Center for
Studies in
Demography
and Ecology

THE LINK BETWEEN EXPONENTIAL RANDOM GRAPH MODELS AND LOGLINEAR MODELS FOR NETWORKS

By

Laura M. Koehly

Texas A&M University

Steven M. Goodreau

University of Washington

Martina Morris

University of Washington

ABSTRACT

Much progress has been made on the development of statistical methods for network analysis in the past ten years. Building on the general class of exponential random graphs, a range of new statistical models have been proposed, including Markov graphs, "p^*" models, and actor-oriented models, to capture the systematic patterns of association and dyadic dependence in networks. This class of models is directly related to the loglinear models used in earlier work to analyze mixing patterns in local network data. Both approaches are based on the exponential family, and the link between them turns out to have a number of interesting implications for network analysis generally. Random graphs model the probability that two actors are relationally tied given their attributes and structural characteristics in the data, while loglinear methods model the probability that two actors have specific attributes given that they are relationally tied. Under dyadic independence the two probabilities are related via Bayes' rule. The modeling frameworks do not yield equivalent predicted values except when fully saturated, however, due to alternate forms of conditioning. In practice, the differences are unlikely to be large, but the alternate conditioning helps to clarify the strengths and weaknesses of each modeling approach, as well as the behavioral assumptions. Understanding the relationship between the two models sheds light on the relationship between local and complete network data, and the role that models can play in bridging the traditional gap between them.

1. INTRODUCTION

For many years the methodology for network analysis has developed along two distinct paths, one mathematical, the other statistical. The mathematical approach is based in graph theory and has largely defined the field, providing a conceptual framework for thinking about networks and a wide range of summary measures to represent network position and structure. Almost all of the classic network measures – like density, centrality, structural equivalence, and cliques -- owe their development to researchers working in this tradition. Textbooks and computer packages for network analysis typically have these measures at their core. They have become the common language for network analysis, and have helped to develop our intuitions about the complex relational structures we seek to understand.

The statistical approach to network methodology is distinguished by the additional concern with measuring the variation and uncertainty in the quantities that are estimated. This approach is rooted in probability theory. While it has given rise to a number of different techniques in network analysis, it has, until recently, played a relatively minor role in the field. Traditional statistical methodology, with tractable likelihood-based inference, requires observations to be independent. In network analysis, the whole point is that observations are not independent. As non-likelihood-based re-sampling methods like the bootstrap and jackknife were developed in statistics during the 1980s, a number of techniques were adopted for network analysis, including quadratic assignment (QUAP) and permutation tests for matrix regression. These made their way into a number of computer packages and have been widely used. In addition, some non-model-based techniques like multi-dimensional scaling, correspondence analysis, and cluster analysis have been adapted for network analysis. Model-based approaches to network estimation and inference, on the other hand, have taken longer to develop.

Progress in model-based statistical methods for network analysis has been based on both theoretical developments and innovations in data collection. On the theoretical side, the goal of modeling the probability that a tie exists between two persons has given rise to the sequence of *p*-models, from the p_1 models first proposed in the late 1970s by Holland and Leinhardt, to the *p** models developed during the 1990s by Wasserman and Pattison. The key statistical advances have involved the definition of the class of exponential random graph (ERG) models that can be used to represent these types of processes (Besag, 1974), and the development of estimation techniques, first maximum pseudolikelihood (Strauss and Ikeda 1990) and then Markov-Chain Monte Carlo methods (Geyer and Thompson 1992, Gilks et al. 1996).

The earliest models in this tradition began by representing modest forms of dependence between the links: reciprocity and transitivity. Holland and Leinhardt (1970, 1981) made impressive progress exploring the effects of these forms of dependence on network structure given the limited computational methods available at the time. Frank and Strauss (1986) took the next logical step, proposing the Markov random graph as a general model for local dependence. The dependence is called Markovian because it extends only one step out from each network tie: two ties are dependent if they share a node, and independent otherwise. Following the work of Besag (1975, 1977), Strauss and Ikeda (1990) solved the problem of estimation by proposing the use of maximum pseudo-likelihood (MPLE), thus making it possible for the first time to estimate these models using standard statistical software. In the last few years, Pattison, Wasserman and Robins (Wasserman and Pattison 1996, Pattison and Wasserman 1999, Robins et al. 1999) have pointed out that the class of exponential random graph models is not restricted to Markovian forms of dependence. This class of models is in fact very flexible and general, capable of representing such things as propensities for larger cycles, small world

patterns and latent groups. For the first time, we have statistical models for generalized spatial or temporal dependence in networks. Estimation issues continue to pose challenges, and we will need to develop a new set of intuitions about graph parameterization, but the main limitation will soon be the availability of data, and that is a big change.

The other mainstream of statistical models developed for networks during the past twenty years was driven by a search for pragmatic approaches to network data collection. Almost all of the methods reviewed above, both mathematical and statistical, require the equivalent of a network census – data on every node and every link in the network of interest. This has been a serious obstacle to network data collection. In the early 1970's, however, a number of studies were designed to collect egocentric or local network data using slight modifications of standard sample survey methods: sample the nodes (egos), and ask them to report on their partners (alters) and the relations they have with these partners (ties). The most ambitious and well-known studies were the Northern California Communities Study (Fischer 1982) and the core discussion partners network module used in the General Social Survey (Burt 1984). Fischer used the local network data to create network attributes (e.g. size, composition), which could then be treated as either response variables or covariates in a traditional linear model. More in keeping with the spirit of the random graph models, other researchers have used local network data to examine biases in the patterns of network partners – what network analysts call homophily, and other fields refer to as non-random mixing, with specific examples being assortative or disassortative mixing (Marsden 1987 and 1988, Burt 1983). Typically, this type of analysis is conducted by forming a "mixing matrix" from the data – a contingency table that cross tabulates the attributes of the respondent (ego) by the attributes of their alter – and using a loglinear model to capture the degree of homophily in the matrix. Net of the dependence induced when

respondents contribute multiple partners to the matrix, the statistical methods here are straightforward—a generalized linear model with a log link function and Poisson errors—and estimation is routine. Overall, this local network approach has proven quite practical, and questionnaire modules have been adapted for studies in sociology (Granovetter 1973), demography (Massey 1990) and epidemiology (Laumann et al. 1989, Morris and Dean 1994, Buve et al. 2001).

To date, the statistical models for complete and local network data have been developed independently. Both, however, are generalized linear models based on the exponential family. For saturated models, the two methods are equivalent, and their fitted probabilities can be directly related via Bayes' rule**.** For non-saturated models they are not perfectly equivalent, but their fitted values when linked via Bayes' rule are likely to be highly similar. The conditions under which equivalence holds, and the reasons for similarity when it does not, help to illuminate the links between the two models, to bridge the gap between local and complete network data, and to make the first steps towards a single coherent statistical framework for modeling networks.

In this paper, we explicate the relationships between random graph models and loglinear mixing models for network data. We then illustrate these using data on a network of school friendships from the National Longitudinal Study of Adolescent Health (Add Health).

2. TERMINOLOGY AND NOTATION

Social network data include a set of social entities, generally referred to as *actors* or *nodes*, and a set of relational measurements, also known as *ties, links, arcs, lines, edges* or *partnerships,* that exist between pairs of those actors on some social relation. In the example we will be using below, actors are individual people and the relation is friendship. The number of

actors in the network will be denoted by n; the fixed set of network actors will be represented by N, where $N = \{1,2,3,\ldots, n\}$. One generally measures some attribute variables on the actors such as sex, ethnic origin, religious affiliation, geographic location, or age. For simplicity, we shall assume for this paper that actors are coded according to a single nominal attribute that can take on K values; the results are easily generalizable to multiple and ordinal attributes. We define the sets C_k for $k = 1$ *to* K, whose elements are all those nodes possessing the k^{th} value of the attribute. (The ordering of attribute values is arbitrary for nominal attributes). The number of actors with attribute k is denoted n_k, so that $n = \sum_{k=1}^{K} n_k$.

Pairs of actors, whether or not they share a relational tie, are referred to as *dyads*. The value of the tie between two actors is denoted by X; for specific actors i,j the random variable is denoted X_{ij}. In the current discussion, we will assume that the tie relation is dichotomous, such that $X_{ij} = 1$ if actors i and j share a tie and $X_{ij} = 0$ if they do not. We define T as the total number of ties in the network.

