_{ij} = 1$, given C_{ij}) for two models, and the parameter estimates in each case. Outside the three diagonal blocks there are just two values for the probabilities, corresponding to $y_{ji} = 0$ and $y_{ji} = 1$. Within the blocks the Markov conditional probabilities differ because they depend on neighboring dyads, whereas the other model predicts just two probabilities. The ability of these models to provide reasonable fits with so few parameters is worth noting; we have found much the same pattern in other data sets that we have examined.

5. CONCLUDING REMARKS

We have seen that maximum pseudolikelihood estimation is easy to implement and can be used for a large class of models. The class includes dyad-dependent (Markov) models, for which maximum likelihood is not generally feasible; in our experience, Markov models frequently provide satisfactory fits to data showing clustering or transitivity. The fitting of a wide variety of models is often a useful way to gain insight into the network structure.

We have restricted attention to the case of a single binary variable $\{y_{ij}\}$. The logit form for models and the associated pseudolikelihood scheme can be generalized to multivariate and polytomous cases. We conclude with some brief remarks on these. For the polytomous case, let a be a generic value of y_{ij}, and suppose that the value 0 has positive probability. For log-linear models of form (2.1) we can write

$$\log\{\Pr(y_{ij} = a \mid C_{ij})/\Pr(y_{ij} = 0 \mid C_{ij})\} = \theta'\Delta x,$$

a logit form analogous to those of Section 2. We can again define a pseudolikelihood by

$$\prod_{i \neq j} \Pr(Y_{ij} = y_{ij} \mid C_{ij}). \tag{5.1}$$

As in the binary case, maximization of (5.1) can be shown to be equivalent to the MLE for a polytomous logistic regression; this can be performed by the CATMOD procedure in the SAS computer package. In the multivariate case, if D_{ijr} denotes the dyad (i, j) with respect to relation r, one can define a pseudolikelihood as

$$\prod_{i \neq j} \Pr\left\{ \bigcap_r (D_{ijr} = d_{ijr}) \mid C_{ij} \right\}, \tag{5.2}$$

for example, where C_{ij} denotes $\{(Y_{klr}, Y_{lkr})$: $(k, l) \neq (i, j)$; $r = 1, 2, \ldots\}$. Maximization of this is not equivalent to logistic regression, but it will be free of the troublesome normalizing function. We can even reduce the problem to a logistic regression by defining instead a "doubly-pseudo"-likelihood

$$\prod_{i \neq j} \prod_r \Pr(y_{ijr} \mid C_{ijr}), \tag{5.3}$$

where C_{ijr} denotes $\{y_{k,l,s}$: $(k, l, s) \neq (i, j, r)\}$. We are currently studying the properties of such estimators.

[*Received June 1987. Revised January 1989.*]

REFERENCES

Arnold, B. C., and Press, S. J. (1989), "Compatible Conditional Distributions," *Journal of the American Statistical Association,* 84, 152–156.

Arnold, B. C., and Strauss, D. (1988), "Pseudolikelihood Estimation," Technical Report 164, University of California, Riverside, Dept. of Statistics.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society,* Ser. B, 36, 192–236.

——— (1975), "Statistical Analysis of Non-Lattice Data," *The Statistician,* 24, 179–195.

Darroch, J. N., and Ratcliff, D. (1972), "Generalized Iterative Scaling of Loglinear Models," *The Annals of Mathematical Statistics,* 43, 1470–1480.

Fienberg, S. E., Meyer, M., and Wasserman, S. (1981), "Analyzing Data From Multivariate Directed Graphs: An Application to Social Networks," in *Interpreting Multivariate Data,* ed. V. Barnett, London: John Wiley, pp. 289–306.

——— (1985), "Statistical Analysis of Multiple Sociometric Relations," *Journal of the American Statistical Association,* 80, 50–67.

Frank, O., and Strauss, D. (1986), "Markov Graphs," *Journal of the American Statistical Association,* 81, 832–842.

Geman, S., and Graffigne, C. (1987), "Markov Random Field Image Models and Their Application to Computer Vision," in *Proceedings of the 1986 International Congress of Mathematicians,* ed. A. M. Gleason, Providence, RI: American Mathematical Society.

Holland, P. W., Laskey, K., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks,* 5, 109–137.

Holland, P. W., and Leinhardt, S. (1976), "Local Structure in Social Networks," *Sociological Methodology,* 1–45.

——— (1978), "An Omnibus Test for Social Structure Using Triads," *Sociological Methods and Research,* 7, 227–255.

——— (1981), "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association,* 77, 33–50.

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models,* London: Chapman & Hall.

Sampson, S. F. (1968), *A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships,* unpublished Ph.D. dissertation, Cornell University, Dept. of Sociology.

Strauss, D. (1986), "On a General Class of Models for Interaction," *Society for Industrial and Applied Mathematics Review,* 28, 513–527.

Strauss, D. J., and Freeman, L. C. (1989), "Stochastic Modelling and

the Analysis of Structural Data," in *Research Methods in Social Network Analysis,* eds. L. Freeman, A. Romney, and D. White, Fairfax, VA: George Mason University Press, pp. 135–183.

Wang, Y. Y., and Wong, G. Y. (1987), "Stochastic Blockmodels for Directed Graphs," *Journal of the American Statistical Association,* 82, 8–19.

Wasserman, S. (1987), "Conformity of Two Sociometric Relations," *Psychometrika,* 52, 3–18.

Wasserman, S., and Anderson, C. (1987), "Stochastic A Posteriori Block-Models: Construction and Assessment," *Social Networks,* 9, 1–36.

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

Wasserman, S., and Iacobucci, D. (1986), "Statistical Analysis of Discrete Relational Data," *British Journal of Mathematical and Statistical Psychology,* 39, 41–64.