Random Effects Models for Network Data

Peter D. Hoff[1]
Working Paper no. 28
Center for Statistics and the Social Sciences
University of Washington
Seattle, WA 98195-4320

January 14, 2003

[1]Department of Statistics, University of Washington, Seattle, Washington. Email:hoff@stat.washington.edu, Web:www.stat.washington.edu/hoff. This research was supported by Office of Naval Research grant N00014-02-1-1011.

Abstract

One impediment to the statistical analysis of network data has been the difficulty in modeling the dependence among the observations. In the very simple case of binary (0-1) network data, some researchers have parameterized network dependence in terms of exponential family representations. Accurate parameter estimation for such models is quite difficult, and the most commonly used models often display a significant lack of fit. Additionally, such models are generally limited to binary data. In contrast, random effects models have been a widely successful tool in capturing statistical dependence for a variety of data types, and allow for prediction, imputation, and hypothesis testing within a general regression context. We propose novel random effects structures to capture network dependence, which can also provide graphical representations of network structure and variability.

1 Network Dependence

Network data typically consist of a set of n nodes and a relational tie $y_{i,j}$, measured on each ordered pair of nodes $i, j = 1, \ldots, n$. This framework has many applications, including the study of war, trade, the behavior of epidemics, the interconnectedness of the World Wide Web, and telephone calling patterns.

It is often of interest to relate each network response $y_{i,j}$ to a possibly pair-specific vector valued predictor variable $x_{i,j}$. A flexible framework for doing so is the generalized linear model (see, for example McCullagh and Nelder 1983), in which the expected value of the response is modeled as a function of a linear predictor $\beta' x_{i,j}$, where β is an unknown vector of regression coefficients to be estimated from the data. The ordinary regression model $E(y_{i,j}) = \beta' x_{i,j}$ is perhaps the most commonly used model of this type. A generalized linear model for binary (0-1) data is logistic regression, which relates the expectation of the response to the regression variable via the relation $g(E[y_{i,j}]) = \beta' x_{i,j}$, where $g(p) = \log \frac{p}{1-p}$.

As an example of the use of such statistical models, consider the analysis of strong friendship ties among 13 boys and 14 girls in a sixth-grade classroom, as collected by Hansell (1984). Each student was asked if they 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." Also recorded is the sex of each student. The data, presented in Figure 1, suggest a general preference for same-sex friendship ties. Of potential interest is statistical estimation of this preference, as well as a confidence interval for its value. One approach for such statistical analysis would be to formulate the logistic regression model $g(E[y_{i,j}|x_{i,j}, \beta]) = \beta_0 + \beta_1 x_{i,j}$, where $x_{i,j}$ is one if children i and j are of the same sex, and zero otherwise, and $\beta = (\beta_0, \beta_1)$ are parameters to be estimated.

Estimation of regression coefficients β typically proceeds under the assumption that the observations are conditionally independent given β and the $x_{i,j}$'s. However, this assumption is often violated by many network datasets. For example, the data on friendship ties display several types of dependence:

Within-node dependence: 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). This node level variability suggests that responses from the same individual are positively dependent, in that the probability that $y_{i,j} = 1$ (i sends a tie to j), is high if we know $y_{i,k} = 1$ for lots of other nodes k, and lower if $y_{i,k}$ is mostly zero. More formally, we may wish to have a model in which $\Pr(y_{i,j} = 1 | y_{i,1}, \ldots, y_{i,j-1}, y_{i,j+1}, \ldots, y_{i,n})$ is an increasing function of $y_{i,k}, k \neq j$.

