ELSEVIER

Social Networks 21 (1999) 239–267

SOCIAL NETWORKS

www.elsevier.com/locate/socnet

The interaction of size and density with graph-level indices

Brigham S. Anderson [a], Carter Butts [b,c,1], Kathleen Carley [b,c,d,*]

[a] *Department of Robotics, Carnegie Mellon University, Pittsburgh, PA, USA*
[b] *Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA*
[c] *Center for the Computational Analysis of Social and Organizational Systems, Carnegie Mellon University, Pittsburgh, PA, USA*
[d] *H.J. Heinz III School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA, USA*

Abstract

The size and density of graphs interact powerfully and subtly with other graph-level indices (GLIs), thereby complicating their interpretation. Here we examine these interactions by plotting changes in the distributions of several popular graph measures across graphs of varying sizes and densities. We provide a generalized framework for hypothesis testing as a means of controlling for size and density effects, and apply this method to several well-known sets of social network data; implications of our findings for methodology and substantive theory are discussed. © 1999 Elsevier Science B.V. All rights reserved.

JEL classification: C150 (Statistical simulation and Monte Carlo methods)

1. Introduction

In the study of social networks, positional (or nodal) indices are often employed in order to understand particular features of social positions; likewise, higher level measures like centralization are useful for gaining an understanding of social networks in their entirety. The need to quantify phenomena at the network level has given rise to Graph-level indices (GLIs) such as degree centralization, connectedness, and hierarchi-

[*] Corresponding author. Department of Social and Decision Sciences, Porter Hall 208, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Tel.: +1-412-268-3225; fax: +1-412-268-6938; e-mail: carley@centro.soar.cs.cmu.edu

[1] This material was based upon work supported under a National Science Foundation Graduate Fellowship.

0378-8733/99/$19.00 © 1999 Elsevier Science B.V. All rights reserved.
PII: S0378-8733(99)00011-8

calization. These measures quantify various features of graphs.[2] For instance, in measuring Krackhardt hierarchicalization, one obtains the fraction of connected pairs that are asymmetric in their ability to reach one another. This quantity by itself is informative and useful in that one gains insight into a graph's structure by knowing it; furthermore, in various substantive contexts the value of such a measure may have theoretical significance for some phenomenon of interest per se (the spread of rumors, for instance). Problems arise, however, when one is interested in constructing a sociological interpretation of a graph-level measure. Rather than being concerned with *what* the GLI is for a particular graph, one may wonder *why* it takes on a particular value. For example, examining a very sparse relation, such as ''x is the mentor of y'', will usually result in the observation of an extremely high hierarchicalization value.[3] One can certainly conclude that the network is very hierarchical. However, is this hierarchicalization the result of the nature of ''mentorship'' or merely of the network's sparseness? Similarly, we may run into difficulties when attempting to use this measure as a predictive or classificatory variable. If we examine the hierarchicalization values of mentorship in a variety of populations and find little variance, does this suggest that there is something inherent to mentorship per se which makes it uniformly hierarchical in a wide range of cases, or is it simply the case that such is a necessary result of studying any sparse relation? If we attempt to predict other variables from hierarchicalization and find positive results, should we assume that it is the hierarchy which matters, or the sparseness of the relation? The problem is an important one, and affects our research whether our use of the GLI is independent or dependent, classificatory or motivated by substantive theory.

The reason for this particular difficulty, as we have suggested, lies in the subtleties of the distributions of GLIs across the space of possible structures. In the case described above, problems arise because sparse digraphs have disproportionately high hierarchicalization scores; this follows from the fact that a far greater proportion of reachability relations are asymmetric within this set of graphs than within the set of all directed graphs. When faced with one or more observations of ''high'' hierarchicalization, then, one cannot immediately distinguish between the possibility that the observations follow from the mathematical necessity of sparseness, and the possibility that the observations reflect a network formation process which is biased towards hierarchicalization per se. Without a baseline model (Mayhew, 1984) — that can tell one what one should expect from the most basic parameters of graph structure — one is therefore quite limited in the conclusions one can draw from the GLI values alone.

The hierarchy example demonstrates that density can be a powerful covariate of GLIs, and that failure to carefully consider its effects can lead to difficulties in the analysis of network data. Unfortunately, many if not most GLIs are also quite sensitive to graph size as well. How can we take such factors into account when analyzing network data containing GLIs? Here, we shall attempt to characterize GLI behavior with

[2] Throughout this document, the term ''graph'' will be used generically to refer to both simple graphs and digraphs; results which apply only to simple or directed graphs will be specifically identified as such.

[3] Relative to the maximum.

respect to these most basic of structural parameters, and to suggest how these behaviors impact network theory and methodology. In order to control for size and density effects, hypothesis tests using baseline models on size and density are also described. As an illustration of the use of the technique, simple null-hypotheses will be tested against 20 of the social networks found in the database of the network software UCINET IV (Borgatti et al., 1991) for six common GLIs. Finally, some implications of our findings regarding the interaction of GLIs with size and density for network theory and methodology will be discussed, along with directions for future research.

1.1. Background

Although size and density clearly affect GLIs, efforts to control their effects have often been limited to adjusting the GLIs themselves. To partially remove the effects of size, for instance, measures are often normalized by the maximum attainable value for a graph of a given size. Below is an example of one common normalized measure of graph degree centralization (Freeman, 1979).

$$C_{\mathrm{D}} = \frac{\sum_{i=1}^{g}\left(C_{\mathrm{D}}(n^{*}) - C_{\mathrm{D}}(n_i)\right)}{\max\sum_{i=1}^{g}\left(C_{\mathrm{D}}(n^{*}) - C_{\mathrm{D}}(n_i)\right)} = \frac{\sum_{i=1}^{g}\left(C_{\mathrm{D}}(n^{*}) - C_{\mathrm{D}}(n_i)\right)}{(g-1)(g-2)}. \qquad (1)$$

The $C_{\mathrm{D}}(n_i)$ in the numerator are the individual sums of each of the g actors' in- and out-degrees, or links, while $C_{\mathrm{D}}(n^{*})$ is the largest of these values among all actors. [4] The denominator contains the normalizing term, which is equal to the maximum possible value of the numerator (occurring when the graph has a star configuration). This type of normalization, while limiting the measure to the range of 0–1, does not usefully control for size. As can be seen, the maximum (non-normalized) degree centralization increases at a rate proportional to the square of g. However, there is no reason to believe that raw degree centralization in real social networks will increase at the same rate, or even that the median normalized degree centralization over the population of all possible graphs will fall at the 0.5 point.

In addition to size, density is not controlled for by maximum-value normalization either. It is often the case that the maximum GLI scores are not even attainable for all densities. For instance, the degree centralization for a sparse network with fewer links than that needed for a star configuration can never equal $(g-1)(g-2)$; the same, of course, holds true for extremely dense networks. Recognizing that many GLIs attain maximum values on a very small set of special case graphs which are unlike the larger population of graphs in a variety of respects, one is given to wonder whether other

[4] Throughout this paper, the term ‘‘link’’ will be used to refer to both arcs in digraphs and to edges in symmetric graphs, since statements will often apply to both types.

statistics on these measures are similarly skewed. As we shall see presently, this is in fact the case for a number of commonly employed GLIs.

The revelation that density is interwoven with other GLIs is not a new one. Friedkin (1981) showed that, in the set of GLIs he examined, attempts to control for graph size encounter problems of non-linearity and heteroscedasticity. In the same study, Friedkin also found that density has a strong effect, though his conclusions primarily concerned the merits of the measure of density itself. This paper uses a similar approach to examine the distributions of several other structural measures, and also examines the usefulness of a simple hypothesis test in controlling for both size and density.

A great deal of research along a different vein has gone into controlling not only for density and size, but also the number of mutual, asymmetric, and null ties (Holland and Leinhardt, 1970), and the in- and out-degrees of individual nodes (Snijders, 1991). In general, no analytical methods are known for deriving either means or variances for GLIs under these conditions, much less their distributions. Monte Carlo sampling methods are used by Snijders to control for in- and out-degrees, and a similar approach is taken here for the simpler case of controlling only density and size. This study diverges in purpose from that of Snijders by focussing on the use of actual GLIs and directly illustrating their distributions and the usage of their distributions in hypothesis testing.

2. Graph-level indices

The six GLIs examined are either in common use or illustrate distributions of theoretical interest. Though not included in the list which follows, it must be emphasized that both size and density are also GLIs. Size is defined here as the number of nodes in the graph, and density is given as the average number of links per node. [5] These measures are ‘‘privileged’’ here for the following reasons: first, size and density are often largely determined by exogenous factors such as choice of population, demographics, and spatial layout of actors; second, both size and density of populations may be estimated using network sampling methodologies (Marsden, 1988, 1990; Bernard et al., 1991), and their relationships with other GLIs (which often cannot be estimated in this fashion) are of special interest; and third, size and density are commonly thought of as ‘‘basic’’ dimensions of network structure, and thus provide an intuitive starting point for our analysis. Given this motivation, then, our choice of ‘‘other’’ GLIs is as follows.

(1) Degree centralization (Freeman, 1979) is a measure of the dispersion in vertex degree. This GLI can be defined on both simple graphs and digraphs, and is used here in both cases. [6]

[5] A common expression of density which is not used here is the number of links divided by the number of possible links. Our measure of density, however — equivalent to the mean degree centrality — is useful when comparing against size effects, and is of theoretical interest in models which are founded at the microlevel (where number of ties per node is a common constraint — see also Mayhew et al., 1995).

[6] Although our exploratory simulations consider exclusively the directed case, results for simple graphs are used in hypothesis testing (see Tables 1 and 2 and Appendix A).

(2) Betweeness centralization (Freeman, 1977; Gould, 1987) is a measure of the dispersion in betweeness centrality (as per the general form of Eq. (1) above). As with degree centralization, betweeness centralization is defined on both simple and directed graphs.

(3) The hierarchy of Hummon and Fararo (1995) (H–F) is the length of the longest directed path in the graph minor formed by condensing all maximal strongly connected subgraphs into single vertices, normalized by the maximum possible path length. Because the H–F hierarchy of any simple graph is trivially zero, this measure is employed only on directed graphs.

(4) The hierarchy of Krackhardt (1994) is the fraction of dyads in the graph which are neither strongly connected nor strongly disconnected (i.e., one vertex can reach the other through some path, but the other vertex cannot reach it). This measure is defined only for directed graphs.

