A Graph-Theoretic Approach to Goodness-of-Fit in Complete-Link Hierarchical Clustering

FRANK B. BAKER and LAWRENCE J. HUBERT*

The complete-link hierarchical clustering strategy is reinterpreted as a heuristic procedure for coloring the nodes of a graph. Using this framework, the problem of assessing goodness-of-fit in complete-link clustering is approached through the number of "extraneous" edges in the fit of the constructed partitions to a sequence of graphs obtained from the basic proximity data. Several simple numerical examples that illustrate the suggested paradigm are given and some Monte Carlo results presented.

1. INTRODUCTION

Within the last twenty years an enormous number of strategies for the hierarchical clustering of a set of objects have been proposed [1, 8, 22]. Since the main emphasis in these efforts has been on the creation or application of clustering techniques, relatively little attention has been given to methods for evaluating the goodnesss-of-fit of an obtained partition hierarchy to the basic data. Within the context of hierarchical cluster analysis, the standard approach to goodness-of-fit is to calculate a coefficient of agreement between the inter-object proximity values, or their ranks, and an index of the level at which two objects are first merged into a common cluster. Most commonly, correlational statistics of one type or another are employed, such as the usual Pearson product-moment index used by Sneath and Sokal [22] or associational indices for ordinal data defined by Kendall's Tau or Goodman-Kruskal's Gamma [3, 6, 11].

None of these goodness-of-fit indices merely adopted from another field, however, are entirely satisfactory in practice and alternative approaches are needed that generate more detailed information and relate directly to the clustering criterion employed. Moreover, some notion of a sampling distribution is required for any index that is chosen. Unfortunately, since each clustering method will require a separate development of a sampling theory, the basic problem of measuring goodness-of-fit in clustering is difficult to solve with any generality.

The well-documented relationship that exists between certain types of cluster analyses and graph theory [10, 14] appears to hold considerable promise in characterizing exactly what the term "goodness-of-fit" should mean. While not attacking this problem directly, Ling [15] has made an important step by indicating how exact values may be obtained for the cumulative probability distribution for the minimum number of edges in a connected random graph. Tables of these distributions, prepared by Ling and Killough [16], provide one way of evaluating whether an observed result yielded by the single-link strategy could conceivably correspond to a random assignment of the proximity measures to the object pairs. An exactly similar point of view has been implemented by Lingoes and Cooper [17] in what they call Probability Evaluated Partition Analysis, but without the aid of exact probability values.

As yet, no graph-theoretical analysis similar to Ling's has been attempted for the complete-link procedure even though this strategy, compared to the single-link method, is much more widely used in the behavioral sciences and also has a clear graph-theoretic reinterpretation related to the problem of coloring the nodes of a graph. Consequently, one of the main purposes of this paper is to discuss the complete-link method from a graph-theoretic point of view as a way of clarifying what is involved in assessing goodness-of-fit for this particular method. The conclusions that may be drawn from the later Monte Carlo results are still very tentative and incomplete; nevertheless, the orientation developed here is capable of extensive exploitation and provides some perspective on the overall task of evaluating a complete-link result similar to that now partially available for the single link.

From a practical point of view the statistical problems encountered in assessing the adequacy of a constructed partition hierarchy are very important to recognize, both for the substantive theoretician and for the applied practitioner. For example, in developing a theory of memory organization and the structuring of what is called the subjective lexicon, one of the most common motivating hypotheses states that information is stored hierarchically. Consequently, clustering techniques have been used as the most natural way of imposing a hierarchical form on a set of "words." Usually, the basic data are proximity measures collected by a variety of experimental procedure, e.g., free-associate, free-recall, free-sort, and so on [2, 19]. In most instances, however, the obtained hierarchies have been evaluated only in a very loose heuristic sense. If the general form of the constructed hierarchies appears to be "close" to the assumed structure, then the initial hypotheses regarding

* Frank B. Baker and Lawrence J. Hubert are both Professor, Department of Educational Psychology, The University of Wisconsin, Madison WI 53706. Order of authorship is alphabetical. The authors wish to thank a referee for several helpful comments.