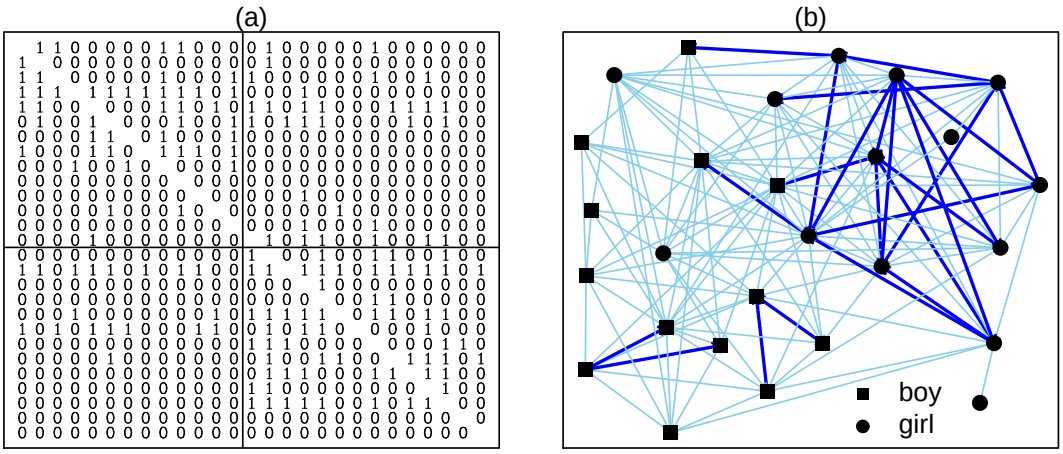

Figure 1: (a) Sociomatrix for friendship data: Rows and columns 1-13 are boys, 14-27 are girls; (b) Graphical representation of friendship data: Dark blue lines are reciprocated ties.

Reciprocity: For directed relations, it is often expected that $y_{i,j}$ and $y_{j,i}$ are positively dependent. The classroom data exhibit a sizable degree of reciprocity, in that the number of pairs in which $y_{i,j} = y_{j,i} = 1$ is 24, which is more than we would expect due to just random chance: In only 11 of 500 (2.2%) random permutations of the network data, holding constant the number of ties sent by each student, did the number of such reciprocal dyads exceed 24. The average number of reciprocal dyads in the 500 permutations was 17.2. This suggests an appropriate model would be one which allowed for positive dependence between $y_{i,j}$ and $y_{j,i}$.

Transitivity and Balance: In many situations we expect that two nodes with a positive relation will relate similarly to other nodes. For relations which are positive or negative, this has led to the concept of "balance" in which a positive value for $y_{i,j}$ implies $y_{i,k}$ and $y_{j,k}$ are likely to be of the same sign for other nodes k. A related concept is transitivity, in which a large value of $y_{i,j}$ together with a large value of $y_{j,k}$ implies a large value of $y_{i,k}$ (see Wasserman and Faust, 1994, Chapter 6).

The classroom data exhibit a large degree of transitivity, in that the number of non-vacuously transitive ordered triples (see Wasserman and Faust, page 244), is 400. In 500 random permutations of the network data, holding constant the number of ties sent by each student, the largest observed number of transitive triples was 347. This indicates the data exhibit significantly more transitivity than would be expected due to just random chance and node-level variability, and an appropriate model should allow for some form of transitive dependence.

In this article, we discuss statistical regression models which can describe such types of network dependence. This is done by incorporating random effects structures in a generalized linear model setting. We discuss parameter estimation for these models in a Bayesian framework, and provide example statistical analyses of the classroom data described above and of a dataset on alliances and conflict among New Guinean tribes.

2 Network Random Effects Models

Generalized linear models, or glm's, are ubiquitous tools which extend linear regression models to non-normal data and transformably additive covariate effects (McCullagh and Nelder, 1983). A standard glm assumes the expectation of the response variable $y_{i,j}$ can be written as a function of a linear predictor $\eta = \beta' x_{i,j}$. Assuming observations are conditionally independent given the $x_{i,j}$'s and β, the model is:

$$\begin{aligned} \Pr(y_{1,2}, \ldots, y_{n,n-1}|X, \beta) &= \prod_{i \neq j} p(y_{i,j}|x_{i,j}, \beta) \\ g(E[y_{i,j}|x_{i,j}, \beta]) &= \eta_{i,j} = \beta' x_{i,j}. \end{aligned}$$

Examples of glm's include ordinary linear regression, logistic regression, Poisson regression, and quasilikelihood methods.