Relations may be either *directed* or *nondirected*. The relation is *nondirected* if a tie is either present or absent between each actor pair ($X_{ij} = X_{ji}$ for all i,j pairs). A *directed* relation consists of measurements where the orientation of the ties between actors is meaningful. In this case X_{ij} need not equal X_{ji}. An example of a nondirected relationship would be "has sex with"; a directed relationship would be "sells drugs to". With local network data there is often a directionality implied by the study design (separate from the relationship itself), such that respondents may be viewed as "sending" the relationship and their nominated partners as "receiving". We will use an example based on directed data here, but the approach is easily extended to undirected relationships.

Figure 1 depicts three common forms of representation for network data, using a hypothetical directed network containing ten actors identified by location (urban/rural). The first representation is a *graph*, $\boldsymbol{G}$, consisting of a set of nodes joined by lines or arcs. The actors in N are the nodes in the graph. Relational ties are represented graphically by connecting two nodes with a directed line, $i \rightarrow j$, indicating that actor i initiates a relationship towards, or *chooses,* actor j. (Nondirected relationships are typically represented by a nondirected line, i—j.) The network can also be represented in a two-dimensional array called a *sociomatrix* or *adjacency matrix*, denoted by $\boldsymbol{X}$ with elements X_{ij}. If self-relations are disallowed, the main diagonal of the sociomatrix is ignored. For a nondirected relation one may assume that $X_{ji} = X_{ij}$ for all (i,j) pairs, or ignore the lower triangle, restricting analysis to those X_{ij} for which $i < j$. The sociomatrix can be collapsed into a *mixing matrix* or *contact matrix.* Rows and columns of the sociomatrix are aggregated within attribute classes, resulting in a smaller matrix in which cell entries t_{ab} indicate the total number of ties in a network among actor pairs with attributes a and b:

$$t_{ab} = \sum_{i \in a} \sum_{j \in b} X_{ij} \qquad [1]$$

Information about the specific actors involved in the relationships is ignored in this matrix. As with the sociomatrix, the contact matrix is square for a directed relationship and triangular for a nondirected one. Square contact matrices are only used for nondirected data when the population can be divided into two classes and all partnerships are between classes. This is called a bipartite graph (e.g. heterosexual relationships with male race and female race as the margins).

The contact matrix ignores information about the absence of ties. This information can be represented in another matrix, which we will call the "non-contact" matrix. The contact and non-contact matrices form a three-dimensional array, with the three dimensions implying three sets of marginals. We follow the standard notation representing the margins with a plus symbol in the relevant subscript, and we refer to the two attribute dimensions as A and B and the tie/non-tie dimension as Y. Since we assume a directed graph, attribute dimensions A and B refer to the attribute classes of the sender (i) and receiver (j) of the relational tie, respectively.

The marginal table t_{ab+} represents the total number of dyads between two actors with a given attribute combination. In this marginal table A and B are always independent since t_{ab+} represents the number of possible a,b dyads and is simply the product of n_a and n_b for all a,b.[1] This constraint turns out to have important implications both for the patterns of mixing that occur in practice and for the model.

3. MODELING THE GRAPH

Both of the modeling approaches we compare are probabilistic, treating the X_{ij} ties as random variables with realizations x_{ij}. For dichotomous relations, the expected value of X_{ij} is thus equal to $P(X_{ij} = 1)$. A graph in which every potential partnership is independent and has an identical expected value is known as a *Bernoulli graph*. A graph obeys *conditional tie independence* (CTI) if its tie probabilities do not depend on one another given the attributes of the nodes; this model is sometimes referred to as an independence model in the network literature, dropping the "conditional" since complete independence models are rarely of interest. For directed relationships, *dyadic independence (*or more correctly, *conditional dyadic independence)* refers

[1] This is true for all bipartite graphs, while for non-bipartite graphs it is only exactly true in the case where actors are allowed to share a tie with themselves. Otherwise, the number of homophilous dyads (those on the main diagonal of the contact matrix) in a group with n actors equals n^2-n rather than n^2. As n gets large, however, this difference becomes negligible. Since modeling as if on-diagonal relationships were allowed simplifies the analysis considerably and since its effects in large populations are small, we will do so throughout the paper.

to a model in which tie probabilities are dependent on the value of the tie between the same two actors in the opposite direction, but not on other ties given the actor attributes. Otherwise, ties are said to be *conditionally dependent*, analogously shortened to *dependent* in common usage. We will retain the longer but more accurate terms here for clarity. See Frank (1988) for a full discussion.

Nodes *i* and *j* are said to be *homogeneous* if they can be interchanged without affecting the probability of the graph. All nodes are homogeneous in a Bernoulli graph, while the definition of CTI implies that nodes with the same attributes are homogeneous. Homogeneity constraints allow for a more parsimonious representation, but they represent substantive hypotheses that should be considered part of the model.

For the remainder of the paper we assume CTI. We will also assume that the size of the graph (the number of nodes) and its overall attribute composition are fixed, and we will leave these conditions out of our probability statements for simplicity. In the discussion, we will review the ability of different models to relax these assumptions.

3.1 *(Conditional) loglinear models for social mixing*

Loglinear models have long been used to explore mixing matrices in contexts where dyadic independence is assumed, and actors are considered homogeneous by attributes. Conditioning on the presence of a tie means that this approach ignores information about either the non-contact matrix or the size of the population or attribute groups as a whole. We refer to these models as *conditional loglinear models* (CLLs), to distinguish them from the more general approach we discuss below. In this context each tie is a Bernoulli trial whose probability depends only on the attributes of the two actors involved. The cell counts t_{ab1} are the sum of these trials; since we have assumed a fixed population and attribute composition, these cell

counts have a Poisson (if the total number of ties T is not fixed) or multinomial distribution (if it is).

Let π_{ab} denote the fitted probability of a tie falling in cell ($a,b,1$) where $a,b = \{1 \ldots k\}$ are attribute classes.[2] This is the expected count of (a,b,1) divided by the total number of ties (t_{++1}). The saturated CLL model can be expressed as:

$$\log \pi_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_{ab}^{AB} \qquad [2]$$

The first term represents a reference level for tie formation, the next two terms are main effects for the relative levels of tie formation for each group, and the last is an interaction effect for specific attribute pairings. Interaction effects can be used to saturate the model, or they can be constrained to index groups of cells. A simple one-parameter interaction effect is uniform homophily (also known as quasi-independence in the statistical literature), which splits the cells into on- and off-diagonal groups. This parameter can be used to estimate the extent to which actors have a differential tendency to choose partners from their own attribute class (and to test whether this model fits the data). Examples of other interaction effects include differential homophily factors for each diagonal cell, linear or non-parametric distance off the diagonal (e.g., for mixing by age), and single-cell interaction terms (cf. Morris, 1991 for examples).

Identification requires constraints, and the two most common parameterizations are symmetric (or ANOVA) constraints, and first-level constraints. The latter set the first level or category effects for each variable and their interactions equal to zero, thus acting as a baseline

[2] Since loglinear models for partnership data only examine the contact matrix, they generally possess two subscripts. However, there is a third implied subscript representing tie value $y = \{0,1\}$, which is unnecessary in the conditional loglinear framework since it always equals 1. We will follow standard notation here, but it may useful to remember the implied presence of this third subscript.

for interpretation of the parameters associated with the remaining categories (Agresti 2002). The first-level parameterization is used for the examples below. As with all generalized liner models, fitted probabilities and parameter values cannot generally be expressed in terms of the observed cell counts; finding values for them generally requires iterative solutions. If we ignore the potential dependence induced by actors contributing multiple partnerships to the data, the model can be fit using a generalized linear model with a log link and Poisson errors. Otherwise, the model can be fit using generalized estimating equations (GEE, Liang and Zeger 1986) or non-linear mixed effects models.

A shorthand notation is often used to identify specific loglinear models, which will prove useful in later sections. In this notation, a single variable in brackets suggests that a full set of terms for the levels of that variable are included in the right-hand side of the model formula. Two or more variables in brackets implies a full set of interaction terms for those variables, as well as all lower-order terms. Thus the model in Equation [2] can be abbreviated as [AB], since this signifies a full set of AB interaction terms as well as marginal A terms and B terms.

For the saturated first-level constraints model the parameters are defined as :

$$\begin{aligned} \lambda &= \log(\pi_{111}) \\ \lambda_a^A &= \log(\pi_{a11} / \pi_{111}) \\ \lambda_b^B &= \log(\pi_{1b1} / \pi_{111}) \\ \lambda_{ab}^{AB} &= \log\left(\frac{\pi_{ab1}\pi_{111}}{\pi_{a11}\pi_{1b1}}\right) \end{aligned} \qquad [3]$$

A model for only marginal effects of A and B in this framework sets the λ_{ab}^{AB} terms to 0. The remaining parameter values are adjusted accordingly, and the odds ratios for the fitted cell probabilities must satisfy:

$$\frac{\pi_{a_1 b_1 1}\pi_{a_2 b_2 1}}{\pi_{a_1 b_2 1}\pi_{a_2 b_1 1}} = 1, \quad \forall a_1, a_2, b_1, b_2 \qquad [4]$$

This model fits the margins perfectly, but not necessarily the individual cell values.

