Statistics and Its Interface Volume 8 (2015) 267–276

Exponential random graph models for networks resilient to targeted attacks

Jingfei Zhang and Yuguo Chen*,†

One important question for complex networks is how the network's connectivity will be affected if the network is under targeted attacks, i.e., the nodes with the most links are attacked. In this paper, we fit an exponential random graph model to a dolphin network which is known to be resilient to targeted attacks. The fitted model characterizes network resiliency and identifies local structures that can reproduce the global resilience property. Such a statistical model can be used to build the Internet and other networks to increase the attack tolerance of those networks.

Keywords and phrases: Exponential random graph model, Global efficiency, Markov chain Monte Carlo, Maximum likelihood estimation, Network robustness, Random graphs.

1. INTRODUCTION

In recent years, there has been an increasing interest in studying the effects of attacks on real-world networks. For example, terrorist attacks on physical networks, such as power networks, transportation networks, or the Internet, can traumatize modern societies (Arianos et al., 2009; Albert et al., 2000; Schneider et al., 2011). Attacks by hackers on computer networks can lead to security breaches in the cyber space. Two types of attacks have been studied in the literature: random attacks and targeted attacks. Random attacks mean the nodes are attacked at random, and targeted attacks mean the nodes with the highest degrees (i.e., with the most links) are attacked. Many complex networks are quite robust to random attacks, but are highly vulnerable to target attacks (Albert et al., 2000). In social networks and technology networks, the degree distributions are mostly disclosive, so the high degree nodes are often easy to identify and are exposed to targeted attacks. In some networks, the high degree nodes are more likely to fail due to their load. Since the most connected nodes play an important role in maintaining the network's connectivity, it is important to understand how to design networks that are robust to targeted attacks.

*Corresponding author.
†Partially supported by the NSF grant DMS-1106796.

It is well known that the Internet and the World Wide Web (WWW) are highly vulnerable to targeted attacks (Albert et al., 2000). If the 2.5% most connected nodes in the Internet are removed, then the diameter (average length of the shortest paths between any pair of nodes) of the Internet more than triples (Albert et al., 2000). Surprisingly Lusseau (2003) found that for the network of 62 bottlenose dolphins in a community at Doubtful Sound, New Zealand, the diameter only increases by 5.78% when 5% of the dolphins with the most links are removed. Figure 1 is the social network of these bottlenose dolphins. Two dolphins are linked if they were seen together more frequent than expected. The data were collected between 1995 and 2001. Every time a school of dolphins was sighted, all adult members were photographed and identified based on the natural markings on their dorsal fins. These data provide information on the frequency that two dolphins were seen together. Based on these data, Lusseau (2003) implemented a permutation test to determine whether two dolphins were seen together more often than by chance, which is referred to as 'preferred companionship' in Lusseau (2003). The network in Figure 1 was constructed based on preferred companionship.

The dolphin network has a very small increase in diameter even when the community is under targeted attacks. Zhang and Chen (2013) developed an efficient sequential sampling algorithm to compare the dolphin network with random networks with the same degree sequence, and they concluded that such a small change in diameter is statistically significant. This indicates that the dolphin network is formed in a particular way (instead of randomly linked with each other) that is resilient to targeted attacks.

Most of the existing approaches for studying attack tolerance rely on analyzing one or two statistics of the network, such as the diameter, global efficiency, local efficiency, clustering coefficient, or the size of the largest connected cluster (Albert et al., 2000; Crucitti et al., 2003; Schneider et al., 2011). In order to better understand the resilience property of the dolphin network and find statistical models that characterize network resiliency, we fit the dolphin network with the exponential random graph model (ERGM), which is one of the most widely used models for social network analysis (Wasserman and Pattison, 1996; Robins et al., 2007a). The ERGM involves a set of local structures of the network, so fitting an ERGM can help us understand what kind of local features can contribute to the global resilience property.

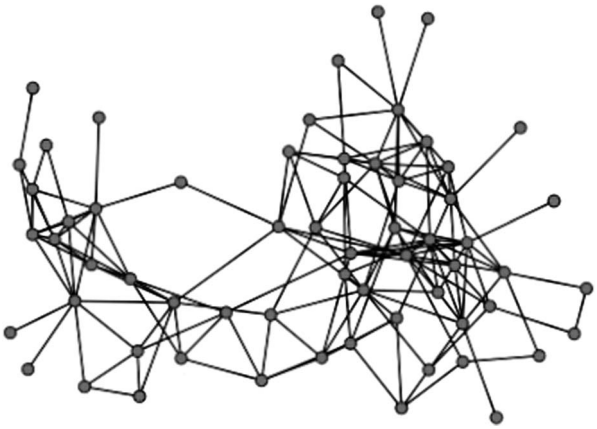

Figure 1. Social network of 62 bottlenose dolphins in Doubtful Sound, New Zealand.

Such a statistical model can also be used to build the Internet and other networks to increase the attack tolerance of those networks.

The paper is organized as follows. Section 2 introduces the global efficiency measure and explains the resilience property of the dolphin network. Section 3 reviews the ERGM and the model fitting procedure. Section 4 discusses the constraints we put on the network space. Section 5 fits the ERGM to the dolphin network. Section 6 studies the attack tolerance of the fitted model, and Section 7 provides concluding remarks.

2. RESILIENCE PROPERTY OF THE DOLPHIN NETWORK

In the discussion of the attack tolerance of the Internet, the World Wide Web, and the bottlenose dolphins in the Introduction, we used the change of the diameter after the attack as the measure. The diameter of a network is the average of the minimum distances between any pair of nodes in the network. It is a well-studied, important network metric because it is one of the metrics that characterize the small world property of networks. However, the diameter is not well defined for networks that are not connected. This is problematic because in practice some networks consist of several isolated fragments or become disconnected after attacks. In that case, it is often up to the researchers to redefine diameters.

Recently, another measure called global efficiency was proposed to characterize the small world property of networks (Latora and Marchiori, 2001). For a network G with n nodes, its global efficiency is defined as

$$E(G) = \frac{1}{n(n-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}, \tag{1}$$

where d_{ij} is the length of the shortest path between nodes i and j. The global efficiency is closely related to the diameter because the diameter is the average of d_{ij} instead of $1/d_{ij}$. If nodes i and j are disconnected, then $d_{ij} = \infty$ and $1/d_{ij} = 0$, so global efficiency is well defined for disconnected networks as well. The global efficiency is always between 0 and 1, with $E(G) = 0$ for an empty graph with no edges and $E(G) = 1$ for a complete graph with all $n(n-1)/2$ possible edges.

