

**An Algorithmic Approach to the Comparison  
of Partially Labeled Graphs**

by

**Kathleen M. Carley  
Carnegie Mellon University  
Department of Social and Decision Sciences  
and  
H. J. Heinz III School of Policy and Management  
and**

**Carter Butts  
Carnegie Mellon University  
Department of Social and Decision Sciences**

**in Proceedings of the 1997 International Symposium on Command and  
Control Research & Technology. June, Washington, DC**

# An Algorithmic Approach to the Comparison of Partially Labeled Graphs

**Kathleen M. Carley**<sup>o</sup>  
Social and Decision Sciences  
and

H.J.Heinz III School of Policy and  
Management

**Carter Butts**  
Social and Decision Sciences

Carnegie Mellon University  
Pittsburgh, PA 15213

## Abstract

Aspects of  $C^2$  structures can be represented as graphs. In order for these graphs to be contrasted and compared, however, they must be labeled in the same way. This is rarely the case. Thus, there is a need for an algorithmic approach for labeling and aligning graphs in order to statistically contrast those elements of the  $C^2$  structure that can be represented as graphs. A color-splitting algorithm for labeling unlabeled graphs is presented. It is demonstrated that graphs that are labeled in this way can be aligned and the central graph located. Further, this alignment appears to minimize distance between the graphs.

## 1. Introduction and Motivation

Each unit or task force has its own somewhat unique  $C^2$  structure. Imagine for the moment that we have data on a dozen JTFs. On paper, the  $C^2$  architectures look somewhat different. The issues we want to address are "Are these architectures really different?" and "If these architectures are different, how different are they?". A related question is "How will we know when a new architecture has emerged?" Many aspects of the  $C^2$  structure can be represented as a series of

networks or graphs<sup>1</sup> with relationships among these graphs. Among such networks are those that represent the command or authority relations, the access to resources relations, the task assignment relations, the precedent ordering among tasks, and so forth. Once a  $C^2$  architecture has been chosen, then each of these sub-structures can be measured and represented as a network or graph. Using a variety of metrics differences or similarities in these graphs can be measured. Using such metrics it should be possible to classify  $C^2$  structures.

Unfortunately, there does not currently exist a commonly accepted taxonomy for classifying  $C^2$  architectures; indeed, within the organizational theory community debate rages over whether or not such a taxonomy is possible, let alone useful. McKelvey [1982] sees a need for such a taxonomy. Hannan and Freeman [1989], by contrast, argue that categories of organizational designs should be specified according to the interests of the researcher. Some schemes for classifying organizations have been based on strategy [Romanelli, 1989b] or product service [Fligstein, 1985]. Other researchers have classified organizations using multiple dimensions. For example, Aldrich and Mueller [1982] categorize organizations using the dimensions of technology, coordination, and control. Similarly, in contrast with these previous efforts, what we wish to suggest is a graph theoretic approach to this problem. Specifically, we here conceptualize organizational form, i.e., the  $C^2$  architecture as a set of interlinked graphs. We then attempt to develop a method of distinguishing alternative, and possible new, forms by locating those

---

<sup>o</sup> This work was supported by Grant No. N00014-97-1-0037 from the Office of Naval Research (ONR), United States Navy.

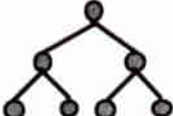

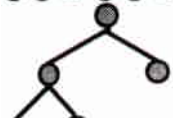
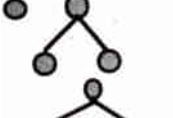
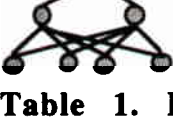
<sup>1</sup> We use the term network and graph interchangeably. However, in some literatures graph are equivalent to binary matrices and networks to weighted matrices.

structures which differ (statistically) significantly from others. We develop formal criteria for determining whether two  $C^2$  structures, given a set of measures of structure, are significantly different under a null hypothesis. Particular attention will be paid to developing a graph theoretical procedure for locating common networks and sub-networks. This is critical, as the ability to locate a common network is necessary for developing a mathematical criterion for determining whether the measures of structure for two different graphs are meaningfully different. In order to determine whether the difference between structures is significant we will need to be able to define the distributions on sets of graphs, and measure the similarity between graphs. Thus, by being able to define distributions on sets of graphs, in terms of such things as their central tendency, we hope to answer fundamental methodological and theoretical questions regarding  $C^2$  architectures. Our intent is to address the questions, given a set of graphs: "Can we define their distribution?", "Do these graphs exhibit some central tendency?" and "When can we say that two graphs are distinct?"

