Inference on Edge Density in Undirected Binary Networks

Marcus B. Perry [1]
Department of Information Systems, Statistics, and Management Science
The University of Alabama
Tuscaloosa, AL 35487-0226 USA

Richard F. Deckro [2][3]
Department of Operational Sciences
Air Force Institute of Technology
Wright Patterson AFB OH 45433-7765 USA

Abstract

Undirected networks are used in a plethora of applications across many disciplines. For example, they are often used to model communication networks, financial networks, transportation networks, protein networks, social networks, and many more. This paper considers *binary* undirected networks, where the links (or edges) can only take on values of 0 or 1. It is assumed that the edge set of the network is not directly observable and must be estimated via an observable (but noisy) adjacency matrix. The proposed model assumes that changes in the edge set probabilities are due in large part to changes in a one or more exogenous nodal attribute variables. Using a log-likelihood ratio approach, we develop a hypothesis testing framework useful for detecting differences in the edge set probabilities given settings of the nodal attribute variables. Results of the hypothesis test can be used to draw inference on the unknown edge density of the network. We show that the proposed framework is equivalent to logistic regression with a categorical input variable modeled via dummy variables in the logit. Finally, application of the proposed hypothesis testing framework is demonstrated using both a simulated network and an open-source terrorist collaboration network.

Keywords: Hypothesis testing, log-likelihood ratio, binary networks, logistic regression.

[1] Corresponding author; Email: mperry@cba.ua.edu

[2] Voice: 937-255-6565 x4325; Email: richard.deckro@afit.edu

[3] The views expressed in this paper are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

1 Introduction

Undirected networks are used in a plethora of applications across many disciplines. For example, they are often used to model communication networks, financial networks, transportation networks, biological networks, social networks, and many more. This paper considers *binary* undirected networks where the links (or edges) can only take on values of 0 or 1, indicating that an edge does not exist or does exist between two nodes, respectively. One can imagine several types of networks that can be modeled this way, including (but not limited to) collaboration networks, acquaintance networks, dating networks, computer networks, and protein interaction networks.

To motivate the problem, consider an acquaintance network consisting of a study group of known male and female actors. Suppose we are interested in determining the effect of gender on the acquaintance pattern of this group. Let us denote the number of male-to-male, female-to-female, and male-to-female acquaintance relationships by N_m, N_f, and N_{mf}, respectively. Suppose that actors in the study group only acquaint themselves with the opposite gender, then we would expect $N_{mf} \geq 0$ and $N_m = N_f = 0$. On the other hand, if actors in the study group only acquaint themselves with individuals of the same gender, we might expect $N_m \neq N_f > 0$ and $N_{mf} = 0$. The methodology developed in this paper will permit the network analyst to draw inference on the acquaintance pattern of the study group. If this study group is a sample from a larger population, then the methodology proposed in this paper permits the analyst to draw statistical inference on the general study group population.

Much of the literature on *statistical models* for networks is rooted in the social network analysis literature. In particular, statistical models for social network analysis have been proposed by several authors; the most sophisticated of these currently being the exponential random graph models (ERGM) (e.g., see Frank & Strauss, [1]; Wasserman & Pattison, [11]; Pattison & Wasserman, [4]; Robins et. al., [6]). ERGMs are commonly referred to as the class of p^* models in the social network literature. In general, these models consider dyadic dependencies, and thus, permit the construction of fairly realistic models of complex social systems. However, the use of these models in practice has been quite limited due to their intractability with regards to parameter estimation. In particular, to obtain parameter estimates with known statistical properties, network analysts must rely on complicated Markov Chain Monte Carlo (MCMC) algorithms (e.g., see Snijders, [9]). Handcock [2] points out that the application of these models in their traditional specification to observed networks can often lead to model degeneracy problems and instability of the MCMC algorithm, and hence, a poor fit to empirical data (i.e., the algorithm will not converge). More recently, however, Snijders *et. al.* [10] developed new specifications for the ERGMs that involve more complicated network statistics, and these authors conclude that these new specifications "push back" some of the degeneracy problems of commonly promoted models in the literature. Unfortunately, these new network statistics are of a very high order, and thus, interpretation is not as straightforward as the traditional specifications (Traditional specifications for undirected networks often involved 2-star, 3-star, and triangle configurations). Additionally, Smyth [8] points out that scalability is another major concern of the ERGMs. In particular, parameter estimation using MCMC methods is intractable for large networks due to the inherent computational issues.

It should be noted that the hypothesis testing framework proposed in this paper is a special case of the ERGMs mentioned above under edge independence and vertex block homogeneity assumptions. Further, we show in a later section that, under certain conditions, the proposed method is equivalent to performing a logistic regression. One major advantage of the proposed approach relative to existing statistical models in the literature is that it does not rely on complicated computational algorithms as a means to estimate unknown model parameters[4]. This should provide a more user friendly tool for the network analyst, while providing useful information and avoiding problems such as interpretation, model degeneracy, scalability, and instability of computational algorithms.

The remainder of this manuscript is organized as follows. In Sec. 2 we introduce the proposed model and derive the test statistic. In Sec. 3 the proposed methodology is applied to a real data network of collaboration ties amongst a group of known Al Qaeda terrorists. An alternative parameterization of the proposed method using logistic regression is discussed in Sec 4, where both methods are applied to a simulated network to illustrate the practical differences. Finally, Sec. 5 closes with a summary and discussion.

[4] However, perhaps at the expense of model adequacy depending on the *type* and *magnitude* of correlation that may exist between the edges.

2 Model Definition

Under the proposed framework, network edge density is studied via an observed adjacency matrix, which is assumed to be subject to sampling variability (i.e., the true edges are not directly observable). The methodology involves partitioning network actors into k mutually exclusive and collectively exhaustive subsets, based upon available exogenous nodal attribute information[5]. Edge density within and across the subsets is then measured, and the observed differences between these measurements are evaluated statistically using the log-likelihood ratio. The proposed model can be parameterized as follows.

Consider the graph $G = \{V, E\}$, where V and E denote the vertex set and edge-set, respectively. We assume that V is known; however, E is not directly observable and needs to be estimated from an observed adjacency matrix A. The test procedure involves partitioning V into $m > 1$ mutually exclusive and collectively exhaustive subsets. We assume that edge probabilities between vertices contained in the same subset V_h $(h = 1, 2, ..., m)$ are all equal, and thus can be denoted by a single parameter, p_h. In addition, edge probabilities between vertices contained in V_i and those contained in V_j, for a given $i < j = 2, 3, ..., m$, are all equal and thus denoted by p_{ij}. Thus, for a partition of V resulting in m mutually exclusive and collectively exhaustive subsets, where each subset contains at least two vertices, there are a total of $g = \frac{1}{2}m(m + 1)$ parameters.