© Journal of the American Statistical Association
December 1976, Volume 71, Number 356
Applications Section

the organization of the subjective lexicon are considered supported. More significantly, very few serious attempts have been made to proceed further and evaluate statistically the degree to which the clustering procedure has imposed a form that is inappropriate for the given proximity data. Conceivably, information in the subjective lexicon may be organized in other ways, say, through clusters that overlap, and therefore, nonhierarchical methods (see [12]) would be more useful for eliciting the unknown structure.

Similar difficulties in assessing the "disjointness" assumptions of hierarchical clustering are important as well for the field of abnormal psychology, where it is of interest to subdivide a population of individuals into discrete groups, e.g., schizophrenic, manic-depressive, etc., as a first step in providing a differential therapy appropriate for the group in which an individual is placed (see [21]). However, if individuals could just as well belong to more than a single group, the labeling assumption on which much of abnormal psychology is based would lack convincing empirical support; more importantly, a clustering strategy that by necessity produces nonoverlapping subsets would fail to identify the appropriate classifications. In short, statistical procedures that test the degree of fit between the proximity data and the obtained hierarchy are important since they aid in the interpretation of the data as well as in assessing the legitimacy of a given clustering technique.

2. BACKGROUND

As an introduction to the basic problem attacked by hierarchical clustering, suppose S is a set of n objects, $o_1, \ldots, o_n$, and between each pair of objects o_i and o_j a symmetric proximity measure s_{ij} is defined.[1] For convenience, we use the interpretation that smaller proximity values correspond to the more similar objects pairs. The complete-link procedure produces a sequence or hierarchy of partitions of S, denoted by $\ell_0, \ldots, \ell_{n-1}$, from the ordinal information present in the matrix $\|s_{ij}\|$, i.e., from the rank order of all object pairs in terms of increasing dissimilarity. In particular, the partition ℓ_0 contains all objects in separate classes, ℓ_{n-1} consists of one all-inclusive object class, and ℓ_{k+1} is defined from ℓ_k by uniting a single pair of sets in ℓ_k.

As a way of characterizing what sets are chosen to unite in defining ℓ_{k+1} from ℓ_k, several elementary concepts from graph theory are helpful. Using G to represent an arbitrary graph defined by the node set S and certain unordered node pairs joined by undirected edges, the terms given here are standard in graph theory; for a more complete discussion, the reader should consult [7] or [20].[2]

[1] For our purposes, a proximity function is assumed to be a nonnegative real-valued function on $S \times S$ that indexes the degree of similarity between any two objects. Various substantive interpretations of the proximity concept are given in [1, 8, 12, 22].

[2] A pictoral representation of a graph with 8 nodes in S and 12 edges is given in Figure B.

1. A graph G is *complete* if and only if an edge exists between each pair of distinct nodes in G.
2. An *induced subgraph* of G defined by the subset D, $D \subseteq S$, is a graph consisting of the nodes in D and where $o_i, o_j \in D$ are linked, i.e., an edge exists between o_i and o_j, if and only if they are linked in G.
3. The *complementary graph* $\bar{G}$ is a graph with the same node set as G but in which two nodes are linked if and only if they are unlinked in G.
4. A graph G is said to be *m-colorable* if and only if the nodes of G can be assigned m different colors in such a way that no two distinct nodes with the same color are linked.
5. A graph G has *chromatic number* $\chi(G)$ if G is $\chi(G)$-colorable but not $(\chi(G) - 1)$-colorable.

Given these elementary concepts, the complete-link clustering method can be characterized as follows: suppose G_c is a graph consisting of the nodes $o_1, \ldots, o_n$, where o_i and o_j are linked by a single edge if and only if $s_{ij} \leq c$; $i \neq j$, $0 \leq c \leq \max\{s_{ij} | 1 \leq i, j \leq n\}$. If ℓ_k consists of the sets $L_1, \ldots, L_{n-k}$, then L_u and L_v contained in ℓ_k are united to form ℓ_{k+1} when a diameter measure $Q(\cdot, \cdot)$ is a minimum on the two sets L_u and L_v; the diameter measure is defined as follows:

$$Q(L_s, L_t) = \min\{s_{ij} | \text{the induced subgraph of } G_{s_{ij}} \text{ defined by the node set } L_s \cup L_t \text{ is complete}\} .$$

For technical convenience, it is assumed that the proximity measures are all distinct, and consequently, the minimums used here are attained uniquely. This assumption is taken up in more detail in [13].

