Latent Space Approaches to Social Network Analysis

Peter D. HOFF, Adrian E. RAFTERY, and Mark S. HANDCOCK

Network models are widely used to represent relational information among interacting units. In studies of social networks, recent emphasis has been placed on random graph models where the nodes usually represent individual social actors and the edges represent the presence of a specified relation between actors. We develop a class of models where the probability of a relation between actors depends on the positions of individuals in an unobserved "social space." We make inference for the social space within maximum likelihood and Bayesian frameworks, and propose Markov chain Monte Carlo procedures for making inference on latent positions and the effects of observed covariates. We present analyses of three standard datasets from the social networks literature, and compare the method to an alternative stochastic blockmodeling approach. In addition to improving on model fit for these datasets, our method provides a visual and interpretable model-based spatial representation of social relationships and improves on existing methods by allowing the statistical uncertainty in the social space to be quantified and graphically represented.

KEY WORDS: Conditional independence model; Latent position model; Network data; Random graph; Visualization.

1. INTRODUCTION

Social network data typically consist of a set of n actors and a relational tie $y_{i,j}$, measured on each ordered pair of actors $i, j = 1, \dots, n$. This framework has many applications in the social and behavioral sciences including, for example, the behavior of epidemics, the interconnectedness of the World Wide Web, and telephone calling patterns. Quantitative research on social networks has a long history, going back at least to Moreno (1934). The development of log-linear statistical models by Holland and Leinhardt (1981); Fienberg, Meyer, and Wasserman (1985); Wang and Wong (1987); and others represent major advances.

In the simplest cases, $y_{i,j}$ is a dichotomous variable indicating the presence or absence of some relation of interest, such as friendship, collaboration, transmission of information or disease, and so forth. The data are often represented by an $n \times n$ sociomatrix Y. In the case of binary relations, the data can also be thought of as a graph in which the nodes are actors and the edge set is $\{(i, j) : y_{i,j} = 1\}$. If (i, j) is in the edge set, then we write $i \to j$. If ties are undirected, in that $y_{i,j} = y_{j,i}$ for all $i \neq j$ by logical necessity, then we write $i \sim j$ if $y_{i,j} = 1$. However, even in the case of directed relations, ties often tend to be reciprocal ($y_{i,j} = y_{j,i}$ with high probability) and transitive ($i \to j, j \to k \Rightarrow i \to k$ with high probability). As such, probabilistic models of network relations have typically allowed for some sort of dependence between ties. For example, the p_1 model of Holland and Leinhardt (1981) includes parameters for the propensity of ties to be reciprocal, as well as parameters for the number of ties and individual tendencies to give or receive ties. However, these models are restrictive, as they assume the $\binom{n}{2}$ dyads $(y_{i,j}, y_{j,i})$ to be independent.

Frank and Strauss (1986) characterized the exponential family of random graph models by elaborating work of Besag (1974) developed in the context of spatial statistics. These have been referred to as the p^* class of models in the psychology and sociology literature (Wasserman and Pattison 1996). Given their general nature and applicability, we refer to them simply as (exponentially parameterized) random graph models. Frank and Strauss (1986) also proposed models with Markov structure that allow for forms of dyad dependence, often referred to as homogeneous monadic Markov models. Recent work of Corander, Dahmström, and Dahmström (1998); Crouch, Wasserman, and Trachtenberg (1998); Besag (2000); Handcock (2000); and Snijders (2002) has developed likelihood-based inference for these models based on Markov chain Monte Carlo (MCMC) algorithms. Approximate maximum likelihood approaches have been developed by Frank and Strauss (1986). Alternative estimating procedures based on the concept of pseudolikelihood (Besag 1974) have been proposed by Strauss and Ikeda (1990) and Wasserman and Patterson (1996); however, the statistical properties of pseudolikelihood estimators in this context have been criticized by Besag (2000) and Snijders (2002).

Recent works have explored the properties of homogeneous monadic Markov models. Results of Besag (2000), Handcock (2000), and Snijders (2002) suggest that commonly used models are more global than local in structure and that this contributes to model degeneracy and instability problems (Ruelle 1968). These issues are not resolved by alternative forms of estimation, but rather represent defects in the models themselves—at least to the extent that they are useful for modeling realistic social networks. These factors have motivated the development of alternative models without these restrictions.

For networks in which actors belong to prespecified groups Wang and Wong (1987) developed a stochastic blockmodel, an extension of the p_1 model that includes parameters describing differential rates of between-group and within-group ties. For cases in which group membership is not observed, Nowicki and Snijders (2001) presented a model in which the dyads in a social network are conditionally independent, given the latent class membership of each actor. In such a model, actors within a latent class are treated as *stochastically equivalent*; that is, the events $(i_1 \to j_1)$ and $(i_2 \to j_2)$ have the same probability if actors i_1 and j_1 are in the same respective latent classes

Peter D. Hoff is Assistant Professor of Statistics (E-mail: *hoff@stat.washington.edu*), Adrian E. Raftery is Professor of Statistics and Sociology (E-mail: *raftery@stat.washington.edu*), and Mark S. Handcock is Professor of Statistics and Sociology, University of Washington, Seattle, WA 98195 (E-mail: *handcock@stat.washington.edu*). Raftery's research was supported by Office of Naval Research grant N00014-96-1-1092, and Handcock's research was supported by National Institutes of Health grant R01 DAI2831-01. The authors thank two referees for helpful comments.

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Theory and Methods
DOI 10.1198/016214502388618906

as i_2 and j_2. Such a model may prove useful in identifying clusters of individuals for whom stochastic equivalence holds, that is, clusters of individuals who relate to all other actors in the system in a similar way. However, models based on distinct clusters may not fit well when many actors fall between clusters, or when relations are transitive yet there is no strong clustering.

