Arid Lands Resource Sciences

ARL 930

Applicability of ERG Models (P*) to a hypothetical case of Irrigation System users: testing socio economic status homophily

Luis Alan Navarro Navarro[1,*]

[1] *Arid Lands Resource Sciences Graduate Multidisciplinary Program Doctoral Student, University of Arizona*

* *E-mail: alann@email.arizona.edu*

Date: June 9th 2010

ABSTRACT

This paper represents an exploratory research analysis intended to find whether an Exponential Random Graph (ERG) model fits a hypothetical social network (HSN). This HSN forms a graph "G" which mathematically is represented by two subsets $G\{N, E\}$ where "N" is the number of nodes (actors) $N = \{1, 2, 3, \dots, n\}$ and $E = \{1, 2, 3, \dots, l\}$ is the number of directed ties (arcs). The adjacency matrix (or Sociomatrix 1) of "G" has binary [0,1] off-diagonal entries, where $x_{ij} = 1$ if there is a tie $i \rightarrow j$ from actor "i" to actor "j", and 0 otherwise. Thus, the HSN is a 20x20 non-symmetric sociomatrix, with row marginals = 5 ∀ "i", and main diagonal $(x_{ii} = 0)$ not defined but set to 0 by convention. Sociomatrix 1 was manually created resembling egocentric data, where actor (farmer) "i" (called "ego") was asked to name five other farmers (usually called "alters") with whom s/he would be willing to collaborate and partner up for a hypothetical (bu possible) project related with the improvement and maintenance of the irrigation system's infrastructure. Framers were randomly assigned into one out of three categories of an ordinal variable which represents farmer's socioeconomic status (SES), coded [1,2,3] from low to high SES, then ties were formed to intentionally match farmer's SES crating thus homophily ties. Sociomatrix 1 was once randomly permuted to form Sociomatrix 2. We test on both Sociomatrices a single null hypothesis Ho: Irrigation System (IS) users' willingness to cooperate is less likely to occur within individuals sharing a similar SES (no SES heterophily); Ha: IS user's willingness to cooperate is more likely to occur within individuals sharing a similar SES (SES homophily).
The goodness of fit of the data to the ERG model proposed provided insights for the analytic potentials of a future empirical social network derived from case studies.

| Irrigation | Commons | Social Networks | ERGM|

1. Introduction

Statistical exponential family models (Wasserman and Pattison 1996) are a generalization of the Markov random graph models introduced by Frank and Strauss (1986), which in turn are derived from developments in spatial statistics (Besag 1974) [taken from: Handcock (2003)].

ERG modeling is a statistical technique for modeling structural properties of networks (Snijders et al., 2006).

The importance of statistical network analysis resides in the fact that whatever we conclude from a social network we need to evaluate its significance. A statistical model allows a researcher to perform comparisons of the observed effects to hypothesized effects, as well as significance tests to determine whether an effect is due

to sampling variability (Wasserman and Faust, 1994). The model should provide theoretically plausible interpretations about the type of effects that might have produced the network

2. Software

ERG Models of social network structure are fit using R package statnet. Statnet is a set of statistical network analysis routines in the R environment (Handcock et al, 2004).

3. Model specification: cross-sectional models for ERG

There are different authors which mathematically specify the ERG models in different ways here we try to adapt and unify some of them (Knoke and Yang (2008); Pattison and Robin (2008); Goodreau et al. (2009)). In ERG models there are a fixed number of actors and the possible arcs among nodes of a network are regarded as random variables (Robins et al. 2007).

The range of substantially motivated network statistics that might be included in the model is vast. Our choice of observable variables will be determined by what we want to predict or learn about, what model's goodness-to-fit we want, how powerful our computers are, and what is possible to measure. Here we consider only a few key statistics (Hunter et al. 2008). We have a uniform distribution conditional on 2 network statistics (structural configurations) and one node covariates:

$$\mathbf{U}|\mathbf{Z}_{\boldsymbol{\theta}}, \mathbf{Z}_{\boldsymbol{\rho}}, \mathbf{Z}_{\boldsymbol{\theta w}}$$

Where:

$\mathbf{Z}_{\boldsymbol{\theta e}} = \sum \boldsymbol{x_{ij}}$ is the choice (total number of arcs) explanatory variable.