In general, we can consider $r > 1$ partitions, where the i^{th} partition $(i = 1, 2, ..., r)$ has $k_i > 1$ levels. At this point it is convenient to define the notion of a *combined* partition. Suppose that $r = 2$ and $k_1 = k_2 = 2$, so that we have two partitions each at two levels. For the first partition, let us arbitrarily assign those vertices in V having attribute x a value of "1" and "0" otherwise. Similarly, for the second partition, let us assign those vertices in V having attribute z a value of "1" and "0" otherwise. Then the *combined* partition would divide V into $k_1 k_2 = 2^2 = 4$ mutually exclusive and collectively exhaustive subsets[6]. When the condition $k_i = k$ for all i holds, the combined partition will produce k^r mutually exclusive subsets, resulting in a model with $g = \frac{1}{2}k^r(k^r + 1)$ parameters[7].

[5] Clearly, a *fundamental* assumption of the proposed method is that actor attribute information exists and is available for each actor in the network. Later we discuss an alternative when this information is missing or does not exist.

[6] These would include vertices: 1) having both attributes x and z 2) having only attribute x 3) having only attribute z 4) having neither attribute x or z.

[7] i.e., the p_h's $(h = 1, 2, ..., k^r)$ and p_{ij}'s $(i < j = 2, 3, ..., k^r)$

More generally, if we have r partitions and the i^{th} partition has k_i levels, then we can define the total number of mutually exclusive and collectively exhaustive subsets of V by

$$w = \prod_{i=1}^{r} k_i \tag{1}$$

and thus the total number of model parameters is given more generally as $g = \frac{1}{2}w(w+1)$.

It is convenient to note here that the analyst may not be interested in comparing differences between all g parameters; rather, interest might lie in comparing differences between a subset of these parameters (depending on the objectives of the analysis). Although not to be overlooked, we will refrain from any further discussion on this now and return to it in a later section.

It should be noted that the number of partitions r is bounded for any given n and k_i $(i = 1, 2, ..., r)$, where n denotes the cardinality of V. In general the following condition must hold

$$w(w+1) \leq n(n-1) \tag{2}$$

so that if $k_i = k$ $(i = 1, 2, ..., r)$ we obtain

$$r \leq \lfloor \frac{\ln[n(n-1)]}{3\ln(k)} \tag{3}$$

where "$\lfloor$" denotes the floor function.

Under the above described model, our interest lies in estimating the p_h's and the p_{ij}'s and subsequently drawing inference on their true values via a formal hypothesis testing framework. It should be noted that, in general, the goal of any hypothesis test is to measure the plausibility of H_0 (null hypothesis) relative to H_1 (alternative hypothesis). The model specified under H_0 postulates a model with *fewer* parameters than that specified by H_1. Therefore, rejecting H_0 might suggest the "better" model is that specified under H_1. On the other hand, if H_0 cannot be rejected, this would suggest that the model specified under H_1 is no "better" than that specified under H_0.

In general, several forms for H_0 and H_1 could be specified; however, in this paper interest lies in testing the hypotheses: $H_0 : p_h = p_0 \cap p_{ij} = p_0$ (for all $h = 1, 2, ..., m$ and $i < j = 2, 3, .., m$) versus $H_1 : p_h \neq p_0 \cup p_{ij} \neq p_0$ (for at least one $h = 1, 2, ..., m$ or $i < j = 2, 3, .., m$). Failing to reject H_0 in this case would suggest that all network edge probabilities are equal and thus can be parameterized by a single parameter p_0. However, if H_0 is rejected, this would suggest that a more plausible model can be obtained by a partitioning of the vertex set into m distinct subsets.

2.1 Edge Set: Probability Mass Function

Let y_h denote the number of observed edges between vertices contained in V_h, and y_{ij} denote the number of observed edges between vertices contained in V_i and those contained in V_j, for a given $i \neq j$. Suppose that the observations are conditionally independent given the $m-$level partition $z(m)$. Under this assumption, if the vertex set V is partitioned into m mutually exclusive and collectively exhaustive subsets, then the probability mass function for $\mathbf{y}$ is given by

$$f(\mathbf{y}|\mathbf{p}, z(m)) = \prod_{h=1}^{m} \left(\begin{array}{c} N_h \\ y_h \end{array} \right) p_h^{y_h}(1-p_h)^{N_h - y_h} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} \left(\begin{array}{c} N_{ij} \\ y_{ij} \end{array} \right) p_{ij}^{y_{ij}}(1-p_{ij})^{N_{ij}-y_{ij}} \tag{4}$$

for $0 \leq p_h, p_{ij} \leq 1$, where $y_h \in [0, 1, \ldots, N_h]$, $N_h = \frac{1}{2}n_h(n_h - 1)$ and n_h is the total number of vertices contained in subset V_h. Also, $y_{ij} \in [0, 1, \ldots, N_{ij}]$ $(i < j = 2, 3, ..., m)$, where N_{ij} denotes the total number of possible edges between vertices contained in V_i and those contained in V_j. Also, it should be noted that

$$y = \sum_{h=1}^{m} y_h + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} y_{ij} \tag{5}$$

and

$$N = \sum_{h=1}^{m} N_h + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} N_{ij}, \tag{6}$$

where y denotes the total number of observed edges in V and N denotes the total number of possible edges in V.

