The Complexity of Social Networks: Theoretical and Empirical Findings[1][2]

Carter Butts[3]

Department of Social and Decision Sciences

Center for the Computational Analysis of

Social and Organizational Systems

Carnegie Mellon University

August 20, 2000

[1]The author would like to thank Kathleen Carley and John Miller for their input and encouragement.

[2]A previous version of this paper was presented at Sunbelt XIX (1999) in Charleston, SC.

[3]This material is based upon work supported under a National Science Foundation Graduate Fellowship, and was supported in part by the Center for the Computational Analysis of Social and Organizational Systems and the Institute for Complex Engineered Systems at Carnegie Mellon University.

Abstract

A great deal of work in recent years has been devoted to the topic of "complexity,' its measurement, and its implications. Here, the notion of algorithmic complexity is applied to the analysis of social networks. Structural features of theoretical importance - such as structural equivalence classes - are shown to be strongly related to the algorithmic complexity of graphs, and these results are explored using analytical and simulation methods. Analysis of the complexity of a variety of empirically derived networks suggests that many social networks are nearly as complex as their source entropy, and thus that their structure is roughly in line with the conditional uniform graph distribution hypothesis. Implications of these findings for network theory and methodology are also discussed.

Keywords: complexity, entropy, social networks, equivalence, graph distribution

1 Introduction

At least since the development of information theory in the 1950s and 1960s (Shannon, 1948; Rényi, 1961), scientists and mathematicians working in a variety of areas have used notions of "complexity" in order to describe the properties of objects and processes. In addition to this formal usage, complexity has been used informally and semi-formally for decades in conjunction with the study of large, multidimensional systems in many fields (including biology, organization theory, and physics[1]). More recently, interdisciplinary work in a number of areas[2] has brought a renewed interest in complexity as a formal concept with wide applicability to systems across domains. Despite (or perhaps because of) the diversity of scientific efforts involved in this recent work, little agreement has been reached on what, precisely, "complexity" entails, or how a general notion of complexity may be systematically applied. Multiple definitions and measures of complexity – derived from a variety of basic assumptions – have been proposed by researchers working in various fields, and it is suspected (at the very least) that there is no single conceptual dimension which captures the notion of "complexity" in a general, useful fashion. In this paper, then, an attempt is made not to provide an overarching statement regarding the complexity of social networks, but to examine how a particular type of complexity – algorithmic complexity (Kolmogorov, 1965; Chaitin, 1975; Li and Vitányi, 1991) – is related to a variety of structural features which are of substantive theoretical interest to network analysts, and to consider the behavior of one such complexity measure on a range of network data sets. It is hoped that this approach will highlight a potential (concrete) application of formal complexity theory to the substantive concerns of network analysis, rather than to simply exhibit the use of complexity measures for their own sake[3] (in the spirit of Fararo (1978; 1984)).

1.1 Complexity and Its Definitions

As has been noted, the natural language concept of "complexity" has been related to a variety of formal notions[4]. A thorough going review of these notions (and their respective assumptions, histories, and applications) is beyond the scope of this paper; nevertheless, we shall briefly note a few examples from the literature in order to provide some context for what is to follow. It should be emphasized that the selection which follows is neither exhaustive nor necessarily representative, but rather reflect some of the more formally developed notions of complexity which have been employed in the study of natural systems.

Perhaps the earliest (and simplest) notion of complexity is simply that of the cardinality and/or differentiation of an object's subcomponent sets. Intuitively, organisms, organizations, societies, or mechanical

[1] Although formal usage has occurred in these fields as well.

[2] Examples include the study of self-organized criticality in sandpiles, earthquakes, etc., emergent computation and evolutionary approaches to artificial intelligence, and spin-glass models of loosely coupled thermodynamic systems.

[3] Although the study of complexity per se is an interesting and worthy cause (in a mathematical or information theoretic context), our use of the concept in a social scientific context must be predicated upon its value in addressing substantive concerns.

[4] Some of which are inversely related; see Wolpert and Macready (1998).

systems with many parts – or types of parts – are more "complex" than those with fewer components or varieties thereof. Such a straightforward conception of complexity has been widely employed (particularly on an informal or semi-formal basis) in organization theory, biology, and mental model theory, and shows up in various guises in some of the earliest of sociological works (e.g., Durkheim 1933/1893; Spencer, 1874). Size and differentiation have also served as some of the most common targets of structural theory (Blau, 1986; Mayhew et al., 1972), and continue to be of interest at the micro level to network analysts studying personal networks (Burt, 1997; Valente and Foreman, 1998).

Complexity as cardinality is strongly related to another conception which is more closely tied to physics (and to the study of dynamics in general systems); in particular, it is common in these fields to speak of *systems* of high dimensionality as being "complex", particularly when coupling between dimensions is high[5]. Examinations of many families of such systems have revealed often surprising behaviors, such as unexpected convergence to stable fixed points (e.g., the fairly stable slope angle of a sandpile (Christensen et al., 1991)), systematic and robust relationships between event size and frequency (e.g., the "1/f" distributions of earthquakes (Sornette and Sornette, 1989), firm size (Simon, 1955), and word usage (Zipf, 1949)), and emergent intelligent behavior (e.g., the ability of unintelligent simulated ants to collectively locate, transport, and store food (Koza, 1992)). Indeed, it was largely the popularization of these findings in books like those if Waldrop (1992) and Kauffman (1993) which fueled widespread interest in complexity in the larger community[6], and phenomena such as self-organized criticality (Christensen et al., 1991; Bhowal, 1997), emergent computation (Crutchfield and Young, 1990), and evolutionary dynamics (Kauffman and Johnsen, 1991; Wolfram, 1994) continue to have a strong association with complexity despite the sometimes tenuous link between the concept and its applications.

In contrast with these relatively informal notions, complexity has been treated far more rigorously in computer science, information theory, computational mechanics, and combinatorics. A wide range of formal definitions of complexity have been formulated within these fields to deal with particular problems, and some have found widespread application in a number of areas. One such notion (or collection thereof) is that of *computational complexity* which, in informal terms, is the number of basic operations required to execute a particular algorithm (Cook et al., 1998; West, 1996). Computational complexity is of critical importance in algorithm design and heuristic optimization (to name two areas), where it permits the comparison of the computational "cost" of problem solving algorithms; computational complexity can also be used in a variety of engineering contexts to help determine feasibility of information processing systems, to predict execution time, and to assess possible gains due to parallelization (Cook et al., 1998). More theoretically, one notion of

[5] Or, more properly, when the dynamics along any given phase dimension are tightly coupled with the dynamics along other phase dimensions, particularly when such coupling is highly nonlinear.

[6] It is also important to recall that these discoveries (and their popular treatments) followed the realization that "simple," low-dimensional systems could give rise to extremely complicated behaviors (e.g., chaos); the knowledge that the "simple" could give rise to the "complex" may have given the finding that the "complex" could also give rise to the "simple" more salience than it might otherwise have had.

computational complexity (known as *logical depth*) is based on the running time of the shortest algorithm[7] which will reproduce a given sequence (Bennett, 1985; 1990). Logical depth is closely related to *algorithmic complexity*, which we shall consider in more detail presently; simply put, algorithmic complexity is based on the length of the shortest algorithm (for some machine type) which will reproduce a given sequence (Kolmogorov, 1965; Chaitin, 1975; Li and Vityáni, 1991). This notion of complexity is important in coding theory and data compression, and has application in a number of proof techniques (Kircherr, 1992; Li and Vityáni, 1991). Numerous complexity measures exist which are based on information content[8] (e.g., mutual information (Li, 1991), Kullbach information (Cover and Thomas, 1991)), with the general presumption that sequences or systems are more complex to the degree to which they contain more information or to which information in various subcomponents/subsequences is not shared[9] (Wolpert and Macready, 1998). Closely related to the information theoretic measures are those such as *thermodynamic depth* which are based on differences in entropy between macrostates and microstates of thermodynamic systems (Lloyd and Pagels, 1988). In yet another twist, Feldman and Crutchfield (1998a) measure complexity in terms of the computational dynamics of automata called ϵ-machines which contain both stochastic and deterministic components; their aim is to provide a unifying framework for measuring complexity which separates out "statistical" simplicity (i.e., randomness) and "deterministic" simplicity (i.e., repeated patterns) from a presumed genuine[10] complexity (Feldman and Crutchfield, 1998a; 1998b; Crutchfield and Young, 1989). As with much of this literature, work on these more recent measures is ongoing, and consensus on the "proper" complexity formalism for many scientific applications has not yet been reached (see Feldman and Crutchfield, 1998b; Wolpert and Macready, 1998).

1.1.1 Complexity in Network Analysis

In network analysis, relatively little work has been done on complexity per se, though network analysis "inherits" to an extent some of the large body of work on the subject in graph theory (Cook et al., 1998; Kircherr, 1992; Li and Vityáni, 1991)). Freeman (1983) discusses structural complexity of social networks in terms of dimensions of classification. Everett (1985), considering previous work by Mowshowitz (1968a; 1968b; 1968c), provides a specific measure of network complexity related to the number of positions spanned by orbits of a graph; Everett, importantly, demonstrates the substantive meaningfulness of this notion in terms of the total potential interchangeability of positions, arguing that such a measure is more interpretable (and scientifically useful) than that of Mowshowitz (1968a; 1968b; 1968c) (Everett, 1985). Recent work by Butts (2000) has attempted to deal with the interpretability issue by axiomatizing the notion of graph

[7] In the Kolmogorov-Chaitin sense; see below.

[8] Although, arguably, all algorithmic complexity measures are also informational; the division used here is admittedly an informal one.

[9] Though there seems to be some disagreement on this issue; some have asserted that complex systems tend to be self-similar, while others argue that self-*dis*similarity is emblematic of complexity (Wolpert and Macready, 1998).

[10] In that the authors do not regard randomness per se as being complex, which is not in accordance with an algorithmic notion of complexity (see below).

complexity. In this work, Butts shows that no single measure of graph complexity simultaneously satisfies all of several intuitive requirements for a general notion of structural "complexity," and relates the measures of Everett (1985), Mowshowitz (1968a; 1968b; 1968c), and others[11] via a set of axioms which discriminates between them. In the graph theoretic literature, proof techniques (primarily probabilistic methods) have been developed which are based on the Kolmogorov-Chaitin complexity of graphs (Li and Vityáni, 1991). Work in this area has shown a relationship between complexity and graph-level indices (for instance, Kircherr (1992) has shown that almost all trees do not possess a vertex of high degree (i.e., are not highly degree centralized) by way of algorithmic complexity); Anderson et al. (1999) consider the information content of graph-level index (GLI) distributions, though they do not explicitly link this to complexity theory. Mowshowitz (1968a; 1968b; 1968c) also discusses the information content of graphs at some length, and relates this to orbit structure. Some aspects of the relationship between organizational network structure and complexity vis a vis the theory of random graphs are also discussed by Morel and Ramanujan (1998).

1.1.2 Summary

In general, then, even a very cursory overview of the complexity literature identifies a wide range of notions, created for different purposes under differing conceptions of what "complexity" might mean, and employed in differing ways. While some have used "complexity" in a fairly intuitive fashion to describe systems which *seem* complex, others have developed extremely specific formalisms capturing particular aspects of the natural language term. Like the aspects captured, the intended "targets" of these definitions are not all alike: in some cases, complexity is framed in terms of processes or algorithms; in others, systems, sequences, or static objects are assumed to be under evaluation. Some definitions of complexity are constructed around stochastic assumptions (e.g., thermodynamic depth), while others are deterministic (e.g., Kolmogorov-Chaitin complexity). Even in terms of operationalization, one finds definitions of complexity which are largely of theoretical interest (such as the above-mentioned Kolmogorov-Chaitin measure), and others which are framed in measurable terms (such as that of Everett (1985)). Given the immense diversity of definitions and measures of complexity which exist, how is the network analyst to proceed? Here, as has been mentioned, the strategy which shall be followed is one which is motivated by substantive theoretical concerns: as Fararo (1978; 1985) suggests, we shall seek to understand how a particular notion of complexity – algorithmic complexity – relates to standing problems in social network analysis, and shall then attempt to use a particular measure to assess a number of empirical networks.

1.2 Algorithmic Complexity and Social Networks: A Theoretical Motivation

Modern network analysis has devoted considerable effort towards identifying simplifying features of social networks, and towards using these features to represent social structures in a succinct fashion. The notion

[11] Cardinality of node and arc sets, source entropy of the arc set, induced subgraph complexity, and Lempel-Ziv complexity (used here) are also treated.

of human social structures as being representable in terms of interactions among roles (each of which may be composed of multiple positions, each of which may in turn be occupied by multiple actors) implies that a reducibility exists within social structures; formally, these theoretically important reductions are represented by the standard family of algebraic equivalences (structural (Lorrain and White, 1971), automorphic (Everett, 1985; Everett and Borgatti, 1988), regular (White and Reitz, 1983), and their variants) used throughout network research, and by the blockmodels which are produced by applying these reductions to social networks. As the term "reduction" implies, all of the above methods aim to (relatively losslessly) *compress* a large, "complicated" social structure into a smaller, "simpler" one. Insofar as this is possible, then, it must be the case that human social structures are compressible; this, in turn, implies that human social structures must be of low algorithmic complexity[12].

Turning to a somewhat different theoretical tack, it is commonly held (and there exists some evidence to argue; see (Krackhardt, 1987; Romney and Faust, 1982; Freeman, 1992)) that human cognitive representations of social structure conform to a number of biases which are shaped by inherent cognitive processing limitations. Effects of imposed balance, assumptions of reciprocity, assumed interaction due to mutual association[13], and the like can act to constrain actors' cognitive social structures, inhibiting the perception of structural elements which do not fit the easily regenerated pattern. Insofar as this is the case, then, it follows that cognitive social structures must also be of low algorithmic complexity – lower, even, than behavioral social structures – due to the fact that actors' complexity-reducing cognitive mechanisms should be expected to omit relations which violate simple, easily regenerated patterns and to confabulate relations whose absence would likewise cause problems for cognitive representation[14].

In addition to the above, numerous other arguments have been made regarding constraints on social network structure. Social interaction is constrained by physical space (Latané et al., 1995; Wellman, 1996), which in turn can imply the existence of spatially defined stochastic equivalence classes (Butts and Carley, 1999) among groups of actors. Group membership may heavily structure interaction (Blau, 1977), and even in otherwise unconstrained settings shared practices, behaviors, and beliefs can severely constrain practical possibilities for meaningful social interaction (Carley, 1990a, 1990b, 1991; Goffman, 1963). Insofar as these factors are in operation, then, it would seem reasonable to presume that social networks are not highly complex objects in the algorithmic sense: they are heavily constrained, simple structures which may be summarized relatively easily by a smaller system of key variables.

Here, then, we have a clear motivation for considering the complexity of social structures, a motivation which furthermore leads us to consider a particular notion of complexity: numerous theoretical arguments

[12] Relative to what, one should ask. We shall come to this presently.

[13] E.g., actors who are associated with the same "group" may be thought to interact more than they actually do.

[14] Indeed, cognitive social structures imply a very direct argument for considering algorithmic complexity: since human brains are in fact information processors who must regenerate information from stored algorithms, it follows that their capacity to accurately represent social structure *must* be constrained by the interaction of their storage capacity with the algorithmic complexity of the structure to be stored.

from the social network literature suggest that social networks in general (and cognitive social structures in particular) should be algorithmically simple. Furthermore, to turn the point on its head, because many of the network features of interest to theorists – non-trivial structural equivalence classes, balance, etc. – necessarily imply algorithmic simplicity relative to graphs which lack these features, it likewise follows that a general means of screening data for reductions consists of determining whether said data is too algorithmically complex to contain them[15]. Given that substantial effort (human and computational) can be required to manually screen data for the wide range of equivalences and the like which can occur, a technique such as this which can provide a general sense of the most which one might find in a data set with a single computation may prove quite useful. Furthermore, by comparing the observed complexity of a particular social network with the distribution of complexity values produced by a random null model, one may (albeit crudely) gain a sense of whether any reductions one happens to find in a given network are noteworthy, or whether one would be expected to find similar levels of reducibility under a baseline model[16] (Mayhew, 1984).

1.3 The Kolmogorov-Chaitin Complexity

Given that we are interested in algorithmic complexity, it then behooves us to consider the concept more closely. The most widely known definition of algorithmic complexity is the Kolmogorov-Chaitin measure (often referred to simply as the *Kolmogorov complexity*), and which may be defined as follows[17]:

Definition 1 (Kolmogorov-Chaitin Complexity) *Let W be the set of all words over some finite alphabet S, let the length of a word $w = s_1 s_2 ... s_n$, $w \in W$, $s_i \in S$ be given by $l(w)$, and let A be an algorithm transforming binary sequences into words for some finite alphabet. The* complexity of an element w with respect to A *is then the length of the shortest program which computes it, and is given by $K_A(w) = \min_{A(p)=w} l(p)$, $p \in \{0,1\}^*$. $K_A(w)$ for A asymptotically optimal is simply referred to as the* complexity *of w, where an asymptotically optimal algorithm is any algorithm A such that, for any algorithm B, $K_A(w) \leq K_B(w) + c$ for some constant c which does not depend on w. (The existence of such an algorithm is guaranteed by the theorem of Kolmogorov and Solomonoff (1964).)*

Put in more prosaic terms, the Kolmogorov-Chaitin (K-C) complexity of a sequence is the length of the shortest self-terminating program which will produce that sequence. While this seems intuitive enough at first blush, the details of the above definition (and the literature surrounding the measure) suggest that much is hidden beneath the surface. For instance, it happens that an algorithm which would return the K-C complexity would have to solve the halting problem, which is not possible; hence the K-C complexity

[15] Of course, the data in question might be simple in ways other than those of interest as well, but it is nevertheless true that the data cannot be highly reducible if it is algorithmically complex.

[16] At present, few tools exist to enable researchers to determine whether blockmodels they identify are typical of baseline models; given that an equivalence analysis may take the researcher through multiple algorithms, data treatments, relaxations, etc., the potential for inadvertent self-deception is substantial (Dawes, 1988).

[17] The particular formulation used here follows Martin-Löf (1966), which is essentially that of Kolmogorov (1965) but slightly clearer; numerous other, equivalent expressions are possible.

cannot, in general, be computed with certainty[18]. Furthermore, questions exist regarding the possibility of canonical minimum encoding in general, so the question of which asymptotically optimal A to use may not be trivial even at a theoretical level. Nevertheless, for all its uncertainty the K-C complexity measure turns out to be a very important one, with many useful properties. Despite the fact that it cannot be exactly computed, it may be estimated under various constraints (or indexed), and it is often possible to make *theoretical* arguments about the relative K-C complexities of various sequences without knowing the exact values; this has proven to be of great importance in (for instance) the use of K-C complexity as a tool for probabilistic proofs[19] (Li and Vityáni, 1991).

