# Multiplicative latent factor models for description and prediction of social networks

Peter D. Hoff [1]
Working Paper no. 54
Center for Statistics and the Social Sciences
University of Washington
Seattle, WA 98195-4320

January 17, 2006

[1]Departments of Statistics, Biostatistics and the Center for Statistics and the Social Sciences, University of Washington, Seattle, Washington 98195-4322, U.S.A.. 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 and National Science Foundation grant SES-0417559.

**Abstract**

We discuss a statistical model of social network data derived from matrix representations and symmetry considerations. The model can include known predictor information in the form of a regression term, and can represent additional structure via sender-specific and receiver-specific latent factors. This approach allows for the graphical description of a social network via the latent factors of the nodes, and provides a framework for the prediction of missing links in network data.

*Some key words*: eigenvalue decomposition, exchangeability, prediction, singular value decomposition, social network, visualization.

# 1 Introduction

Social network data are characterized by a set of binary link variables $y_{i,j}$ measured on pairs of a set of $n$ nodes, in which $y_{i,j}$ indicates the presence of a link from node $i$ to node $j$. Frequently accompanying such data are vectors of predictor variables $\mathbf{x}_{i,j}$ that may be specific to nodes or pairs. For example, in a set of $n$ schoolchildren $y_{i,j} = 1$ could indicate that child $i$ claims child $j$ as a friend, and $\mathbf{x}_{i,j}$ could contain demographic information about the pair such as age, sex, SES and co-residence in the same neighborhood.

Writing the network as an $n \times n$ matrix $\mathbf{Y}$ and the predictor information as an $n \times n \times p$-dimensional array $\mathbf{X}$, a common approach to the analysis of such data is to fit a statistical model relating $\mathbf{Y}$ to $\mathbf{X}$. A statistical model can provide many forms of inference, such as a description of the relationship between $\mathbf{Y}$ and $\mathbf{X}$, a measure of the uncertainty in this relationship via confidence intervals, and predictions about missing or future network data. Perhaps the simplest model that relates $\mathbf{Y}$ to $\mathbf{X}$ is the ordinary logistic regression model:

$$\begin{aligned} \Pr(\mathbf{Y}|\beta, \mathbf{X}) &= \prod_{i \neq j} \frac{\exp\{\theta_{i,j}\}}{1+\exp\{\theta_{i,j}\}} \\ \log \text{odds}(y_{i,j} = 1) &= \theta_{i,j} = \beta' \mathbf{x}_{i,j}. \end{aligned} \tag{1}$$

Equation (1) indicates that the observations $\{y_{i,j}\}$ are assumed to be statistically independent. Inference for the regression coefficient $\beta$ in this situation is fairly straightforward (see, for example, McCullagh and Nelder 1983), and is provided by most statistical software packages. However, this assumption is strongly violated in most social network datasets. In many networks there is heterogeneity in activity levels across nodes. For example, in social settings some people are more active than others, and on the web some pages are more heavily linked. This across-node heterogeneity leads to within-node homogeneity of ties: The relationships $\{y_{i,1}, y_{i,2}, \ldots, y_{i,n}\}$ will often be more similar to each other than they are to other network measurements because they all share something in common: they all involve node $i$. This typically is manifested statistically by a strong within-node dependence of ties, a violation of the independence assumption in equation (1). Other manifestations of network dependence include reciprocity and clustering. Reciprocity is the notion that $y_{i,j}$ and $y_{j,i}$ will be statistically dependent. For example, friendship ties are generally positively correlated. Clustering is the phenomenon in which a subset of nodes exhibit a large number of within-group ties and relatively few ties outside of the group. This is related to the notion of transitivity ("a friend of a friend is a friend"). See Wasserman and Faust (1992) for more on common structures of network data.

One approach to statistical inference and modeling of such dependence patterns has been the use of exponentially parameterized random graph models, or "$p^*$" models (Wasserman and Pattison 1996). In these models the log-probability of a network $\mathbf{Y}$ is given by a linear regression term plus

a linear combination of network statistics. For a single pair of nodes, the log-odds of a link from $i$ to $j$ can be written as

$$\log \text{odds}(y_{i,j} = 1) = \beta' \mathbf{x}_{i,j} + \alpha_1 t_1(\mathbf{Y}_{-(i,j)}) + \cdots + \alpha_m t_m(\mathbf{Y}_{-(i,j)}). \tag{2}$$

The log-odds written above is the *conditional* log-odds, which indicates how the probability of a link from $i$ to $j$ may depend on $\mathbf{Y}_{-(i,j)}$, the data from other pairs of nodes. The statistics $t_1(\mathbf{Y}_{-(i,j)}), \ldots, t_p(\mathbf{Y}_{-(i,j)})$ are related to the sufficient statistics and are typically functions of such things as the total number of ties in the network, the number of reciprocal ties and the number of transitive triangles. It is often quite difficult to estimate the parameters $\beta$ and $\alpha$, and the resulting models often display considerable lack-of-fit (Snijders 2002, Handcock 2003). However, these models are conceptually straightforward and the parameter estimates can provide a representation of the global features of the network.

