

Modeling Indegree Centralization in NetSAS: A SAS Macro Enabling Exponential Random Graph Models

M. Francis Johnston¹

Center for East-West Medicine; University of California Los Angeles

Xiao Chen

Academic Technology Services; University of California Los Angeles

Phillip Bonacich

Department of Sociology; University of California Los Angeles

Silvia Swigert

School of Education; University of California Los Angeles

The dual purpose of this paper is to (1) introduce SAS computer code (NetSAS) facilitating ERGM analysis of network data and (2) empirically investigate estimation and interpretation of the parameter for indegree centralization. NetSAS directly transforms square-matrix network data into rectangular-matrix dyadic data, thereby eliminating the need for computations exogenous to SAS and extensive data management. The macro is illustrated through estimation on 7 graphs of 21 nodes that vary from 0 to 100% on the conventional graph theoretic measure of indegree centralization. ERGM in a conventional statistical package may facilitate wider use of and further dialogue about the meaning, interpretation, and advancement of the ERGM framework.

INTRODUCTION

Exponential random graph modeling (ERGM, also known as p-star) is a statistical technique for modeling structural properties of networks (Snijders, Pattison, Robins, & Handcock, 2004). Wasserman and Pattison (1996) provide a rationale for modeling dyadic, triadic, subgroup, and entire network characteristics approximately via maximum pseudolikelihood (MP) methods in logistic regression (Wasserman and Pattison 1996, p. 417). Crouch and Wasserman (1998) introduce the PREPSTAR program to calculate preliminary output, along with fairly extensive code for transferring the input into and managing it in SAS. Here, we introduce NetSAS, a macro that enables statistical analysis of networks in SAS. NetSAS directly produces dyadic network data from which SAS can immediately produce basic statistics about the network and carry out ERGM.

To illustrate use of NetSAS, we engage an issue of long-standing importance in the field of network analysis -- centralization (Wasserman & Faust, 1994, pp. 175-7). For directed graphs, there are several operationalizations: indegree, outdegree,

betweenness directed, closeness directed, eigenvector centrality, radiality and integration (Costenbader & Valente, 2003, p. 285). Following Crouch and Wasserman (1998), NetSAS provides the ability to model outdegree centralization and indegree centralization.

We review the graph theoretic and ERGM definitions of indegree centralization and show conceptually the issue of cross-dyadic dependency, which we illustrate with an example. We then empirically investigate the estimation and interpretation of the indegree centralization parameter on 7 graphs, each composed of 21 nodes. Empirically, our primary finding is a modest correspondence between ERGM estimation of the indegree centralization parameter and the conventional graph theoretic measure of indegree centralization. This relationship appears to be mediated somewhat by the effects of cross-dyadic dependency. By enabling analysis in a conventional statistical program, we aim to facilitate wider dialogue about the meaning, interpretation, testing, and advancement of ERGM.

¹ Address correspondence to: Michael Francis Johnston, Ph.D.; Assistant Researcher; UCLA Department of Medicine; Division of General Internal Medicine & Health Services Research; Center for East-West Medicine; 2428 Santa Monica Blvd., Suite 308; Santa Monica, CA 90404; Telephone: 310-453-7679; Fax: 310-315-1856. For an electronic copy of the SAS program, email Dr. Michael Johnston at Johnston@ucla.edu. We thank staff of the UCLA ATS Statistical Consulting group for providing statistical advice. We also thank Paulette Lloyd for commenting on a draft. We bear exclusive responsibility for any errors.

NetSAS

Hitherto, analysts wishing to experiment with ERGM have relied either on PREPSTAR, or highly specialized computer programs such as StOCNET and PSPAR, or even computer languages such as R. Of these ERGM-enabling options, Crouch and Wasserman (1998) created PREPSTAR to facilitate computations in SAS by using a C+ environment to calculate a range of network parameters and then providing extensive SAS code for data input, merging, management, and finally analysis. The procedure is somewhat cumbersome and the PREPSTAR algorithms are not easily interpretable to those unfamiliar with C+.

Inspired by Crouch and Wasserman, we have developed a macro we call NetSAS. NetSAS is a set of self-contained programming statements that shape conventional network data into a rectangular dyadic data matrix format that also provides a range of standard network statistics and ERGM network statistics. The data output by the macro is immediately analyzable by logistic regression in SAS. The macro is in Appendix 1 and includes some additional comments in the program itself.

NetSAS is comprised of two macro programs. The first, NetSAS Part I, produced basic network statistics. The second, NetSAS Part II creates ERGM statistics. Each macro program begins with the line “%macro” and ends with the line “%mend;”. To activate the macro, simply highlight the entire macro and press run (either the SAS running person icon or the Function 3 key [F3]). To obtain results of basic network statistics, run the line “%netstat(5, d:network.txt, netstats);” where “network.txt” refers to the input data set and “netstats” refers to the output dataset. To obtain the ERGM statistics, run the line “%pstar(21, d:network.txt, tdyadic);”. The macro is written with the assumption that the txt file is a square matrix located on the D drive.

NetSAS Part II outputs a SAS file titled “tdyadic”, which is a rectangular-shaped dyadic data matrix composed of one row for each of the directed nodal pairs. The macro transforms the input matrix, a square $g \times g$ network matrix where g is the number of nodes, into an output matrix, a rectangular dyadic data matrix in which each dyad is one row. There are a total of $(g) * (g-1)$ rows in the rectangular dyadic data matrix (following convention, the diagonal of the original network matrix, node-to-itself relations, is excluded). The number of dyads (rows) in a rectangular dyadic data matrix is the number of observations, for which we reserve the symbol “ n ”. The dyadic data matrix includes column vectors for all network statistics produced in PREPSTAR: density, mutual, outstars, instars, mixed stars, transitivity, cycles, outdegree centralization (also known as degree centralization) and indegree centralization (also known as group prestige).

