Cluster Analysis of Unlabeled Structures[1]

Carter Butts[2]
Department of Social and Decision Sciences
and
Center for the Computational Analysis of Social and Organizational Systems
Carnegie Mellon University

April 24, 1998

[1]The author would like to thank Kathleen Carley for her input and encouragement.
[2]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 Organizational Systems and the Institute for Complex Engineered Systems at Carnegie Mellon University.

1 Introduction

A central problem for the structural analyst is the comparison of disparate social structures in order to identify underlying commonalties between them. This is especially true when large populations of structures are concerned; in such cases, it is often reasonable to ask whether there may be groups of structures which are especially similar to others within the population. Such questions, however, raise fundamental issues regarding the meaning of underlying structure, and of similarity between structures. These issues, ultimately, are central to the problem of describing the distribution of structures within a population.

In a variety of problems of interest, there is no theoretical justification for treating structural elements as *á priori* distinct from one another. This lack of distinction is equivalent to a notion of *interchangeability*: certain elements may be exchanged within the structure without doing violence to the theoretical basis for comparison. This situation, common as it is, poses serious difficulties for comparative work. In particular, the methods which exist for directly assessing the differences between structures are based on oriented representations such as labeled graphs; unlabeled or unoriented structures (the sort implied above) cannot be treated in this way.

Previous work by Banks and Carley (1994) and Butts and Carley (1998) has used the Hamming distance (Hamming, 1950) as the fundamental basis for comparison of directed graphs. The reasons for this choice have been detailed elsewhere, and will not be considered

at length here; however, it is worth noting that the Hamming distance forms a metric on the set of labeled digraphs, and that the properties of the Hamming distance with respect to various structural measures (such as the central graph) are reasonably well-understood (see (Banks and Carley, 1994; Butts and Carley, 1998) for more details). Work by Butts and Carley (1998) has further shown that the observed Hamming distance is an unreliable indicator of the difference between labeled structures, and have defined a emphstructural distance measure (closely related to the Hamming distance) which is applicable to unlabeled or partially labeled structures. Identifying the structural distance between directed graphs can be accomplished in a number of ways, including heuristic search and canonical labeling approaches; the identified pattern of distances may then be used in identifying the central graph, or in other analyses (Butts and Carley, 1998).

While previous research, then, has given us a valuable set of tools for the dyadic comparison of social structures and for the identification of central tendencies within collections of such structures, we do not have, as yet, techniques for addressing the more general question of weaker tendencies towards similarity within large graph sets. In particular, given some large collection of social structures we would like to be able to determine the extent to which the larger set may be expressed in terms of a smaller set of archetypal structures. In traditional data analysis, such questions are commonly answered via methods of *cluster analysis* (Johnson, 1967), which identify reductions of data which can be accomplished with minimal

introduction of error. Here, we will attempt to follow a similar approach, using techniques derived from traditional cluster analysis to characterize social network data. Of course, in our analyses each "data point" will in fact represent an entire social structure (rather than a vector of attribute values), a fact which will introduce some special concerns. Nevertheless, we expect to be able to demonstrate that a cluster analysis of underlying (i.e., unlabeled or unoriented) social structures is both possible and informative, using a sample analysis of data collected on work teams in a Carnegie Mellon University information systems program.

2 Methodology

Following earlier work by Banks and Carley (1994) and Butts and Carley (1998), we here employ the Hamming distance (Hamming, 1950) as our basic measure of difference between structures. In particular, given two labeled digraphs, $\mathbf{H_i}$ and $\mathbf{H_j}$ with vertex sets $V_i = V_j = V_U$ and edge sets E_i and E_j respectively, we may define a metric distance between them as per Hamming (1950). First, we define an indicator function $\delta_h(x, y)$ such that

$$\delta_h(x,y) = \begin{cases} 1 & if \exists e\,(v_x, v_y) \in E_h \\ 0 & otherwise \end{cases} \tag{1}$$

The function δ permits us to count directed edges within a given labeled digraph. To derive the Hamming distance between our two labeled digraphs, then, we simply count the number of directed edges which exist in one graph and not the other. This gives us the following

expression for the Hamming distance:

$$D\left(\mathbf{H_i}, \mathbf{H_j}\right) = \sum_{x=1}^{|V_U|} \sum_{y=1}^{|V_U|} \left|\delta_i\left(x, y\right) - \delta_j\left(x, y\right)\right| \tag{2}$$