Somewhat more intuitively, the complete-link (and the single-link) procedure unites those two classes in ℓ_k to form ℓ_{k+1} that are the "closest" together. For the complete-link method the concept of "closeness" is defined as the maximum proximity value attained for pairs of objects within the union of the two sets. The single-link procedure, on the other hand, employs a notion of "closeness" defined by the minimum proximity value attained for pairs of objects, where the two objects from the pair belong to the separate object classes. Consequently, from a graph-theoretic point of view the complete-link technique has to be reinterpreted in terms of complete subgraphs, whereas the single-link technique requires the much weaker concept of a connected subgraph, i.e., a subgraph in which each pair of nodes can be joined by a sequence of contiguous edges. Unfortunately, it is not possible to use Ling's [16] analysis of the single-link connectivity criterion in dealing with the concept of complete subgraphs for the alternative complete-link criterion. An extensive development of these graph-theoretic ideas within the clustering context is given in [10].

3. A REINTERPRETATION OF COMPLETE-LINK CLUSTERING

As a way of defining a more precise context in which to discuss the complete-link scheme and develop the necessary relationship with node colorability, the actual sequence of partitions formed by a complete-link clustering will be ignored for the present, and instead only a single graph G_c for some proximity value c will be considered.

It is assumed that P is the set of all partitions of S and we wish to choose certain elements p in P that satisfy reasonable homogeneity requirements, i.e., that are related in some way to the partitions formed by the complete-link procedure. In particular, suppose B is the set of partitions such that for all $p \in B$, o_i and o_j are linked if o_i and o_j belong to the same object class in p. Alternatively, no two dissimilar objects can belong to the same object class in p. Explicitly, the elements of B can be partially ordered[3] with respect to partition refinement, where one specific partition p is a refinement of a second partition p' if and only if p' can be formed by uniting certain sets contained within p. No universal least upper bound is available in B, and consequently, if it is necessary to select members of B for a more thorough substantive analysis, two specific classes are intuitively natural to consider:

1. The maximal elements of B, i.e.,

   $B_1 = \{p \in B \mid$ there is no element $p' \in B$ such that p is a proper refinement of $p'\}$.

2. The maximal elements of B with the smallest number of object classes, i.e.,

   $$B_2 = \{p \in B_1 \mid p \text{ has } f_1 \text{ object classes}\} ,$$

   where

   $$f_1 = \min \{f \mid p \in B_1 \text{ and } p \text{ has } f \text{ object classes}\} .$$

The set B_2 appears to be the most useful alternative to consider given the common substantive concern of interpreting a minimum number of object classes for meaning.

As a simple illustration that will be developed throughout the discussion, suppose S is $\{o_1, o_2, o_3, o_4\}$ and has an associated proximity matrix $\|s_{ij}\|$ of the form

	o_1	o_2	o_3	o_4
o_1	0	1	4	5
o_2	1	0	2	6
o_3	4	2	0	3
o_4	5	6	3	0

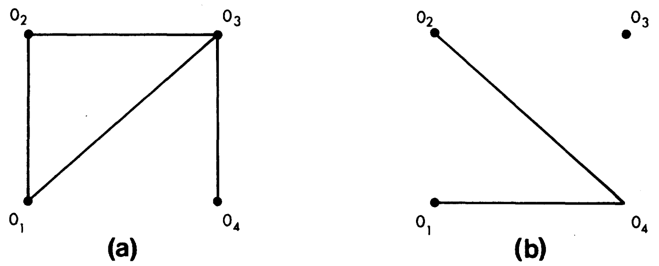
If we let $c = 4$, then the graph G_4 (Figure A) can be obtained by drawing undirected edges between those object pairs whose proximity values are less than or equal to four. The graph $\bar{G}_4$ (Figure A) is obtained by connecting only those object pairs whose proximity values are greater than four.