## 2. Background and Working Definitions

Network measures can be used to characterize graphs; similarly, statistically significant differences in these measures can be used to indicate differences in structure. At the network level, measures such as density, hierarchy, and graph connectivity are available for characterizing graphs [Krackhardt, 1994; Wasserman and Faust, 1994]. While most of these measures can be applied to any data that can be represented as graphs, whether or not they are meaningful depends on what data it is. For example, while span of control makes sense if the graph represents the command structure it makes less sense if the graph represents the precedence ordering among tasks.. Each aspect of the  $C^2$  architecture that can be represented as a graph will then have its own set of measures. We can then contrast graphs on the basis of differences in these measures. We illustrate this approach in Table 1. In Table 1, four hypothetical command structures are

shown, and their difference on a number of dimensions is indicated.

<i>Structures</i>	<i>Density</i>	<i>Span of Levels</i>	<i>Control</i>
	0.29	2.00	3
	0.29	6.00	2
	0.29	2.00	4
	0.48	3.33	3
			

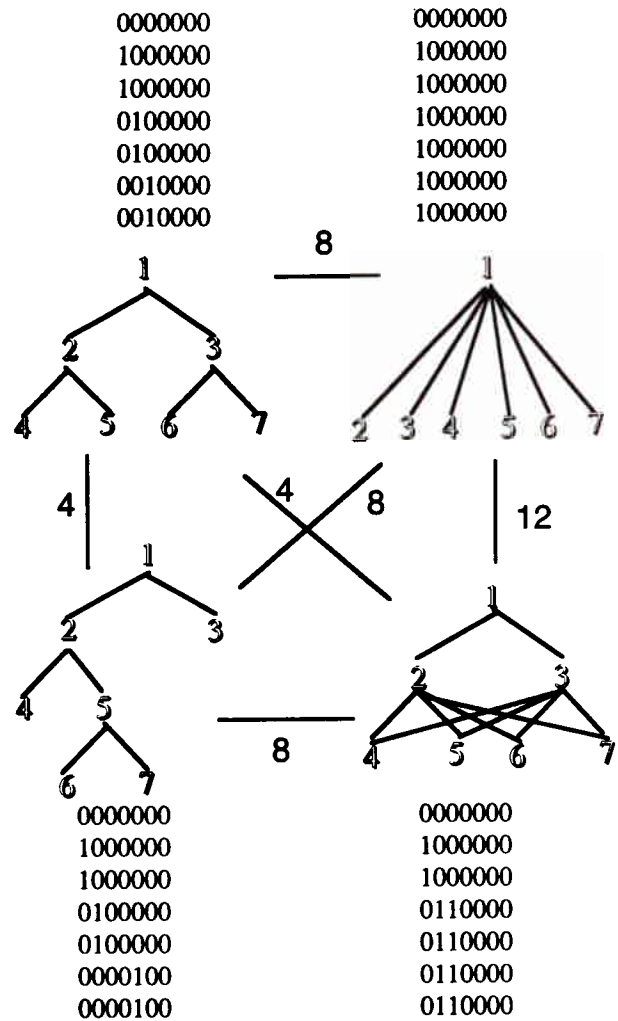
**Table 1. Illustrative Structures & Measures**

It is not our intent in Table 1 to exhaustively list all measures that are possible for these graphs. Rather, we have merely illustrated the types of measures possible for only one aspect of the  $C^2$  structure. The first point we wish to make is that this metric based approach is a possible approach for characterizing differences between graphs. Exactly what the metrics are for each aspect of the  $C^2$  architecture and whether or not any of these metrics are predictive of performance or other unit level behaviors is a separate issue. Indeed, we expect that  $C^2$  architectures with very different performance characteristics will be identical on some dimensions and different on others depending on which metrics are examined and which of the substructures the metrics are based on. Thus, for example, two  $C^2$  structures that have identical command structures but different task precedence orderings will exhibit different performance characteristics. While this is assuredly true, it does not detract from our second point that if we want to contrast and compare structures, or some single aspect of structure, that can be represented as graphs, then we will need to be

able to characterize the underlying distribution of these graphs.

Unfortunately, graphs with widely disparate configurations can look very similar given a set of network measures. This makes statistical comparison difficult. For example, all of the graphs in Table 1 are distinct structures yet each pair appear identical on a number of dimensions. On the one hand, this problem can be somewhat resolved by using a suite of measures for each structure that cover the range of implicit dimensions. Assuming that all of the dimensions have been characterized, structures that are really different should show up as different on one or more of these measures. The difficulty here, however, is defining a set of measures that exhaust all possibilities and so ensure coverage of all of the ways in which the graphs could be different. Further, this approach is unsatisfying as it does not let us simultaneously capture all differences.