In general, eq.(4) is defined for all $m \in [1, 2, ...n]$, where n is the cardinality of V. However, when $m = 1$, all vertices are contained in the same set, which is just V. Further, when $m = n$, each vertex defines its own subset, or $V \equiv V_1 \cup V_2 \cup \cdots \cup V_n$. For these two special cases, eq.(4) can be written explicitly as

$$f(\mathbf{y}|p, z(1)) = \left(\begin{array}{c} N \\ y \end{array} \right) p^y (1-p)^{N-y} \tag{7}$$

and

$$f(\mathbf{y}|\mathbf{p}, z(n)) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \left(\begin{array}{c} N_{i,j} \\ y_{i,j} \end{array} \right) p_{i,j}^{y_{i,j}} (1-p_{i,j})^{N_{i,j}-y_{i,j}}, \tag{8}$$

respectively. Notice that in eq.(7), there is only a single parameter to estimate since all vertices are contained in the same set. However, in eq.(8), there are $\frac{1}{2}n(n-1)$ parameters to estimate since each vertex defines a unique subset of V. To ensure that at least two vertices are assigned to each level of the partition $z(m)$, a necessary condition for m is

$$2 \leq m \leq \lfloor \frac{n}{2} \tag{9}$$

where $\lfloor$ denotes the floor function. Also, the number of partitions of V into q levels resulting in all q levels having at least two vertices is given by

$$d = \begin{pmatrix} \lfloor \frac{n}{2} \\ q \end{pmatrix} \tag{10}$$

for $q = 2, 3, \ldots, \lceil \frac{n}{2}$, where $\lceil$ denotes the ceiling function. For example, consider a network of $n = 20$ vertices. Suppose that you partition the $n = 20$ vertices into $m = 3$ mutually exclusive and collectively exhaustive subsets. Then, there are $d = 120$ unique ways to do this so that at least 2 vertices are contained in each level.

2.2 Hypothesis Test: Derivation of the Test Statistic

The likelihood function is proportional to eq. (4) and is given by

$$L(\mathbf{p}|\mathbf{y}, z(m)) = \prod_{h=1}^{m} p_h^{y_h}(1 - p_h)^{N_h - y_h} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} p_{ij}^{y_{ij}}(1 - p_{ij})^{N_{ij} - y_{ij}} \tag{11}$$

and taking the natural log of this function then produces the log-likelihood function, or

$$\begin{aligned} \ell(\mathbf{p}|\mathbf{y}, z(m)) = & \sum_{h=1}^{m} [y_h \ln(p_h) + (N_h - y_h) \ln(1 - p_h)] + \\ & + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} [y_{ij} \ln(p_{ij}) + (N_{ij} - y_{ij}) \ln(1 - p_{ij})] \end{aligned} \tag{12}$$

It is easily shown that the values of p_h and p_{ij} $(h = 1, \ldots, m; i < j = 2, \ldots, m)$ that maximize the log-likelihood function in eq.(12) are given by $\hat{p}_h = \frac{y_h}{N_h}$ and $\hat{p}_{ij} = \frac{y_{ij}}{N_{ij}}$. Since these are maximum likelihood estimates, we can exploit their asymptotic properties and derive approximate $100(1 - \gamma)\%$ confidence bounds on each p_h and p_{ij}. This will allow network analysts to gain some insight into the quality of these point estimates.

It is well known that a maximum likelihood estimator of θ, say $\hat{\theta}$, is asymptotically distributed as approximately multivariate normal with $\mathrm{E}[\hat{\theta}] = \theta$ and $\mathrm{Var}(\hat{\theta}) = [\mathbf{I}(\theta)]^{-1}$, where $\mathbf{I}(\theta)$ denotes the Fisher information matrix with element $I_{p,q}$ given by (see Rice, [5])

$$I_{p,q} = \mathrm{E}[\frac{d}{d\theta_p}\ell(\mathbf{x}|\theta) \cdot \frac{d}{d\theta_q}\ell(\mathbf{x}|\theta)]. \tag{13}$$

In eq. (13), $\ell(\mathbf{x}|\theta)$ denotes the log-likelihood function and the expectation is taken over the random variables $\mathbf{x}$. For any unbiased estimator, the diagonal elements of the inverse of $\mathbf{I}(\theta)$ are then the Cramer-Rao lower bounds on the variances of the parameter estimates. Thus, since maximum likelihood estimators are asymptotically unbiased, $\mathbf{I}(\theta)^{-1}$ can be viewed as the asymptotic variance-covariance matrix of the estimator $\hat{\theta}$.

Suppose $\hat{\mathbf{p}} = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m, \hat{p}_{1,2} \ldots, \hat{p}_{1,m}, \hat{p}_{2,m}, \ldots, \hat{p}_{m-1,m}]'$ denotes the $\left(\frac{m(m+1)}{2}\right) \times 1$ dimensional parameter vector estimate. Under the assumed model in eq. (4), the variance-covariance matrix of $\hat{\mathbf{p}}$ is diagonal, suggesting the covariances between the parameter estimates are zero for any given partition $z(m)$ of V. It follows that the asymptotic variances of the parameter estimates are given explicitly as $\mathrm{Var}(\hat{p}_h) = \frac{p_h(1-p_h)}{N_h}$ $(h = 1, 2, ..., m)$ and $\mathrm{Var}(p_{ij}) = \frac{p_{ij}(1-p_{ij})}{N_{ij}}$ $(i < j = 2, 3, ..., m)$. Since $\hat{\mathbf{p}}$ is a maximum likelihood estimator, the asymptotic distribution of $\hat{\mathbf{p}}$ is approximately multivariate normal with mean vector $\mathbf{p}$ and variance-covariance matrix $\mathrm{Var}(\hat{\mathbf{p}})$. Therefore, a large sample $100(1-\gamma)\%$ confidence interval on p_h is given by $\hat{p}_h \pm z_{\gamma/2}\sigma_{\hat{p}_h}$, where $z_{\gamma/2}$ denotes the upper $(\gamma/2)^{th}$ quantile of the standard normal distribution and $\sigma_{\hat{p}_h} = \sqrt{Var(\hat{p}_h)}$. Large sample confidence intervals on the p_{ij}'s are computed similarly.

The likelihood function given in eq. (11) can be used to derive a test for equality of the parameters. This is more formally stated in terms of the null hypothesis (H_0) and the alternative hypothesis (H_1) given as

$$H_0: \quad p_h = p_0 \cap p_{ij} = p_0 \quad \text{(for all } h = 1, ..., m \text{ and } i < j = 2, 3, ..., m) \tag{14}$$

and

$$H_1 : p_h \neq p_0 \cup p_{i,j} \neq p_0 \quad \text{(for at least one } h = 1, ..., m \text{ or } i < j = 2, .., m) \tag{15}$$

for some partition $z(m)$ $(m > 1)$. Conclusions from such a test can be used to mitigate some of the variability in the sampled edge-set of the network, thus, providing insight into its true underlying structure.