An alternative approach to modeling dependencies among relational data is the use of random effects models. Random effects models are a cornerstone of many statistical methods for the analysis of dependent data, although their application to the analysis of relational network data has been fairly recent. In the context of logistic regression, such models take the form

$$\begin{aligned} \Pr(\mathbf{Y}|\theta_{i,j}) &= \prod_{i \neq j} \frac{\exp\{\theta_{i,j}\}}{1 + \exp\{\theta_{i,j}\}} \\ \theta_{i,j} &= \beta' \mathbf{x}_{i,j} + z_{i,j} \end{aligned}$$

As in (1), the data are modeled as conditionally independent given the $\theta_{i,j}$'s, but the $\theta_{i,j}$'s depend on the set of $z_{i,j}$'s, the unobserved random effects. The $z_{i,j}$'s are then modeled to account for potential dependencies in the data. For example, we might want to allow for heterogeneity in the total volume of sending and receiving activity across nodes. This could be accomplished by a model of the form $z_{i,j} = a_i + b_j$. Although such an additive random effects model for the $z_{i,j}$'s generally provides a drastic improvement in model fit over ordinary logistic regression, it is unable to represent higher-order network structure, such transitivity or clustering of nodes. Recently, some authors have taken a non-additive approach to modeling the $z_{i,j}$'s. Nowicki and Snijders (2001) represent the probability of a link between nodes $i$ and $j$ as depending on their membership to a set of unobserved latent classes. Hoff, Raftery and Handcock (2002) and Hoff (2005) model the probability of a link as depending on the similarity of nodes $i$ and $j$ in a space of unobserved latent characteristics. These types of models are able to represent standard network behavior such as clustering and transitivity, and estimation for these types of models is generally less problematic than estimation for models of the form (2).

In the remainder of this article we motivate a multiplicative latent factor effects model for social network data. In this model, structure in the network is represented by the form $z_{i,j} = \mathbf{u}_i' \mathbf{D} \mathbf{v}_j + \epsilon_{i,j}$, where $\mathbf{u}_i$ and $\mathbf{v}_j$ represent vectors of sender-specific and receiver-specific latent nodal attributes.

This is a more general form of the bilinear mixed-effects model discussed in Hoff (2005), in that the characteristics of a node as a sender may differ from its characteristics as a receiver. In the next section we motivate this type of model based on matrix representations and invariance properties. In Section 3 we briefly discuss a method of parameter estimation using a Markov chain Monte Carlo algorithm. Section 4 presents a data analysis of international conflict data. Using this example, it is shown how the latent factor model can be used to graphically represent patterns in the data and make predictions about missing network data. Undirected network data is considered in Section 5, and a discussion follows in Section 6.

# 2 Latent variable models

Consider a model for binary network data of the form

$$\begin{aligned} \log \text{odds}(y_{i,j} = 1) &= \theta_{i,j} \\ \theta_{i,j} &= \beta' \mathbf{x}_{i,j} + z_{i,j}. \end{aligned} \tag{3}$$

Here the regression term $\beta' \mathbf{x}_{i,j}$ represents patterns in the data related to known predictor variables $\mathbf{x}_{i,j}$ and $z_{i,j}$ represents any additional patterns in the data unrelated to those of the predictors. As discussed above, one simple model for social network data is obtained by restricting $z_{i,j} = a_i + b_j$, i.e. letting $z_{i,j}$ represent only additive row effects and column effects. In this section we motivate a more general non-additive row and column effects model based on matrix decomposition methods, and further justify the model by an invariance assumption for the distribution of the $z_{i,j}$'s known as exchangeability.

## 2.1 Models via matrix decompositions

Let $\mathbf{Z}$ be an $n \times n$ random matrix of effects representing deviations of the log-odds $\theta_{i,j}$ from the linear predictor $\beta' \mathbf{x}_{i,j}$. We can write $\mathbf{Z} = \mathbf{M} + \mathbf{E}$, where the mean matrix $\mathbf{M}$ represents systematic patterns in the effects and $\mathbf{E}$ represents noise. A basic result from matrix theory is that every $n \times n$ matrix $\mathbf{M}$ has the representation

$$\mathbf{M} = \mathbf{UDV}'$$

where

- $\mathbf{U}$ is an $n \times n$ matrix with orthonormal columns;
- $\mathbf{V}$ is an $n \times n$ matrix with orthonormal columns;
- $\mathbf{D}$ is an $n \times n$ diagonal matrix, with diagonal elements $\{d_1, \ldots, d_n\}$.

The triple $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$ is called the singular value decomposition of $\mathbf{M}$. The squared elements of the diagonal of $\mathbf{D}$ are the eigenvalues of $\mathbf{M}'\mathbf{M}$ and the columns of $\mathbf{V}$ are the corresponding eigenvectors. The matrix $\mathbf{U}$ can be obtained from the first $n$ eigenvectors of $\mathbf{M}\mathbf{M}'$. The number of non-zero elements of $\mathbf{D}$ is the rank of $\mathbf{M}$.