What, then, are the properties of the Kolmogorov-Chaitin complexity (other than its incomputability)? First, and perhaps most obvious, the K-C complexity is minimal for pure repeated sequences such as "00000000" or "11111111"; it is higher for sequences such as "1011010110", and higher still for sequences with even fewer repetitions. Interestingly, it happens that sequences with maximum K-C complexity are, for all intents and purposes, *random*[20] (Martin-Löf, 1966). Of course, when finite sequences are involved, this clearly raises certain questions about what one means by a "random" sequence (see Lempel and Ziv (1976)); after all, even a uniform draw on the thousand-digit binary integers *can* still produce a pure repeated sequence, so the best that can be done in any case is to discriminate between sequences which are "typical" of a certain random process and those which are not[21]. In any event, the connection between the K-C complexity and randomness is an important one which has earned the measure some controversy in certain circles. Some researchers (e.g., Lloyd and Pagels, 1988) have simply equated algorithmic complexity with a "measurement of randomness" and, by reasoning (on intuitive grounds) that randomness should not be complex, have argued that this is not properly a complexity measure. This line of reasoning, however, misses the point: randomness *is* complex, in the sense that there is no "simple" deterministic representation of a random sequence. Randomness is incompressible; it is, indeed, definable in this fashion[22]. Insofar, then, as we are lead by our basic assumptions to a complexity measure such as the Kolmogorov-Chaitin complexity, we must realize that these two notions are intimately interconnected.

2 The Lempel-Ziv Complexity Measure

As we have seen, there exists a strong theoretical rationale for examining the algorithmic complexity of social networks. The most basic definition of algorithmic complexity – the Kolmogorov-Chaitin complexity

[18] Kolmogorov (1965) poses the incomputability argument in somewhat different terms, but the implication is similar.

[19] For example, one may show that a given property of interest implies a K-C complexity below a certain bound, and then show that the probability of any given sequence's having such a complexity is vanishingly small, thereby demonstrating that almost no sequences possess the property of interest.

[20] As Martin-Löf (1966) puts it, they "possess all conceivable statistical properties of randomness."

[21] Which, of course, is exactly what a null hypothesis test does; the complexity literature does not generally frame the issue in this fashion, however.

[22] This connection is also clearer if one notes that the Kolmogorov-Chaitin complexity can be related to information content (Kolmogorov, 1965).

– has a strong axiomatic basis, but is also fundamentally uncomputable. What, then, are we to do as a practical matter if we are to investigate the algorithmic complexity of empirical networks? One solution to this dilemma is not to attempt to find the K-C complexity directly, but instead to find another measure which acts as an *index* of the K-C complexity, in the general sense that it can be shown to behave as the K-C complexity does over a range of inputs. Just such an approach has been taken by Lempel and Ziv (1976), who introduced an algorithmic complexity measure based on a particular encoding scheme whose properties could be shown to be very similar to those of the Kolmogorov-Chaitin complexity. Though not a true measure of the K-C complexity, this index is readily computable (an $O(n^2)$ algorithm exists) on finite sequences, and, as we shall see, it can be shown to detect reductions which are of interest to social scientists.

2.1 Rationale

The basic approach taken by Lempel and Ziv (1976) to the problem of measuring algorithmic complexity of finite sequences consists of the formulation of a self-delimiting production process which is used to generate the input sequence; the number of steps in this sequence (which corresponds to the number of unique elements in the vocabulary of the process) is then used as a measure of sequence complexity[23]. Because the Lempel-Ziv (L-Z) complexity is thus derived from a particular encoding of a sequence, it is not (nor is it intended to be) an "absolute" measure of complexity; for our purposes, this is significant in that it implies that there may be features of graphs which imply algorithmic simplicity, but whose impact will not be assessed by the L-Z measure (which may be unable to exploit such reductions). That said, it happens that the L-Z complexity measure can be shown to detect certain features of interest (such as structural equivalence), and its successful application in other areas suggest that its capacity for pattern recognition is reasonably robust (Kaspar and Schuster, 1987). The L-Z measure as applied to graphs has also been shown to satisfy several of the structural complexity axioms put forth in Butts (2000)[24], including complementarity (the complexity of a graph is the complexity of its complement) and sample monotonicity (the complexity of a graph is at least as large as the complexity of its most complex induced subgraph); the variant developed in Section 2.5.1 below satisfies labeling invariance as well. Given, then, that the measure can be shown to satisfy the basic requirements of a structural complexity measure, to be readily computable, to detect features of interest, and to have other desirable properties (as shown by Lempel and Ziv (1976)), we here take it as a reasonable starting point for an empirical investigation of network complexity.

[23] In some sense, then the Lempel-Ziv complexity is related simultaneously to computational complexity and to algorithmic complexity; its properties of interest (to us), however, correspond to the latter interpretation.

[24] These axioms are argued to form a reasonable basis for the measurement of structural complexity in digraphs, and can be used to discriminate between different measures of structural complexity.

2.2 Formal Derivation

Having considered something of the rationale behind the Lempel-Ziv complexity measure, then, let us proceed to its formal definition[25]. First, let us define A^* to be the set of all finite length sequences over a finite alphabet A with the null-sequence, Λ, in A^*. Let the length of a sequence $S \in A^*$ be given by $l(S)$, and let $A^n = \{S \in A^* | l(S) = n\}, n \geq 0$.

Given the above, a sequence $S \in A^n$ may be specified by $S = s_1 s_2 ... s_n$; we denote the substring of S which starts at position i and ends at position j by $S(i,j)$ ($S(i,j) = s_i s_{i+1} ... s_j$ for $j \geq i$, otherwise $S(i,j) = \Lambda$). Two such sequences $Q \in A^m$ and $R \in A^n$ may be concatenated to form a new sequence $S = QR$, $S \in A^{m+n}$, in which case $S(1,m) = Q$ and $S(m+1, m+n) = R$. A sequence Q is more generally called a *prefix* of a sequence S if there exists an integer i such that $Q = S(1,i)$; in this case, S is also referred to as an *extension* of Q. Prefixes of S may also be denoted via the operator π, which is defined such that $S\pi^i = S(1, l(S) - i)$ for $i = 0, 1, ...$ (Note that $S\pi^0 = S$, and $S\pi^i = \Lambda$ for $i \geq l(S)$.) The *vocabulary* of a sequence S ($v(S)$) is the subset of A^* formed by all substrings ("words") of S; hence, $v(S) = \bigcup_{i,j} S(i,j)$.

Having established some basic notation, we may now proceed to consider the process upon which the Lempel-Ziv complexity is based. Given a sequence S and a word $W \in v(S)$, let $R = SW$ be the extension formed by the concatenation of S and W. Obviously, R can be produced by a simple algorithm which sets $r_i = s_i$ for $1 \leq i \leq l(S)$, and $r_i = s_{i-l(S)+a}$ for $i > l(S)$ where a is that integer satisfying $W = S(a,b)$ (the starting position of W). Similarly, the same procedure may be used to construct a much longer extension $R = SQ$, for $Q \in v(SQ\pi)$. (This works because the first portion of Q must be in $v(S)$, and the subsequent portion may be "bootstrapped" after the first characters are copied; Q may itself be a concatenation of several substrings of S.) For such an extension R, we say that R is *reproducible* from S (denoted $S \rightarrow R$). If, on the other hand, we take non-null sequence S with prefix $Q = S(1,j)$ such that $Q \rightarrow S\pi$ and $j < l(S)$, we say that S is *producible* from Q (denoted $Q \Rightarrow S$). Note that these two concepts are distinct: reproduction demands that the entire output sequence be the result of the recursive copying process, while production permits some "innovation" at the end of the process.

This, then, forms the basis of the Lempel-Ziv production process. The algorithm in question begins with the null sequence $\Lambda = S(1,0)$, and performs $S(1,0) \Rightarrow S(1,1)$ by adding the innovation s_1. After this first step, $v(S_1) = \Lambda, s_1$, which may in turn be used to generate a longer prefix of S, and so on. More formally, at each iteration the algorithm performs $S(1,h_i) \Rightarrow S(1,h_{i+1})$ (beginning with $S(1,0) = S(1,h_0) = \Lambda$) until (after at most $l(S)$ steps), S has been produced, at which point the algorithm terminates. If we consider the prefix $S(1,h_i)$ produced at iteration i to be the ith *state* of the m-step production process, then we may parse S into $H(S) = S(1,h_1)S(1+h_1, h_2)...S(1+h_{m-1}, h_m)$, called the *production history* of S. $S(1+h_{i-1}, h_i)$, denoted $H_i(S)$, is referred to as the ith *component* of $H(S)$; such a component is said to be *exhaustive* iff $H_i(S) \not\rightarrow H_{i+1}(S)$[26], and a history is said to be exhaustive iff each of its components is likewise exhaustive.

[25] This discussion follows Lempel and Ziv (1976, sections II and III), though some details have been omitted for brevity.

[26] Note that, by definition, $H_i(S) \Rightarrow H_{i+1}(S)$, but it does not follow that $H_i(S) \rightarrow H_{i+1}(S)$. In essence, a component is

Given this, the Lempel-Ziv complexity is as follows:

Definition 2 (Lempel-Ziv Complexity) *Given a sequence S with history $H(S)$, let $C_H(S)$ denote the number of components of $H(S)$. The* Lempel-Ziv complexity *of S is then given by $C_{L-Z}(S) = \min_{H(S)} C_H(S)$.*

As it happens, it is possible to prove that C_{L-Z} is equal to the number of components of the exhaustive history of S, which both exists and is unique for all non-null sequences S; this proof will not be given here, but may be found in Lempel and Ziv (1976). Thus, C_{L-Z} exists and is unique for all sequences. Representing the L-Z complexity in terms of the number of steps in the exhaustive production process also leads to the deduction that (because all steps are productions in an exhaustive history) the representation of S in terms of a parsing given by the exhaustive history must contain at least C_{L-Z} symbols. A variety of other deductions are also made by Lempel and Ziv (1976), which shall not be considered at length. It will, however, be noted that an upper bound on the L-Z complexity is given by

$$C_{L-Z}(S) < \frac{n}{\left(1 - 2\frac{1+\log_\alpha \log_\alpha (\alpha n)}{\log_\alpha (n)}\right) \log_\alpha (n)} \tag{1}$$

where $S \in A^n$ and α denotes the size of the alphabet A. A somewhat weaker (but more useful) constraint is given by the observation that, for a random sequence $S \in A^n$ with $|A| = \alpha$ and source entropy h,

$$\lim_{n\to\infty} C_{L-Z}(S) \to \frac{hn}{\log_\alpha (n)} \tag{2}$$

(Lempel and Ziv, 1976; Kaspar and Schuster, 1987).

2.3 Sample Application of the Lempel-Ziv Measure

To gain a stronger intuition for the Lempel-Ziv measure (and the production process upon which it is based), it may be useful to consider a few simple examples. To begin with, let us consider the sequence

$$S = 0000000000000000$$

The exhaustive history of this sequence can be given by the following decomposition:

$$S = 0 \cdot 000000000000000$$

and hence $C_{L-Z}(S) = 2$. The somewhat more complex sequence given by

$$S = 0101010101010101$$

yields the decomposition

$$S = 0 \cdot 1 \cdot 01010101010101$$

exhaustive when there is no way in which it could be extended by a production.

and hence here $C_{L-Z}(S) = 3$. Finally, a yet more sophisticated sequence

$$S = 0001101001000101$$

produces the exhaustive history

$$S = 0 \cdot 001 \cdot 10 \cdot 100 \cdot 1000 \cdot 101$$

for which $C_{L-Z}(S) = 6$. Thus, as one can see, the L-Z complexity measure provides fairly intuitive outputs for a range of binary sequences; more formal results demonstrating the C_{L-Z} behaves as an algorithmic complexity measure "should" are not treated here, but can be found in Lempel and Ziv (1976).

2.4 The Kaspar and Schuster Algorithm

As has been mentioned, the L-Z complexity measure has a variety of properties which make it especially suitable for our purposes. Another advantage of the L-Z measure is that an ($O(n^2)$) algorithm to calculate the L-Z complexity of a given sequence has been demonstrated by Kaspar and Schuster (1987), who have also explored some of the measure's behaviors in an applied setting. In particular, Kaspar and Schuster examine the convergence of the L-Z measure to the theoretical expected value for random sequences[27] and find that errors generally fall within the $\pm 5\%$ range for sequences of length 1000 and greater. (Given the encoding used here (described below), this corresponds to networks of size ≥ 32.) Smaller sequences were associated with larger errors, although even a few hundred elements were sufficient to bring errors to within $\pm 10\%$. Kaspar and Schuster also utilize the L-Z complexity measure to analyze several specific systems (including several cellular automata and the logistic map), and show that the L-Z measure can discriminate between pattern formation and source entropy change, between chaotic and periodic behavior[28], and between multiple pattern types. The success of the Lempel-Ziv complexity measure in performing these tasks on a number of systems with known properties lends credence to the robustness of the measure across a range of inputs, though it obviously does not imply that the L-Z complexity is a perfect match for the Kolmogorov-Chaitin measure.

2.5 Application to the Measurement of Graph Complexity

Having defined the Lempel-Ziv complexity (and having identified an algorithm which will produce it), we may now proceed to apply this measure to directed graphs. Given a labeled digraph $G = (V, E)$, we may define an adjacency matrix $\mathbf{A}$ such that $\mathbf{A_{ij}} = 1$ if and only if $(v_i, v_j) \in E(G)$, and 0 otherwise. From $\mathbf{A}$,

[27] Which are not, of course, fully random. The problem of pseudorandom number generation – always an issue in research which depends heavily on the assumption of true stochasticity – is obviously amplified when using measures which can presumably tell the difference between "real" randomness and pseudorandomness for some sequence length. Kaspar and Schuster (1987) employ sequences from the Numerical Algorithm Group, Ltd., but do not discuss any experiments with other methods.

[28] Within a limited periodic window, of course.

we may in turn define a unique bit string $\mathbf{S}$ such that $\mathbf{S_i} = \mathbf{A}_{\lfloor i/|V| \rfloor, i - \lfloor i/|V| \rfloor}$ (that is, a concatenation of the rows of $\mathbf{A}$). This string is a direct representation of the labeled digraph G; it also happens to be suitable for encoding using the algorithm of Kaspar and Schuster (1987). The Lempel-Ziv complexity of a labeled digraph G, then, is simply $C_{LZ}(S)$ where C_{LZ} is as given in Lempel and Ziv (1976) and described above.

It should be noted at this point that this encoding scheme can easily be extended to incorporate integer valued digraphs, or even categorically valued digraphs[29]. Since the Lempel-Ziv production process assumes a finite but arbitrary alphabet, it is not required that sequences be binary (although more complex alphabets may require more convergence time, and their complexity ceilings are obviously different); finding C_{LZ} for non-binary alphabets is thus fairly trivial. While many of the potential applications of this fact are fairly obvious, one in particular bears mentioning: a non-binary alphabet may be employed to permit a *dyadic* digraph encoding, in place of the *arc* encoding here employed. Although this will not be discussed in depth here, the procedure is fairly straightforward; for each dyad of the labeled digraph, one assigns each possible state (4 for simple digraphs) to a given element of the L-Z alphabet. The complexity of the resulting sequence of dyad states may then be determined using standard methods, with the usual interpretation. One advantage of the dyadic approach is that it is potentially much more sensitive to reciprocity effects than the arc encoding[30]; on the other hand, because the number of dyads is only half the number of arcs (and because the alphabet is larger), this encoding is only appropriate for relatively large networks[31]. For this reason, only the arc encoding will be employed in the analyses which follow.

2.5.1 Complexity of Unlabeled Structures

The above procedure suffices to give us an estimate of the complexity of a labeled digraph; in many circumstances, however, this is not a sensible quantity. To understand why this is so, let us consider the following two sociomatrices:

(Insert Figure 1 Here)

Note that Structure A and Structure B are isomorphic: there exists a bijection from A to B or, to put it another way, there exists a *relabeling* (or, here, reordering) of the vertices of A which produces B, and vice versa[32]. In this sense, then, A and B are structurally identical; while they may be written a bit differently (perhaps due to different data collection procedures, naming conventions, etc.) there is no purely

[29] I.e., a digraph in which each tie belongs to a given category, but in which those categories may have neither ordinal nor cardinal significance (e.g., contacted by phone versus interacted in person).

[30] Preliminary experiments have suggested that reciprocity is a graph feature of theoretical interest which is *not* readily picked up by the L-Z measure.

[31] Precisely how large is hard to say, because the Kaspar and Schuster (1987) convergence results apply only to binary sequences; this is a potentially important question for future research. We also do not here attempt to investigate column-wise, rather than row-wise, concatenation, though informal examination suggests that results using each are generally similar.

[32] This particular relabeling happens to be given by the vertex ordering (1,8,2,7,3,6,4,5).

structural measure which would vary between these graphs. Unfortunately, however, the above application of the Lempel-Ziv measure to these two matrices does not yield identical results. If one considers the binary representations of these sociomatrices, one can see why:

Structure A: 011100001011000011010000111000000000011100001011000011010000 1110

Structure B: 001010100001010110001010010001011010001001010001101010000 1010100

Observe that Structure A has more "runs" of 0's and 1's than does Structure B. When we recall that long strings of 1's and 0's can be very succinctly expressed using the Lempel-Ziv algorithm, we can clearly see that Structure A can be expected to have a somewhat lower complexity than Structure B. Just as it is easier for the human brain to quickly grasp this simple underlying pattern from matrix A above than from matrix B, so too is it easier for the Lempel-Ziv algorithm to express the neatly blocked pattern of A than the more periodic patterns of B; in the Lempel-Ziv sense, the second representation *is* more complex.

This fact is no fault of the method itself – the measure is indeed accurately characterizing complexity – but reveals a slight disjuncture between the structures whose complexity we would like to measure[33] (the *underlying* structures associated with unlabeled graphs) and those we are actually measuring (labeled graphs). What is needed, then, is a way to use the Lempel-Ziv measure to determine the complexity of unlabeled graphs. The approach which shall be presented here is similar in spirit to that used by Butts and Carley (1998) for comparison of unlabeled structures: we shall attempt to decompose observed complexity values into a fundamental, structural component and a variable, labeling component, and shall attempt to estimate the latter from the former. Although this procedure will not permit us to utilize some of the more elegant results from Lempel and Ziv (1976), we shall endeavor to use computational techniques to compare our findings with those of various null models.