$\mathbf{Z}_{\boldsymbol{\rho}} = \sum_{i<j} \boldsymbol{x_{ij} x_{ji}}$ is the mutuality variable (number of mutual ties).

$\mathbf{Z}_{\boldsymbol{\theta w}} = \sum \boldsymbol{x_{ij} \delta_{ij}}$ is the mutuality variable within a type.

Let's consider $\theta = \{\theta e, \rho, \theta w\}$ a vector of "k" parameters to be estimated, δ_{ij} is a binary indicator that equals 1 if i and j are the same type and 0 otherwise. The binary random variable x_{ij} indicates whether there exists a tie between nodes i and j $(x_{ij} = 1)$ or not $(x_{ij} = 0)$.

$Z(x_{ij}^{+})$ and $Z(x_{ij}^{-})$ are vectors of raw data for various independent or explanatory variables, the superscript + or – sign indicates whether a tie is present or not. The superscript "c" stands for complement, x_{ij}^{C} denotes the status of all dyads in X except x_{ij}, so that x_{ij}^{C} is in a sense "everything else"(do not confuse it with the normalizing constant "c").

$$\frac{\mathbf{Pr}\left(\mathbf{x_{ij}} = \mathbf{1}|\mathbf{x_{ij}^{C}}\right)}{\mathbf{Pr}\left(\mathbf{x_{ij}} = \mathbf{0}|\mathbf{x_{ij}^{C}}\right)} = \frac{\mathbf{exp}\{\boldsymbol{\theta_k}\}\boldsymbol{Z_k}(\boldsymbol{x_{ij}^{+}})}{\mathbf{exp}\{\boldsymbol{\theta_k}\}\boldsymbol{Z_k}(\boldsymbol{x_{ij}^{-}})} = \mathbf{exp}\{\boldsymbol{\theta_k}[\boldsymbol{Z_k}(\boldsymbol{x_{ij}^{+}}) - \boldsymbol{Z_k}(\boldsymbol{x_{ij}^{-}})]\}$$

Taking the natural log of both sides:

$$\mathbf{log}\left[\frac{\mathbf{Pr}\left(\mathbf{x_{ij}} = \mathbf{1}|\mathbf{x_{ij}^{C}}\right)}{\mathbf{Pr}\left(\mathbf{x_{ij}} = \mathbf{0}|\mathbf{x_{ij}^{C}}\right)}\right] = \boldsymbol{\theta_k}[\boldsymbol{Z_k}(\boldsymbol{x_{ij}^{+}}) - \boldsymbol{Z_k}(\boldsymbol{x_{ij}^{-}})]$$

The values of the independent variables (graph statistics) are the differences of the aggregated values of those variables of the entire graph when a given tie from i to j forced to be changed from being present to being absent. Thus:

Change in choice: $\varphi Z_{\theta e} = \sum x_{ij}^{+} - \sum x_{ij}^{-}$

Change in choice within a type: $\varphi Z_{\theta w} = \sum x_{ij}^{+} \boldsymbol{\delta_{ij}} - \sum x_{ij}^{-} \boldsymbol{\delta_{ij}}$

Change in mutuality:$\varphi Z_{\rho} = \sum_{i<j}^{+} x_{ij} x_{ji} - \sum_{i<j}^{-} x_{ij} x_{ji}$

For an example of how these statistics are calculated see Knoke and Yang (2008).