An alternative approach is to map two graphs onto each other and then look for discrepancies in their overall structures. For the special case in which the two graphs to be compared are uniquely labeled, the difference between them can be readily captured by the Hamming distance [Hamming, 1950]. A graph is labeled if the nodes have names. Note any graph can be equivalently represented as a binary matrix with the number of rows/columns equal to the number of nodes and a cell with a 1 representing the presence of the link. The Hamming distance is simple the number of cells whose value needs to be flipped so that the two matrices come to be identical; i.e., the number of links that need to be added/dropped to make the two graphs identical. In Figure 1, the matrices corresponding to the networks in Table 1, and their hamming distance from each other is shown. Each of these matrices is an adjacency matrix showing what edges or links are present between nodes. In order to create Figure 1 we labeled each of the nodes in the structures in Table 1 using the proposed color-splitting algorithm we will describe. All edges are assumed to be uni-directional - from the lower level to the upper level.



**Figure 1. Hamming Metric and Illustrative Structures**

Previous work has demonstrated that, for sets of graphs in which all nodes are labeled, it is possible to derive a structural distribution and to locate its central graph. Banks & Carley [1994] developed a non-parametric network based statistical technique for locating the central graph<sup>2</sup>, the standard deviation, and confidence intervals. Banks and Carley measure the distance between networks using the hamming distance [Hamming, 1950]. The

<sup>2</sup> We use here the term central graph as we want to emphasize the relationship between this graph and the mean that one gets for variable level data. In other contexts, the terms consensus structure (Krackhardt, 1987), cultural consensus (Romney, Weller and Batchelder's, 1986), and majority intersection structure (Carley, 1984, 1986) have been used to denote the same basic idea.

central graph is that network containing the union of the node sets of the graphs from which it is constructed and in which two nodes are adjacent if and only if they have been adjacent in 50% or more of the graphs in the set. Whether or not a graph is significantly different than the central graph is assessed by comparing its hamming distance from the central graph with that which would be expected under the null hypothesis.

Currently, it is possible to determine whether or not two graphs are significantly different only for the special case in which all nodes are labeled (each node has a unique id) and in which both graphs share the same label set. This technique assumes a null hypothesis in which all links between nodes are independent and identically distributed [Banks and Carley, 1994]. Given this assumption it is possible to generate a distribution of networks from the sample population using non-parametric bootstrapping techniques, determine the first moment of this distribution by locating the central graph, and then calculate the distribution of distances from this central graph using the hamming metric. A non-parametric t-test (essentially) can then be used to determine whether the distance of the network(s) in question from the central graph are sufficiently large to reject the hypotheses that they are the same as the central graph. The central graph, hence, is the network equivalent of a mean for a non-network variable.

In many situations, however, the nodes are not labeled, or the labels are not relevant. For example, In one JTF two nodes may be MEU1 and MEU2; whereas in another these might be labeled MEU-A and MEU-B. These differences in labels may be due simply to documentation differences and not reflect real underlying structural differences. Given a set of unlabeled graphs it might still be possible to produce a partial labeling of nodes by "coloring" them: one might say that those nodes possessing some given characteristic are colored yellow, while those nodes with another characteristic are red, and so on. For example, all nodes representing resources are yellow and all nodes representing personnel are blue. Regardless, given a set of unlabeled graphs, or at best colored graphs, we might still want to identify any central tendencies present among these graphs.

For colored networks, however, there are many ways in which two networks can overlay, thus complicating the process of locating the central graph. For example, in Figure 2, a1 and a2 are the same color and therefore interchangeable. Two different matches may be found (simply in terms of node a) by either lining up the a1's and a2's or by lining up the left sides and the right sides. Still other matches are possible when the nodes of other colors are considered. Thus there are multiple ways in which the central graph can be calculated.