The likelihood function under H_1 is given by eq. (11), while the likelihood function under H_0 can be written as

$$L_0(y|p_0) = L_0(y|p_0) = p_0^y(1-p_0)^{N-y} \tag{16}$$

where $0 \leq p_0 \leq 1$ might be some hypothesized value of interest and is used to denote the probability that an edge exists between any two vertices in the complete vertex set V. The variables N and y were defined previously. Clearly, eq. (16) suggests that the edge probabilities between vertices in V are best explained by a single parameter, p_0. Notice if $p_0 = 0.50$, then H_0 postulates that the observed variability in the sampled edge-set is strictly due to random chance. Often in practice the parameter p_0 is unknown and needs to be estimated. In these cases, the test is conducted by computing the difference between the log-likelihood function maximized under H_0 and the log-likelihood function maximized under H_1, or $\Delta = \ell_0^* - \ell_1^*$, where

$$\ell_0^* = y\ln(\frac{y}{N-y}) + N\ln(\frac{N}{N-y}) \tag{17}$$

and

$$\ell_1^* = \sum_{h=1}^{m}\left(y_h\ln(\frac{y_h}{N_h-y_h}) + N_h\ln(\frac{N_h}{N_h-y_h})\right) + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\left(y_{ij}\ln(\frac{y_{ij}}{N_{ij}-y_{ij}}) + N_{ij}\ln(\frac{N_{ij}}{N_{ij}-y_{ij}})\right) \tag{18}$$

Under the null hypothesis H_0, the test statistic

$$\hat{r} = -2\Delta \tag{19}$$

is then approximately distributed as χ_D^2, where D denotes the difference in dimensionality of the parameter spaces under H_1 and H_0. Assuming all levels of the partition contain at least two vertices, then $D = \frac{1}{2}m(m+1) - 1$. For example, when $m = 2$, and if all levels of the partition contain at least two vertices, then $D = 3 - 1 = 2$. If $\hat{r} > \chi_{\alpha,D}^2$, then H_0 is rejected at the $1-\alpha$ significance level.

If the null hypothesis above is rejected, then the analyst should set out to perform multiple comparisons of the $\hat{p}$'s. In general, if there are m levels of the vertex set partition, then we will have as the total number of possible comparisons

$$\nu = \left(\begin{array}{c} \frac{1}{2}m(m+1) \\ 2 \end{array}\right) \tag{20}$$

In general, for any given p_h and $p_{h'}$ $(h \neq h')$, we can compute the test statistic $z_0 = \frac{\hat{p}_h - \hat{p}_{h'}}{se(\hat{p}_h - \hat{p}_{h'})}$ for all $h \in [1, 2, ..., \nu]$, where

$$se(\hat{p}_h - \hat{p}_{h'}) = \sqrt{\frac{p_h(1-p_h)}{N_h} + \frac{p_{h'}(1-p_{h'})}{N_{h'}}} \tag{21}$$

A major concern when performing multiple comparisons is that the false-alarm rate (or type I error) of the test will increase with the number of comparisons. Suppose that $\nu' \leq \nu$ comparisons are performed, then p_h and $p_{h'}$ are deemed significantly different if $|z_0| > z_{\alpha_0}$ where z_{α_0} is the upper α_0^{th} percentile point of the standard normal distribution and α_0 is determined by

$$\alpha_0 = 1 - \exp\left\{\frac{\ln(1-\alpha)}{\nu'}\right\} \tag{22}$$

where α is an acceptable upper bound on the "experiment-wide" false-alarm rate[8]. This will ensure that the type I error for the test is *at most* α.

As mentioned earlier, the analyst may not be interested in comparing differences between all g model parameters. To address this, let us consider the following. Suppose that $m = 2$ so that we have the parameters p_1, p_2, and p_{12}. Suppose further that the analyst is only interested in testing the hypotheses H_0: $p_1 = p_2$ versus H_1: $p_1 \neq p_2$. For the $m = 2$ case, the probability mass function is given by

$$f(\mathbf{y}|\mathbf{p}) = \binom{N_1}{y_1} p_1^{y_1}(1-p_1)^{N_1-y_1} \binom{N_2}{y_2} p_2^{y_2}(1-p_2)^{N_2-y_2} \binom{N_{12}}{y_{12}} p_{12}^{y_{12}}(1-p_{12})^{N_{12}-y_{12}} \tag{23}$$

and to find the joint distribution of y_1 and y_2, we can sum eq.(23) over all possible values of y_{12} to obtain

$$f(\mathbf{y}|\mathbf{p}) = \binom{N_1}{y_1} p_1^{y_1}(1-p_1)^{N_1-y_1} \binom{N_2}{y_2} p_2^{y_2}(1-p_2)^{N_2-y_2} \tag{24}$$

suggesting that one needs only to drop the terms associated with those parameters not considered for analysis from the log-likelihood function in eq.(18). Thus, we would drop the terms associated with p_{12} to obtain

$$\ell_1^* = y_1 \ln(\frac{\hat{p}_1}{1-\hat{p}_1}) + N_1 \ln(1-\hat{p}_1) + y_2 \ln(\frac{\hat{p}_2}{1-\hat{p}_2}) + N_2 \ln(1-\hat{p}_2) \tag{25}$$

and then conduct the hypothesis test in the same manner as described above[9].

[8] The Bonferroni inequality suggests that the "experiment-wide" error rate is less than α when the comparisons are not mutually independent. As a result, one can view α as an acceptable *upper bound* on the type I error probability for any ν' differences compared.

[9] Note that the degrees of freedom for the test will change due to a change in the dimensionality of the parameter space specified under H_1.

3 Application of Proposed Hypothesis Test

In this section we demonstrate how the proposed hypothesis test is applied to a real network. Specifically, we consider the open-source Al Qaeda data compiled by Sageman [7]. The data consists of several possible network contexts, including acquaintanceship, friendship, teacher-student, family, collaboration, and so on. Additionally, the data compiled by Sageman [7] contains a variety of information on each actor (e.g., age, marital status, education type, etc.). In this paper, we consider the collaboration network consisting of the first 50 actors (i.e., known terrorists) listed in the data set. Then, in the observed sample, a link present between any two actors implies that these two actors were observed to collaborate on at least one terrorist-related activity (e.g., Sept. 11, Embassy 98, etc.). The network is shown in Figure 1.