Between the marginal and saturated models lie a range of non-saturated interaction models, which involve grouping cells into categories representing layers of a given effect. Non-saturated interactions can be thought of as fitting a generalized margin, in the sense that the cells sharing a value for the interaction term will have their sum fit by the model, but not individual cells. The values of these categories I_{ab} can be placed into a *design matrix*, which helps to clarify their relationship to standard marginal models. For example, Table 1 contains the design matrix for a uniform homophily parameter with first level constraints in a four-value attribute, along with the design matrices for the marginal effects parameters. Together these would yield the model:

$$\log \pi_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_{a,b}^{HOM} \qquad \begin{cases} \lambda_{a,b}^{HOM} = \lambda^{HOM}, & a = b \\ \lambda_{a,b}^{HOM} = 0, & a \neq b \end{cases} \qquad [5]$$

The odds-ratios for the fitted cell values must then equal:

$$\frac{\pi_{a_1 b_1 1}\pi_{a_2 b_2 1}}{\pi_{a_1 b_2 1}\pi_{a_2 b_1 1}} = \left(\lambda^{HOM}\right)^{(I_{a_1 b_1 1}+I_{a_2 b_2 1}+I_{a_1 b_2 1}+I_{a_2 b_1 1})}, \quad \forall a_1, a_2, b_1, b_2 \qquad [6]$$

For instance, for the four cells defined by a_1=2, a_2=3, b_1=1, b_2=2, only $I_{a_1 b_2 1}$ is on the diagonal, and the odds ratio would equal

$$\frac{\pi_{a_1 b_1 1}\pi_{a_2 b_2 1}}{\pi_{a_1 b_2 1}\pi_{a_2 b_1 1}} = \left(\lambda^{HOM}\right)^{(0+0-1-0)} = \left(\lambda^{HOM}\right)^{-1} \qquad [7]$$

Although the statistical literature on loglinear models is extensive, there is comparatively little on non-saturated interaction models, despite widespread use of such models in the social sciences.

All of the above model parameterizations provide estimates for the π_{ab1} values, which represent $P(i \in C_a, j \in C_b \mid x_{ij} = 1)$. Fitted cell counts for the tie matrix (m_{ab1}) are found by multiplying these probabilities by T.

3.2 *Random Graph Models for Social Networks*

Conditional tie independence ERG models reverse the conditioning of CLL, modeling the probability that actors share a tie given that they possess certain attributes. ERG models use both the tie matrix and the non-tie matrix, treating the tie dimension as an outcome variable and modeling the log-odds that it is present. Population size and attribute composition are exogenously given in this model so the total number of dyads of each attribute combination is fixed.

The ERG model represents the probability function of the random graph ***G,*** defined by the sociomatrix ***X,*** as a linear combination of network statistics:

$$P(\mathbf{X} = \mathbf{x}) \;=\; c^{-1} \exp\{\theta' \mathbf{z}(\mathbf{x})\} \qquad [8]$$

with

$$c = \sum_{\text{all } G} \exp\{\theta' \mathbf{z}(\mathbf{x})\} \qquad [9]$$

(Besag 1974). The vector **z**(**x**) represents a set of network configurations. The θ parameters represent the unknown weights of the linear function of network properties. The normalizing constant c is needed to ensure a proper probability distribution. Any dyad-based measure from the network may be included in **z**(**x**), although typically sums or sums of products of X_{ij} are used.[3]

For CTI with homogeneity constraints, the model statistics are the total number of ties and the number of ties between members of the attribute classes:

$$P\left(\mathrm{X} = \mathrm{x}\right) = c^{-1} \exp\left\{\theta z + \sum_{a=1}^{K} \theta_a^A z_a^A + \sum_{b=1}^{K} \theta_b^B z_b^B + \sum_{a=1}^{K}\sum_{b=1}^{K} \theta_{ab}^{AB} z_{ab}^{AB}\right\} \qquad [10]$$

where z = the total number of ties in the network, z_a^A = the number of ties initiated by actors in attribute class C_a, z_b^B = the number of ties received by actors in attribute class C_b, and z_{ab}^{AB} = the

[3] Examples include nodal degrees, the number of within-group ties (analogous to uniform homophily), or the number of transitive triads ($X_{ij} = X_{jk} = X_{ki} = 1$).

number of ties initiated by actors in C_a and received by actors in C_b. The θ, θ_a^A, θ_b^B and θ_{ab}^{AB} are the coefficient on each term.

By definition, the probability of the graph under CTI is simply the product of the probability of the value of each dyad:

$$P\left(\mathbf{X}=\mathbf{x}\middle|C_1,...C_K\right)=\prod_{i,j}P\left(X_{ij}=x_{ij}\middle|i\in C_a,j\in C_b\right) \qquad [11]$$

To derive these individual dyad probabilities, we define $\mathbf{x}_{ij}^{+}$, $\mathbf{x}_{ij}^{-}$, and $\mathbf{X}_{ij}^{C}$ respectively as the realization of $\mathbf{X}$ with x_{ij} set equal to 1, the realization of $\mathbf{X}$ with x_{ij} set equal to 0, and the realization of $\mathbf{X}$ with X_{ij} coded as missing. The conditional log-odds of a tie between actor *i* and actor *j,* given the rest of the data, is represented in this framework as:

$$\log\left(\frac{P(X_{ij}=1\,|\,\mathbf{X}_{ij}^{C})}{P(X_{ij}=0\,|\,\mathbf{X}_{ij}^{C})}\right)=\log\left(\frac{\exp\left\{\theta'\mathbf{z}(\mathbf{x}_{ij}^{+})\right\}}{\exp\left\{\theta'\mathbf{z}(\mathbf{x}_{ij}^{-})\right\}}\right)=\theta'\boldsymbol{\delta}_{ij} \qquad [12]$$

where $\mathbf{z}(\mathbf{x}_{ij}^{+})$ and $\mathbf{z}(\mathbf{x}_{ij}^{-})$ represent the vector of network statistics evaluated from $\mathbf{x}_{ij}^{+}$ and $\mathbf{x}_{ij}^{-}$, respectively. The $\boldsymbol{\delta}_{ij}$ terms represent the difference between $\mathbf{z}(\mathbf{x}_{ij}^{+})$ and $\mathbf{z}(\mathbf{x}_{ij}^{-})$, the change in the network statistics when the tie between actor *i* and actor *j* is toggled from 1 to 0. These conditional logit P(X_{ij}) values can then be converted to $P(X_{ij}=1\,|\,\mathbf{X}_{ij}^{C})$ or $P(X_{ij}=0\,|\,\mathbf{X}_{ij}^{C})$.

Under CTI, unbiased estimates for the θ's can be obtained from logistic regression with the observed x_{ij} values as the outcome variable and the δ_{ij}'s as the predictors (Strauss and Ikeda 1990). This is a generalized linear model with a logit link function and binomial errors.

When reframed in logit form for an individual tie, Eq. [10] reduces to:

$$\text{logit } P\left(X_{ij}=1 \middle| i \in C_a, j \in C_b\right) = \theta + \theta_a^A + \theta_b^B + \theta_{ab}^{AB} \qquad [13]$$

Identifiability again requires setting some parameters equal to zero. This model can also be abbreviated as [AB], indicating that the right-hand side of the equation contains a similar set of terms as in the saturated CLL model. The left-hand side of the equation is different, however.

A marginal effects model in this context involves setting the AB interaction terms to 0:

$$\text{logit } P\left(X_{ij}=1 \middle| i \in C_a, j \in C_b\right) = \exp\left\{\theta + \theta_a^A + \theta_b^B\right\} \qquad [14]$$

This is commonly referred to as a model of independence for A and B, but it is not the same as the independence model for the CLL. While the logit is now an additive function of row and column effects alone, *A* and *B* are not independent conditional on Y. The model instead implies:

$$\frac{\pi_{a_1 b_1 1}\pi_{a_2 b_2 1}}{\pi_{a_1 b_2 1}\pi_{a_2 b_1 1}} = \frac{\pi_{a_1 b_1 0}\pi_{a_2 b_2 0}}{\pi_{a_1 b_2 0}\pi_{a_2 b_1 0}} \qquad [15]$$

We will draw out the implications further below.