Goodreau et al. (2009) specify the model as:

$$\mathbf{P}(\mathbf{X} = \mathbf{x}) = \left(\mathbf{1}/\mathbf{c}\right) \boldsymbol{exp}\left(\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{k}} \boldsymbol{\theta_k Z_k}(\boldsymbol{X})\right)$$

Where:

X is a random network.

Z_k model's covariates, this term can be expanded as:

$Z_k\{x, W\}$, W = where the ijk[th] element is the covariate of the k[th] attribute of the ij[th] dyad; and "x" = attributes specific to the individuals.

$\theta = \{\theta e, \rho, \theta w\}$ a vector of "k" unknown parameters to be estimated.

The denominator "c" represents the quantity from the numerator summed over all possible networks of order "n", constraining the probability to sum 1.

$$\mathbf{log}\left[\frac{\mathbf{Pr}\left(\mathbf{x_{ij}} = \mathbf{1} | \mathbf{x_{ij}^{C}}\right)}{\mathbf{Pr}\left(\mathbf{x_{ij}} = \mathbf{0} | \mathbf{x_{ij}^{C}}\right)}\right] = \mathbf{logit}\left[\mathbf{Pr}\left(\mathbf{x_{ij}} = \mathbf{1} | \mathbf{x_{ij}^{C}}\right)\right]$$

$$\mathbf{logit}\left[\mathbf{Pr}\left(\mathbf{x_{ij}} = \mathbf{1} | \mathbf{x_{ij}^{C}}\right)\right] = \sum_{\mathbf{k}=\mathbf{1}}^{\mathbf{k}} \boldsymbol{\theta_k \varphi Z_k}(\mathbf{X})$$

The logit formulation clarifies the interpretation of the θ vector: if forming a tie increases Z_k by 1, then ceteris paribus the log-odds of that tie forming increases by θ_k. Note that a single tie may affect multiple Z_k statistics (Goodreau et al. 2009). θ_k, it can be interpreted as the increase in the conditional log-odds of a partnership between individuals "i" and "j" induced by the formation of the tie and conditional on all other ties remaining unchanged (Handcock, 2003b).

Expanding the model above:

$$\mathbf{Pr}\left(\mathbf{x_{ij}} = \mathbf{1} | \mathbf{x_{ij}^{C}}\right) = \left(\mathbf{1}/\mathbf{c}\right) \mathbf{exp}\left[\boldsymbol{\theta_{\theta e} \varphi Z_{\theta e}}(\mathbf{X}) + \boldsymbol{\theta_{\theta w} \varphi Z_{\theta w}}(\mathbf{X}) + \boldsymbol{\theta_{\rho} \varphi Z_{\rho}}(\mathbf{X})\right]$$

Expressed in words, the equation above states that the probability that an ERG model generated graph is identical to an observed graph is equal to a (generally small but intractable to calculate) constant (1/c) multiplied by the exponent of the sum of the parameters multiplied by the graph statistics (counts) of all the components in the model (Harriagan, 2007).

4. Estimation

The aim when attempting to generate an ERG model is to find the set of parameters which maximize the probability that any random graph generated by simulating the ERG model will be identical to the observed network (Harriagan, 2007).

The constant (1/c) is the sum $exp\{\theta_k Z_k(X)\}$ for the $2^{(n(n-1))}$ graphs (the whole sample space of allowable networks). For graphs n≤6, it is feasible to compute (1/c) explicitly, but for large graphs in the dyad dependent

case the 1/c constant is intractable (because "c" is an astronomically large number). As a result, Maximum Likelihood Estimates (MLE) for dyad dependent models is not available (Strauss and Ikeda, 1990). The MPLE is the product of the probabilities of x_{ij} with each probability conditional on the rest of the data. This model does not depend on the normalizing constant 1/c:

$$\mathbf{Pr} = \left(\mathbf{x_{ij}} = \mathbf{1} | \mathbf{x_{ij}^{C}}\right) = \frac{\left(\mathbf{Pr} = \mathbf{x_{ij}^{+}}\right)}{\left(\mathbf{Pr} = \mathbf{x_{ij}^{+}}\right) + \left(\mathbf{Pr} = \mathbf{x_{ij}^{-}}\right)}$$

The MPLE has the disadvantage that it is an approximate representation of the true joint distribution. These models use local conditional probability distribution so they are not required to factor the full joint distribution (Xiang and Neville, 2008). Because the properties of the MPLE are not well-understood, the more recent body of work has developed the previously mentioned MCMCMLE (Wasserman and Robins, 2005).