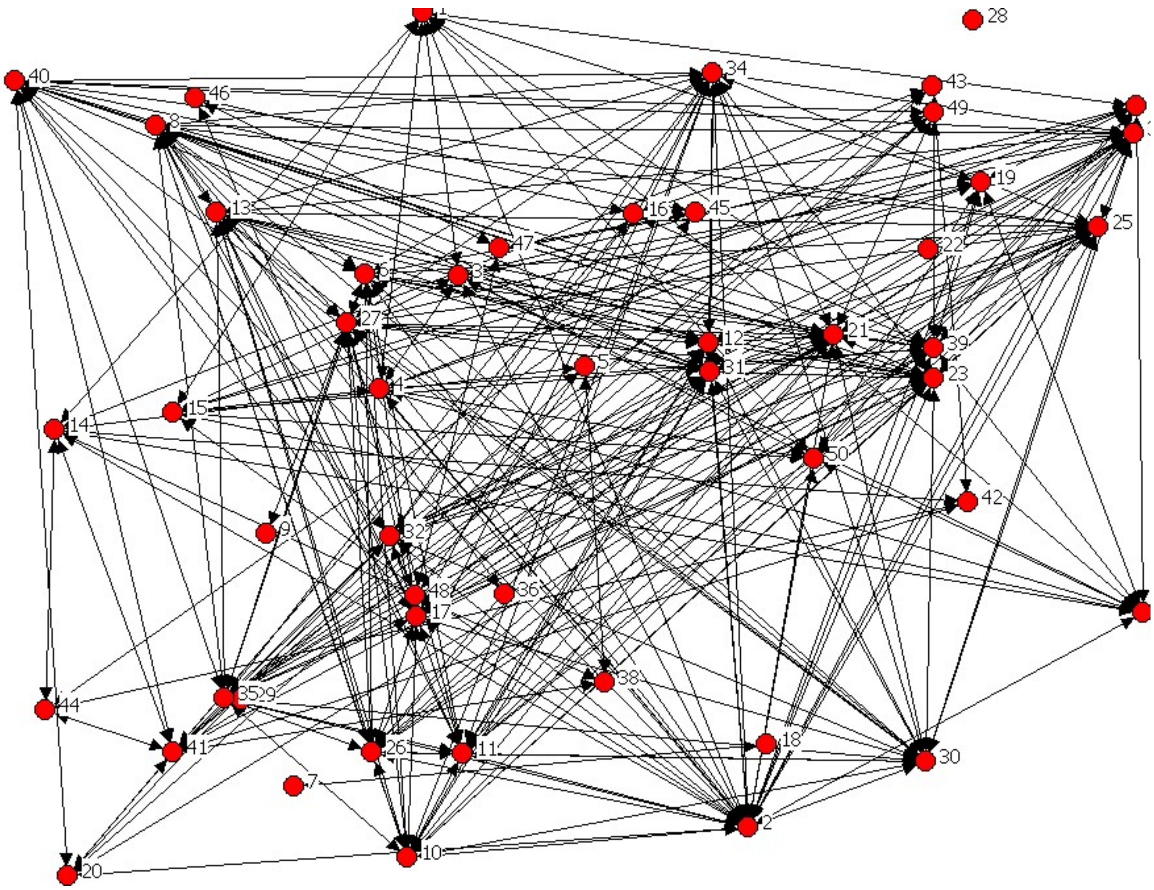


Figure 1: Open source Al-Qaeda collaboration network.

To begin, we first need to identify those nodal attributes that are most relevant to the study objectives. In doing so, it is important to keep in mind that there is a constraint on the number of factors that can be studied. Suppose that all g parameters are of interest to the analyst. Then with $n = 50$, and if we assume that each attribute has only two levels (e.g., black/white, on/off, etc.), then according to the expression in (3), the total number of nodal attributes that can be studied is $r = 3$. On the other hand, suppose we are only interested in comparing differences amongst the p_h's and have no interest in drawing inference on the p_{ij}'s. Then, assuming each attribute considered has only 2 levels, we can study up to $r \leq \lfloor \frac{\ln[0.5n(n-1)]}{\ln(2)} = 10$ attributes.

Suppose we are interested in knowing whether or not the probability of collaboration between any two actors is heavily influenced by their education level. We can determine the marginal effect of 'education level' on the probability of collaboration by partitioning the vertex set so that those actors NOT possessing a college degree are assigned to group 1 and those possessing a college degree are assigned to group 2. The results of the test are shown in Table 1 and indicate 'education level' yields a highly significant model. Notice that those actors who are educated tend to collaborate more frequently with each other. Similarly, those actors who are not educated also tend to collaborate more frequently with each other. It is interesting to note that $\hat{p}_{12} = 0.03$, which suggests that educated and uneducated actors rarely collaborate with each other on terrorist activities. Perhaps this could suggest the existence of multiple independent operating cells amongst the study group; with one having more sophistication (e.g., with respect to planning and execution of the activities) than the others due to the level of education of its group members.

Suppose that we choose two more attributes: 'criminal background' and 'age joined Jihad', as well as their combined partition. For the attribute 'criminal background', actors who do NOT possess a criminal background are assigned to group 1 and those possessing a criminal background assigned to group 2. Similarly, for the attribute 'age joined Jihad', actors who joined the Jihad before age 25 are assigned to group 1 and those who joined the Jihad at age 25 or beyond are assigned to group 2. Notice that the combined partition has $2^2 = 4$ unique subsets and thus the number of parameters required for this partition is $g = 10$. The results of the tests are also given in Table 1, and suggest that all partitions considered yield significant models. We should note that since 'education level', 'criminal background', and 'age joined Jihad' are all 2-level attributes, we can compare their $\hat{r}$ statistics directly. This would suggest that the attribute 'education level' does relatively best amongst the 2-level factors in explaining the change in collaboration probabilities between actors. Unfortunately, we cannot compare directly the $\hat{r}$ values from a 2-level partition and a $m > 2$ level partition. To permit correct comparison, we can use the difference in deviance explained below.

Table 1: Results of hypothesis tests.