(5) Connectedness (Krackhardt, 1994) is the fraction of all vertex pairs which are not strongly disconnected. This is considered on both directed and simple graphs.

(6) Efficiency (Krackhardt, 1994) is, essentially, the degree to which the graph uses as few links as possible to connect the nodes which are already connected in the graph. This is defined on both directed and simple graphs.

These measures fall into three broad categories which capture typical social network dimensions: measures (1) and (2) are of centralization, (3) and (4) relate to hierarchicalization, and (5) and (6) measure aspects of connectedness. Such GLIs are interesting both from a substantive theoretical and a classificatory point of view; degree of hierarchicalization, for instance, is thought to relate to task performance (Cohen, 1962; Aldrich, 1978; Carley, 1992), and is also a key dimension along which organizations have often been classified (Blau, 1972; Krackhardt, 1994).

2.1. Generating GLI distributions

Characterizing the distributions of GLIs, will be accomplished through the use of Monte Carlo simulation. This method requires drawing large numbers (here, thousands) of graphs from a given distribution and obtaining the GLIs for each. The frequency with which particular values arise allows for the inference of an approximation of the true distribution for each GLI on a given region in the space of graphs.

Of critical importance to generating the GLI distribution is the nature of the relevant graph population, i.e., how are these networks created (or sampled)? Unfortunately, a general characterization of the actual probability distributions of social networks over the space of all graphs is not currently available.[7] Knowing the actual probability distribution of social networks would require a means of identifying a probability distribution over all possible graphs, which is a decidedly non-trivial problem given the nature of the space in question, and choosing a distribution which is consonant with a given substantive context. Though this is an exciting frontier area in the field of network

[7] Although some foundational work in this area has been undertaken. See, for instance, Banks and Carley (1994), Pattison et al. (1997) and Butts and Carley (1998).

analysis — in which progress is being made — we currently lack a base of reliable results from which to generalize.

Lacking a more precise criterion, a first approximation to the graph probability distribution is a uniform distribution, conditional on size and density. (That is, a distribution in which all graphs of a given size and density are equally probable.) Such a distribution has several advantages as a baseline model. First, the conditional uniform distribution here reflects prior knowledge (size and density) without making additional assumptions which could prove inadvisable in particular contexts: it is highly general without presuming total ignorance. Second, this distribution is easy to implement, permitting us to examine a greater range of conditions, and facilitating the replication, extension, and use of our methods by others. Third, the uniform distribution of graphs has been carefully characterized by others (Erdos and Renyi, 1960, Friedkin, 1981, Donninger, 1986) and the generation of the uniform graph probability distribution used here follows their method.

The set of labeled digraphs of size n and M links will be called $G_{n,M}$. Obtaining a sample from the set $G_{n,M}$ requires creating an empty graph with n vertices, then randomly inserting M links from the $n(n-1)$ possible links without replacement. This generates a digraph, and it is an unbiased sample from $G_{n,M}$. The set $G_{n,M}$ has $n(n-1)$ choose M elements. In the uniform distribution each element has an equal chance of occurring. The probability of any single given graph g being generated is therefore

$$p(g) = \binom{n(n-1)}{M}^{-1}. \tag{2}$$

The size and density parameters, n and M, are constrained in this study to take values representative of social networks. Since most social networks are small and generally sparse, n takes values of 6 to 42 in increments of 6 and the density, M/n, ranges from 0 to 5 in increments of 0.5. This range of sizes and values will be sufficient to characterize the GLIs for many networks of interest to social network research. [8] The $G_{n,M}$ distributions are produced by creating 10,000 graphs for each combination of n and M. For each of these directed graphs, the score for each of this study's six GLIs is recorded and later used to construct a probability distribution for each GLI. Thus, probability distributions for every GLI for each size and density combination are created. As an illustration, a small subset of degree centralization's distributions is shown in Fig. 1. (Estimated 5% and 95% quantiles of the GLI distributions used here, for a range of sizes and densities, are given in Appendix A.)

2.2. Characterizing GLI behavior

After creation of the GLI distributions, data interpretation becomes an issue due to the large amount of data generated. With six different GLIs, six different graph sizes, and 10 densities, there are now 360 distributions with 10,000 observations in each (for a

[8] Extension of these findings to size and density values other than those considered here would be a valuable addition to the present work; this is left as a topic for future research.

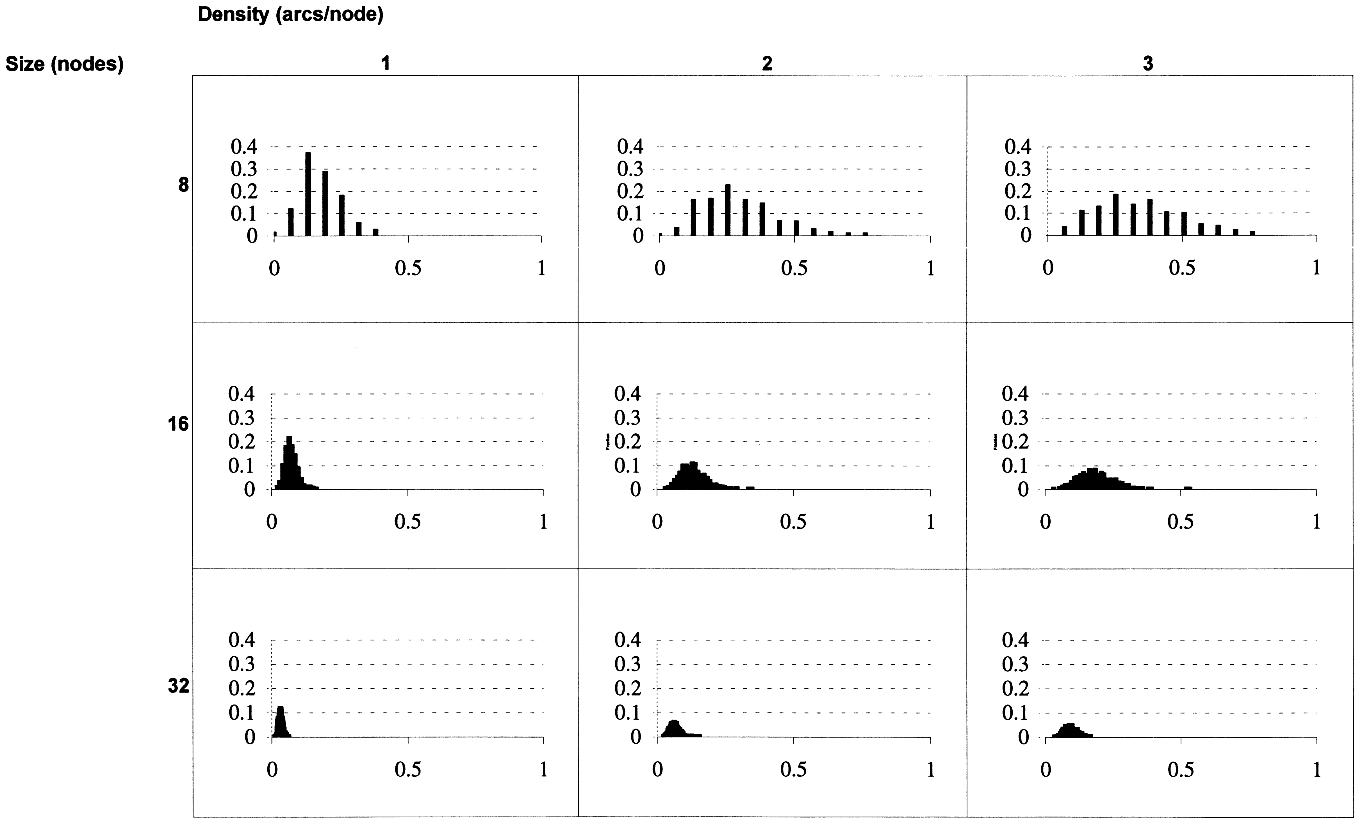

Fig. 1. Degree centralization distribution for random graphs under nine size and density conditions.

total of 3.6 million data points). Thus, each of the 360 distributions are summarized by three statistics: mean, standard deviation, and information content. The mean and standard deviation of the distribution is defined conventionally. Information content (Shannon and Weaver, 1949), a less conventional index, measures both the ‘‘discreteness’’ and the uniformity of the distribution.

More specifically, a distribution’s information content characterizes 1) the number of different values the measure can (actually) take, and 2) the proportion of the distribution taken up by each value. For instance, if a GLI can only assume two values, 0 and 1, with frequencies of 1% and 99%, respectively, then it will have a low information content. Essentially, information content is high when the GLI tends to produce many different values. [9]

The calculation of information content is a straightforward equation, given that

$$\sum_{i=1}^{k} p_i = 1 \tag{3}$$

where p_i is the proportion of the distribution that equals a particular value. In a fair coin toss, $p_1 = 0.5$, $p_2 = 0.5$, and $n = 2$, for instance. The following formula is the general formula for information:

$$I = -\sum_{i=1}^{k} p_i \log_2 p_i . \tag{4}$$

So what does information mean? In the context of information theory (Shannon and Weaver, 1949), information content literally is the expected length of an optimally encoded message transmitting the results of a draw from the distribution (e.g., ‘‘the flip was heads’’ or ‘‘the centralization was 0.32’’). For example, transmitting the result of a fair coin toss requires the transmission of 1 bit: heads or tails. If both sides of the coin are heads, of course, 0 bits are required because $p_{\text{heads}} = 1$.

For the purposes of this study, information is a rough approximation of the discreteness of a distribution. Assuming that the distribution of graphs is uniform, then high information in a GLI distribution indicates that the GLI tends to take on many values. Low information indicates that the GLI tends to take on only a few values.

Information content can also be interpreted as the expected amount of ‘‘surprise’’ in a distribution. When low information is found, we can say that the GLI for this combination of size and density is on average less surprising or informative. This would follow from the fact that the GLI is more predictable (due to a clumped distribution of probability mass).