In some social network data, the probability of a relational tie between two individuals may increase as the characteristics of the individuals become more similar. A subset of individuals in the population with a large number of social ties between them may be indicative of a group of individuals who have nearby positions in this space of characteristics, or "social space." Various concepts of social space have been discussed by McFarland and Brown (1973) and Faust (1988). In the context of this article, social space refers to a space of unobserved latent characteristics that represent potential transitive tendencies in network relations. A probability measure over these unobserved characteristics induces a model in which the presence of a tie between two individuals is dependent on the presence of other ties. Relations modeled as such are probabilistically transitive in nature. The observation of $i \to j$ and $j \to k$ suggests that i and k are not too far apart in social space, and therefore are more likely to have a tie. In Section 2 we develop a latent variable model for such transitive relations in which it is assumed each actor i has an unknown position z_i in social space. The ties in the network are assumed to be conditionally independent given these positions, and the probability of a specific tie between two individuals is modeled as some function of their positions, such as the distance between the two actors in social space. Estimation of positions is simplified by the use of a logistic regression model, and confidence regions for latent positions are computable using standard MCMC algorithms, as we describe in Section 3. In Section 4 we fit these latent-space models to a number of standard datasets and compare their performance in terms of model fit to alternative stochastic blockmodels. In addition to improving on model fit, the results from our approach are relatively easy to interpret, and modeling the positions as belonging to a low-dimensional Euclidean space provides a model-based means of graphically representing social network data.

2. LATENT POSITION METHODS

The data we model in this article consist of an $n \times n$ sociomatrix Y, with entries $y_{i,j}$ denoting the value of the relation from actor i to actor j, and possibly additional covariate information X. We focus on binary-valued relations, although the methods in this article can be extended to more general relational data using ideas from generalized linear models. Both directed and undirected relations can be analyzed with our methods, although the features of the model are slightly different in the two cases, as described below.

We take a conditional independence approach to modeling by assuming that the presence or absence of a tie between two individuals is independent of all other ties in the system, given the unobserved positions in social space of the two individuals,

$$P(Y|Z, X, \theta) = \prod_{i \neq j} P(y_{i,j}|z_i, z_j, x_{i,j}, \theta),$$

where X and $x_{i,j}$ are observed characteristics which are potentially pair-specific and vector-valued and θ and Z are parameters and positions to be estimated.

2.1 Distance Models

A convenient parameterization of $P(y_{i,j}|z_i, z_j, x_{i,j}, \theta)$ is the logistic regression model in which the probability of a tie depends on the Euclidean distance between z_i and z_j, as well as on observed covariates $x_{i,j}$ that measure characteristics of the dyad,

$$\begin{aligned}\eta_{i,j} &= \log\text{odds}(y_{i,j} = 1|z_i, z_j, x_{i,j}, \alpha, \beta) \\ &= \alpha + \beta' x_{i,j} - |z_i - z_j|. \end{aligned} \qquad (1)$$

This model has a simple interpretation: For two actors j and k equidistant from i, the log odds ratio of $i \to j$ versus $i \to k$ is $\beta'(x_{i,j} - x_{i,k})$.

Note that the $|z_i - z_j|$'s could be replaced by an arbitrary set of distances $\{d_{i,j}\}$, satisfying the triangle inequality, $d_{i,j} \leq d_{i,k} + d_{k,j}\ \forall \{i, j, k\}$. A semiparametric modeling approach would impose no further constraints on the distances, and so the parameter space would include $\binom{n}{2}$ distances to estimate, subject to the inequality constraints. Generally, we prefer to model the $d_{i,j}$'s as distances between actors in some low-dimensional Euclidean space for reasons of parsimony and ease of model interpretability.

The latent position model is inherently reciprocal and transitive: If $i \to j$ and $j \to k$, then $d_{i,j}$ and $d_{j,k}$ are probably not too large, making the events $j \to i$ (reciprocity) and $i \to k$ (transitivity) more probable. One interesting feature of the model is that it provides an essentially perfect model fit for many social network datasets with undirected relations, in a parameter space of much lower dimension than that of the data $[nk$ vs. $n(n-1)]$. To explore this feature further, consider the following reparameterization of (1) in the case of no covariate information and an undirected relation $y_{i,j} = y_{j,i}$:

$$\log\text{odds}(y_{i,j} = 1|d_{i,j}, x_{i,j}, \alpha) = \alpha(1 - d_{i,j}). \qquad (2)$$

We say a set of distances $\{d_{i,j}\}$ *represents* the network Y if

$$\{d_{i,j} > 1\ \forall i, j : y_{i,j} = 0\}$$

and

$$\{d_{i,j} < 1\ \forall i, j : y_{i,j} = 1\}. \qquad (3)$$

For such a set of distances, the probability of the data under parameterization (2) will converge to unity as $\alpha \to \infty$. As we model the distances as being Euclidean distances in some k-dimensional space, we say that a network is d_k *representable* if there exist points $z_i \in \Re^k$ such that the distances $d_{i,j} = |z_i - z_j|$ satisfy (3). In such a space, d_k representability is equivalent to being able to find a set of points for the actors such that $i \sim j$ if and only if i and j lie within k-dimensional unit balls centered around each other.

It is interesting to note there are many examples of social networks which are d_k representable for k much smaller than

n, and even for $k=2$. Consider, for example, an *n-star network* composed of one central actor having ties to $n-1$ otherwise unconnected actors. Such a network is trivially $d_{\frac{n}{2}-1}$ representable for any n, by positioning pairs of noncentral actors on either sides of the central actor along one of the $n/2$ coordinate axes. As another example, consider an *n-chain network*, in which there is an ordering of n actors so that $1 \sim 2 \sim 3 \sim \cdots \sim n \sim 1$. This network is d_2 representable for all n by placing the actors equidistant from the origin but separated by equal angles. Such results suggest that distance-based models may provide a good method of data reduction and presentation for undirected relational data. Although the foregoing examples may seem contrived, in Section 4.2 we analyze a real-life 15-actor network that is d_2 representable.

2.2 Projection Models

The distance model presented earlier is inherently symmetric in that $p(i \to j) = p(j \to i)$. However, in many networks such symmetry is not achieved. For example, perhaps actor i sends a large number of ties, whereas j sends ties to a small subset of the actors receiving ties from i. In this case we want to model that i and j are "similar" but i is more "socially active." Such a model could be achieved by including actor-specific activity parameters, an approach used by Wang and Wong (1987) to allow for actor-level variability in their stochastic blockmodel.