The corresponding ERG uniform homophily model is:

$$\text{logit } P\left(X_{ij}=x_{ij} \middle| i \in C_a, j \in C_b\right) = \exp\left\{\theta + \theta_a^A + \theta_b^B + \theta_{ab}^{HOM}\right\}. \qquad [16]$$

As with loglinear models, design matrices can be used to construct and interpret non-saturated interaction models. The odds ratios here are:

$$\frac{\pi_{a_1b_1 1}\pi_{a_2b_2 1}}{\pi_{a_1b_2 1}\pi_{a_2b_1 1}} = \frac{\pi_{a_1b_1 0}\pi_{a_2b_2 0}}{\pi_{a_1b_2 0}\pi_{a_2b_1 0}} * \left(\theta^{HOM}\right)^{(I_{a_1b_1}+I_{a_2b_2}-I_{a_1b_2}-I_{a_2b_1})} \qquad [17]$$

3.3 *Linking ERG models and conditional loglinear models*

Loglinear models predict $P(i \in C_a, j \in C_b \mid X_{ij} = 1)$, while ERG models predict $P(X_{ij} = 1 \mid i \in C_a, j \in C_b)$. They are related by Bayes' rule:

$$P(i \in C_a, j \in C_b \mid X_{ij} = 1) = \frac{P(X_{ij} = 1 \mid i \in C_a, j \in C_b)\, P(i \in C_a, j \in C_b)}{P(X_{ij} = 1)} \qquad [18]$$

The two conditional probabilities are linked by the two marginal probabilities for ties and attributes: P(X_{ij} = 1) is the fraction of all dyads in the network that have a tie, and $P(i \in C_a, j \in C_b)$ is the joint distribution of nodal attributes for all dyads. Bayes' rule thus provides a simple explicit expression for transforming the predicted conditional probabilities from one model to that of the other.[4]

We will define models as *equivalent* when this transformation yields identical probabilities, and therefore identical fitted cell counts. Due to the nature of the conditioning in

[4] The distinction in conditioning is similar to those observed in a series of papers by Robins, Pattison, and Elliott (2001a, 2001b) that distinguish between social influence and social selection. Social influence models assume that network structure can influence individual characteristics (like beliefs). The probability of the attribute is conditional on the tie, as with the CLL. A social selection models assumes that individuals select partners based on attributes. This is analogous to the process underlying the CTI-based ERGMs.

each model, it turns out that the only equivalent models by this definition are fully saturated models.

Since both CLLs and ERG models are generalized linear models, it would be natural to expect that models in one class would have an explicit representation in the other. However, non-saturated models from each class that appear comparable in terms of predictors in fact yield different outcomes. Intuitively, this is because non-saturated ERGM use information from the non-tie layer to fit values in the tie layer, and vice versa. The CLL ignores the information in the non-tie layer, so in general a non-saturated ERG model will result in different fitted cell values than any CLL.

Unconditional loglinear models (ULLs), which form a bridge between the CLL and the ERG model, help to make this clearer. The ULL does not condition on the presence of a tie, but rather considers all three dimensions (A,B,Y) as predictors with cell counts or probabilities as outcome. The saturated ULL [ABY] is represented as:

$$\log \pi_{aby} = \gamma + \gamma_a^A + \gamma_b^B + \gamma_y^Y + \gamma_{ab}^{AB} + \gamma_{ay}^{AY} + \gamma_{by}^{BY} + \gamma_{aby}^{ABY} \qquad [19]$$

Its parameters under first-level constraints are[5]:

[5] Note that whereas *A* and *B* take values {1…k}, *Y* takes values {0,1}. Thus the levels in which parameter values are set to zero in the first-level constraints model are *A*=1, *B*=1, and *Y*=0.

$$
\begin{aligned}
&\gamma = \log\left(\pi_{110}\right) \\
&\gamma_a^A = \log\left(\pi_{a10}/\pi_{110}\right),\ \gamma_b^B = \log\left(\pi_{1b0}/\pi_{110}\right),\ \gamma_1^Y = \log\left(\pi_{111}/\pi_{110}\right) \\
&\gamma_{ab}^{AB} = \log\left(\pi_{ab0}\pi_{110}/\pi_{a10}\pi_{1b0}\right) \\
&\gamma_{a1}^{AY} = \log\left(\pi_{a11}\pi_{110}/\pi_{a10}\pi_{111}\right) \\
&\gamma_{b1}^{BY} = \log\left(\pi_{1b1}\pi_{110}/\pi_{1b0}\pi_{111}\right) \\
&\gamma_{ab1}^{ABY} = \log\left[\left(\pi_{ab1}\pi_{111}/\pi_{a11}\pi_{1b1}\right)/\left(\pi_{ab0}\pi_{110}/\pi_{a10}\pi_{1b0}\right)\right]
\end{aligned}
\qquad [20]
$$

All ERG models with CTI correspond to a 3-way ULL that contains the following terms and no other (Agresti 2002, p. 332):

- a full set of AB interaction terms;
- a Y marginal term;
- every term in the ERG model;
- every term in the ERG model crossed by Y.

The AB interaction terms in the ULL ensure that the cells in the t_{ab+} marginal matrix are fit exactly. These establish the population size, marginal attribute composition, and the number of dyads (not ties) among attribute groups. Any equivalent ULL must have this [AB] term in the model, because population size and composition are exogenous to the ERG model.

Each CLL corresponds to a ULL containing the following terms:

- a Y marginal term;
- every term in the CLL model;
- every term in the CLL model crossed by Y.

Here the *Y* term sets the number of ties, and each of the terms crossed by *Y* allows the CLL terms to be represented in the ULL tie layer independently of the non-tie layer.

Table 2 lays out the set of equivalencies among ERG models, ULLs, and CLLs. We use the symbol U_{AB} as a general symbol representing any set of non-saturated interaction terms between *a* and *b*. Note that non-saturated ERG models have no corresponding CLL. We can now formalize our earlier intuition about this fact by arguing deductively from the two sets of equivalence rules above.[6]

To make these differences clearer, we highlight the most familiar model to many, that of marginal effects, in the two frameworks; Table 2 makes explicit how these differ. While this is the model that we commonly think of as implying that A and B are independent, independence clearly means different things in the two models. For CLL, it means that A and B are independent conditional on Y. For ERG models, independence means "no 3-way association"; all three variables are pairwise dependent, but each pair is conditionally independent given the third. This does not mean that A and B are independent in either layer of Y; instead, the pattern of dependence is the same in each layer. The difference is also evident when comparing the fitting constraints, Eq. [4] for the CLL, and Eq. [15] for the ERG model.

The model of "no 3-way association" is one of the most difficult to interpret in practice, yet it corresponds to the basic marginal effects ERG model. There is no simpler definition in the ERG context because there is an implicit constraint that A and B are independent in the marginal matrix of all dyads, m_{ab+}. Since the tie and non-tie matrices must sum to this marginal matrix, the cell values in one layer determine the other when the totals are fixed. The two layers can

[6] Imagine that there exists some ERGM with an equivalent CLL. The ULL that is equivalent to this ERGM must contain an [AB] interaction term. If the ULL contains [AB] then its CLL equivalent must also contain [AB]. If the CLL contains [AB] then the ULL must contain [ABY], which means it is fully saturated.

only exhibit conditional independence of A and B given Y for both Y=1 and Y=0 under a narrow range of conditions: if either all of the sender or all of the receiver attribute groups are homogeneous with respect to tie formation. (When both groups are, or one group is and ties are undirected, this degenerates into the Bernoulli model). In essence, the additional implicit constraint m_{ab+} creates an inverse form of Simpson's paradox; two attributes are independent in the marginal table, but when stratified by a third variable (here, tie value), they are not independent in each stratified table.

In practice, however, the fitted values from the two marginal effects models are likely to be similar. Social networks for populations of reasonable size are generally quite sparse, because the number of ties in a population generally scales roughly with the population size, while the number of dyads varies with the square of population size. If almost all dyads have $Y = 0$, then we can assume $\pi_{ab0} \cong \pi_{ab+}$ for all a,b. Thus

$$\frac{\pi_{ab0}\pi_{110}}{\pi_{a10}\pi_{1b0}} \approx \frac{\pi_{ab+}\pi_{11+}}{\pi_{a1+}\pi_{1b+}} \approx 1. \qquad [21]$$

This means the right hand side of Eq. [15] is approximately equal to 1 for sparse matrices, reducing it to Eq. [4], and implying that the two models will yield approximately equal results. Bayes' rule can be used to transform the results from one model to the other to determine the magnitude of the difference. In our experience, sparse matrices of at least a few hundred people yield marginal models in which cell counts differ by at most a tenth of a partnership. For non-sparse matrices from small settings such as an office or classroom, the differences will be small as long as the sizes and activity levels of the different attribute classes are roughly equal.