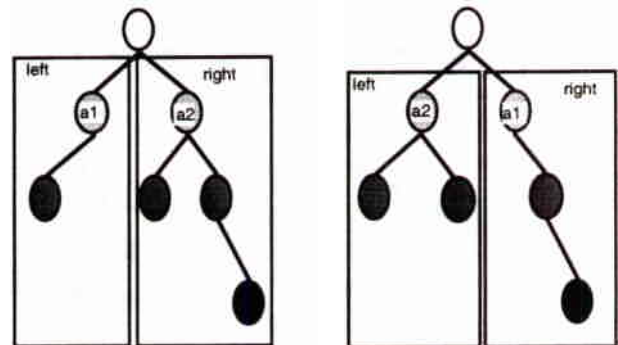
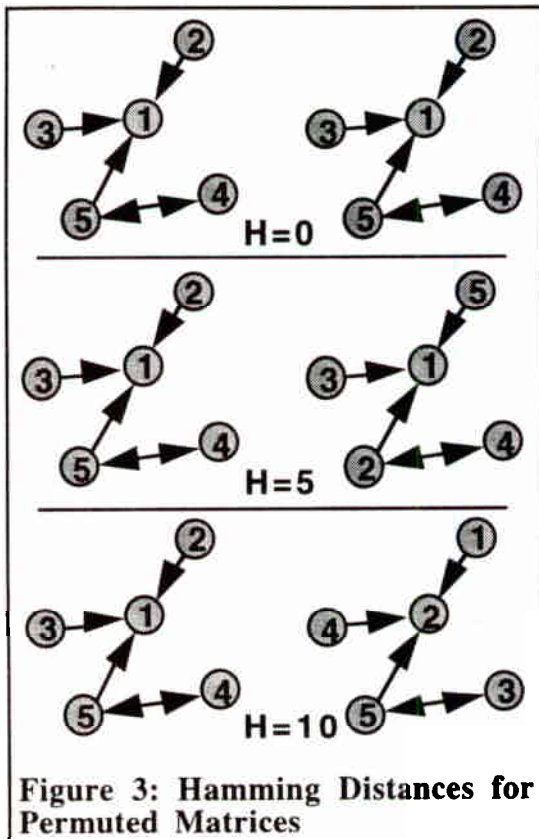


Figure 2. Illustrative Colored Networks

One approach to the comparison of unlabelled (or partially labeled) graphs is to re-label the nodes based on their network properties. Once a unique and complete labeling has been established, it is possible to use traditional methods for the assessment of structural distance. The choice of algorithm for labeling the nodes is critical however. The reason for this is that the Hamming metric is sensitive to minor permutations of nodal labels. An example of this may be seen in Figure 2. For these graphs the three comparisons differ from the minimum distance of 0 to the maximum distance of 10 - despite the fact that all six graphs are perfectly isomorphic! This phenomenon suggests that using arbitrary label choices may lead to poor inferences regarding structural distance: a critical problem for cases in which not all nodes (e.g., people, resources or tasks) are interestingly unique (that is, non-interchangeable in terms of the theory of interest). Imagine, for instance, two JTFs, each with an officer, M, two staff members (A & B), and two tasks (t1 & t2) such that in the first JTF A works on t1 and B works on t2 whereas, in the second JTF, A works on t2 and



B works on t1. Though these JTFs are functionally identical, a straightforward application of the Hamming metric would indicate a structural difference between the two. In order to correctly assess the difference between the two graphs, then, nodes of the same "color" (such as , in this case, staff members working under the same officer) must be treated as interchangeable; treating the structure as a conventional, pre-labeled graph is not a viable option.

If arbitrary labels yield arbitrary distances, what sort of labelings might prove more useful? In general, it would seem reasonable to seek a method of labeling nodes such that A) the Hamming distance between labelings of any two graphs will remain constant across (pre-labeled) permutations of those graphs, and B) the Hamming distance between any two labeled graphs will be minimized<sup>3</sup>. Finding a general

3 This follows from the fact that A) we would prefer for isomorphic graphs to have a Hamming distance of 0, and B) the Hamming distance can never fall below the minimum number of tie additions/subtractions needed to make the two graphs under comparison isomorphic. By minimizing the Hamming distance, we ensure that our comparison is as

method of achieving this goal, however, poses several problems. The first is simply one of combinatorics: for an unlabeled graph with N nodes, the number of possible labelings is equal to the number of node permutations, or N!. While, in theory, one could exhaustively search the space of labelings for the one which minimizes the Hamming distance between graphs, this method would be prohibitively costly for all but the smallest networks<sup>4</sup>. Genetic algorithms or other adaptive search mechanisms could be employed as well, but would be unlikely to yield uniform performance. As an alternative to either of these approaches, we propose a heuristic method of node labeling which exploits structural features of the networks on which it operates. A simple heuristic technique for locating structural similarity on colored nodes has been developed and applied to locating organizational structures [Carley, 1995c]. In this paper we intend to refine this technique, and to explore alternative graph theoretic approaches for coloring and re-arranging matrices in order to calculate the statistical significance of structural differences using non-parametric network based statistical techniques [Banks & Carley, 1994].

### 3. Implementation

Our intent is to expand the Banks & Carley [1994] technique for locating a central graph to the case in which the set of networks is constrained, and to the case where the nodes are colored rather than labeled. We will use Monte Carlo simulation techniques to generate the distribution of possible structures given the known organizational constraints on those structures.

There is no known technique for locating the optimal match between two colored networks, let alone locating the central structure based on the optimal match on a set of structures. This problem is at least NP hard for networks of more than two colors, consequently heuristic techniques are called

close as possible to the smallest number of changes required to convert one graph into another.