A major caveat in interpreting these information measures is that the actual graph probability distribution for social networks may in some cases differ from the conditional uniform probability distribution. Any ‘‘clumps’’ in the graph distribution — regions of the space of graphs which are over-represented in actual social structures —

[9] A signal from the uniform distribution is of maximum information content, while one from a degenerate distribution conveys the minimum information; the information measure may thus also be thought of as a concentration measure.

will affect the ‘‘clumpiness’’ of the GLI distribution, and thereby affect the latter’s information content. [10] The information values reported in this study are thus of primary utility in understanding the behavior of GLIs across the space of graphs generally, rather than in predicting their specific behavior on social networks per se. That said, the results presented here indicate powerful trends in information content as size and density change, and suggest strong constraints on GLI behavior. While it is always possible that the particular distributions which occur in various substantive contexts deviate from these findings, then, there is good reason to believe that many of the effects noted here will generalize to other cases. [11]

2.3. Findings

The Monte Carlo results clearly indicate that variation on both size and density has powerful effects on the distribution of GLI values. Indeed, it is apparent from the distributions that GLIs of graphs with different sizes/or densities cannot be directly compared. Sampling uniformly across the space of possible structures, one can observe that graphs of different sizes and densities are strongly ‘‘predisposed’’ toward different distributions of GLI values.

Figs. 1 and 2 display selected probability distributions for degree centralization and Krackhardt hierarchy (respectively) for three different sizes and three different densities of uniformly sampled directed graphs. Each plot in Fig. 1 represents a different distribution of degree centralization for a particular size and density of uniformly generated random graphs. (For instance, the middle plot in Fig. 1 is a Monte Carlo approximation of the distribution of degree centralization when measured on digraphs drawn from a uniform sample conditional on size 16 and density 2.0.) As can be seen, the degree centralization distributions are fairly well-behaved, but the measure’s sensitivity to size is observable; as graph size is increased, both the mean and the variance of degree centralization decrease (though degree centralization is supposedly a size-normalized measure). Fig. 2 uses the same sampling distributions of graphs, but shows the distribution of Krackhardt hierarchy, which is clearly sensitive to both density and size. As density increases, mean hierarchy decreases dramatically. This is due to the fact that, as density increases, the likelihood of a pair being connected (the denominator of this hierarchy measure) increases, and the probability of a connection being asymmetric (the numerator) decreases. At extreme density values, this powerful constraining effect becomes a mathematical necessity for *any* distribution on the space of graphs. [12]

[10] From this, one can infer that any deviations between the information values given here and those on naturally occurring distributions are likely to produce lower information content in the empirical case (due to the ‘‘clumping’’ of cases). While it is in principle possible that more concentrated distributions on the space of graphs could produce less concentrated distributions on the space of GLI values, this is not probable.

[11] Thus, results on real-world data may not be radically different. Tests where the number of mutual, asymmetric, and null pairs were controlled for in the graph distribution, for instance, resulted in a similar pattern of information content.

[12] Indeed, recall that there are many fewer graphs of extreme density (high or low), and fewer graphs of small size, than there are other graphs. Such differences are a simple and direct consequence of combinatorics, but may play havoc with statistical techniques which assume consistent distributional behavior across parameters (OLS regression and Student’s *t*-test being two ubiquitous examples).

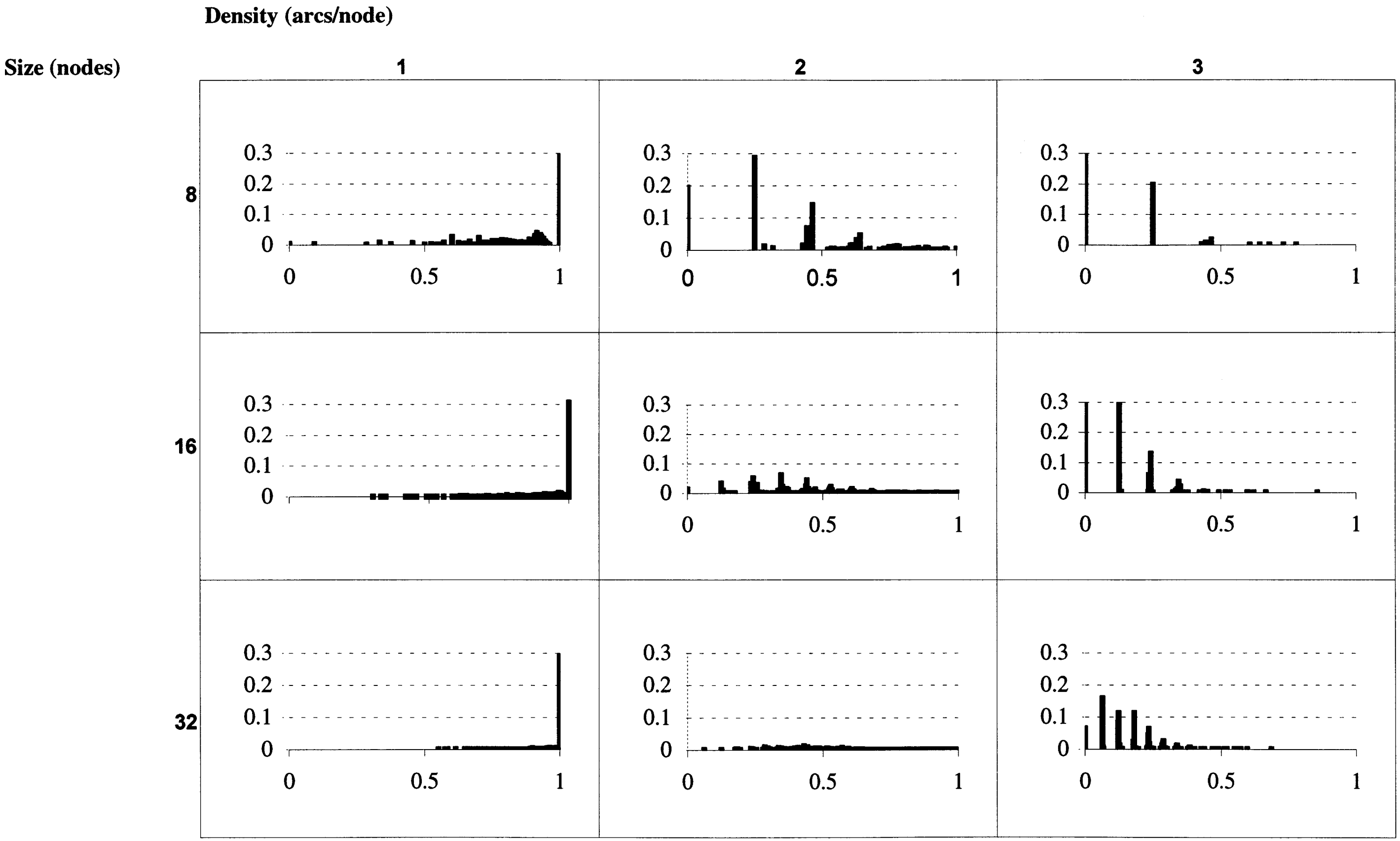

Fig. 2. Krackhardt hierarchy distribution for random graphs under nine size and density conditions.

Figs. 1 and 2 represent only 18 of the 360 distributions generated. Contour plots showing summary statistics on all of the tested distributions for each GLI are used to present the rest of the results. For each such contour plot, the horizontal axis represents density, the vertical axis represents size, and ''height'' (or shading) represents either the mean, standard deviation, or information content of the GLI (depending on the particular summary statistic in question). Fig. 3 displays the contour plots for means.[13] In these contour plots, it is clear that the mean values of GLIs are sensitive to the size and the density of the graphs they measure, but the pattern of sensitivity depends on the GLI. For instance, mean of betweeness centralization is a non-linear function of both size and density. In all cases, however, we fail to observe the stable mean which would be expected a priori if GLI behaviors were consistent across the space of graphs. Fig. 4, similarly, summarizes the standard deviations for each GLI at different sizes and densities, showing where the greatest variations in GLI value occur. Taken together, then, Figs. 3 and 4 demonstrate strong, general, and non-trivial interactions between these GLIs and both the sizes and densities of the graphs they measure. With this in mind, it is apparent that researchers confronted with graph-level data *must* account for the effect of the size of the network and the density of its relations when attempting to compare networks or draw substantive inferences on the basis of GLI values.

Referring back to Figs. 1 and 2, one will also notice that continuity is not a safe assumption to make when interpreting GLIs. This is an inherent feature of all GLIs: since there are only a finite number of graphs of a given size and density, a finite number of possible values for a GLI exist. This trait is clearly visible in the distributions of degree centralization taken from size 8 graphs (Fig. 1), as well as in many of the distributions of Krackhardt hierarchy (Fig. 2). In general, as graph size decreases, the number of possible graphs goes down dramatically, and therefore the discreteness of the distribution should tend to increase. This trend is supported by Fig. 5. Another expectation resulting from the connection between information content and number of possible graphs is that one should see reduced information (increased concentration) at both extremes of low and high densities where fewer distinct graphs are possible. This is also seen in Fig. 5, but the effect is not entirely visible in the centralization measures due to the fact that the highest densities for graphs larger than size 6 were not measured. The number of possible graphs is not the only determinant of information content, however, since the number of possible graphs is generally extremely large for almost all sizes and densities of graphs examined here. An illustration of this is in the information content plot for betweeness centralization, which represents a floor for achievable information content. None of the other GLIs possess close to the same level of information content as betweeness centralization. Clearly, different GLIs react differently to density in their information content as well. As an illustration, the information content of the Krackhardt hierarchy distribution increases as density increases, then decreases past a certain threshold. One reason for this is that when density is either low

[13] These plots are only for directed graphs. However, results for undirected graphs show the same patterns as those for directed graphs.

