The Link Probability Model: A Network Simulation Alternative to the Exponential Random Graph Model

Ian McCulloh*, Joshua Lospinoso* and Kathleen M. Carley

December, 2010
CMU-ISR-10-130

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Center for the Computational Analysis of Social and Organizational Systems
CASOS technical report.

This research is part of the Dynamics Networks project in CASOS (Center for Computational Analysis of Social and Organizational Systems, http://www.casos.cs.cmu.edu) at Carnegie Mellon University.
This work was supported in part by:
• The Army Research Institute for the Behavioral and Social Sciences, Army Project No. 611102B74F
•The Army Research Labs for Assessing C2 structures, Collaborative Technology Alliance,
•Alion Science and Technology,
•ARL Telecordia: Communications and Networks Technology Collaborative Alliances,
•Additional support on measures was provided by the DOD and the NSF IGERT 9972762 in CASOS.
• The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation or the U.S. government.

* Ian McCulloh, School of Information Science, Curtin University of Australia, GPO Box U1987 Perth, Western Australia 6845. Email: cusum6@gmail.com
* Joshua Lospinoso, Department of Statistics, Oxford University, Oxford England. Email: jalospinoso@me.com.

Keywords: Exponential random graph models; *p** models; Statistical models for social networks; Degeneracy; Longitudinal social network analysis.

Abstract

The Link Probability Model (LPM) can be used as an alternative to Exponential Random Graph Models (ERGM) to simulate network data. The LPM characterizes the networks in terms of link probabilities based on historical frequencies. In this paper, the LPM is presented, compared and contrasted with the ERGM. The relative utility of the two approaches is examined by applying both to four longitudinal data sets. The relative strengths and weaknesses of the two approaches in terms of data requirements, scalability, and assumptions are described.

Table of Contents

Introduction

Social networks often exhibit stochastic behavior. For example, an agent in a network might communicate with a friend several times during a given day and not at all during another day. In this example, the underlying relationship remains the same; however, the observed network ties fluctuate. This is an intuitive example, however the accuracy of observed network data has been well documented in the literature (Killworth, et al, 1976, 1979; Bernard, et al, 1977, 1980, 1982; Krackhardt, 1990; Kashy and Kenny, 1990; Wasserman and Faust, 1994). Furthermore, it is possible that the underlying relationships in a social network may change (Carley, 1991; Doreian and Stokman, 1997; Snijders, 2007). This relatively common behavior will also cause fluctuations in observed network data. Therefore statistical models of social networks are necessary for any kind of meaningful inference on network data.

A necessary prerequisite for statistical inference of social networks is an underlying probability structure for the presence of links in the network. Detecting changes over time, comparing multiple networks, or evaluating a wide range of potential hypothesis all depend upon a method to estimate the probability of links occurring in an observed network. Several statistical models have been proposed. The *p** model was introduced by Frank and Strauss (1986). This model describes the distribution of a Markov random graph. Many others have contributed to developing this family of models (Strauss and Ikeda, 1990; Wasserman and Pattison, 1996; Anderson, et al, 1999; Wasserman and Faust, 1994), especially in the area of parameter estimation. A common approach to describe the link probability is the Exponential Random Graph Model (ERGM) (Krackhardt, 1998; Handcock, 2002, 2003; Hunter, 2006; Goodreau, 2007; Robins, et al, 2007; Hunter, et al, 2008). The ERGM is based on a regression of structural variables in the network that may explain the probability of links occurring in the network. Several have used the ERGM to simulate many instances of a given network and then estimate statistical properties of various network measures (Handcock et al, 2006, 2007, 2008; Handcock, 2008). I introduce an alternative approach with the Link Probability Model (LPM) that uses the historical presence of links to estimate the link probability. I demonstrate both simulation approaches on a range of empirical data and show that for a limited number of longitudinal data sets, the LPM provides a better fit to the data than the ERGM.

The ERGM is a family of statistical models that describe the probability of a link being present between two nodes and is a common statistical model for social network analysis. The models are based on logistic regression, where model terms are usually structural variables in the network. The model is used to explore statistically significant properties of networks. The ERGM notation is also flexible, allowing it to represent a wide range of network variables. Unfortunately, many ERGM models are degenerate, meaning that observed data might be highly improbable given the model (Handcock, 2003, 2002). The ERGM is not typically used for over-time network analysis, however Mark Handcock presented an application of the ERGM for simulating networks at the 28th Sunbelt Conference (2008).

The Link Probability Model (LPM) is not a statistical model, but rather a matrix of probabilities of a link being present between ordered pairs of nodes. The LPM is estimated from longitudinal networks based on the frequency of links being present over time. The LPM avoids issues of model degeneracy because the model is not dependent upon highly correlated terms and

there are more data points than parameter estimates. The LPM is particularly useful for our application, because we are only interested in modeling over-time data.