Global efficiency has been shown to be a better measure than the diameter for describing the global properties of complex networks, especially when a large number of nodes are removed (Crucitti et al., 2003). Therefore we will use the percentage of global efficiency change after the attack to measure network resilience. It is shown by Crucitti et al. (2003) that scale-free networks are extremely vulnerable to targeted attacks in terms of the global efficiency. We looked at the global efficiency change for two real data sets. The first is the Internet router-level network based on the ITDK0304 skitter data between April 21 and May 8 of 2003. The data is available at the web page of the Cooperative Association for Internet Data Analysis (http://www.caida.org/tools/measurement/skitter/router_topology/). This network contains 192,244 nodes and 609,066 undirected edges. After removing the 2.5% most connected nodes, the global efficiency reduced from 0.1501 to 0.0696, which is a decrease of 53.63%. The second data set is a subset of the World Wide Web containing 325,729 nodes and 1,090,108 undirected edges (Albert et al., 1999). After a 2.5% targeted attack, the global efficiency reduced from 0.1535 to 0.0189, which is a decrease of 87.69%. Such vulnerability to targeted attacks is also observed in simulated scale-free networks similar to the Internet and the World Wide Web (Crucitti et al., 2003).

The bottleneck dolphin network has 62 nodes and 159 edges with a global efficiency of 0.3792. After removing the three most connected individuals (about 5% of the community), the global efficiency becomes 0.3585 which only decreases by 5.459%. This is a very small change comparing to the behavior of other complex real world networks. It also shows that under the global efficiency measure, the dolphin network is still resilient to targeted attacks. To test the statistical significance of this small change in global efficiency, we compared the dolphin network with random networks having the same degree sequence as the dolphin network. Totally 1,000 random networks were generated using the sequential importance sampling algorithm developed in Zhang and Chen (2013), and for each network the percentage of global efficiency change is computed after the removal of the three most connected nodes. The histogram of the 1,000 values of the percentage of global efficiency change is given in Figure 2, and the probability of having a change of global efficiency less than or equal to 5.459% is estimated to be 0.0152 with standard error 0.0039. This shows that the dolphin network is formed in a way that has a very high attack tolerance comparing to other random networks with the same degree sequence. In this paper, we fit a statistical model to the dolphin network to understand its resilience property.

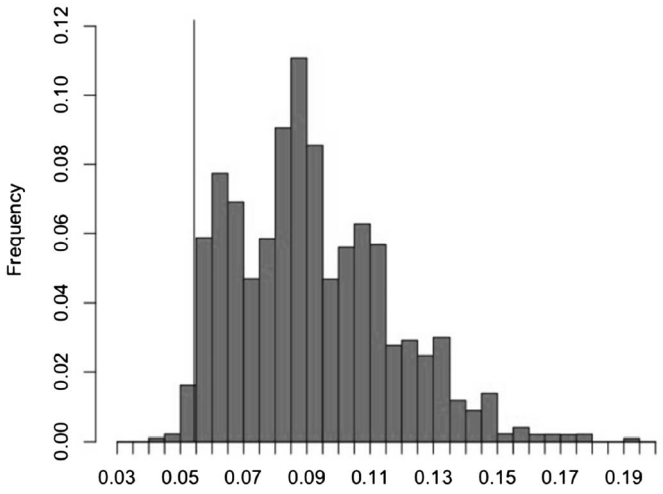

Figure 2. The histogram of the percentage of global efficiency changes based on 1,000 random samples. The vertical line indicates the value calculated from the observed dolphin network.

3. EXPONENTIAL RANDOM GRAPH MODELS

A network (or graph) G with n vertices (or nodes) V and a set of edges (or links) E can be represented by its adjacency matrix $\mathbf{y}$, where $y_{ij} = 1$ if there is an edge from node i to node j and 0 otherwise. The degree of a node is the number of edges incident to the node. We use $\{i, j\}$ to denote an edge between node i and node j. In this paper, we are mainly concerned with simple undirected graphs (no loops or multiple edges) because the dolphin network is of this type. Therefore $\mathbf{y}$ is an $n \times n$ symmetric 0–1 matrix with a zero diagonal.

The exponential random graph model (ERGM) specifies a probability distribution on the space $\mathcal{Y}$ of all graphs under consideration

$$P_\theta(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\theta^T g(\mathbf{y})\}}{\kappa(\theta)}, \tag{2}$$

where $\kappa(\theta)$ is the normalizing constant, $\theta = (\theta_1, \ldots, \theta_p)$ is the parameter, and the statistics $g(\mathbf{y}) = (g_1(\mathbf{y}), \ldots, g_p(\mathbf{y}))$ are counts of graph structures or features of the network. Sometimes $g(\mathbf{y})$ can also incorporate additional covariates of the network. The normalizing constant $\kappa(\theta)$ usually cannot be computed explicitly even for a moderate size graph because it involves the summation over all $\mathbf{y} \in \mathcal{Y}$.

ERGMs have been used extensively in the study of networks (Wasserman and Pattison, 1996; Robins et al., 2007a, 2007b). The statistics $g(\mathbf{y})$ often include a set of local structures of the networks. Some local rules can describe the transitivity of the network, such as the ratio of the number of triangles to the number of two-stars. Some can provide information on how well the network conveys information, such as the total number of edges. To fit an ERGM, we need to identify a subset of local measures that can concisely summarize the global property of a network. A well fitted ERGM can help us understand how the global structure can be reproduced by local metrics and how local rules could affect the global property of a network. We can also use the fitted ERGM to design networks with certain properties.