4 Just how prohibitive are these costs? To search the labeling space of a fairly modest (N=20) graph at a rate of 100,000 comparisons per second would take over 77,000 years; this is probably a bit long to wait for a single data point.

for. We will develop a “matching” or “alignment” heuristic for locating the optimal match between a set of unlabeled or colored networks. Once the reordering for the networks is determined, the previous techniques for locating the central graph and determining differences among networks can be applied. The approach and tools for generating these distributions and locating the central graphs should generalize for networks with other constraints and colorings than those we examine.

### 3.1 A Color-Splitting Algorithm for the Labeling of Graphs

The basic approach to the labeling problem which shall be considered here consists of a recursive algorithm which splits sets of identically colored nodes into subsets of identically colored nodes such that the colorings are unique between subsets. Nominally, this process terminates when no subsets with more than one member remain, thus producing a complete graph coloring.. (As we shall see, the algorithm will be unable to split certain color sets - the impact of this fact for algorithmic performance will be discussed.) The method by which subsets are recolored is based on an ordering principle which sorts nodes based on both local (e.g., degree) and global (e.g., connection to high-degree alters, number of directed walks to other nodes) network features; hence, the algorithm may be applied to individual graphs, and works independently of any within-color preliminary ordering.

This color-splitting algorithm works by “feeling out” a graph’s structure. One of the key elements used in this process is the set of paths between nodes. In particular, we shall be interested in the number of directed walks between nodes in the graph given by adjacency matrix  $\mathbf{A}$ <sup>5</sup>. Formally,

$$[1] \quad W_{d(i)} = \left( \mathbf{I} + \mathbf{A} + \mathbf{A}^T + \mathbf{A}\mathbf{A}^T + \mathbf{A}^T\mathbf{A} + \dots + \mathbf{A}^d + \mathbf{A}^T{}^d \right)_{i,j}$$

gives us the total number of directed walks from  $i$  to  $j$  which are of length  $d$  or less.

As has been noted, the basic problem of finding a reasonable labeling rests on using

structural properties of the network to produce a unique ordering of nodes. One obvious property which might be considered in this context is the row sum, or outdegree. In an authority or reporting structure the outdegree of a node would be the number of others that each node reports to. Because outdegree follows directly from the graph structure, it is invariant under permutation; furthermore, it is simple to calculate and compare.. Unfortunately, however, most networks involve numerous nodes with identical outdegrees<sup>6</sup>; hence, we cannot rely on this measure alone. For example, in many JTFs all nodes have an outdegree or row sum of 1 in the reporting structure. A simple extension of the row sum concept, however, can be effected by adding in the row sums of adjacent nodes. In particular, we shall here consider the structural row characteristic,  $SR_i(d)$ , of node  $i$  for distance  $d$ , to be the sum of the outdegrees of all nodes within  $d$  steps, weighted by the number of walks between the two nodes. More explicitly, we define  $SR_i(d)$  as

$$[2] \quad SR_i(d) = \begin{cases} \sum_{j=1}^N \left( W_{d(i)} \sum_{k=1}^N A_{jk} \right) & 0 \leq d \leq N-1 \\ 0 & d \geq N \end{cases}$$

Similarly, we can apply the same principle to the column sum (or in degree), yielding the structural column characteristic ( $SC_i(d)$ ):

$$[3] \quad SC_i(d) = \begin{cases} \sum_{j=1}^N \left( W_{d(i)} \sum_{k=1}^N A_{kj} \right) & 0 \leq d \leq N-1 \\ 0 & d \geq N \end{cases}$$

Note that, in both cases, the characteristics are only meaningful for  $d < N$ . This follows from the fact that the maximum path length for any graph is  $N-1$ ; walks longer than this distance are cyclical, and hence structurally redundant.

Now, we declare our structural characteristic coloring of an  $N$ -member network  $\mathbf{A}$  ( $SK(i)$  for all  $i$  in  $\mathbf{A}$ ), to be the ordered coloring which is determined by the following recursive process:

1) Let  $d=0$

<sup>5</sup> Throughout this discussion, we shall treat directed graphs and their sociomatrices as interchangeable.

<sup>6</sup> This is even more true of outdegree than indegree, due to the presence of time and energy constraints on actors’ nominating ability for many relationships.