2.5.2 Formal Presentation of a Method for Estimating the Lempel-Ziv Complexity of an Unlabeled Digraph

Given an unlabeled digraph G, let the labeling λ represent a unique ordering (or labeling) of the vertex set V(G). The labeling function $L(\lambda, G)$, then, returns a labeled digraph G_λ with vertex labels λ. If $C(L(\lambda, G))$ is the L-Z complexity of such a digraph, then we may define a minimum value for C on a given unlabeled digraph as follows:

$$C^S(G) = \overset{min}{\lambda} C(L(\lambda, G)) \tag{3}$$

$C^S(G)$ shall henceforth be referred to as the *structural complexity* of the unlabeled digraph G. Note that this function depends *only* on the unlabeled digraph; by minimizing over all possible labelings, we remove any λ effects. From the definition of C^S above, we may furthermore define a complementary quantity C^λ:

[33] In fact, there are circumstances in which we might very well care about the complexity of labeled structures. One obvious case is illustrated by the matrices above: less complex representations of graphs appear to be easier to comprehend, and thus the L-Z complexity may be applied to problems of graph visualization.

$$C^{\lambda}(L(\lambda,G)) = C(L(\lambda,G)) - C^{S}(G) \tag{4}$$

$C^{\lambda}(L(\lambda,G))$ is the *labeling complexity* of the labeled digraph $L(\lambda,G)$, and represents that portion of the observed Lempel-Ziv complexity of $L(\lambda,G)$ accounted for by the labeling λ alone. This term, combined with the structural complexity, permits us to re-express the L-Z complexity in the following decomposition:

$$C(L(\lambda,G)) = C^{\lambda}(L(\lambda,G)) + C^{S}(G) \tag{5}$$

Thus, we see that the observed L-Z complexity can be thought of as being a combination of some fundamental structural complexity which depends solely on the unlabeled digraph being examined, and some additional labeling complexity which is dependent on the way in which the digraph happens to be presented. Furthermore, it is trivial to show that both the structural and labeling complexities are positive integers, and that the labeling complexity is bounded below at 0[34]. If one can find some labeling of a given digraph for which $C^{\lambda} = 0$, then, the Lempel-Ziv of complexity of that labeled digraph will be the structural complexity of its corresponding unlabeled digraph.

Unfortunately, computing $C^{S}(G)$ is not a trivial task; in the worst case, it would require searching over all $|V(G)|!$ labelings of G for the minimum complexity! In the analyses which follow, then, we are content to merely *estimate* $C^{S}(G)$ by taking the minimum observed value over a reasonable (1,000) uniform sample from the set of all labelings. Clearly, alternative procedures are possible – specialized canonical labeling algorithms somewhat like those of Butts and Carley (1998) or heuristic labeling search as is used in Butts (1998) (applying genetic algorithms, simulated annealing, or other forms of heuristic optimization) are obvious choices – and their investigation is left as a task for future research.

3 Complexity and Structural Equivalence

Structural equivalence, first defined formally by Lorrain and White (1971), is perhaps the best known of an array of network analysis techniques which attempt to simplify graphs by means of identifying sets of nodes with similar structural properties. Structural equivalence has, further, served as an important theoretical tool for understanding the diffusion of innovations (Burt, 1987), competition between actors (Burt, 1992), and the structure of scientific fields (Breiger et al., 1975); it has also encouraged a long line of research into similar methods of position, role, and class analysis (White and Reitz, 1983; Everett and Borgatti, 1988; Everett, 1985). Interestingly, it also happens that (as we shall see) the presence of structural equivalence is intimately related to graph complexity.

Members of a given structural equivalence class must, by definition, have precisely the same neighborhoods. As a result, we can treat such nodes as "copies" of each other; if we select one copy from each class,

[34] This follows from the fact that the L-Z complexity is a positive integer, that C^{S} is the L-Z complexity of some labeled digraph (see Equation 4) and that C^{S} is a minimum over all possible labelings of G.

then, we can reduce the larger graph to a much smaller structure which, in an important sense, represents the entire graph. Such reduced structures are commonly referred to as *blockmodels*[35], and are important tools for distilling large, difficult to comprehend structures into smaller, more easily inspected forms. More salient for our purposes, however, is the fact that such blockmodels represent an *encoding* of the larger graph: given a graph of relations between equivalence classes and a vector indicating how many members are contained within each class, we can reproduce the original graph by "copying" the relevant nodes. This suggests that graphs which contain nontrivial structural equivalence classes may be compressible, and thus that their Kolmogorov complexity may be lower than graphs which do not contain these equivalences.

To see how this may be the case, let us consider a graph, $\mathbf{A}$, with SE blockmodel[36] $\mathbf{B}$ and multiplication vector $\mathbf{m}$ consisting of the number of nodes in each respective equivalence class. Given this, it is obvious that there exists some algorithm which can reproduce $\mathbf{A}$ by repeatedly copying the nodes of $\mathbf{B}$ as per the vector $\mathbf{m}$. An upper bound on the Kolmogorov complexity of $\mathbf{A}$, then, must be given by

$$C_K(\mathbf{A}) \leq C_K(\mathbf{B}) + C_K(\mathbf{m}) + k \tag{6}$$

where k is a constant associated with the complexity of the copying algorithm itself. Obviously, it is always possible that a simpler algorithm exists; however, since one can always reconstruct any given graph from the above, it clearly serves as an upper bound.

How much does this tell us? This depends on the complexity of the blockmodel $\mathbf{B}$ and the associated multiplication vector. If these can themselves be reproduced with relatively minimal algorithms, then the above could constitute a significant improvement over considering $\mathbf{A}$ in the absence of structural equivalence. As we have already noted, however, we cannot easily find C_K for an arbitrary structure; thus, while the above result provides us with some insight into the connection between structural equivalence, blockmodels, and algorithmic complexity, it is of little practical value. But what of our empirical complexity measure, C_{LZ}? Provided that C_{LZ} is a reasonable index of algorithmic complexity, it is not unreasonable to assume that the same logic used in equation 6 might be applied to the Lempel-Ziv measure as well. Given this, we may hypothesize that

$$\begin{aligned} C_{LZ}(\mathbf{A}) &\approx C_{LZ}(\mathbf{B}) + C_{LZ}(\mathbf{m}) + k & (7) \\ &\approx \frac{|V(\mathbf{B})|}{log_2\left(|V(\mathbf{B})|\right)} + \frac{|V(\mathbf{B})|}{log_{\substack{max \\ \forall i}} |\mathbf{B_i}|\left(|V(\mathbf{B})|\right)} + k & (8) \end{aligned}$$

which, for large $\mathbf{B_i}$, gives us

[35] Blockmodels are used for a variety of purposes, and summarizing the relations between structural equivalence classes is only one of these; for the moment, however, we shall restrict ourselves to this application.

[36] In the text which follows, the term "blockmodel" will often be used to refer to what is also called the "image matrix" or "blockmodel reduction" of a graph. The simple assignment of nodes to blocks without reduction is here referred to as a "blocking," as distinct from the model which a given blocking produces.

$$\lim_{\substack{max \\ i}} \mathbf{B}_i \to \infty \; C_{LZ}(\mathbf{A}) \to \frac{|V(\mathbf{B})|}{log_2\left(|V(\mathbf{B})|\right)} + k \tag{9}$$

Clearly, this (if correct) is a powerful result: algorithmic complexity of the larger graph is, in the limit, on the order of the complexity of its structural equivalence-induced blockmodel. Where $|V(\mathbf{G})| \gg |V(\mathbf{B})|$, this may be differ substantially from a random structure of equal size and density. Does this argument, however, actually hold for the Lempel-Ziv measure? To find out, we consider the complexity of expanded blockmodels.

3.1 Complexity of Expanded Blockmodels

To examine the degree of algorithmic simplicity found in graphs containing non-trivial structural equivalence classes, it is first necessary to obtain a criterion; in particular, we must have graphs which are known to contain the desired structural features. One method of generating such graphs, which will be employed here, is that of *blockmodel expansion*. In a blockmodel expansion, one takes a (randomly generated) block image matrix, and "expands" it by making multiple copies of each node. Here, we examine expanded images for a variety of initial image sizes, maximum class sizes, and under additional constraints (random permutation and noise). By evaluating the complexity of the resulting graphs, we can test the hypothesis that expanded blockmodels are indeed of very low complexity under the Lempel-Ziv measure.

Our analysis will take the form of a virtual experiment. In this experiment we shall generate random uniform blockmodels of fixed size, expand them via a class membership vector (uniform random with controlled maximum), and in some cases add symmetric noise and/or randomly permute the resulting graph. These last two steps should allow us to test the robustness of the measure: if the Lempel-Ziv measure is extremely robust, it should continue to detect the underlying structural simplicity even when that structure is permuted and randomly modified[37]. As a further handicap, we shall not search the permutation space of the graphs under consideration, instead examining only the complexity of the labeled graph. Here again, the question is whether the Lempel-Ziv measure can still notice underlying structure under extremely harsh conditions; if it can, in fact, do so, then we can be confident of its ability to identify structural equivalence under more favorable conditions.

The conditions for our virtual experiment are as follows:

(Insert Table 1 Here)

Generating random matrices in each condition, we then apply the L-Z complexity measure to each resulting graph, and tabulate the results. Do the structures appear to be algorithmically simple? For an overall

[37] Of course, adding *too* much noise will truly reduce any equivalence present to random levels, defeating the purpose of the test. For this reason, noise is here limited to 0, 5, and 10 percent of all arcs.

result, we examine the histogram of observed complexity values across all conditions:

(Insert Figure 2 Here)

As figure 2 demonstrates, even across all conditions (including noise and permutations, and using only the labeled graph encoding) there is clearly a great deal of simplicity present. While many graphs are fairly close to their asymptotic maxima, much of the weight of the distribution is sharply less, with the bulk well below 1.0 and many at 0.5 or less. Clearly, these structures are correctly identified by the Lempel-Ziv measure as being algorithmically simple, even under adverse conditions, thus proving the feasibility of the measure for the identification of structural features of interest. What, however, of the relationship between image size, class size, and complexity? For this we turn to figure 3, below.

(Insert Figure 3 Here)

Across all conditions, there is a clear curvilinear relationship between normalized complexity and maximum class size: as classes become larger, more reduction is possible, resulting in a lower normalized complexity score. Block image size has a negligible effect on normalized complexity, which is as expected; image size should dominate the raw complexity measure, but the *normalized* complexity should be a function only of the degree of reduction (discounting noise and permutations). Here again, then, we find that the measure is behaving as expected.

The above distributional analyses were taken across noise and permutation conditions; as noted, a great deal of compressibility was identified despite these factors. What, however, are the effects of noise and permutation on normalized complexity? How robust is the Lempel-Ziv measure to these influences? For this, we examine the normalized complexity distribution by noise and permutation (figure 4):

(Insert Figure 4 Here)

Clearly, the effect of both factors on complexity is dramatic. With no noise nor random permutation, the normalized complexity values overwhelmingly fall around 25% of their asymptotic maximum. Permutation increases this to near 40%, indicating that while the L-Z measure is affected by random labeling effects (when permutation sampling is not used), these effects are not crippling. Noise, however, very quickly degrades algorithmic performance: a 5% chance of change per tie raises median observed complexity above 50%, and permutation combined with this places the distribution in the 75% range. While even 10% noise cannot completely prevent the algorithm from exploiting SE reductions, it alone is able to raise observed complexity to about three-fourths of the theoretical limit. In combination with permutations, this level of noise can quite effectively mask structural equivalence.

In general, then, the implication of this would seem to be that while the Lempel-Ziv measure is reasonably robust, fairly noisy data is likely to mask any underlying order in the criterion structure[38]. Labeling effects can compound this problem; hence, it is clearly advisable to utilize a technique such as that described above to attempt to estimate the unlabeled graph complexity where possible, particularly when noise is involved. Clearly, the L-Z complexity measure behaves as expected on marginally "clean" data, and can be employed to identify structural equivalence (even approximate structural equivalence, to a degree) in social networks.

4 Complexity of Empirical Networks

We have, in the above sections, argued that the algorithmic complexity of social networks is a matter of theoretical interest. Likewise, we have shown how one such complexity measure may be applied to digraphs, and have demonstrated that at least one property of special interest – structural equivalence – is easily detected by the complexity measure. What, however, of the actual complexity of empirical networks? To gain some idea, we now proceed to apply the Lempel-Ziv complexity measure to a variety of social networks taken from a wide range of empirical studies.

4.1 Data and Procedure

The networks which are here examined are taken from a collection of sample data sets included in Wasserman and Faust (1994) and with the UCINET IV network analysis package (Borgatti, Everett, and Freeman, 1996). These data sets span a range of relation types, collection contexts, and collection methods, and are often used as a testbed for new methods (see, for instance, Breiger et al., Boorman and White). Although it should be emphasized that this does represent a sample of convenience, rather than a representative sample of all social networks, the total number of networks (n=112) is relatively evenly divisible into three categories (observational/behavioral (n=36), simple self-report (n=34), and cognitive social structure (n=42)) which represent the majority of data types used in the field. It is hoped, then, that the data set employed here will be broad (and representative) enough to permit some preliminary conclusions regarding the complexity of social networks, conclusions which may then be tested more extensively by future research.

The specific data sets included here are listed by a shortened coding system. The first element(s) of the code identify the set (and in some cases the collector), along with some further identifying information where necessary. These codes (and the corresponding sets) are as follows:

(Insert Table 2 Here)

Additional clarifying codes are as follows. For sociometric or valued behavioral data, the dichotomization

[38] Again, it is not clear to what extent this is a fault of the measure, and to what extent the introduction of random noise truly obliterates structure in the data. This would seem to be an important question for future research.

employed is indicated by a three letter code starting with the letter "d", followed either by "g" (if values "greater than or equal to" were used) or "l" (for "less than or equal to"), and then by an "m" (if the dichotomization was about the mean over all arcs) or a "d" (if about the median). Hence, a code of "dld" indicates that the data was coded such that values less than or equal to the median arc strength were mapped to 1, while those greater than the median were mapped to 0. When choosing dichotomizations, the direction of strength which was theoretically appropriate given the data was used (e.g., greater than for number of interactions, less than for rankings). Where possible, dichotomizations were performed using the median arc strength; if (and only if) this resulted in a degenerate matrix, the mean was used instead. In addition to these two codes (the set identifier and dichotomization codes), each data set considered here bears a "nos" code[39] (a data format reference which has no effect on the analyses conducted here), and in some cases a trailing number. The trailing number, where present, identifies the specific matrix in a matrix stack; hence, krackoff.fr.nos.8 refers to the 8th matrix in the Krackhardt office CSS data stack, friendship relation. Since each matrix is analyzed separately, the number of matrices in any given stack has no relation to the complexity of the data set. However, being able to identify individual matrices with particular properties may be useful in some cases[40].

For each of these matrices, then, the following procedure was performed. First, each adjacency matrix was transformed into its arc encoding, as per the method of Section 2.5. Second, the Lempel-Ziv complexity of each such encoding was determined using the Kaspar and Schuster (1987) algorithm, and the asymptotic complexity maximum was determined using the encoding length and source entropy[41]. Third, this process was repeated on 10,000 random sequences which were constrained to have length and entropy identical to the original graph encoding; these were used to obtain a 95% Monte Carlo confidence interval for the distribution of C_{LZ}, and to obtain specific p-values for C_{LZ} observed. Finally, after this was performed, steps two and three were repeated with the following modifications: the number of Monte Carlo trials was restricted to 100, and each matrix (original and random) was tested under 1,000 random permutations, with the minimum complexity draw being used as an estimate of the unlabeled graph complexity[42]. The above data – complexity values, theoretical maxima, p-values, and 95% Monte Carlo confidence intervals – for both the labeled and unlabeled cases were then divided by data type (observational, standard self-report, or cognitive social structure) for subsequent analysis and presentation.

[39] For those who are interested, it refers to "*N*eo-*O*rg*S*tat" format.

[40] For instance, to identify individuals with unusually complex or simple cognitive network representations.

[41] This is merely the first order Shannon entropy (h) of the sequence; in the binary case, it is simply $-(p\log_2 p + (1-p)\log_2(1-p))$, where p is the graph density. Note that $0 \leq h \leq 1$, and h is symmetric about its maximum (which is, in this case, at $p = 0.5$).

[42] The number of Monte Carlo trials was restrained for computational reasons: since the number of conditions to be run scales by the product of the number of permutations and the number of random sequences, it was not feasible to consider the same number of trials in the unlabeled graph analysis.

4.2 Preliminary Hypotheses

While we are, to some extent, simply interested here in discovering what patterns of complexity are found in social networks, it is also the case that our prior theoretical motivation permits us to pose some preliminary hypotheses regarding network complexity. Although the present study cannot be expected to yield firm confirmations or denials of these hypotheses[43], we may nevertheless use this initial, semi-exploratory test to serve as the basis for future, more rigorous analyses. With this in mind, then, the following hypotheses are suggested[44]:

Hypothesis 1 (The Simple Network Hypothesis) *Social networks, in general, will be substantially algorithmically simpler than random graphs of equivalent size and density.*

Hypothesis 2 (The Cognitive Representation Hypothesis) *Cognitive social structures, in general, will be substantially algorithmically simpler than either random graphs or non-cognitive social networks of equivalent size and density.*

Hypothesis 3 (The Physical Constraint Hypothesis) *Behavioral networks, in general, will be algorithmically simpler than self-report networks of equivalent size and density.*

While the basic rationale for each of these hypotheses was stated in the introduction, a brief reconsideration will be given here. **H1** follows from the argument that social networks in general – whether of behavioral or cognitive origin – are heavily constrained by numerous factors, contain non-trivial equivalences, and deviate in other ways which cause them to be algorithmically simple relative to random graphs with identical sizes and densities. The motivation for **H2** is similar: cognitive social structures should be *especially* constrained by associative thinking (which will tend to cause actors to link people who have otherwise similar associations), pressures towards balance (which will promote the confabulation of ties between actors with positive ties towards shared alters, among other things), framing of interaction in terms of groups rather than relations (Freeman, 1992), etc., and hence should be much simpler than would be expected either from the behavioral network or from comparison with a random graph. **H3** is obviously a bit more presumptive, but follows an ironically similar logic; the assumption of **H3** is that (as per Bernard et al., 1980) each individual within a self-report network will add his or her own biases to the total network, resulting in a structure which obscures the actual constraints which structure behavior (Mayhew et al., 1972; Latané et al., 1995) (but not always the memory of it)[45].

[43] The primary reasons being that the methods employed are still "new" and require further testing, and that the data set being examined may or may not be biased in some fashion.

[44] Hypotheses 1 and 2 were explicitly chosen (and 3 implicitly) prior to the data analysis; notably, they are also (largely) incorrect.

[45] Contrast this with **H2**: in the latter case, one is eliciting a complete set of biased data, which should result in compression. In the former, it is assumed that differences in individual circumstances will *collectively* result in random noise, despite the fact that each such bias is part of an overall simplification scheme at the intraindividual level.

In general, then, prior theory (and practice) in network analysis suggests to us that the hallmark of social networks will be their simplicity, rather than their complexity. With these three hypotheses in mind, then, we now turn to the data.

4.3 Observed Behavioral Networks

The set of observed behavioral networks contains 36 separate matrices, running the gamut from anthropological observation (kapmine) to international trade (cntrytrade). What unites all data in this collection, however, is that all sets are derived from some third-party observer (human or otherwise), and that all concern observed behavior of some sort or other (rather than self-reports of affect, for instance). While many of these networks are defined on physically proximate populations, this is not true of the entire set (freeei and cntrytrade being two obvious examples); likewise, some but not all of the sets concern actors embedded in a formal organizational context (e.g., krackhiman.rep and bkoff). This set thus represents a fairly diverse group of studies, which should (presumably) test the robustness of our hypotheses.