First, I briefly review the ERGM. Then the LPM is described and presented as an alternative model to the ERGM. Then the LPM and ERGM are both used to model four data sets: the Sampson (1969) Monastery data, the Newcomb (1961) Fraternity data, and then two sets of data from Fort Leavenworth (Graham, 2005; Baller, et al, 2008). These are four interesting data sets because they all have a temporal component and have been well documented in the literature. The fit of each of these models is compared to the data. I find that the ERGM is degenerate for the Fort Leavenworth data and that the LPM provides a better fit in the other two data sets under certain conditions. I conclude by discussing the strengths and limitations of LPM and its general usefulness to network analysts.

Exponential Random Graph Model

The ERGM is used in social network analysis as a statistical model that enables an analyst to conduct inference on dependent relational data (Goodreau, 2007; Robins, et. al., 2007). The ERGM is therefore less restrictive than the Holland and Leinhardt (1981) p_1 model that assumed dyadic independence. In many social network applications the relationship between two individuals depends on relationships between the individual and others in the network; cognitive limits on the number of relationships that can be maintained; similarity between individuals; and more. The ERGM framework for relaxing the dyadic independence assumption is thus essential for accurate inference in many data sets.

Exponential random graph models (ERGM) have been studied a great deal in the literature as a model for the probability of links occurring in a social network. The ERGM was first proposed in 1986 (Frank and Strauss) as a very general model. The ERGM can thus be used to model a wide range of explanatory variables. The basic ERGM is given by,

$$P(Y) \propto \theta_1 g_1(y) + \theta_2 g_2 + \ldots + \theta_k g_k(y) \qquad (1)$$

where Y is a graph, θ's are model coefficients, and $g(y)$ is a covariate or term in the model. Covariate terms are general and can represent many features of a graph. These terms are often structural properties of the graph such as the number of links, dyadic relations, and transitive properties, among others.

Estimating ERGM terms and parameters can be computationally challenging in large networks (Snijders, 2002; Pattison and Robins, 2002). Markov chain Monte Carlo estimation of ERGM has been used to fit these models to data (Goodreau, 2007; Robins, et. al., 2007; Handcock, 2003, 2002; Snijders, 2002; Pattison and Robins, 2002). The Markov dependence in these models leads to problems of degeneracy, which is discussed in detail by Handcock (2003, 2002). Essentially, model degeneracy occurs when the observed data is almost impossible under the specified model. This often occurs when explanatory terms are highly correlated and there is insufficient data to construct an appropriate model. Many of the terms used in ERGM are

correlated and it is difficult to define enough terms to preclude networks that do not represent the data, when they spuriously satisfy the ERGM terms. Several advances in ERGM have been proposed to include curved exponential family models (Hunter and Handcock, 2006) and neighborhood models (Pattison and Robins, 2004). However, these advances have not completely removed issues of model degeneracy.

Link Probability Model Formulation

The LPM framework for viewing the probability space of a social network avoids issues of model degeneracy, while preserving flexibility for modeling dyadic relationships. It provides researchers with an improved means to understand the probability space of the network, under certain conditions. The LPM is a square matrix where the rows and columns correspond to the nodes in a social network. The entries are the link probabilities of the directed link from the row node to the column node. This is not to be confused with an adjacency matrix, where the entries are either zero or some number representing the strength of a relationship between nodes. The link probability is a number between 0 and 1, and determines the likelihood of a link being present in an observed adjacency matrix.

The link probabilities can be derived from empirical data in several ways. Given network data collected over multiple time periods on a group of subjects, the link probabilities can be estimated by the proportion of link occurrences, l_{ij}, for each cell in the adjacency matrix, $a_{ij.}$ In the case of communication networks, statistical distributions can be fit to the time between messages for each potential link in the network. For a specified period of time, t, the link probability p for each set of entities i and j can be found. Let x_{ij} be the time between messages in a communication network. The probability density function for any x can then be defined as $f_{ij}(x \mid \theta_{ij})$, where θ_{ij} is the set of parameters for the density function. Then, the probability, p, of a link occurring within some time period t is the probability that $x < t$, which can be expressed as,

$$p = \int_0^t f_{ij}(x \mid \theta_{ij})dx \tag{2}$$

In practice, the function $f_{ij}(x \mid \theta_{ij})$ must be estimated using techniques such as maximum likelihood estimation from empirical data collected on the group being studied. It may be desirable to construct a network based on a restriction such as, "two emails within a time period demonstrate a relationship, but one does not." In this case, it is necessary to compose a function of random variables. If $h_{ij}(2 \mid t, \theta_{ij})$ represents the probability density function of time between two sets of two emails and $f_{ij}(x \mid \theta_{ij})$ represents the probability density function of time between one set of two emails, then the following is true under certain assumptions:

$$h_{ij}(2 \mid \theta_{ij}) = \left(\int_0^t f_{ij}(x \mid \theta_{ij})dx \right)^2 \tag{3}$$