As noted above, previous work by Butts and Carley (1998) has shown that the observed Hamming distance between two labeled graphs may be decomposed into a minimal, structural distance which depends only on the underlying unlabeled graphs, and an additional labeling distance which is a function both of the underlying unlabeled graphs and their respective labelings. For $\mathbf{H_i} = L_i(\mathbf{G_i})$ and $\mathbf{H_j} = L_j(\mathbf{G_j})$ (where L represents a labeling on the unlabeled graph $\mathbf{G}$), this decomposition gives us

$$D_O\left(L_i\left(\mathbf{G_i}\right), L_j\left(\mathbf{G_j}\right)\right) = D_S\left(\mathbf{G_i}, \mathbf{G_j}\right) + D_L\left(L_i\left(\mathbf{G_i}\right)\right), L_j\left(\mathbf{G_j}\right)) \tag{3}$$

where D_L represents the labeling distance and the structural distance D_S is given by

$$D_S\left(\mathbf{G_i}, \mathbf{G_j}\right) = \min_{\forall a,b} \left(D_O\left(L_a\left(\mathbf{G_i}\right), L_b\left(\mathbf{G_j}\right)\right)\right) \tag{4}$$

Minimization of the labeling distance (D_L) between pairs of graphs can be achieved as described above via a canonical labeling algorithm, but other approaches are available for problems of dyadic comparison. As Butts and Carley (1998) suggest, heuristic search techniques such as Monte Carlo sampling and genetic algorithms provide possible alternatives for finding the structural distance between a given pair of graphs, and have the distinct advantage of being tunable to adjust performance based on the problem under study.

With the above in mind, we here employ a genetic algorithm (Holland, 1975) to search the space of labelings to find the structural distance between unlabeled graphs. The particular algorithm in question operates via sequences of paired "switches" (e.g., $1 \mapsto 2$, $2 \mapsto 1$) which may vary in length up to a maximum of (for this sample) 100 such exchanges. The population for the algorithm is of size 50, with 50 generations permitted per search and 5 searches attempted per distance measured. The selection mechanism for the algorithm is double tournament selection[1], with fixed-point crossover and a flat mutation rate[2] of 0.10. The algorithm in question was evaluated on a number of sample graph sets, and was generally found to identify the structural distances within those sets[3].

Given the estimates of the structural distances (D_S) between graphs obtained via the genetic algorithm, we may construct a two-dimensional matrix of distances between all graphs within the set under analysis. This distance matrix, because it reflects a metric on the space

[1]That is, for each member of the new population, two pairs from the old population are selected (with replacement), the better of each acting as input for the crossover routine.

[2]Each mutation replaces an exchange instruction with one drawn from a uniform distribution of possible instructions.

[3]Although the search heuristic did produce some error in identifying the minimal structural distance, these errors were generally small in comparison to the correct distances. In the results presented here, these errors have been further minimized by applying a minimum symmetry rule to all distance matrices prior to final analysis.

of unlabeled graphs[4], can be unproblematically used as input data to standard cluster analysis routines (as might be Euclidean distances in a more conventional analysis). While a wide range of analyses could be employed (including multidimensional scaling, hierarchical clustering, K-means clustering, etc.) in order to characterize the extant similarities among graphs, we shall here utilize only hierarchical clustering. The algorithm used in this case is that of Johnson; only average-link clusterings will be examined in detail, although single-link and complete-link clusterings have also been considered. Extension of these methods to other algorithms should be obvious.

3 The IDS Data Set

The data for this analysis is drawn from an ongoing study of IDS[5] majors at Carnegie Mellon University (see Carley et al., 1993). As part of the IDS training program, advanced students are assigned to four to seven member project teams which must over a one year period design and implement an information system in order to meet the needs of an external client[6]. These groups are graded based on their ability to meet client requirements, and successful project completion is critical both to the awarding of degrees and to future job placement. For

[4]This result is shown in Butts and Carley (1998).

[5]Information and Decision Systems

[6]These clients are external to the IDS program, and often to Carnegie Mellon University itself. Local businesses, administrative departments, and the like are typical examples.

these reasons, the IDS work teams are highly motivated to be successful, and have strong incentives to accurately report their team structure[7].