As discussed above, one feature that distinguishes network data is the likely dependence among the $y_{i,j}$'s. This lack of independence makes standard glm models inappropriate. In other settings which involve dependent data, a common approach to parameter estimation has been the generalized linear mixed-effects model (McCulloch and Searle, 2002) in which it is assumed the network observations can be modeled as *conditionally* independent, given appropriate random effects terms which can be incorporated into the glm framework. The model above becomes

$$\begin{aligned} \Pr(y_{1,2}, \ldots, y_{n,n-1}|X, \beta, \gamma) &= \prod_{i \neq j} p(y_{i,j}|x_{i,j}, \beta, \gamma_{i,j}) \\ g(E[y_{i,j}|X, \beta, \gamma_{i,j}]) &= \eta_{i,j} = \beta' x_{i,j} + \gamma_{i,j}, \end{aligned} \tag{1}$$

where $\gamma_{i,j}$ is an unobserved random effect. The distribution of and dependence among the $\gamma_{i,j}$'s determines the dependence among the $y_{i,j}$'s. For many kinds of network data, we may wish to find a form for the $\gamma_{i,j}$'s that induces the kinds of dependence described above, such as within-node dependence, reciprocity, transitivity, and balance.

A simple approach to modeling the node variability that gives rise to within-node dependence is the use of random intercepts, that is, to let $\gamma_{i,j} = a_i + b_j + \epsilon_{i,j}$, where a_i and b_j

represent independently distributed sender- and receiver-specific effects. Such a distribution on the a_i's and b_j's induces a positive dependence among responses involving a common node. Typically, the distribution of these effects are taken to be normal distributions with means equal to zero, and variances to be estimated from the data.

Modeling other forms of network dependence is not as straightforward. In the case of binary logistic regression, Hoff, Raftery, and Handcock (2002) propose using a latent-variable approach as a means of modeling balance, transitivity, and reciprocity in network data. As applied to the glm above, such an approach presumes the error $\epsilon_{i,j}$ can be written as a function f of independent k-dimensional latent variables $z_i, z_j \in \mathbb{R}^k$ so that $\epsilon_{i,j} = f(z_i, z_j)$, $i, j = 1, \ldots, n$. The function f is chosen to be simple and to mimic the forms of network dependence described above. Incorporating both the random intercepts and the z_i's into the model, and assuming independent normal distributions, (1) becomes

$$\begin{aligned} \eta_{i,j} &= \beta' x_{i,j} + a_i + b_j + f(z_i, z_j) \\ a_1, \ldots, a_n &\sim \text{i.i.d. Normal}(0, \sigma_a^2) \\ b_1, \ldots, b_n &\sim \text{i.i.d. Normal}(0, \sigma_b^2) \\ z_1, \ldots, z_n &\sim \text{i.i.d. Normal}(0, I_k \times \sigma_z^2), \end{aligned}$$

where $\beta, \sigma_a^2, \sigma_b^2$, and σ_z^2 are parameters to be estimated, and I_k is the $k \times k$ identity matrix. Additionally, if the researcher is interested in local network structure, it may be desirable to estimate a_i, b_i, z_i for each node.

It remains to choose a suitable function f. One approach is to presume reciprocity, transitivity, and balance arise due to the existence of unobserved node characteristics, and that nodes relate preferentially to other nodes with similar values of those characteristics. This motivates letting f be a measure of "similarity" between the random effects z_i and z_j, which gives rise to a "latent position" interpretation as discussed in Hoff et al. (2002). For example, consider the following forms for f :

- (distance model) $f(z_i, z_j) = -|z_i - z_j|$;
- (inner product model) $f(z_i, z_j) = z_i' z_j$.

In the case of directed responses, each of the above functions induces a degree of reciprocity as $\epsilon_{i,j} = f(z_i, z_j) = f(z_j, z_i) = \epsilon_{j,i}$ due to the symmetry of f. The common error term induces a positive dependence between $y_{i,j}$ and $y_{j,i}$.

The above functions also give rise to higher-order dependence. For example, the distance model gives an error structure that is inherently transitive, since $|z_i - z_j| \leq |z_i - z_k| + |z_k - z_j|$

by the triangle inequality. The observation of strong ties from i to k and k to j suggests that $|z_i - z_k|$ and $|z_k - z_j|$ are small, and therefore $|z_i - z_j|$ cannot be too large and we might expect strong ties from i to j. The inner product model satisfies a similar but more complicated relation: in the special case that the vectors z_i are of unit length, $z_i'z_j \geq z_i'z_k + z_k'z_j - (1 + 2\sqrt{(1 - z_k'z_i)(1 - z_k'z_j)})$.