It is possible to generalize this idea; if h_{ij} (x | t, θ_{ij}) is the probability that x or more communications occur within time t, then the following is true:

$$h_{ij}(x \mid \theta_{ij}, t) = \left(\int_0^t f_{ij}(y \mid \theta_{ij}) dy \right)^x \quad (4)$$

The LPM is an important improvement over some traditional models. Individuals in a social network are not connected to other individuals with uniform random probability. The probability structure is much more complex. Intuitively, there are some people whom a person will communicate with or be connected more closely than others. In a study of email communication conducted at the U.S. Military Academy (McCulloh et al, 2007) one subject emailed his wife more than ten times per day on average, while other people that he worked with received an email from him once or twice per month. For this reason, real-world networks tend to have clusters or cliques of nodes that are more closely related than others (Newman, 2003; Topper and Carley, 1999; Carley, 1996). This can be simulated by varying the probabilities that certain nodes will communicate. In this way, stochastic behavior in dynamic social networks can realistically be simulated.

The LPM is a desirable model due to its ability to accurately model empirical data and its ability to avoid degeneracy. The accuracy of the LPM will be discussed in the Results section. The LPM can avoid issues of model degeneracy because the only parameters for the model are the link probabilities. As long as there are at least two time periods for estimating parameters, there are more data points than there are parameters. Each link is treated independently of other links in the model; therefore, none of the terms are correlated. The naïve assumption of independence between links is corrected by the historic presence of links over time. Intuitively, links have some dependence. For example, if an individual chooses to communicate with another, the likelihood of that person reciprocating the communication increases. If we assume a dynamic equilibrium in the underlying relationships of individuals in the network, these patterns of dependent communication will be apparent over time. If node *i* has a high link probability with node *j*, it may be likely that node *j* has a reciprocal high link probability with node *i*. It is not necessary to directly account for this in the model. If the relationship is true, there will be a high expected occurrence of *i* to *j* and *j* to *i* links in the networks over time. The LPM will model these links with high link probability due to their over time frequency, and not directly from their structural dependency. In this way, the LPM can never be over specified, have high variance inflation, or be degenerate. Thus, the LPM may provide an attractive alternative to the ERGM for modeling longitudinal degenerate networks.

Data for Comparison

Four data sets are used to demonstrate the efficacy of the LPM. The first and second are longitudinal data sets that are well established in the SNA literature, namely the Sampson (1969) Monastery data and the Newcomb (1961) Fraternity data. The third and fourth data sets are larger in size. For the reader's convenience, Table 1 summarizes the similarity and difference among the data sets. All four are explained in more detail.

Table 1. Data Summary.

Name of data set	Monastery	Fraternity	Leavenworth '05	Leavenworth '07
Author	Sampson	Newcomb	Graham	Schrieber
Number of nodes	18	17	156	68
No. of time periods	3	15	8	9
Method of collection	Observation	Survey	Survey& Observation	Survey
Link weight	Dichotomous	Weighted	Dichotomous	Dichotomous
Link Relationship	Interpersonal relationship	Preference ranking	Self Reported Communication	Self Reported Communication
Change in density	0.17974-0.18301	0.50000-0.50000	0.01431-0.02906	0.04473-0.04628
Change in average betweenness	0.05556-0.05556	0.33574-0.41176	0.00880-0.00994	0.02009-0.01909
Change in average closeness	0.40158-0.02485	0.66510-0.39859	0.03759-0.05172	0.05739-0.08186
Change in average eigenvector cent	0.23428-0.23247	0.79907-0.74891	0.23591-0.22963	0.2125-0.22243

The first data set was collected in a monastery by Samuel F. Sampson (1969). The participants included 18 monks, and data was recorded on their interpersonal relationships. This is a directed network, where relationships are not necessarily reciprocal. Data was collected over three time periods, representing the time in which a new cohort joined the monastery.

The second data set was collected by Theodore Newcomb (1961) at the University of Michigan. The participants included 17 incoming transfer students, with no prior acquaintance, who were housed together in fraternity housing. The participants were asked to rank their preference of individuals in the house from 1 to 16, where 1 is their first choice. Data was collected each week for 15 weeks, except for week number nine. The relational data recorded between agents were ranks. Both the ERGM and LPM require dichotomous networks to construct a model. I chose to adopt the binarization scheme proposed by David Krackhardt (1998). He dichotomized the network data by assigning a link to preference ratings of 1-8 and having no link for ratings of 9-16. Krackhardt also fit an ERGM to the Newcomb Fraternity data which will be used for comparison with the LPM.