- 2) Given a set of identically colored nodes,  $G$ , recolor members as follows:
- 3) (For all  $i$  and  $j$  in  $G$ ): If  $SR_i(d) > SR_j(d)$  then  $SK(i) > SK(j)$ ; if  $SR_i(d) < SR_j(d)$  then  $SK(i) < SK(j)$
- 4) (For all  $i$  and  $j$  in  $G$  such that  $SR_i(d) = SR_j(d)$ ): If  $SC_i(d) > SC_j(d)$  then  $SK(i) > SK(j)$ ; if  $SC_i(d) < SC_j(d)$  then  $SK(i) < SK(j)$
- 5) (For all  $i$  and  $j$  in  $G$  such that  $SR_i(d) = SR_j(d)$  and  $SC_i(d) = SC_j(d)$ ): If  $d = N-1$  then  $SK(i) = SK(j)$ , else let  $d = d+1$  and goto (2) for all sets  $G_1 \dots G_m$  such that  $SK(k) = SK(l)$  for all  $k$  and  $l$  in  $G_n$  and  $SK(k) < SK(l)$  for all  $k$  in  $G_n$  and  $l$  in  $G_{n+h}$  ( $h_0, m-nh > -n$ )

Once the structural characteristic coloring has been found, we can easily label the network in question by assigning node numbers in descending color order. If the graph still contains some identically-colored nodes, their specific ordering is irrelevant (so long as they are properly ordered with respect to all differently-colored nodes). In many cases, such non-degenerate color sets are due to structural equivalence [Lorraine and White, 1971]; hence, their orderings will not affect the Hamming metric<sup>7</sup>. It is possible, however, for algorithmic failures to cause non-unique labelings of some non-equivalent nodes<sup>8</sup>. The degree to which this affects assessments of Hamming distance can vary, but (as with any heuristic method) caution is advised.

### 3.2 Illustrative Application of the Color-Splitting Algorithm

Figure 4 presents a simple directed graph. Given that the graph is uncolored (or 1-colored), how would the foregoing algorithm determine a unique labeling?

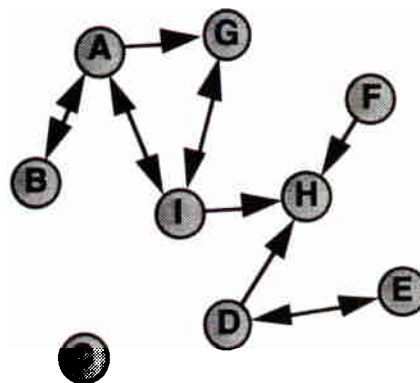


Figure 4: A 1-Colored Graph with Arbitrary Labels

To determine this, let us "run through" the instructions and observe the results. The intermediate results are also shown in Figure 5. Initially, our  $G$  consists of the nodes A-I (2), and  $d$  is equal to 0 (1). Proceeding to step (3), we note that:  $SRA(0)=3$ ;  $SRB(0)=1$ ;  $SRD(0)=2$ ;  $SRE(0)=1$ ;  $SRF(0)=1$ ;  $SRG(0)=1$ ;  $SRH(0)=0$ ;  $SRI(0)=3$ . Thus, we can already split  $G$  into  $\{A,I\}$ ,  $\{D\}$ ,  $\{B,E,F,G\}$ , and  $\{C,H\}$ .

In (4), we now attempt to split the identically-colored subsets, observing that:  $SCA(0)=2$ ;  $SCI(0)=2$ ;  $SCB(0)=1$ ;  $SCE(0)=1$ ;  $SCF(0)=0$ ;  $SCG(0)=2$ ;  $SCC(0)=0$ ; and  $SCH(0)=3$ . Applying the same ordering rule as we used in (3), we are able to arrive at the division  $\{A,I\}$ ,  $\{D\}$ ,  $\{G\}$ ,  $\{B,E\}$ ,  $\{F\}$ ,  $\{H\}$ ,  $\{C\}$ .

At this point (5), we note that  $d < N-1$  and our ordering is not degenerate; thus, we let  $d=1$  and return the two sets  $G_1=\{A,I\}$  and  $G_2=\{B,E\}$  to step (2).

	1	2	3	4	5	6	7	8	9
Initial	A	B	C	D	E	F	G	H	I
Outdegree ( $SR_i(0)$ )	A	I	D	B	E	F	G	C	H
Indegree ( $SC_i(0)$ )	A	I	D	G	E	B	F	H	C
$SR_i(1)$	A	I	D	G	B	E	F	H	C

Figure 5: A Sample Color-Splitting Process

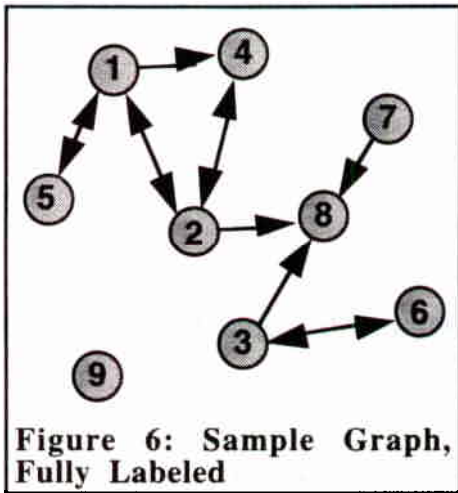
For (3)  $G_1$ , we can see that:  $SRA(1)=8$ ; and  $SRI(1)=7$ . This splits  $G_1$  into  $\{A\}$ ,  $\{I\}$ .