3.1 Network statistics

Although basic local structures, such as star counts, triangle counts, and the degree distribution, are traditional candidates for the local measures (Frank and Strauss, 1986), it is pointed out in Snijders et al. (2006) that including such basic terms could result in a probability model which concentrates its mass at either the full graph or the empty graph. This so called "degeneracy" phenomenon makes it very difficult to have reasonable parameter estimation, and places a serious barrier between specifying a reasonable ERGM and making reliable parameter estimation. However, the degeneracy issue is caused not by the ERGM itself, but by the network statistics chosen to be included in the model (Snijders et al., 2006). Hunter (2007) discussed three new network metrics: geometrically weighted degree (GWD), geometrically weighted edgewise shared partner (GWESP), and geometrically weighted dyadwise shared partner (GWDSP). These new statistics not only help avoid the degeneracy problem, but also provide insight on network structures from a different perspective. The definitions of these statistics are given below.

For a network with n nodes and an $n \times n$ adjacency matrix $\mathbf{y}$, let $D_i(\mathbf{y})$ be the number of nodes in $\mathbf{y}$ with i edges. Then $D_i(\mathbf{y})$, $i = 0, \ldots, n-1$, are the degree distribution of $\mathbf{y}$, and they satisfy the linear constraint $D_0(\mathbf{y}) + \cdots + D_{n-1}(\mathbf{y}) = n$. For a given i, let $EP_i(\mathbf{y})$ be the number of edges $\{k, l\}$ such that nodes k and l are linked through an edge (i.e., nodes k and l are neighbors) and they share exactly i partners in common (i.e., there are exactly i nodes that are linked to both nodes k and l). Then $EP_i(\mathbf{y})$, $i = 0, \ldots, n-2$, are the edgewise shared partner distribution of $\mathbf{y}$, and the sum $EP_0(\mathbf{y}) + \cdots + EP_{n-2}(\mathbf{y})$ equals the total number of edges in the graph. For a given i, let $DP_i(\mathbf{y})$ be the number of dyads (k, l) such that nodes k and l share exactly i partners in common. Here the dyad k and l do not need to be neighbors of each other. Then $DP_i(\mathbf{y})$, $i = 0, \ldots, n-2$, are the dyadwise shared partner distribution of $\mathbf{y}$, and the sum $DP_0(\mathbf{y}) + \cdots + DP_{n-2}(\mathbf{y})$ equals the total number of dyads in the graph. For a given i, define the non-edgewise shared partner $NSP_i(\mathbf{y})$ as $DP_i(\mathbf{y}) - EP_i(\mathbf{y})$, which equals the number of dyads in the network that are not connected but share exactly i partners in common.

Based on the above terms, the statistics GWD, GWESP, GWDSP and geometrically weighted non-edgewise shared

partner (GWNSP) are defined as:

$$\text{GWD} = e^{\tau_1}\sum_{i=1}^{n-1}\{1-(1-e^{-\tau_1})^i\}D_i(\mathbf{y}), \quad (3)$$

$$\text{GWESP} = e^{\tau_2}\sum_{i=1}^{n-2}\{1-(1-e^{-\tau_2})^i\}EP_i(\mathbf{y}), \quad (4)$$

$$\text{GWDSP} = e^{\tau_3}\sum_{i=1}^{n-2}\{1-(1-e^{-\tau_3})^i\}DP_i(\mathbf{y}), \quad (5)$$

$$\text{GWNSP} = e^{\tau_4}\sum_{i=1}^{n-2}\{1-(1-e^{-\tau_4})^i\}NSP_i(\mathbf{y}). \quad (6)$$

Here $\tau_i \geq 0$, $i = 1, \ldots, 4$, are decay parameters. When $\tau_2 = \tau_3 = \tau_4$, it is easy to see that GWDSP = GWESP+GWNSP. The intuition behind the four geometrically weighted metrics is to constrain the effect of higher order terms in the summation and control the degeneracy problem. As explained in Snijders et al. (2006), a model with these geometrically weighted metrics can avoid the model degeneracy problem and capture the higher order dependency structure in the network. More details about these four metrics can be found in Snijders et al. (2006) and Hunter (2007).

3.2 Model fitting

We consider estimating the parameters in the ERGM by the maximum likelihood method. Because the analytical form of the maximum likelihood estimate (MLE) is not available for the ERGM, finding the MLE is normally done with either a Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) (Geyer and Thompson, 1992; Snijders, 2002) or maximum pseudo-likelihood estimation (MPLE) (Frank and Strauss, 1986; Strauss and Ikeda, 1990). Although the MPLE procedure is easier to implement, it can produce unreliable estimates. In this paper, we use MCMCMLE to make inference on the ERGM. See Hunter and Handcock (2006) for the details of the MCMCMLE method. In this paper, fitting and simulating from the ERGMs are done through the R package "ergm" (version 2.4–3) (Hunter et al., 2008b).

4. CONSTRAINTS ON THE NETWORK SPACE

In this section, we specify the space $\mathcal{Y}$ of all networks of interest in the ERGM (2). Usually the number of nodes n is fixed. In that case, the number of edges in the network plays an important role in network resilience. In general, if we add more edges to the network, it will increase the global efficiency of the network. An extreme case is the complete graph whose global efficiency is 1 before and after targeted attacks, so it is most resilient to targeted attacks. However the complete graph is not of particular interest here.

If we fix both the number of edges and the number of nodes in the network to control for the effect of edge density, it seems networks with evenly distributed degrees tend to have a high attack tolerance. The importance of the degree distribution in attack tolerance is discussed in Albert et al. (2000). They showed that a network from the Erdös-Rényi model (Erdös and Rényi, 1960), in which the expected degree of each node is the same, tends to have high tolerance to targeted attacks. On the other hand, for some scale-free networks with inhomogeneous power-law degree distribution, they are vulnerable to targeted attacks. To control for both the effect of edge density and the effect of degree variation, we fix the degree sequence $(d_1, \ldots, d_n)$ in this paper. Of course this implies that the number of nodes and the number of edges are fixed as well. So the space $\mathcal{Y}$ consists of all networks with the same degree sequence as the observed network.

Fixing the degree sequence has been considered in the literature for various reasons. Schneider et al. (2011) argued that in practice we cannot keep adding edges to increase the robustness of the network because the cost of adding links between every pair of nodes is too expensive in the context of power grids or the Internet. They also assumed that changing the node degree can be much more expensive than changing the links between nodes. In some other situations, fixing the degree sequence may create a basis for exact inference because they are sufficient statistics for the unknown parameters (Chen, 2007). This is sometimes related to random graphs with given degrees which have been used to model complex networks. Another reason to fix the degree sequence is that the degree of a node may reflect certain inherent characteristics of an individual, such as the capacity of a machine or the friendliness of a person. These characteristics may not be changeable, and we may need to model the network with these quantities fixed.