The third data set was collected from an Army war fighting simulation at Fort Leavenworth, Kansas in 2005, by Craig Schreiber and Lieutenant Colonel John Graham. The participants were mid-career U.S. Army officers taking part in a brigade level staff training exercise. This data set contains 156 individual agents that were monitored over the course of four and a half days. Data consists of communication ties between individuals as measured from self reported communications surveys. Surveys were completed at the end of each morning and at the end of the day before the officers went home. Therefore there are nine longitudinal time periods.

The fourth data set was also collected from an Army war fighting simulation at Fort Leavenworth, Kansas by Craig Schreiber; this time in April, 2007. There were 68 participants in

this data set, who served as staff members in the headquarters of the brigade conducting a simulated training exercise. The data contains the communication between agents in the network which were collected through self reported communications surveys. Data was collected over a period of four days, twice per day. Thus, there were eight time periods.

Method of Comparison

The ERGM and LPM are investigated for their strengths and weakness in modeling longitudinal data. For the Sampson (1969) Monastery data, I use the ERGM that was fit to the data by Hunter et al (2008). The Akaike Information Criterion (AIC) is 302.61 and the Bayesian Information Criterion (BIC) is 436.65. The Hunter (2008) ERGM of the Sampson (1969) data was chosen for this study based on its more favorable AIC and BIC compared to other models found in the literature. I feel that this model is therefore an appropriate benchmark for comparison with the LPM. An ERGM is also fit to the Newcomb (1961) fraternity data. Again, I have chosen an ERGM accepted in the literature; this time the model proposed by Krackhardt (1998). An LPM is fit to both the Sampson and Newcomb data sets. Monte Carlo simulation is used to generate instances of the Sampson Monastery social network and the Newcomb Fraternity social network under the ERGM and LPM. In addition, an LPM is also fit to the two Fort Leavenworth data sets (Graham, 2005; Baller, et. al., 2008). For the two Fort Leavenworth data sets, the ERGM was degenerate. The ERGM were not degenerate for the Sampson or Newcomb data sets. The LPM is successfully used to model all data sets.

A distance measure is required to compare the similarity between the dichotomous networks generated using the ERGM, the LPM, and the empirical data. Hamming distance (1950) evaluates a distance between dichotomous networks. If the data were weighted networks and the models generated weighted networks as well, then a Euclidean distance would be appropriate. The quadratic assignment procedure (QAP) (Krackhardt, 1987b) could be used to compare the correlation between networks; however, we focus on network distance, because we intend to demonstrate that the LPM can generate simulated models that are very similar to the original networks in terms of actual distance and not simply a structural isomorphism.

The ERGM and LPM are evaluated on how well they model empirical data using a t-test. I illustrate the method with the Sampson Monastery data. Let the three networks in the Monastery data be labeled N1, N2, and N3. An ERGM is used to simulate networks and they are labeled E1, E2, E3, … E100,000. The LPM is also used to simulate networks and they are Labeled L1, L2, L3, … L100,000. The Hamming distances are calculated between each empirical data set to every simulated ERGM network and I use the following notation,

$$\text{Dist}_{\text{ERGM},1,1} = \text{Hamming}(\text{N1},\text{E1})$$
$$\text{Dist}_{\text{ERGM},1,2} = \text{Hamming}(\text{N1},\text{E2})$$
$$\dots$$
$$\text{Dist}_{\text{ERGM},i,j} = \text{Hamming}(\text{N}i,\text{E}j)$$
$$\dots$$
$$\text{Dist}_{\text{ERGM},3,100000} = \text{Hamming}(\text{N3},\text{E100000}).$$

The Hamming distances are also calculated between each empirical data set and every simulated LPM network and its notation is given by,

$$\text{Dist}_{\text{LPM},i,j} = \text{Hamming}(\text{N}i,\text{L}j).$$

The Hamming distances are calculated between each empirical data set and every other empirical data set and its notation is given by,

$$\text{Dist}_{\text{empirical},i,j} = \text{Hamming}(\text{N}i,\text{N}j), \text{ where } i \neq j.$$

This last set of Hamming distances are a measure of noise or observation error inherent in the data.

The ERGM and LPM are compared using a two-sample T-test between the Hamming distances from the empirical network, N*i*, and all of the simulated networks from the ERGM and the LPM. The test statistic is given by,

$$T_i = \frac{\mu_{ERGM,i} - \mu_{LPM,i}}{S_{P,i}/\sqrt{100{,}000}}$$

where,

$$\mu_{ERGM,i} = \frac{1}{100{,}000}\sum_j Dist_{ERGM,i,j}$$

$$\mu_{LPM,i} = \frac{1}{100{,}000}\sum_j Dist_{LPM,i,j}$$

and $S_{P,i}$ is the pooled standard deviation between the ERGM and LPM Hamming distances (Montgomery, 1991). This is repeated for each time period, *i*.