For (3)  $G_2$ , we find that:  $SRB(1)=4$ ; and  $SRE(1)=3$ . This final division splits  $G_2$  into  $\{B\}$ ,  $\{E\}$ .

<sup>7</sup> Structurally equivalent actors have identical relations with identical (in this case, identically-colored) alters. Thus, their ordering cannot change the value of any entry in a sociomatrix so long as they are a part of a coherent block which B) is in the same position vis a vis all other nodes.

<sup>8</sup> Informal observation seems to indicate that the algorithm is particularly vulnerable to regular equivalence.





At this point, the ordering {A}, {I}, {D}, {G}, {B}, {E}, {F}, {H}, {C} is degenerate, and we are finished. The descending-order labeling which results from this process can be seen in Figure 6.

### 3.3 Empirical Evidence for the Validity of the Color-Splitting Algorithm

The proposed color-splitting algorithm is generally effective at producing unique labelings of  $N$ -colored graphs. However, as we noted earlier, if the goal is to compare and contrast graphs it is not sufficient to just label the nodes. Rather, the labelings that are produced must be well-behaved with respect to the Hamming metric. This is especially true for cases in which the graphs to be compared are known to differ in some way; ideally, there will be a strong linear relationship between this underlying structural difference and the Hamming distance between the two networks.

To effect a preliminary test of the validity of the color-splitting algorithm for this specialized purpose, we ran a virtual experiment in which we generated a large number of graphs which were more or less typical of what might be expected for  $C^2$  structures with a small number of nodes. Each of these graphs was then copied and “tweaked”: that is, some number of ties (chosen at random) were “flipped”, so as to produce a slight difference between the two networks. The size of the tweak is the number of flipped ties. After being modified, the copied graphs were randomly permuted, and the color-splitting algorithm was executed on

each. Once the graphs were labeled, the Hamming metric was used to find the distance between the two graphs (the original and the permuted tweaked graph). More specifically, we generated graphs with either 5, 7, or 10 nodes, that were and were not hierarchies, with a mean tie probability of 0.01, 0.17333, or 0.36667, and a tweak between 0 and  $2N - (2N/5)$  by  $2N/5$  where  $N$  is the number of nodes. Thus for graphs of size 5 the tweaks were 0,2,4,6, or 8; for graphs of size 7 the tweaks were 0,2,4,6,8,10, or 12; and for graphs of size 10 the tweaks were 0,4,8,12, or 16. Each of these 306 cells was replicated twenty times. The mean hamming distance was recorded for all 20 cases within a cell. For this research hierarchy is defined as an upper triangular matrix; i.e., there are no cycles in the reporting structure.

Based on this data, for each size of graph a linear regression was run on the mean hamming distance (the results are shown in Table 2). Of the various graph parameters tweak has a big impact on the observed average Hamming distance. Specifically, the larger the tweak (the more ties that are flipped) the greater the average Hamming distance, even when other factors (such as the graph’s being a hierarchy) are controlled for. Furthermore, the optimal value (1.0) of the tweak coefficient is within the 95% confidence interval<sup>9</sup> for the parameter; hence, we have reason to suspect that the coloring algorithm labels graphs in a way which gives reasonable results.

	Coefficients	P-value
Intercept	-24.746748	1.6819E-18
N	3.684900	6.1998E-22
Hierarchy	7.660784	1.7749E-10
Density	29.540780	6.7093E-11
Tweak	0.774872	1.4544E-08
Adjusted R <sup>2</sup>	0.789153	
Observations	102	

**Table 2: Regression Statistics for the Hamming Test**

There are, however, a few cautionary notes. First of all, the test presented here is preliminary, and may not be fully indicative of the challenges to which the method may be put in the field. That is, the set of networks we