Attribute	$\hat{r}$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_{12}$	$\hat{p}_{13}$	$\hat{p}_{14}$	$\hat{p}_{23}$	$\hat{p}_{24}$	$\hat{p}_{34}$
Ed Level	123.5	0.21	0.29	0.03	-	-	-	-	-	-	-
Age Joined Jihad (AJJ)	83.7	0.19	0.31	0.09	-	-	-	-	-	-	-
Crim Back (CB)	48.0	0.13	0.35	0.19	-	-	-	-	-	-	-
(CB) × (AJJ)	131.9	0.25	0.16	0.13	0.46	0.13	0.09	0.08	0.15	0.07	0.26

In general, for any given partition, the deviance is defined as $\lambda = -2[\ell(\mathbf{p}|\mathbf{y}, z(m))]$, where $\ell(\mathbf{p}|\mathbf{y}, z(m))$ was defined earlier and denotes the log-likelihood function given $\mathbf{y}$ and $z(m)$. The deviance can be viewed as a measure of lack-of-fit between the postulated model and observed data. Suppose that z_0 denotes a model with g_0 parameters and z_1 denotes a model with $g_1 > g_0$ parameters. That is, it is assumed that z_0 excludes any effects that are hypothesized to be null, while z_1 includes these effects. Thus, z_0 denotes a model specified under H_0 and z_1 denotes a model specified under H_1. The difference in deviance between these two models is then given by

$$\delta = \lambda_0 - \lambda_1 \tag{26}$$

where λ_0 denotes the deviance statistic computed under model z_0 and λ_1 denotes the deviance statistic computed under model z_1. It is easily shown that δ is a log-likelihood ratio statistic, and it is well known that under H_0 eq.(26) is asymptotically χ^2_D, with $D = g_1 - g_0$. Therefore, one should reject H_0 and conclude in favor of H_1 if $\delta > \chi^2_{\alpha,D}$, suggesting that the more plausible model is z_1.

In applying this concept to the problem in hand, we need to compute the difference in deviance between the model produced by the attribute 'education level' and the model produced by the *combined* partition since these seem to be the two best models. Doing so produces $\delta = \lambda_0 - \lambda_1 = 8.5162$, and since $g_1 - g_0 = 7$, we compare δ to the χ^2_7 distribution. At the 95% significance level, the appropriate critical value is then $\chi^2_{0.95,7} = 14.0671$. Therefore, since $\delta = 8.5162 < \chi^2_{0.95,7} = 14.0671$, we can conclude that the combined partition is no more plausible than the model based upon 'education level'. Thus to maintain model parsimony, we should choose the $g = 3$ parameter model since it has fewer parameters.

At this point in the analysis we need to perform multiple comparisons. Since the $g = 3$ parameter model was chosen as the final model, we are interested in comparing differences $(\hat{p}_1 - \hat{p}_2)$, $(\hat{p}_1 - \hat{p}_{12})$, and $(\hat{p}_2 - \hat{p}_{12})$. Since there are three tests to be performed, to maintain an upper bound of 0.05 on the "experiment-wide" false-alarm rate, the appropriate critical value for the test is $z_{0.0170} = 2.1201$. Table 2 shows the results of the multiple comparison and suggests that p_1 and p_2 are both significantly different from p_{12}; however, there is no evidence that p_1 is different than p_2.

It is convenient to note that the proposed method lends to simple analysis via odds ratios. For example, using the results in Table 1, the estimated ratio of the odds of observing a collaboration tie between two actors who are both educated, to the odds of observing a collaboration tie between two actors, where only one of the actors is educated, is

$$\widehat{OR} = \frac{0.2658}{0.0309} = 8.6019 \tag{27}$$

suggesting that a collaboration tie is almost 9 times more likely to occur when both the actors are educated, relative to the case where only one actor is educated.

Table 2: Results of the multiple comparisons analysis.

Comparison	$\lvert z_0 \rvert$	P
$\hat{p}_1 - \hat{p}_2$	0.9215	0.1784
$\hat{p}_1 - \hat{p}_{12}$	10.706	0.0000
$\hat{p}_2 - \hat{p}_{12}$	3.0118	0.0013

In this section we demonstrated application of the proposed hypothesis testing framework on a real-world network. In the next section we show that the proposed methodology is a special case of logistic regression when the predictor variable is categorical and dummy (or indicator) variables are used to model the effects of this variable in the logit. We then conduct another analysis on a simulated network to illustrate any practical differences between the two approaches. Finally, we close with a discussion and summary.

4 Alternative Parameterization using Logistic Regression

In this section, we show that the proposed framework is equivalent to logistic regression when the predictor variable under study is categorical and indicator variables are used in the logit to model its effect. If exogenous information is available on each dyad (i, j) for $i < j = 2, 3, ..., m$, then application of logistic regression is straightforward. That is, each dyad would be an observation with model covariates (or factors) determined from available exogenous information at the *dyad level.* However, for the problem in hand, exogenous information is assumed to be available at the *node level*, and so for logistic regression to be applied appropriately, we need to incorporate dummy variables into the logit. To illustrate, consider a two-level nodal attribute denoted by z. At the dyad level, each (i, j) can either (1) both have attribute z (2) either i or j have attribute z, but not both (3) neither have attribute z. In general, in any empirical model building exercise, if we have a single factor with v levels, then $v - 1$ dummy variables are required to model the factor effect. In our illustration, $v = 3$, and therefore we require $v - 1 = 2$ dummy variables in the logit, or

$$g(x_1, x_2) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{28}$$

where the x_i's are coded according to Table 3.

Table 3: Coding scheme for dummy variables.

	x_1	x_2
Both i and j have attribute z	0	0
Either i or j has attribute z but not both	1	0
Both i and j do not have attribute z	0	1

To see that the proposed method is equivalent to logistic regression in this case, note that the log-likelihood specified under H_0: $\beta_1 = \beta_2 = 0$ can be written as

$$\ell_0 = N(\beta_0 \hat{p}_0 - \ln[1 + \exp\{\beta_0\}]) \tag{29}$$

and taking the derivative of eq.(29), setting to zero, and solving for β_0 we obtain

$$\hat{\beta}_0 = \ln\left(\frac{y}{N-y}\right) \tag{30}$$

so that

$$\ell_0^* = y \ln\left(\frac{y}{N-y}\right) + N \ln\left(\frac{N}{N-y}\right) \tag{31}$$

which is exactly that given in eq.(17). Note also that the log-likelihood function specified under H_1: $\beta_j \neq 0$ (for at least 1 $j = 1, 2$) can be written as

$$\ell_1 = \beta_0 y_1 - N_1 \ln(1 + e^{\beta_0}) + \gamma_2 y_2 - N_2 \ln(1 + e^{\gamma_2}) + \gamma_1 y_{12} - N_{12} \ln(1 + e^{\gamma_1}) \tag{32}$$

where $\gamma_1 = (\beta_0 + \beta_1)$ and $\gamma_2 = (\beta_0 + \beta_2)$. Taking partial derivatives of ℓ_1 with respect to β_0, γ_1 and γ_2, setting equal to zero, and solving for these parameters yields $\hat{\beta}_0 = \ln\left(\frac{y_1}{N_1 - y_1}\right)$, $\hat{\gamma}_1 = \ln\left(\frac{y_{12}}{N_{12} - y_{12}}\right)$, and $\hat{\gamma}_2 = \ln\left(\frac{y_2}{N_2 - y_2}\right)$, producing

$$\ell_1^* = \sum_{h=1}^{2} \left[y_h \ln\left(\frac{y_h}{N_h - y_h}\right) + N_h \ln\left(\frac{N_h}{N_h - y_h}\right) \right] + y_{12} \ln\left(\frac{y_{12}}{N_{12} - y_{12}}\right) + N_{12} \ln\left(\frac{N_{12}}{N_{12} - y_{12}}\right) \tag{33}$$