The specific observations on each data set in the observational data group follow. Note that $C_{\max}$ here refers to the asymptotic random sequence maximum, adjusted for sequence length and source entropy. All p-values and confidence intervals reported are derived from the Monte Carlo test procedure described above.

(Insert Table 3 Here)

As can be seen, the above groups vary in their theoretical maximum complexity; in general, however, one finds little support for hypotheses **H1** or **H3** here. Surprisingly, only around half of the networks considered are below the 95% confidence interval[46], and in the labeled case approximately 86% of all graphs are within a confidence interval of the asymptotic random maximum[47]! To better understand the data's global behavior, let us now examine a visualization of the observational data set:

(Insert Figure 5 Here)

In figure 5 above, the plotted bars represent the 95% Monte Carlo confidence intervals for the labeled graph complexities. The asymptotic maximum complexity for each graph is represented by a left-pointing triangle; note that this value may not always be contained within the confidence interval, due to the previously mentioned fact that entropy and graph size can affect convergence rates. The observed L-Z complexity of each graph is given by a right-facing triangle. By comparing the distance between the right-facing point of this triangle and the confidence interval (or the asymptotic maximum), one can infer how much simpler the

[46] 44% in the labeled case, versus 56% in the unlabeled case.

[47] The width of the confidence interval is used here in place of the standard deviation, due to the nature of the distributions in question. Additionally, the asymptotic maximum is only a very rough guideline in the unlabeled case, since it does not take into account the complexity reduction due to permutation on random sequences.

graph in question is than a typical random graph with similar properties. Looking across the irregular line of intervals and complexity values, one can at a glance obtain a sense of their overall distribution; as noted above, while many networks *are* simpler than would be expected, many are not, and those which display unusual levels of simplicity are not generally drastic in the differences they evince.

Given the above results in the labeled case, let us now visualize the unlabeled data:

(Insert Figure 6 Here)

Figure 6, above, is much like figure 5 in layout, and can be read in much the same fashion. The asymptotic maximum, which is defined for labeled sequences, has been omitted here due to the fact that it does not correct for the unlabeled case. (A numerical comparison is still available in the above data, however.) In comparison with figure 5, several differences are noticeable in figure 6. First, and most obvious, far more graphs are significantly simple in the unlabeled case than in the labeled case; despite this, however, one finds that most absolute differences remain small (with the exception of sets such as cntrytrade, freeei, and bk*). Second, one is also struck by the fact that the confidence intervals appear to be far narrower in the unlabeled case than in the labeled case. While this could in some cases be due to the smaller sample size used, it would also seem that the range of common complexity values is far narrower for unlabeled structures than for their labeled counterparts. While this partially accounts for the increased number of significant differences in the unlabeled case, it is noteworthy that none of the graphs examined show the substantial simplicity found in the expanded SE blockmodels; a strong claim for **H1**, then, is substantially undermined by this data.

4.4 Self-Report Networks

Where the observational/behavioral data sets consist of data collected by third parties, the self-report networks considered here consist of structures created from a composition of individually reported ties. Generally, the relations sampled in this way focus on reported interaction, affect, or participation in a socially-defined relationship (e.g., friendship, advice-giving). Also unlike the previous set, the self-report networks considered here consist exclusively of relations among human subjects[48]; hence, this set is less broad than the observational data set.

As before, we begin by examining the Lempel-Ziv complexity data for each matrix directly:

(Insert Table 4 Here)

Out of the total data set, 32% and 29% (labeled and unlabeled respectively) of all graphs had L-Z complexity ratings below the 95% confidence interval. While this would appear to be a substantial difference

[48] Note that this is not necessarily true of self-reports in general: self-reports of institutional ties are common in interorganizational research.

from the observational data, the asymptotic limit comparison tells a somewhat different story: approximately 88% of the matrices considered here were within one confidence interval of $C_{\max}$, which is nearly the same as 86% in the observational case. What is going on here? To get a better sense, let us turn to the visualizations:

(Insert Figure 7 Here)

Figure 7 shows the L-Z complexity scores (and associated measures) for the self-report data in the labeled case. Clearly, some of the labeled structures – such as certain of the freeei.* sets – are much simpler than would be expected, though most of the others are not. Further elaboration may be seen in the unlabeled case:

(Insert Figure 8 Here)

Here, it is evident that some sets (such as freeei) are relatively simple in general, while other sets (such as newfrat) show no signs of simplification even in the unlabeled case. Clearly, then, we do not see a ringing endorsement for **H1** or **H3** here, although there is continued evidence of some consistent deviation from the random limit.

4.5 Cognitive Social Structures

Our final data set is, in many ways, the most unusual of the three. Unlike the other two collections, all of the matrices in this set come from a single study, on a single population of human actors. The ties which are reported are relatively subjective ones, and the data is "self-report" in the sense that network participants are used, but here each network consists of an individual's full elicited representation of connections among all actors in the set; this is thus quite different from self-report data, in which subjective perceptions of many actors may be joined to form a single network[49]. As with the other sets, this one contains a fair number of matrices (n=42) of multiple types, though the multiple in this case is only two (ascribed friendship and advice-giving). Examining this CSS stack, then, gives us a chance to look more deeply into a particular data set, at the expense of a broader view.

Let us now proceed to examine the complexity data for the cognitive social structure set. Note that all of the data presented here comes from the krackoff study, on two relations. Each matrix number corresponds to a single subject from the study; hence, one may compare across relations to seek individual differences in representational complexity[50]. The data is as follows:

[49] Though this has more to do with our use of the data (in "slices," to use the CSS term) than with the data itself. Krackhardt (1987) rightly regards CSS stacks as being more than simply a collection of independent networks, and has demonstrated how elements of the CSS may be pooled to gain new observations. For our purposes, however, we shall be interested in the complexities of the individually perceived networks.

[50] Though, with an n of 2, this is ill-advised....

(Insert Table 5 Here)

Far from what **H2** would predict, in general we do not find strong evidence for substantial structural simplicity within the CSS stack. Every graph examined was within one confidence interval of the asymptotic maximum in the labeled case, and only 19% of the labeled graphs were significantly different from the Monte Carlo sample at the 0.05 level. The unlabeled case, however, is a bit more interesting: 57% of all CSS slices were below the 95% confidence interval here, a higher ratio than either of the other groups. For more information, we look to the visualizations:

(Insert Figure 9 Here)

In figure 9, we can clearly see that, while most of the graphs have complexity values near the bottom of their respective confidence intervals, they are still within acceptable bounds. One can also see in figure 9 a clear demarcation between the friendship networks (which are both low-entropy and low complexity) and the advice networks (which are higher on both counts). Substantial variance within relations exists as well, for instance, indicating high variability in network density by respondent. How does this change so drastically in the unlabeled case? To see this, we must turn to figure 10:

(Insert Figure 10 Here)

Here, in the unlabeled case, we can see that matrix complexity values are still close to their confidence intervals; the intervals, however, have diminished in magnitude (a common finding across all three sets) and graph complexities have also fallen slightly. The uniform result of this across the stack is a large number of graphs which are slightly less complex than the random samples, but not greatly so. The basic pattern of differences across relations persists in the unlabeled case, as does the pattern of individual differences; given that sequence length is fixed, the fact that L-Z complexity scores can vary from under 5 to over 50 indicates the dramatic effect of entropy in constraining possibilities for complexity. Indeed, this variability makes clear the important fact that most of the algorithmic simplicity which is present in these graphs is not due to "deep" factors, but simply to low entropy. In a sense, then, some of our initial intuitions seem to be correct: social networks are indeed simple, compared to *unconditional* random graphs. Once density is controlled for, however, the picture changes sharply.

4.6 General Comparison

Having examined each of our three empirical subsets in turn, what can we now say regarding their relations? While we hypothesized in **H1** that all three of the groups would be substantially simpler than a conditional uniform random sample, this does not seem to be borne out in practice. While it is true that a fair number

of our structures are *significantly* simple at the 0.05 level (31% in the labeled case, and 48% in the unlabeled case), the differences are rarely drastic[51]. In the labeled case, it is useful to examine the distribution of normalized complexity values:

(Insert Figure 11 Here)

Figure 11, above, shows us these normalized values (values less than 1 signify complexity levels below the asymptotic maximum) for the labeled case. Despite some tail weight towards the bottom of the graph in the observational case, none of the three data groups appear to have most of their distributions below the 1.0 level. More importantly, few values indeed range below 0.7, and none broaches half of the asymptotic maximum value. On the other hand, we have already seen that the asymptotic maximum – while a useful benchmark – does not always reflect accurately the actual complexity of relatively small, entropy-constrained graphs. As another means of comparison, then, we can examine the distribution of an index given by $(C_{LZ} - CI_L)/(CI_U - CI_L)$, where CI_U and CI_L are the upper and lower bounds (respectively) of the 95% Monte Carlo confidence interval. This index provides us with a sense of how many intervals below (or above) the lower bound of the confidence interval the typical observation lies, and is given in figure 12:

(Insert Figure 12 Here)

Here again, we are examining labeled graphs, and here again we find that much of the distributional weight is above the 0 mark (i.e., at or above the minimum of the interval). Long, fat negative tails are present, however, particularly within the observational data set[52], indicating that a few networks are some intervals below the Monte Carlo lower bound. (To get a sense of what this means in absolute terms, it is helpful to note that the mean confidence interval in the labeled case was 4.9 units wide, with a standard deviation of 1.13 units.)

What about the unlabeled case? After all, if graphs are unimpressively simple when labeled, it may simply be that the structures are "scrambled" in a way which adversely affects the Lempel-Ziv measure. To examine this possibility, we turn in figure 13 to the distribution of the same index as figure 12 above, but for unlabeled graphs:

(Insert Figure 13 Here)

Clearly, it would seem that there is more downside weight in the unlabeled case (though a few outliers, such as the freeei.acq2.dgm data set, make the effect appear more substantial than it is). At the same time,

[51] In the labeled case, 80% of the total sample was within one confidence interval of the asymptotic maximum.

[52] While the self-report data does not have a fat lower tail, it does have a number of highly negative outliers.

however, the distributional medians remain very close to the 0.0 mark, indicating continued weight at the top of the scale[53]. Also, the fact that the mean confidence interval has a width of only 1.77 (stddev 0.70) in the unlabeled case suggests that larger tail multipliers may actually mean *less* for unlabeled graphs (in absolute terms), since the total variability of the random samples decreases in this case[54].

What of differences between distributions? Statistical comparison under these circumstances is somewhat problematic, but non-parametric tests (Kruskal-Wallace and medians test) indicated generally significant differences[55] on both indicator variables across the three samples. **H3**, which argues that observational networks should be simpler than self-report networks, has some support; however, this result clearly rests on lower tail weight within the observational complexity distribution, and does not reflect substantial differences. The cognitive social structures, far from being the simplest graphs (as per **H2**) are if anything *more* complex relative to random structures than the self-report and observational networks! (Here again, however, the total differences are not large.) In terms of general comparison, then, it would seem that these three types of social network data are more alike than they are different (though some differences exist), and likewise it would seem that at least two of our three hypotheses are either contradicted or only weakly supported by the empirical evidence.

5 Discussion and General Implications

What have we learned thusfar regarding the complexity of social networks, and what are the implications of these findings for social scientific research? To begin with, it would seem that the assumption that cognitive and other limitations will cause certain networks – especially cognitive social structures – to be extremely simple beyond what would be expected from density alone is not supported by the data. Whatever differences exist between data types appear to be subtle, and there is no evidence here to suggest that these network types will vary greatly in their possession of unusual features (although *which ones* each possesses is not addressed by this work). What, however, of network complexity *in general*? In **H1** it was argued that social networks would (by virtue of their containing non-trivial equivalences, unusual GLI distributions, etc.) be extremely simple relative to random graphs. By any reasonable measure, this seems not to be the case. Clearly, some social networks do seem to be less complex than random structures, but not vastly so[56].

If this is true, however, what does it mean? While this study is only a preliminary, one possibility is that

[53] And, it should be noted, it is difficult to have long tails on the high end, due to the relative improbability of being *more* complex than a sample of random graphs.

[54] The mean reduction in graph complexity due to permutation was 5.1 units (sd=2.4); CI lower bounds decreased by 3.38 (sd=1.01) units on average, but CI upper bounds fell by an average of 6.52 (sd=1.51), more than making up for the difference.

[55] In the unlabeled case, $\frac{C}{C_{max}}$ p values were $p < 0.001$ and $p < 0.014$ for KS and medians tests, respectively; for the confidence interval index, respective p values were $p < 0.052$ and $p < 0.012$.

[56] Or, at least, this is true of our *observations* of these networks. If errors in network data are primarily random rather than simplifying, then the present result could argue that our data contains so much noise that we cannot perceive the underlying structural simplicity. Such an alternative hypothesis is interesting (albeit distressing), but will not be treated further here.

a competing hypothesis is actually a better depiction of social network structure than one might think at first blush. Specifically, we must consider the following:

Hypothesis 4 (Conditional Uniform Graph Distribution Hypothesis) *The* conditional uniform graph distribution hypothesis *is that hypothesis which states that the aggregate distribution of empirically realized social networks is isomorphic with a uniform distribution over the space of all graphs, conditional on graph size and density.*

This hypothesis has been used in the past as a baseline model of network formation (Mayhew et al., 1972; Mayhew, 1983), and work by Anderson et al. (1999) have found that many networks do not vary significantly from this hypothesis on a range of graph-level indices. Insofar as this hypothesis is correct, then, what are the implications? First, insofar as graphs obey the CUGDH, network analysts should be extremely cautious when using tools such as approximate equivalence class detection algorithms. By searching across definitions of equivalence, detection algorithms, and relaxation levels, researchers are nearly certain to "find" some blockmodel which fits with their intuitions; however, it may be the case that many approximate blockmodels identified in this fashion are simply random artifacts. Without a solid null-hypothesis testing apparatus, it is difficult to be certain that one has not simply found "phantom" blockmodels. Use of complexity measures as a preliminary screening technique may alleviate this problem somewhat, but a more specialized set of tools for discerning between unusual and random equivalences is still required.

Of course, if the CUGDH is valid, it also follows that much of what will be found – or not found – in any given graph will be driven heavily by density and graph size. Anderson et al. have already shown that GLI distributions are both extremely poorly behaved and heavily influenced by size and density; given the substantial constraints on possible values imposed by combinatoric considerations, it seems likely that these influences are active on social networks as well. Considerations of algorithmic complexity and source entropy suggest, likewise, that graphs of extreme density will be simple relative to unconditional uniform graphs, and that certain graph properties may be more common in these structures due to these factors alone. Of course, this is quite consistent with a long line of sociological theory (Spencer, 1874; Durkheim, 1893; Mayhew, 1983) which has argued that social density *is* a key determinant of social life. If so, it would seem critical that social scientists ensure that their measurement of these two structural variables is correct, lest they be lead astray by apparent effects which are in fact the result of density (or size) misestimation. Given the evidence seen here for the effects of changing dichotomization levels on algorithmic complexity, better dichotomization procedures might be a good place to start.

All of this said, to what extent *can* we say that the CUGDH is representative of social networks? This study, to be sure, cannot resolve this issue. Clearly, many actual social networks do appear to have levels of complexity which are not entirely consistent with the CUGDH, but neither do the networks examined here display the level of simplicity required for **H1**. A researcher who attempted to model the data presented here using the CUGDH would have a reasonably low magnitude of error (in terms of L-Z complexity), but

that error would be both persistent and in one direction. One suspects, then, that CUGDH is a baseline model rather than a "finish-line model[57];" the question of how far off that finish-line may yet be is obviously a question for subsequent research.

6 Conclusions

In this paper, an approach to the use of algorithmic complexity in the analysis of social networks has been introduced, based on a theoretical motivation regarding constraints on graph structure. A specific measure of algorithmic complexity developed by Lempel and Ziv (1976) was introduced, and its application to the measurement of complexity in directed graphs was discussed. Examination of the algorithmic complexity of expanded structural equivalence blockmodels demonstrated the ability of the Lempel-Ziv measure to detect features of structural interest, and it was shown that graphs constructed from random SE blockmodels are highly algorithmically simple. Application of the Lempel-Ziv complexity measure to a large set of social networks revealed (contrary to expectations) that most networks are nearly as complex as would be expected from their size and source entropy; some persistent deviation from a random baseline was detected, however. Comparison of multiple data types – observational/behavioral, self-report, and cognitive social structures – failed to identify strong differences in complexity between types, also contrary to a priori expectations (though some differences were detected). Some implications of the apparent complexity of graphs for network theory (and the conditional uniform graph distribution hypothesis in particular) were discussed, and evidence for the importance of density as a determinant of social structure was presented. It is hoped that this preliminary work will lead to further development in the application of algorithmic complexity to social network analysis, and that it will encourage the development and use of new tools for the identification of social structure.

7 References

Anderson, Brigham, Butts, Carter, and Carley, Kathleen. (1999). "The Interaction of Size and Density with Graph Level Indices." *Social Networks*, 21(3), 239-267.

Bennett, C.H. (1985). "Dissipation, Information, Computational Complexity and the Definition of Organization." In D. Pines (ed.), *Emerging Syntheses in Science*, 215-133. Redwood City, CA: Addison-Wesley.

Bennett, C.H. (1990). "How to Define Complexity in Physics, and Why." In W.H. Zurek (ed.), *Complexity, Entropy and the Physics of Information,* 137-148. Redwood City, CA: Addison-Wesley.

Bernard, H., Killworth, P., and Sailer, L. (1980). "Informant Accuracy in Social Network Data IV." *Social*

[57] See the discussion surrounding Mayhew (1984) in *JMS*, 9.

Networks, 2, 191-218.

Bhowal, Ajanta. (1997). "Damage Spreading in the 'Sandpile' Model of SOC." *Physica A,* 327-330.

Blau, Peter M. (1968). "The Hierarchy of Authority in Organizations.' *American Journal of Sociology*, 73(4), 453-467.

Borgatti, Everett, and Freeman. (1996). *UCINET IV*, v1.64. Natick, MA: Analytic Technologies.

Breiger, R., Boorman, S., and Arabie, P. (1975). "An Algorithm for Clustering Relational Data, With Applications to Social Network Analysis and Comparison With Multidimensional Scaling." *Journal of Mathematical Psychology,* 12, 328-383.

Burt, Ronald. (1987). "Social Contagion and Innovation: Cohesion Versus Structural Equivalence." *American Journal of Sociology,* 92, 1287-1335.

Burt, Ronald. (1992). *Structural Holes: The Social Structure of Competition.* Cambridge: Harvard University Press.

Burt, Ronald. (1997). "A Note on Social Capital and Network Content." *Social Networks,* 19(4).