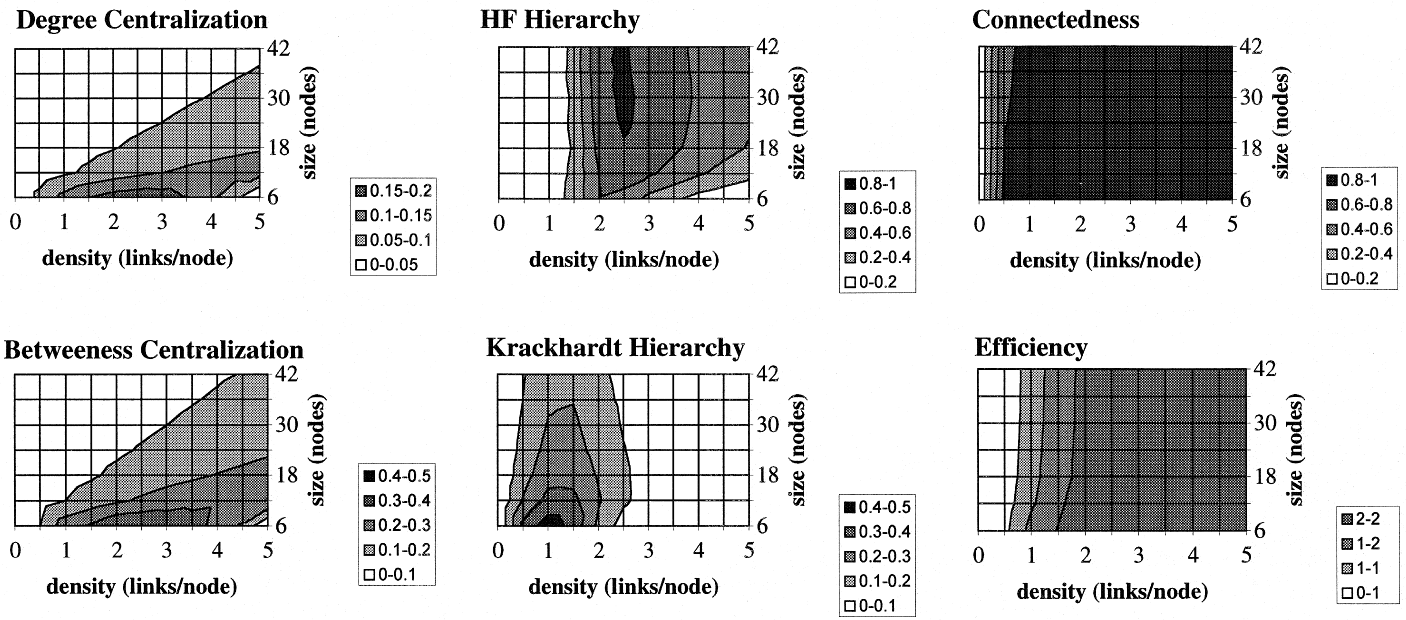

Fig. 3. Means of six GLM distributions as functions of size and density.

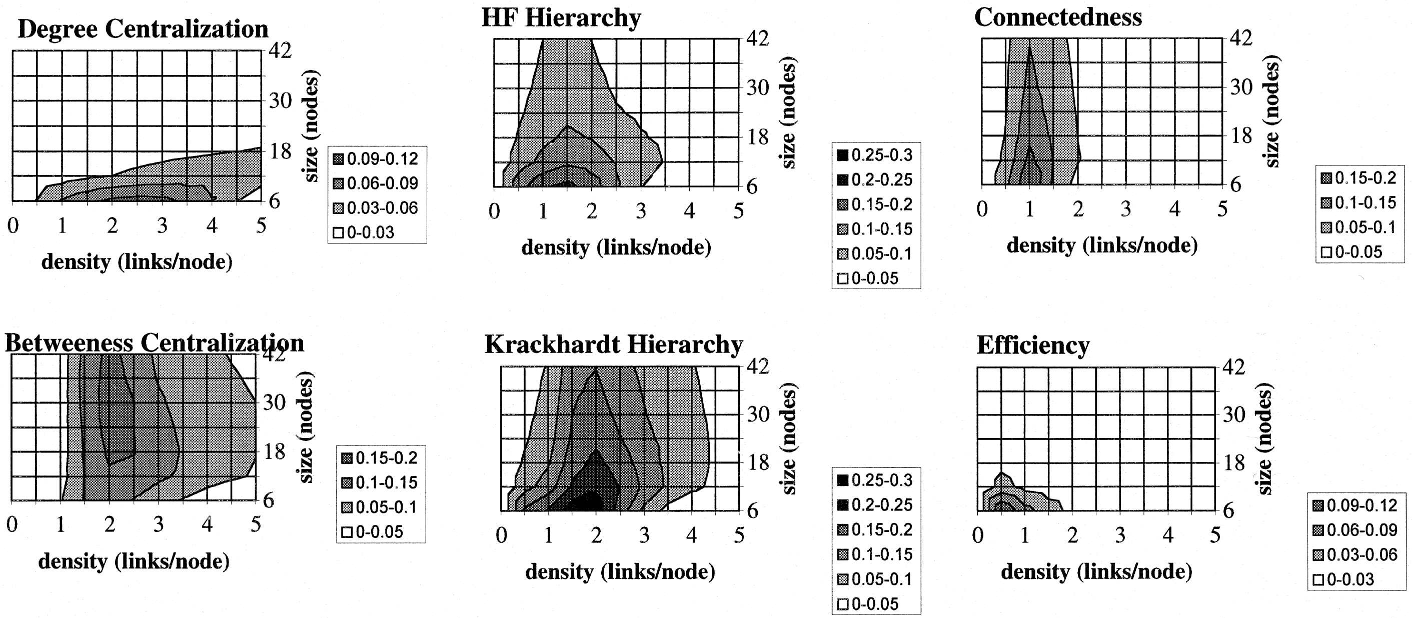

Fig. 4. Standard deviations of six GLM distributions as functions of size and density.

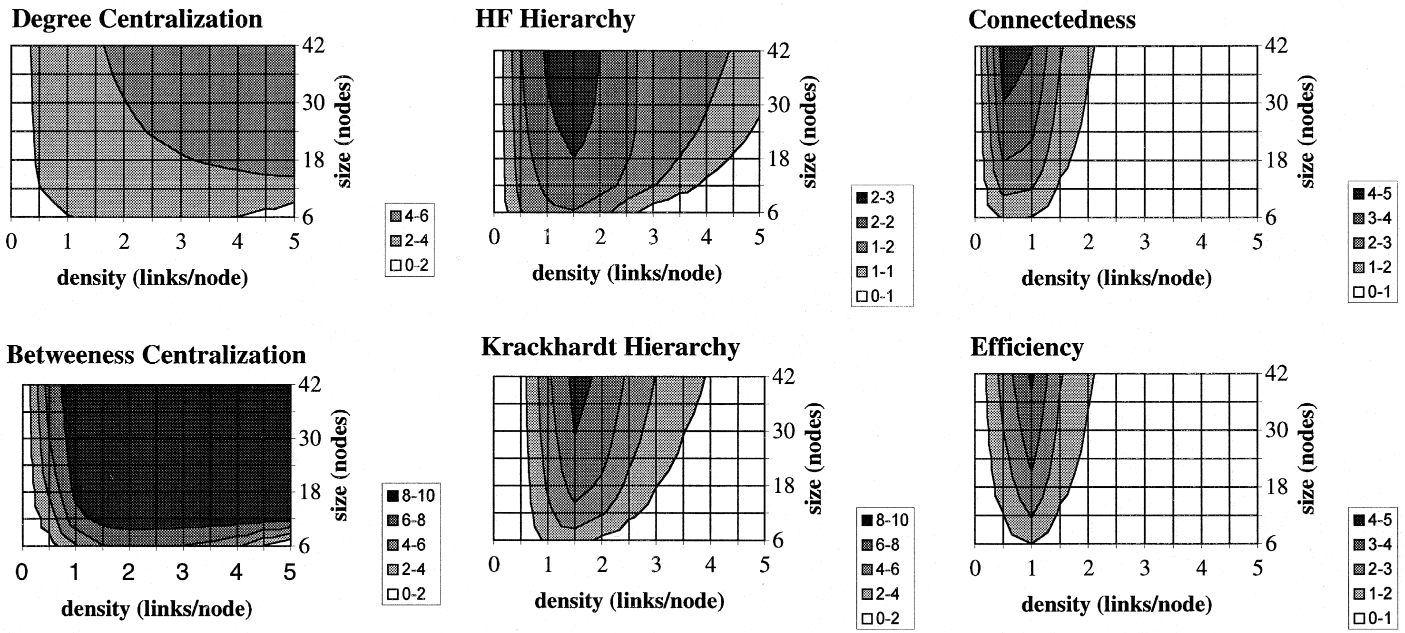

Fig. 5. Average information content (in bits) of six GLM distributions as functions of size and density.

or high, both the expected number of asymmetric pairs and the variance in number of asymmetric pairs is low, so the measure tends to take on fewer values.

For all GLIs in Fig. 5, there are areas with extremely low information and areas with very high information. These large differences indicate that measures taken from graphs of some sizes and densities may tend to take on very few values, and are thus more discrete than other distributions. It is interesting to note the dramatic differences in discreteness when measuring graphs of different sizes and densities. Once again, one should take note that the information content of measures is used here only to characterize these particular distributions, and should not be understood to be synonymous with ‘‘usefulness’’ or ‘‘informativeness’’, as the terms are generally used. Because GLIs are used both for classificatory and for substantive theoretical purposes, there may be cases in which lower levels of information content are particularly helpful to researchers, as well as situations in which more informative GLIs are more desirable. Specific discussion of the relative merits of each extreme is included below.

3. Methodological applications

In Section 2, we considered a number of GLIs, showed a simple framework for examining GLI distributions, and examined GLI behavior across graphs of varying sizes and densities. In this section, we present some simple applications of these findings to network methodology. (Quantiles for use with the null hypothesis testing procedure described below have been included in Appendix A.)

3.1. Controlling for size and density

Having shown that density and size interact powerfully with most GLIs in terms of average value, variance, and continuousness, how should the researcher interpret a GLI value? One possibility is to use the distributions of GLIs produced via a simple baseline model to define a criterion for rejection of a null hypothesis that the observed value is ‘‘typical’’ of those graphs with the aforementioned characteristics. Here, we use a Monte Carlo sampling process to generate our baseline probability distribution, and derive p-values associated with null hypothesis tests of particular GLI scores against the baseline model. Such a procedure is comparable to those employed in standard hypothesis testing, and to network methods such as the quadratic assignment procedure (Hubert, 1987; Krackhardt, 1988), and the inferences thus produced are of similar form. For instance, if one observed a degree centralization score of 0.5 for a graph on 10 vertices and 20 links, one could use this method to infer that this value is higher than the degree centralization scores of 98% of all possible graphs with the same number of vertices and links ($p < 0.02$). Using a p-value in this way to characterize GLI observations has a clear interpretation, and controls for size and density as well.

To consider the above more formally, we wish to test the following null hypothesis.

$\mathbf{H}_0$. *The observed GLI value was drawn from a distribution isomorphic to that of the GLI on the uniform distribution on $G_{n,M}$.*

To accomplish this, we employ the following algorithm.