Alternatively, variable levels of activity can be modeled parsimoniously in the context of a latent position model which allows for probabilistic transitivity in the relations, as well as individual-specific levels of social activity. Suppose each actor i has an associated unit-length k-dimensional vector of characteristics v_i. These characteristics can be thought of as points on a k-dimensional sphere of unit radius. We might imagine that i and j are prone to having ties if the angle between them is small, neutral to having ties if the angle is a right angle, and averse to ties if the angle is obtuse. These three situations correspond to $v_i'v_j > 0$, $v_i'v_j = 0$, and $v_i'v_j < 0$, respectively. In other words, i and j are more likely to have a tie if the characteristics of i and j are in the same direction, and less likely to have a tie if they have characteristics in opposite directions. Adding a parameter for each node to allow for different levels of activity is equivalent to having latent vectors of various lengths: Letting $a_i > 0$ be the activity level of actor i, we can model the probability of a tie from i to j as depending on the magnitude of $a_i v_i' v_j$ or, equivalently, $z_i'z_j/|z_j|$, where $z_i = a_i v_i$. This is the signed magnitude of the projection of z_i in the direction of z_j and can be thought of the extent to which i and j share characteristics, multiplied by the activity level of i. For convenience, we parameterize the probability of a tie from i to j using the logistic regression model as before,

$$\log \operatorname{odds}(y_{i,j} = 1 | z_i, z_j, x_{i,j}, \alpha, \beta) = \alpha + \beta' x_{i,j} + \frac{z_i' z_j}{|z_j|}.$$

In some situations we may wish to model differential rates of accepting ties. In this case, the foregoing probability could depend on the latent vectors through $z_i'z_j/|z_i|$.

3. ESTIMATION

In contrast to the p^* and Markov random graph models, the log-likelihood of a conditional independence model is relatively simple,

$$\log P(Y|\eta) = \sum_{i \neq j} \{\eta_{i,j} y_{i,j} - \log(1 + e^{\eta_{i,j}})\}, \tag{4}$$

where η is a function of parameters, unknown positions, and perhaps known explanatory variables. As such, likelihood-based estimation methods, such as maximum-likelihood and Bayesian inference, are feasible.

The likelihood (4) is strictly concave in the matrix $\eta = \{\eta_{i,j}\}$. Consider first the semiparametric model $\eta = \alpha 1 1' - D$, where D is constrained only to be a positive symmetric matrix of values satisfying the triangle inequality. As the parameter space $\{\alpha, D\}$ is convex and $\eta(\alpha, D)$ is affine, there is a unique value of $\alpha 1 1' - D$ maximizing the likelihood. (Note, however, α is confounded with D, as the addition of a positive constant to a set of distances gives a set of distances.) Unfortunately, the log-likelihood is *not* generally concave in $\{\alpha, Z\}$ for either the distance model or the projection model, because the function $\eta = \eta(\alpha, Z)$ is not affine. This makes identification of a global maximum likelihood estimator (MLE) problematic. However, one approach is to first identify a set of distances, not necessarily Euclidean, that maximize the likelihood (a convex minimization problem). A set of positions in $\Re^k$ approximating the distances can then be found using multidimensional scaling methods. This set of positions can be used as a starting point in a nonlinear optimization routine. A simpler approach that works well in the examples in this article is to obtain a set of dissimilarities between nodes based on an ad hoc measure, such as the Euclidean distances between rows or columns of the sociomatrix or the geodesic distance (path length) between the nodes (Wasserman and Faust 1994). Starting values for the positions can then be found using multidimensional scaling.

Distances between a set of points in Euclidean space are invariant under rotation, reflection, and translation. Therefore, for each $k \times n$ matrix of latent positions Z, there is an infinite number of other positions giving the same log-likelihood. More specifically, $\log \Pr(Y|Z, \alpha) = \log \Pr(Y|Z^*, \alpha)$ for any Z^* that is equivalent to Z under the operations of reflection, rotation, or translation. A confidence region that includes two equivalent positions Z_1 and Z_2 is in a sense overestimating the variability in the unknown positions (although not overestimating the variability in distances or relative positions, because these are identical for Z_1 and Z_2). Fortunately, this problem can be resolved by basing inference on equivalence classes of latent positions: Let $[Z]$ be the class of positions equivalent to Z under rotation, reflection, and translation. For each $[Z]$, there is one set of distances between the nodes. We call this class of positions a *configuration*.

We make inference on configurations via inference on particular elements of configurations that are comparable across configurations. For a given configuration $[Z]$, we select for inference $Z^* = \arg\min_{TZ} \operatorname{tr}(Z_0 - TZ)'(Z_0 - TZ)$, where Z_0 is a fixed set of positions and T ranges over the set of rotations, reflections, and translations. Z^* is a "Procrustean" transformation of Z, being the element of $[Z]$ closest to Z_0 in terms of the sum of squared positional difference and is unique if Z_0Z' is nonsingular (Sibson 1979). Z^* is relatively easy to compute: Assuming Z and Z_0 are both centered at the origin, Z^* is given by $Z^* = Z_0Z'(ZZ_0'Z_0Z')^{-1/2}Z$. We typically take $Z_0 = \widehat{Z}$, an MLE of the latent positions centered at the origin.

Given prior information on α, β, and Z, our procedure for sampling from the posterior distribution is as follows:

1. Identify an MLE $\widehat{Z}$ of Z, centered at the origin, by direct maximization of the likelihood.
2. Using $Z_0 = \widehat{Z}$ as a starting value, construct a Markov chain over model parameters as follows:
   a. Sample a proposal $\check{Z}$ from $J(Z|Z_k)$, a symmetric proposal distribution.
   b. Accept $\check{Z}$ as Z_{k+1} with probability $\frac{p(Y|\check{Z},\alpha_k,\beta_k,X)}{p(Y|Z_k,\alpha_k,\beta_k,X)} \frac{\pi(\check{Z})}{\pi(Z_k)}$; otherwise, set $Z_{k+1} = Z_k$.
   c. Store $\widetilde{Z}_{k+1} = \arg\min_{TZ_{k+1}} \text{tr}(\widehat{Z}-TZ_{k+1})'(\widehat{Z}-TZ_{k+1})$.
3. Update α and β with a Metropolis–Hastings algorithm.