Since S contains four objects, P consists of fifteen partitions. Moreover, the set B, defined by all partitions of S that decompose G_4 into node sets defining complete subgraphs, contains the seven elements

$$\{\{o_1, o_2, o_3\}, \{o_4\}\}, \{\{o_1, o_2\}, \{o_3, o_4\}\} ,$$
$$\{\{o_1, o_2\}, \{o_3\}, \{o_4\}\}, \{\{o_1, o_3\}, \{o_2\}, \{o_4\}\} ,$$
$$\{\{o_2, o_3\}, \{o_1\}, \{o_4\}\}, \{\{o_3, o_4\}, \{o_1\}, \{o_2\}\} ,$$
$$\{\{o_1\}, \{o_2\}, \{o_3\}, \{o_4\}\}.$$

In this simple case, B_1 and B_2 happen to be the same; and each consists of the two partitions $\{\{o_1, o_2\}, \{o_3, o_4\}\}$ and $\{\{o_1, o_2, o_3\}, \{o_4\}\}$. The definition for membership in B_1 is satisfied by both partitions since uniting any pair of sets in either partition generates a new decomposition of the set S that fails to be a member of B. The definition of B_2 is also satisfied by both partitions since $f_1 = 2$ is the minimum number of object classes obtainable for any partition within B.

A. Graphs G_4 and $\bar{G}_4$ for a Four Member Object Set



The relationship between B_1 and the complete-link hierarchy can now be identified since some element of B_1 has to appear within the sequence of partitions generated by the complete-link method; for instance, in our simple example, the partition $\{\{o_1, o_2\}, \{o_3, o_4\}\}$ would be obtained at level 2 in the hierarchy. Even though it is not generally true that an element of B_2 has to appear in the complete-link sequence, the set B_2 has some very interesting connections to the problem of coloring the nodes of a graph that help to clarify the nature of complete-link clustering. In particular, f_1 is the chromatic number of the graph $\bar{G}_c$, i.e., $\chi(\bar{G}_c) = f_1$, and furthermore, finding all the elements of B_2 is equivalent to finding all the different ways of coloring the nodes of $\bar{G}_c$ with f_1 colors.

To be more specific, suppose we are given a graph G_c for some proximity value c with, say, t edges. Next, the first t distinct ranks are arbitrarily assigned to the object pairs that are linked in G_c and ranks larger than t assigned to the remainder. If a complete-link clustering is performed until the measure Q exceeds t, the last element formed in the hierarchy must be a member of B_1. In fact, if all $t!$ possible assignments of rank are made, all members of B_1 will be constructed. Finally, the set B_2 can be obtained from B_1 by choosing those partitions with f_1 object classes. These relationships can be seen rather easily using $\bar{G}_4$ given in Figure A. The colorings of $\bar{G}_4$ with the minimum of two colors are defined by the nodes o_1, o_2, and o_3 being assigned one color and o_4 a second, or alternatively, o_1 and o_2 being assigned one

[3] If C is any collection and $\leq$ a binary relation between certain pairs of elements of C, then C is said to be partially ordered with respect to $\leq$ if and only if the following axioms hold: 1. for all x in C, $x \leq x$; 2. if $x \leq y$, $y \leq x$, then $x = y$; 3. if $x \leq y$, $y \leq z$, then $x \leq z$.

color and o_3 and o_4 a second. The number of distinct ranks t is 4; consequently, since B_1 and B_2 happen to be the same in this example, each of the $t! = 24$ possible assignments of ranks to the edges that exist in G_4 along with the complete-link routine would lead to one of these two colorings of $\bar{G}_4$.