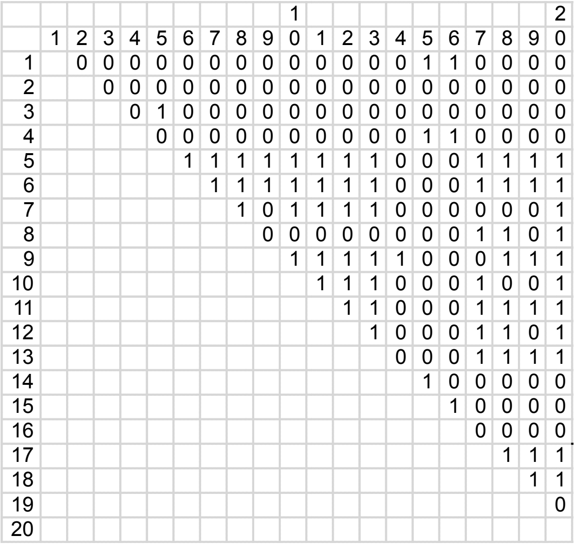
which is exactly equivalent to that given in eq.(18) for $m = 2$. It is straightforward then to show this equivalence for the more general case where z has more than two levels.

4.1 Example using Simulated Data

In this subsection, we use both parameterizations (i.e., proposed method and logistic regression) to analyze a simulated realization of a sampled network. Suppose we have a single nodal attribute with two-levels and that the network consists of $n = 20$ vertices. Suppose further the vertex set is partitioned such that $V = V_1 \cup V_2 = \{1, 2, 3, 4, 14, 15, 16\} \cup \{5, 6, 7, 8, 9, 10, 11, 12, 13, 17, 18, 19, 20\}$. The simulated network was generated by independently assigning a link between vertices contained in V_1 with probability $p_1 = 0.25$. Similarly, links between vertices contained in V_2 are independently assigned with probability $p_2 = 0.75$. Finally, links between those vertices contained in V_1 and those contained in V_2 are independently assigned with probability $p_{12} = 0.05$. The simulated adjacency matrix is shown in Figure 2.

Applying the proposed hypothesis testing framework to the simulated adjacency matrix shown in Figure 2 yields a test statistic value of $\hat{r} = 130.413$. Since H_0 postulates a $g_0 = 1$ parameter model and H_1 postulates a $g_1 = 3$ parameter model, $D = g_1 - g_0 = 2$, and the appropriate critical value at the 95% significance level is $\chi^2_{0.95,2} = 5.9915$. Thus, H_0: $p_1 = p_2 = p_{12} = p_0$ is rejected and we conclude in favor of H_1.

At this point we need to perform multiple comparisons. Since there are $g = 3$ parameters, we will be making $\nu = 3$ comparisons; namely $(\hat{p}_1 - \hat{p}_2)$, $(\hat{p}_1 - \hat{p}_{12})$, and $(\hat{p}_2 - \hat{p}_{12})$. The critical value for each



										1										2
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0
1		0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
2			0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3				0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4					0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
5						1	1	1	1	1	1	1	1	0	0	0	1	1	1	1
6							1	1	1	1	1	1	1	0	0	0	1	1	1	1
7								1	0	1	1	1	1	0	0	0	0	0	0	1
8									0	0	0	0	0	0	0	0	1	1	0	1
9										1	1	1	1	1	0	0	0	1	1	1
10											1	1	1	0	0	0	1	0	0	1
11												1	1	0	0	0	1	1	1	1
12													1	0	0	0	1	1	0	1
13														0	0	0	1	1	1	1
14															1	0	0	0	0	0
15																1	0	0	0	0
16																	0	0	0	0
17																		1	1	1
18																			1	1
19																				0
20																				

Figure 2: Simulated adjacency matrix.

of the three tests (assuming an upper bound on the "experiment-wide" false alarm rate of $\alpha = 0.05$) is given by $z_{0.0170} = 2.1201$. Results of the multiple comparison are given in Table 4, and suggests that all the parameters are significantly different from each other. The fitted values are then given by

$$\hat{y}_{ij} = \begin{cases} \hat{p}_1 = 0.2857 & \text{If } i \cap j \in V_1 \\ \hat{p}_2 = 0.8077 & \text{If } i \cap j \in V_2 \\ \hat{p}_{12} = 0.0200 & \text{If } i \in V_1 \cap j \in V_2 \end{cases} \tag{34}$$

Table 4: Results of the multiple comparisons analysis.

Comparison	$\lvert z_0 \rvert$	P
$\hat{p}_1 - \hat{p}_2$	4.8240	0.0000
$\hat{p}_1 - \hat{p}_{12}$	2.6659	0.0038
$\hat{p}_2 - \hat{p}_{12}$	16.7684	0.0000

As an alternative, consider the logistic regression parameterization discussed previously. Since the nodal attribute contains two-levels, we need to create two dummy variables according to the coding scheme shown in Table 3. Results of the analysis are shown in Table 5 and, as expected, suggest that the fitted model is highly significant. Recall that the test statistic in logistic regression for testing that all slopes are equal to zero is given by

$$G = -2\left[\ell_0^* - \ell_1^*\right] = -2\Delta \tag{35}$$

which is minus 2 times the difference in the log-likelihoods maximized under H_0 and H_1. As shown previously, the G statistic is exactly the same as the $\hat{r}$ statistic derived under the proposed framework. Thus, for the simulated adjacency matrix in Figure 2, we find that $G = \hat{r} = 130.413 > \chi^2_{0.05,2}$, so that the null hypothesis is rejected.

Table 5: Results of logistic regression analysis.