In applications such as signal processing, image analysis and more recently large-scale gene expression data, researchers often represent the main patterns of a noisy, matrix-valued dataset with the first few singular vectors and values of the matrix. The goal of such reduced-rank approximations is to represent the main patterns in the data matrix while eliminating the lower-order noise. For model (3), this motivates the reduced-rank representation $\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}' + \mathbf{E}$ where $\mathbf{U}, \mathbf{D}$ and $\mathbf{V}$ all have $K$ columns, with $K << n$, thus restricting $\mathbf{M}$ to be of rank $K$. Letting $\mathbf{u}_i$ and $\mathbf{v}_j$ be the $i$th and $j$th rows of $\mathbf{U}$ and $\mathbf{V}$ respectively, the entries of $\mathbf{Z}$ have the representation

$$z_{i,j} = \mathbf{u}_i'\mathbf{D}\mathbf{v}_j + \epsilon_{i,j},$$

which has an appealing interpretation as a multiplicative model based on row and column factors. In this model the relationship between $i$ and $j$ is a product of a vector of $K$ latent sender-specific factors $\mathbf{u}_i$ with latent receiver-specific factors $\mathbf{v}_j$, weighted by $\mathbf{D}$. The model for binary network data thus becomes

$$\begin{aligned} \log \text{odds}(y_{i,j} = 1) &= \theta_{i,j} \\ \theta_{i,j} &= \beta'\mathbf{x}_{i,j} + \mathbf{u}_i\mathbf{D}\mathbf{v}_j' + \epsilon_{i,j}. \end{aligned} \tag{4}$$

### 2.2 Models via exchangeability

The matrix $\mathbf{Z}$ discussed above represents structures in the data that are not associated with known covariate information $\mathbf{X}$. The model (4) represents this structure as a function of latent row-specific and column-specific factors, plus noise. Such a representation has a justification and interpretation as a type of random effects model, via a concept from probability theory known as exchangeability. An infinite sequence of random variables $z_1, z_2, \ldots$ is said to be exchangeable if for every integer $n$, the distribution of $\{z_1, \ldots, z_n\}$ is equal to that of $\{z_{\pi 1}, \ldots, z_{\pi n}\}$ for any permutation $\pi$ of $\{1, \ldots, n\}$. Exchangeability is more general than independence: a sequence of independent and identically distributed random variables is exchangeable, but not necessarily vice versa. Exchangeable models are commonly used in statistics for sequences of random variables which have something in common, but are indistinguishable in that their labels carry no information. A remarkable theorem of de Finetti makes this concept precise: Any sequence of exchangeable random variables has the representation $z_i = f(\mu, \epsilon_i)$, where $\mu$ and the $\epsilon_i$'s are independent random variables. Relating to the above discussion, $\mu$ is the quantity shared by all members of the exchangeable sequence and the $\epsilon_i$'s determines the patternless way in which they differ.

The concept of exchangeability extends to matrices. A matrix $\mathbf{Z}$ is said to be row-and-column exchangeable (RCE) if the random variables $\{z_{i,j}\}$ are equal in distribution to the set $\{z_{\pi_1 i,\pi_2 j}\}$ for all finite permutations $\pi_1$ and $\pi_2$. Exchangeability in this case can be interpreted as saying that the row labels and the column labels carry no information about $\mathbf{Z}$. The analog of de Finetti's theorem in this case is as follows:

**Theorem 1 (Aldous 1981)** *If* $\mathbf{Z}$ *is an RCE matrix, then there exists a function* $f$ *and independent random variables* $\mu, \{u_1, u_2, \ldots\}, \{v_1, v_2, \ldots\}, \{\epsilon_{i,j}, i = 1, \ldots, j = 1, \ldots\}$ *such that* $z_{i,j} \stackrel{d}{=} f(\mu, u_i, v_j, \epsilon_{i,j})$.

(the "$\stackrel{d}{=}$" above means "equal in distribution"). This theorem says that *any* statistical model for an RCE matrix can be expressed in terms of a "grand mean" $\mu$, row effects $\{u_i\}$, column effects $\{v_j\}$, and independent disturbance terms $\{\epsilon_{i,j}\}$.

Returning to our model for social network data, the effects $\{z_{i,j}\}$ are meant to represent any patterns in the data beyond any known covariate information $\mathbf{X}$. In this sense, the $z_{i,j}$'s are unrelated to any node-specific information we may have, and so it may be appropriate to model the $z_{i,j}$'s as being the components of an RCE array. By virtue of Theorem 1, this justifies the use of a model of the form $z_{i,j} = \mathbf{u}_i' \mathbf{D} \mathbf{v}_j + \epsilon_{i,j}$, i.e., modeling the effects $\{z_{i,j}\}$ as functions of row-specific, column-specific and dyad-specific random effects.

A less restrictive form of matrix exchangeability is the concept of weak exchangeability, under which the random variables $\{z_{i,j}\}$ are equal in distribution to the set $\{z_{\pi i,\pi j}\}$ for any simultaneous permutation $\pi$ of both the row and the column labels. Such a modeling assumption might be desirable if one wanted to relate the outcomes involving node $i$ as a sender $\{z_{i,1}, z_{i,2}, \ldots\}$ to the outcomes involving $i$ as a receiver $\{z_{1,i}, z_{2,i}, \ldots\}$. In this case, it would be appropriate to develop a statistical model in which the row effect $u_i$ of each node is correlated with its column effect $v_i$, and to allow $\epsilon_{i,j}$ to be correlated with $\epsilon_{j,i}$. Additive statistical models of this form for normally distributed data have been called "social relations models" (Warner, Kenny and Stoto 1979, Wong 1982), and have been extended to the analysis of binary social network data in Hoff (2005).