As has been noted, data is collected on the IDS teams throughout the course of their training program. For our purposes, the data which is the most relevant are the responses to the social network questionnaires which are administered at three times during the training period. These instruments focus on three basic relations: ego's reporting being dependent on alter; ego's reporting alter's being dependent on ego; and ego reporting having worked with alter. Responses to these surveys are indicated on a scale ranging from 0 (lowest value) to 5 (highest value), and are collected for each member of each group with respect to each other member of that group.

The sample we will consider here consists of the set of junior-year work teams for the year 1990. During this year, there were eight such teams, six consisting of five members and two consisting of four. These eight groups were measured using each of the three instruments at three points during the project development cycle, resulting in a total set of seventy-two social structures. Although extensive information for each team (including performance, allocation of responsibility, and the like) is available, we shall here restrict ourselves to a simple exploratory examination of the unlabeled team structures within the

[7]Peer review is a part of the student evaluation process; students are familiarized with these practices in earlier IDS courses.

nine conditions. While this is far from adequate to characterize this data set, it is hoped that such an illustrative analysis will demonstrate the manner in which clustering methods may be applied to the examination of social structural data.

4 Data Analysis

In order to analyze the 1990 junior-year IDS work team structures, it was first necessary to dichotomize the survey data; this necessity stemmed from the definition of the Hamming distance, and from the conceptual tools described above. In order to reduce the data to adjacency form in a uniform fashion, all ties were defined to exist if and only if the reported tie strength from the initial survey was 3 or greater. This dichotomization divided the data at the midpoint of the reporting scale, and thus seemed to be a theoretically reasonable simplification. Other thresholds, of course, could be chosen for this analysis, but as our goal is to illustrate the methodology of graph clustering rather than to perform an in-depth analysis of the IDS data set, the simplicity of the midpoint threshold appears a sensible choice. After dichotomizing the network structures about the midpoint of the survey reporting scale, the aforementioned genetic algorithm was employed on each dyadic comparison within every condition in order to find the structural distances between the unlabeled team structures. From these distance counts, it was possible to construct generalized distance matrices repre-

senting the relationships between all graphs within each set. Three such matrices are given in Table 1 below:

Structural Distance Matrix: "Depends On" Relation																									
	Time 0								Time 1								Time 2								
G	0	1	2	3	4	5	6	7	G	0	1	2	3	4	5	6	7	G	0	1	2	3	4	5	6	7
0	0	4	3	3	5	2	3	3	0	0	5	6	4	3	6	4	5	0	0	3	3	7	5	2	4	4
1	4	0	5	5	4	4	6	4	1	5	0	3	5	7	4	7	5	1	3	0	5	6	6	4	3	5
2	3	5	0	4	8	0	2	3	2	6	3	0	4	2	6	6	4	2	3	5	0	3	4	3	3	4
3	3	5	4	0	3	6	2	5	3	4	5	4	0	5	6	6	6	3	7	6	3	0	3	8	4	3
4	5	4	8	3	0	8	7	4	4	3	7	2	5	0	7	5	4	4	5	6	4	3	0	5	4	5
5	2	4	0	6	8	0	4	3	5	6	4	6	6	7	0	2	5	5	2	4	3	8	5	0	5	4
6	3	6	2	2	7	4	0	2	6	4	7	6	6	5	2	0	4	6	4	3	3	4	4	5	0	5
7	3	4	3	5	4	3	2	0	7	5	5	4	6	4	5	4	0	7	4	5	4	3	5	4	5	0

Table 1: Structural Distance Matrix for the "Depends On" Relation, Reporting at Three Time Steps

Table 1 presents the distance matrices for the first network survey question at all three time points. Note that, as distance matrices, all are forced to be symmetric, with 0 distance on the diagonal (all groups are isomorphic to themselves). Note too that there are isomorphic structures within the sample: groups two and five, for instance, have identical dependency relations[8] at time zero. As can be seen, distances vary appreciably both between pairs of groups and over time. While there appears to be some tendency for very similar or disparate groups to remain so over time, this is not always the case: teams six and seven, for instance, have quite similar dependency relations in the first time step, and become dissimilar over

[8]Based on the underlying, unlabeled structures.

time.