An undirected signed graph is said to be balanced if the product of the relations in all cycles is nonnegative, i.e. $y_{i_1,i_2} \times y_{i_2,i_3} \times \cdots \times y_{i_{k-1},i_k} \geq 0$ for all sequences of indices for which the corresponding data are available (Wasserman and Faust 1994, Chapter 6). As $f(z_i, z_j)$ exists in the model for each pair i, j, balance in terms of this random effect is equivalent to the balance of the complete graph formed by the sociomatrix with i, jth entry equal to $f(z_i, z_j)$. For a complete signed graph, all cycles are balanced if and only if each triad is balanced, i.e. $f(z_i, z_j) \times f(z_j, z_k) \times f(z_k, z_i) \geq 0$ for all triples i, j, k. Interestingly, this is satisfied by the inner product model in one dimension ($z_i \in \mathbb{R}$), as $(z_i'z_j) \times (z_j'z_k) \times (z_k'z_j) \geq 0$. For $z_i \in \mathbb{R}^k, k > 1$, these terms are not necessarily balanced, although they are "probabilistically" balanced in the following sense: if the directions of the z_i's are uniformly distributed, then the expected number of balanced triads exceeds the number of imbalanced triads, with the difference decreasing with increasing k. An additional feature of the inner product model is that if the directions of the $z_i's$ are uniformly distributed, then in general $E(z_i'z_j) = 0$. In particular, if each z_i is a vector of k independent normal random variables with mean 0 and variance σ_z^2, then $z_i'z_j$ will have mean 0 and variance $k\sigma_z^4$, furthering the interpretation of $z_i'z_j$ as an error term.

On the other hand,$-|z_i - z_j|$ is always negative, and so we lack this interpretation for the distance model. However, the distance model may be easier to interpret as a spatial representation of network structure: The z_i's can be interpreted as positions in a latent "social space," with nodes having strong ties to one another being estimated as close together, and subsets of nodes with strong within-group ties being estimated as clusters in this social space. Additionally, plotting estimates and confidence regions for the z_i's gives a graphical, model-based representation of the network data.

3 Parameter Estimation

Given network data $Y = \{y_{i,j}\}$ and possible regressor variables $X = \{x_{i,j}\}$, the goal is to make statistical inference on the unknown model parameters, which we generically denote as θ. The parameter θ may include the regression coefficients β, the variances of the random effects, and possibly the random effects themselves. We take a Bayesian approach to param-

eter estimation, in that we posit a (potentially diffuse) prior probability distribution $p(\theta)$, and base our inference on the posterior, or conditional distribution of the parameters given the information in the data, which is given by Bayes' rule, $p(\theta|Y) = p(Y|\theta) \times p(\theta)/p(Y)$. A closed form expression for the desired conditional distribution is generally unavailable, however we can make approximate random samples from this distribution using Markov chain Monte Carlo (MCMC) simulation (Gelfand and Smith 1990, Besag, Green, Higdon, and Mengersen 1995). MCMC-based inference constructs a dependent sequence of θ-values as follows: Given the lth-value θ_l in the sequence,

- sample a parameter value θ^* from a proposal distribution $J(\theta|\theta_l)$;
- compute the acceptance probability

$$r = \max\left(1, \frac{p(Y|\theta^*)p(\theta^*)J(\theta_l|\theta^*)}{p(Y|\theta)p(\theta)J(\theta^*|\theta_l)}\right);$$

- set $\theta_{l+1} = \theta^*$ with probability r, otherwise set $\theta_{l+1} = \theta_l$.

The particular details, such as the choice of the proposal distribution J, will depend on the model and the data. See Hoff et al. (2002) for MCMC algorithms designed specifically for such latent variable models.

The result of the algorithm is a sequence of θ values having a distribution that is approximately equal to the target distribution $p(\theta|Y)$. Statistical inference can be based on these samples. For example, a point estimate of θ is often taken to be the posterior mean, which is approximated by the average of the sampled θ-values. Posterior confidence intervals can be based on the sample quantiles.

4 Example Data Analyses

We now apply the methods described above to the statistical analysis of two example datasets. In the first example, we use the inner product model as a means of making inference on the preference for same sex friendship ties in Hansell's classroom data. In the second example, we use the distance model to make inference on the network of alliances among sixteen New Guinean tribes studied by Read (1954). Both datasets involve binary network data, although the methods are easily adapted to other types of network data via an appropriate generalized linear model.