In addition to the problem of the normalizing constant "c", we have variables which are highly correlated. For instance we expect having dependence within each row (outgoing), dependence within each column (ingoing) where popularity choices lead to more choices; structural configurations are highly correlated by definition, for instance: an edge is the same as 1-star and many 1-stars form k-stars substructures.

In Statnet, when the ERG model is a dyadic independent model not containing the mutual or asymmetric terms, the true likelihood (MLE) and the MPLE are the same, which is to say that the true MLE may be found via the MPLE computation (Hunter et al. 2008).

The most advanced method used to estimate the parameters of an ERG model in current software is the Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMCMLE). The basic method of the MCMCMLE estimation of parameters for the ERG models involves the simulation of a set of random graphs from a starting set of estimated parameter values and then the refinement of the parameter values by comparing the simulated graphs with the observed graph. A computer program using MCMCMLE procedure generally repeats this process until the parameter estimates stabilize (converge) (Harrigan, 2007).

Table 1. Social Networks statistical models estimators

Estimators	Important features
Maximum Pseudo Likelihood Estimate (MPLE)	• May be poor for models with strong global dependence (Handcock, 2003). • In Statnet, MPLE only=TRUE, it's a function to return the MPLE estimates.
MCMCMLE	• Recommended for dyadic dependence models (Goodreau, 2007). • If we included network specific (dependent) statistics and once we attempt to account for this dependence we can no longer estimate the model using simple logistic regression.
Maximum Likelihood Estimate (MLE)	• For edge and dyad independence models. Dyadic independence: $P(X_{ij} = x_{ij})$ is independent of $P(X_{kl} = x_{kl})\ \forall\ (i,j) \neq (k,l)$ • An inadequate but simple model can be fitted by applying logistic regression, assuming independent tie variables (Snijders).

5. Data

The sociomatrix (Matrix 1) below was manually created on an Excel spreadsheet to intentionally show a tendency of individuals of the same family (show as three different colors at the header) to form a tie. The sociomatrix is a digraph simulating the answer to a name generator as:

"If you had to form a 5-member working group (e.g. to work in a Water User Association project to improve and maintain the irrigation system infrastructure, acquire a tractor), based on your experience, mention at least 5 persons in a descending preference with whom you would like to partner up".

Thus we have a 20x20 asymmetric X sociomatrix, and a 20x1 vector of nodal covariates. The entries of the X matrix, termed adjacency matrix (or sociomatrix) are all 0s and 1s, with $x_{ij}=1$ indicating the presence of an edge between i and j. Because self-nomination was disallowed $x_{ii}=0 \ \forall$ i.

Table 2. Hypothetical sociomatrix (Matrix 1), where 20 farmers were asked to nominate 5 alters (header colors indicates "families" and SES indicates the socio economic status) See this sociomatrix graph on Figure 1

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

Figure 1 (2) depicts Matrix 1(2) graphically, where the size and color of the nodes represent covariates measured at the individual's level (SES).