Variable	$coef$	$se(coef)$	z	P	Odds Ratio
Constant	-0.9163	0.4830	-1.90	0.058	
x_1	-2.8792	0.8629	-3.34	0.001	0.056
x_2	2.3514	0.5620	4.18	0.000	10.50

Recall in logistic regression that odds ratios play a fundamental role in interpretation of the fitted model. An odds ratio is defined by the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. For example, the odds ratio corresponding to x_1 in Table 5 estimates

$$\exp\{\beta_1\} = \frac{\pi_{10}(1-\pi_{00})}{(1-\pi_{10})\pi_{00}} = \frac{p_{12}(1-p_1)}{(1-p_{12})p_1}$$

which is the ratio of the odds of observing a link between any two vertices not both contained in V_1 or V_2, to the odds of observing a link between any two vertices that are both contained in V_1. The notation π_{uv} $(u, v \in [0,1])$ is used to denote the expected probability of success given the u^{th} level of the first variable and the v^{th} level of the second variable. Thus, we are approximately twenty times more likely to observe a link between any two vertices contained in V_1, relative to that of any two vertices not both contained in the same set. Similarly, the estimated odds ratio corresponding to x_2 estimates

$$\exp\{\beta_2\} = \frac{\pi_{01}(1-\pi_{00})}{(1-\pi_{01})\pi_{00}} = \frac{p_2(1-p_1)}{(1-p_2)p_1}$$

which is the ratio of the odds of observing a link between any two vertices both contained in V_2, to the odds of observing a link between any two vertices that are both contained in V_1. The estimated odds ratio then suggests that we are approximately 10 times more likely to observe a link between any two vertices contained in V_2, relative to that of any two vertices contained in V_1.

Recall that, in general, the logistic regression model is given by

$$E(y_i) = \pi_i = \frac{1}{1+\exp\{-\mathbf{x}_i'\boldsymbol{\beta}\}} \tag{36}$$

for $i = 1, 2, ..., N$, so that the fitted logistic regression model for the simulated network in Figure 2 is then given as

$$\hat{y}_{ij} = \begin{cases} (1 + \exp\{0.9163\})^{-1} = 0.2857 & \text{If } i \cap j \in V_1 \\ (1 + \exp\{-1.435\})^{-1} = 0.8077 & \text{If } i \cap j \in V_2 \\ (1 + \exp\{3.7955\})^{-1} = 0.0200 & \text{If } i \in V_1 \cap j \in V_2 \end{cases} \tag{37}$$

where, as expected, the fitted values obtained from the fitted logistic regression model are exactly equal to those obtained from the proposed analysis framework.

5 Summary and Discussion

Undirected networks are used in a plethora of applications across many disciplines. For example, they are often used to model communication networks, financial networks, transportation networks, biological networks, social networks, and many more. In this paper, a tractable hypothesis testing framework for application to undirected "noisy" binary networks was presented. The methodology involves partitioning network actors into m mutually exclusive and collectively exhaustive subsets (or levels), based upon available exogenous nodal attribute information. Edge density within and between partition levels is then measured, and the observed differences between these measurements are then evaluated statistically using the log-likelihood ratio statistic. It was shown that the proposed approach is equivalent to logistic regression when the regressor variable is categorical, and dummy variables are used in the logit to model the effect of this variable.

Although the proposed approach is a special case of logistic regression, its parameterization permits more easily interpretable analysis results when applied to networks. This is because the proposed method works with the p_h's and p_{ij}'s directly, whereas logistic regression works with the p_h's and p_{ij}'s indirectly via the β's (which are the coefficients of the indicator variables). Therefore, in logistic regression, a transformation is needed to compute and subsequently interpret the odds ratios. Using the proposed method, estimated odds ratios can be computed directly from the estimated p_h's and p_{ij}'s. Further, analysis using the proposed method does not require a logistic regression software package; it can be performed using even simple spreadsheet software.

As a final note, the proposed method also lends nicely to combinatorial optimization algorithms in cases where exogenous *nodal* attribute information does not exist. That is, in the absence of nodal attribute data, suppose the analyst is interested in finding that partition that best explains the variability in the edge set. Let Z denote the set of all $m = 2$ level partitions, then the analyst is interested in

$$z^* = \arg\max_{z \in Z}\{\hat{r}(z)\} \tag{38}$$

where z^* denotes the maximum likelihood estimator for the 2-level partition of V; that is, the partition in Z that maximizes $\hat{r}$ in eq.(19). For smaller networks, total enumeration is possible. However, for larger networks, total enumeration is not feasible in real time. For example, if the vertex set has cardinality $n = 25$, and if a 2-level partition is considered, then there are a total of $33,554,380$ partitions in the set Z. Unfortunately, this number will grow exponentially with the cardinality of V. As a result, alternative solution methods are needed. The first author of this paper is currently studying the application of meta-heuristics to the optimization problem in eq.(38) with promising results.

References

[1] Frank, O. and Strauss, D. (1986). "Markov Graphs." *Journal of the American Statistical Association* 81: 832-842.

[2] Handcock, M. S. (2003). "Statistical Models for Social Networks: Degeneracy and Inference." Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers, edited by R. Breiger, K. Carley, and P. Pattison. National Research Council of the National Academies. Washington, DC: The National Academies Press.

[3] McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001). "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415-444.

[4] Pattison, P. E. and Wasserman, S. (1999). "Logit and Logistic Regression for Social Networks II: Multivariate Relations." *British Journal of Mathematical and Statistical Psychology* 52: 169-194.

[5] Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*, Duxbury Press, Belmont, CA.

[6] Robins, G. L., Pattison, P. E., and Wasserman, S. (1999). "Logit models and logistic regression for Social Networks III: Valued Relations." *Psychometrika* 64: 371-394.

[7] Sageman, M. (2004). *Understanding Terror Networks*, University of Pennsylvania Press.

[8] Smyth, P. (2003). “Statistical Modeling of Graph and Network Data.” *IJCAI Workshop on Learning Statistical Models from Relational Data.*

[9] Snijders, T. A. B. (2002). “Markov Chain Monte Carlo Estimation of Exponential Random Graph Models.” *Journal of Social Structure* 3:2.

[10] Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). “New Specifications for Exponential Random Graph Models.” *Sociological Methodology* 36:(1) 99-153.

[11] Wasserman, S. and Pattison, P. E. (1996). “Logit Models and Logistic Regression for Social Networks I: An Introduction to Markov Random Graphs and p*.” *Psychometrika*, 61: 401-425.