4.1 Classroom Friendships

Hansell's (1984) data exhibit a tendency of children to form same sex friendship ties, in that 72% of the ties are same-sex. We consider a statistical analysis of this preference, in which we estimate the log odds of a same-sex tie, as well as make a confidence interval for its value. This is done via the logistic regression model with random effects described above,

$$g(E[y_{i,j}|\beta, x_{i,j}, \gamma_{i,j}]) = \beta_0 + \beta_1 x_{i,j} + \gamma_{i,j},$$

where $x_{i,j}$ is the indicator that i and j are of the same sex, $\beta = \{\beta_0, \beta_1\}$ are parameters to be estimated, and $\gamma_{i,j}$ is a random effect. In this parametrization, β_0 is the log odds of a friendship between children of opposite sexes, and $\beta_0 + \beta_1$ is the log odds for children of the same sex.

As described in the introduction, Hansell's (1984) classroom data exhibit several forms of network dependence, including node-level variability, reciprocity, and transitivity. This suggests we model the data with node-specific rates of sending and receiving ties, as well as a term which captures reciprocity and transitivity. We choose the following inner-product model with random sender and receiver effects:

$$\begin{aligned}
\text{log odds}(y_{i,j} = 1) &= \beta_0 + \beta_1 x_{i,j} + a_i + b_j + z_i' z_j \\
a_1, \ldots, a_n &\sim \text{i.i.d. Normal}(0, \sigma_a^2) \\
b_1, \ldots, b_n &\sim \text{i.i.d. Normal}(0, \sigma_b^2) \\
z_1, \ldots, z_n &\sim \text{i.i.d. Normal}(0, \sigma_z^2)
\end{aligned}$$

The parameters in this model are the regression coefficients β_0 and β_1, as well as the variance terms $\sigma_a^2, \sigma_b^2, \sigma_z^2$, which determine the dependencies between ties.

A Bayesian analysis was performed using the methods outlined in Section 3. The prior distributions for β_0 and β_1 were taken to be independent, diffuse normal distributions, both having mean zero and variance 100. The variance terms $\sigma_a^2, \sigma_b^2, \sigma_z^2$ were given diffuse inverse-Gamma(2,1) distributions, having an expectation of one but an infinite variance. An MCMC algorithm was used to obtain the 500,000 approximate samples from the posterior distribution $p(\beta_0, \beta_1, \sigma_a^2, \sigma_b^2, \sigma_z^2|Y)$. Marginal posterior distributions of $\beta_1, \sigma_a^2, \sigma_b^2, \sigma_z^2$ are presented in Figure 2. The results suggest a significant preference for same sex friendship ties, in that the posterior distribution for β_1 is centered around a median of 1.49, and a 95% quantile-based confidence interval for β_1 is (0.84, 2.11), which does not contain zero. The posterior distributions of σ_a^2 and σ_z^2 have deviated from their prior distributions and have moved to the right, giving evidence for sender-specific variability as well as the need for the latent variables

$z_1, \ldots, z_n$. The posterior for σ_b^2 concentrates mass on low values, and is not much different from the prior distribution, indicating little evidence for strong receiver-specific variability.