# 3 Parameter estimation

The unknown quantities in our model include

- $\Theta = \{\theta_{i,j}\}$, the set of predictors;
- $\beta$, the vector of regression coefficients;
- $\mathbf{U}$ and $\mathbf{V}$, both $n \times K$ matrices with orthonormal columns, denoted $\mathbf{U} = \{\mathbf{U}_{[,1]}, \ldots, \mathbf{U}_{[,K]}\}$ and $\mathbf{V} = \{\mathbf{V}_{[,1]}, \ldots, \mathbf{V}_{[,K]}\}$;

- $\mathbf{D} = \text{diag}\{d_1, \ldots, d_K\}$, an $n \times n$ diagonal matrix.

Estimation of the model parameters is most easily done in a Bayesian context: Given a prior distribution on the model parameters, obtain their posterior distribution via Bayes rule, $p(\Theta, \beta, \mathbf{U}, \mathbf{D}, \mathbf{V}|\mathbf{Y}) \propto p(\mathbf{Y}|\Theta, \beta, \mathbf{U}, \mathbf{D}, \mathbf{V}) \times p(\Theta, \beta, \mathbf{U}, \mathbf{D}, \mathbf{V})$. Various posterior quantities of interest, such as posterior means, confidence intervals and predicted values are functions of this posterior distribution. Although these quantities cannot be derived directly, they can be approximated via Markov chain Monte Carlo sampling, a type of stochastic algorithm that generates a dependent sequence of realizations of the parameters. Such algorithms can be constructed so that the empirical distribution of the generated parameters approximates the desired posterior distribution (see, for example, Tierney, 1994). Given starting values $\psi_0 = \{\Theta, \beta, \mathbf{U}, \mathbf{D}, \mathbf{V}\}$ one such MCMC scheme iteratively generates a sequence $\psi_1, \psi_2, \ldots$ as follows:

1. sample $\beta$ from its full conditional distribution $p(\beta|\Theta, \mathbf{U}, \mathbf{D}, \mathbf{V})$;
2. for $k \in 1 \ldots, K$,
   (a) sample $\mathbf{U}_{[,k]}$ from $p(\mathbf{U}_{[k]}|\Theta, \mathbf{U}_{[,-k]}, \mathbf{D}, \mathbf{V})$;
   (b) sample $\mathbf{V}_{[,k]}$ from $p(\mathbf{V}_{[k]}|\Theta, \mathbf{U}, \mathbf{D}, \mathbf{V}_{[,-k]})$;
   (c) sample $\mathbf{D}_{[k,k]}$ from $p(\mathbf{D}_{[k,k]}|\Theta, \mathbf{U}, \mathbf{D}_{[-k,-k]}, \mathbf{V})$;
3. sample $\mathbf{\Theta}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{D}\mathbf{V}' + \mathbf{E}^*$, where $\mathbf{E}^*$ is a matrix of standard normal noise. Replace $\theta_{i,j}$ by $\theta^*_{i,j}$ with probability $\frac{p(y_{i,j}|\theta^*_{i,j})}{p(y_{i,j}|\theta_{i,j})} \wedge 1$.

From a current state of the parameters $\psi_s$, one run of the above procedure generates a new set of parameters $\psi_{s+1}$. Run iteratively, the distribution of the generated samples converges to the desired posterior distribution $p(\psi|\mathbf{Y})$, and posterior quantities of interest can be approximated from the empirical distribution of the generated samples. For example, the posterior mean of $\beta$ can be approximated by the empirical mean of the MCMC samples, and the endpoints of a 95% confidence interval can be obtained from the 2.5% and 97.5% sample quantiles. More details on the above sampling scheme, and software to run such an algorithm, is available at my website: `www.stat.washington.edu/hoff`.

## 4 Example: international conflict data

In this section we illustrate the use of model (4) with an analysis of a social network of international conflicts among 130 nations from 1990-2000, compiled by Mike Ward and Xun Cao of the University of Washington Political Science department. For these data, $y_{i,j} = 1$ if country $i$ initiates one or more conflicts with country $j$ sometime during 1990-2000. Standard practice in the international

relations literature is to relate the entries $y_{i,j}$ of the sociomatrix $\mathbf{Y}$ to vectors of explanatory variables $\mathbf{x}_{i,j}$ via logistic regressions of the form $\log \text{odds}(y_{i,j} = 1|\beta, \mathbf{x}_{i,j}) = \beta' \mathbf{x}_{i,j}$. For the analysis in this section, $\mathbf{x}_{i,j}$ is an eight-dimensional vector of regressor variables including an intercept and the following seven predictors :

1. log populations of the aggressor nation;
2. polity score of the aggressor (a measure of democracy);
3. log populations of the target nation;
4. polity score of the target;
5. geographic distance between aggressor and target;
6. product of the aggressor polity score and the target polity score (an interaction term);
7. number of intergovernmental organization having both nations as members.