Once the macro has produced the dyadic data matrix, take a few moments to examine the data. One step is to examine the dyadic structure of the new dataset by printing out the nodal relations, which entails the “From” node of the directed relation, the “to” node of the directed relation, and the value of

the relation (1 if there is a relation between the nodes and a 0 otherwise). SAS code to do so is provided underneath “Comment 1”. A second step is to examine the network statistic values in the rows (see “Comment 2”). A third step is to examine the frequencies for variables of interest (see “Comment 3”).

The next step is to fit a logistic regression model to the data (see code under “fitting the model”). When entered, the SAS code will generate output, from among which a few pieces of information are vital. Towards the top, “number of observations read” indicates the total number of directed node-to-node relationships. Further on down, under “Analysis of Maximum Likelihood Estimates,” is a listing of the parameters in the model, their point estimates, standard errors, Wald Chi-Square Value, and probability of significance. Finally, there is a suite of statistical procedures for assessing model fit, which, as we describe in greater depth below, are very important in ERGM. Allison (1999) provides an excellent description of how to use SAS to carry out preliminary data characterization methods, the logistic regression procedure, and diagnose any model specification problems.

Defining Indegree Centralization: Graph Theoretic and ERGM

Indegree centralization is, roughly, a measure of the variability of actor scores on indegree centrality (Wasserman & Faust, 1994, pp. 176). When one actor’s degree centrality score is high compared to the rest, the centralization score for the network as a whole will be high. Conversely, when actors have relatively equal degree centrality scores, centralization will be low. Freeman (1979) provides the conventional graph theoretic measure of indegree centralization (Formula 1). Note that indegree centralization is normalized so that scores range from 0% (a circle graph) to 100% (a star graph). In Formula 1, C_{FID} stands for a measure of centralization as defined by Freeman based upon vertex indegree, $L_{ID}(v^*)$ denotes the vertex with the largest indegree, $L_{ID}(v_i)$ refers to the indegree of a vertex, and g refers to the number of vertices in the original square matrix (Wasserman & Faust, 1994, p. 180, 177).

$$C_{FID} = \left[\sum_{i=1}^g L_{ID}(v^*) - L_{ID}(v_i) \right] / (g-1)^2 \quad (1)$$

In ERGM, indegree centralization and other network statistics are calculated via change score statistics. The general formula for change score statistics is Formula 2 (Anderson et al, 1999, p 48), where $z(x_{ij}^+)$ refers to the situation in which the tie from node i to node j is forced to be present, and $z(x_{ij}^-)$ refers to the situation in which the tie from node i to node j is forced to be absent. Formula 2 indicates that change scores are actually calculated in one of two ways: (1) Existent Relation Present – Existent Relation Hypothetically Absent, or (2) Non-Existent Relation Hypothetically Present – Non-Existent Relation Absent. Essentially, change scores measure how a particular network statistic would differ if the social network under scrutiny were to change by either the addition or subtraction of one

social network tie. In the rectangular dyadic data matrix, there is one column vector for each network statistic so that the effect of adding or subtracting a tie is carried out for each dyadic relationship (that is, each row). Those readers who wish to review a detailed example of how change scores are constructed may find Crouch and Wasserman (1998) to be helpful.

$$\varpi_{ij} = \log \left\{ \frac{\Pr(X_{ij} = 1 | X_{ij}^c)}{\Pr(X_{ij} = 0 | X_{ij}^c)} \right\} = \theta' [z(x_{ij}^+) - z(x_{ij}^-)] \quad (2)$$

The formula used to estimate indegree centralization is based on a measure of the number of choices received (Anderson et al., 1999, p. 57), which is a variance-based measure. In Formula 3 (Wasserman & Faust, 1994, page 180), C_{VID} is the variance-based definition of indegree centralization, $I(v_i)$ represents the indegree of the i^{th} node, \bar{I} denotes the average nodal indegree

$$C_{VID} = \left[\sum_{i=1}^g (I(v_i) - \bar{I})^2 \right] / (g-1) \quad (3)$$

One of the strengths of the variance-based measure of indegree centralization in comparison to the conventional graph theoretic measure of indegree centralization is that the variance-based measure allows for a larger number of change score values.³

Variance-Based Indegree Centralization Reveals Cross-Dyadic Dependency in ERGM

Unique to the calculation of network statistics in a change score framework is what we refer to as cross-dyadic dependency. To discuss this in depth with reference to indegree centralization, we first note that there will be, at most, “n” distinct values for the indegree centralization change scores. Consider a 10x10 square matrix will become a rectangular matrix consisting of 90 rows. For such a matrix, there are $[g*(g-1) = 10*9 =]$ 90 dyadic relations. If the dyads were completely independent of each other, there would potentially be 90 distinct values for the indegree centralization change scores.

Even with independence, there might be less than 90 distinct values for the indegree centralization change scores. One reason is very common, namely that in any dataset some values might occur more than once. Imagine that final grades for a class of 90 undergraduate students could potentially range from 0 to 100 total possible points. In this individualistic example, undergraduates would be considered as independent of each other but it is likely that a few might have the same number of

total points. Despite independence among observations in this example, there would be less than 90 distinct values for the final numeric grade. The network equivalent of this individualistic example is to note that some vertices might have the same indegrees, which would result in a fewer number of indegree centralization change scores than the possible maximum. This is not what we mean by cross-dyadic dependency.

By cross-dyadic dependency, we are referring to the realization of a much fewer number of values for indegree centralization change scores (and other network statistics) than the maximum possible because of dependencies among the dyads which arise because individual vertices are involved in more than dyad. This becomes obvious when the rectangular matrix of dyads is arranged by the “to” vertices. For example, consider output from an analysis of a size 10 network from the Knoke bureaucracies in UCINET, the matrix titled Money. Table 1 shows all of the node-to- node relations that involve Node 5 (indegree=1) and Node 8 (indegree=6). Node 5 only receives money from one organization, Node 1, which is reflected in the column labeled Y. There is only a single 1 which is located in the first row-the row that corresponds to the directed relationship FROM node 1 TO Node 5. Since Node 5 does not receive money from any of the other organizations, all of the other rows have a 0 in the column labeled Y. In contrast, Node 8 receives money from six other organizations.