Butts, Carter. (1998). "Cluster Analysis of Unlabeled Structures." ICES Research Report 88-04-98, Carnegie Mellon University.

Butts, Carter. (2000). "An Axiomatic Approach to Network Complexity." *Journal of Mathematical Sociology*, in press.

Butts, Carter, and Carley, Kathleen. (1998). "Canonical Labeling to Facilitate Graph Comparison." ICES Research Report 88-06-98, Carnegie Mellon University.

Butts, Carter, and Carley, Kathleen. (1999). "Spatial Models of Network Formation." Center for the Computational Analysis of Social and Organizational Systems Working Paper, Carnegie Mellon University.

Carley, Kathleen. (1990a). "Group Stability: A Socio-Cognitive Approach." *Advances in Group Processes*, 7.

Carley, Kathleen. (1990b). "On the Persistence of Beliefs." Working Paper, Department of Social and

Decision Sciences, Carnegie Mellon University.

Carley, Kathleen. (1991). "Co-Evolution of Structure and Culture: A Socio-Cognitive Approach." Prepared for the *Small Groups Conference*, Cincinnatti, Ohio, August 1991.

Chaitin, Gregory J. (1975). "A Theory of Program Size Formally Identical to Information Theory." *Journal of the Association for Computing Machinery,* 22(3), 329-340.

Christensen, K., Fogedby, H.C., Jensen, H.J. (1991). "Dynamical and Spatial Aspects of Sandpile Cellular Automata." *Journal of Statistical Physics,* 63(3-4).

Cook, W.J., Cunningham, W.H., Pullyblank, W.R., and Schrijver, A. (1998). *Combinatorial Optimization.* New York: John Wiley and Sons.

Cover, T.M., and Thomas, J.A. (1991). *Elements of Information Theory.* New York: John Wiley and Sons.

Crutchfield, James P., and Young, Karl. (1989). "Inferring Statistical Complexity." *Physical Review Letters,* 63(2).

Crutchfield, James P., and Young, Karl. (1990). "Computation at the Onset of Chaos." In W. Zurek (ed.), *Complexity, Entropy, and the Physics of Information: SFI Studies in the Sciences of Complexity, Vol VIII.* Reading: Addison-Wesley.

Dawes, Robyn. (1988). *Rational Choice in an Uncertain World.*

Durkheim, Emile. (1933) [1893]. *The Division of Labor in Society.* New York: Free Press.

Everett, Martin G. (1985). "Role Similarity and Complexity in Social Networks." *Social Networks,* 7, 353-359.

Everett, Martin G. and Borgatti, Steve. (1988). "Calculating Role Similarities: An Algorithm That Helps Determine the Orbits of a Graph." *Social Networks*, 77-91.

Fararo, Thomas J. (1978). "An Introduction to Catastrophes." *Behavioral Science*, 23, 291-317.

Fararo, Thomas J. (1984). "Critique and Comment: Catastrophe Analysis of the Simon-Homans Model."

Behavioral Science, 29, 212-216.

Feldman, David P., and Crutchfield, James P. (1998a). "Discovering Noncritical Organization." Santa Fe Institute Working Paper, 98-04-026.

Feldman, David P., and Crutchfield, James P. (1998b). "Measures of Statistical Complexity: Why?" *Physics Letters A,* 238, 244-252.

Freeman, Linton C. (1983). "Spheres, Cubes, and Boxes: Graph Dimensionality and Network Structure." *Social Networks,* 5, 139-156.

Freeman, Linton C. (1992). "Filling in the Blanks: A Theory of Cognitive Categories and the Structure of Social Affiliation." *Social Psychology Quarterly*, 55(2), 118-127.

Freeman, S.C. and Freeman, L.C. (1979). "The Networkers Network: A Study of the Impact of a New Communications Medium on Sociometric Structure." Social Science Research Reports No. 46. Irvine, CA: University of California.

Galaskiewicz, J. (1985). *Social Organization of an Urban Grants Economy.* New York: Academic Press.

Goffman, Erving. (1963). *Behavior in Public Places.* New York: Free Press.

Kapferer, B. (1969). "Norms and the Manipulation of Relationships in a Work Context." In J. Mitchell (ed.), *Social Networks in Urban Situations.* Manchester: Manchester University Press.

Kapferer, B. (1972). *Strategy and Transaction in an African Factory.* Manchester: Manchester University Press.

Kaspar, F., and Schuster, H. G. (1987). "Easily Calculable Measure for the Complexity of Spatiotemporal Patterns." *Physical Review A*, 36(2), 842-848.

Kauffman S. A. (1993) *The Origins of Order: Self-Organization and Selection in Evolution.* New York:Oxford University Press, New York.

Kauffman, S.A., and Johnsen, S. (1991). "Co-Evolution to the Edge of Chaos: Coupled Fitness Landscapes, Poised States, and Co-Evolutionary Avalanches." In C.G. Langton, C. Taylor, J.D. Farmer, and S. Ras-

mussen (eds.), *Artificial Life II: SFI Studies in the Sciences of Complexity, Vol X.* Addison-Wesley.

Killworth, P., and Bernard, H. (1976). "Informant Accuracy in Social Network Data." *Human Organization,* 35, 269-286.

Kircherr, W. (1992). "Kolmogorov Complexity and Random Graphs." *Information Processing Letters,* 41(3).

Kolmogorov, A. N. (1965). "Three Approaches to the Definition of the Concept 'Quantity of Information'." *Problems in Information Transmission*, 1, 3-11.

Koza, John R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* Cambridge, Mass: MIT Press.

Krackhardt, David. (1987). "Cognitive Social Structures." *Social Networks,* 9, 104-134.

Latané, B., Liu, J.H., Nowak, A., Bonevento, M., and Zheng, L. (1995). "Distance Matters: Physical Space and Social Impact." *Personality and Social Psychology Bulletin,* 21(8), 795-805.

Lempel, A. and Ziv, J. (1976). "On the Complexity of Finite Sequences." *IEEE Transactions on Information Theory*, 22, 75.

Li, Meng, and Vitányi, Paul M.B. (1991). "Combinatorics and Kolmogorov Complexity." In *Proceedings of the Sixth Annual IEEE Conference on Structure in Complexity Theory.*

Li, Wentian. (1991). "On the Relationship Between Complexity and Entropy for Markov Chains and Regular Languages." *Complex Systems,* 5, 381-399.

Lloyd, Seth, and Pagels, Heinz. (1988). "Complexity as Thermodynamic Depth." *Annals of Physics,* 188, 186-213.

Lorrain, F. and White, Harrison C. (1971). "Structural Equivalence of Individuals in Social Networks." *Journal of Mathematical Sociology*, 1, 49-80.

MacRae, J. (1960). "Direct Factor Analysis of Sociometric Data." *Sociometry,* 23, 360-371.

Martin-Löf, Per. (1966). "The Definition of Random Sequences." *Information and Control,* 9, 602-619.

Mayhew, Bruce. (1983). "Hierarchical Differentiation in Imperatively Coordinated Associations." *Research in the Sociology of Organizations*, 2, 153-229.

Mayhew, Bruce. (1984). "Baseline Models of Sociological Phenomena." *Journal of Mathematical Sociology*, 9, 259-281.

Mayhew, Bruce, McPherson, Miller, Levinger, Roger, and James, Thomas. (1972). "System Size and Structural Differentiation in Formal Organizations: A Baseline Generator." *American Sociological Review*, 37, 629-633.

Morel, Benoit, and Ramanujan, Rangaraj. (1998). "Complex Systems Theory and Organization Theory." Working paper, Carnegie Mellon University.

Mowshowitz, A. (1968a). "Entropy and the Complexity of Graphs I: An Index of the Relative Complexity of a Graph." *Bulletin of Mathematical Biophysics,* 30, 175-204.

Mowshowitz, A. (1968b). "Entropy and the Complexity of Graphs II: The Information Content of Graphs and Digraphs." *Bulletin of Mathematical Biophysics,* 30, 225-240.

Mowshowitz, A. (1968c). "Entropy and the Complexity of Graphs III: Graphs with Prescribed Information Content." *Bulletin of Mathematical Biophysics,* 30, 387-414.

Newcomb, T. (1961). *The Acquaintance Process.* New York: Holt, Reinhard, and Winston.

Padgett, J.F. and Ansnell, C.K. (1993). "Robust Action and the Rise of the Medici, 1400-1434." *American Journal of Sociology,* 98, 1259-1319.

Read, K. (1954). "Cultures of the Central Highlands, New Guinea." *Southwestern Journal of Anthropology,* 10, 1-43.

Rényi, A. (1961) "On Measures of Entropy and Information." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press.

Roethlisberger, F., and Dickson, W. (1939). *Management and the Worker.* Cambridge: Cambridge University Press.

Romney, A. Kimball, and Faust, Katherine. (1982). "Predicting the Structure of a Communications Network from Recalled Data." *Social Networks*, 4, 285-304.

Sampson, S. (1969). *Crisis in a Cloister.* Unpublished doctoral dissertation, Cornell University.

Schwimmer, E. (1973). *Exchange in the Social Structure of the Orokaiva.* New York: St. Martins.

Shannon, Claude. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal,* 27, 379-423.

Simon, Herbert. (1955). "On a Class of Skew Distribution Functions." *Biometrica,* 42, 425-440.

Solomonoff, R.J. (1964). "A Formal Theory of Inductive Inference, I." *Information and Control,* 7(1), 224-254.

Sornette, A. and Sornette, D. (1989). "Self-Organized Criticality and Earthquakes." *Europhysics Letters,* 9(3), 197-202.

Spencer, Herbert. (1874). *Principles of Sociology*, vol 1. New York: Appleton.

Stokman, F., Wasseur, F., and Elsas, D. (1985). "The Dutch Network: Types of Interlocks and Network Structure." In F. Stokman, R. Ziegler, and J. Scott (eds.), *Networks of Corporate Power.* Cambridge: Polity Press.

Thurman, B. (1979). "In the Office: Networks and Coalitions." *Social Networks,* 2, 47-63.

Valente. T.W. and Foreman, R.K. (1998). "Integration and Radiality: Measuring the Extent of an Individual's Connectedness and Reachability in a Network." *Social Networks,* 20(1).

Waldrop, M. Mitchell. (1992). *Complexity: The Emerging Science at the Edge of Order and Chaos.* New York: Simon and Schuster.

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

Wellman, B. (1996). "Are Personal Communities Local? A Dumptarian Reconsideration." *Social Networks,* 18, 347-354.

West, D.B. (1996). *Introduction to Graph Theory.* Upper Saddle River, NJ: Prentice-Hall.

White, D.R. and Reitz, K.P. (1983). "Graph and Semigroup Homomorphisms on Networks of Relations." *Social Networks,* 5, 193-235.

Wolfram, S. (1994). *Cellular Automata and Complexity.* Reading, MA: Addison-Wesley.

Wolpert, David H., and Macready, William G. (1998). "Self-Dissimilarity: An Empirical Measure of Complexity." Santa Fe Institute Working Paper.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort.* New York: Hafner.

Table 1: Conditions for the Structural Equivalence Experiment

Experimental Conditions	
Permuted?	Yes, No
Block Image Size	4, 8, 12
Maximum Class Size	5, 8, 11, 14
Noise Level	0.0, 0.05, 0.1
Replications per Condition	5
Total Runs	360

Table 2: Empirical Data Sets, with Identification Codes

Set Code	Description
krackhiman.ad.*	Krackhardt's Hi-Tech Managers, Advice (Krackhardt, 1987)
krackhiman.fr.*	Krackhardt's Hi-Tech Managers, Friendship (Krackhardt, 1987)
krackhiman.rep.*	Krackhardt's Hi-Tech Managers, A Reports to B (Krackhardt, 1987)
padffam.bus.*	Padgett's Florentine Families, Business Relations (Padgett, 1987)
padffam.mar.*	Padgett's Florentine Families, Marriage Relations (Padgett, 1987)
freeei.acq1.*	Freeman's EIES Data, Acquaintanceship (T1) (Freeman and Freeman, 1979)
freeei.acq2.*	Freeman's EIES Data, Acquaintanceship (T2) (Freeman and Freeman, 1979)
freeei.mes.*	Freeman's EIES Data, Messages (2 Mats) (Freeman and Freeman, 1979)
cntrytrade.bmg.*	Country Trade Data, Basic Manufactured Goods (Wasserman and Faust, 1994)
cntrytrade.fla.*	Country Trade Data, Food and Live Animals (Wasserman and Faust, 1994)
cntrytrade.cm.*	Country Trade Data, Crude Materials (Wasserman and Faust, 1994)
cntrytrade.mfe.*	Country Trade Data, Mineral Fuels, Etc. (Wasserman and Faust, 1994)
cntrytrade.ed.*	Country Trade Data, Exchange of Diplomats (Wasserman and Faust, 1994)
galaceocl.com.*	Galaskiewicz's CEOs and Clubs, CEOs and Clubs (Galaskiewicz, 1985)
bkfrat.beh.*	Bernard and Killworth's Fraternity Data, # Obs Interactions (Bernard et al., 1980)
bkfrat.cog.*	Bernard and Killworth's Fraternity Data, Cog Rankings of Interaction (Bernard et al., 1980)
bkham.beh.*	Bernard and Killworth's Ham Radio Data, # Obs Interactions (Killworth and Bernard, 1976)
bkham.cog.*	Bernard and Killworth's Ham Radio Data, Cog Rankings of Interaction (Killworth and Bernard, 1976)
bkoff.beh.*	Bernard and Killworth's Office Data, # Obs Interactions (Killworth and Bernard, 1976)
bkoff.cog.*	Bernard and Killworth's Office Data, Cog Rankings of Interaction (Killworth and Bernard, 1976)
bktec.beh.*	Bernard and Killworth's Technical Data, # Obs Interactions (Bernard et al., 1980)
bktec.cog.*	Bernard and Killworth's Technical Data, Cog Rankings of Interaction (Bernard et al., 1980)
gama.pos.*	Read's Highland Tribes Data, Positive Relations (Read, 1954)
gama.neg.*	Read's Highland Tribes Data, Negative Relations (Read, 1954)
kapmine.uni.*	Kapferer's Mine Data, Uniplex (Kapferer, 1969)
kapmine.mul.*	Kapferer's Mine Data, Multiplex (Kapferer, 1969)
kaptail.soc.*	Kapferer's Tailor Shop, Social Relations (2 Mats) (Kapferer, 1972)
kaptail.ins.*	Kapferer's Tailor Shop, Instrumental Relations (2 Mats) (Kapferer, 1972)
krackoff.ad.*	Krackhardt's Office CSS Data, Advice Relation (21 Mats) (Krackhardt, 1987)
krackoff.fr.*	Krackhardt's Office CSS Data, Friendship Relation (21 Mats) (Krackhardt, 1987)
newfrat.*	Newcomb's Fraternity Data (15 Mats) (Newcomb, 1961)
prison.*	Gagnon and Macrae's Prison Data (MacRae, 1960)
sampson.lk.*	Sampson's Monestary Data, Liking (3 Mats) (Sampson, 1969)
sampson.dk.*	Sampson's Monestary Data, Disliking (Sampson, 1969)
sampson.es.*	Sampson's Monestary Data, Esteem (Sampson, 1969)
sampson.des.*	Sampson's Monestary Data, Disesteem (Sampson, 1969)
sampson.in.*	Sampson's Monestary Data, Positive Influence (Sampson, 1969)
sampson.nin.*	Sampson's Monestary Data, Negative Influence (Sampson, 1969)
sampson.pr.*	Sampson's Monestary Data, Praise (Sampson, 1969)
sampson.bl.*	Sampson's Monestary Data, Blame (Sampson, 1969)
szcid.*	Stokman-Ziegler's Corporate Interlock Data, Netherlands (Stokman et al., 1985)
szcig.*	Stokman-Ziegler's Corporate Interlock Data, West Germany (Stokman et al., 1985)
taro.*	Schwimmer's Taro Exchange Data (Schwimmer, 1973)
thuroff.org.*	Thurman's Office Data, Organizational Structure (Thurman, 1979)
thuroff.inf.*	Thurman's Office Data, Informal Relations (Thurman, 1979)
wiring.gam.*	Roethlisberger and Dickson's Bank Wiring Room Data, Horseplay (Roethlisberger and Dickson, 1939)
wiring.arg.*	Roethlisberger and Dickson's Bank Wiring Room Data, Arguments (Roethlisberger and Dickson, 1939)
wiring.fr.*	Roethlisberger and Dickson's Bank Wiring Room Data, Friendship (Roethlisberger and Dickson, 1939)
wiring.neg.*	Roethlisberger and Dickson's Bank Wiring Room Data, Negative Behavior (Roethlisberger and Dickson, 1939)
wiring.hlp.*	Roethlisberger and Dickson's Bank Wiring Room Data, Helping Others (Roethlisberger and Dickson, 1939)
wiring.job.*	Roethlisberger and Dickson's Bank Wiring Room Data, # Times Traded Job Assignments (Roethlisberger and Dickson, 1939)
wolfe.kin.*	Wolfe's Primate Data, Kinship (Borgatti et al., 1996)
wolfe.int.*	Wolfe's Primate Data, # Interactions (Borgatti et al., 1996)