3.4 *Parameter equivalencies*

By combining Eqs. [3] and [20], the parameters for the saturated CLL can be represented as sums of the parameters from its corresponding ULL:[7]

$$
\begin{aligned}
\lambda &= \log(\pi_{111}) = \log\left(\pi_{110}\right) + \log\left(\pi_{111}/\pi_{110}\right) = \gamma + \gamma_1^Y \\
\lambda_a^A &= \log(\pi_{a11}/\pi_{111}) = \log(\pi_{a10}/\pi_{110}) + \log\left(\pi_{a11}\pi_{110}/\pi_{a10}\pi_{111}\right) = \gamma_a^A + \gamma_{a1}^{AY} \\
\lambda_b^B &= \log(\pi_{1b1}/\pi_{111}) = \log\left(\pi_{1b0}/\pi_{110}\right) + \log\left(\pi_{1b1}\pi_{110}/\pi_{1b0}\pi_{111}\right) = \gamma_b^B + \gamma_{b1}^{BY} \\
\lambda_{ab}^{AB} &= \log\left(\frac{\pi_{ab1}\pi_{111}}{\pi_{a11}\pi_{1b1}}\right) = \log\left(\frac{\pi_{ab0}\pi_{110}}{\pi_{a10}\pi_{1b0}}\right) + \log\left[\left(\frac{\pi_{ab1}\pi_{111}}{\pi_{a11}\pi_{1b1}}\right)\Big/\left(\frac{\pi_{ab0}\pi_{110}}{\pi_{a10}\pi_{1b0}}\right)\right] = \gamma_{ab}^{AB} + \gamma_{ab1}^{ABY}
\end{aligned}
\qquad [22]
$$

We can also reformulate the saturated ULL into a logit model that models the log odds of a tie:

$$
\log\frac{\pi_{ab1}}{\pi_{ab0}} = \log\pi_{ab1} - \log\pi_{ab0} = \gamma_1^Y + \gamma_{a1}^{AY} + \gamma_{b1}^{BY} + \gamma_{ab1}^{ABY} \qquad [23]
$$

Each ERG model parameter in the fully saturated model is equivalent to the parameter on that term crossed by Y in the ULL. These relationships between parameter values also hold for non-saturated models, as will be seen in the following example.

4. EXAMPLE: THE ADD HEALTH STUDY

[7] If we had formulated the ULL parameters in reverse so that *Y=1* was the reference category instead of *Y=0*, then the CLL parameters would have been identical to a subset of the ULL parameters.

We use the friendship nomination data from the first wave of the National Longitudinal Study of Adolescent Health (Add Health) to demonstrate the results above. Add Health is a nationally representative study of students in grades 7 through 12, and the first wave was conducted in 1994-1995. The study was school-based, and students were provided with a roster of all students in the school and asked to select up to five close male friends and five close female friends. Complete details of this and subsequent waves of the study can be found in Resnick et al. (1997) and Udry and Bearman (1998) and at *http://www.cpc.unc.edu/projects/addhealth*.

We will use friendship data from one school comprising 71 students divided into six grades (with 15, 13, 16, 10, 13 and 4 students in grades 7 through 12, respectively). The ties are directional since it is possible person A could name B as a friend without B nominating A. The limit on nominations means that the data are not complete, but we will assume for convenience that a lack of nomination in these data means that there is no friendship.

Figure 2 provides a graph of the data, while Table 3 shows the corresponding contact and non-contact matrices. We begin by fitting a CLL and an ERG model with main effects only and the ULL that corresponds to each. A quick glance at Figure 2 makes it clear that there is a strong preference for students of all grades to nominate friends in their own grade; we thus also run the CLL and ERG models for main effects with uniform homophily. In each case, a first-level constraint was used, and both were fit using the *glm* macro in R (Ihaka and Gentleman 1996).

The parameter estimates for the marginal effects models are shown in Table 4. These allow for a comparison of the CLL and ERG model to their respective ULL parameterizations. In the case of the ERG model, the ULL parameter values in the first column of Table 4 are those that fit the t_{ab+} cells exactly; there are no corresponding parameters on the ERG model side

because these values are conditioned on. Note that all 25 (=[a-1]*[b-1]) of the AB interaction terms in the ULL list are very close to 0 and are described by summary statistics rather than enumerated; they are modeling the log of the ratio of odds ratios between the tie and non-tie matrix, and those ratios are all very close to 1 for the reason explained in the previous section (see Eq. [21]). The ULL parameters in the second column (those that represent the patterns in the Y = 1 layer) have values equal to the ERG model parameters.

For the CLL marginal effects model, the corresponding ULL model does not have any [AB] interaction parameters. There are exactly twice as many parameters in the ULL parameterization as in the CLL model, since the ULL model is fitting both layers; the first column of the ULL values fits the appropriate independence model in the non-tie layer, while the second column then fits independence in the tie layer. With the first-level parameterization, each CLL parameter equals the sum of the two parameters in the ULL in the same row in Table 4.

Note the strong similarity between the ULL parameters for the two models, despite the fact that they are not identical. We can also see the similarity between the models by applying Bayes' rule to calculate the fitted probabilities of having a nodal attribute composition given the presence of a tie. Take the example of a 7th grade sender and 8th grade receiver; under the marginal effects ERG model, Eq. [18] would yield:

$$P\left(i \in C_7, j \in C_8 \mid X_{ij} = 1\right) = \frac{\dfrac{\exp\left(-2.895 - 0.102\right)}{1 + \exp\left(-2.895 - 0.102\right)} * \left(\dfrac{5 + 190}{305 + 4736}\right)}{\left(\dfrac{305}{305 + 4736}\right)} = 0.0304 \qquad [24]$$

The marginal effects CLL model gives this value directly as:

$$P\left(i \in C_7, j \in C_8 \mid X_{ij} = 1\right) = \frac{\exp\left(2.468 - 0.240\right)}{305} = 0.0304 \qquad [25]$$

Other combinations are also similar; despite being different models, the fitted cell probabilities obtained from the two are generally equal down to the fourth decimal place.

Figure 2 made clear the strong tendency for ties to be homophilous by grade. A likelihood ratio test confirms that adding a single uniform homophily parameter significantly improves the model fit in either modeling framework. Using the uniform homophily parameters accentuates the differenced between the fitted cell probabilities of the two frameworks, although the differences are still on the order of the third decimal place. Thus, even for relatively small, dense social networks, the practical differences between the two modeling frameworks are not very large.

5. DISCUSSION

Conditional loglinear models and ERG models are both generalized linear models based on the exponential family that can be used to represent attribute mixing in networks, but they condition on different aspects of the data. CLLs condition on the tie being present, and model the patterns in partner selection, while CTI ERG models condition on the attribute composition of the population, and model the distribution of ties and non-ties. Which of these is a better model of social behavior? It depends on the application. For large populations in which people form ties with only a small fraction of possible partners, it seems reasonable to assume that the non-ties (or at least the great majority of them) are not explicitly chosen. In this case, the CLLs would be a reasonable choice. In small settings such as school or offices or isolated populations, the patterns of non-ties (do not collaborate, do not get along, can not marry) may be as

intentionally chosen as the ties. Here ERG models may be a better choice. Whatever the relative theoretical merits of each model, however, the similarity of the fitted values in practice suggests that there is little to be gained by selecting the model based on the form of conditioning. That leaves one free to choose on other grounds.

One of the other grounds to consider is the flexibility of the modeling framework. In this paper, attention has been limited to the restricted set of "comparable" models, so that the similarities and differences between the frameworks can be clearly identified. The set of comparable models share two assumptions: fixed population attribute composition and dyadic (tie) independence. In practical applications these assumptions may be a severe handicap. For example, when network models are used in a dynamic context, such as modeling the transmission of HIV through a network of partnerships, it is often desirable to allow for population composition changes (for example, to allow for group specific infection and mortality rates). The CLLs make it relatively easy to relax fixed attribute composition, and replace it with the much weaker assumption that population sizes and preferences are separable. This is because the CLL mixing parameters are specified in terms of odds ratios, which allows the margins to change independently of selection patterns, making it straightforward to integrate exogeneously changing population sizes into a dynamically changing mixing matrix (Morris, 1991). Currently, it is not clear how this would be accomplished in the ERG modelling framework. On the other hand, it is often important to be able to relax the dyadic independence assumption to be able to model such things as temporal dependence in dyads. A good example in the same HIV transmission context is the rule of serial monogamy in sexual partnerships, which imposes a very strong form of dyadic temporal dependence: the probability of a link between two nodes is zero if either node is already linked to another. Here ERG models have the

clear advantage, as they can model forms of dyadic dependence explicitly (Wasserman and Pattison, 1996; Pattison and Wasserman, 1999). Because non-ties cannot be modeled explicitly in the CLL, there is no way to represent this form of dependence.