As a way of highlighting the relationship between the complete-link procedure and node coloring, the complete-link method can be viewed as a generalization of a common heuristic used in graph theory and referred to as sequential node coloring. Following Matula, Marble, and Isaacson [18], any ordered sequence of the n nodes, say $o_{h_1}, \ldots, o_{h_n}$ can be used to produce a coloring of $\bar{G}_c$ through the following procedure:

1. o_{h_1} is assigned to color class 1;
2. if the nodes $o_{h_1}, \ldots, o_{h_{i-1}}$ have been assigned to j color classes, then, if possible, o_{h_i} is assigned to the color class m, where m is the minimum positive integer less than or equal to j such that o_{h_i} is not linked to any of the objects in the mth class. If no such integer exists, then o_{h_i} is used to define the new $(j+1)$st color class.

The complete-link procedure can be used to effect the same partitioning of G_c. Specifically, the existing edges of G_c, defined by the node pairs $\{o_{h_i}, o_{h_j}\}$, $h_i < h_j$, are ordered lexicographically according to the index sequence $h_1, \ldots, h_n$ and the "nonexisting" edges are all given an arbitrarily large rank, say, $n(n-1)/2$. If the complete-link procedure is used, as long as the diameter measure Q is less than the arbitrary rank $n(n-1)/2$, the last partition so constructed is the appropriate decomposition of G_c, i.e., the sequential coloring of $\bar{G}_c$ defined by the node sequence $o_{h_1}, \ldots, o_{h_n}$. As an illustration using Figure A, the node sequence o_1, o_2, o_3, o_4 produces the partition $\{\{o_1, o_2, o_3\}, \{o_4\}\}$ when the heuristic scheme discussed in [18] is applied. Equivalently, suppose the node pairs are ordered in the lexicographic manner suggested previously: $(\{o_1, o_2\}, \{o_1, o_3\}, \{o_2, o_3\}, \{o_3, o_4\}, \{o_1, o_4\}, \{o_2, o_4\})$, and given ranks of 1, 2, 3, 4, 6, and 6, respectively, where the first four node pairs define existing edges in G_4 and the last two pairs define nonexisting edges. The complete-link procedure continued as long as Q is still less than 6, produces exactly the same partition.

Obviously, a method that uses a random assignment of proximity ranks for constructing B_1 and B_2 is computationally inefficient since the same partition will be obtained again and again for different assignments for the first t ranks. What is of interest, however, is that at least theoretically the complete-link method does lead to a well-defined procedure for obtaining B_2. For convenience, the use of the complete-link method and any assignment of the first t ranks will be referred to as a "naive" coloring of $\bar{G}_c$, where the term "naive" means there is usually no guarantee that an element of B_2 will be obtained.

In summary, by monotonically varying the criterion or threshold c, the complete-link procedure can be considered a technique for generating a sequence of naive colorings for the set of graphs $\{\bar{G}_c\}$, i.e., for each value of c, a partition is constructed that satisfies the definition for membership in the set of partitions labeled B_1. Since the complete-link method has this particular interpretation, alternative coloring schemes can be viewed as competitors to the complete-link routine. For purposes of historical perspective, see [18].

The single-link strategy discussed by Ling [15] has a related graph-theoretic characterization. However, the difficulties encountered in the complete-link scheme for possibly having multiple elements in the set B_1 or B_2 have no single-link analogues. Specifically, in the single-link method a homogeneous set of partitions B' can be considered in which p belongs to B' if and only if the following condition holds: if o_i and o_j are linked, then o_i and o_j belong to the same object class in p. Clearly, B and B' are defined by converse conditions, and also, the elements of B' can be partially ordered with respect to partition refinement. In this case, however, the greatest lower bound is the set of connected components of G_c and forms a natural single representative member of B' that may be interpreted substantively, and moreover, must be within the partition hierarchy generated by the single-link routine.