Table 3: Observational Data

Observational Data		**Labeled**						**Unlabeled**					
Set Name	$C_{\max}$	C	$\frac{C}{C_{\max}}$	$p < C$	$p > C$	$^{95\%}_{LB}$	$^{95\%}_{UB}$	**C**	$\frac{C}{C_{\max}}$	$p < C$	$p > C$	$^{95\%}_{LB}$	$^{95\%}_{UB}$
bkfrat.beh.dgd.nos	282.463	276	0.977119	0	1	289	298	264	0.934636	0	1	282	286
bkham.beh.dgm.nos	111.6406	104	0.93156	0	1	110	117	91	0.815115	0	1	105	107
bkoff.beh.dgm.nos	132.0143	130	0.984742	0.0008	0.9992	134	141	127	0.962017	0	1	129	132
bktec.beh.dgm.nos	100.5072	101	1.004903	0.015	0.985	102	109	96	0.955155	0	1	97	100
cntrytrade.bmg.nos	62.54938	51	0.815356	0	1	65	70	42	0.67147	0	1	61	63
cntrytrade.cm.nos	62.61669	64	1.022092	0.0234	0.9766	65	70	54	0.86239	0	1	61	63
cntrytrade.ed.nos	59.18107	59	0.99694	0.0042	0.9958	61	66	51	0.861762	0	1	57	59
cntrytrade.fla.nos	62.61669	59	0.942241	0	1	65	70	54	0.86239	0	1	61	63
cntrytrade.mfe.nos	49.34421	48	0.972758	0.0056	0.9944	50	55	39	0.790366	0	1	46	48
freeei.mes.dgm.nos.1	63.04508	43	0.682052	0	1	62	68	39	0.618605	0	1	58	60
freeei.mes.dgm.nos.2	59.7317	45	0.753369	0	1	59	65	41	0.686403	0	1	55	57
galaceocl.com.nos	42.93704	45	1.048046	0.56	0.44	42	48	39	0.908307	0.27	0.73	39	41
gama.neg.nos	24.70306	27	1.092982	0.4616	0.5384	25	30	23	0.931059	0.71	0.29	22	24
gama.pos.nos	24.70306	26	1.052501	0.2096	0.7904	25	30	23	0.931059	0.79	0.21	22	24
kapmine.mul.nos	22.00548	26	1.181524	0.912	0.088	23	27	19	0.863421	0.03	0.97	20	21
kapmine.uni.nos	18.86542	23	1.219162	0.9528	0.0472	19	23	17	0.90112	0.64	0.36	16	18
kaptail.ins.nos.1	53.5404	53	0.989907	0.3996	0.6004	51	57	46	0.859164	0.01	0.99	47	49
kaptail.ins.nos.2	65.93985	65	0.985747	0.279	0.721	63	70	58	0.87959	0.01	0.99	59	61
kaptail.soc.nos.1	106.0699	101	0.952203	0.0006	0.9994	106	113	99	0.933347	0	1	101	104
kaptail.soc.nos.2	125.5924	124	0.987321	0.001	0.999	127	134	121	0.963434	0.02	0.98	122	125
krackhiman.rep.nos	13.36921	14	1.047182	0.3438	0.6562	13	17	10	0.747987	0.01	0.99	11	12
padffam.bus.nos	16.67911	19	1.13915	0.6826	0.3174	17	21	15	0.899329	0.92	0.08	14	16
padffam.mar.nos	20.0084	22	1.099538	0.5292	0.4708	20	25	19	0.949601	1	0	17	19
szcid.dgm.nos	31.90977	34	1.065504	0.1722	0.8278	34	38	30	0.940151	0.12	0.88	30	32
szcig.dgm.nos	25.75599	31	1.203603	0.98	0.02	27	31	24	0.931822	0.35	0.65	24	25
taro.nos	34.5719	37	1.070233	0.6052	0.3948	35	40	32	0.925607	0.77	0.23	31	33
thuroff.inf.nos	25.13794	28	1.113854	0.619	0.381	26	30	23	0.914952	0.25	0.75	23	25
thuroff.org.nos	17.64039	11	0.623569	0	1	18	22	8	0.453505	0	1	15	17
wiring.arg.nos	18.26262	15	0.82135	0	1	19	23	14	0.766593	0	1	16	17
wiring.fr.nos	14.5344	16	1.100837	0.4016	0.5984	15	19	12	0.825627	0.13	0.87	12	14
wiring.gam.nos	22.21643	22	0.990258	0.018	0.982	23	27	18	0.810211	0	1	20	22
wiring.hlp.nos	13.80571	15	1.086507	0.3408	0.6592	14	18	12	0.869205	0.73	0.27	12	13
wiring.job.dgm.nos	5.721535	7	1.223448	0.4056	0.5944	6	10	5	0.873891	0.89	0.11	5	6
wiring.neg.nos	18.26262	21	1.14989	0.695	0.305	19	23	16	0.876107	0.2	0.8	16	17
wolfe.int.dgd.nos	44.16705	46	1.0415	0.1074	0.8926	46	51	42	0.950935	0.1	0.9	42	44
wolfe.kin.nos	10.67625	12	1.12399	0.5152	0.4848	11	15	9	0.842993	0.98	0.02	9	9

Table 4: Self-Report Data

Self-Report Data		**Labeled**						**Unlabeled**					
Set Name	$C_{\max}$	C	$\frac{C}{C_{\max}}$	$p < C$	$p > C$	$^{95\%}_{LB}$	$^{95\%}_{UB}$	**C**	$\frac{C}{C_{\max}}$	$p < C$	$p > C$	$^{95\%}_{LB}$	$^{95\%}_{UB}$
bkfrat.cog.dgd.nos	275.8342	265	0.960722	0	1	282	290	262	0.949846	0	1	276	279
bkham.cog.dgm.nos	130.3134	118	0.905509	0	1	130	138	110	0.844119	0	1	125	128
bkoff.cog.dgd.nos	150.2971	153	1.017984	0.0164	0.9836	155	162	145	0.964756	0	1	150	152
bktec.cog.dgd.nos	113.5881	118	1.038841	0.1544	0.8456	117	123	111	0.977215	0	1	113	115
freeei.acq1.dgm.nos	182.1962	144	0.790357	0	1	185	193	133	0.729982	0	1	179	182
freeei.acq2.dgm.nos	194.4871	137	0.704417	0	1	199	206	127	0.653	0	1	193	196
krackhiman.ad.nos	49.50621	51	1.030174	0.052	0.948	51	56	43	0.868578	0	1	48	50
krackhiman.fr.nos	39.16994	39	0.995661	0.0338	0.9662	40	45	33	0.842483	0	1	36	38
newfrat.dld.nos.1	35.26371	39	1.105953	0.5736	0.4264	37	41	34	0.964164	0.22	0.78	34	36
newfrat.dld.nos.2	35.26371	38	1.077595	0.2772	0.7228	37	41	35	0.992522	0.96	0.04	34	35
newfrat.dld.nos.3	35.26371	40	1.13431	0.8336	0.1664	37	41	35	0.992522	0.95	0.05	34	35
newfrat.dld.nos.4	35.26371	40	1.13431	0.8252	0.1748	37	41	34	0.964164	0.27	0.73	34	36
newfrat.dld.nos.5	35.26371	36	1.020879	0.0246	0.9754	37	41	35	0.992522	0.92	0.08	34	36
newfrat.dld.nos.6	35.26371	37	1.049237	0.0914	0.9086	37	41	34	0.964164	0.25	0.75	34	35
newfrat.dld.nos.7	35.26371	36	1.020879	0.0208	0.9792	37	41	33	0.935806	0.01	0.99	34	35
newfrat.dld.nos.8	35.26371	39	1.105953	0.569	0.431	37	41	35	0.992522	0.94	0.06	33	36
newfrat.dld.nos.9	35.26371	37	1.049237	0.0948	0.9052	37	41	34	0.964164	0.27	0.73	33	36
newfrat.dld.nos.10	35.26371	39	1.105953	0.5734	0.4266	37	41	35	0.992522	0.96	0.04	34	35
newfrat.dld.nos.11	35.26371	39	1.105953	0.5756	0.4244	37	41	34	0.964164	0.26	0.74	34	36
newfrat.dld.nos.12	35.26371	37	1.049237	0.0874	0.9126	37	41	34	0.964164	0.32	0.68	34	35
newfrat.dld.nos.13	35.26371	37	1.049237	0.094	0.906	37	41	34	0.964164	0.24	0.76	34	35
newfrat.dld.nos.14	35.26371	38	1.077595	0.2926	0.7074	37	41	35	0.992522	0.97	0.03	34	35
newfrat.dld.nos.15	35.26371	39	1.105953	0.5804	0.4196	37	41	34	0.964164	0.34	0.66	34	35
prison.nos	90.57006	86	0.949541	0.336	0.664	84	91	80	0.883294	0.51	0.49	79	81
sampson.bl.dg1.nos	21.281	21	0.986612	0.0778	0.9222	21	26	18	0.845667	0.09	0.91	18	20
sampson.des.dg1.nos	26.33661	26	0.987219	0.0482	0.9518	27	31	23	0.873309	0.06	0.94	23	25
sampson.dk.dg1.nos	23.20655	22	0.948008	0.0168	0.9832	23	28	21	0.904917	0.81	0.19	20	22
sampson.es.dg1.nos	25.25312	24	0.950377	0.0102	0.9898	25	30	22	0.871179	0.1	0.9	22	24
sampson.in.dg1.nos	24.97278	26	1.041134	0.2494	0.7506	25	30	23	0.921003	0.92	0.08	22	24
sampson.lk.dg1.nos.1	25.53	29	1.35935	0.8306	0.1694	26	30	24	0.940084	1	0	22	24
sampson.lk.dg1.nos.2	26.07132	29	1.112334	0.7204	0.2796	26	31	24	0.920552	0.79	0.21	23	25
sampson.lk.dg1.nos.3	25.80234	28	1.085173	0.537	0.463	26	31	24	0.930148	0.92	0.08	23	25
sampson.nin.dg1.nos	24.10808	24	0.995517	0.0746	0.9254	24	29	20	0.829597	0	1	21	23
sampson.pr.dg1.nos	20.60679	23	1.116137	0.707	0.293	20	25	19	0.922026	1	0	17	19

Table 5: Cognitive Social Structure Data

CSS Data		Labeled						Unlabeled					
Set Name	$C_{\max}$	C	$\frac{C}{C_{\max}}$	$p < C$	$p > C$	$^{95\%}_{LB}$	$^{95\%}_{UB}$	**C**	$\frac{C}{C_{\max}}$	$p < C$	$p > C$	$^{95\%}_{LB}$	$^{95\%}_{UB}$
krackoff.ad.nos.1	47.79696	50	1.046092	0.1372	0.8628	50	54	45	0.941482	0.04	0.96	46	48
krackoff.ad.nos.2	40.68204	44	1.081558	0.649	0.351	41	46	37	0.909492	0.03	0.97	38	40
krackoff.ad.nos.3	49.55118	52	1.04942	0.1426	0.8574	52	56	47	0.948514	0.01	0.99	48	50
krackoff.ad.nos.4	49.85646	53	1.063052	0.2584	0.7416	52	57	48	0.962764	0.02	0.98	49	50
krackoff.ad.nos.5	49.50621	55	1.110972	0.841	0.159	52	56	48	0.969575	0.21	0.79	48	50
krackoff.ad.nos.6	26.27206	22	0.837392	0	1	26	31	21	0.799328	0	1	23	24
krackoff.ad.nos.7	47.1551	47	0.996711	0.0058	0.9942	49	54	40	0.848265	0	1	45	47
krackoff.ad.nos.8	44.47594	46	1.034267	0.1148	0.8852	46	51	40	0.899363	0	1	42	44
krackoff.ad.nos.9	46.85574	52	1.109789	0.8446	0.1554	48	53	45	0.960395	0.25	0.75	45	47
krackoff.ad.nos.10	46.43363	48	1.033734	0.0942	0.9058	48	53	43	0.926053	0	1	44	46
krackoff.ad.nos.11	33.54085	38	1.132947	0.9336	0.0664	34	39	29	0.864617	0	1	30	32
krackoff.ad.nos.12	44.05954	45	1.021345	0.067	0.933	45	50	40	0.907862	0.02	0.98	42	43
krackoff.ad.nos.13	25.26422	26	1.029124	0.295	0.705	25	30	21	0.831215	0.01	0.99	22	23
krackoff.ad.nos.14	45.00652	46	1.022074	0.0564	0.9436	46	51	36	0.799884	0	1	42	45
krackoff.ad.nos.15	38.97165	37	0.949408	0.0022	0.9978	40	45	32	0.82111	0	1	36	38
krackoff.ad.nos.16	25.26422	27	1.068705	0.5378	0.4622	25	30	21	0.831215	0.03	0.97	22	23
krackoff.ad.nos.17	23.14473	26	1.123366	0.8286	0.1714	23	28	20	0.864128	0.45	0.55	20	21
krackoff.ad.nos.18	45.75056	46	1.005452	0.0214	0.9786	47	52	42	0.918021	0	1	43	45
krackoff.ad.nos.19	39.75231	44	1.106854	0.8376	0.1624	40	45	36	0.905608	0.04	0.96	37	39
krackoff.ad.nos.20	35.73518	34	0.951443	0.0048	0.9952	36	41	31	0.867493	0	1	32	34
krackoff.ad.nos.21	49.82358	48	0.963399	0	1	52	57	47	0.943329	0	1	48	50
krackoff.fr.nos.1	28.80609	33	1.145591	0.9686	0.0314	28	33	25	0.867872	0.12	0.88	25	27
krackoff.fr.nos.2	13.86537	18	1.298199	0.983	0.017	13	18	12	0.865466	0.99	0.01	11	12
krackoff.fr.nos.3	5.903404	8	1.35515	0.8922	0.1078	6	9	5	0.846969	1	0	5	5
krackoff.fr.nos.4	20.47741	25	1.220858	0.9846	0.0154	20	25	17	0.830183	0.1	0.9	17	19
krackoff.fr.nos.5	31.41493	30	0.95496	0.014	0.986	31	36	27	0.859464	0	1	28	30
krackoff.fr.nos.6	17.56525	22	1.252473	0.9876	0.0124	17	22	14	0.797028	0.03	0.97	15	16
krackoff.fr.nos.7	33.79467	34	1.006076	0.1036	0.8964	34	39	29	0.858123	0	1	31	32
krackoff.fr.nos.8	4.494889	10	2.224749	1	0	5	8	4	0.8899	0.61	0.39	4	5
krackoff.fr.nos.9	5.213085	7	1.342775	0.6078	0.3922	6	9	5	0.959125	1	0	5	5
krackoff.fr.nos.10	23.86728	25	1.047459	0.3942	0.6058	24	28	20	0.837967	0.04	0.96	21	22
krackoff.fr.nos.11	31.96272	33	1.032453	0.2648	0.7352	32	37	28	0.876021	0.04	0.96	29	30
krackoff.fr.nos.12	16.68122	21	1.258901	0.9828	0.0172	16	21	14	0.839267	0.63	0.37	14	15
krackoff.fr.nos.13	19.26198	20	1.038315	0.34	0.66	19	24	18	0.934484	1	0	16	17
krackoff.fr.nos.14	24.92082	26	1.043305	0.3822	0.6178	24	29	22	0.882796	0.59	0.41	21	23
krackoff.fr.nos.15	19.67231	21	1.06749	0.499	0.501	19	24	16	0.813326	0.05	0.95	16	18
krackoff.fr.nos.16	15.77125	19	1.204724	0.9652	0.0348	15	19	13	0.824285	0.78	0.22	12	14
krackoff.fr.nos.17	22.40524	21	0.937281	0.024	0.976	22	27	16	0.714119	0	1	19	21
krackoff.fr.nos.18	13.86537	18	1.298199	0.9846	0.0154	14	18	12	0.865466	1	0	11	12
krackoff.fr.nos.19	32.76379	33	1.00721	0.1228	0.8772	33	38	29	0.885124	0.1	0.9	29	31
krackoff.fr.nos.20	7.84201	8	1.020147	0.1464	0.8536	8	12	7	0.892628	1	0	6	7
krackoff.fr.nos.21	25.60385	29	1.132642	0.9288	0.0712	25	30	22	0.859246	0.36	0.64	22	23

Figure 1: Two Isomorphic Structures

Two Isomorphic Structures															
Structure A								Structure B							
0	1	1	1	0	0	0	0	0	0	1	0	1	0	1	0
1	0	1	1	0	0	0	0	0	0	0	1	0	1	0	1
1	1	0	1	0	0	0	0	1	0	0	0	1	0	1	0
1	1	1	0	0	0	0	0	0	1	0	0	0	1	0	1
0	0	0	0	0	1	1	1	1	0	1	0	0	0	1	0
0	0	0	0	1	0	1	1	0	1	0	1	0	0	0	1
0	0	0	0	1	1	0	1	1	0	1	0	1	0	0	0
0	0	0	0	1	1	1	0	0	1	0	1	0	1	0	0

Figure 2: Histogram of Normalized Complexity for the Structural Equivalence Experiment

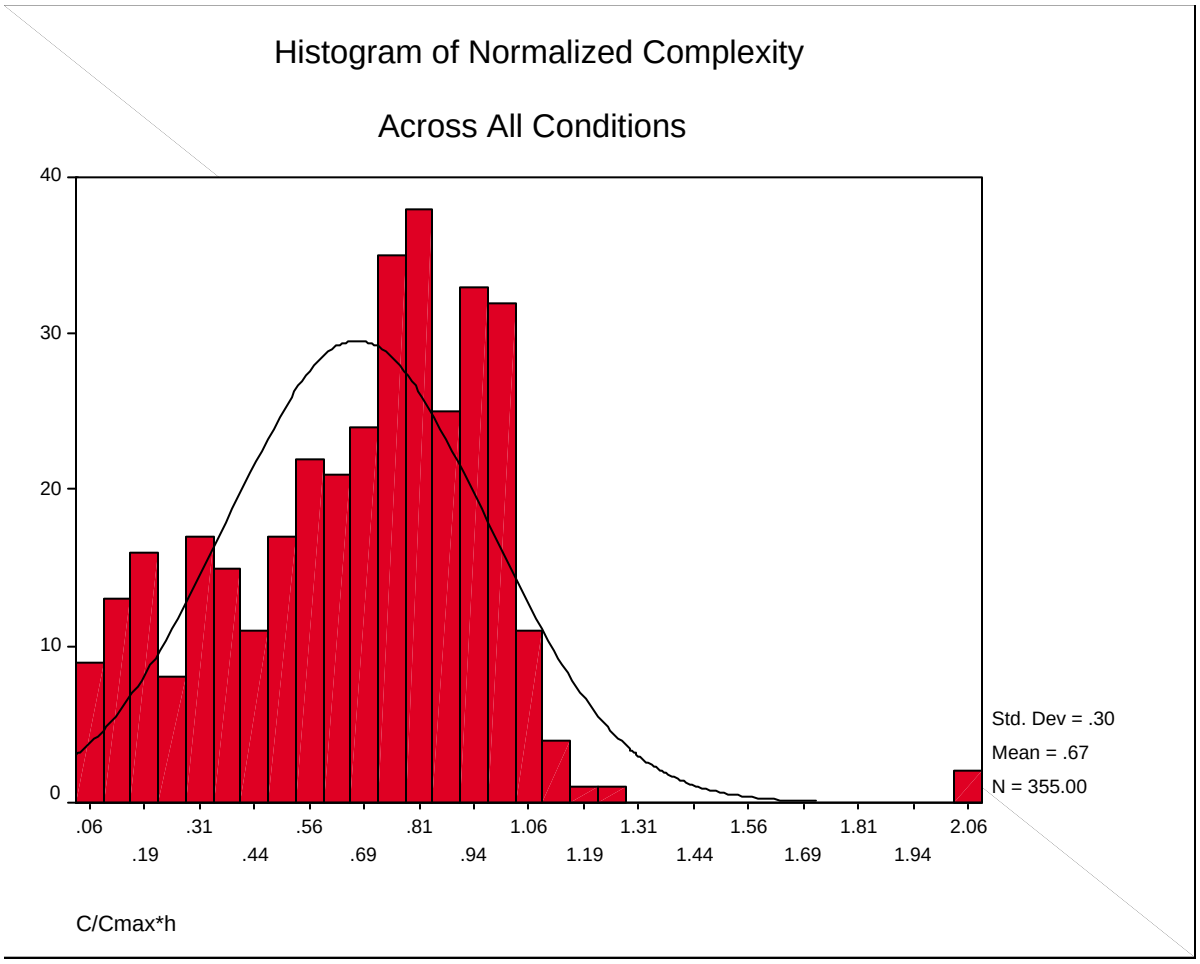

Figure 3: Boxplots of Normalized Complexity for the Structural Equivalence Experiment, by Block and Class Size