Because each configuration can be represented by its unique Procrustean statistic, the posterior distribution of the configuration around $\widehat{Z}$ is represented by samples of $\widetilde{Z}$ from the Markov chain.

The computational details for the projection model are the same as before, except that the likelihood is invariant under rotation and reflection of positions, but not under translation. Therefore, the only modification to the foregoing is to let $\widetilde{Z}_{k+1} = \arg\min_{TZ_{k+1}} \text{tr}(\widehat{Z} - TZ_{k+1})'(\widehat{Z} - TZ_{k+1})$, where T ranges over the set of rotations and reflections.

4. EXAMPLES

Here we analyze three standard datasets from the social networks literature: Sampson's (1968) Monk data, Padgett and Ansell's (1993) data on marriage relations between Florentine families, and Hansell's (1984) classroom data.

4.1 Monk Data

Sampson (1968) collected data on a variety of interpersonal relations among 18 monks. Of particular interest has been the data on positive affect relations, in which each monk was asked whether he had positive relations to each of the other monks. Based on the network and other data, Sampson originally classified each monk as belonging to one of four groups: the *loyal opposition* (monks 2–6), the *young Turks* (monks 8–14), the *outcasts* (monks 16–18), or the *waverers* (monks 1, 7, and 15). Subsequent data analyses have placed monks 1 and 7 with the loyal opposition and monk 15 with the outcasts.

These data are standard in the social network analysis literature, having been modeled by Holland and Leinhardt (1981), Reitz (1982), Holland, Laskey, and Leinhardt (1983), and Fienberg et al. (1981). Wang and Wong (1987) extended these models by allowing for individual-level variation in relations as well as group-level preferences for ties, and obtained a substantially improved fit. Specifically, their stochastic blockmodel modeled each pair $\{y_{i,j}, y_{j,i}\}$ as depending on parameters for actor-specific rates of sending and receiving ties, a parameter representing mutuality of ties, and a parameter representing the preference of actors to send ties to members of their own group. Note that Wang and Wong took the group membership information as given, even though it was derived to some extent from the data.

The relations between the monks are somewhat transitive. The number of nonvacuously transitive ordered triples ($i \to j$, $j \to k$, $i \to k$) is 49. In 500 random reallocations of ties, holding the number of ties sent by each actor constant, the largest number of nonvacuously transitive triads was 35. The distance model takes advantage of this transitivity and can achieve a better fit than Wang and Wong's model, using fewer parameters and not presuming the a priori existence of distinct groups. Our model is the distance model presented in Section 2.1,

$$P(Y|\alpha, Z) = \prod_{i \neq j}^{n} p(y_{i,j}|\alpha, z_i, z_j)$$

and

$$\text{logit}\ p(y_{i,j} = 1|\alpha, z_i, z_j) = \alpha - |z_i - z_j|, \tag{5}$$

where the z_i's lie in $\Re^2$. Note that the probability of the data depends only on the distances, which are invariant under reflection, rotation, and location shift. As a result, 3 of the 18×2 model parameters can be fixed, so this model has $33 + 1 = 34$ parameters (including α).

The distance between each pair of nodes was first calculated as the average of the two directed path lengths between each pair. Crude estimates of latent positions were then found using multidimensional scaling, and the results were used as starting values for the nonlinear minimizer optim in the R statistical programming environment. Random sampling of starting values from a normal distribution produced identical results.

As shown in Table 1, the maximized log-likelihood is -66.02 with 34 parameters, compared with the maximized log-likelihood of the stochastic blockmodel fit of -82.12 with 37 parameters (Wang and Wong 1987). The improvement of the position-based model over the stochastic blockmodel of Wang and Wong suggests that because relationships are indeed transitive to some extent, modeling them as such leads to an improvement in model fit. The maximum likelihood estimates of monk positions from the distance model are shown in Figure 1(a).

The conditional independence model lends itself relatively easily to a Bayesian analysis. Priors can be formulated for α and Z, and posterior inference can be made about each. In particular, this provides a means of making confidence regions for the positions of the actors in social space. Using diffuse independent normal priors for α and Z with mean 0 and standard deviation 100, we performed a Bayesian analysis via 2.5×10^6 scans from a Markov chain as described in Section 3. The chain mixes reasonably quickly in the z_i's and the pairwise

Table 1. Model Fitting Results for the Monk Data

Model	*Maximized log-likelihood*	*Number of parameters*
Distance model ($\Re^3$)	−34.04	50
Distance model ($\Re^2$)	−66.02	34
Stochastic blockmodel	−82.12	37

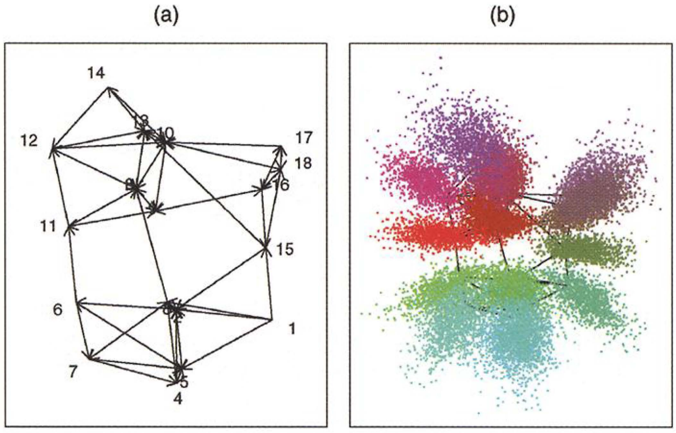

Figure 1. Maximum Likelihood Estimates (a) and Bayesian Marginal Posterior Distributions (b) for Monk Positions. The direction of a relation is indicated by an arrow.

distances between nodes, but quite slowly in α, as shown in Figure 2(b). Output from the chain was saved every 2,000 scans, and positions of the different monks are plotted for each saved scan in Figure 1(b) (the plotting color for each monk is based on their mean angle from the positive x-axis and their mean distance from the origin). The categorization of the monks given at the beginning of this section is validated by the distance model fitting, as there is little between-group overlap in the posterior distribution of monk positions. Additionally, this model is able to quantify the extent to which some actors (such as monk 15) lie between other groups of actors.