Using R a vector of numbers from 1 to 20 was randomly reordered. The matrix above was randomly permuted in Ucinet 6 (using the previous vector). The entries of the randomly permuted matrix were replaced on the original Matrix 1, so the SES attribute column, headers and row labels remained the same. The randomly permuted sociomatrix can be seen below:

Table 3. Matrix 2 randomly permuted sociomatrix. See its Graph on Figure 2

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

Hypothetical Social Network,
20 Farmers were asked to nominate 5 Alters,
node's color = Socio Economic Status

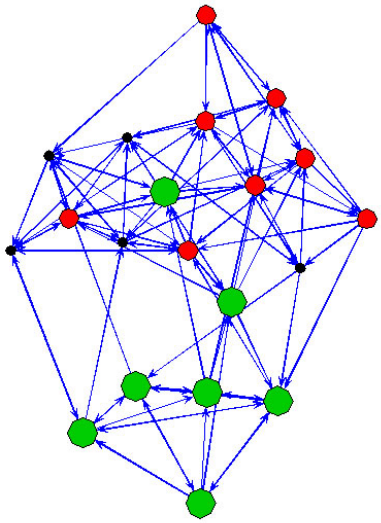

Figure 1

Hypothetical Social Network,
20 Farmers Randomly Nominate 5 Alters,
node's color = Socio Economic Status

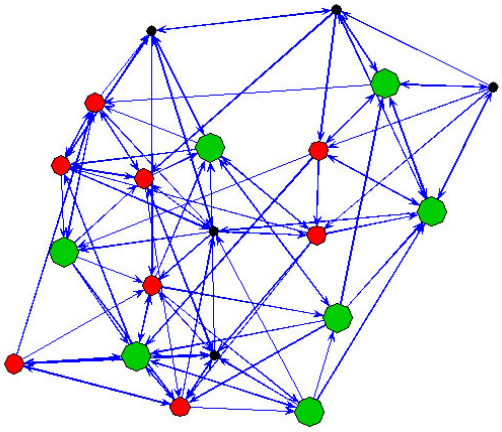

Figure 2

6. Results

Table 4. Approximate maximum likelihood results for a standard ERG model, Matrix 1: hypothetical sociomatrix

Network Statistics	Model 01[A]	Model 02[B]	Model 03[C]
Density	**-1.0296*****	**-1.5268*****	**-1.8573*****
Mutuality	**---**	**1.5266*****	**1.2262*****
Homophily			
Same SES	**---**	**---**	**1.1388*****
AIC	**440.01**	**405.39**	**388.62**
BIC	**443.95**	**413.27**	**400.44**

[A] MLE=MPLE
[B] Monte Carlo MLE, warning: standard errors are suspect due to possible poor convergence
[C] Monte Carlo MLE

The probability of a configuration being present in a network is expressed in an ERG model in the form of parameters θ_k. The parameter values can be read like parameter values of a standard logistic regression, on a log-odds scale controlled for structural effects (Snijders et al. 2006). Statistics with positive values express greater than chance probability of being present in a graph produced by the model ($exp^{\theta_k}/(1 - exp^{\theta_k})$ when θ_k>0 the probability > 0.50), statistics with a negative parameter ($exp^{\theta_k}/(1 - exp^{\theta_k})$ when θ_k<0 the probability < 0.50) have a less than a change of being present. A close to 0 statistics represent a 50/50 chance.

The significance test performed on each of the parameters, the results of which are expressed by asterisks (three asterisks p<0.001), involve the usual null hypotheses that the true parameter values equal 0.

In this example, the negative density parameter indicates that arcs occur relatively rarely, especially if they are not part of a reciprocated relationship or if they are not formed within the same SES.

Table 5. Approximate maximum likelihood results for a standard ERG model, Matrix 2: hypothetical permuted sociomatrix

Network Statistics	Model 01[A]	Model 02[B]	Model 03[C]
Density	**-1.0296*****	**-1.5285*****	**-1.5230*****
Mutuality	**---**	**1.5244*****	**1.5276*****
Homophily			
Same SES	**---**	**---**	**-0.001**
AIC	**440.01**	**405.39**	**407.39**
BIC	**443.95**	**413.27**	**419.21**

[A] MLE=MPLE
[B] Monte Carlo MLE, warning: standard errors are suspect due to possible poor convergence
[C] Monte Carlo MLE

The positive parameter for SES (homophily) (see Table 4) reflects that actors who are similar on SES have a higher tendency to collaborate with each other, which contribute to a positive network autocorrelation on SES.