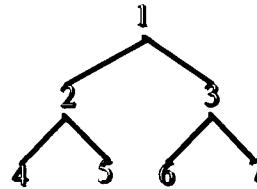
<sup>9</sup> 95% confidence interval for tweak: (0.526,1.023)

tested it on are not exhaustive of the set of all networks. In particular, we did not examine very large networks nor networks with a large number of bi-directional links. Secondly, it is important to note that factors other than tweak can also influence the Hamming distance. In Table 2 we see a massive effect due to density; however, this effect is primarily an artifact of the low real densities used for the simulation. Nonetheless, it is worth recognizing that density does produce some linear distortion in the Hamming metric under the coloring algorithm. Likewise, the presence or absence of strict hierarchy and the size of the network size can alter Hamming outcomes (albeit, apparently, in a fairly straightforward fashion<sup>10</sup>). How important are these effects for research purposes? While it is too early to be sure, there is reason to believe that the distortions are minimally problematic when comparing structures within a basic category (e.g., hierarchies, low-density networks, etc.), simply because the primary source of variance within categories is tweak (which is what we wish to capture in the first place). Comparison of differences across categories may be somewhat riskier. Further research must be done to determine the precise strengths and weaknesses of this heuristic approach. However, initial results suggest that it is both uniquely labeling non-isomorphic nodes and is doing so in a fashion that it becomes possible to locate the central graph.

### 3.4 Central Graphs

The central graph [Banks & Carley, 1994] is that graph which contains those edges in 50% or more of the set of graphs on which it is calculate. It is analogous to the mean for variable level data. After labeling the nodes for the illustrative graphs, as shown in Figure 1, we calculated the central graph. This central graph is shown in Figure 7. For the four original graphs, their hamming distance from this graph is (clockwise in Figure 1) 0, 8, 4, and 4. This is the minimal set of Hamming

distances of these graphs from the central graph.



**Figure 7: Central Graph for Illustrative Graphs**

### 4. Conclusion

The proposed color-splitting algorithm is a heuristic based algorithm for labeling unlabeled or colored graphs. Labeling the graphs basically aligns the structures. Once the graphs are labeled similarities and differences can be measured and the central graph calculated. Given the central graph, it is then possible to generate other features of the distribution and to statistically evaluate differences and similarities between graphs. This technique will be useful in determining differences in those aspects of C<sup>2</sup> structures that can be graphically represented.

### References

- [Aldrich & Mueller, 1982] H.E. Aldrich and S. Mueller, The evolution of organizational forms: technology, coordination, and control. In *Research in Organizational Behavior*, vol. 4, edited by B.M. Staw & L.L. Cummings. Greenwich, CN: JAI, 33-87, 1982/
- [Banks & Carley, 1994] D. Banks and K. Carley, Metric Inference for Social Networks. *Journal of Classification*, 11: 121-149, 1994.
- [Carley, 1995] K.M. Carley, Automatic Restructuring of the C<sup>2</sup> Structure and Performance. *Proceedings of the 1995 International Symposium on Command and Control Research and Technology*. June. Washington D.C., 1995.
- [Carley, 1986] K.M. Carley, "An Approach for Relating Social Structure to Cognitive

<sup>10</sup> Many of these effects may be traced to the fact that the Hamming distance counts the number of tie differences between matrices; hence, any ordering distortion will have an effect which is proportional to density and to network size.

Structure," *Journal of Mathematical Sociology*, 12, 137-189, 1986.

[Carley, 1984] K.M. Carley, Constructing Consensus. Unpublished doctoral dissertation, Harvard, 1984.

[Fligstein, 1985] N. Fligstein, 1985, The spread of the multi-divisional form among large firms, 1919-1979. *American Sociological Review* 50: 377-391, 1985.

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

[Hannan & Freeman, 1989] M.T. Hannan, and J. Freeman, *Organizational Ecology*. Cambridge, MA: Ballinger, 1989.

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

[Krackhardt, 1994] D. Krackhardt, (1994).. Graph Theoretical Dimensions of Informal Organizations. In *Computational Organization Theory*, edited by K.M. Carley and M.J. Prietula. Hillsdale, NJ: Lawrence Erlbaum Associates.

[McKelvey, 1982] B. McKelvey, , *Organizational Systematics: Taxonomy, Evolution, Classification*. Berkeley, CA: University of California Press, 1982

[Romanelli, 1989] E. Romanelli, Environments and Strategies of Organizational Start-up: Effects on Early Survival. *Administrative Science Quarterly*, 34: 369-387, 1989.

[Romney et al., 1986] A.K. Romney , S.C. Weller and W.H. Batchelder, "Culture as consensus: A theory of culture and informant accuracy," *American Anthropologist*, 88(2), 313-338, 1986.

[Wasserman & Faust, 1994] S. Wasserman, and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[Lorrain & White, 1971] F. Lorrain, and H. C. White, Structural Equivalence of Individuals in Social Networks, *Journal of Mathematical Sociology* 1: 49-80, 1971.