The extent to which model fit can be improved by increasing the dimension of the latent space was examined by fitting the distance model in $\Re^3$, that is, $z_i \in \Re^3$ for $i = 1, \ldots, n$. The maximum likelihood for this model is -34.04 in 50 parameters, a substantial improvement over the fit in $\Re^2$ at a cost of 16 additional parameters. It is interesting to note that the fit cannot be improved by going into higher dimensions. This can be seen as follows. For a given dataset Y, the best-fitting symmetric model ($p_{i,j} = p_{j,i}$) has the property that $p_{i,j} = 1$ for $y_{i,j} = y_{j,i} = 1$, $p_{i,j} = 0$ for $y_{i,j} = y_{j,i} = 0$, and $p_{i,j} = 1/2$ for $y_{i,j} \neq y_{j,i}$. The log-likelihood of such a fit is thus $-a \log 4$, where a is the number of asymmetric dyads. For the monk

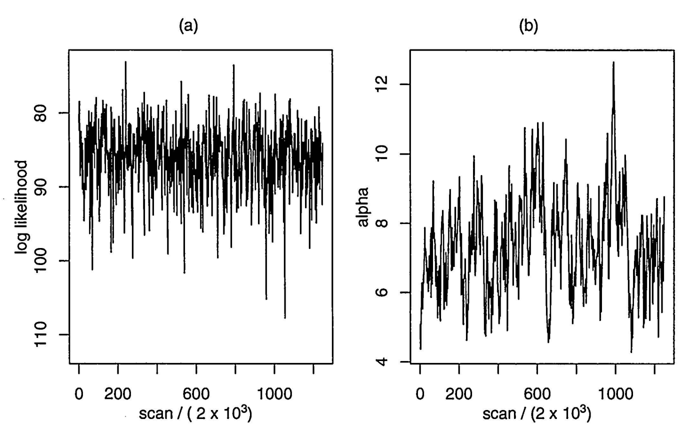

Figure 2. MCMC Diagnostics for the Monk Analysis. (a) Log-likelihood; (b) alpha.

dataset, the number of asymmetric dyads is 26, and so the maximum possible log-likelihood under a symmetric model is $-26 \log 4 = -36.04$, which is achieved by the distance model in $\Re^3$. More precisely, there exists a set of positions $\hat{z}_i \in \Re^3$ and a rate parameter $\hat{\alpha}$ such that $\lim_{c \to \infty} \log P(Y|c\hat{\alpha}, c\widehat{Z}) = -26 \log 4$.

4.2 Florentine Families

Padgett and Ansell (1993) compiled data on marriage and business relations between 16 historically prominent Florentine families, using a history of this period given by Kent (1978). We analyze data on the marriage relations taking place during the 15th century. The actors in the population are families, and a tie is present between two families if there is at least one marriage between them. This is an undirected relation, as the respective families of the husband and wife in each marriage were not recorded. One of the 16 families had no marriage ties to the others, and was consequently dropped from the analysis. If included, this family would have infinite distance from the others in a maximum likelihood estimation and a large but finite distance in a Bayesian analysis, as determined by the prior.

Modeling $d_{i,j} = |z_i - z_j|$, $z_i, z_j \in \Re^2$ and using the parameterization $\eta_{i,j} = \alpha(1 - d_{i,j})$ as described in Section 2, the likelihood of (α, Z) can be made arbitrarily close to 1 as $\alpha \to \infty$ for fixed $Z = \widehat{Z}$; that is, the data are d_2 representable. Such a representing $\widehat{Z}$ is plotted in Figure 3(a). Family 9 is the Medicis, whose average distance to others is greater only than that of families 13, the Ridolfis and 16, the Tornabuonis. Another d_2 representation is given in Figure 3(b). This configuration is similar in structure to the first, except that the segments 9-1 and 9-14-10 have been rotated. This is somewhat of an artifact of our choice of dimension: When modeled in three dimensions, 1 and 14 are fit as being relatively equidistant from 6.

One drawback of the MLEs presented earlier is that they overfit the data in a sense, as the fitted probabilities of ties are all either 0 or 1 (or nearly so, for very large α). Alternatively, a prior for α can be formulated to keep predictive probabilities more in line with our beliefs; for example, that the probability of a tie rarely goes below some small, but not infinitesimal value. Using the MCMC procedure outlined in Section 3, the marriage data were analyzed using an exponential prior with mean 2 for α and diffuse independent normal priors for the components of Z (mean 0, standard deviation 100). The MCMC algorithm was run for 5×10^6 scans, with the output saved every 5,000 scans. This chain mixes faster than that of the monk example, as can be seen in the diagnostic plots of Figure 4 and in plots of pairwise distances between nodes (not shown). Marginal confidence regions are represented by plotting samples of positions from the Markov chain, shown in Figure 3(c). Note that the confidence regions include both the configurations given in Figure 3(a) and (b). Actors 14 and 10 (in red and purple) are above or below actor 1 (in green) for any particular sample; the observed overlap of these actors in the figure is due to the bimodality of the posterior and that the plot gives the *marginal* posterior distributions of each actor.

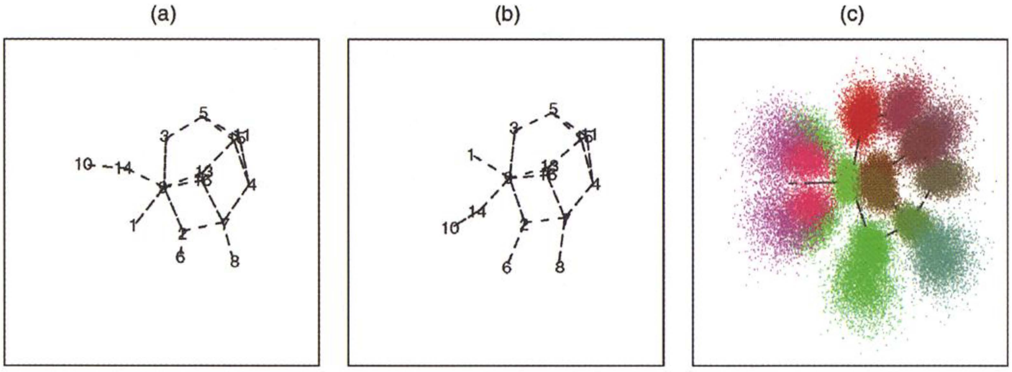

Figure 3. (a) and (b) Alternate d_2 Representations of the Florentine Family Data. (c) Marginal posterior distributions of family positions.