1. Let $P_{\mathrm{H}}=0$, $P_{\mathrm{L}}=0$, $N=0$; define O_{GLI} to be the observed GLI value, α to be the desired level of significance, and $N_{\max}$ to be the maximum number of iterations.
2. Draw a graph, g, from the conditional uniform distribution on $G_{n,M}$.
3. If $\mathrm{GLI}(g) \geq O_{\mathrm{GLI}}$, increment P_{H}. If $\mathrm{GLI}(g) \leq O_{\mathrm{GLI}}$, increment P_{L}.
4. If $N < N_{\max}$, increment N and return to 2.
5. If $P_{\mathrm{H}}/N_{\max} < \alpha/2$ or $P_{\mathrm{L}}/N_{\max} < \alpha/2$, reject H_0.

(In the above, $P_{\mathrm{L}}/N_{\max}$ and $P_{\mathrm{H}}/N_{\max}$ can be interpreted as the proportion of all simulated observations less than or equal to or greater than or equal to the observed value (respectively). These can be used to infer specific p-values in the usual fashion. [14])

Note that the above does *not* imply that the generating mechanism for the *graphs* in question was necessarily random, or even that it was uniform per se; what is tested is the hypothesis that the observed GLI *value* was typical of what would be expected assuming a uniform draw from the space of graphs with the same size and density values to that observed. By comparing the actual value, then, to the tails of the GLI distribution, we obtain a p-value, which can be interpreted variously as: the probability of observing such a high (or low) value on a uniform random graph with the appropriate parameters; the probability of observing such a high (or low) value on a graph drawn from a uniform sample conditional on size and density; the proportion of all graphs with matching sizes and densities having values which are lower (or higher) than that observed; or the probability of observing such a high (or low) value from a GLI with distribution isomorphic to that of the baseline model.

As an illustration of the application of this test, Tables 1 and 2 summarize GLI attained significance levels for 20 of the binary-valued social network data sets in UCINET IV (Borgatti et al., 1991). [15]

Each cell in Tables 1 and 2 denote whether a particular GLI for a network was outside the range expected from a random graph of the same size and density. In the Padgett_marriage network from Padgett (1987) and Padgett and Ansell (1993), for example, none of the GLI values were outside of this range. Thus, one cannot reject the hypothesis that a simple random link formation model explains all this network's centralization, hierarchy, and connectedness scores. However, that assumes that the data consists of nothing but unlabeled nodes, which is not always true. Padgett in fact has access to node attributes, which allow a more sophisticated analysis. The Kapferer Tailor Shop data, from Kapferer (1972), on the other hand, have centralization scores that are all significantly large. One can infer from this that non-random link formation created the network, and that any social model explaining the data will have to produce centralization scores higher than those of a random model. Kapferer could have

[14] This procedure is standard, and is employed in well-known programs such as UCINET (Borgatti et al., 1991).

[15] Closeness centralization (Freeman, 1979; directed and simple) and common deferent (Krackhardt, 1994; directed only) were also computed for these data sets. The significant results were as follows: for closeness, Kapf_TS_instr_t1 and t2, Kapf_TS_soc_t1 and t2, Thur_office_chart, and Wiring_neg were significantly high, while Wiring_con was significantly low (0.05 or better). For common deferent, only Prison_friendship was significant (low, $p < 0.05$).

Table 1
Significance tests on undirected graphs

	Size	Density (ties per node)	Betweeness centralization	Degree centralization	Krackhardt connectedness	Krackhardt efficiency
NG_tribes_neg	16	3.8	–	–	–	–
NG_tribes_pos	16	3.8	–	–	L***	L***
Kapf_mine_multiplex	15	2.7	–	–	–	–
Kapf_mine_uniplex	15	3.6	–	–	–	–
Kapf_TS_soc_t1	39	8.2	H***	H***	–	–
Kapf_TS_soc_t2	39	11.7	H***	H***	–	–
Padgett_marriage	16	2.7	–	–	–	–
Padgett_financial	16	2.1	–	H*	–	–
Wiring_con	14	2.9	L***	H***	L***	L***
Wiring_games	14	4.3	H***	–	L***	L***
Wiring_neg	14	2.9	–	H***	–	–
Wiring_pos	14	2	–	–	–	L*
Taro_exchange	22	3.7	–	L**	–	–
Thur_office_multiplex	15	4.7	H**	H**	–	–

$^{*}p < 0.05$ H = Significantly high value observed.
$^{**}p < 0.01$ L = Significantly low value observed.
$^{***}p < 0.001$ – = Observed value not significant.

extended his analysis of the tailor shop's network as a whole by including these types of results in his study.

Appendix A contains tables of the (5%) rejection regions for all of the six GLIs used in this paper, each one generated from 10,000 data points. They are available for both directed and undirected graphs of several different sizes and densities. [16]

3.2. Extension to a general function on GLIs

The simple hypothesis test described above for differences between an observed GLI and a baseline model can easily be extended to any arbitrary function of one or more GLIs defined on a more general set of models. More precisely, consider the observation $O = f(o_1, \dots, o_n)$, where f is some function of the n GLI observations $o_i = \mathrm{GLI}_i(g_i)$. Now, let $M_1, \dots, M_n$ represent the set of n baseline models believed to serve as the generators of the distributions of $o_1, \dots, o_n$. We may then test the null hypothesis

$\mathbf{H}_0$. *The observation $O = f(o_1, \dots, o_n)$ was drawn from a distribution isomorphic to that of $f(M_1, \dots, M_n)$.*

by means of the following algorithm.

1. Let $P_{\mathrm{H}} = 0$, $P_{\mathrm{L}} = 0$, $N = 0$; define O to be the observed value of $f(o_1, \dots, o_{\mathrm{n}})$, α to be the desired level of significance, and N_{max} to be the maximum number of iterations.

[16] The two hierarchy measures and the common deferent measure are only available for directed graphs.

Table 2
Significance tests on directed graphs

	Size	Density (ties per node)	Betweeness centralization	Degree centralization	Krackhardt connectedness	Krackhardt efficiency	Krackhardt hierarchy	HFD hierarchy
Kapf_TS_Instr_t1	39	2.7	–	H***	L***	L***	–	–
Kapf_TS_Instr_t2	39	3.90	H**	H***	L***	L***	L*	–
Prison_friendship	67	2.7	H**	H*	–	–	H***	–
Wiring_help	14	1.8	–	–	L*	L*	–	–
Thur_office_chart	15	2.4	L***	H***	–	–	H***	–
Wolf_kinship	20	0.8	–	–	–	–	–	–

$^{*}p < 0.05$ H = Significantly high value observed.
$^{**}p < 0.01$ L = Significantly low value observed.
$^{***}p < 0.001$ – = Observed value not significant.

2. Draw a set of graphs, $g_1, \ldots, g_n$, from the distributions $M_1, \ldots, M_n$ (respectively).
3. If $f(\text{GLI}_1(g_1), \ldots, \text{GLI}_n(g_n)) \geq O$, increment P_{H}. If $f(\text{GLI}_1(g_1), \ldots, \text{GLI}_n(g_n)) \leq O$, increment P_{L}.
4. If $N < N_{\max}$, increment N and return to 2.
5. If $P_{\text{H}}/N_{\max} < \alpha/2$ or $P_{\text{L}}/N_{\max} < \alpha/2$, reject H_0.

This general procedure can be used to test arbitrary functions of GLIs with arbitrary baseline models; [17] it is not even necessary to assume the independence of $g_1, \ldots, g_n$, so long as all dependencies are specified within the model set. [18] While this permits a wide range of useful applications, we shall here present as an example the simple case of differences between GLIs on two graphs drawn independently from the uniform baseline model. Let G_1 and G_2 be the two observed graphs, with edge sets denoted by $E(G)$ and vertex sets denoted by $V(G)$. M_1 is then equal to the conditional uniform distribution on $G_{|V(G_1)|,|E(G_1)|/(|V(G_1)|2-|V(G_1)|)}$, and M_2 is equal to the conditional uniform distribution on $G_{|V(G_2)|,|E(G_2)|/(|V(G_2)|2-|V(G_2)|)}$. For some GLI, then, we are interested in testing the hypothesis that $O = f(G_1,G_2) = \text{GLI}(G_1) - \text{GLI}(G_2)$ was drawn from a distribution isomorphic to that which would result from conditional uniform selection of G_1 and G_2 alone. To perform the test, we simply use the above algorithm, inserting the above definitions as appropriate. A significant result, if one occurs, suggests that the difference in GLI values deviates substantially from that which would be expected from size and density effects alone; failure to reject the null hypothesis, by contrast, indicates that we cannot rule out the notion that a difference of the observed magnitude and direction could have been produced by the baseline model.

[17] Indeed, the basic algorithm provides a generalized Monte Carlo based hypothesis testing procedure which can in principle be applied to almost any sort of data for which baseline models can be identified; this broader usage is beyond the scope of this paper, however.

[18] For instance, M_1 might be the conditional uniform distribution on $G_{2,0.5}$, and M_2 might be g_1, the realization of M_1. Thus, it is possible to test for significant differences between multiple GLIs on the same set of relations.

3.3. Generating baseline-controlled variables

In addition to simple hypothesis testing, it is often desirable to have variables which reflect the variability of a particular graph-level measure independent of baseline effects. While (as has been noted) GLIs are inherently non-independent, it is possible in some cases to construct artificial variables which control out at least some effects of other GLIs (such as size and density) with respect to a particular baseline model.

One family of such baseline-controlled variables consists of the quantities P_L/N_{max} and P_H/N_{max} described in the algorithms above. Interpretable as the (estimated) proportions of all graphs under the baseline model with values lesser/greater than or equal to the observed value (respectively), these variables are both standardized to the [0,1] interval and dependent only conditionally on size, density, and the baseline model. As such, they may be used in other analyses in which a measure of relative GLI magnitude is desired (provided that the proportional interpretation is contextually reasonable). It should be noted, however, that several caveats apply to the use of the normalized P_L/P_H scores as variables in traditional statistical analyses. First, and foremost, the distributional properties of these measures are not well-understood at this time; given what we have already seen of GLI distributions, however, there is every reason to believe that these variables will be poorly behaved (and thus possibly unsuitable for some applications [19]). Secondly, it must be remembered that the values which are obtained are not without error, and hence it is inadvisable to base a detailed analysis on normalized P scores for which N_{max} is small. Finally, it is important to reiterate that the use of normalized P scores as variables in and of themselves depends upon the applicability of the interpretation of the variables to the problem at hand: where one is interested in comparing (or otherwise making use of) the *relative* proportions of suitably conditioned graphs with GLI values above/below those observed, these variables are quite appropriate. If one is interested in absolute GLI values or the like, or if one's procedures demand variables which are known to be statistically well-behaved, then one must look to other options.