Results

An ERGM was fit to the Sampson (1969) Monastery data according to the model specification laid out by Hunter, et. al. (2008). Four model terms were used: links, sender, receiver, and mutual. A summary of the model fit is shown in Table 2.

Table 2. Fit Summary for Sampson ERGM.

Model Parameter	Coefficient	Standard Error	MCMC S.E.	p-value
Links	-2.5131	0.3361	0.005	0.0000
sender2	-0.7356	0.6854	0.015	0.2842
sender3	-0.2146	0.7274	0.017	0.7682
… output edited for length …				
receiver17	-1.2015	0.8191	0.018	0.1436
receiver18	-1.0562	0.7193	0.015	0.1432
Mutual	3.6816	0.6731	0.011	0.0000

The Hamming distance from each of the three empirical data sets to each of the ERGM simulated networks was calculated. The Hamming distance from each of the empirical data sets to each of the LPM simulated networks was calculated. The mean and standard deviation of these Hamming distances are displayed in Table 3. A two-sample t-test for each time period illustrates that the networks simulated using the LPM have a smaller average hamming distance to the empirical data sets than the networks simulated using the ERGM. This indicates that the LPM models the Sampson data more accurately than the ERGM model.

Table 3. Sampson Data Hamming Distances and T-test for ERGM and LPM.

Time Period	$\mu_{ERGM,i}$	ERGM Hamming Distance Standard Deviation	$\mu_{LPM,i}$	LPM Hamming Distance Standard Deviation	T_i t-test	p-value
1	98.70	5.6970	27.67	3.5922	39.43	0.0006
2	99.10	6.2263	24.99	3.5935	37.64	0.0007
3	103.70	6.2902	24.66	3.5945	39.74	0.0006

The Newcomb (1961) Fraternity data was also fit with an ERGM. Three model terms were used: mutual, Simmelian ties, and balance. A summary of the model fit is shown in Table 4. The AIC is 308.93 and the BIC is 319.75, which are more favorable than similar variations of the ERGM.

Table 4. Fit Summary for Newcomb ERGM.

Model Parameter	Coefficient	Standard Error	MCMC S.E.	p-value
Mutual	-1.5745	0.2304	0.0070	0.0000
Simmelian Ties	0.6581	0.0006	0.0001	0.0000
Balance	0.2333	0.0364	0.0010	0.0000

The Hamming distances from each of the fourteen empirical data sets to each of the ERGM simulated networks and each of the LPM simulated networks were calculated. The mean and standard deviation of these Hamming distances are displayed in Table 5. A two-sample t-test for each empirical data set illustrates that the networks simulated using the LPM have a smaller average hamming distance to the empirical data sets than the networks simulated using the ERGM. This indicates that the LPM models the Newcomb fraternity data more accurately than the ERGM model.

Table 5. Newcomb Data Hamming Distances and T-test for ERGM and LPM.

Time Period	$\mu_{ERGM,i}$	ERGM Hamming Distance Standard Deviation	$\mu_{LPM,i}$	LPM Hamming Distance Standard Deviation	T_i t-test	p-value
1	139.7	8.3938	91.9	5.1913	18.0147	0.0353
2	138.9	8.1847	75.1	5.2128	24.6573	0.0258
3	137.3	8.2872	48.3	5.2226	33.9732	0.0187
4	135.5	9.3363	49.7	5.2340	29.0460	0.0219
5	134.1	8.9870	50.1	5.2319	29.5558	0.0215
6	136.3	8.5251	45.5	5.2440	33.6983	0.0189
7	133.9	9.0609	47.3	5.2397	30.2202	0.0211
8	134.1	7.2946	51.9	5.2591	35.6377	0.0179
10	133.7	5.1865	64.2	5.2223	42.3990	0.0000
11	132.7	6.0562	53.4	5.2074	41.4119	0.0006
12	136.3	8.4466	51.1	5.2147	31.8930	0.0200
13	134.9	9.0117	46.6	5.2311	30.9989	0.0205
14	133.9	5.4457	46.1	5.2230	50.9574	0.0000
15	133.1	5.7242	47.2	5.2378	47.4518	0.0004

The LPM is further investigated using the Fort Leavenworth data. ERGM's with only a single term were found to be degenerate for several common parameter choices; therefore, they are not included in the analysis of this section. For both of the Fort Leavenworth data sets, the Hamming distance between the simulated LPM networks and each empirical network, $Dist_{LPM,i,j}$ = Hamming(Ni,Lj), was compared to the Hamming distance between each empirical network to the other empirical networks within the data set, $Dist_{empirical,i,j}$ = Hamming(Ni,Nj), where $i \neq j$. Two-sample t-tests were used to determine if there was a significant difference in mean Hamming distance between the empirical networks and the LPM. The t-tests were properly adjusted for heteroscedasticity and unequal sample sizes. Table 6 displays the Hamming distances and the results of the two-sample t-tests for the 2005 Fort Leavenworth data, and Table 7 displays this information for the 2007 Fort Leavenworth data. In all cases the Hamming distance is less for the LPM. The low p-values show a statistically significant difference in mean