4.3 Classroom Data

Hansell's (1984) data measure the existence of strong friendship ties between 13 boys and 14 girls in a sixth-grade classroom. Each student was asked whether he or she liked each other student "a lot," "some," or "not much." A strong friendship tie is considered present if a student likes another student "a lot."

The number of ties sent by each student varies considerably, ranging from 0 to 19 with a mean of 5.8 and a standard deviation of 4.7. (The standard deviation of the number of ties received was 3.2.) For this reason, we choose to analyze the data using the projection model described in Section 2.2, which allows for a variable rate in sending ties across students. Additionally, 72% of the ties are same sex, indicating that the friendship relation is more prevalent within sex. Finally, the relations are transitive, in that the number of nonvacuously transitive ordered triples is 400, compared to a maximum of 347 in 500 random reallocations of ties, holding constant the number of ties sent by each student.

To illustrate the features of the projection model, we fit models both with and without covariate information on the sex of the students, that is, we consider both of the following formulations:

- Projection model, no covariate: $\mathrm{logit}(p_{i,j}) = \alpha + z_i' z_j / |z_j|$
- Projection model, one covariate: $\mathrm{logit}(p_{i,j}) = \alpha + \beta x_{i,j} + z_i' z_j / |z_j|$.

The covariate $x_{i,j}$ is the indicator of actors i and j being of the same sex. We also compare these models to the stochastic blockmodel fit of Wang and Wong (1987).

We first obtained distance estimates for both models by calculating the average of the directed path lengths between each pair. Crude positions in a single dimension were found using Sammon's (1969) nonlinear mapping. These positions were converted into positions on a circle, which became the starting values of the latent vectors in the optimization routine. Randomly sampled starting values give the same optimum fit (Table 2). The projection model with sex as a covariate gives the best fit, with the coefficient β being nominally significant based on a likelihood ratio test.

Fitting the model without the covariate information on sex gives the estimates of positions shown in Figure 5(a). Here the students are plotted along the circumference of a circle according to the angle of their latent vector, and the size of the

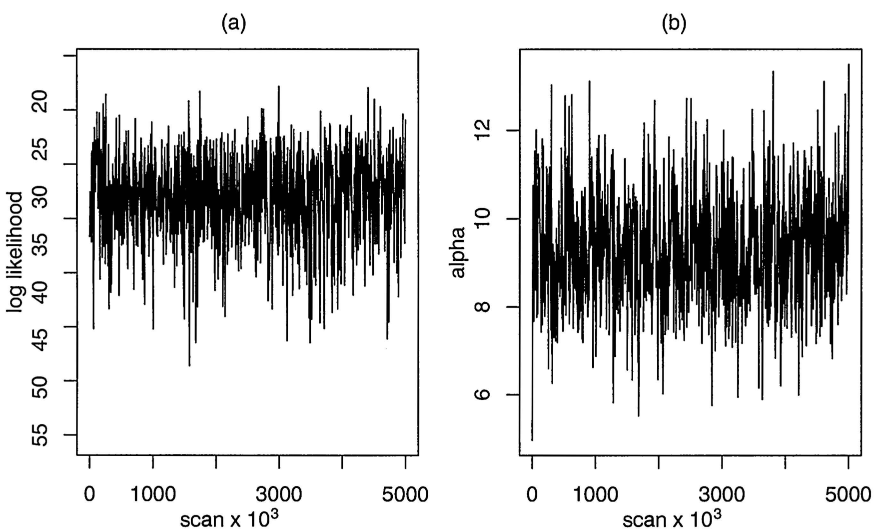

Figure 4. MCMC Diagnostics for the Florentine Family Analysis. (a) Log-likelihood; (b) alpha.

Table 2. Model-Fitting Results for Classroom Data

Model	Maximized log-likelihood	Number of parameters
Projection, with covariate	−224.58	55
Stochastic blockmodel	−227.57	55
Projection, no covariate	−229.05	54

plotting character for a student is increasing in the magnitude of their vector. The model identifies two somewhat orthogonal groups of actors falling on vectors emanating from the origin, one consisting of mostly boys (□); the other, girls (○). The difference between the boys' and girls' median angles, plotted in dashed lines, is 76 degrees.

Note that if the sexes were separated by 180 degrees, then based on the model, it would be improbable for actors to have ties to both boys and girls, which is not completely uncommon in the data. By having the group vectors separated by 76 degrees, the model predicts ties between the sexes as rare, although it allows for a nonnegligible probability of some actors sending ties to both groups or even sending ties primarily to members of the opposite group.

A further application of the projection model is as a means of identifying boys and girls who may be in similar social groups, after having accounted for the fact that the frequency of between-sex friendship ties is low. The estimated positions after having partially accounted for this known covariate structure are shown in Figure 5(b). Note there is still considerable separation of the sexes, although the difference in median angles has been reduced to 60 degrees. This suggests that the single covariate $x_{i,j}$ does not fully explain the different rates of within-sex and between-sex friendship ties. A "full" model would have different baseline rates for the four different types of ties (boy → boy, boy → girl, girl → girl, and girl → boy). Indeed, inclusion of these parameters reduces the median angle between the sexes to 13 degrees. We present only the model with the single covariate, as this data analysis is meant primarily as an illustrative example.