The ability to represent dyadic dependence is the reason that ERG models have attracted so much interest in contemporary network analysis. Dependence among social ties has always been the theoretical heart of network analysis – from balance theory and cognitive networks to kinship structure and role algebras. While most applications of ERG models have focused on locally connected subsets of the graph (Markov graphs), ERG models can incorporate a much wider range of interdependence, including global network properties such as connectivity, centrality and distance. Using Markov Chain Monte Carlo algorithms for estimation, ERG models also place the problem of inference for conditional dependence on a firm statistical footing. This has important practical implications, as it enables one to test whether dyadic independence, or limited forms of dependence, provide a reasonable fit to the data in particular contexts. When the answer to this question is yes, network structure may still be present (as in attribute mixing), but the data requirements for estimating such structural parameters are much simpler.

This raises another important difference between the two models: the data required for estimation. Data requirements have played an important role in limiting the use and development of network analytic methods (Morris, *in press*). ERG models currently require data on the complete network: measurements on all individuals and ties within the bounded group of inference. This "census" requirement, also a feature of traditional network analytic methods, places a huge burden on data collection. It severely limits the range of potential applications because complete network data are rarely available, except in cameo settings (like

the Sampson Monastary). The Add Health data we use here is an example of nearly complete network data collected in a large scale applied setting, but there are few studies of this kind (but cf. also Rindfuss et al., *in press*). The interest in email and internet networks is in part a reflection of this constraint. Electronic information exchange is one of the few large-scale settings in which complete network data collection is feasible. But it would be a shame if the data constraints of the modeling framework limited network analysis to a few unusual settings like this.

In most contexts where network approaches could be applied to real world problems, sampling cannot be avoided. Sampling options for networks range from the local "egocentric" network designs used to study friendship networks in the GSS and the Northern California Communities Study (Fischer, 1982), to various forms of partial network sampling based on adaptive sample designs (snowball samples, random walks, chain-link samples, respondent referral samples, etc.). The distinctive feature of network samples is that the sampling is done through the questionnaire. Local network designs have three central components: alter elicitation ("name generators"), questions asked about each alter ("name interpreters"), and links among the alters (the "egonet" matrix). This allows for a wide range of designs that are well described in the Methodological Appendix of Fischer's book. Partial network designs have an additional two components: selection of alters for tracing and enrollment, and the number of waves of enrollment. Here, too, there are many possible designs, from every possible alter in one wave, to one alter in each wave until there are no more links to follow.

Local network data can be analyzed fairly easily if based on a random sample of respondent "egos", because the sample design can be incorporated through the use of weights and appropriate adjustments for clustering. If the network information is summarized at the

respondent level (e.g., network size, heterogeneity, density, mutuality), then the issue of dyadic dependence is avoided, and any standard statistical technique can be employed. Examples include the regression-based analyses in Fischer's study, and many of the studies based on the Add Health survey. If the data are analyzed so that the dyad is the unit of analysis, then the individual respondent may contribute multiple dyads to the sample. If this dependence is not strong, it may be ignorable. For example, assortative age mixing in friendships may be due to a general preference for similarity, rather than a sequential effect that makes one's preference for the second partner depend on one's experience with the first. If this dependence instead needs to be modeled, techniques like generalized estimating equations (GEE) or mixed-effect models can be used. Partial network data pose more of a problem for analysis, since it is often difficult to determine the inclusion probability for the nodes and links, and therefore difficult to make proper inference from the sample to the population. Design-based approaches work for some sampling schemes known as "adaptive samples" (Thompson and Seber, 1996). Progress is beginning to be made on the general problem of model-based inference for network sampling. A model-based approach integrates the sampling mechanism into the likelihood function, using the methods similar to those developed for missing data analysis.

It is interesting to note that the conditioning on partnerships in the CLL is similar to the idea presented by Pattison and Robbins (2002) applying neighborhood constraints on ERG models. These neighborhood constraints may provide a more realistic representation of social behavior and the patterns of social relations. Thus, using the sampling process to define social neighborhoods, or setting structures, may be a fruitful avenue of future investigation (cf Koehly, Peterson, Watts, et al., 2003).

One of the primary strengths of the CLL models is that they can be used on local network samples. Local network data collection is the simplest and least expensive, and it is easily integrated into standard survey methodology. As a result, local network data have now been collected in a wide range of applied settings, from discussion networks in the GSS to sexual partnership networks in countries around the world. While the CLLs are limited to analyzing attribute-based mixing patterns, local network data contain more information than this. For example, one can observe the timing and sequence of partnerships by asking about their start and end dates. The limited temporal dyadic dependence embedded in these data will become accessible to modeling once the model-based sampling approaches referred to above have been fully developed.

The difference between these two statistical frameworks for modeling networks currently poses a difficult choice for network analysts: flexibility to model population changes while constrained to dyadic independence, or flexibility to model dependence constrained by static population composition. This would be a sorry state of affairs if it were not such an improvement over the recent past. Statistical models for networks have been long in coming, and the choice now afforded, even if not ideal, is a sign of what is soon to come. In short order, the constraints of the current frameworks will be eclipsed by developments in both network sampling and modeling. Much of the impetus for current work is coming from the need for network models in applied settings. Especially in the field of HIV/AIDS prevention, the value of a network perspective is tremendous, providing a unique and powerful perspective on the dynamics of transmission and the opportunities for prevention. The spillover effects of resources devoted to this aim are making it possible for basic network methods to develop at a remarkable pace. The theoretical perspective afforded by network analysis has always been regarded by

some as the heart of a true social theory – because it makes the relation, and the positions defined by relations, the unit of analysis. It has taken some time for the methods of empirical analysis to reach the level of the theory. While we are still some years from having a viable statistical framework that can pose a challenge to the linear regression model, it now seems like just a matter of time. The links with developments in social psychology, political science and economics, and in particular the growing literature in evolutionary game theory and the general study of agent-based interacting dynamic systems (Bowles, 2003; Gintis, 2000; Macy, 2002; Majeski, 1999; Padgett, 1993; Parker, 2003; Watts, 2003), suggest that the social sciences are converging in a remarkable way to the empirical study of dependent relations.

Author Note

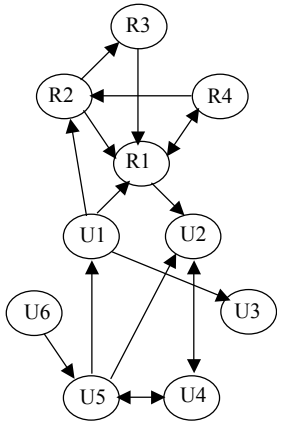
Laura M. Koehly, Department of Psychology, Texas A&M University; Steven M. Goodreau, Center for AIDS and STD and Center for Statistics and the Social Sciences, University of Washington; Martina Morris, Department of Sociology, Center for Statistics and the Social Sciences, University of Washington. The first two authors contributed equally to the intellectual content of this manuscript.

This research was supported by NIH grants HD34957 and DA12831. Portions of this paper were presented at the 2000 IUSSP Conference on Partnership Networks and the Spread of HIV and Other Infections, Chiang Mai, Thailand; the 2000 Annual Meeting of the International Network for Social Network Analysis, Vancouver, BC; and the 2001 annual meeting of the methodology section of the American Sociological Association, Minneapolis, MN.

Correspondence concerning this article should be addressed to Laura M. Koehly, Department of Psychology, Texas A&M University, College Station, TX 77843-4235. e-mail: koehlyl@tamu.edu.