The network of 62 bottlenose dolphins has low edge density and unevenly distributed degrees $\mathbf{d} = (6, 8, 4, 3, 1, 4, 6, 5, 6, 7, 5, 1, 1, 8, 12, 7, 6, 9, 7, 4, 9, 6, 1, 3, 6, 3, 3, 5, 5, 9, 5, 1, 3, 10, 5, 1, 7, 11, 8, 2, 8, 5, 6, 7, 4, 11, 2, 6, 1, 2, 7, 10, 4, 2, 7, 2, 2, 9, 1, 5, 1, 3)$, see Figure 3 for the histogram of the degree sequence. Conditioning on the degree sequence allows us to make a conditional inference on how and to what extent configurations of local rules could affect attack tolerance. The set of all graphs with the same degree sequence as the dolphin network is the space $\mathcal{Y}$ for the ERGM (2). The space $\mathcal{Y}$ is still enormous, containing about 1.826×10^{167} networks (Zhang and Chen, 2013). We hope to find an appropriate model for the dolphin network under these constraints.

5. MODEL FITTING FOR THE DOLPHIN NETWORK

In this section, we fit ERGMs to the network of 62 bottlenose dolphins. A list of network statistics can be potential candidates for $g(\mathbf{y})$ in the ERGM (2). However since the degree sequence is fixed, metrics such as the number of edges,

*Table 1. Parameter estimates and their standard errors (in parentheses) for the three ERGMs. Here * means significant at the 0.05 level*

Coefficients	Model I (GWESP)	Model II (GWNSP)	Model III (GWESP and GWNSP)
θ_1	1.468 (0.129)*	–	0.058 (0.112)
θ_2	–	−0.313 (0.033)*	−0.421 (0.015)*

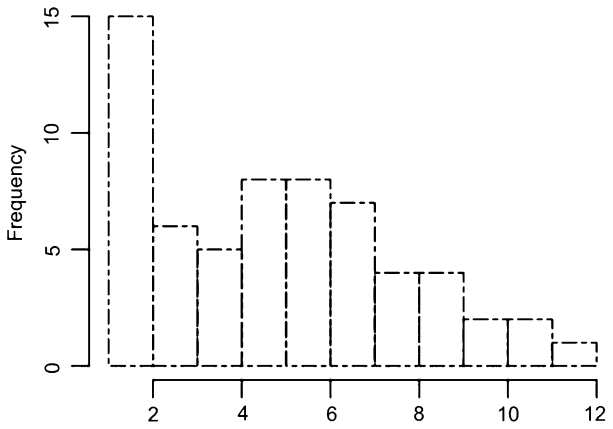

Figure 3. The histogram of the degree sequence of the dolphin network.

the number of nodes with degree k, the number of two-paths, k-star counts, and the GWD all become fixed numbers. Although for networks with fixed degree sequence, the degeneracy issue discussed in Section 3.1 should not occur, we find that the ERGM with simple local structures, such as triangle count or the number of open triads, does not fit the data well. It may be necessary to include network statistics that can capture the higher order dependency structure in the network, such as the geometrically weighted metrics defined in Section 3.1. In the following, we mainly look at four statistics that are not constants in the set $\mathcal{Y}$: k-cycle counts, GWESP, GWDSP, and GWNSP.

In our study of the model fitting, we found that there is no particular advantage to choose different decay parameter τ for the three geometrically weighted metrics. When τ is chosen to be the same, we have GWESP+GWNSP = GWDSP, and this linear relation implies there is no need to consider GWDSP. Based on the definition of edgewise shared partners and dyadwise shared partners, we can write the number of cycles as (Hunter et al., 2008a)

$$(7)\qquad \text{number of 3-cycle} \;=\; \frac{1}{3}\sum_{i=1}^{n-2} iEP_i(\mathbf{y}),$$

$$(8)\qquad \text{number of 4-cycle} \;=\; \frac{1}{2}\sum_{i=2}^{n-2} \binom{i}{2} DP_i(\mathbf{y}).$$

Since GWESP and GWDSP are also weighted sums of $EP_i(\mathbf{y})$ and $DP_i(\mathbf{y})$, there is a subtle connection between k-cycle counts and geometrically weighted metrics. We found through model selection that when GWESP and GWNSP are included, adding k-cycle counts does not improve the fitting of the model. Therefore only two metrics GWESP and GWNSP will be considered for $g(\mathbf{y})$. That leads to three possible ERGMs with the exponent $\theta^T g(\mathbf{y})$ being $\theta_1 \cdot$ GWESP (Model I), $\theta_2 \cdot$ GWNSP (Model II) and $\theta_1 \cdot$ GWESP $+ \theta_2 \cdot$ GWNSP (Model III), respectively.

We fitted these three models to the observed dolphin network. We found that for the decay parameter τ ranging from 0.1 to 0.5, MCMCMLE gave similar estimates for parameters θ_1 and θ_2. The approximate AIC (Akaike information criterion) values for fitted models with different τ are also similar with τ around 0.4 being slightly better than others. Therefore we fix $\tau = 0.4$ in the model fitting. Table 1 gives the estimates of the parameters for the three models, and each model is fitted and diagnosed with R package "ergm".