The foregoing model could also be used as a means of making inference on the preference for within-sex friendship ties. A naive approach to inference would be to treat each possible tie as a Bernoulli random variable, independent of the other ties. Using logistic regression, we would estimate the log-odds ratio of a between-sex pair being friends compared with that of a within-sex pair being friends as 1.3, with a standard error of .2. Of course, we would expect a confidence interval based on such an analysis to be too small, because ties between individuals are not independent, unconditional on the latent positions. As an alternative, we performed a Bayesian analysis as outlined in Section 3. We constructed a Markov chain of length 5×10^6 scans, starting at the MLE. Output was saved every 1,000 scans, which we then used to make marginal posterior inference on β. The marginal posterior density of β is given in Figure 5(c), in which the solid vertical line represents the MLE from the projection model and the dashed lines represent the MLE plus and minus two standard errors, based on an ordinary logistic regression. As we would expect, a 95% confidence region from the Bayesian analysis would be longer than the one based on the ordinary logistic regression.

5. DISCUSSION

This article proposes a new model for social networks based on spatial representation for which maximum likelihood and Bayesian inference are practical to implement. The approach has some advantages over existing social network models and inferential procedures. First, the proposed method provides a visual and interpretable model-based spatial representation of network relationships. Second, it improves on existing methods by allowing the statistical uncertainty in the social space to be quantified and graphically represented. Third, it is flexible and can be easily generalized to allow for multiple relationships, ties with varying strengths (using generalized linear models), and time-varying relations (by modeling the latent positions as stochastic processes). Fourth, it deals easily with missing data, at least if information on ties is missing at

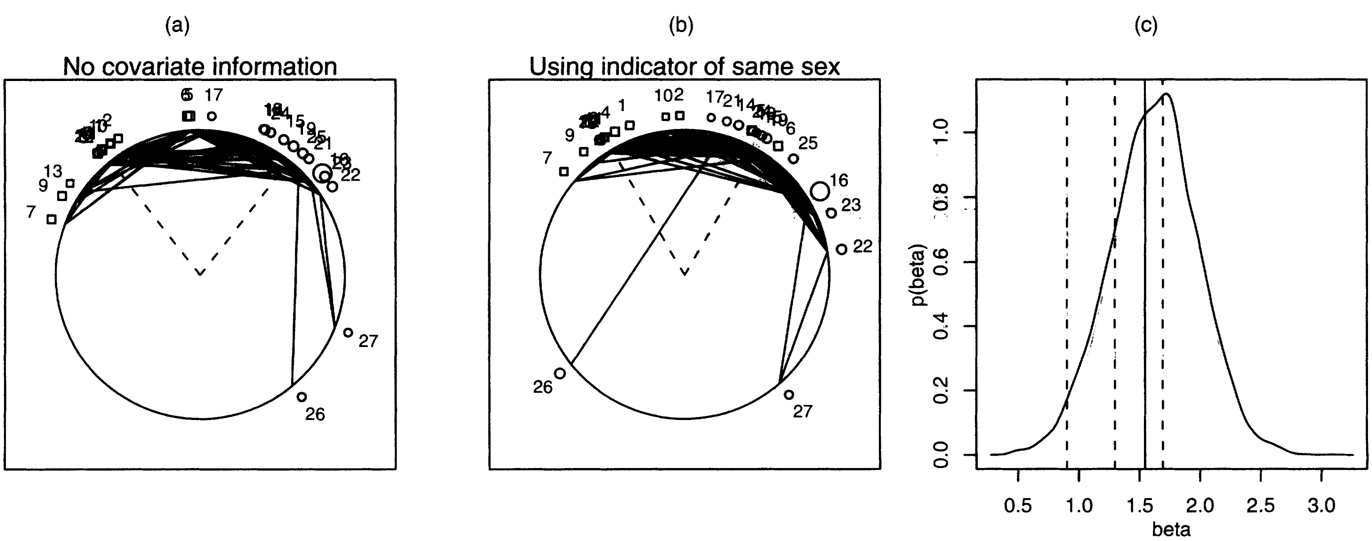

Figure 5. Maximum Likelihood Estimates of Student Positions (a) No covariate information; (b) using indicator of same sex, and (c) the posterior of β.

random: The likelihood includes only terms corresponding to observed ties. Finally, the model is inherently transitive, and so we can expect an improved fit over models lacking such structure (such as the stochastic blockmodel) when the relations are transitive in nature.

It may be desirable to allow for further dependence in the model. For example, the data might exhibit more or less reciprocity ($y_{i,j} = y_{j,i}$) than is estimated by the model. In such a case, the model can be extended by treating each dyad $(y_{i,j}, y_{j,i})$ as independent of other dyads, given positions Z and parameters α and γ, where γ models the dependence between ties within a dyad. In the absence of such a model, some aspects of potential lack of fit can be checked. For example, using the fitted probabilities of ties and assuming the conditional independence model, we can calculate the expected number of reciprocal, nonreciprocal, and null-reciprocal dyads in a dataset and compare these with the observed numbers. For the monk data using the distance model in $\Re^2$, the expected counts are $(17.3, 21.4, 114.3)$, and the observed counts are $(15, 26, 112)$. For the classroom data, the model with one covariate gives expected counts of $(26.7, 103.6, 220.7)$, whereas the observed counts are $(24, 109, 218)$, indicating a reasonable fit according to this criteria (although both datasets exhibit slightly less reciprocity than suggested by the estimated models).

The nonconcavity of the log-likelihood as a function of the latent positions presents difficulties in finding MLEs of the parameters. The likelihood surface needs to be carefully examined to differentiate between global and local maxima. MCMC methods can be a useful tool in this regard. For example, in the analysis of the Florentine family data, the MCMC procedure was able to identify and mix over two separate maxima. Additionally, we reiterate that the log-likelihood is concave in (α, D), where D is the set of distances not constrained to be Euclidean. One could potentially do model fitting in this high-dimensional space and then examine (graphically or otherwise) pairs or subnetworks in lower dimensions.

We have not discussed in detail the choice of a prior distribution for latent positions in this article. Although simple, the diffuse independent normal priors presented in the examples may not accurately represent prior beliefs about the structure of social networks. More appropriate might be clustered point processes or mixtures of normals with an unknown number of components. Such priors could allow one to incorporate prior information on tendencies for clustering, without specifying cluster membership. This would add another level of hierarchy to the analysis, although the resulting model would be more flexible and perhaps more accurately represent any tendencies of populations to form segregating groups.