More details on the data are available in Ward and Hoff (2005). Models such as these are often used in the political science literature to evaluate various hypotheses about the nature of international relations. For example, it is often hypothesized that rates of conflict are lower among democratic countries, and among countries that are co-members of intergovernmental organizations.

As discussed in Section 1, ordinary logistic regression models for such data are limited in their ability to describe patterns in the data and typically have poor predictive performance. In the next two subsections we show how the multiplicative latent factor model can represent higher-order network patterns, and how predictive performance is drastically improved upon by the inclusion of these factors.

### 4.1 Data description and visualization

The dataset contains information on the presence or absence of conflicts among 130 nations, giving $130 \times 129 = 16770$ binary observations. The network is very sparse, with only 1.2 % of the pairs having links ($\frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} = 0.012$). We fit the model

$$\log \text{odds}(y_{i,j} = 1|\beta, \mathbf{x}_{i,j}) = \beta' \mathbf{x}_{i,j} + \mathbf{u}_i \mathbf{D} \mathbf{v}_j' + \epsilon_{i,j}$$

to the data, with the dimension of the latent factors being $K = 2$. This was done by constructing a Markov chain of length 100,000 using the algorithm described in Section 3. For simplicity, we used independent diffuse normal(0,1000) prior distributions for the regression coefficients $\beta$ and the diagonal elements of $\mathbf{D}$. The priors distributions on $\mathbf{U}$ and $\mathbf{V}$ were taken to be the uniform distributions on the space of orthonormal $n \times 2$ matrices.

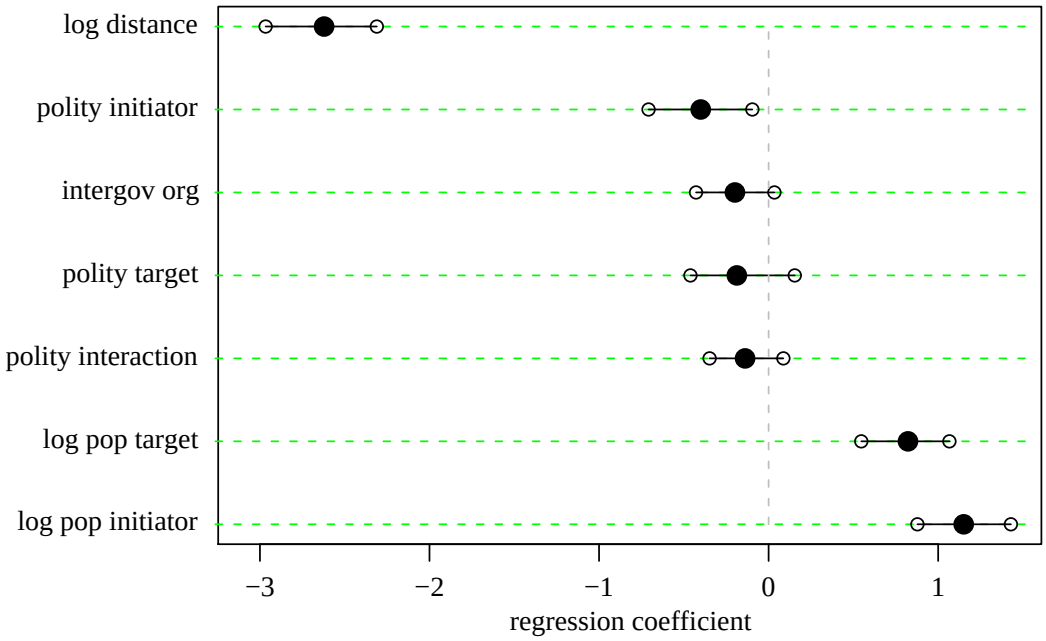


Figure 1: Regression coefficients. Closed dots are posterior medians, open dots are the 2.5% and 97.5% quantiles.

Figure 1 displays posterior $2.5, 50$ and $97.5\%$ quantiles of the regression coefficients of the seven predictor variables, providing a posterior estimate and a 95% confidence interval for each. The results indicate strong relationships between conflict and geographic distance (with higher conflict rates between geographically proximate nations), and between conflict and population (with higher conflict rates between countries with large populations). Less strong are the relationships between conflict and the other predictors, but the results indicate lower rates of conflicts among democratic countries and among countries that are co-members of intergovernmental organizations

Posterior estimates of the multiplicative latent factors are shown in Figure 2. These were obtained from the first two left- and right-singular vectors of the posterior mean of the matrix $\mathbf{UDV}'$, yielding two two-dimensional vectors $\{\hat{\mathbf{u}}_i, \hat{\mathbf{v}}_i\}$ for each nation. The directions of these vectors are plotted in Figure 2 for each country involved in a conflict, with the direction of $\hat{\mathbf{u}}_i$ indicated in red and $\hat{\mathbf{v}}_i$ in blue. The size of the text for each country is related to the magnitude of their vectors. Finally, links between countries are shown by green lines.

These latent factors indicate a large amount of structure in the data beyond that which can be represented by a simple ordinary logistic regression with a small number of predictors. For example, conflicts in the Middle East and Persian Gulf show up clearly as clusters of aggressors and targets at roughly 180 and 270 degrees from the horizontal axis. Conflicts among African countries appear on the opposite side of the circle. Note that very few links cross through the center of the circle.