Hamming distance of the empirical to empirical comparison versus the LPM to empirical comparison. Additionally, since $\mu_{emperical,i} - \mu_{LPM,i} > 0$ it is shown that the simulated LPM networks have, on average, less Hamming distance from each of the empirical data sets than the empirical data sets have from each other. This means that networks generated using the LPM are closer to the original data than the observed empirical networks are to each other. While the t-tests for 2005 Fort Leavenworth time periods 6, 8, and 9 are only marginally significant, they have the same positive trend as the other 14 empirical networks in the 2005 and 2007 data sets.

Table 6. 2005 Fort Leavenworth Data Hamming Distances and T-test for LPM.

Time Period	$\mu_{empirical,i}$	Empirical Hamming Distance Standard Deviation	$\mu_{LPM,i}$	LPM Hamming Distance Standard Deviation	T_i t-test	p-value
1	1445.000	84.774	1284.338	23.747	3.467	0.001
2	1394.750	67.487	1239.647	23.703	3.765	0.000
3	1296.125	85.436	1151.946	23.671	3.287	0.001
4	1315.875	153.533	1169.665	23.718	2.421	0.015
5	1191.250	112.324	1058.990	23.667	2.732	0.006
6	1204.875	207.944	1071.116	23.623	1.912	0.056
7	1167.375	190.431	1037.713	23.695	1.980	0.048
8	1159.625	204.465	1030.815	23.732	1.888	0.059
9	1170.125	195.266	1040.142	23.618	1.953	0.051

Table 7. 2007 Fort Leavenworth Data Hamming Distances and T-test for LPM.

Time Period	$\mu_{empirical,i}$	Empirical Hamming Distance Standard Deviation	$\mu_{LPM,i}$	LPM Hamming Distance Standard Deviation	T_i t-test	p-value
1	409.286	38.560	358.094	12.775	3.755	0.00
2	365.857	18.298	320.097	12.739	7.073	0.00
3	365.857	29.043	320.164	12.793	4.450	0.00
4	377.857	38.247	330.674	12.773	3.489	0.00
5	375.286	36.100	328.377	12.796	3.675	0.00
6	349.857	38.159	306.078	12.785	3.245	0.00
7	373.857	48.451	327.073	12.826	2.731	0.01
8	362.429	55.635	317.151	12.775	2.302	0.02

Discussion

The LPM has been used to model longitudinal social network data for four different data sets. In those data sets, the LPM generates simulated networks that are more like the original data than networks generated using the ERGM. In addition, it is generally the case that the networks generated using the LPM are more similar to the original data than any prior time period. The LPM avoids issues of model degeneracy due to its formulation. The probability of link occurrence is based on the historic presence of links and does not use a Markov assumption or over specify a statistical model. For these reasons, the LPM provides an alternative method for modeling and conducting longitudinal social network analysis.

Monte Carlo simulations can be generated using the LPM. Each cell, $a_{ij,}$, in the LPM can be compared to a uniform (0,1) random variable to determine the presence of a link in a simulated adjacency matrix. As demonstrated earlier, these simulated adjacency matrices are very similar to the empirical data as demonstrated by the low Hamming distance between simulated networks and empirical networks. Statistical distributions can then be fit to any social network measures calculated on the simulated networks. These statistical distributions can then be used for inference using traditional statistical methods.

The LPM cannot be used in place of the ERGM in all situations, however. Multiple networks are required to estimate the LPM for a given empirical data set. The ERGM on the other hand, can be estimated from a single observed network. The approach to adding and removing nodes is different for the ERGM and LPM. For the LPM, a missing node would be included in the model with a 0 recorded for all column and row entries of the missing node. Finally, the LPM is formulated based on the assumption that there are fixed probability structures under-laying social networks that do not change significantly over time. The observed social networks based on the LPM will fluctuate between time periods, but the general patterns of connections remain the same. Table 8 illustrates some differences and similarities between LPM and ERGM data requirements.

Table 8. Comparison of LPM and ERGM.

Data characteristics	LPM	ERGM
Link weighting	Dichotomous	Dichotomous
Number of links	No limit	Probability of degeneracy increases with number of links
Min. no. time period	2	1
Practical no. time period	5+	1
Assumed cause of stochasticity	Dynamic equilibrium	Evolves due to structural properties of the network.