The estimate of θ_1, the coefficient for GWESP, is positive for both Models I and III. This indicates that two neighboring individuals are encouraged to share partners. Consider the hypothetical situation that the number of shared partners for a pair of neighboring nodes with k shared partners is increased to $k+1$, and assume this only results in a change of (EP_k, EP_{k+1}) to $(EP_k - 1, EP_{k+1} + 1)$ and all other EP_i and NSP_i are not affected. This assumption is difficult to satisfy in practice because the increase of the shared partner for one neighboring pair typically will affect the edgewise shared partner distribution as well as the non-edgewise shared partner distribution. However, studying the probability change under this seemingly unrealistic assumption can provide some insight on what kind of networks the model favors. Let P_{before} and P_{after} denote the probability of the network before and after the change $(EP_k, EP_{k+1}) \to (EP_k - 1, EP_{k+1} + 1)$ occurs. We have

$$\begin{aligned}(9)\qquad &\log\left(\frac{P_{after}}{P_{before}}\right)\\ &= \left(\theta_1 e^{\tau}\left[(EP_k - 1)\left(1-\rho^k\right) + (EP_{k+1}+1)\left(1-\rho^{k+1}\right)\right]\right)\\ &\quad - \left(\theta_1 e^{\tau}\left[EP_k\left(1-\rho^k\right) + EP_{k+1}\left(1-\rho^{k+1}\right)\right]\right)\\ &= \theta_1\rho^k,\end{aligned}$$

where $\rho = 1 - e^{-\tau}$. Therefore the change of (EP_k, EP_{k+1}) to $(EP_k - 1, EP_{k+1} + 1)$ will result in a log probability change of $\theta_1\rho^k$. Since θ_1 is positive in our models, we can see that the probability increases when a neighboring pair obtain one more shared partner, but the additional gain in probability due to the increase of one shared partner decreases as the number of shared partners k increases.

The estimate of θ_2, the coefficient for GWNSP, is negative for both Models II and III. This indicates that if two individuals are not connected, the model discourages them to have shared partners. In other words, two individuals are discouraged to have distance two. Interestingly, similar properties have been discussed for the brain network, i.e., two nodes have a direct connection if needed, but otherwise prefer a longer path between them to maintain efficiency or stability of the network (Simpson et al., 2011). We can also look at how the probability of a network changes when a pair of unconnected nodes increase their shared partner count by one, assuming this does not affect other terms in the model. Similar analysis as (10) suggests that the reduction in probability due to the increase of one shared partner decreases as the number of shared partners k increases. In Model III, the estimate of θ_1 is small and it is not significant at the 0.05 level. This shows that the GWESP is a less important term than the GWNSP.

5.1 Goodness of fit test

To select an appropriate ERGM from the three models under consideration, traditional criteria that involve the likelihood function have their limitations because the intractable normalizing constant $\kappa(\theta)$ cannot be computed directly and some approximation will be necessary. It is also hard to use the traditional criteria to answer the central question in fitting ERGMs, i.e., can the global structures be reproduced by the local rules? To emphasize this special aspect of ERGM fitting, Hunter et al. (2008a) proposed to simulate a number of samples from the fitted model and compare the values of a set of network statistics in the observed network to those calculated from sampled networks. If the comparison shows that one or more of the observed network statistics are not typical, it indicates that the model does not fit well.

The set of network statistics used in the comparison should characterize different aspects of network structures. Hunter et al. (2008a) proposed using the degree distribution, the minimum geodesic distance distribution, and the edgewise shared partner distribution as the statistics. Since the degree distribution is fixed in our network space, it is not of interest to consider that in our study. The minimum geodesic distance for any pair of nodes is the length of the shortest path connecting them. It is one of the most important metrics of networks and many useful characteristic metrics, such as the diameter and vertex betweenness, are calculated based on the minimum geodesic distance. The edgewise shared partner can quantify the clustering of the network and give triangle counts and other high order metrics. In this paper, we select models based on Hunter et al.'s (2008a) graphical goodness of fit method using the minimum geodesic distance distribution and the edgewise shared partner distribution.

We generated 100 samples from each fitted model and the goodness of fit plots for each model are given in Figure 4. We can see that for Model III, the observed network statistics always fall in the 95% confidence intervals formed by the simulated networks, but that is not the case for Models I and II. In terms of the minimum geodesic distance, both Models I and II overestimate the number of dyads with minimum distances 2 and 3, but underestimate the number of dyads with minimum distance 5, 6, 7, etc. This shows that comparing to the observed network, the distance between a pair of nodes tends to be shorter in the simulated networks from Models I and II. In terms of edgewise shared partners, Model I underestimates $EP_0(\mathbf{y})$ which denotes the number of neighboring pairs that share no partners in common. Because the sum of $EP_i(\mathbf{y})$ equals the total number of edges which is a fixed constant here, we can see that comparing to the observed network, more neighboring pairs in the networks generated from Model I share common partners. On the contrary, Model II overestimates $EP_0(\mathbf{y})$. Both Models I and II seem to overestimate $EP_1(\mathbf{y})$ which denotes the number of neighboring pairs that share one partner in common. Therefore, based on the goodness of fit plots, Model III has the best fit among the three.

6. ATTACK TOLERANCE OF THE FITTED MODELS

In this section, we study the attack tolerance of the samples from three fitted models. For each model, we generated 5,000 samples from the model and computed the percentage of global efficiency change for each sample under the same targeted attack (removing three most connected individuals). The histogram of the percentage of global efficiency change for samples generated from each model is given in Figure 5.

The global efficiency for the dolphin network decreases 5.459% after the 5% targeted attack. Based on the 5,000 samples from Model III, the probability of seeing less than or equal to 5.459% global efficiency change is estimated to be 0.6398 with standard error 0.0152. This indicates that Model III does capture the resilience property of the dolphin network. Comparing with the random networks shown in Figure 2, we can see that samples from Model III are more resilient to targeted attacks than random networks. The samples from Model I, however, estimated the probability of seeing less than or equal to 5.459% global efficiency change to be 0.0116 with standard error 0.0034. This shows Model I does not capture the resilience property of the dolphin network. For Model II, the estimate for the same probability is 0.4054 with standard error 0.0155. This indicates that Model II also does pretty well in terms of capturing the resilience property of the dolphin network. Both Models II and III share the statistic GWNSP, and the simulation shows that the GWNSP is important for reproducing the resilience property of the dolphin network.