As an alternative to the models presented in this article, multiple-dimensional scaling (MDS) is a widely used method of representing the spatial structure of a social network (Breiger, Boorman, and Arabie 1975; Faust and Romney 1985). In this context, MDS is a class of methods that can be used to produce a spatial representation of individuals based on similarity or dissimilarity measures between pairs of individuals. Such applications of MDS differ from the models presented here in that MDS is used primarily as a data-analytic means of visualizing given dissimilarities, whereas our method is a model-based representation of the measured relations and latent positions, although recently DeSarbo, Kim, and Fong (1999) and Oh and Raftery (2001) developed model-based MDS applicable to two-mode networks within a Bayesian framework. Our model has a number of advantages over MDS. First, our method directly models the response, whereas the usual choices for dissimilarities in MDS are ad hoc and do not reflect the stochastic nature of the sociomatrix. Second, current versions of MDS use maximum likelihood or other optimization methods over large numbers of parameters (e.g., linear in the number of individuals). The asymptotic properties of these methods are largely unknown, and the uncertainty in the latent positions is difficult to quantify. To avoid this, some versions of MDS assume that individuals can be grouped into homogeneous clusters—so-called latent class MDS (Lazarsfeld and Henry 1968; DeSarbo et al. 1994). However, individual-specific variability in relative position is often the primary focus in the social network context—something that can be quantified in an interpretable way via a Bayesian analysis of one of the position-based models discussed in this article.

R-code for implementing the proposed methods will be available through the first author's website, *www.stat.washington.edu/hoff*.

[Received November 2001. Revised April 2002.]

REFERENCES

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

——— (2000), "Markov Chain Monte Carlo for Statistical Inference," working paper, University of Washington, Center for Statistics and the Social Sciences.

Breiger, R. L., Boorman, S., and Arabie, P. (1975), "An Algorithm for Clustering Relational Data With Applications to Social Network Analysis and Comparison With Multidimensional Scaling." *Journal of Mathematical Psychology*, 12, 328–382.

Corander, J., Dahmström, K., and Dahmström, P. (1998), "Maximum Likelihood Estimation for Markov Graphs," research report, University of Stockholm, Dept. of Statistics.

Crouch, B., Wasserman, S. W., and Trachtenberg, F. (1998), "Markov Chain Monte Carlo Maximum Likelihood Estimation for p^* Social Network Models," presented at the XVIII International Sunbelt Social Network Conference, Sitga, Spain.

DeSarbo, W. S., Kim, Y., and Fong, D. (1999), "A Bayesian Multidimensional Scaling Procedure for the Spatial Analysis of Revealed Choice Data," *Journal of Econometrics*, 89, 79–108.

Faust, K. (1988), "Comparison of Methods for Positional Analysis: Structural and General Equivalence," *Social Networks*, 10, 313–341.

Faust, K., and Romney, A. K. (1985), "Does Structure Find Structure?: A Critique of Burt's Use of Distance As a Measure of Structural Equivalence," *Social Networks*, 7, 77–103.

Fienberg, S. E., Meyer, M. M., and Wasserman, S. S. (1981), "Analysing Data From Multivariate Directed Graphs: An Application to Social Networks," in *Interpreting Multivariate Data*, New York: Wiley, pp. 289–306.

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

Handcock, M. S. (2000), "Progress in Statistical Modeling of Drug User and Sexual Networks," unpublished manuscript, University of Washington, Center for Statistics and the Social Sciences.

Hansell, S. (1984), "Cooperative Groups, Weak Ties, and the Integration of Peer Friendships," *Social Psychology Quarterly*, 47, 316–328.

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

Holland, P. W., and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs" (with discussion), *Journal of the American Statistical Association*, 76, 33–65.

Kent, D. (1978), *The Rise of the Medici: Faction in Florence, 1426–1434*, Oxford, UK: Oxford University Press.

Lazarsfeld, P. F., and Henry, N. W. (1968), *Latent Structure Analysis*, Boston: Houghton Mifflin.

McFarland, D., and Brown, D. (1973), "Social Distance as Metric: A Systematic Introduction to Smallest Space Analysis," in *Bonds of Pluralism: The Form and Substance of Urban Social Networks*, ed. E. Laumann, New York: Wiley, pp. 213–253.

Moreno, J. L. (1934), "Who Shall Survive? A New Approach to the Problem of Human Interrelations," in *Psychological Abstracts*, Vol. 8, Washington, DC: Nervous and Mental Disease Publishing, p. 5153.

Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and Prediction for Stochastic Block Structures," *Journal of the American Statistical Association*, 96, 1077–1087.

Oh, M. S., and Raftery, A. E. (2001), "Bayesian Multidimensional Scaling and Choice of Dimension," *Journal of the American Statistical Association*, 96, 1031–1044.

Padgett, J. F., and Ansell, C. K. (1993), "Robust Action and the Rise of the Medici," *American Journal of Sociology*, 98, 1259–1319.

Reitz, K. P. (1982), "Using Log-Linear Analysis With Network Data: Another Look at Sampson's Monastery," *Social Networks*, 4, 243–256.

Ruelle, D. (1969), *Statistical Mechanics*, New York: Wiley.

Sammon, J. W. (1969), "A Non-Linear Mapping for Data Structure Analysis," *IEEE Transactions on Computers*, C-18, 401–409.

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

Sibson, R. (1979), "Studies in the Robustness of Multidimensional Scaling: Perturbational Analysis of Classical Scaling," *Journal of the Royal Statistical Society*, Ser. B, 41, 217–229.

Snijders, T. A. B. (2002), "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure*, (Vol. 3, No. 2, April 19).

Strauss, D., and Ikeda, M. (1990), "Pseudolikelihood Estimation for Social Networks," *Journal of the American Statistical Association*, 85, 204–212.

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

Wasserman, S., and Faust, K. (1994), *Social Network Analysis: Methods and Applications,* Cambridge, UK: Cambridge University Press.

Wasserman, S., and Pattison, P. (1996), "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p^*," *Psychometrika*, 61, 401–425.