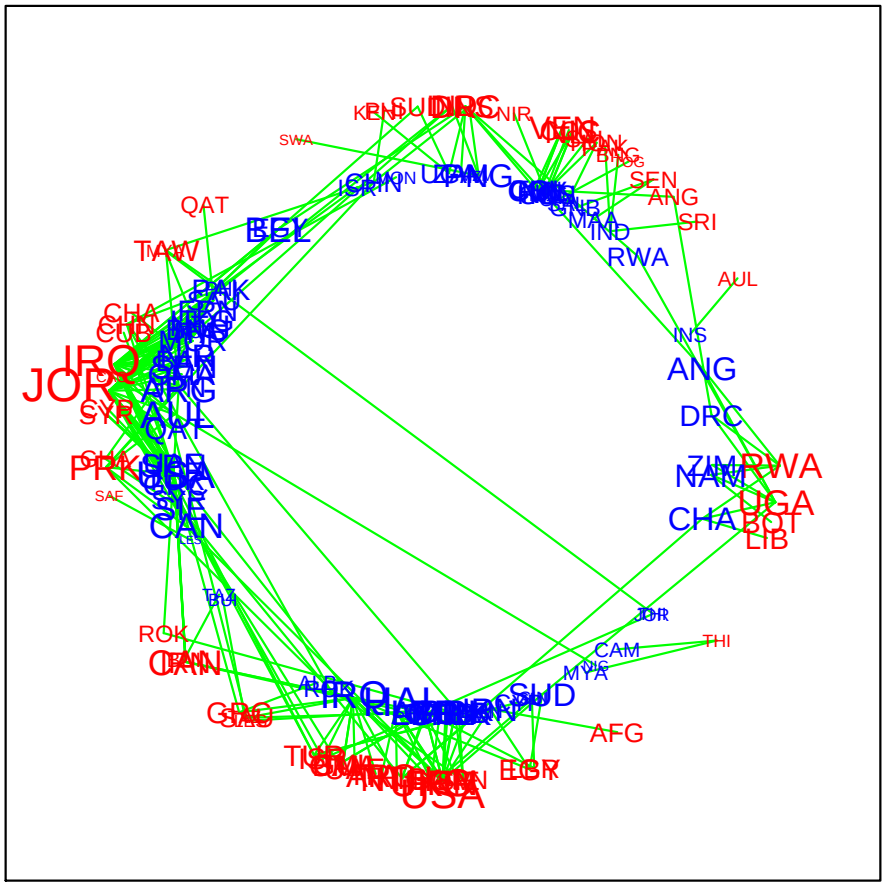


Figure 2: Aggressor and target-specific latent variables. Directions of aggressor-specific latent variables are given on the outer circle, directions of target-specific latent variables on the inner. Conflicts between countries are given in green lines.

This indicates that the pattern of conflicts can be well represented by these two-dimensional latent factors.

### 4.2 Link prediction

When the number of nodes in a social network is large it may be difficult or impossible to make measurements on each pair of nodes, resulting in a large number of missing values in the dataset. In these situations it may be desirable to make predictions about the unobserved, missing data based on the observed data. To evaluate the ability of the multiplicative latent factor model to predict missing links, we performed the following prediction experiment on the international conflict data:

1. Randomly divide the set of $n \times (n-1)$ ordered pairs of indices $(i, j)$ into two parts, a training set $T$ and a test set $M$.

2. Estimate the parameters of model (4) using the MCMC algorithm, using the data $\mathbf{Y}_T = \{y_{i,j} : (i, j) \in T\}$ and treating the data in $M$ as missing.

3. Based on the results of the Markov chain, obtain fitted values $\hat{p}_{i,j} = E[\frac{\exp\{\theta_{i,j}\}}{1+\exp\{\theta_{i,j}\}}|\mathbf{Y}_T]$ for each $\{y_{i,j} : (i, j) \in M\}$

4. Compare $\hat{p}_{i,j}$ to $y_{i,j}$ for each pair $(i, j) \in M$.

The above steps constitute one-half of a two-fold cross-validation procedure, a procedure often used to evaluate the predictive performance of statistical models. In general, an $m$-fold cross validation procedure consists of dividing the dataset into $m$ parts, and making predictions for each part based on parameters estimated from the remaining $m-1$ parts. Note that in cross-validation, increasing the model complexity doesn't always improve the model fit, as the predictive performance of the model is evaluated by predictions for data that were not used to obtain parameter estimates.