Table 1: Indegree Centralization Scores (Variance)

FROM Node	TO Node	Y	Change Score CID
1	5	1	-0.36667
2	5	0	-0.16667
3	5	0	-0.16667
4	5	0	-0.16667
6	5	0	-0.16667
7	5	0	-0.16667
8	5	0	-0.16667
9	5	0	-0.16667
10	5	0	-0.16667
1	8	1	0.74444
3	8	1	0.74444
4	8	1	0.74444
5	8	1	0.74444
7	8	1	0.74444
9	8	1	0.74444
2	8	0	0.94444
6	8	0	0.94444
10	8	0	0.94444

Cross-dyadic dependency arises from calculating indegree centralization by applying a variance-based operationalization

² Although Wasserman and Faust use “g” as the denominator, we use (g-1) for the sake of consistency with the conventional way of computing variance.

³ The conventional formula does not distinguish differences between the nodes in terms of indegree centralization, and, as a result, a large number of node-node relations will cluster into an insufficient number of categories to employ the resulting vector as a variable in a logistic regression analysis.

within a change score procedure. Note that all dyads which involve Node 5 as the “to” node has either one of two values for the indegree centralization change score, either -0.36667 or -0.16667. Note furthermore the pattern organizing these realizations. All dyadic relations involving Node 5 as the “to” when the tie is actually existent in the data (Y=1) have an indegree centralization change score of -0.36667. When the tie is actually non-existent in the network (Y=0), the indegree centralization change score is -0.16667. This pattern also holds for all couples involving Node 8 as the “TO” node (0.7444 when Y=0 or 0.94444 when Y=1) and each of the other nodes.

If the dyads were independent of each other, there could be as many as 90 distinct indegree centralization change score values. Because of cross-dyadic dependency, however, these 90 dyadic relations would fall into at most $2g = 2 * 10 = 20$ indegree centralization scores. As discussed above, calculating indegree centralization by applying a variance-based operationalization within a change score procedure for matrices of size 10 will oftentimes result in less than 20 indegree centralization scores, because nodes with the same indegree will have the same value for their indegree centralization change score.

Table 2. 13 Categories of Indegree Centralization Scores

FROM Node	TO Node	Y	Change Score CID	# node-node relations
10	6	0	-0.38889	30
4	7	1	-0.36667	2
10	7	0	-0.16667	16
8	10	1	-0.14444	2
2	10	0	0.05556	7
7	2	1	0.07778	3
10	2	0	0.27778	6
8	9	1	0.30000	4
10	9	0	0.50000	5
9	3	1	0.52222	5
10	3	0	0.72222	4
9	8	1	0.74444	6
10	8	0	0.94444	3

Variance-Based Indegree Centralization in ERGM

Of the possible 20 indegree centralization change score values, there are only 13 in Money. Each realization corresponds to a particular kind of node-node relation that is based upon the “to” node and the value of “Y” (see Table 2). Notice that the most negative indegree centralization change score category is -0.38889, which involve nodes with a zero indegree as the “to” node. This signifies that if a tie were to be added to a node with zero indegree, there would be a decrease in the amount of indegree centralization in Money. The next smallest change

score category, -0.36667, involves a node that has an indegree of 1. This signifies that if a tie were to be eliminated to a node with indegree of one, there would be a decrease in the amount of indegree centralization in Money. The third smallest change score category, -0.16667, shows that if a tie were to be added to a node with indegree one, there would be a decrease in the amount of indegree centralization in Money.⁴ Informing the calculation of change scores for indegree centralization is the general idea that if all the nodes had exactly the same indegree, the graph would be entirely non-centralized.

The two largest change score values are associated with the node with the largest indegree, Node 8: 0.94444 and 0.74444. The largest occurs when Node 8 is changed from a node with indegree of six to a node with indegree of seven, thereby increasing the amount of indegree centralization in the graph, even compared to the second largest which occurs when Node 8 is changed from a node with indegree of six to a node with indegree of five. Table 2 shows one of the desirable properties of using the variance-based measure of indegree centralization to calculate change scores, namely that when the node-node relations are ordered by magnitude of the change score values, the dyadic relations with the largest indegrees score the highest.

Estimating Indegree Centralization in ERGM

In the previous section, we suggested that the method of calculating change scores, though it may account for the non-independence among dyads, also brings about cross-dyadic dependency. Specifically, we showed that those node-node observations with the same “to” node will have either one or two values for the change score of indegree centralization. Recall that a primary assumption of generalized linear models, of which logistic model is a specific example, is that observations are independent of each other (Agresti, 2002, p. 116, 455).⁵ What is the impact of violating this assumption of statistical independence?

A first order of concern prompts the question: Does cross-dyadic dependency bias the coefficient estimate for indegree centralization? One way to approach this question is to conceptualize cross-dyadic dependency as a type of clustering similar to students nested in a classroom — dyadic relations with the same “to” node can be grouped together as being part of the same setting. In this way, those who take a standard approach to statistical modeling would seem to argue no, the coefficient

⁴ We find this negative value to be mildly counter-intuitive. We had expected that taking away a tie to a node with one indegree would increase the amount of indegree centralization. However, we do not consider this to be strongly counter-intuitive because the decrease in indegree centralization is much greater when a tie is taken away from a one-degree node than when a tie is added.