Another look at dependency can be seen from the data presented in Table 2, below. This set of distance matrices is derived from ego reports of alters' depending on ego (as opposed to ego reports of depending on alters, which we have seen above). Naïvely, one would expect for the graphs of this second question to be close to those of the first; after all, the relation being reported continues to be dependency, and one might suppose that actors would have a fairly high degree of agreement as to who depends upon whom. As it happens, however, this is not the case. Comparison of the distance matrices in Table 2 with those in Table 1 reveals a substantially different distribution of structural distances between groups, which in turn implies that the underlying dependency structures of many (if not all) IDS teams are reported quite differently under the two instruments[9]. While interesting, this finding is not overly surprising; numerous previous studies (e.g., (Bernard et al., 1984), (Krackhardt, 1987)) have demonstrated that social structures derived from disparate positions are unlikely to be similar. A perhaps more subtle result in the same vein, however is that there is no obvious relation between trends on the two measures over time: teams six and seven, for instance, which appear to become increasingly dissimilar based on ego reports of dependency, appear to maintain roughly similar distances with respect to ego

[9]This could be further investigated by estimating the structural distances between team reports for the two instruments at identical time steps.

reports of alters' dependency on ego. While more rigorous, quantitative techniques would need to be applied in order to verify these initial observations, then, we can immediately see that the matrix of structural distances is a potentially powerful tool for answering a variety of questions regarding structure and structural change.

Structural Distance Matrix: "Other Depends On" Relation																								
Time 0									Time 1								Time 2							
G	0	1	2	3	4	5	6	7	G	0	1	2	3	4	5	6	7	G	0	1	2	3	4	5	6	7
0	0	4	5	5	4	7	6	5	0	0	4	4	3	5	5	6	3	0	0	5	2	8	6	2	3	1
1	4	0	5	2	3	6	2	4	1	4	0	3	3	9	2	4	3	1	5	0	3	6	7	5	4	4
2	5	5	0	3	6	3	2	1	2	4	3	0	3	6	2	3	2	2	2	3	0	4	5	5	4	2
3	5	2	3	0	5	4	4	4	3	3	3	3	0	6	3	7	4	3	8	6	4	0	3	4	7	5
4	4	3	6	5	0	8	6	6	4	5	9	6	6	0	11	5	7	4	6	7	5	3	0	5	6	6
5	7	6	3	4	8	0	3	2	5	5	2	2	3	11	0	3	3	5	2	5	5	4	5	0	1	3
6	6	2	2	4	6	3	0	3	6	6	4	3	7	5	3	0	4	6	3	4	4	7	6	1	0	2
7	5	4	1	4	6	2	3	0	7	3	3	2	4	7	3	4	0	7	1	4	2	5	6	3	2	0

Table 2: Structural Distance Matrix for the "Depends On" Relation, Reporting at Three Time Steps

With respect to variance in distance both within and between time steps, the data of Table 2 appears to show roughly the same behaviors as that of Table 1. The "Works With" relation reported in Table 3 (below) is not entirely novel, by comparison, but nonetheless has a few interesting features which bear pointing out. The first of these is the high prevalence of isomorphism on this relation: in the initial time step, two clusters of isomorphic groups are clearly present, with a smaller number of unique team structures distributed about them. This highly homogeneous state appears to break down over time: while some isomorphism

is present in later time steps, there is a clear reduction in the number of isomorphic team structures as early as time step 1.

Structural Distance Matrix: "Works With" Relation																									
Time 0									Time 1								Time 2								
G	0	1	2	3	4	5	6	7	G	0	1	2	3	4	5	6	7	G	0	1	2	3	4	5	6	7
0	0	4	3	4	5	4	1	4	0	0	7	7	3	3	6	7	6	0	0	3	3	6	6	6	1	6
1	4	0	3	6	6	5	4	4	1	7	0	3	3	5	0	4	3	1	3	0	4	5	6	3	2	2
2	3	3	0	0	10	0	0	2	2	7	3	0	2	6	1	4	0	2	3	4	0	4	5	5	4	3
3	4	6	0	0	4	4	2	4	3	3	3	2	0	2	4	6	1	3	6	5	4	0	2	4	6	4
4	5	6	10	4	0	8	8	8	4	3	5	6	2	0	5	3	5	4	6	6	5	2	0	10	6	8
5	4	5	0	4	8	0	2	3	5	6	0	1	4	5	0	3	1	5	6	3	5	4	10	0	5	0
6	1	4	0	2	8	2	0	2	6	7	4	4	6	3	3	0	5	6	1	2	4	6	6	5	0	4
7	4	4	2	4	8	3	2	0	7	6	3	0	1	5	1	5	0	7	6	2	3	4	8	0	4	0