The following argument provides some connection between the GWNSP and the resilience property. We start

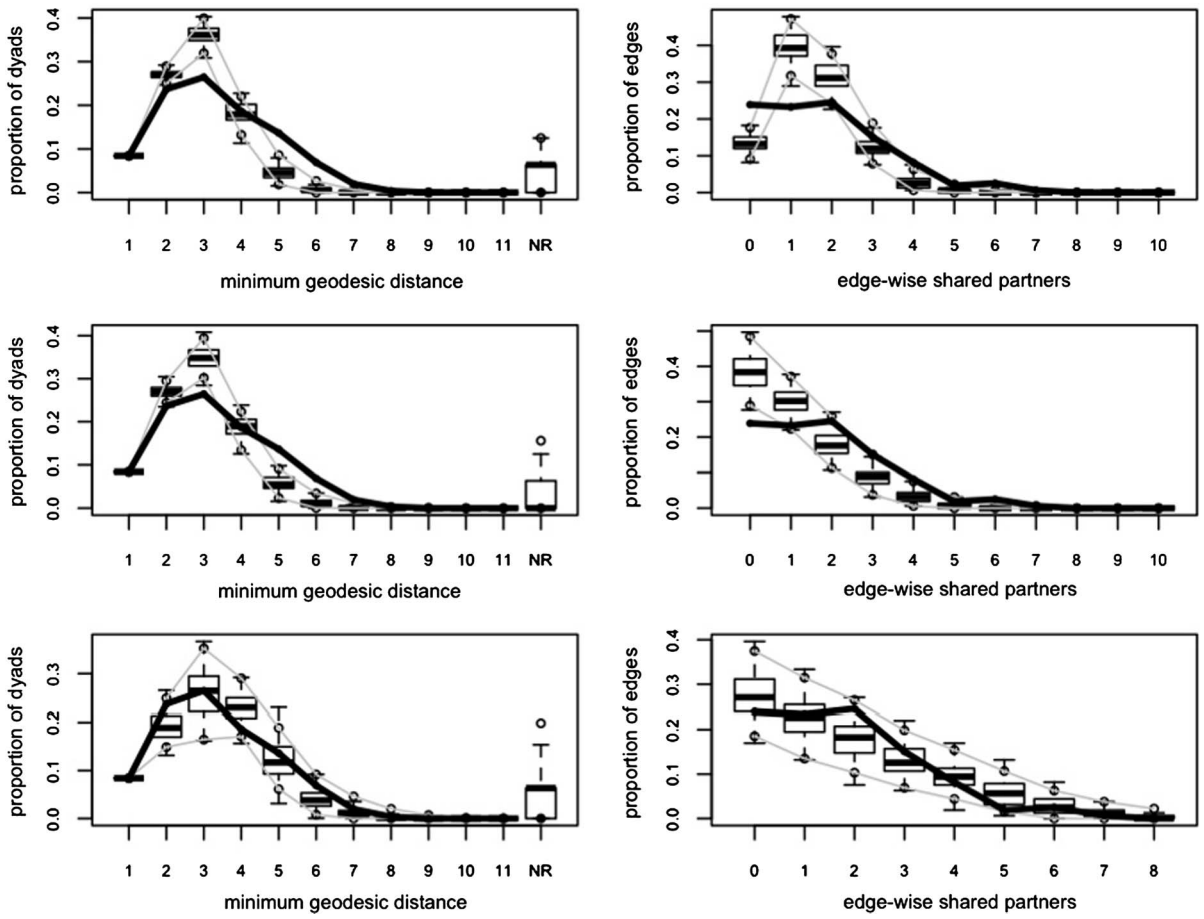

Figure 4. Goodness-of-fit plots for model I (top), Model II (middle), and Model III (bottom). In each plot, the black solid line indicates the statistics computed from the dolphin network. The grey lines indicate the range that covers 95% of the statistics computed from 100 sampled networks. The boxplot indicates the median and the interquartile range.

by rewriting the expression of the global efficiency as

(10)
$$E(G) = \frac{1}{n(n-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}} = \frac{2}{n(n-1)} \sum_{k=1}^{n-1} \frac{s_k}{k} \triangleq c \sum_{k=1}^{n-1} \frac{s_k}{k},$$

where s_k is the number of dyads (i, j) in the network with $d_{ij} = k$, and $c = 2/n(n-1)$. After a targeted attack of l most connected nodes, the global efficiency of the remaining graph $\tilde{G}$ becomes

(11)
$$\begin{aligned} E(\tilde{G}) &= \frac{1}{(n-l)(n-l-1)} \sum_{i \neq j \in \tilde{G}} \frac{1}{d_{ij}} \\ &= \frac{2}{(n-l)(n-l-1)} \sum_{k=1}^{n-l-1} \frac{\tilde{s}_k}{k} \\ &\triangleq \tilde{c} \sum_{k=1}^{n-l-1} \frac{\tilde{s}_k}{k}, \end{aligned}$$

where $\tilde{s}_k$ is the number of dyads (i, j) in $\tilde{G}$ with $d_{ij} = k$, and $\tilde{c} = 2/(n-l)(n-l-1)$. If $\tilde{G}$ is connected, a simple lower bound for $E(\tilde{G})$ is

(12)
$$E(\tilde{G}) \geq \tilde{c}\tilde{s}_1 + \tilde{c} \sum_{k=2}^{n-l-1} \frac{\tilde{s}_k}{n-l-1} = \tilde{c}\tilde{s}_1 + \tilde{c} \frac{\binom{n-l-1}{2} - \tilde{s}_1}{n-l-1}.$$

If $\tilde{G}$ is not connected, then $E(\tilde{G})$ can be simply bounded below by $\tilde{c}\tilde{s}_1$. Assume that $\tilde{G}$ is connected (the argument for the disconnected case is similar). Then the change of the global efficiency after the attack is

(13)
$$\begin{aligned} & E(G) - E(\tilde{G}) \\ &= c \sum_{k=1}^{n-1} \frac{s_k}{k} - \tilde{c} \sum_{k=1}^{n-l-1} \frac{\tilde{s}_k}{k} \\ &\leq c s_1 - \tilde{c}\tilde{s}_1 + c \sum_{k=2}^{n-1} \frac{s_k}{k} - \tilde{c} \frac{\binom{n-l-1}{2} - \tilde{s}_1}{n-l-1}. \end{aligned}$$

Since s_1 and $\tilde{s}_1$ equal to the total number of edges in G and $\tilde{G}$ respectively, they are fixed numbers. The last term in the upper bound is also fixed. Therefore the only term