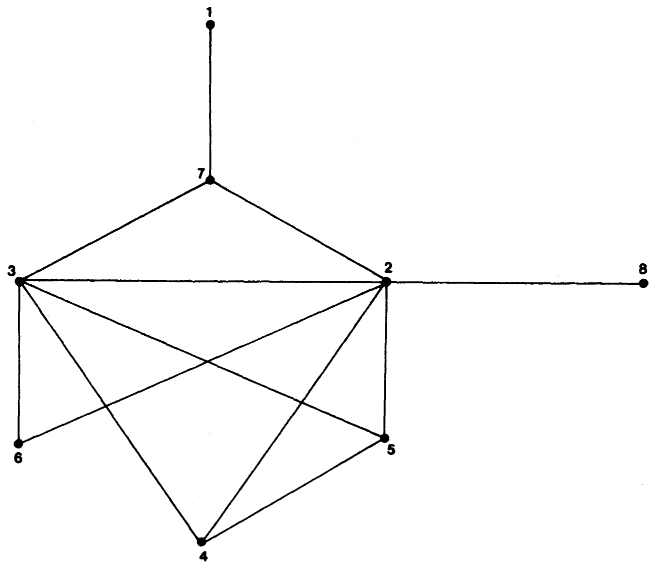
As a more realistic numerical example of the preceding discussion, Table 1 presents data given by Holzinger and Harman [9] on the first eight psychological tests of a battery of 24 used by Holzinger and Harman in their well-known presentation of various factor analytic procedures. The proximity values listed in Table 1 are one minus the product-moment correlations given by Holzinger and Harman [9, p. 30]. Figure B is a diagram of $\bar{G}_c$, where c was arbitrarily chosen as .682. This particular value of c, however, does induce a connected graph G_c.

1. Proximity Values for Holzinger and Harman's Eight Psychological Tests

Object pair	*Proximity value*	*Object pair*	*Proximity value*
{6,7}	.278	{1,5}	.679
{5,7}	.344	{1,2}	.682
{5,6}	.378	{2,3}	.683
{7,8}	.381	{3,4}	.695
{5,8}	.422	{1,7}	.696
{6,8}	.473	{2,5}	.715
{1,4}	.532	{3,6}	.732
{1,3}	.597	{3,5}	.753
{4,8}	.609	{2,6}	.766
{3,8}	.618	{2,4}	.770
{4,7}	.665	{4,5}	.773
{1,6}	.665	{3,7}	.777
{1,8}	.668	{2,7}	.843
{4,6}	.673	{2,8}	.843

It can be shown using bounds on the chromatic number of $\bar{G}_c$ (see [7]) that $\chi(\bar{G}_{.682}) = 4$, which is the minimum number of subsets that can be found at $c = .682$ that decompose $G_{.682}$ into complete subgraphs. In an attempt to obtain a specific coloring of $\bar{G}_c$ with four colors, the complete-link hierarchy was constructed using the proximity values given in Table 1 to the point where the criterion Q was equal to .682. Fortunately, an element of

B. For the Holzinger and Harman Data; c = .682



B_2 was yielded by the complete-link strategy in this case since the last partition in the sequence before the clustering was curtailed contained four object classes, i.e., {{1, 4}, {2}, {3}, {5, 6, 7, 8}}. It is interesting to note that this same partition is also obtained much "earlier" at a proximity value of .532. The total complete-link hierarchy is given in Table 2 and is used in the later discussion.

2. Complete-Link Partition Hierarchy for Holzinger and Harman's Eight Psychological Tests

Level	c_k	*Partition*	A_k	T_k
1	.278	{{6,7},{1},{2},{3},{4},{5},{8}}	1	1
2	.378	{{5,6,7},{1},{2},{3},{4},{8}}	3	3
3	.473	{{5,6,7,8},{1},{2},{3},{4}}	6	6
4	.532	{{5,6,7,8},{1,4},{2},{3}}	7	7
5	.683	{{5,6,7,8},{1,4},{2,3}}	17	8
6	.843	{{5,6,7,8},{1,2,3,4}}	24	12