The LPM has several advantages over the ERGM for longitudinal social network analysis; however the ERGM has advantages over the LPM for other types of analysis. Table 9

displays advantages and disadvantages of the LPM and ERGM models. The LPM requires multiple observed networks to estimate model parameters, where the ERGM can be estimated using a single observed network. At a minimum, two observed networks are required to estimate an LPM, however, in practice; the variance of the estimate is proportionate to $1/\sqrt{n}$, where n is the number of observed networks. We nominate five observed networks as a rule of thumb for fitting the LPM as most of the estimate variance is eliminated with this number. The LPM is more computationally efficient than the ERGM. The number of link probabilities for a network is quadratic with the number of nodes. The LPM estimates are then linear with the number of observed networks. The ERGM parameter estimates can be n^n with number of nodes for each term. Heuristics are often used to estimate ERGM model parameters. In addition, the ERGM has problems with model degeneracy as previously discussed. The LPM has been shown to provide a model that can be used to simulate data that is more similar to empirical data than data generated with ERGM simulations. An additional benefit for the LPM is the ability to use link probabilities as dependent variables in regression models for homophily. Homophily is an expression to describe the similarity between individuals in terms of certain attributes that the individuals have. In more complex models, the parameters of link probability densities can serve as dependent variables in homophily regression. Unfortunately, the LPM does not provide any explanation of likely structural causes for the stochastic behavior of networks. Significant terms in an ERGM can be interpreted as the underlying mechanism for network evolution over time. It

Table 9. Advantages and Disadvantages of LPM and ERGM.

Considerations	LPM	ERGM
Required no. of observed networks	***Disadvantage***: The LPM requires multiple observed networks to estimate the link probability of a network based on historic frequency of occurrence.	***Advantage***: The ERGM requires only a single network
Computational efficiency	***Advantage***: The computational speed is quadratic with the number of nodes in the network.	***Disadvantage***: The computational speed is n^n which requires heuristic approximations of model parameters.
Model quality	***Advantage***: Stable and consistent model estimates.	***Disadvantage***: Prone to degenerate models.
Accuracy to real data	***Advantage***: Shown to more closely resemble empirical data as measured by Hamming distance.	***Disadvantage***: Has not been shown to consistently model empirical data accurately as measured by Hamming distance.
Explanation of social dynamics	***Disadvantage***: Does not attempt to explain underlying social dynamics of the group or organization.	***Advantage***: Model terms can be interpreted as underlying mechanisms for social dynamics within the modeled group or organization.

may be possible to develop similar explanations of behavior through future research in homophily regression using the LPM. Further research is needed on both the ERGM and the LPM to illuminate strengths and limitations. In the interim, there is strong evidence to suggest the use of the LPM whenever degeneracy is a problem among ERGM's, or when the goal is to estimate the normal behavior of a social group that is in dynamic equilibrium.

Another important area for future research is network periodicity. Intuitively, social networks are subject to periodic trends. An average person's communication patterns may be different during the week, while they are at work, than during the weekend, when they are at home with their family. Future research will hopefully expand both the LPM and ERGM to handle periodic trends in longitudinal data. It will be interesting to compare the performance of the LPM and ERGM for modeling time dependent longitudinal social network data sets.

This paper has introduced the Link Probability Model (LPM) for longitudinal social network analysis. The primary strength of the LPM is its ability to accurately model longitudinal network behavior with better goodness of fit than competing models. The LPM also avoids issues of model degeneracy due to the method of its construction. Finally, the LPM is more computationally efficient than the ERGM for both estimation and simulation. Using the LPM, accurate simulation of longitudinal social network data can be performed. This opens the door for researchers to explore an entirely new approach for inference on social networks.

References

Anderson, C.J., Wasserman, S., Crouch, B., 1999. A *p** primer: logit models for social networks. *Social Networks* 21, 37–66.

Baller, D., Lospinoso, J., and Johnson, A.N. 2008. An Empirical Method for the Evaluation of Dynamic Network Simulation Methods. In Proceedings, The 2008 World Congress in Computer Science Computer Engineering and Applied Computing, Las Vegas, NV.

Bernard, H. R., & Killworth, P. D. 1977. Informant accuracy in social network data II. *Human Communication Research, 4,* 3-18.

Bernard, H. R., Killworth, P. D., & Sailer, L. 1980. Informant accuracy n social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks, 2*, 191-218.

Bernard, H. R., Killworth, P. D., & Sailer, L. 1982. Informant accuracy in social network data V: An experimental attempt to predict actual communication from recall data. *Social Science Research, 11*, 30-66.

Carley, K.M. 1991. A Theory of Group Stability. *American Sociological Review*, 56(3): 331-354.

Doreian, P. & Stokman, F. 1997. *Evolution of social networks*, London: Gordon & Beach.