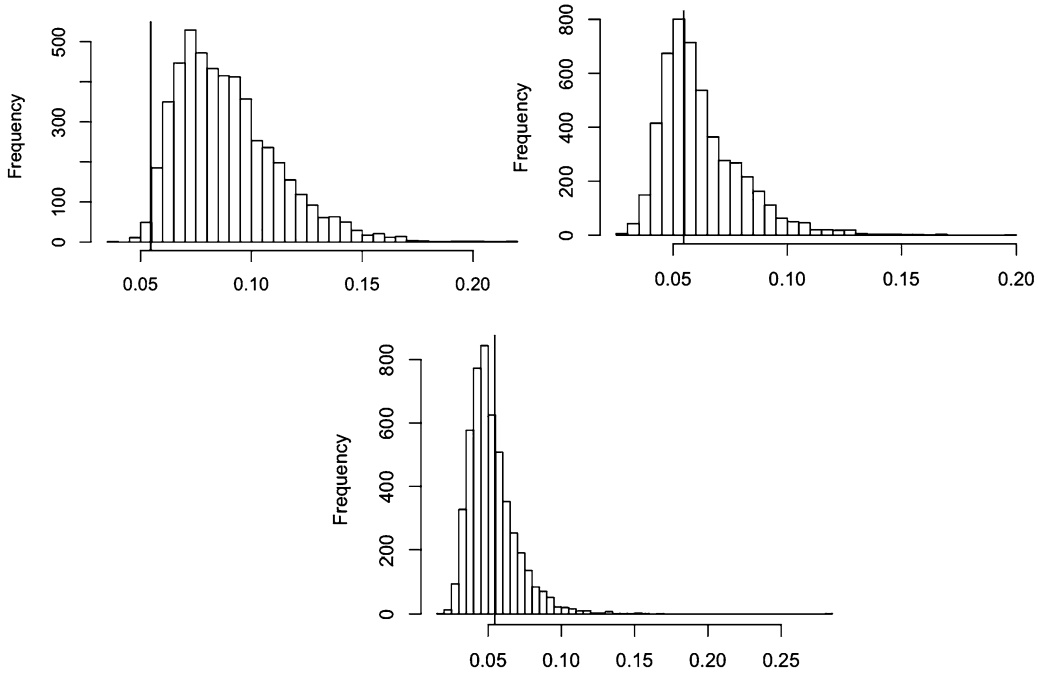

Figure 5. The histograms of the percentage of global efficiency change for samples generated from Model I (top left), Model II (top right), and Model III (bottom). The vertical line indicates the percentage of global efficiency change for the dolphin network.

that may vary is $\sum_{k=2}^{n-1} s_k/k$. Since $\sum_{k=1}^{n-1} s_k$ equals to the total number of dyads in G which is a fixed number, so $\sum_{k=2}^{n-1} s_k = \sum_{k=1}^{n-1} s_k - s_1$ is fixed as well. If we decrease s_2 and increase $s_3, \ldots, s_{n-1}$ correspondingly, then $\sum_{k=2}^{n-1} s_k/k$ would decrease because the leading term s_2 has the largest coefficient $1/2$.

There is a subtle connection between s_2 and the GWNSP. Notice that the GWNSP is defined as $e^{\tau_4} \sum_{i=1}^{n-2} \{1 - (1 - e^{-\tau_4})^i\} NSP_i(\mathbf{y})$. As i increases, the coefficient $1-(1-e^{-\tau_4})^i$ becomes close to 1. If we replace this coefficient by 1, we have an approximation to the GWNSP as $e^{\tau_4} \sum_{i=1}^{n-2} NSP_i(\mathbf{y}) = e^{\tau_4} s_2$. In other words, the GWNSP is approximately proportional to s_2. The term GWNSP is in Models II and III. For example, Model II takes the form of $P(\mathbf{Y} = \mathbf{y}) \propto \exp\{\theta_2 \cdot \text{GWNSP}\}$. When $\theta_2 < 0$, the model discourages networks with a large value of the GWNSP. Since the GWNSP is approximately proportional to s_2, Model II also tends to discourage networks with large value of s_2. As we argued earlier, a small s_2 will lead to a small $\sum_{k=2}^{n-1} s_k/k$ and a small upper bound of the global efficiency change in (14). Although a small upper bound does not necessarily mean a small global efficiency change, it still sheds light on why Models II and III with $\theta_2 < 0$ for the GWNSP tend to favor networks resilient to targeted attacks.

Because the term GWNSP seems to play an important role in producing networks with high attack tolerance, in the next simulation study, we look at how the attack tolerance of the model changes as we vary the coefficient θ_2 for GWNSP. Consider Model II

$$P(\mathbf{Y} = \mathbf{y}) \propto \exp\{\theta_2 \cdot \text{GWNSP}\}, \qquad \mathbf{y} \in \mathcal{Y}, \tag{14}$$

where the space $\mathcal{Y}$ consists of all networks with the same degree sequence as the dolphin network. We chose $\tau = 0.5$ and five different values for θ_2: -0.5, -0.25, 0, 0.25, and 0.5. For each θ_2, we used the R package "ergm" to generate 1,000 samples from the corresponding model. These samples are from an MCMC output with 50,000 burn-in steps and every thousandth sample in the chain is kept for inference. To study the attack tolerance of these five models, we computed the percentage of the global efficiency change after 5% targeted attacks for the 1,000 samples from each model, and then used kernel density estimation to obtain the plots in Figure 6. We can see that the density curve gradually shifts to the right as θ_2 increases, which indicates that the attack tolerance gradually decreases as θ_2 increases. Another interesting observation is that some of the 1,000 sampled networks are disconnected (before the attack), and the percentage of disconnected networks is 0.611, 0.302, 0.1, 0.017, and 0.008 for samples from the five models, which again shows a decreasing pattern as θ_2 increases.

We did the same simulation study for another degree sequence following the power-law distribution $P(k) \propto k^{-2.5}$ with 100 nodes and 179 edges. We used the same τ and θ_2 values as before, generated 1,000 samples from each model using the "ergm" package, and computed the percentage of global efficiency change after 5% targeted attacks. Figure 7

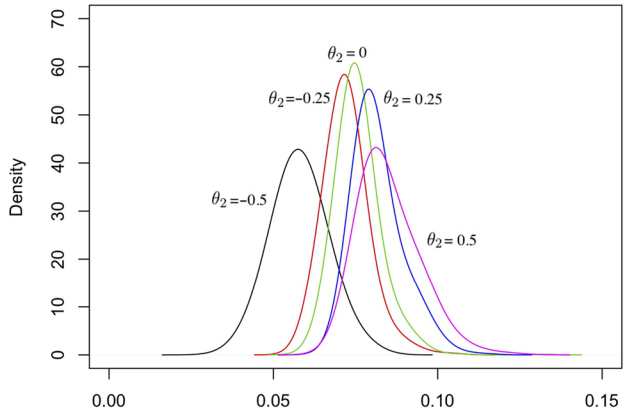

Figure 6. Densities of the percentage of global efficiency change after 5% targeted attacks for Model II on the space of networks with the same degree sequence as the dolphin network. Five different values for the parameter θ_2 in Model II are considered: -0.5, -0.25, 0, 0.25, and 0.5.