⁵ See also Hardin & Hilbert, 2003, p. vii : “...[B]eing likelihood based, [Generalized Linear Models] assume that individual rows in the data are independent from one another. However, in the case of longitudinal and clustered data, this assumption may fail. The data are correlated.”

estimate is not biased.⁶ We hasten to add, however, that this issue is now being debated in a large and rapidly growing area of statistical literature addressing what is variously labeled as cluster-level covariates, correlated binary data, or random effects modeling. In this area, some statisticians advocate for a more complicated model that includes a cluster-specific random effect term within the logit model (for a discussion, see Hosmer and Lemeshow 2000, pp. 308-330). Beyond the scope of our paper is another special branch of statistical modeling known as Generalized Estimating Equation (GEE), which adjusts both parameter estimates and standard errors for clustering by using a population average model (Hardin & Hilbert, 2003). Both random effects and GEE may provide much traction for modeling correlated binary data. But they are still relatively new areas of research, and many modeling details are in the process of being worked out. After reviewing much of this research, Hosmer and Lemeshow (2000, p. 327) write: “we think it best to proceed cautiously when fitting cluster-specific models.”

A second order of concern prompts the question: Does clustering affect the standard error estimate for indegree centralization? The answer appears to be yes. From the perspective of those utilizing a conventional logistic regression modeling framework, when clustering impacts variance, it will almost always inflate the variance of the binomial response variable and only rarely in practice deflates the variance (Collett, 2003, p. 195). Various models have been proposed to weigh the data to compensate for inflated variance (Collett, 2003, pp. 202-213). Since variance is an important component in the calculation of standard errors in logistic regression (see Collett 2003, Chapter 3 for details), it is likely that problems with the variance would lead to bias in the standard errors for indegree centralization. This might be the factor that motivated Wasserman and Pattison (1996, p. 415, 424) to advocate for testing overall model fit (by comparing model fit with and without the parameter) instead of examining inferential tests for particular parameters in their original p-star paper.

More recently, Snijders and colleagues (2004, p. 7) have claimed that the chi-squared likelihood ratio tests, which logistic regression packages automatically compute to evaluate the statistical significance of particular coefficient parameters, are problematic.⁷

⁶ For example, Long (1997, p. 50), after a mathematical proof specifically on the impact of clustering on coefficient estimation writes: “Consequently, the probability of an event is unaffected by the identifying assumption regarding $\text{Var}(\mathcal{E} | x)$. While the specific value assumed for $\text{Var}(\mathcal{E} | x)$ is arbitrary and affects the β 's, it does not affect the quantity that is of fundamental interest, namely, the probability that an event occurred... The critical point is that while the β 's are not affected by the arbitrary scale assumed for \mathcal{E} , the probabilities are not affected. Consequently, these probabilities can be interpreted without concern about the arbitrary assumption that is made to identify the model. That is to say, the probabilities are estimable functions. Further, any function of the probabilities is also estimable. Importantly, we can interpret changes in probabilities and odds, which are ratios of probabilities.”

We summarize our understanding of parameter estimation for indegree centralization with the following five points.⁷

1. Conceptually, the parameter estimates the extent to which indegree centralization contributes to a graph's overall structure by computing the extent to which “the actual network” differs from “the set of all hypothetical networks distinguished by just a one tie.”
2. Computationally, the indegree centralization parameter is estimated in a change score format with a variance-based operationalization.
3. Because of cross-dyadic dependency in the data, observations with the same “to” node will have at most two distinct values for the indegree centralization change score.
4. Cross-dyadic dependency may bias the coefficient estimate for indegree centralization, but this point is debated.
5. Cross-dyadic dependency likely biases estimation of standard error.

We now turn to empirically examine the estimation of the indegree centralization parameter. To maximize insight into the basic workings of inferential statistics in ERGM, and avoid the issue of biased standard errors, we carry out this work out in a bivariate framework, where testing a coefficient parameter is equivalent to testing overall model fit (Hays, 1963, pp. 354, 375, 465).

Data and Analysis

In this section, we begin the process of testing parameter estimation of indegree centralization in the ERGM framework with selected graphs that have twenty-one nodes. We choose to start with networks of size 21 for two primary reasons. First, this is a network size of interest to those who carry out research in education in that many classrooms have approximately 20 students, as is the case for data analyzed in Anderson et al. (1999, pp. 42-44). Second, there is well-known data available with 21 nodes (Krackhardt, 1987). The first graph we choose to examine is Circle, in which each node chooses two others. Circle is considered the most non-centralized, or most egalitarian, of graph structures. On the other side of the spectrum, we have chosen Hierarchy, a graph in which one node receives ties from each of the other 20 nodes but this node does not choose the other nodes and the other nodes do not select each other (in other words, this is a directed star graph). Additionally, we analyze three well-known graphs collected by Krackhardt (1987) concerning relations between 21 managers in a company, manufacturing high-tech equipment on the west coast of the United States. Each manager was asked two questions. Answers to the first question (“To whom do you go to for advice?”) are recorded in a graph we label as “Advice.”

⁷ “To estimate the parameters, the pseudo-likelihood method continued to be used, although it was acknowledged that the usual chi-squared likelihood ratio tests were not warranted here...” (Snijders et al., 2004, p. 7).

Information from the second question (“Who is your friend?”) is in the overall graph, “Friendship.” Also, collected from company documents was information about a third type of tie: “To whom do you report?” We label this overall graph as “Reports.” We do not provide graphics for Circle, Hierarchy, Advice, and Reports because these networks are very straightforward.