Step 1 above divided the dataset into sets $T$ and $M$, each containing data on 8385 pairs of countries. The number of missing links in the test set was $\sum_{(i,j)\in M} y_{i,j} = 103$. Steps 2, 3 and 4 were performed for $K = 0, 1, 2$ and 3 to see how the predictive performance changed as the complexity of the model was increased via the dimension of the latent factors. The results of this prediction study are shown in two plots in Figure 3. The first plot refers to the following scenario: Imagine that a dataset with missing information is obtained. Let the set $M$ consist of all pairs for which $y_{i,j}$ is missing. The task is to identify pairs $(i, j) \in M$ for which it is likely that $y_{i,j} = 1$, i.e. the task is to find the missing links. One strategy would be to fit a statistical model with the available data $\mathbf{Y}_T$, obtain predicted probabilities $\hat{p}_{i,j} = \Pr(y_{i,j} = 1|\mathbf{Y}_T)$ for all node pairs $\{i, j\} \in M$, and then investigate the node pairs having the highest predictive probabilities of having a link. The first panel of Figure 3 plots the results of this exercise for models with $K = 0, 1, 2$ and 3-dimensional latent vectors. For each value $K$ being considered, predictive probabilities were made

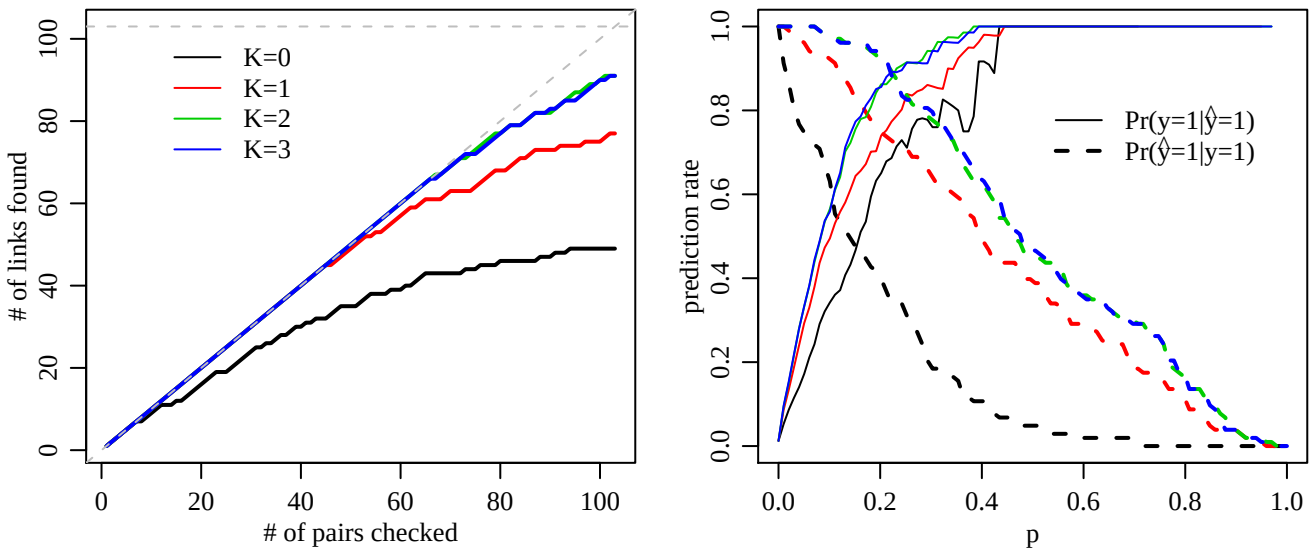


Figure 3: Predicting links. The left panel indicates the number of missing links uncovered as a function of the number of pairs checked. The right panel gives prediction error rates as a function of a prediction threshold $p$.

for each $(i, j)$-pair in $M$ based on the parameter estimates from $\mathbf{Y}_T$. The first panel of Figure 3 tells us how many links we would find if we were to investigate the pairs having the highest predictive probabilities. For example, if we were to investigate the pairs in $M$ having the top 100 predictive probabilities using the $K = 2$ estimates, we would uncover 90 links, or almost 90% of the total number of missing links. In contrast, if we didn't use the latent effects model ($K = 0$) we would uncover only 49.

The second panel of Figure 3 looks at the predictive performance in a different way. Consider predicting $y_{i,j}$ for $(i, j) \in M$ as $\hat{y}_{i,j} = 0$ or $\hat{y}_{i,j} = 1$ according to whether or not $\hat{p}_{i,j} < p$ or $\hat{p}_{i,j} \geq p$ for some threshold $p$. The plot displays the true positive rates $\Pr(y_{i,j} = 1|\hat{y}_{i,j} = 1)$ in solid lines, as well as the fraction of links that are recovered $\Pr(\hat{y}_{i,j} = 1|y_{i,j} = 1)$ in dashed lines, for various prediction criteria $p$. For example, if $p = .25$, a model with $K = 2$ yields a set of predicted links in which 91.5% are true links, and the set of predicted links will include 83.5% of the actual missing links. In contrast, if we don't use a latent effects model these numbers will be $\Pr(y_{i,j} = 1|\hat{y}_{i,j} = 1) = .711$ and $\Pr(\hat{y}_{i,j} = 1|y_{i,j} = 1) = .311$ respectively. These results show that a two-dimensional latent factor model has dramatically better predictive performance than a model lacking such structure, and that a three-dimensional factor model is unnecessary, given that it has similar predictive performance to that of a two-factor model.

We suggest the following general approach to link prediction in the presence of missing data:

1. Let $M = \{(i,j) : y_{i,j} \text{ is missing}\}$ and $T = \{(i,j) : y_{i,j} \text{ is observed}\}$.

2. Using data on pairs in $T$, use a cross validation procedure to obtain an optimal $K$ and the associated prediction and error rates.