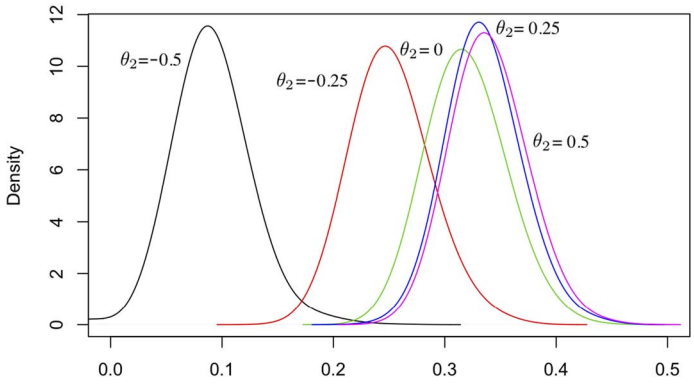

Figure 7. Densities of the percentage of global efficiency change after 5% targeted attacks for Model II on the space of networks with the same degree sequence as the one generated from a power-law with 100 nodes and 179 edges. Five different values for the parameter θ_2 in Model II are considered: -0.5, -0.25, 0, 0.25, and 0.5.

shows the same pattern as Figure 6, i.e., the density curve gradually shifts to the right as θ_2 increases. We did not observe disconnected networks from models with $\theta_2 = 0.25$ and 0.5. The percentage of disconnected networks from models with $\theta_2 = -0.5$, -0.25, and 0 are 0.109, 0.032, and 0.003, respectively. Again the percentage of disconnected networks decreases as θ_2 increases.

The simulation results suggest that if we need to build a network that is resilient to targeted attacks and the degree sequence is already given, we may sample a network from Model II with a negative parameter value θ_2. Or we may run an MCMC algorithm with Model II as the stationary distribution, and then pick a network from the MCMC output that has the highest attack tolerance. The idea of simulated annealing can be used as well. This could be useful for building the Internet, the World Wide Web, or some other networks to achieve high attack tolerance. Models with small θ_2 seem to have high attack tolerance, but they also tend to generate some disconnected graphs. If connectivity is a basic requirement, we can focus on the sampled networks that are connected.

7. DISCUSSION

In this paper, we fit ERGMs to a dolphin network to study its resilience to targeted attacks. To control for the effect of edge density and degree variation and focus on how the nodes are connected with each other to make the network resilient, we consider networks having the same degree sequence as the dolphin network. The local structures we identified that play an important role in the resilience property are GWNSP and GWESP, with GWNSP being the most important one. The samples generated from the fitted model show that the model captures the resilience property and fits the dolphin network well. Such a statistical model can be used to build the Internet and other networks with the same resilience property.

The attack tolerance is measured by the percentage of global efficiency change in our study. The conclusion still holds when the absolute change of global efficiency is used as the measure. Most figures, such as Figures 2 and 5, are similar under these two measures.

Received 3 May 2013

REFERENCES

Albert, R., Jeong, H. and Barabási, A.-L. (1999). Diameter of the World-Wide Web. *Nature* **401** 130–131.

Albert, R., Jeong, H. and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* **406** 378–382.

Arianos, S., Bompard, E., Carbone, A. and Xue, F. (2009). Power grids vulnerability: A complex network approach. *Chaos* **19** 013119.

Chen, Y. (2007). Conditional inference on tables with structural zeros. *Journal of Computational and Graphical Statistics* **16** 445–467. MR2370949

Crucitti, P., Latora, V., Marchiori, M. and Rapisarda, A. (2003). Efficiency of scale-free networks: Error and attack tolerance. *Physica A-Statistical Mechanics and Its Applications* **320** 622–642.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5** 17–60. MR0125031

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association* **81** 832–842. MR0860518

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society*, Series B **54** 657–699. MR1185217

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social Networks* **29** 216–230.

Hunter, D. R., Goodreau, S. M. and Handcock, M. S. (2008a). Goodness of fit of social network models. *Journal of the American Statistical Association* **103** 248–258. MR2394635

Hunter, D. R. and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics* **15** 565–583. MR2291264

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. and Morris, M. (2008b). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software* **24**.

Latora, V. and Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical Review Letters* **87** 198701.

Lusseau, D. (2003). The Emergent properties of a dolphin social network. *Proceedings of the Royal Society of London*, Series B (Suppl.) **270** S186–S188.

Robins, G. L., Pattison, P. E., Kalish, Y. and Lusher, D. (2007a). An introduction to exponential random graph ($p*$) models for social networks. *Social Networks* **29** 173–191.

Robins, G. L., Snijders, T. A. B., Wang, P., Handcock, M. S. and Pattison, P. E. (2007b). Recent development in exponential random graph models for social networks. *Social Networks* **29** 192–215.

Schneider, C. M., Moreira, A. A., Andrade, J. S. Jr., Havlin, S. and Herrmann, H. J. (2011). Mitigation of malicious attacks on networks. *Proceedings of the National Academy of Sciences* **108** 3838–3841.

Simpson, S. L., Hayasaka, S. and Laurienti, P. J. (2011). Exponential random graph modeling for complex brain networks. *PLoS ONE* **6** e20039.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure* **3**.

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** 99–153.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association* **85** 204–212. MR1137368

Wasserman, S. and Pattison, P. E. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov models and $p*$. *Psychometrika* **61** 401–425. MR1424909

Zhang, J. and Chen, Y. (2013). Sampling for conditional inference on network data. *Journal of the American Statistical Association* **108** 1295–1307.

Jingfei Zhang
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820
USA
E-mail address: zhang197@illinois.edu

Yuguo Chen
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820
USA
E-mail address: yuguo@illinois.edu