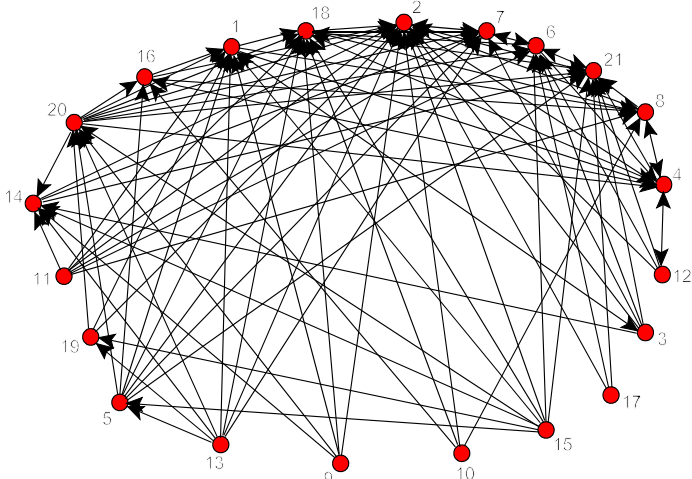


Figure 1. Graphic for Actor 2 (Advice)

Krackhardt also asked each of the managers to indicate what he or she perceived to be the relations among all other managers. So, for each actor, there is a graph for advice relations among the 21 actors and a graph for the friendship relations among each of the 21 actors. From these 42 matrices, we selected the perception of the second actor of the advice relation among the 21 managers, because it has a relatively high amount of indegree centralization (see Figure 1). We also selected the perception of Actor4 of the friendship relations among the 21 managers, because it has a low amount of indegree centralization (see Figure 2).

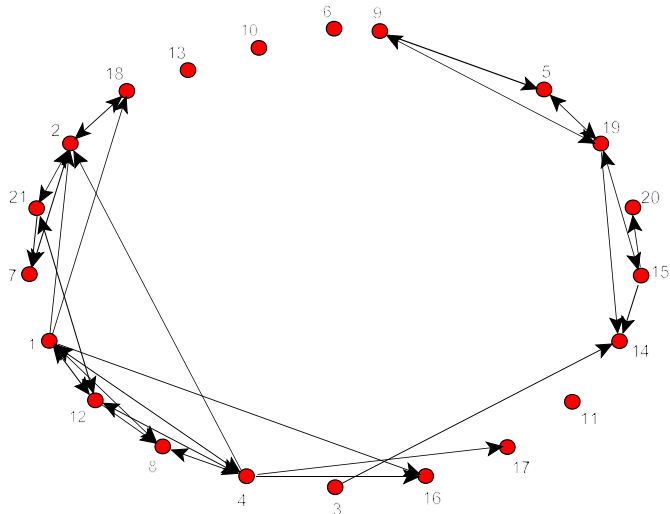


Figure 2. Graphic for Actor 4 (Friendship)

We begin by identifying the amount of indegree centralization in each of the networks. To provide a sense for the level of centralization in these networks, we first compute indegree

centralization scores in UCINET (Borgatti, Everett, & Freeman, 1999). Other measures of centralization could have been used, for example a variance-based measure of centralization. However, we chose the standard calculation (graph indegree centralization) because it is widely used and recognized.⁸ Scores are shown in Table 3. Ordered from most to least centralized, the graphs are: Hierarchy, Actor2, Advice, Reports, Friendship, Actor4, and then Circle.

Table 3 also contains coefficient and standard error estimates for the indegree centralization parameter as computed in SAS. First, note that the parameter estimates for Hierarchy and Circle are both very high and in the expected direction: Hierarchy is highly positive and Circle is highly negative. But, corresponding standard errors are also extremely high, and therefore the p-values show the structure to be insignificant, though in truth the structure is very significant. The standard errors are inflated because the logistic regression model is being fit to data with a very small number of change score values.

Next, we turn to examine Actor2, which is a relatively centralized graph. As expected, there is a relatively strong important coefficient value and a low standard error. Moreover, the chi-square statistic identifies the amount of centralization in the graph as statistically significant.

Consider now the graph for Advice, which is a less centralized graph. The estimated coefficient is smaller (1.663, compared to 2.820 for Actor2). Also, the estimated standard error is small, so that the p-value is statistically significant ($p < .0001$). Surprisingly, Reports has a much higher coefficient value than is the case for Actor2 and Advice. There is also a much higher standard error. In part, we attribute this higher standard error to sparsity of ties in the matrix: there are only 20 ties present from a total possible of 420 and only 9 different indegree centralization change score values, which are skewed right.

The next most centralized graph is Friendship. Though similar in centralization score to Reports, Friendship has many more ties (102 vs. 20) and more change score values (16 vs. 9). Moreover, the structure of relations is less skewed: two nodes each have an indegree of 1, 2, and 3. Three nodes have an indegree of 4, five nodes an indegree of 5, and four an indegree of 6. At the other tail, two nodes have an indegree of 8 and one an indegree of 10. The parameter estimate is smaller than the estimates for Actor2, Advice, and Reports (0.730 vs. 2.82, 1.663, 5.335). Note, however, in contrast to the previous matrices, that the p-value is non-significant.

Finally, we turn to the graph for Actor4, which has the second lowest indegree centralization score. There are a slightly higher number of ties compared to Reports (36 vs. 20). Yet, more ties

⁸ Additionally, the variance measure of centralization is not very standardized. Suppose, for example, that we examine the probability (p) that any pair of individuals is connected in a random graph. The variance indegree will be $Np(1-p)$, the variance of the binomial distribution where N is the number of alternative partners for each person (one less than the group size).

does not translate into more spread—the number of nodes with distinct indegree scores is the same (5), as is the number of distinct change score values (9). Indeed, the nodal indegree ranges from 0-4 for Actor4, as compared to 0-7 for Reports, which suggests that the change score values will also have a much more restricted range. One quality of the graph for Actor4, when compared to Reports, is that it is less skewed: five nodes have an indegree of 0, four with an indegree of 1, five with an indegree of 2, six with an indegree of 3, and one with an indegree of 4. All the nodes fall close around the average indegree of 1.7. The parameter coefficient is very low (0.195), indicating a very small slope. Also, the standard error is a bit high (1.441), perhaps because of the lack of spread in nodal indegrees. Similar to Actor4, the p-value is non-significant (0.8924).