4. Discussion

As we have seen, both size and density have powerful — and complex — interactions with other GLIs. These interactions stem from fundamental constraints on the space of graphs, constraints that severely limit the combinations of GLI values which can be realized on a given graph. Across the space of graphs, such constraints further alter GLI distributions, causing some values to be vastly more common than others and to generally affect the ranges of realizations which are possible in particular regions of the space. Unlike many interactions familiar to data analysts in the social sciences, these

[19] E.g., these quantities should *not* be used as dependent variables in OLS regressions, due to the fact that there is no reason to assume that they follow a normal perturbation pattern.

interactions are both *fundamental* and *non-trivial*: generally speaking, they can neither be removed through judicious experimental design, nor can they be accounted for simply by adding a covariate to a regression equation. Data analysis in the presence of such fundamental interaction effects often requires new procedures, and demands that researchers pay close attention to the theoretical foundations of their choice of methods. Given this, we shall now briefly consider some of the broader implications of our findings for social network analysis.

4.1. Size and density as explanatory variables

As has been demonstrated, many commonly used GLIs are ill-behaved in general, and their distributions vary substantially across various levels of size and density. Does this mean, however, that such researchers should immediately ‘‘control out’’ the effects of size and density, or else abandon GLIs altogether? Indeed, it does not. Many GLIs — Krackhardt connectedness, betweenness centralization, and H–F hierarchy, for instance — are motivated by substantive theoretical concerns, and are postulated to capture aspects of social structure which are pertinent to the prediction of empirical phenomena. While it may be the case that such measures are strongly related to the size and density of social structures, this in no way invalidates their relationship to phenomena of interest; it merely suggests that size and density will be critical determinants of social phenomena.

By way of example, it is useful to consider the role of number of components [20] in understanding the behavior of epidemiological networks. Clearly, total population exposure to disease is related to this graph-level measure, as diseases cannot cross between components; the number of components in a graph, however, is well-known to be heavily influenced by density. Should a researcher seeking to assess population exposure vis-a-vis an epidemiological network then seek to control out the effects of density before examining component count? Clearly not: it is the number of components per se which are of interest to the researcher in this case, not the number of components *relative* to some baseline model. On the other hand, if said researcher were interested in determining whether some particular factor *accounted* for the particular number of observed components, he or she would be quite wise to consider first the effects of size and density under a simple baseline model (such as that chosen here) before resorting to more esoteric explanations. In the former case, the goal of the researcher is to extract particular information from a graph’s structure; in the latter, size and density are themselves important explanatory variables, whose effects must be accounted for prior to consideration of secondary influences.

The basic notion that population size and social density are critical determinants of social phenomena is not a new one; indeed, it was one of sociology’s first critical insights (Spencer, 1874; Durkheim, 1893, 1897). What has been less clear, however, is

[20] Obviously, this is a graph-level measure, though it may not always be thought of as such by social network analysts.

the manner in which these determinants operate. While earlier theories, such as those of Durhkeim and Spencer, emphasized such notions as ritual solidarity, substitutability, and opportunities for cooperation or aggression, modern structural theory has been more concerned with subtle structural features such as paths of information flow (Festinger et al., 1950; Burt, 1987), positions with ''bargaining power'' over others (Emerson, 1972; Cook et al., 1983), and the presence of ordered relations (i.e., hierarchy) (Krackhardt, 1994; Hummon and Fararo, 1995). Our current findings, then, bring us full circle: we can now understand how the less obvious structural features which define social reality are themselves constrained by the size and density of populations. Since size and density are, in turn, closely related to the constraints of demographics and physical space (Latané et al., 1994; Latané, 1996), this realization presents us with an exciting opportunity to construct general theories of society which are founded on fundamental physical principles, without losing sight of the insights of network analysis. [21]

The strength of these GLI constraints, of course, varies both across measures and across the space of graphs. In some cases, regions of graph space exist within which a great deal of variance is maintained; in others, GLI values are highly predictable for most graphs. In the former areas of ''low information content'', theorists should be aware that GLI variability is strongly connected to variability in size and density, while in the latter areas of ''high information content'', theorists should be aware of the reverse. These facts are neither ''good'' nor ''bad''; they neither indicate that various GLIs are ''broken'' nor ''sensible''. They are merely realities of the fundamental nature of structure, which must be taken into account.

4.2. GLIs as classificatory tools

The same issues, of course, pertain to the use of GLIs for purposes of graph classification. As we have seen, many GLIs are highly concentrated over the space of graphs; such ''low information content'' GLIs are unsuitable for general classification, though they may be useful in special cases. When attempting to discern differences between graphs using GLI comparisons, it *is* possible to use baseline model controls to account for expected differences due to size and density, although it is important not to use such controls inappropriately. [22] The same is true for more complex functions of GLIs, and control under any number of baseline models is possible. This, of course, suggests that a hierarchical approach to graph characterization utilizing hypothesis tests on baseline models of increasing complexity may be useful; similar approaches have been taken by Wasserman and Pattison (1996) and Pattison et al. (1997), and may provide fertile ground for methodological cross-over.

[21] A theoretical program which was strongly endorsed by Mayhew (1980; 1981) among others.

[22] Controlling for size and density effects is appropriate for theories which postulate differences in GLIs (or effects therefrom) above and beyond the baseline; theories which merely postulate differences in or effects from GLIs themselves (and which do not stipulate the origin of these differences or effects) are not usually candidates for such a procedure.

4.3. Baseline GLI distributions as Bayesian priors

Another line of research which could potentially both inform and benefit from this current work is that of Bayesian methods for network analysis. The most obvious such application would be the use of baseline GLI distributions (such as those described here) as priors for purpose of GLI estimation under conditions of measurement error, limited information, or from network samples. [23] Such priors are, as we have seen, much more reasonable than an assumption of uniform GLI distributions, and are logically sensible as reflecting conditional ignorance regarding graph structure. [24] A still more interesting use of GLI distributions for Bayesian network analysis would be in the inference of the distribution of graphs from GLI distributions themselves. This would be particularly useful in cases in which GLIs on graphs may be subject to estimation, but graph structure is not directly observable; by using Bayesian methods, it is in principle possible to estimate a probability function over the space of graphs from such data, which would permit identification of maximum probability structures given GLI observations. Currently, such applications of the current work on GLI distributions pose formidable computational obstacles. In the long run, however, they may prove to be powerful tools for social network analysis.

4.4. Some caveats

While we have uncovered important — and fundamental — relationships between GLIs, some caveats bear reiteration. One limitation of this study, as has been indicated, is that it relies on the baseline model of conditional uniform selection across graph space. While this is not entirely problematic, the fact remains that actual social networks may in reality occupy a small region of graph-space that is not well-represented by a uniform sampling strategy. Indeed, in some cases one can even imagine perverse distributions of social networks in which the relationships between size, density, and other GLIs are the exact opposite of those found in this study (though this is clearly impossible for some results — connectedness will always be high in dense graphs, for instance, no matter how they are selected).

One response the above objection is that it is unlikely that social networks have a distribution constrained enough to invalidate the observed relationships, since the results of Tables 1 and 2 show that for at least 20 actual social networks observed, the GLIs are not significantly different those expected assuming the uniform distribution on multiple measures. [25] Another, which has already been touched upon, is that there are limits to

[23] That is, for situations in which one wishes to infer a statistic regarding the GLIs of a particular population of graphs, having observed only some subset of those graphs (e.g., attempting to determine the average centralization of university administrative offices by observing the centralization scores of a sample of such offices).

[24] I.e., an uninformed prior distribution vis-a-vis the probability of particular structures (conditional, perhaps, on factors such as size and/or density) necessarily implies a non-uniform prior distribution vis-a-vis GLI values. As we have seen, a uniform prior on GLI values is almost never sensible.

[25] Likewise, recent work on algorithmic complexity of social networks suggests that other, more subtle deviations from random structures are uncommon (Butts, 1999).

the degree that even perverse distributions can alter the relationships between size, density, and other GLIs. Since the combinatorics of graph structure dictate the (highly variable) number of graphs which are possible for each size and density condition, as well as the number of graphs which can exist with particular GLI values, there will often be whole ranges of GLI values which are not even *possible* for given size and density levels. Such effects cannot be altered by modifying the distribution by which graphs are selected from the relevant regions of graph space.

Even given that the uniform baseline model may be useful for many purposes, we should not be blinded to the possibility of alternatives. The usage of distributions other than uniform has become increasingly common in social network research; in particular, distributions that condition on reciprocity, in- and out-degrees, transitivity, and other such algebraic constraints are currently being studied (Pattison et al., 1997). A next step in developing this line of research will be testing existing social network data using graph distributions that incorporate some of the known properties of social networks other than size and density. The work by Fararo and Skvoretz (1984; 1987), Skvoretz (1990), Snijders (1991), and Pattison et al. (1997) on random networks has already laid the groundwork for generating these distributions.

5. Conclusion

Using a simplifying assumption regarding graph distribution, this study shows that size and density strongly interact with all graph-level measures (GLIs) examined. Graphs with different sizes and/or densities will often have dramatically different probability distributions for the same GLI, and thus their GLI values will have different interpretations. Because of this, it may be difficult for the researcher to know whether a specific GLI value is the result of a direct structural social phenomenon or simply a secondary effect of the network's size and density.

One partial solution to this problem is to control for size and density via a hypothesis test, which uses a distribution generated from random networks of the same size and density as the observed network. This hypothesis test was demonstrated on 20 of the data sets in UCINET IV (Borgatti et al., 1991) and showed that many of the GLI values measured were not significantly different from the uniform baseline model. The testing framework in question can be generalized to arbitrary functions of GLIs across graphs, and to arbitrary baseline models (which may differ by observation), without adding additional complexity to the base algorithm.

Finally, we have seen that the interaction of size and density with other GLIs is a sort of double-edged sword for the network analyst. On the one hand, the presence of strong constraints on substantively meaningful structural features suggests the potential power of structural theories rooted in demographic and spatial processes. Such constraints also imply that measuring size and density of social structures — features which are relatively amenable to estimation using network sampling techniques — may provide us with a great deal of information from which to construct theory. On the other hand, size and density interactions pose problems for the use of GLIs as classificatory tools, and play havoc with traditional techniques of statistical inference.