The amount of deviance explained by the model was checked using the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion), a model is judged better than another model if it has a smaller AIC (or BIC) value.

For Sociomatrix 2 (see Table 5) the network statistics (density and mutuality) remained the same (as in Table 4) because in the permutation process, all rows and columns of the matrix are permuted identically. That is, if row 3 and row 7 are switched, the columns 3 and 7 are switched in the same manner. This form of permutation is a tantamount to "relabeling" of the matrix (actors switch places) while the structure of the network remains unchanged under each permutation. This can be proven checking the dyad census for both matrices which is the same: {Mutual: 25; Asymmetric: 50; Null: 115}. Matrix 2 is a "relabeled" matrix which has a systematic structure (non-randomness). A Bernoulli graph of the same order (n=20) and the same number of edges (100) would not have had a significant mutuality parameter.

As expected, after Matrix 1 was permuted, the SES homophily parameter was no longer statistically significant and close to 0 (θ_w = -0.001).

7. Goodness of fit: how networks look compared to real life networks?

The graphs below are the distribution of frequencies of the minimum geodesic distances for Model 03 of Matrix 1 (Figure 3) and Matrix 2 (Figure 4). What we basically do is choosing a network statistic that's not in the model and comparing the value of this statistic observed in the original network to the distribution of values we get in the simulated networks from our model (Butts et al 2009). In this case, the geodesic from node i to another node j is a path of minimum length. The geodesic distribution is the distribution of frequencies of the geodesic distances ($d_{ij} = d_{ji} \; \forall$ n(n-1)/2 unordered dyads "d" for a graph "G" of order "n") (Pattison and Robins, 2008). If the model is a good fit to the observed data, then networks drawn from this distribution will be more likely to "resemble" the observed data (Butts et al 2009). For examples interpreting the goodness of fit diagnostic graphs consult Hunter et al. (2008). Thus, the dark black line represents the geodesics of the original network, the boxplots represent the distribution across all simulated networks, and the soft lines connect the 95% bounds on these distributions. If the dark line were off the 95% envelope, we would have concluded a poor fit of data to the model. A particular model could fit properly for some statistics and very poorly on others.

Thus we propose to consider an additional statistic: the edge wise shared partner (EWSP) distribution.

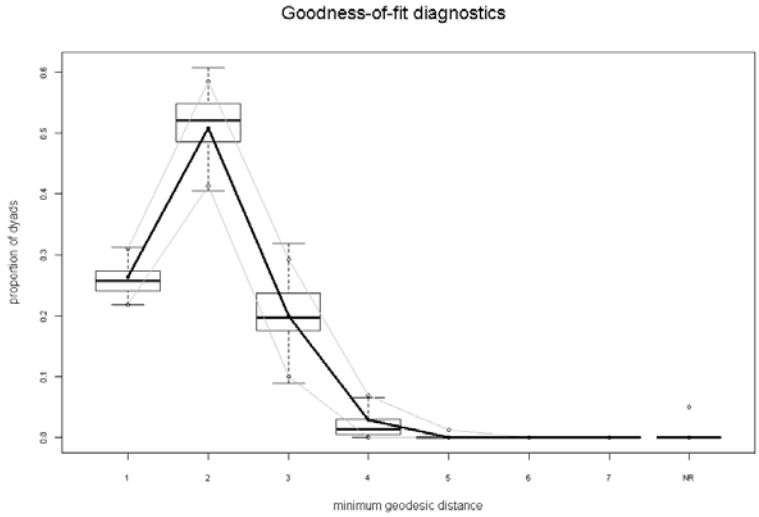

Figure 3. Graphical test of the goodness-of-fit, Matrix 1, Model 03, for geodesic distance

The EWSP distribution consists of the values EP_0/E, ..., EP_{n-2}/E, where E denotes the total number of edges and EP_s equals the number of edges whose endpoints both share edges with exactly "s" other nodes (Hunter et al. 2008).