DISCUSSION

This paper introduces NetSAS, a SAS macro that transforms conventional network data into dyadic data and provides a set of conventional network statistics, both of which facilitate ERGM in SAS. We use the program as a launching point to examine estimation of the indegree centralization parameter in ERGM. We identify and discuss the origin and some consequences of cross-dyadic dependency. We further suggest that the idea of clustering may be one fruitful way to conceptualize cross-dyadic dependency.

We agree with Seary and Richards (2000, p. 87) that it is desirable to proceed with caution when estimating indegree centralization in an ERGM format, though we are by no means calling for researchers to abandon a variance-based operationalization. When we empirically tested parameterization by analyzing seven matrices that span the range of indegree centralization according to the conventional graph theoretic measure, we find that the ERGM framework is generally able to identify those networks with more than a modest amount of indegree centralization. We recognize that our research design is insufficiently rigorous to claim that a variance-based measure of indegree centralization is justifiable in the ERGM framework. However, we claim that these results provide a warrant for further research on this topic, especially that which examines the usefulness of conventional statistical procedures in correcting for clustering. One traditional approach would be to apply a post-hoc overall adjustment to standard errors through a technique often referred to as the sandwich variance estimator (Hardin and Hilbe 2003, p. 5). Another traditional technique, known as fixed effects modeling, would be to include an indicator variable for each node, omitting one from the model. Multilevel modeling would enable adjustments to both the parameters and standard errors in light of clustering. Yet another technique is the method of population average known as GEE (Hardin and Hilberg 2003).

Table 3. Indegree Centralization Network Statistics, Graph and ERGM Parameterization

Graph	Total Number of Ties	Number of Indegree Values	Number of Change Score Values	Graph Indegree Centralization	P-star coef.	Standard Error	P-value
Hierarchy	20	2	2	100.00%	18.941	31568.400	<0.9995
Actor2	110	14	24	77.50%	2.820	0.310	<0.0001
Advice	190	10	20	47.00%	1.663	0.284	<0.0001
Reports	20	5	9	31.75%	5.335	0.929	<0.0001
Friendship	102	8	16	27.00%	0.730	0.540	<0.1764
Actor4	36	5	9	12.00%	0.195	1.441	<0.8924
Circle	42	1	2	0.00%	-264.30	2656.900	<0.9208

Our paper raises another issue worthy of further attention, namely the relation between the ERGM parameter estimate of indegree centralization and the graph theoretic measure. The analyses presented here suggest correspondence, but many details remain to be worked out. For example, what does it mean to hypothesize that the null value of the coefficient is zero? Is this equivalent to hypothesizing that the variance-based graph measure of indegree centralization is 50%? Will the expected value of the distribution vary substantially by graph size? Recent research by Tallberg (2004) suggest possible ways to address these and related questions about model testing using simulation methods.

Efforts to test and compare measures within and between datasets (e.g. Costenbader and Valente 2003) provide a scientific foundation fostering the diffusion of this statistical network methodology. Programs permitting ERGM in conventional statistical packages are critical for enabling a larger number of people to participate in building a more practical foundation with well-understood strengths and limitations. In this paper, we have studied indegree centralization, showing how further scrutiny may uncover issues worthy of further attention. More widespread participation in dialogue about the meaning, interpretation, and testing of ERGM is critically important for further advancing and diffusing this network science innovation.

References

- Agresti A. 2002. *Categorical data analysis*. Hoboken, NJ: Wiley-Interscience.
- Aldrich JH & Nelson FD. 1984. *Linear probability, logit, and probit models*. Thousand Oaks, CA: Sage Publications.
- Allison PD. 1999. *Logistic regression using the SAS system: Theory and application*. SAS Press and John Wiley Sons.
- Anderson C, Wasserman S, & Crouch B. 1999. A p* primer: Logit models for social networks. *Social Networks*, 21: 37-66.
- Borgatti S, Everett M, & Freeman L. 1999. *UCINET 5 for Windows: Software for social network analysis*. Natick: Analytic Technologies.
http://www.analytictech.com/ucinet_5_description.htm
- Collett D. 2003. *Modeling binary data, second edition*. New York: Chapman and Hall/CRC.
- Cook RD & Weisberg S. 1999. *Applied regression including computing and graphics*. New York: Wiley-Interscience.
- Costenbader E & Valente T. 2003. The stability of centrality measures when networks are sampled. *Social Networks*, 25: 283-307.
- Crouch B & Wasserman S. 1998. A practical guide to fitting p* social network models. *Connections*, 31: 87-101.
- Frank K, & Yasumoto J. 1998. Linking action to social structure within a system: social capital within and between subgroups. *American Journal of Sociology*, 104: 642-86.
- Freeman L. 1979. Centrality in social networks: conceptual clarification. *Social Networks*, 1: 215-239.
- Hardin JW & Hilbert JM. 2003. *Generalized estimating equations*. New York: Chapman & Hall/CRC.
- Hays WH. 1963. *Statistics for psychologists*. New York: Holt, Reinhart, and Winston.
- Holland PW & Leinhardt S. 1975. The statistical analysis of local structure in social networks. *Sociological Methodology*, 197: 1-45.
- Hosmer D & Lemeshow S. 1989. *Applied logistic regression*. New York: Wiley-Interscience.
- Hosmer D & Lemeshow S. 2000. *Applied logistic regression* (2nd ed.). New York: Wiley-Interscience.
- Koehly L & Wasserman S. 1996. Classification of actors in a social network based on stochastic centrality and prestige. *Journal of Quantitative Anthropology*, 6: 75-99.
- Knoke D & Wood J. 1981. *Organized for action: Commitment in voluntary associations*. New Brunswick, NJ: Rutgers University Press.
- Knoke D & Kuklinski J. 1982. *Network analysis*. Thousand Oaks, CA: Sage Publications.
- Krackhardt D. 1987. Cognitive social structures. *Social Networks*, 9: 104-134.
- Lazega E & Pattison PE. 1999. Multiplexity, generalized exchange and cooperation in organizations: A case study. *Social Networks*, 21: 67-90.
- Lazega E & van Duijn M. 1997. Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data. *Social Networks*, 19: 375-397.
- Long J. 1997. *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.
- Lubbers M. 2003. Group composition and network structure in school classes: A multilevel application of the p* model. *Social Networks*, 25: 309-332.
- Seary AJ & Richards WD. 2000. Fitting to p* models in Multinet. *Connections*, 23(1): 84-101.
- Skvoretz J & Faust F. 1999. Logit models for affiliation networks. In Sobel ME & Becker MP, (Eds.), *Sociological Methodology* (253-280). Cambridge, MA: Basil Blackwell.
- Snijders TAB, Pattison PE, Robins GL, & Handcock MS. 2004. New specifications for exponential random graph models. Manuscript.
- Tallberg C. 2004. Testing centralization in random graphs. *Social Networks*, 26(3): 189-288.
- Wasserman S & Faust K. 1994. *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.
- Wasserman S & Pattison PE. 1996. Logit models and logistic regressions for social networks: I. An introduction to markov graphs and p*. *Psychometrica*, 61: 401-425.