Acknowledgements

This work was supported in part by the Center for the Computational Analysis of Social and Organizational Systems, the Institute for Complex Engineered Systems, and by grant no. N00014-97-1-0037 from the Office of Naval Research (ONR), United States Navy.

Appendix A

The tables below present Monte Carlo-generated threshold values for two-tailed 0.05 significance, based on 10,000 simulations for each size/density combination. The upper number is the value that the observed GLI must be *higher than* for 0.05 significance, the lower number is the value that the observed GLI must be *lower than* for significance at the 0.05 level. Note that if the upper number is 1.0 or the lower number is 0.0, then the corresponding tail of the estimated distribution is degenerate (which may also be true of the actual distribution, in which case one *cannot* observe a result with the corresponding significance level).

Example: One observes two simple graphs of size 12, the densities 2 and 4 and betweeness centralizations of 0.7 and 0.65, respectively. The rejection region for a size 12 density 2 graph (for betweeness centralization) is given below as $C_B < 0.072$, $C_B > 0.874$; hence, the observed betweeness centralization score is not significant at the 0.05 level. For the second graph, however, the rejection region is $C_B < 0.098$, $C_B >$ 0.533. Since $0.65 > 0.533$, we observe that the betweeness centralization of the second graph is significant at the 0.05 level.

A.1. Binary symmetric graphs

GLI	Size	Density									
		0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Betweeness centralization	6	0.000	0.090	0.280	0.810	0.688	0.360	0.288	0.144	0.072	0.000
		0.000	0.000	0.016	0.048	0.064	0.040	0.048	0.040	0.016	0.000
	12	0.002	0.063	0.349	0.874	1.009	0.838	0.673	0.533	0.427	0.349
		0.000	0.001	0.011	0.072	0.129	0.137	0.121	0.098	0.077	0.054
	18	0.000	0.042	0.305	0.948	1.079	0.937	0.786	0.629	0.551	0.442
		0.000	0.000	0.009	0.098	0.162	0.178	0.166	0.145	0.134	0.114
	24	0.001	0.044	0.360	0.998	1.087	0.904	0.754	0.637	0.539	0.479
		0.000	0.000	0.012	0.117	0.188	0.204	0.193	0.174	0.157	0.139
	30	0.001	0.038	0.339	1.014	1.048	0.868	0.735	0.627	0.546	0.476
		0.000	0.000	0.010	0.136	0.207	0.219	0.209	0.188	0.172	0.157
	36	0.001	0.035	0.390	1.031	0.976	0.820	0.697	0.597	0.525	0.462
		0.000	0.000	0.014	0.151	0.218	0.227	0.216	0.199	0.180	0.166
	42	0.000	0.031	0.366	1.043	0.958	0.776	0.672	0.569	0.501	0.447
		0.000	0.000	0.012	0.162	0.223	0.238	0.225	0.203	0.187	0.168

GLI	Size	Density									
		0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Degree centralization	6	0.100	0.450	0.550	0.750	0.850	0.750	0.700	0.450	0.250	0.000
		0.100	0.150	0.100	0.150	0.100	0.150	0.100	0.150	0.100	0.000
	12	0.076	0.175	0.251	0.305	0.360	0.415	0.447	0.480	0.513	0.524
		0.033	0.044	0.076	0.087	0.098	0.109	0.120	0.131	0.142	0.131
	18	0.043	0.099	0.142	0.182	0.217	0.256	0.283	0.323	0.341	0.372
		0.018	0.033	0.051	0.066	0.076	0.091	0.101	0.108	0.118	0.124
	24	0.034	0.069	0.099	0.129	0.160	0.185	0.211	0.237	0.254	0.280
		0.017	0.026	0.039	0.052	0.065	0.073	0.082	0.091	0.099	0.108
	30	0.023	0.050	0.073	0.100	0.121	0.143	0.166	0.185	0.202	0.224
		0.012	0.024	0.034	0.042	0.052	0.063	0.071	0.079	0.086	0.095
	36	0.020	0.041	0.061	0.080	0.099	0.117	0.134	0.151	0.168	0.185
		0.010	0.020	0.029	0.037	0.047	0.053	0.063	0.069	0.077	0.085
	42	0.017	0.035	0.050	0.067	0.082	0.099	0.112	0.128	0.141	0.155
		0.009	0.018	0.025	0.033	0.040	0.049	0.055	0.061	0.069	0.076
Krackhardt connectedness	6	0.067	0.400	0.667	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.067	0.200	0.400	0.667	0.667	1.000	1.000	1.000	1.000	1.000
	12	0.091	0.318	0.682	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.045	0.121	0.273	0.470	0.682	0.833	0.833	1.000	1.000	1.000
	18	0.046	0.242	0.595	0.889	1.000	1.000	1.000	1.000	1.000	1.000
		0.026	0.092	0.216	0.471	0.686	0.784	0.889	0.889	0.889	1.000
	24	0.058	0.239	0.558	0.841	1.000	1.000	1.000	1.000	1.000	1.000
		0.022	0.076	0.214	0.471	0.688	0.764	0.837	0.917	0.917	0.917
	30	0.039	0.211	0.531	0.869	1.000	1.000	1.000	1.000	1.000	1.000
		0.018	0.064	0.189	0.487	0.690	0.807	0.869	0.871	0.933	0.933
	36	0.038	0.194	0.517	0.838	0.944	1.000	1.000	1.000	1.000	1.000
		0.016	0.057	0.195	0.487	0.690	0.787	0.840	0.890	0.944	0.944
	42	0.029	0.168	0.505	0.816	0.952	1.000	1.000	1.000	1.000	1.000
		0.013	0.051	0.178	0.504	0.691	0.816	0.861	0.906	0.952	0.952
Krackhardt efficiency	6	1.000	1.000	1.000	0.900	0.800	0.600	0.500	0.300	0.200	0.000
		1.000	1.000	0.667	0.667	0.500	0.600	0.500	0.300	0.200	0.000
	12	1.000	1.000	1.000	0.982	0.927	0.873	0.818	0.764	0.709	0.655
		1.000	0.750	0.867	0.857	0.833	0.822	0.756	0.764	0.709	0.655
	18	1.000	1.000	1.000	0.983	0.963	0.926	0.897	0.860	0.831	0.794
		1.000	0.889	0.917	0.924	0.912	0.886	0.875	0.833	0.800	0.794
	24	1.000	1.000	1.000	0.990	0.972	0.949	0.925	0.901	0.877	0.854
		1.000	0.923	0.949	0.949	0.936	0.921	0.900	0.887	0.861	0.835
	30	1.000	1.000	1.000	0.991	0.980	0.961	0.943	0.924	0.906	0.887
		1.000	0.944	0.963	0.963	0.953	0.942	0.929	0.909	0.897	0.876
	36	1.000	1.000	1.000	0.992	0.980	0.968	0.953	0.938	0.923	0.908
		1.000	0.957	0.973	0.971	0.963	0.951	0.940	0.926	0.916	0.900
	42	1.000	1.000	1.000	0.992	0.985	0.973	0.961	0.948	0.935	0.922
		1.000	0.967	0.978	0.977	0.970	0.961	0.950	0.939	0.931	0.917

A.2. Binary directed graphs

Statistic	Size	Density									
		0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Betweeness centralization	6	0.016	0.142	0.328	0.405	0.352	0.250	0.166	0.088	0.036	0.000
		0.000	0.003	0.011	0.028	0.039	0.027	0.016	0.010	0.006	0.000
	12	0.003	0.075	0.311	0.481	0.493	0.425	0.354	0.288	0.228	0.177
		0.000	0.001	0.009	0.036	0.082	0.089	0.075	0.058	0.045	0.033
	18	0.002	0.054	0.314	0.529	0.560	0.487	0.403	0.333	0.280	0.232
		0.000	0.001	0.008	0.044	0.115	0.124	0.110	0.092	0.076	0.064
	24	0.001	0.046	0.323	0.564	0.581	0.492	0.405	0.343	0.289	0.246
		0.000	0.000	0.008	0.057	0.142	0.147	0.127	0.110	0.095	0.083
	30	0.001	0.039	0.342	0.587	0.584	0.482	0.392	0.334	0.283	0.254
		0.000	0.000	0.008	0.074	0.160	0.159	0.139	0.121	0.105	0.091
	36	0.000	0.035	0.339	0.586	0.552	0.453	0.380	0.322	0.279	0.244
		0.000	0.000	0.009	0.090	0.172	0.166	0.145	0.127	0.110	0.097
	42	0.000	0.032	0.334	0.594	0.545	0.435	0.368	0.308	0.268	0.235
		0.000	0.000	0.009	0.105	0.179	0.170	0.151	0.129	0.113	0.101
Degree centralization	6	0.137	0.240	0.309	0.343	0.377	0.343	0.309	0.240	0.137	0.000
		0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.000
	12	0.036	0.068	0.104	0.131	0.153	0.171	0.189	0.203	0.212	0.212
		0.009	0.018	0.027	0.032	0.036	0.041	0.045	0.050	0.050	0.050
	18	0.020	0.040	0.059	0.076	0.091	0.106	0.120	0.131	0.143	0.152
		0.007	0.013	0.020	0.025	0.030	0.035	0.039	0.042	0.047	0.049
	24	0.014	0.028	0.041	0.053	0.065	0.075	0.086	0.096	0.106	0.114
		0.005	0.011	0.016	0.021	0.024	0.029	0.033	0.037	0.040	0.044
	30	0.011	0.021	0.031	0.040	0.050	0.058	0.067	0.075	0.083	0.090
		0.005	0.009	0.013	0.017	0.021	0.024	0.028	0.031	0.035	0.038
	36	0.009	0.017	0.025	0.033	0.040	0.047	0.054	0.061	0.068	0.073
		0.004	0.008	0.011	0.015	0.018	0.022	0.025	0.028	0.031	0.034
	42	0.007	0.014	0.020	0.027	0.033	0.040	0.046	0.052	0.057	0.063
		0.004	0.007	0.010	0.013	0.016	0.019	0.022	0.025	0.028	0.030
HFD hierarchy	6	0.600	0.800	0.800	0.400	0.200	0.200	0.000	0.000	0.000	0.000
		0.200	0.200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	12	0.364	0.545	0.545	0.455	0.273	0.182	0.091	0.091	0.091	0.000
		0.091	0.182	0.091	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	18	0.235	0.412	0.471	0.353	0.235	0.176	0.118	0.059	0.059	0.059
		0.118	0.176	0.118	0.059	0.000	0.000	0.000	0.000	0.000	0.000
	24	0.217	0.348	0.391	0.304	0.217	0.130	0.087	0.087	0.043	0.043
		0.087	0.130	0.130	0.087	0.043	0.000	0.000	0.000	0.000	0.000
	30	0.172	0.310	0.379	0.276	0.172	0.103	0.103	0.069	0.034	0.034
		0.069	0.103	0.103	0.069	0.034	0.000	0.000	0.000	0.000	0.000
	36	0.171	0.286	0.314	0.229	0.143	0.114	0.086	0.057	0.057	0.029
		0.057	0.114	0.086	0.057	0.029	0.029	0.000	0.000	0.000	0.000
	42	0.146	0.268	0.317	0.220	0.122	0.098	0.073	0.049	0.049	0.024
		0.049	0.098	0.098	0.049	0.049	0.024	0.000	0.000	0.000	0.000