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

Graham, J. 2005. *Dynamic Network Analysis of the Network-Centric Organization: Towards an Understanding of Cognition & Performance*. Ph.D. Thesis, Carnegie Mellon University.

Goodreau, S.M., 2007. Advances in exponential random graph (*p**) models applied to a large social network. *Social Networks*, 29, 231-248.

Goodreau, S.M., Hunter, D.R., and Morris, M., 2005. Statistical Modeling of Social Networks: Practical Advances and Results. Center for Studies in Demography and Ecology, University of Washington, Working Paper No. 05-01.

Hamming, R.W. 1950. Error Detecting and Error Correcting Codes, *Bell System Technical Journal* 26(2):147-160

Handcock, M.S., 2008. *Exponential random graph (ERG or p*) models*. Workshop given at Sunbelt XXXVIII Conference, January 24-28, 2008, St Petersburg, Florida

Handcock, M.S., 2003. Statistical models for social networks: degeneracy and inference. In: Breiger, R., Carley, K., Pattison, P. (Eds.), Dynamic Social Network Modeling and Analysis. National Academies Press, Washington, DC, pp. 229–240.

Handcock, M.S., 2002. Statistical models for social networks:degeneracy and inference. In: Breiger, R., Carley, K., Pattison, P. (Eds.), Dynamic Social Network Modeling and Analysis. National Academies Press, Washington, DC, pp 229-240.

Handcock, M., Hunter, D., Butts, C., Goodreau, S., Morris, M., 2008. ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, 24(3): 1-29.

Handcock, M., Hunter, D., Butts, C., Goodreau, S., Morris, M., 2006. Statnet: An R Package for the Statistical Analysis and Simulation of Social Networks. Manual. University of Washington, http://www.csde.washington.edu/statnet.

Handcock, M., Hunter, D., Butts, C., Goodreau, S., Morris, M., 2007. statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *Journal of Statistical Software*, 24(1): 1548–7660.

Holland, P.W., Leinhardt, S., 1981. An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association* 76, 33–65.

Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M., 2008. ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, 24:3, 1-29.

Hunter, D., 2006. Curved Exponential Family Models for Social Networks. *Social Networks*, doi:10.1016/j.socnet.2006.08.005.

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

Kashy, D. A., & Kenny, D. A. 1990. Do you know whom you were with a week ago Friday? A re-analysis of the Bernard, Killworth, and Sailer studies. *Social Psychology Quarterly, 53*, 55-61.

Killworth, P. D. & Bernard, H. R. 1976. Informant accuracy in social network data. *Human Organization, 35*, 269-96.

Killworth, P. D. & Bernard, H. R. 1979. Informant accuracy in social network data III: A comparison of triadic structure in behavioral and cognitive data. *Social Networks*, 2, 19-46.

Krackhardt, D. 1990. Assessing the Political Landscape: Structure, Cognition, and Power in Organizations. *Administrative Science Quarterly*, 35: 342–369

Krackhardt, D. 1998 Simmelian ties: Super strong and sticky. In Power and Influence in Organizations (eds R. Kramer, M. Neale), pp, 21-38. Sage, Thousand Oaks, CA.

McCulloh, I., Garcia, G., Tardieu, K., MacGibon, J., Dye, H., Moores, K., Graham, J. M., and Horn, D. B., 2007. IkeNet: Social network analysis of e-mail traffic in the Eisenhower Leadership Development Program. (Technical Report, No. 1218), U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington, VA.

McCulloh, I., Lospinoso, J., and Carley, K.M., 2007. Social network probability mechanics. In Proceedings, 12th International Conference on Applied Mathematics, World Scientific Engineering Academy and Society, Cairo, Egypt, pp 319-325.

Newcomb, T.N. 1961. *The Acquaintance Process*. Holt, Rinehart and Winston, New York .

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

Pattison, P.E., Robins, G.L., 2004. Building models for social space: neighborhood based models for social networks and affiliation structures. *Mathematiques des Science Humaines* 168, 11–29.

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

Sampson, S.F., 1969. Crisis in a cloister. Ph.D. Thesis. Cornell University, Ithaca.

Snijders, T.A.B., 2002. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure* 3, 2.

Snijders, T.A.B., 2007. Models for Longitudinal Data. In: Carrington, P., Scott, J., Wasserman, S. (Eds.), Models and Methods in Social Network Analysis. Cambridge University Press, NewYork, pp.215-247.

Strauss, D., Ikeda, M., 1990. Pseudo-likelihood estimation for social networks. *Journal of the American Statistical Association* 85, 204–212.

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

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

Wasserman, S., Robins, G.L., 2007. An introduction to random graphs, dependence graphs, and *p**. In: Carrington, P., Scott, J., Wasserman, S. (Eds.), Models and Methods in Social Network Analysis. Cambridge University Press, NewYork, pp. 148–161.