Table 3: Structural Distance Matrix for the "Works With" Relation, Reporting at Three Time Steps

Another notable (if perhaps less encouraging) feature of the data in Table 3 concerns a side effect of graph isomorphism: what might be called "meta-structural" equivalence. Consider two isomorphic graphs; by definition, the structural distance between such graphs is 0. From the metric property of the structural distance, however, we can see that the underlying unlabeled structures associated with these graphs must be identical – as a result, the structural distances between the isomorphic structures and all other structures must be identical as well. This relationship, then, is a distance equivalent of structural equivalence, in that it is a condition which is fulfilled if and only if the relations between the equivalent structures and all other structures are identical.

Such an equivalence relation is an interesting consequence of the nature of isomorphism in this context, and surely poses no problems; what is somewhat discouraging is the fact that this equivalence is not perfectly observed in Table 3. To see this, consider the rows and columns of two isomorphic groups (e.g., teams two and three at time 0). By our above argument, the second and third rows (and columns) of the distance matrix should be identical. In fact, however, there are differences between the distances from group two to other graphs, and from group three to those same alters; this is a theoretical impossibility, and thus indicates that our heuristics for identifying the minimum Hamming distance between structures have not always been successful. On that point, it should be noted that these errors cannot be due to mistakenly identifying dissimilar structures as being identical. From the decomposition of the Hamming distance given in (3), we can see that failure to properly label the graphs can only result in an *increase* in the observed Hamming distance; an observation of zero distance, then, must be correct.

As our observations from Tables 1 through 3 suggest, it is possible to derive a great deal of useful information simply from an examination of the distance matrices in various conditions. Nevertheless, there are limits to the inferences which may be drawn in this fashion: in particular, it can be difficult to detect similarities among more than two groups at any given time, especially when those groups are not isomorphic. In order, then, to more carefully examine the distribution of structures in the IDS data set, we utilize the above

distance matrices to perform an average link hierarchical clustering (Johnson, 1967) on the structures derived from the "Depends On" relation at each time step.

The first such clustering is given below in Table 4. At each step in the clustering algorithm, exactly two existing clusters are merged; the clusters chosen are those with the lowest mean structural distance between them. Examination of the hierarchical clustering at various levels thus suggests the relative similarity of team structures, with the associated stress measure as an indication of the average number of ties which must be changed in each graph from the first group to map it onto the second[10] From Table 4, for instance, it can be seen that two major clusters (groups 0, 2, and 5, and 3, 6, and 7) exist at the 2.0 distance stress level. This suggests that these groups are in fact quite similar at time 0, and that they are more similar to each other than to the members of the other cluster (or to the two outliers). Such a finding of uniformity in the initial time step was not obvious from a simple examination of the distance matrix; it was revealed by consideration of the clustering process.

[10]This follows from the definition of the structural distance.

Hierarchical Clustering "Depends On": Time 0									
Step	Stress	Clusters							
0	0.000	0	1	2	3	4	5	6	7
1	0.000	0	1	5	3	4	5	6	7
2	2.000	5	1	5	3	4	5	6	7
3	2.000	5	1	5	6	4	5	6	7
4	2.000	5	1	5	7	4	5	7	7
5	3.000	7	1	7	7	4	7	7	7
6	4.000	1	1	1	1	4	1	1	1
7	4.000	4	4	4	4	4	4	4	4

Table 4: Average Link Hierarchical Clustering of "Depends On" Relation at Time 0

Continuing with our examination of the dependency relation, we present the clustering at time step 1 in Table 5 below. Immediately, the high stress levels induced by even modest clustering suggest to us that the IDS work teams have become substantially more diverse since the first survey was administered. Furthermore, the particular pattern of similarities has changed: groups 0 and 2 are still together, but now group 5 is more similar to 6 than to its former cluster members, and both 3 and 7 have become highly unusual relative to the distribution of structures. Interestingly, even as much has changed, some things abide: for instance, team 1 is still an outlier at time 1, as it was in time 2.