4. GOODNESS-OF-FIT IN COMPLETE-LINK CLUSTERING

Given a graph G_c and the complete-link partition obtained up to that point, this particular decomposition "fits" the graph perfectly if there are no edges in G_c that link two objects from different sets, i.e., the condition used to define the set of partitions B' is also satisfied. Alternatively, if the partition contains d object classes, then using the terminology of graph theory, $\bar{G}_c$ is a d-partite graph containing the maximum number of possible edges. In this case, $\bar{G}_c$ is uniquely $\chi(\bar{G}_c)$-colorable and the complete-link partition is a member of the set B_2 for the graph G_c. More generally, suppose c_{k+1} is the proximity that caused the partition ℓ_{k+1} to form, i.e., c_{k+1} corresponds to a value of Q defined in (2.1). If ℓ_k contains the sets $L_1, \ldots, L_{n-k}$ with $r_1, \ldots, r_{n-k}$ members, respectively, then the minimum number of edges in G_{c_k} is

$$T_k = \sum_{i=1}^{n-k} r_i(r_i - 1)/2 \ . \tag{4.1}$$

Consequently, using the Holzinger and Harman [9] numerical example of the previous section for $k = 4$, we find that $c_4 = .532$, the graph G_{c_4} is decomposed perfectly by the partition {{5, 6, 7, 8}, {1, 4}, {2}, {3}}, and finally, the minimum number of edges T_4 is 7. To achieve this minimum under the complete-link strategy, the rank ordered proximity values must be scanned up to the same rank order as there are edges in G_{c_k}, generating a perfect fit between the partition at level k and the data. For most real data sets, however, such an ideal situation rarely exists. For example, in the Holzinger and Harman data, the partition at level 5 requires one edge with a proximity rank of 17, and consequently, there are 17 − 8 or 9 "extraneous" edges joining objects from different subsets of the partition at the minimum proximity values required to produce the partition, i.e., at $c_5 = .683$. Specifically, if A_k denotes the total number of edges in the graph G_c, then the difference $E_k = A_k - T_k$ is the number of extraneous edges induced by the partition ℓ_k and indexes the "lack-of-fit" of ℓ_k to the proximity data. Obviously, small values for E_k are desirable since the larger E_k is, the less adequately the partition ℓ_k represents the relationships among the objects in S. This notion of expecting small values of E_k when the partition ℓ_k represents the proximities well, could be developed more precisely using a concept of power under the general approach given by Baker and Hubert [4].

Even though the values $T_1, \ldots, T_{n-1}$ and $A_1, \ldots, A_{n-1}$ can be calculated very easily for any complete-link partition hierarchy, as yet there is no baseline or sampling distribution available for evaluating the E_k's. Ideally, reference distributions for the E_k's could be derived analytically, but in general, the task this poses appears to be extremely difficult for all but the most trivial cases. As a second best alternative for levels above 2 in a partition hierarchy, a reasonable hypothesis on the distribution of the proximity values will be made and approximate empirical distributions constructed for the A_k's for several representative values of n. For an introduction, however, an exact analysis is carried out for the simple example of four objects, or almost equivalently, Level 2 of a complete-link partition hierarchy for any value of n.

Following [14, 11], it is assumed that all $n(n-1)/2$ proximity values are assigned at random to the object pairs, or equivalently, $n(n-1)/2$ ranks are assigned at random since the complete-link strategy merely requires the ordinal information present in the proximity values. For $n = 4$ only the Level 2 partition is nontrivial, and under the assumption that all six proximity ranks are assigned randomly we can proceed as follows: when n is 4, there are six possible pairs and since one pair is already chosen to produce the partition at Level 1, five pairs are left that have to be considered in moving to Level 2. Among these five pairs, four have one object in common with the pair that induced the Level 1 partition and one pair has no such object in common. Thus, the joint distribution of T_2 and A_2 may be found by noting which