FIGURE 1: Representations of network data

Graph



Sociomatrix

	R1	R2	R3	R4	U1	U2	U3	U4	U5	U6
R1	-	0	0	1	0	1	0	0	0	0
R2	1	-	1	0	0	0	0	0	0	0
R3	1	0	-	0	0	0	0	0	0	0
R4	1	1	0	-	0	0	0	0	0	0
U1	1	1	0	0	-	0	1	0	0	0
U2	0	0	0	0	0	-	0	1	0	0
U3	0	0	0	0	0	0	-	0	0	0
U4	0	0	0	0	0	1	0	-	1	0
U5	0	0	0	0	1	1	0	1	-	0
U6	0	0	0	0	0	0	0	0	1	-

Contact matrix and non-contract matrix

Y=1 (tie)

	R	U	t_{i+1}
R	6	1	7
U	2	8	10
t_{+j1}	8	9	17

Y=0 (no tie)

	R	U	t_{i+0}
R	6	23	29
U	22	22	44
t_{+j0}	28	45	73

FIGURE 2: Add Health Friendship Data, by grade

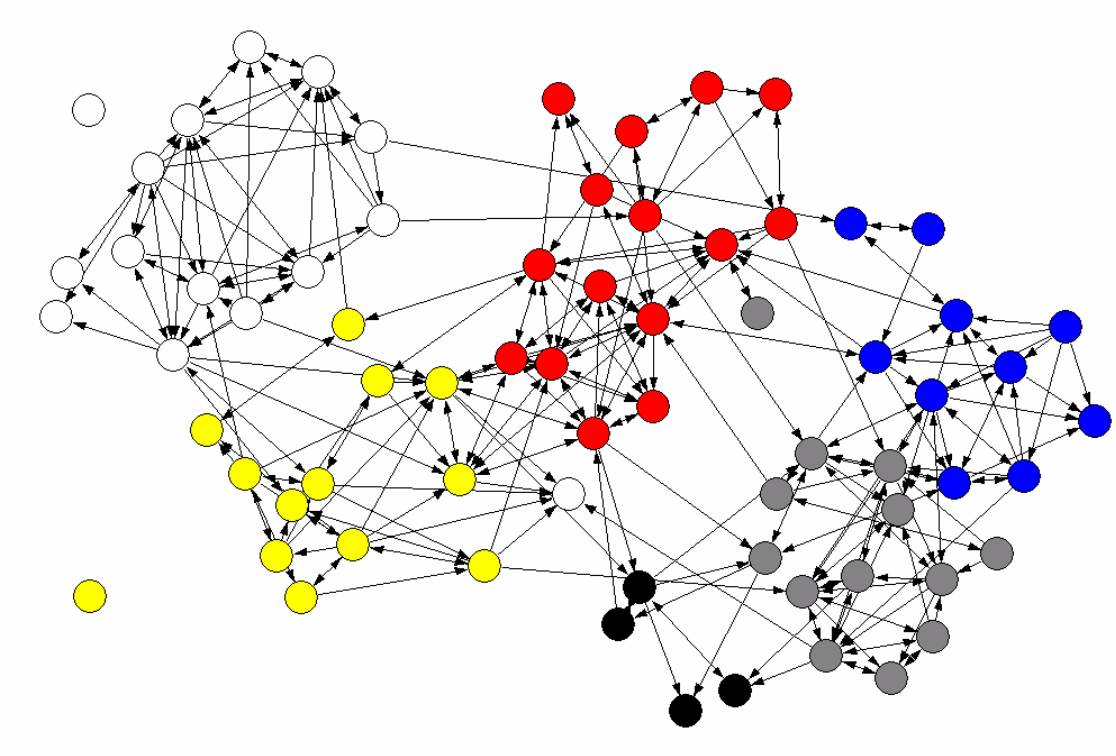


TABLE 1: Design matrices, 4x4 table

Design matrix for uniform homophily

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Implicit design matrices for marginal effects with first-level constraints

A=2

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

A=3

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

A=4

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

B=2

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

B=3

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

B=4

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

TABLE 2: Corresponding models

Name	**Cond. loglinear model** *cell count in layer Y=1 is a function of:*	**Uncond. loglinear model** *cell count is a function of:*	**ERG model** *logit(Y) is a function of:*
Saturated	[AB]	[ABY]	[AB]
Bernoulli graph		[AB][Y]	[-]
Marginal Effects (ERG model) (i.e. no 3-way interaction)		[AB] [AY] [BY]	[A][B]
Marginal Effects (CLL) (i.e. independence of A and B conditional on Y)	[A][B]	[AY][BY]	
Non-saturated interaction (ERG model)		[AB] [AY] [BY] [U_{AB} Y]	[A][B] and [U_{AB}]
Non-saturated interaction (CLL)	[A][B] and [U_{AB}]	[AY] [BY] [U_{AB} Y]	

Notation follows Fienberg (1977) and many others. [X] refers to terms for each value of variable X. [XY] refers to a full set of interaction terms for X by Y, as well as terms for each level of X alone and of Y alone. U_{XY} ("U" for "unsaturated") indicates that some but not all of the set of interaction terms are included in the model (e.g. uniform homophily). Any interaction term implies that all lower order terms are included as well. The model pairs enclosed in each square make clear the lack of equivalence between ERG models and CLLs. In each case, the ULL that corresponds to the ERG model contains an [AB] interaction term that is missing from the ULL corresponding to the CLL. Although all of the other terms are identical between the two models, the presence or absence of the [AB] terms change the values and interpretations of the others.

TABLE 3: Add Health: Reported friendships and imputed non-friendships by grade of nominator and nominee for one school

Friendships

		Grade of nominee						
		7	**8**	**9**	**10**	**11**	**12**	
	7	52	5	*	*	*	*	59
	8	8	33	9	*	*	*	52
Grade	**9**	*	10	70	*	4	*	86
Of	**10**	*	*	3	30	10	*	43
Nominator	**11**	*	*	*	7	43	4	57
	12	*	*	*	*	*	5	8
		61	48	86	39	60	11	305

Non-friendships

		Grade of nominee						
		7	**8**	**9**	**10**	**11**	**12**	
	7	173	190	239	149	195	60	1006
	8	187	136	199	130	168	51	871
Grade	**9**	240	198	186	159	204	63	1050
Of	**10**	150	130	157	70	120	40	667
Nominator	**11**	194	169	206	123	126	48	866
	12	60	52	63	40	50	11	276
		1004	875	1050	671	863	273	4736

* = value < 3

TABLE 4: Parameter values for Add Health, marginal effects models with corresponding ULL models

ERG marginal effects model

ULL				ERG model	
γ	5.362	$\gamma^{Y}_{y=1}$	-2.895	θ	-2.895
$\gamma^{A}_{a=8}$	-0.144	$\gamma^{AY}_{a=8,y=1}$	0.018	$\theta^{A}_{a=8}$	0.018
$\gamma^{A}_{a=9}$	0.044	$\gamma^{AY}_{a=9,y=1}$	0.335	$\theta^{A}_{a=9}$	0.335
$\gamma^{A}_{a=10}$	-0.411	$\gamma^{AY}_{a=10,y=1}$	0.095	$\theta^{A}_{a=10}$	0.095
$\gamma^{A}_{a=11}$	-0.150	$\gamma^{AY}_{a=11,y=1}$	0.116	$\theta^{A}_{a=11}$	0.116
$\gamma^{A}_{a=12}$	-1.295	$\gamma^{AY}_{a=12,y=1}$	-0.706	$\theta^{A}_{a=12}$	-0.706
$\gamma^{B}_{b=8}$	-0.138	$\gamma^{BY}_{b=8,y=1}$	-0.102	$\theta^{B}_{b=8}$	-0.102
$\gamma^{B}_{b=9}$	0.046	$\gamma^{BY}_{b=9,y=1}$	0.299	$\theta^{B}_{b=9}$	0.299
$\gamma^{B}_{b=10}$	-0.403	$\gamma^{BY}_{b=10,y=1}$	-0.044	$\theta^{B}_{b=10}$	-0.044
$\gamma^{B}_{b=11}$	-0.151	$\gamma^{BY}_{b=11,y=1}$	0.135	$\theta^{B}_{b=11}$	0.135
$\gamma^{B}_{b=12}$	-1.304	$\gamma^{BY}_{b=12,y=1}$	-0.411	$\theta^{B}_{b=12}$	-0.411