Statistic	Size	Density									
		0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Krackhardt connectedness	6	0.400	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.200	0.467	0.667	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	12	0.318	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.106	0.439	0.697	0.833	1.000	1.000	1.000	1.000	1.000	1.000
	18	0.242	0.889	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.085	0.438	0.784	0.889	0.889	1.000	1.000	1.000	1.000	1.000
	24	0.210	0.837	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.072	0.446	0.761	0.917	0.917	1.000	1.000	1.000	1.000	1.000
	30	0.211	0.809	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.062	0.455	0.754	0.869	0.933	1.000	1.000	1.000	1.000	1.000
	36	0.192	0.838	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.056	0.479	0.787	0.890	0.944	0.944	1.000	1.000	1.000	1.000
	42	0.168	0.816	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.050	0.482	0.775	0.906	0.952	0.952	1.000	1.000	1.000	1.000
Krackhardt efficiency	6	1.000	0.960	0.840	0.720	0.600	0.480	0.360	0.240	0.120	0.000
		0.750	0.800	0.688	0.720	0.600	0.480	0.360	0.240	0.120	0.000
	12	1.000	0.992	0.942	0.893	0.843	0.793	0.744	0.694	0.645	0.595
		0.889	0.927	0.902	0.860	0.843	0.793	0.744	0.694	0.645	0.595
	18	1.000	0.992	0.965	0.934	0.903	0.872	0.841	0.810	0.779	0.747
		0.944	0.959	0.947	0.922	0.887	0.872	0.841	0.810	0.779	0.747
	24	1.000	0.993	0.975	0.953	0.930	0.907	0.885	0.862	0.839	0.817
		0.966	0.974	0.960	0.946	0.921	0.907	0.885	0.862	0.839	0.817
	30	1.000	0.996	0.981	0.963	0.945	0.927	0.910	0.892	0.874	0.856
		0.975	0.981	0.971	0.955	0.940	0.927	0.910	0.892	0.874	0.856
	36	1.000	0.996	0.984	0.970	0.955	0.940	0.926	0.911	0.896	0.882
		0.980	0.985	0.976	0.964	0.952	0.936	0.926	0.911	0.896	0.882
	42	1.000	0.996	0.987	0.974	0.962	0.949	0.937	0.924	0.912	0.899
		0.984	0.988	0.980	0.970	0.959	0.946	0.937	0.924	0.912	0.899
Krackhardt hierarchy	6	1.000	1.000	1.000	0.800	0.333	0.333	0.000	0.000	0.000	0.000
		0.500	0.400	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	12	1.000	1.000	1.000	0.883	0.613	0.318	0.167	0.167	0.167	0.000
		0.833	0.655	0.345	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	18	1.000	1.000	0.990	0.885	0.613	0.393	0.216	0.111	0.111	0.111
		0.889	0.734	0.441	0.211	0.000	0.000	0.000	0.000	0.000	0.000
	24	1.000	1.000	0.992	0.866	0.554	0.371	0.239	0.163	0.083	0.083
		0.923	0.779	0.506	0.236	0.083	0.000	0.000	0.000	0.000	0.000
	30	1.000	1.000	0.993	0.848	0.544	0.357	0.248	0.131	0.129	0.067
		0.938	0.810	0.524	0.257	0.129	0.000	0.000	0.000	0.000	0.000
	36	1.000	1.000	0.991	0.823	0.529	0.347	0.213	0.161	0.110	0.056
		0.941	0.833	0.551	0.278	0.110	0.056	0.000	0.000	0.000	0.000
	42	1.000	1.000	0.989	0.804	0.510	0.340	0.223	0.139	0.094	0.093
		0.939	0.850	0.574	0.308	0.138	0.048	0.000	0.000	0.000	0.000

References

Aldrich, H., 1978. Centralization versus decentralization in the design of human service delivery systems: a response to Gouldner's lament. In: Sarri, R., Hasenfield, Y. (Eds.), The Management of Human Services. Columbia Univ. Press, New York, pp. 51–79.

Banks, D., Carley, K.M., 1994. Metric inference for social networks. Journal of Classification 11, 121–149.

Bernard, H.R., Johnsen, E.C., Killworth, P.D., Robinson, S., 1991. Estimating the size of an average personal network and of an event subpopulation: some empirical results. Social Science Research 20, 109–121.

Blau, P.M., 1972. Interdependence and hierarchy in organizations. Social Science Research 1, 1–24.

Borgatti, S.P., Everett, M.G., Freeman, L.C., 1991. UCINET, Version IV. Analytic Technology, Columbia, SC.

Burt, R.S., 1987. Social contagion and innovation, cohesion versus structural equivalence. American Journal of Sociology 85, 892–925.

Butts, C., 1999. The complexity of social networks: theoretical and empirical findings. CASOS Working Paper. Carnegie Mellon University.

Butts, C., Carley, K.M., 1998. Canonical labeling to facilitate graph comparison. CASOS Working Paper. Carnegie Mellon University.

Carley, K.M., 1992. Organizational learning and personnel turnover. Organization Science 3 (1).

Cohen, A.M., 1962. Changing small-group communication networks. Administrative Science Quarterly 6, 443–462.

Cook, K.S., Emerson, R.M., Gillmore, M.R., Yamagishi, T., 1983. The distribution of power in exchange networks. American Journal of Sociology 89, 275–305.

Donninger, C., 1986. The distribution of centrality in social networks. Social Networks 8, 191–203.

Durkheim, E., 1893. The Division of Labor in Society. Free Press, New York.

Durkheim, E., 1897. Suicide. Free Press, New York.

Emerson, R.M., 1972. Exchange theory: Part II. Exchange relations and network structures. In: Berger, J., Zelditch, M., Jr., Anderson, B. (Eds.), Sociological Theories in Progress. Houghton Mifflin, New York.

Erdos, P., Renyi, 1960. On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 17–61.

Fararo, T., Skvoretz, J., 1984. Biased networks and social structure theorems: Part II. Social Networks 6, 223–258.

Fararo, T., Skvoretz, J., 1987. Unification research programs: integrating two structural theories. American Journal of Sociology 92, 1183–1209.

Festinger, L., Schachter, S., Back, K.W., 1950. Social Pressures in Informal Groups. Stanford Univ. Press, Stanford.

Freeman, L.C., 1977. A set of measures of centrality based on betweeness. Sociometry 40, 35–41.

Freeman, L.C., 1979. Centrality in social networks: conceptual clarification. Social Networks 1, 215–239.

Friedkin, N.E., 1981. The development of structure in random networks: an analysis of the effects of increasing network density on five measures of structure. Social Networks 3, 41–52.

Gould, R.V., 1987. Measures of betweeness in non-symmetric networks. Social Networks 9, 277–282.

Holland, P.W., Leinhardt, S., 1970. A method for detecting structure in sociometric data. American Journal of Sociology 76, 492–513.

Hubert, L., 1987. Assignment Methods in Combinatorial Analysis. Marcel Dekker, New York.

Hummon, N.P., Fararo, T.J., 1995. Assessing hierarchy and balance in dynamic network models. Journal of Mathematical Sociology 20, 145–159.

Kapferer, B., 1972. Strategy and Transaction in an African Factory. Manchester Univ. Press, Manchester.

Krackhardt, D., 1988. Predicting with networks: nonparametric multiple regression analyses of dyadic data. Social Networks 10, 359–382.

Krackhardt, D., 1994. Graph theoretical dimensions of informal organizations. In: Carley, K.M., Prietula, M.J. (Eds.), Computational Organization Theory. Lawrence Erlbaum, NJ, pp. 89–111.

Latané, B., 1996. Dynamic social impact: the creation of culture by communication. Journal of Communication 46 (4), 13–25.

Latané, B., Nowak, A., Liu, J.H., 1994. Measuring emergent social phenomena: dynamism, polarization, and clustering as order parameters of social systems. Behavioral Science 39, 1–24.

Marsden, P.V., 1988. Homogeneity in confiding relations. Social Networks 10, 57–76.

Marsden, P.V., 1990. Network data and measurement. Annual Review of Sociology 16, 435–463.

Mayhew, B., 1980. Structuralism versus individualism: Part I. Shadowboxing in the dark. Social Forces 59 (2), 335–375.

Mayhew, B., 1981. Structuralism versus individualism: Part II. Ideological and other obfuscations. Social Forces 59 (3), 627–648.

Mayhew, B., 1984. Baseline models of sociological phenomena. Journal of Mathematical Sociology 9, 259–281.

Mayhew, B., McPherson, J.M., Rotolo, T., Smith-Lovin, L., 1995. Sex and race homogeneity in naturally occurring groups. Social Forces 74 (1), 15–51.

Padgett, J.F., 1987. Social Mobility in Hieratic Control Systems, unpublished manuscript.

Padgett, J.F., Ansell, C.K., 1993. Robust action and the rise of the Medici, 1400–1434. American Journal of Sociology 98, 1259–1319.

Pattison, P., Wasserman, S., Robins, G., Kanfer, A., 1997. Statistical Evaluation of Algebraic Constraints for Social Networks, unpublished manuscript.

Shannon, C., Weaver, W., 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana.

Skvoretz, J., 1990. Biased net theory: approximations, simulations, and observations. Social Networks 12, 217–238.

Snijders, T., 1991. Enumeration and simulation methods for 0–1 matrices with given marginals. Psychometrika 56 (3), 397–417.

Spencer, H., 1874. Principles of Sociology. Appleton, New York.

Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov random graphs and p^*. Psychometrika 60, 401–426.