Appendix 1: NetSAS

```

/*****
NetSAS, Part I - produces basic network statistics
Version 1.0 Modified August 26, 2005
*****/

/*****
netstat program takes three parameters:
nnodes: number of nodes
infile: the input file for a network data
outdata: name for the output sas data file
*****/

%macro netstat(nnodes, infile, outtable);
data indata ;
  infile "&infile";
  input a1 - a&nnodes;
run;

proc iml;
  use indata ;
  read all into x;      /*read in the sociomatrix*/
  G = nrow(x);         /*number of nodes, which is also the number of rows*/
  g2 = g*g;            /*g2 is used for computation purposes*/
  N = g2 - g;          /*number of observations, dyadic pairs, is (g^2 - g)*/

  /*number of edges is the sum of all edges in the matrix*/
  L = sum(x);

  /*density is number of edges divided by number of dyadic pairs*/
  D = L/N;

  /*mean indegree is number of edges divided by number of nodes*/
  Mean_Indeg = L/G;

  /*mean outdegree is also number of edges divided by number of nodes*/
  Mean_Outdeg = L/G;

  /*An outstar is the number of nodes that connect outwards to exactly two nodes*/
  Stars_out = (sum(t(x)*x) - trace(t(x)*x))/2;

  /*An instar is a nodes that receives connections from exactly two nodes*/
  Stars_in = (sum(x*t(x)) - trace(x*t(x)))/2;

  /*number of nodes that have an outward connection to exactly one node
  and an inward connections from exactly one node*/
  Stars_mixed = sum(x*x) - trace(x*x);

  /*number of triads out of all possible triads with a transativity*/
  Trans_triads = trace(x*x*t(x));

  /*number of triads out of all possible triads with a cycle*/
  Cyclicity = trace(x*x*x)/3;

  /*number of dyads out of all possible dyads that have a reciprocated relation*/
  Mutual_dyad =sum(x#x)/2;

  t1 = j(1, g, 0);
  t2 = j(1, g, 0);
  do k = 1 to g;

```

```

      t1[1, k] = (sum(x[1:g, k]) - mean_indeg)**2;
      t2[1, k] = (sum(x[k, 1:g]) - mean_outdeg)**2;
    end;
    cen = sum(t1)/(g-1);
    pre = sum(t2)/(g-1);

c = j(14, 1, 0);

c[1, 1] = G;
c[2, 1] = N;
c[3, 1] = L;
c[4, 1] = D;
c[5, 1] = Mean_Indeg;
c[6, 1] = Mean_Outdeg;
c[7, 1] = Stars_out;
c[8, 1] = stars_in;
c[9, 1] = stars_mixed;
c[10, 1] = trans_triads;
c[11, 1] = cyclicity;
c[12, 1] = mutual_dyad;
c[13, 1] = cen;
c[14, 1] = pre;
names={G N L D Mean_indeg Mean_Outdeg Stars_out Stars_in Stars_mixed Transitivity Cyclicity Mutual ind_cen grp_pres};
heading = {N STAT};
print c [rowname=names colname=heading];
quit;
ods output c = &outtable;
%mend;
%netstat(5, d:\network.txt, netstats);

```

 NetSAS, Part II - produces ERGM statistics
 Version 1.0 Modified September 16, 2005

Formulae are almost entirely derived from Table 4 (page 46) from
 "A p* primer: logit models for social networks",
 Social Networks, 21(1999) 37-66.

Expressed in matrices:

dyadic:
 mutual: $\text{sum}(A\#A')/2$

triadic
 2-out-stars: $(\text{sum}(A^*A) - \text{trace}(A^*A))/2$
 2-in-stars: $(\text{sum}(A^*A') - \text{trace}(A^*A'))/2$
 2-mixed-stars: $\text{sum}(A^*A) - \text{trace}(A^*A)$
 transitivity: $\text{trace}(A^*A^*A')$
 cyclicity: $\text{trace}(A^*A^*A)/3$

average indegree per node: $L = \text{sum}(A)/\text{dim}(A)$;
 degree centralization: $(\text{sum}(\text{in}_i - L)^2)/(g-1)$
 with $\text{in}_i = \text{sum}(x[1:g, i])$
 group prestige: $(\text{sum}(\text{out}_i - L)^2)/(g-1)$
 with $\text{out}_i = \text{sum}(x[i, 1:g])$

Formulae for change scores are derived from them.

*****/

```

/*****

```

```

pstar program takes three parameters:
nnodes: number of nodes
infile: the input file for a network data
outdata: name for the output sas data file

```

Output data file contains change statistics on each dyad. Below is the description of each variable in it. Change score by definition is the difference between the statistic when the tie is present and the statistic when the tie is missing.