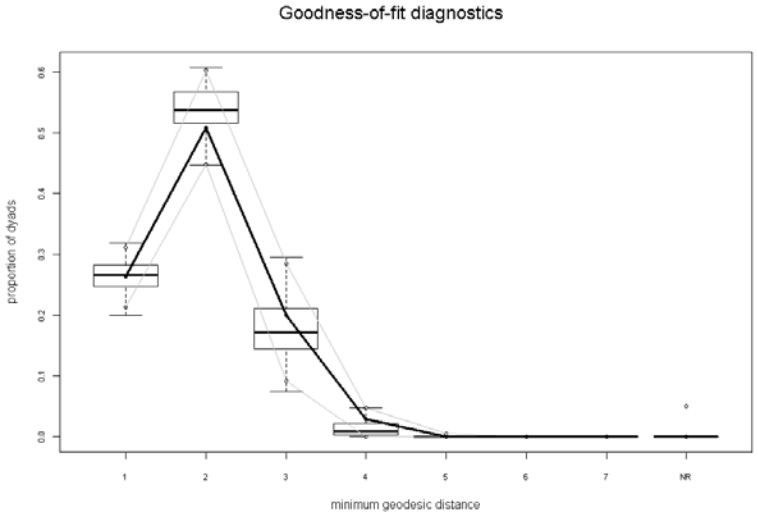

Figure 4. Graphical test of the goodness-of-fit, Matrix 2, Model 03, for geodesic distance

If the observed network is not typical of the simulated network for a particular statistic, then the model is either degenerate (if the statistic is among those included in the ERG model vector θ_k) or poorly fitted (if the statistic is not included) (Hunter et al. 2008).

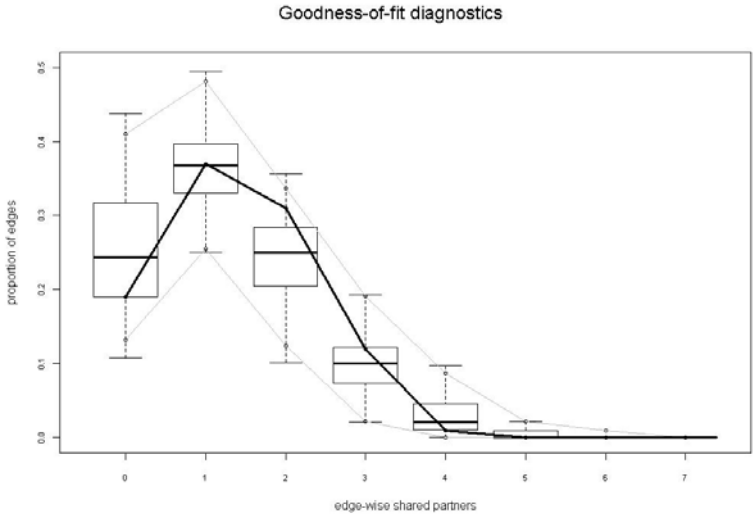

Figure 5. Graphical test of the goodness-of-fit, Matrix 1, Model 03, for edge-wise shared partners

8. Conclusions

It appears that this exploratory model is successful representing structural characteristics of the observed network. The ERG model fit well to the hypothetical data, the Monte Carlo MLE converged satisfactorily and the models produced graphs distributions away from the degenerate region of the parameter space (see figures 4-6).

Despite the fact that we have hypothetical data not meriting a deep interpretation, if we found an empirical network like this, it would show that individuals are not likely to engage in interactions randomly (density = -1.8573, p <0.001). Farmers showed a high tendency to reciprocate their relationships (mutuality = 1.2262, p<0.001). The positive coefficient for SES (p <0.001) showed that individuals are more likely to form ties within the same SES; it increases the odds by 212.3% forming a tie within the same SES than forming one across SES categories ($e^{\theta_\rho} = e^{1.1388}$; $(e^{1.1388} - 1)(100) = 212.3\%$) controlling for the structural mutuality effect, ruling out thus our null hypothesis of SES heterophily.