γ^{AB}_{ab} interaction terms (n=25):

mean = 0.000
max. = 0.009
min. = -0.009
std. dev.= 0.0036

CLL marginal effects model

ULL				CLL	
γ	5.363	$\gamma^{Y}_{y=1}$	-2.894	λ	2.468
$\gamma^{A}_{a=8}$	-0.144	$\gamma^{AY}_{a=8,y=1}$	0.018	$\lambda^{A}_{a=8}$	-0.126
$\gamma^{A}_{a=9}$	0.043	$\gamma^{AY}_{a=9,y=1}$	0.334	$\lambda^{A}_{a=9}$	0.377
$\gamma^{A}_{a=10}$	-0.411	$\gamma^{AY}_{a=10,y=1}$	0.095	$\lambda^{A}_{a=10}$	-0.316
$\gamma^{A}_{a=11}$	-0.150	$\gamma^{AY}_{a=11,y=1}$	0.115	$\lambda^{A}_{a=11}$	-0.034
$\gamma^{A}_{a=12}$	-1.293	$\gamma^{AY}_{a=12,y=1}$	-0.704	$\lambda^{A}_{a=12}$	-1.998
$\gamma^{B}_{b=8}$	-0.138	$\gamma^{BY}_{b=8,y=1}$	-0.102	$\lambda^{B}_{b=8}$	-0.240
$\gamma^{B}_{b=9}$	0.045	$\gamma^{BY}_{b=9,y=1}$	0.299	$\lambda^{B}_{b=9}$	0.343
$\gamma^{B}_{b=10}$	-0.403	$\gamma^{BY}_{b=10,y=1}$	-0.044	$\lambda^{B}_{b=10}$	-0.447
$\gamma^{B}_{b=11}$	-0.151	$\gamma^{BY}_{b=11,y=1}$	0.135	$\lambda^{B}_{b=11}$	-0.017
$\gamma^{B}_{b=12}$	-1.302	$\gamma^{BY}_{b=12,y=1}$	-0.411	$\lambda^{B}_{b=12}$	-1.713

REFERENCES

Agresti, A. 2002. *Categorical Data Analysis*. New York: Wiley-Interscience.

Besag, J. E. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society Series B* 36:192-236.

---. 1975. "Statistical Analysis of Non-lattice Data." *The Statistician*. 24: 179-195.

---. 1977. "Some Methods of Statistical Analysis for Spatial Data." *Bulletin of the International Statistical Association*. 47:77-92.

Bowles S, Choi JK, Hopfensitz A. 2003. "The co-evolution of individual behaviors and social institutions." Journal Of Theoretical Biology 223 (2): 135-147.

Burt, R. S. 1983. "Network Data From Archival Records." In Burt, R. S., and Minor, M. J. (Eds.), *Applied Network Analysis*, pp. 158-174. Beverly Hills: Sage.

---. 1984. "Network Items and the General Social Survey." *Social Networks* 6: 293-340.

Buve, A., M. Carael, R. J. Hayes, B. Auvert, B. Ferry, N. J. Robinson, S. Anagonou, L. Kanhonou, M. Laourou, S. Abega, E. Akam, L. Zekeng, J. Chege, M. Kahindo, N. Rutenberg, F. Kaona, R. Musonda, T. Sukwa, L. Morison, H. A. Weiss and M. Laga. 2001. "Multicentre Study on Factors Determining Differences in Rate of Spread of Hiv in Sub-Saharan Africa: Methods and Prevalence of HIV Infection." *Aids* 15 Suppl 4:S5-14.

Fienberg, S. E. 1977. *The Analysis of Cross-Classified Categorical Data*. Cambridge, Mass.: MIT Press.

Fischer, C. S. 1982. *To Dwell among Friends: Personal Networks in Town and City*. Chicago: University of Chicago Press.

Frank, O. 1988. "Random Sampling and Social Networks: A Survey of Various Approaches." *Mathematiques, Informatique, et Sciences Humaines* 26: 19-33.

Frank, O. and D. Strauss. 1986. "Markov Graphs." *Journal of the American Statistical Association* 81:832-42.

Geyer, C. J. and E. A. Thompson. 1992. "Constrained Monte Carlo Maximum Likelihood for Dependent Data." *Journal of the Royal Statistical Society Series B* 54:657-99.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Gintis, H. 200. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction*. Princeton: Princeton University Press.

Granovetter, M. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78:1360-80.

Holland, P. W. and S. Leinhardt. 1970. "A Method for Detecting Structure in Sociometric Data." *American Journal of Sociology* 72:492-513.

---. 1976. "Local Structure in Social Networks." *Sociological Methodology* 7:1-45.

---. 1981. "An Exponential Family of Probability Distributions for Directed Graphs (with discussion). *Journal of the American Statistical Association* 76: 33-65.

Ihaka, R. and R. Gentleman. 1996. "R: A Langauge for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics* 5:299-314.

Koehly, L.M., Peterson, S.K., Watts, B.G., Kempf, K.G., Vernon, S.W., and Gritz, E.R. 2003. "A Social Network Analysis of Communication about HNPCC Genetic Testing and Family Functioning." *Cancer, Epidemiology, Biomarkers, and Prevention, 12,* 304-313.

Laumann, E. O., J. H. Gagnon, S. Michaels, R. T. Michael and J. S. Coleman. 1989. "Monitoring the AIDS Epidemic in the United States: A Network Approach." *Science* 244:1186-9.

Liang, K.-Y. and S. L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73:13-22.

Macy M.W., Willer R. 2002. From factors to actors: Computational sociology and agent-based modeling. Annual Review of Sociology 28: 143-166.

Majeski S, Sylvan D. 1999. "How foreign policy recommendations are put together: A computational model with empirical applications." *International Interactions* 25 (4): 301-332.

Marsden, P. V. 1987. "Core Discussion Networks of Americans." *American Sociological Review* 52: 122-131.

---. 1988. "Homogeneity in Confiding Relations." *Social Networks* 10: 57-76.

Massey, D. S. 1990. "The Social and Economic Origins of Immigration." *Annals AAPSS* 510: 60-72.

Morris, M. 1991. "A Loglinear Modeling Framework for Selective Mixing." *Mathematical Biosciences* 107:349-77.

---. *In press*. *Network Epidemiology: A Handbook For Survey Design and Data Collection.* Oxford: Oxford University Press.

Morris, M. and L. Dean. 1994. "Effect of Sexual Behavior Change on Long-Term Human Immunodeficiency Virus Prevalence among Homosexual Men." *American Journal of Epidemiology* 140:217-32.

Padgett J.F., Ansell C.K. 1993. "Robust Action and the Rise of the Medici, 1400-1434 *American Journal of Sociology* 98 (6): 1259-1319.

Parker D.C., Manson S.M., Janssen M.A., Hoffmann M.J., Deadman P. Multi-agent systems for the simulation of land-use and land-cover change: A review. 2003. *Annals Of The Association Of American Geographers* 93 (2): 314-337.

Pattison, P. E., and G. L. Robins. 2002. Neighbourhood-based models for social networks. *Sociological Methodology* 32:301-337.

Pattison, P. and S. Wasserman. 1999. "Logit Models and Logistic Regressions for Social Networks: II. Multivariate Relations." *British Journal of Mathematical & Statistical Psychology* 52:169-93.

Resnick, M. D., P. S. Bearman, R. W. Blum, K. E. Bauman, K. M. Harris, J. Jones, J. Tabor, T. Beuhring, R. E. Sieving, M. Shew, M. Ireland, L. H. Bearinger and J. R. Udry. 1997. "Protecting Adolescents from Harm. Findings from the National Longitudinal Study on Adolescent Health." *Journal of the American Medical Association* 278:823-32.

Rindfuss, R., A. Jampaklay, B. Entwisle, Y. Sawangdee, K. Faust, P. Prasartkul *In press*. "The Collection and Analysis of Social Network Data in Nang Rong, Thailand." in M. Morris (ed.) *Network Epidemiology: A Handbook For Survey Design and Data Collection.* Oxford: Oxford University Press.

Robins, G., P. Elliott and P. Pattison. 2001a. "Network Models for Social Selection Processes." *Social Networks* 23: 1-30.

Robins, G., P. Pattison and P. Elliott. 2001b. "Network Models for Social Influence Processes." *Psychometrika* .

Robins, G., P. Pattison and S. Wasserman. 1999. "Logit Models and Logistic Regressions for Social Networks: III. Valued Relations." *Psychometrika* 64:371-94.

Strauss, D. and M. Ikeda. 1990. "Pseudolikelihood Estimation for Social Networks." *Journal of the American Statistical Society* 85:204-12.

Udry, J. R. and P. S. Bearman. 1998. "New Methods for New Research on Adolescent Sexual Behavior." In *New Perspectives on Adolescent Risk Behavior*, edited by R. Jessor. Cambridge ; New York: Cambridge University Press.

Watts, D.J. 2003. *Six Degrees: The Science of a Connected Age.* New York: W.W. Norton.

Wasserman, S. and P. Pattison. 1996. "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and P*." *Psychometrika* 60:401-25.