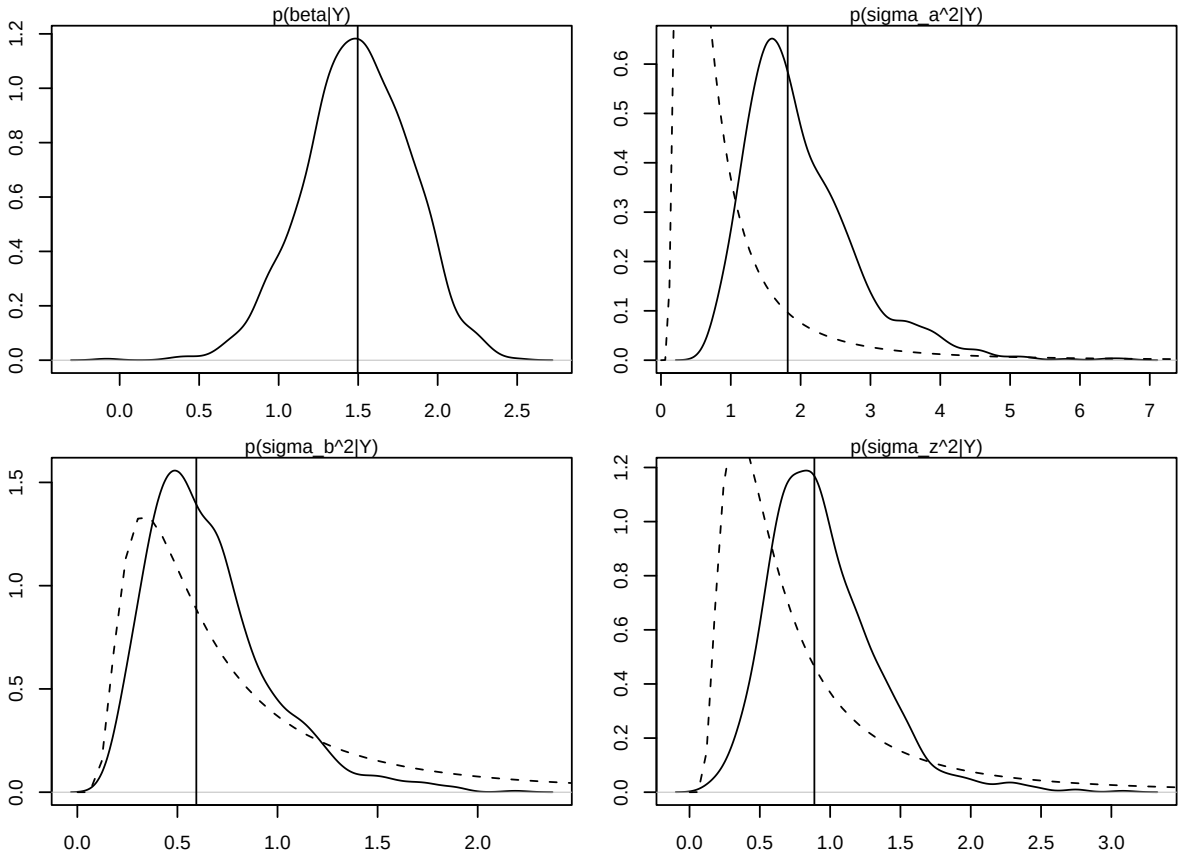

Figure 2: Marginal posterior distributions for the classroom data: Dashed lines represent the prior distributions for the variance parameters, solid lines the posterior. Vertical lines give the posterior median.

In comparison, a naive approach to inference would be to treat each possible tie as a Bernoulli random variable, independent of all other ties. Using standard logistic regression, our estimate of β_1 is 1.3 with a standard error of 0.2, giving an approximate 95% confidence interval of (0.91,1.70), which is of substantially smaller width than the interval obtained with the random effects model. Of course, we might expect the confidence interval based on this naive analysis to be too small, as it incorrectly assumes all ties between individuals are independent and thus overestimates the precision of the parameter estimate.

4.2 Tribal alliances

Read (1954) describes a number of network relations between sixteen New Guinean tribes. Here we consider the network of alliances between tribes, letting $y_{i,j} = 1$ if tribes i and j have an alliance, and $y_{i,j} = 0$ otherwise. We analyze these data using the simple distance model with no covariates or separate sender- and receiver-specific random effects:

$$\text{log odds } \Pr(y_{i,j} = 1|\beta_0, z_i, z_j) = \beta_0 - |z_i - z_j|,$$

where β_0 represents the baseline odds of a tie between two nodes that have the same latent position (i.e. β_0 the maximum log odds of a tie), and the z_i's are latent positions in $\mathbb{R}^2$. Without separate sender- and receiver-specific effects, we may expect that tribes with many alliances will be estimated as being more centrally located, and those with few ties as being on the periphery.

Bayesian estimates and confidence intervals for β_0 and the z_i's are obtained using the methods outlined in Section 3. In particular, samples of latent positions from the posterior distribution $p(z_1, \ldots, z_{16}|Y)$ are plotted in the first panel of Figure 1 (colors are chosen so that nearby node locations will have similar colors). Additionally, a black line drawn between nodes indicates the presence of an alliance.