The mutuality effect was slightly moderated after the SES homophily variable was included. This can be interpreted as evidence that there are organizing principles in this network that go beyond homophilous selection in creating reciprocity (interpretation made by Snijders et al. 2006).

9. References

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Butts, C., Morris, M. Carnegie, N. Krivitsky, P., Handcock, M. S., Hunter, D. R. and Goodreau, S. M. (2009) Statnet Tutorial, presented at: *INSNA Sunbelt workshop February 2009.* Web site: Center for Studies in Demography and

Ecology, University of Washington, Statnet http://csde.washington.edu/statnet/ retrieved 01/31/2010 from http://csde.washington.edu/statnet/Resources/Sunbelt2009/tutorial%20sunbelt%202009%20final.pdf

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

Goodreau S. M., Kitts J. A. and Morris M. (2009) Birds of a feather, or friend of a friend? Using Exponential Random Graph Models to investigate adolescent social networks. Demography, Volume 46-Number 1, 103-125

Handcock, M. S. (2003) Assessing Degeneracy in Statistical Models of Social Networks. Working Paper no. 39 Center for Statistics and the Social Sciences University of Washington, December 31, 2003. Web site: Center for Statistics and the Social Sciences http://www.csss.washington.edu/ retrieved 01/31/2010 from http://www.csss.washington.edu/Papers/wp39.pdf

Handcock, M.S. (2003b) Statistical Models for Social Networks: Inference and Degeneracy. In: Dynamic Social Network Modeling and Analysis: workshop summary and papers. Edited by: Ronald Breiger, Katheleen Carley, and Philippa Pattison. Committee on Human Factors, National Research Council. pp. 229-240

Harrigan N. (2007) Exponential random graph (ERG) models and their application to the study of corporate elites. Web site: Singapore Management University at http://www.smu.edu.sg/ retrieved 02/04/2010 from http://www.mysmu.edu/faculty/nharrigan/index.htm

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

Hunter, D.R., Goodreau, S.M. and Handcock, M.S. (2008) Goodness of fit of Social Network Models. Journal of the American Statistical Association. Vol 103. No. 481 P. 248-258

Knoke D. and Yang S. (2008) Social Network Analysis Second Edition. Sage Publications Inc. pp 133

Pattison P. and Robins G. (2008) Probabilistic network analysis. In: Handbook of probability Theory and Applications Edited by Tamas Rudas. Sage Publications Inc. pp 469

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

Snijders T.A.B, Pattison P., Robins G. and Handcock, M.S. (2006) New specifications for exponential random graph models. Sociological Methodology. Vol. 36 pp 99-153

Strauss D. and Ikeda M. (1990) Pseudolikelihood Estimation for Social Networks. Journal of the American Statistical Association, Vol. 85, No. 409, pp. 204-212.

Wasserman S. and Faust K. (1994) Social network analysis: methods and applications. Cambridge University Press.

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

Wasserman S. and Robins G. (2005) Introduction to Random Graphs, Dependence Graphs, and P*. In: Models and Methods in Social Network Analysis, edited by: Peter J. Carrington, John Scott, Stanley Wasserman. Cambridge University Press, 344 pp

Xiang and Neville (2008) Pseudolikelihood EM for Within-Network Relational Learning. The 2nd SNA-KDD Workshop '08 (SNA-KDD'08), August 24, 2008 , Las Vegas, Nevada , USA. . Web site: College of Science, Department of Computer Science, Purdue University at: http://www.cs.purdue.edu/ retrieved 02/05/2010 from: http://www.cs.purdue.edu/homes/neville/papers/xiang-neville-snakdd2008.pdf