Each row (dyad) is indexed by variable From and to, indicating a dyad (i, j).

Variables created in outdata set are:

Dyad-level variables:

```

var1: From      -- from the ith subject
var2: to        -- to the jth subject
var3: y         -- x[i,j], the link indicator between ith subject and jth subject
var4: density   -- currently defined as 1
var5: mutual    -- x[j,i] (rho)
                 when x[j,i] = 0, it will not be a mutual dyad,
                 so the change score = (0-0) = 0 = x[j,i]
                 when x[j,i] = 1, it will be a mutual dyad when the tie (x[i,j]) is present
                 and not a mutual when the tie is missing so the difference is 1 = x[j,i].

```

Triad-level variables:

```

var6: outs      -- 2-out-stars (sigma_o)
                 number of 2-out-stars when the tie is present
                 minus number of 2-out-stars when the tie is missing.

var7: ins       -- 2-in-stars (sigma_i)
                 number of 2-in-stars when the tie is present
                 minus number of 2-in-stars when the tie is missing.

var8: mixs      -- 2-mixed-stars (sigma_m)
                 number of 2-mixed-stars when the tie is present
                 minus number of 2-mixed stars when the tie is missing.

var9: trans     -- transitivity (tau_t)
                 transitivity when the tie is present
                 minus transitivity when the tie is missing.

var10: cyclic   -- cyclicity (tau_c)
                 cyclicity when the tie is present
                 minus cyclicity when the tie is missing

var11: degree_cen -- degree centralization
                 indegree centralization when the tie is present
                 minus indegree centralization when the tie is missing

var12: group_prestige -- group prestige
                 group prestige when the tie is present
                 minus group prestige when the tie is missing

```

```

*****/

```

```

%macro pstar(nnodes, infile, outdata);
  data  indata ;
      infile "&infile";
      input a1 - a&nnodes;
run;

proc iml;
  use  indata ;
  read all into x;      /* read in the sociomatrix */
  g = nrow(x);         /* number of nodes */
  g2 = g*g;           /* number of pairs = g2 - g*/

  t1 = j(1,g, 0);
  t2 = j(1,g, 0);

  /*****
  calculating change statistics, using matrix calculation
  *****/

  c = j(g2, 12, 0);    /*creating a matrix for all pairs*/

  do i = 1 to g by 1;
  do j = 1 to g by 1;
    tmp = x;
    tmp[i, j] = (x[i,j]=0);
    if i ^=j then do;

      out = (sum(t(x)*x)-trace(t(x)*x))/2 - (sum(t(tmp)*tmp)-trace(t(tmp)*tmp))/2;
      in  = (sum(x*t(x))-trace(x*t(x)))/2 - (sum(tmp*t(tmp))-trace(tmp*t(tmp)))/2;
      mixed = sum(x*x) - trace(x*x) - sum(tmp*tmp) + trace(tmp*tmp);
      trans = trace(x*x*t(x)) - trace(tmp*tmp*t(tmp));
      cyc  = trace(x*x*x)/3 - trace(tmp*tmp*tmp)/3;

      lx  = sum(x)/g;
      ltmp = sum(tmp)/g;

      do k = 1 to g;
        t1[1, k] = (sum(x[1:g, k]) - lx)**2;
        t2[1, k] = (sum(tmp[1:g, k]) - ltmp)**2;
      end;
      grpp = sum(t1)/(g-1);
      grpm = sum(t2)/(g-1);
      grp = (-1)**(1-x[i,j])*grpp + (-1)**(x[i,j])*grpm;

      do k = 1 to g;
        t1[1, k] = (sum(x[k, 1:g]) - lx)**2;
        t2[1, k] = (sum(tmp[k, 1:g]) - ltmp)**2;
      end;
      indp = sum(t1)/(g-1);
      indm = sum(t2)/(g-1);
      ind = (-1)**(1-x[i,j])*indp + (-1)**(x[i,j])*indm;

      c[j + g*(i-1), 1] = i;      /*from */
      c[j + g*(i-1), 2] = j;      /*to */
      c[j + g*(i-1), 3] = x[i,j];
      c[j + g*(i-1), 4] = 1;      /*density*/
      c[j + g*(i-1), 5] = x[j,i]; /*mutual*/
      c[j + g*(i-1), 6] = abs(out);
      c[j + g*(i-1), 7] = abs(in);
      c[j + g*(i-1), 8] = abs(mixed);
    end;
  end;
end;

```

```

    c[j + g*(i-1), 9] = abs(trans);
    c[j + g*(i-1), 10] = abs(cyc);
    c[j + g*(i-1), 11] = ind;
    c[j + g*(i-1), 12] = grp;
  end;
end;
end;
create &outdata var {From to y density mutual outs ins mixs trans cyclic
                    degree_cen group_prestige};
  append from c;
quit;
data &outdata;
  set &outdata;
  if from ~ = to;
run;
%mend;
%pstar(5, d:\network.txt, tdyadic);
options nocenter nodate;

*****;
*Becoming familiar with the data;
*****;

*Coment 1. Examine basic data - two nodes and the relation between them;
proc print data = tdyadic;
var from to y;
run;

* Comment 2: Examine network statistics;
* reminder - degree centrality is also known as outdegree centralization;
* reminder - group prestige is also know as indegree centralization;

proc print data = tdyadic;
var density mutual outs ins mixs trans cyclic degree_cen group_prestige;
run;

* Comment 3: We examine indegree centralization more carefully;
proc freq data = tdyadic;
  tables group_prestige;
run;

*****;
* Fitting the Model;
*****;

* We use a single parameter, indegree centralization;

proc logistic data=tdyadic descending;
  model y = group prestige / lackfit rsq ctable;
  output out=opred1 prob=phat;
run;

```