Hierarchical Clustering "Depends On": Time 1									
Step	Stress	Clusters							
0	0.000	0	1	2	3	4	5	6	7
1	2.000	0	1	4	3	4	5	6	7
2	2.000	0	1	4	3	4	6	6	7
3	3.000	4	1	4	3	4	6	6	7
4	4.000	7	1	7	3	7	6	6	7
5	4.000	6	1	6	3	6	6	6	6
6	5.000	6	3	6	3	6	6	6	6
7	6.000	3	3	3	3	3	3	3	3

Table 5: Average Link Hierarchical Clustering of "Depends On" Relation at Time 1

The outlying behavior of group 1 extends into time 2, as can be appreciated from the clustering in Table 6 below. Here, the teams as a whole appear to have become more similar once again[11]. Minimal clustering is found here; instead, chaining from a single, central structural type appears to dominate the pattern of similarities. One small exception to this lies ironically in groups 1 and 6, which (although fairly resistant to clustering) are more similar to each other than to the central cluster. Otherwise, it would seem that the IDS teams have, by time 2, begun to converge towards variants of a single structural archetype. While further analyses can be done to verify this observation more formally, even a simple perusal of the clustering patterns can tell us much about the distribution of IDS work team structures and their evolution over time.

[11] Although this could simply reflect increased error rates at time 1.

Hierarchical Clustering "Depends On": Time 2									
Step	Stress	Clusters							
0	0.000	0	1	2	3	4	5	6	7
1	2.000	5	1	2	3	4	5	6	7
2	3.000	2	1	2	3	4	2	6	7
3	3.000	3	1	3	3	4	3	6	7
4	3.000	4	1	4	4	4	4	6	7
5	3.000	4	6	4	4	4	4	6	7
6	4.000	6	6	6	6	6	6	6	7
7	5.000	7	7	7	7	7	7	7	7

Table 6: Average Link Hierarchical Clustering of "Depends On" Relation at Time 2

5 Conclusion

As we have seen, a viable approach to characterizing the distribution of social structures exists which draws upon the structural distance metric to establish a matrix of distances between unlabeled graphs. This distance matrix can then be used as an input to various techniques of cluster analysis, many of which have been well-developed in previous work. By examining one such analysis of information systems work teams over time, we were able to draw preliminary inferences regarding the relationships between structures at various points in time, and across various measures, as well as within particular conditions. Further extensions to this methodology might include the use of multidimensional scaling techniques,

K-means clustering, or other methods to obtain more formal characterizations of the distribution of structures. Similarly, the clusters which are discovered in this fashion may be directly characterized via their central graphs, permitting the use of traditional network analysis techniques on the summary structures to better understand network properties which are common across members of a large population. By continuing to develop methods of cluster analysis on unlabeled structures, we enhance our ability to characterize, and ultimately to predict, fundamental features of social structure.

6 References

Banks, D.L. and Carley, K. (1994) "Metric Inference for Social Networks," *Journal of Classification.*

Bernard, H.R., Killworth, P., Kronenfield, D., and Sailer, L. (1984) "The Problem of Informant Accuracy: The Validity of Retrospective Data," *Annual Review of Anthropology,* 13, 495-517.

Butts, C. and Carley, K. (1998) "Canonical Labeling to Facilitate Graph Comparison." Working Paper, Department of Social and Decision Sciences, Carnegie Mellon University.

Carley, K., Kiesler, S., and Wholey, D. (1993) "Learning Teamwork: Studies of Training in Software Development," *Proceedings of the 1993 Coordination Theory and Collaboration*

Technology Workshop. Symposium conducted for the National Science Foundation, Washington, D.C.

Hamming (1950) "Error Detecting and Error Correcting Codes," *Bell System Technical Journal*, 29, 147-160.

Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems*. University of Michigan Press: Ann Arbor.

Johnson, S.C. (1967) "Hierarchical Clustering Schemes," *Psychometrika*, 32, 241-253.

Krackhardt, D. (1987) "Cognitive Social Structures," *Social Networks*, 9, 109-134.