Ad-hoc approaches, or simple point estimates of latent locations, might uncover some of the structure of the network. Our method goes beyond this by providing posterior confidence regions for node locations, which in turn give us a model-based measure of uncertainty about the network structure. Additionally, forms of predictive inference can be obtained from such a model. For example, suppose that the presence or absence of an alliance between pair (i, j) is unobserved or missing. The model can be fit with all available information (excluding the unknown $y_{i,j}$), and from the available information the posterior distributions of z_i and z_j can be obtained. From these, predictive inference about the value of $y_{i,j}$ can be made.

Also collected by Read (1954) were data on conflicts between the tribes. It is interesting to note that, based on a clustering of nodes (1,2,15,16), (3,4,6,7,8,11,12), and (5,9,10,13,14), there were no within-cluster conflicts, even though not every tribe within a cluster had an alliance with every other cluster co-member. Additionally, node 7, towards the center of the alliance structure, had no conflicts with any of the other 15 tribes. We note that both responses (conflict and alliance) could be modeled concurrently by a similar method, in which a multinomial logistic random effects model is employed in place of the binary logistic random effects model above.

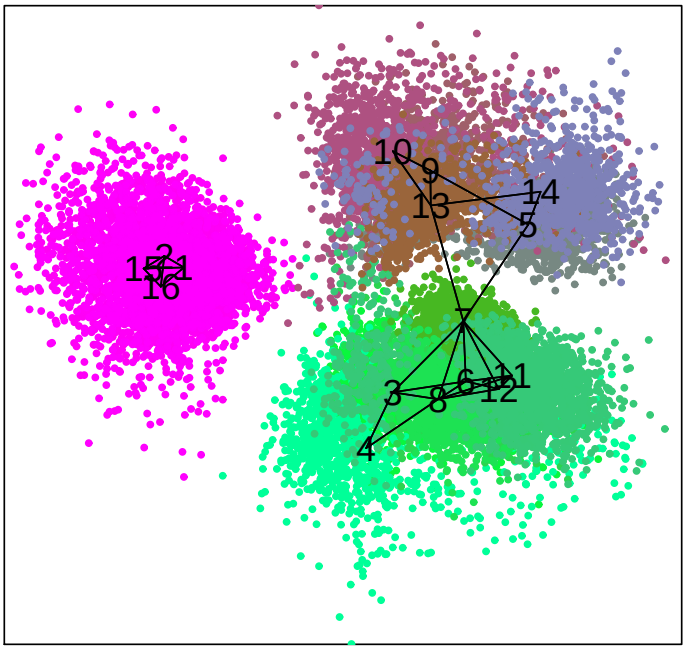

Figure 3: Tribal alliance network and marginal posterior distributions of locations.

5 Discussion

This article proposes a form of generalized linear mixed-effects model for the statistical analysis of network data for which parameter estimation is practical to implement. The approach has some advantages over existing social network models and inferential procedures: the approach allows for prediction and hypothesis testing; lends itself to a model-based method of network visualization; is highly extendable and interpretable in terms of well known statistical procedures such as regression and generalized linear models; and has a feasible means of exact parameter estimation.

The models discussed here can capture some types of network dependence, although it is possible (or even likely) that in many datasets there are types of dependencies that cannot be well-represented with these models. It then becomes important to develop methods for assessing model lack of fit, and determining the effect of lack of fit on the estimation of regression coefficients. Furthermore, it may be useful to combine the types of random effects discussed here with other types of random effects, or latent variables. For example, Nowicki and Snijders (2001) discuss a latent class model, a useful model for identifying clusters of nodes that relate to others in similar ways. Their latent class model, combined with types

of random effects models presented here and possibly other random effects structures, could provide a rich class of models for dependent network data.

References

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian computation and stochastic systems," *Statist. Sci.*, 10, 3–66, With comments and a reply by the authors.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.*, 85, 398–409.

Hansell, S. (1984), "Cooperative groups, weak ties, and the integration of peer friendships," *Social Psychology Quarterly*, 47, 316–328.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, to appear.

McCullagh, P. and Nelder, J. A. (1983), *Generalized linear models*, Chapman & Hall, London.

McCulloch, C. E. and Searle, S. R. (2001), *Generalized, linear, and mixed models*, Wiley Series in Probability and Statistics: Texts, References, and Pocketbooks Section, Wiley-Interscience [John Wiley & Sons], New York.

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

Read, K. (1954), "Cultures of the central highlands, New Guinea," *Southwestern Journal of Anthropology*, 10, 1–43.

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