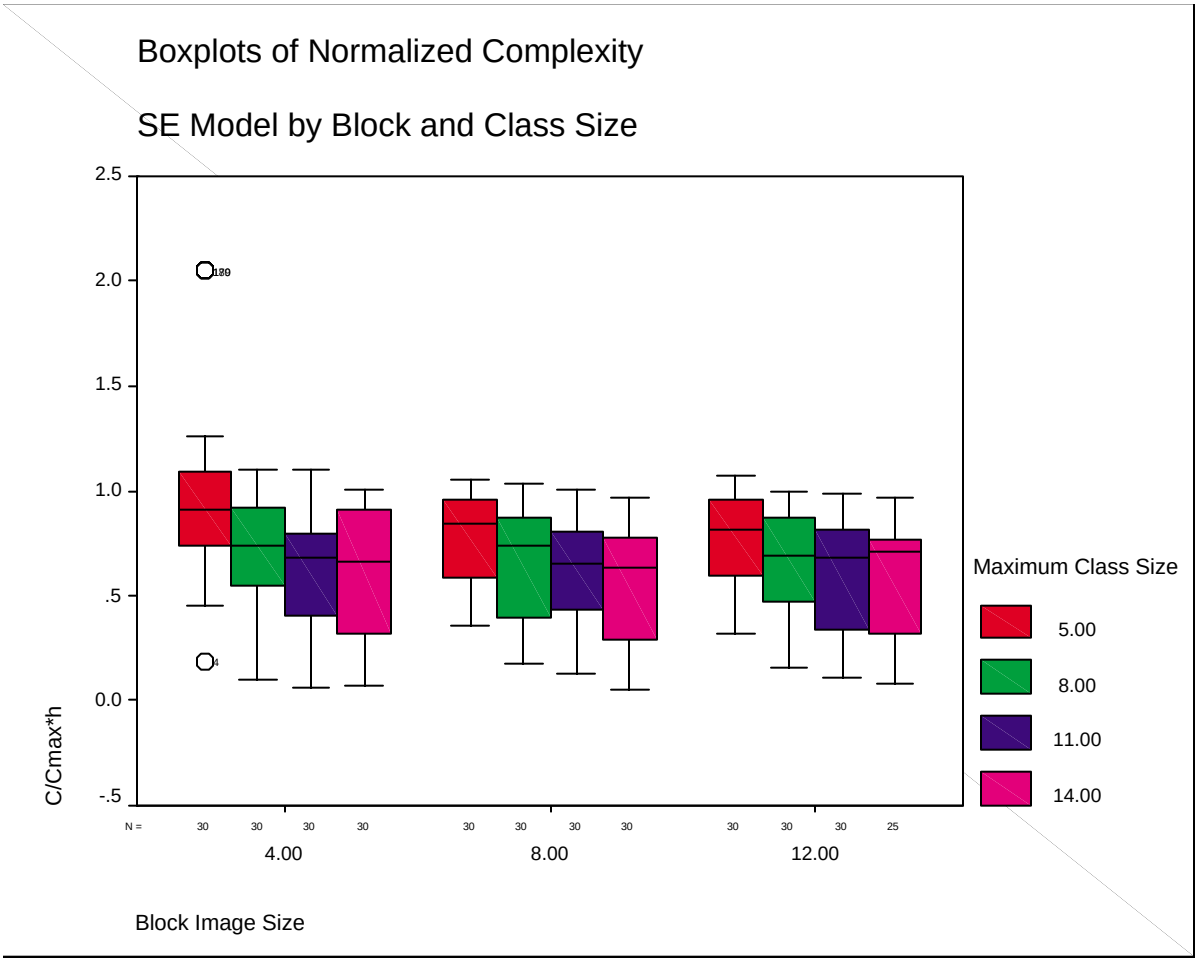

Figure 4: Boxplots of Normalized Complexity for the Structural Equivalence Experiment, by Block and Class Size