3. Use parameters estimates from the optimal model to make predictive probabilities $\hat{p}_{i,j}$ for missing pairs $(i,j) \in M$.

4. Make predictions or search for more links based on the $\hat{p}_{i,j}$'s.

Cross-validation procedures obtain prediction and error rates under the assumption that the data are missing at random. If inclusion in the set $M$ is not random then the estimated prediction and error rates may not be accurate. However, the results might still provide some guidance as to which of the pairs in $M$ are most likely to have a link.

# 5 Extension to undirected data

A social network is called undirected if it consists of binary relationships between nodes in which $y_{i,j} = y_{j,i}$ by design. In this case, a model analogous to the one developed in Section 2 can be constructed using similar results on matrix decompositions and exchangeability. For undirected data, we can write the model in (3) as

$$\begin{aligned} \log \text{odds}(y_{i,j} = y_{j,i} = 1) &= \theta_{i,j} \\ \theta_{i,j} &= \beta' \mathbf{x}_{i,j} + z_{i,j}. \end{aligned}$$

where now the effects $z_{i,j}$ can be represented with a symmetric $n \times n$ matrix $\mathbf{Z}$. We write $\mathbf{Z} = \mathbf{M} + \mathbf{E}$ as before, with all matrices being symmetric. Analogous to the singular value decomposition, every square, symmetric matrix $\mathbf{M}$ has an eigenvalue decomposition of the form $\mathbf{M} = \mathbf{U}\Lambda\mathbf{U}'$, where $\Lambda$ is a diagonal matrix of real numbers and $\mathbf{U}$ is an orthonormal matrix. This motivates a model of the form $m_{i,j} = \mathbf{u}_i' \Lambda \mathbf{u}_j$, giving

$$\log \text{odds}(y_{i,j} = y_{j,i} = 1) = \beta' \mathbf{x}_{i,j} + \mathbf{u}_i' \Lambda \mathbf{u}_j + \epsilon_{i,j} \tag{5}$$

with $\epsilon_{i,j} = \epsilon_{j,i}$. The interpretation is that the relationship between $i$ and $j$ is a function of the observed predictor variables $\mathbf{x}_{i,j}$ and unobserved latent factors $\mathbf{u}_i$ and $\mathbf{u}_j$. As in the case of directed network data, such a model has a justification via exchangeability: By assumption, all known information distinguishing the nodes is contained in $\mathbf{X}$, and so it is reasonable to model $\mathbf{Z}$ such that $\{z_{i,j}\}$ is equal in distribution to $\{z_{\pi i,\pi j}\}$ for any permutation $\pi$. The symmetric matrix $\mathbf{Z}$ is thus weakly exchangeable, and a theorem of Aldous (1985) says that *any* model for a symmetric, weakly exchangeable $\mathbf{Z}$ can be written

$$z_{i,j} \stackrel{d}{=} f(\mu, u_i, u_j, \epsilon_{i,j})$$

for some function $f$ which is symmetric in its second and third arguments.

## 6 Discussion

In the analysis of social network data it is often desirable to make inference about local network structure. For example, it might be of interest to graphically describe regions of the network, or to make predictions about the potential for the presence of unobserved links between a set of nodes. This article has presented a model-based approach to making such inference. Motivated by ideas from matrix decomposition theory and the theory of exchangeable matrices, the approach is based on a latent variable model in which network structure is represented in terms of unobserved latent node-specific factors. This approach allows for the graphical description of the network in terms of the latent factors, and can make predictions about missing observations in social network data. Software and example analyses are available at my website, `www.stat.washington.edu/hoff`.

## References


Aldous, D. J. (1981), "Representations for partially exchangeable arrays of random variables," *J. Multivariate Anal.*, 11, 581–598.

Aldous, D. J. (1985), "Exchangeability and related topics," in *École d'été de probabilités de Saint-Flour, XIII—1983*, vol. 1117 of *Lecture Notes in Math.*, pp. 1–198, Springer, Berlin.

Handcock, M. S. (2003), "Assessing degeneracy in statistical models of social networks," Working paper no. 39, Center for Statistics and the Social Sciences, University of Washington-Seattle.

Hoff, P. D. (2005), "Bilinear mixed-effects models for dyadic data," *J. Amer. Statist. Assoc.*, 100, 286–295.

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, 1090–1098.

McCullagh, P. and Nelder, J. A. (1983), *Generalized linear models*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Nowicki, K. and Snijders, T. A. B. (2001), "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087.

Snijders, T. A. B. (2002), "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure*, 3.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *Ann. Statist.*, 22, 1701–1762, With discussion and a rejoinder by the author.

Ward, M. D. and Hoff, P. D. (2005), "Persistent Patterns of International Commerce," Working paper no. 45, Center for Statistics and the Social Sciences, University of Washington-Seattle.

Warner, R., Kenny, D. A., and Stoto, M. (1979), "A new round robin analysis of variance for social interaction data," *Journal of Personality and Social Psychology*, 37, 1742–1757.

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

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.

Wong, G. Y. (1982), "Round robin analysis of variance via maximum likelihood," *Journal of the American Statistical Association*, 77, 714–724.