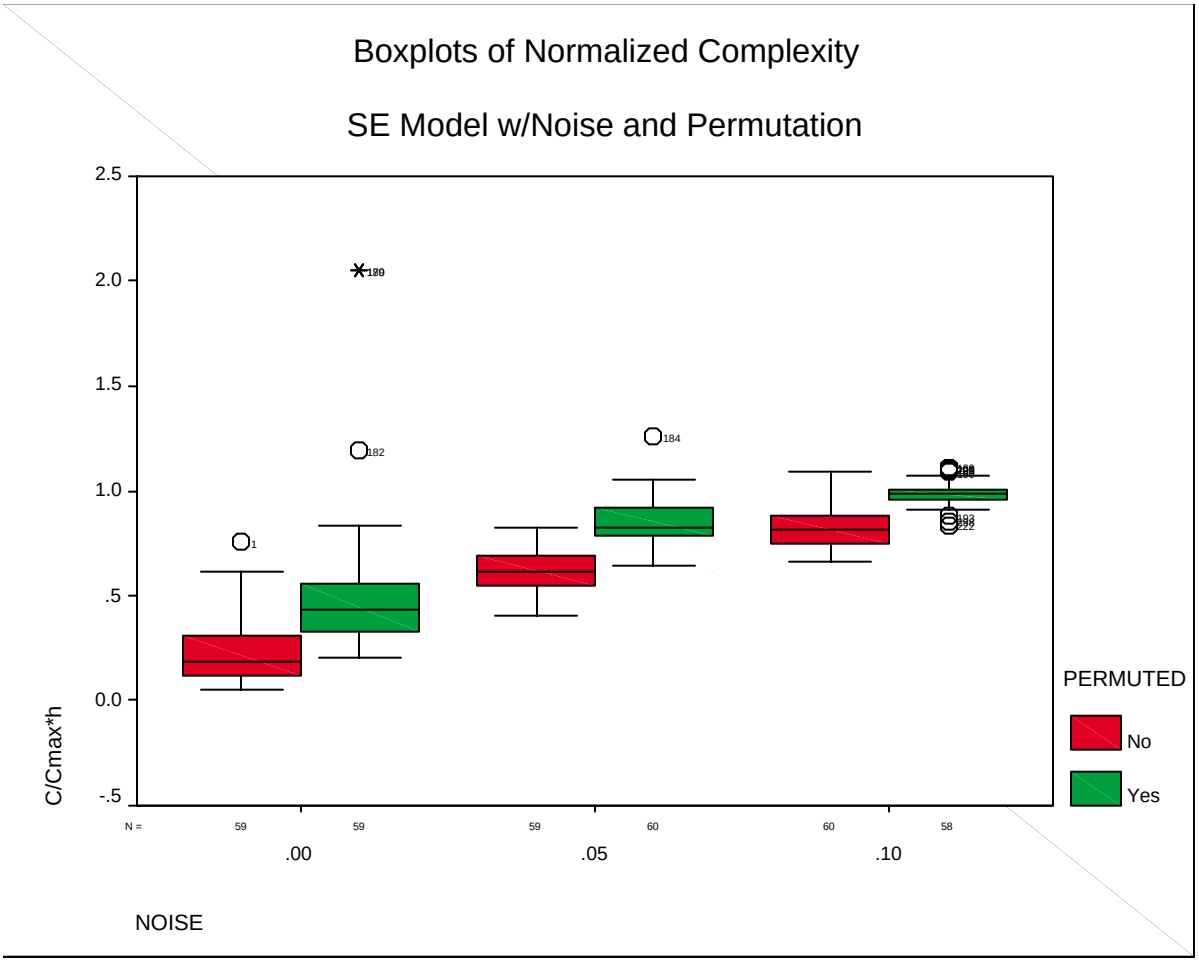

Figure 5: Lempel-Ziv Complexity of Labeled Structures – Observational Data

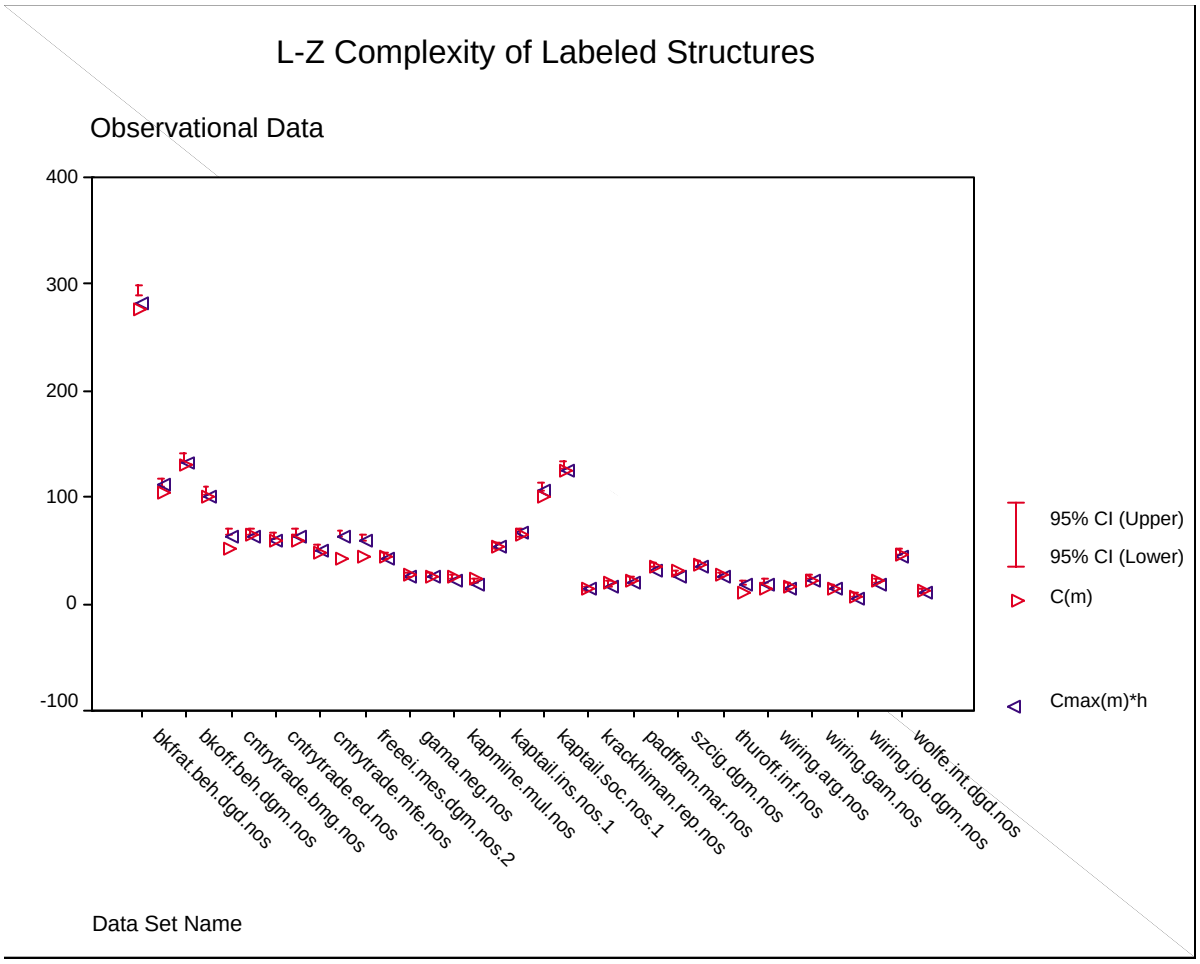

Figure 6: Lempel-Ziv Complexity of Unlabeled Structures – Observational Data

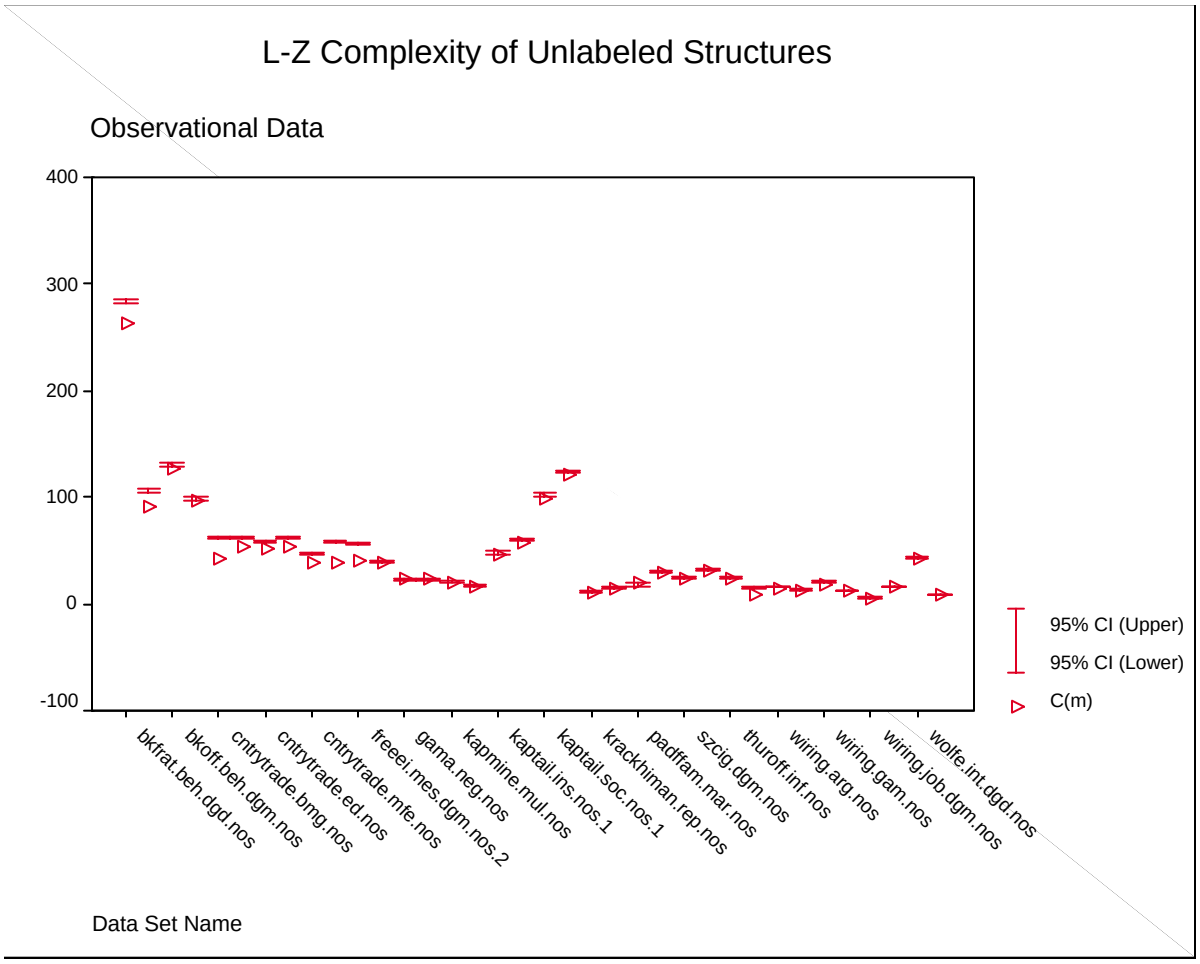

Figure 7: Lempel-Ziv Complexity of Labeled Structures – Self-Report Data

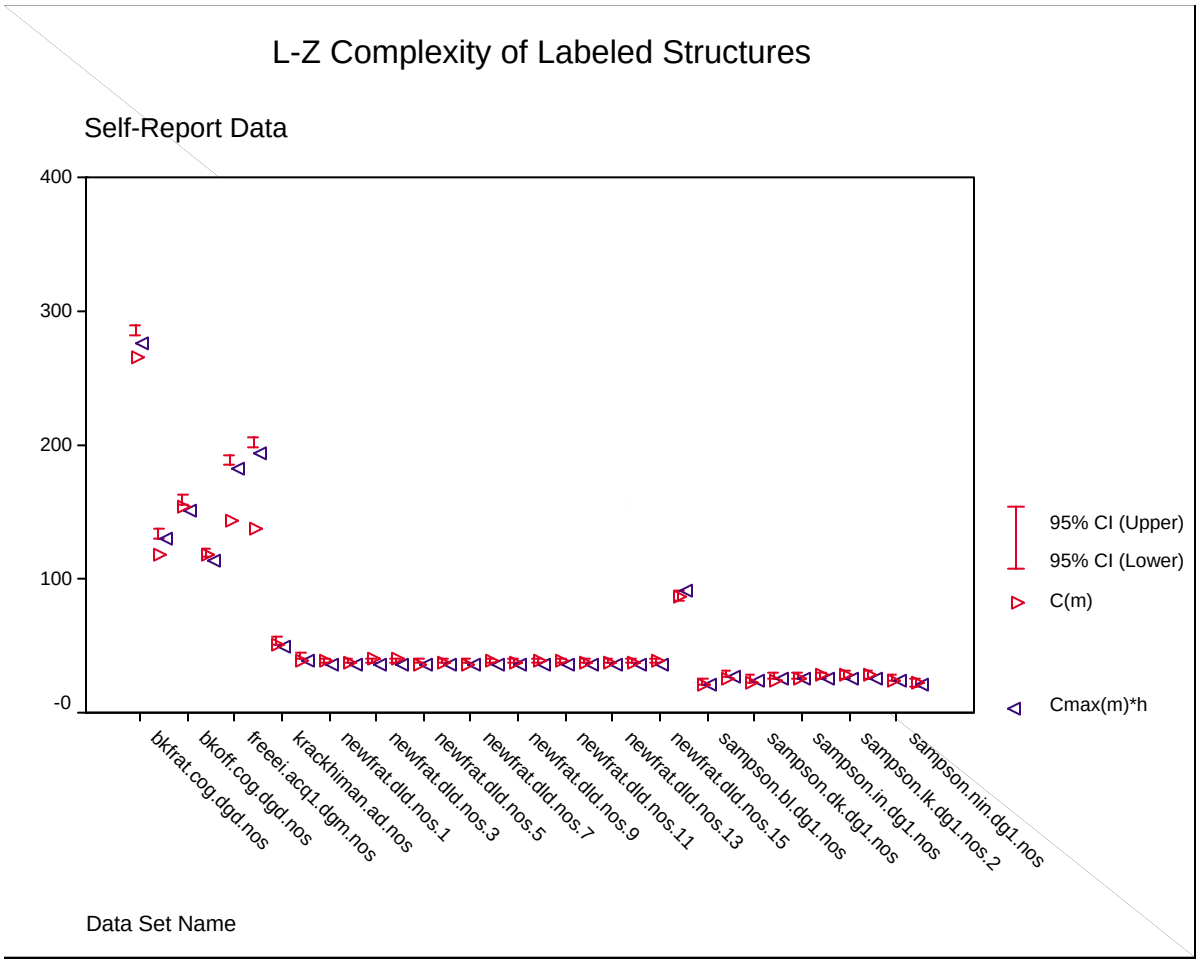

Figure 8: Lempel-Ziv Complexity of Unlabeled Structures – Self-Report Data

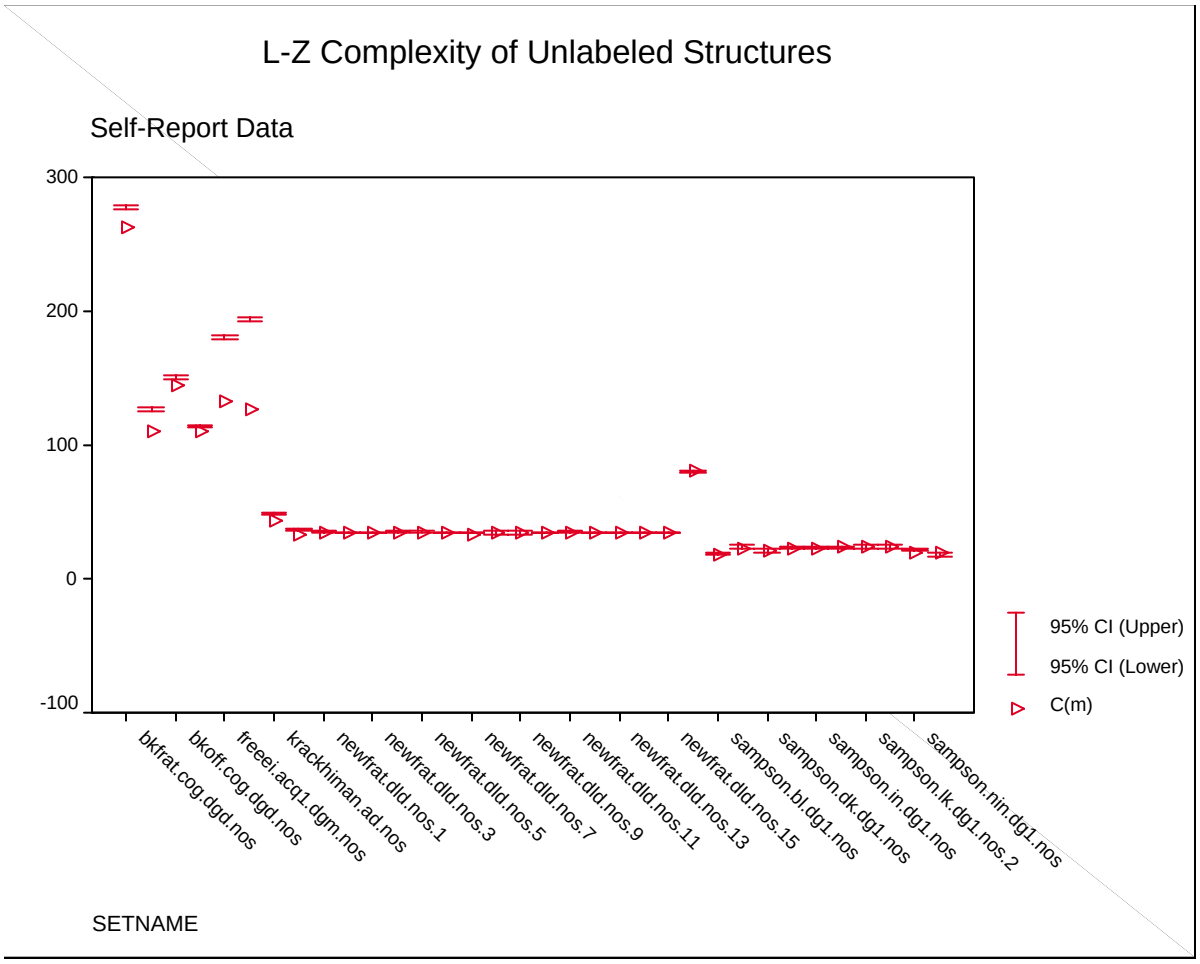

Figure 9: Lempel-Ziv Complexity of Labeled Structures – Cognitive Social Structure Data

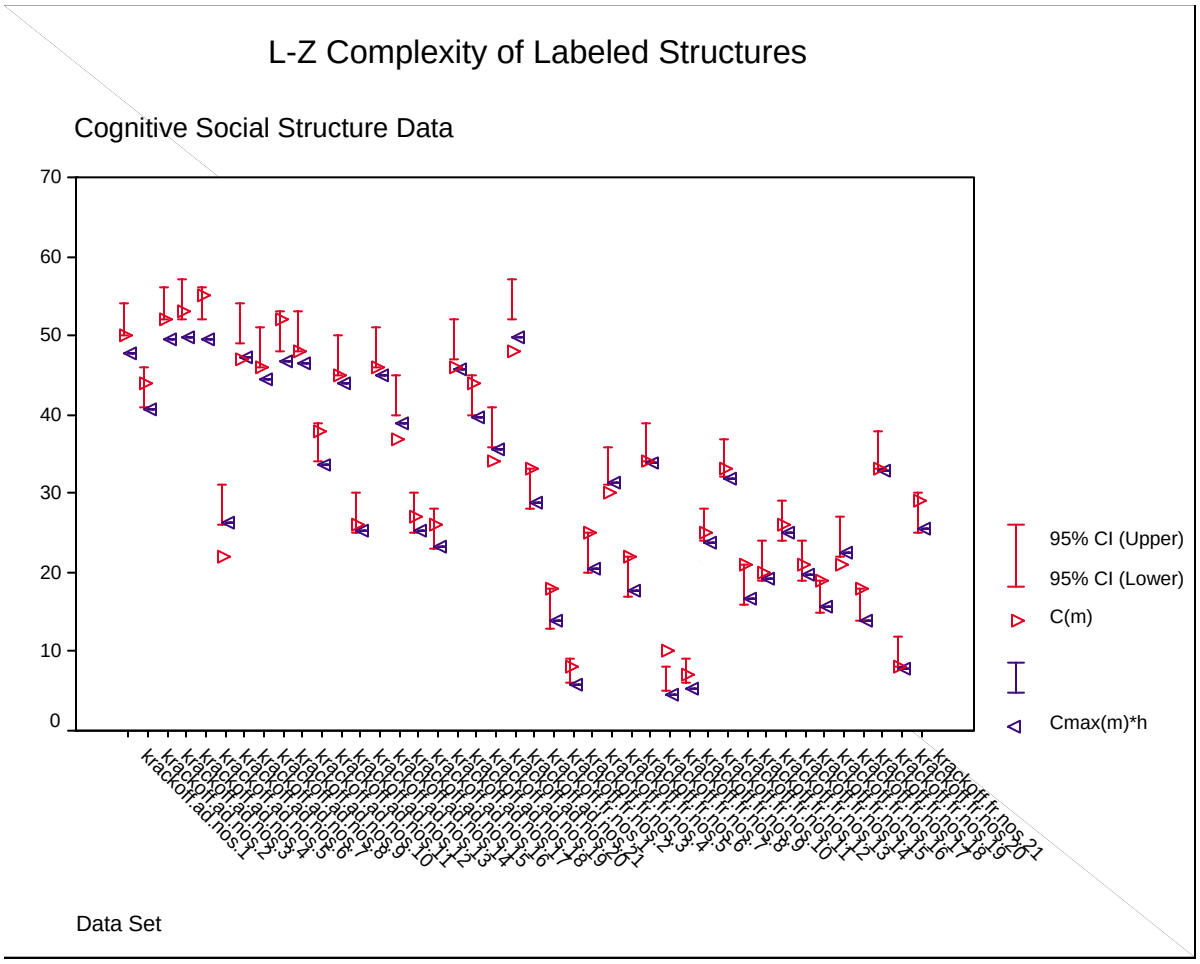

Figure 10: Lempel-Ziv Complexity of Unlabeled Structures – Cognitive Social Structure Data

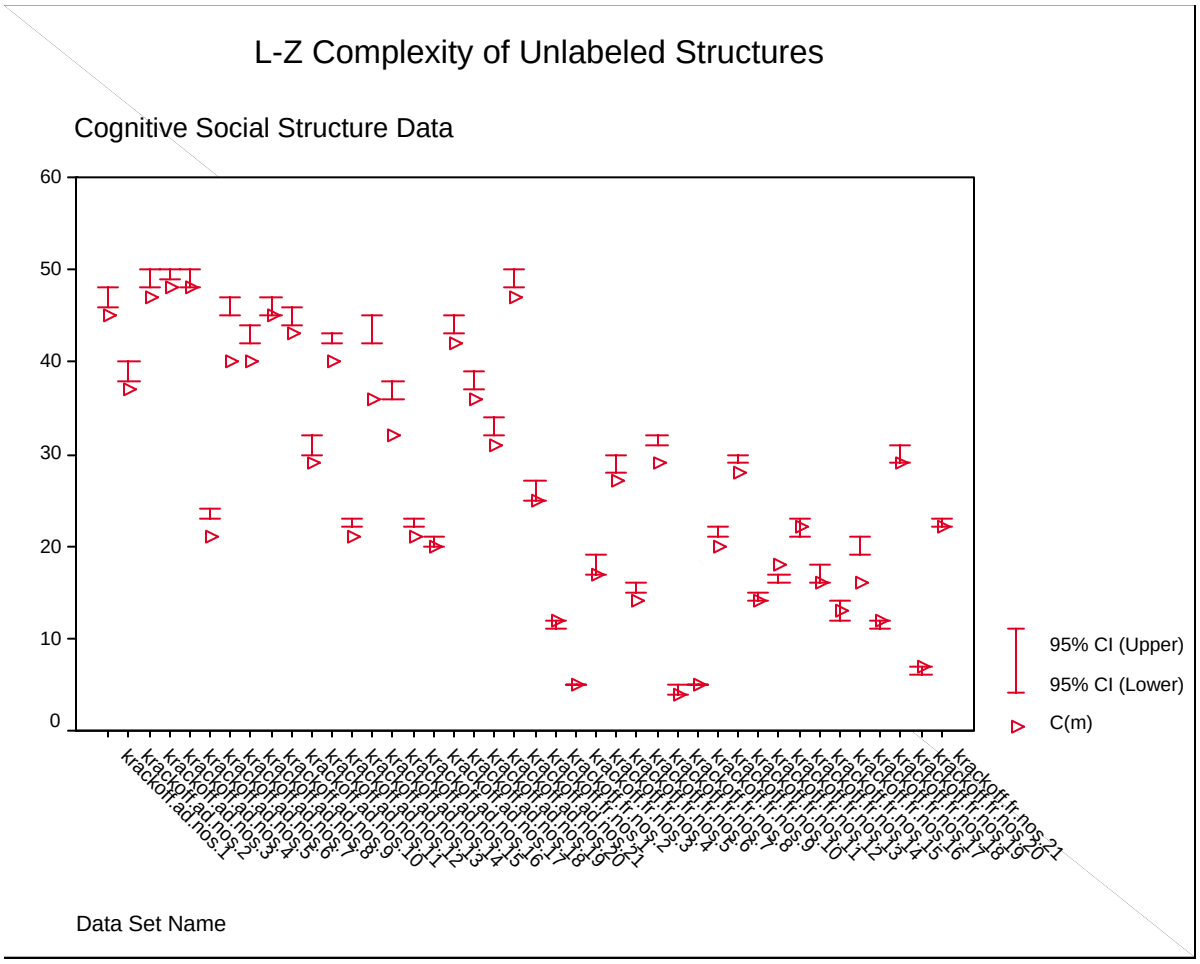

Figure 11: Lempel-Ziv Complexity of Labeled Structures, Normalized by Asymptotic Maximum Complexity

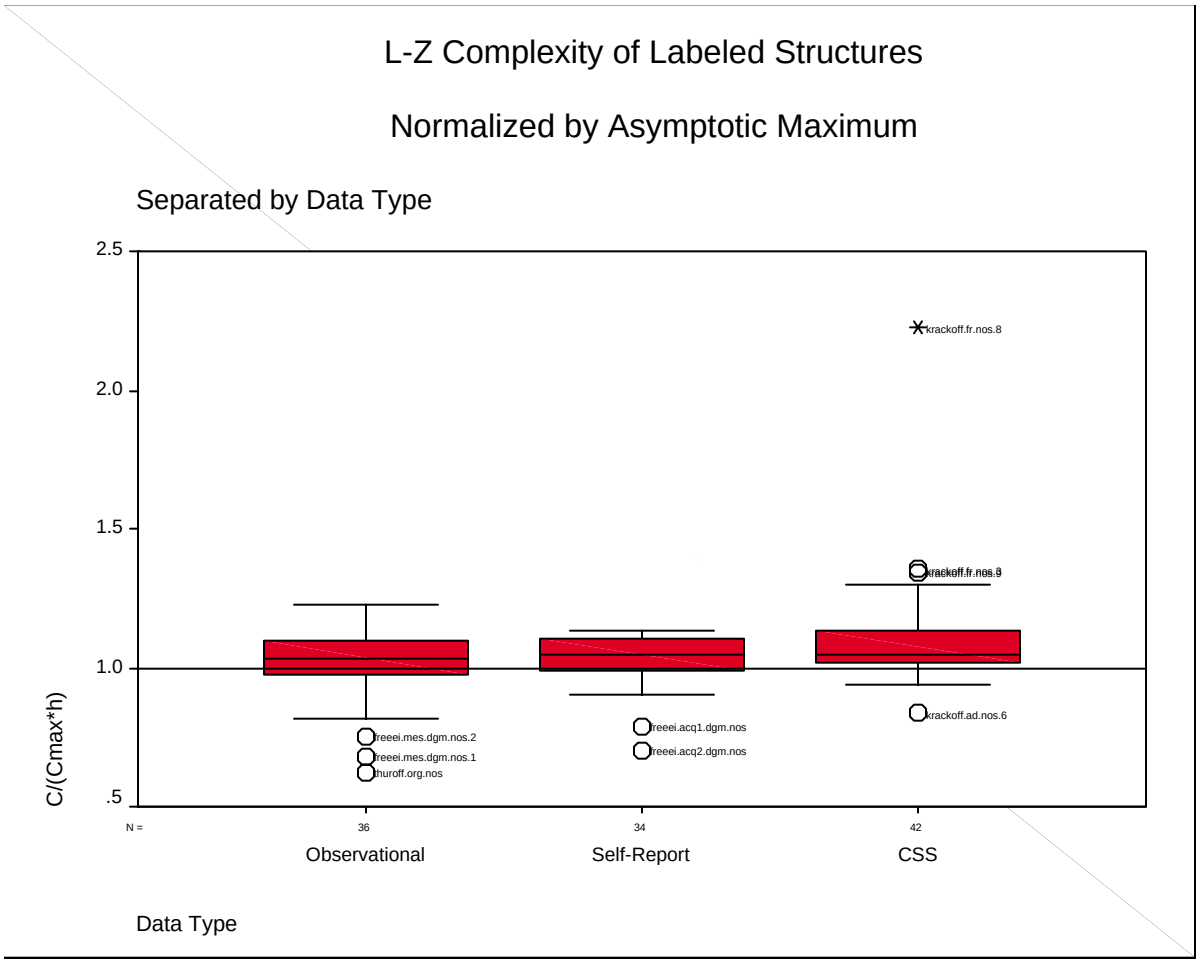

Figure 12: Lempel-Ziv Complexity of Labeled Structures, Relation to 95% Confidence Interval

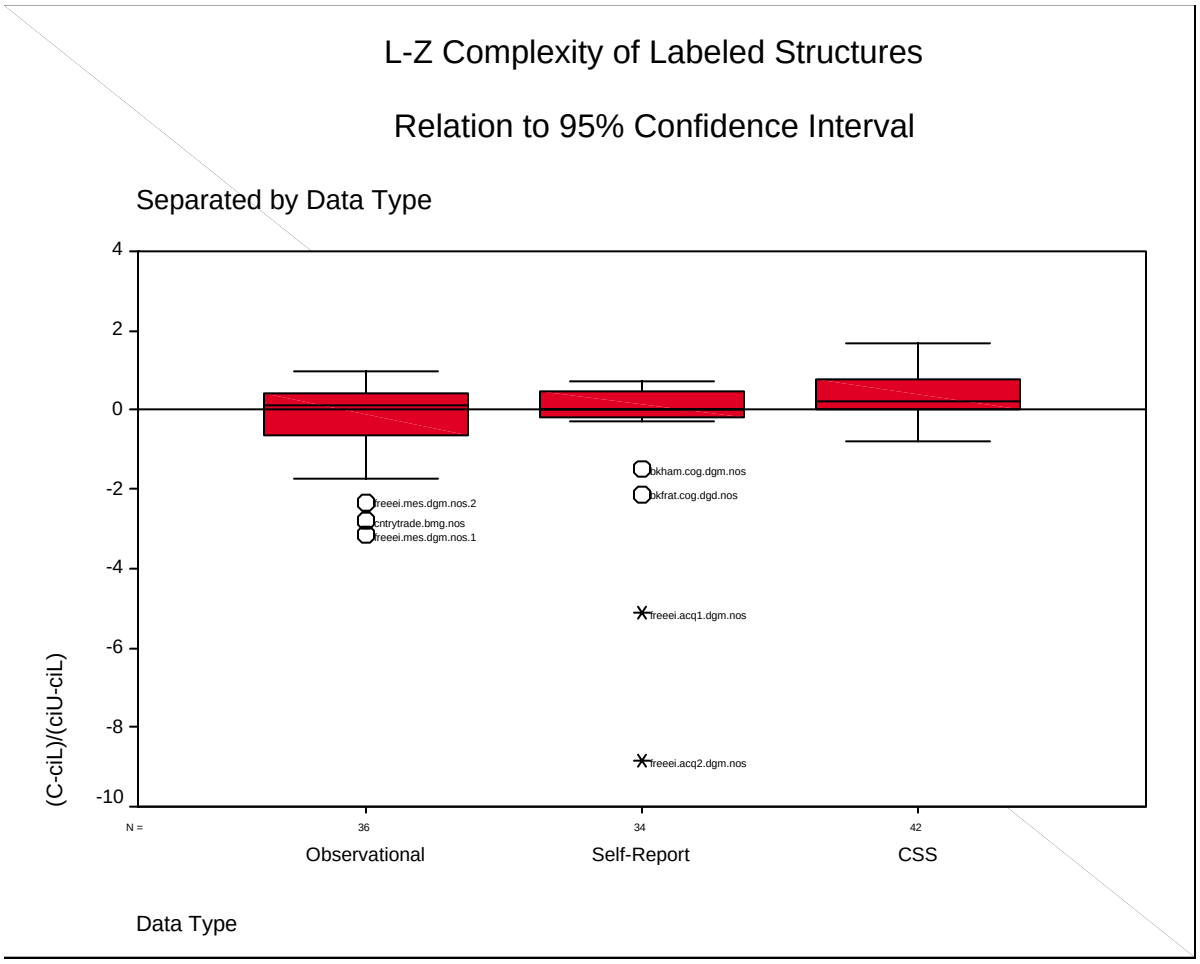

Figure 13: Lempel-Ziv Complexity of Unlabeled Structures, Relation to 95% Confidence Interval