3. Observed Distribution of Values of A_k for Values of T_k at Each Partition Level, $n = 8$, 12, and 16

n	Partition level	T_k	Number of occurrences	A_k Percentage points: Observed minimum A_k	Observed maximum A_k	.05	.25	.50
8	2	2	475	2	7	2	2	2
		3	25	3	5	3	3	4
	3	3	366	3	13	3	4	5
		4	131	4	11	4	5	6
		6	3	9	11	—	—	—
	4	4	110	4	18	5	7	10
		5	308	5	17	6	9	11
		6	43	7	18	7	9	12
		7	38	9	18	10	12	13
		10	1	15	15	—	—	—
	5	7	210	8	23	12	14	17
		8	175	10	23	14	16	18
		9	67	11	22	13	17	18
		11	46	14	23	15	18	20
		15	2	20	23	—	—	—
	6	12	126	18	26	19	23	24
		13	193	16	26	21	23	24
		16	165	20	26	23	24	25
		21	16	23	26	23	24	25
12	2	2	497	2	6	2	2	2
		3	3	3	5	—	—	—
	3	3	481	3	12	3	3	4
		4	19	4	8	4	5	5
	4	4	407	4	17	4	6	7
		5	91	5	17	6	7	9
		6	1	9	9	—	—	—
		7	1	11	11	—	—	—
	5	5	245	5	23	6	9	11
		6	229	6	25	8	11	13
		7	21	9	24	9	12	13
		8	5	13	28	—	—	—
	6	6	43	8	39	9	15	20
		7	274	10	41	12	16	20
		8	123	11	36	14	19	21
		9	48	13	37	14	21	23
		10	11	18	38	18	22	24
	7	9	196	12	47	20	26	30
		10	129	15	46	21	27	33
		11	139	20	48	24	29	33
		12	6	22	41	—	—	—
		13	22	26	44	26	33	35
		14	8	27	38	—	—	—
	8	12	17	26	48	26	32	37
		13	195	26	56	33	39	43
		14	78	33	54	34	40	44
		15	143	28	56	35	41	45
		16	17	34	54	34	40	45
		17	20	35	52	25	41	44
		18	18	38	52	38	44	47
		19	11	39	52	39	42	47
		20	1	48	48	—	—	—
	9	18	36	45	61	46	51	55
		19	170	41	60	44	51	54
		21	125	38	61	46	52	55
		22	94	42	61	47	53	55
		25	53	45	60	50	53	56
		27	4	54	59	—	—	—
		30	13	55	61	55	57	58
		31	4	55	61	—	—	—
		37	1	58	58	—	—	—

Table 3. (Continued)

n	Partition level	T_k	Number of occurrences	A_k Percentage points: Observed minimum A_k	Observed maximum A_k	.05	.25	.50
	10	30	86	55	64	57	60	62
		31	158	55	64	58	60	62
		34	136	54	64	58	61	62
		39	82	59	64	60	62	63
		46	36	59	64	60	63	64
		55	2	63	64	—	—	—
16	2	2	496	2	8	2	2	2
		3	4	3	6	—	—	—
	3	3	491	3	10	3	3	4
		4	9	4	7	—	—	—
	4	4	473	4	14	4	5	6
		5	27	5	14	5	6	7
	5	5	422	5	23	5	7	9
		6	77	6	21	7	9	11
		8	1	19	19	—	—	—
	6	6	342	6	36	8	12	14
		7	146	8	30	10	13	16
		8	10	12	27	—	—	—
		9	2	17	20	—	—	—
	7	7	171	9	43	12	17	22
		8	257	11	48	14	18	23
		9	60	14	41	15	21	26
		10	11	20	37	20	24	29
		11	1	32	32	—	—	—
	8	8	23	20	54	20	30	40
		9	228	15	66	19	27	33
		10	172	19	55	23	28	33
		11	52	25	66	27	33	38
		12	23	26	53	26	34	40
		13	2	33	41	—	—	—
	9	11	165	18	74	28	40	47
		12	146	26	70	31	40	46
		13	128	32	77	37	45	50
		14	38	30	66	34	43	51
		15	15	40	68	40	49	53
		16	8	30	69	—	—	—
	10	14	55	20	81	43	49	58
		15	224	38	87	47	57	65
		16	79	37	83	45	58	64
		17	100	47	84	50	60	66
		18	20	52	81	52	61	68
		19	10	59	86	—	—	—
		20	5	47	75	—	—	—
		21	4	62	81	—	—	—
		22	2	65	79	—	—	—
		25	1	73	73	—	—	—
	11	18	26	59	94	59	67	75
		19	132	45	101	59	70	77
		20	81	65	102	68	74	80
		21	161	59	102	68	76	81
		22	8	62	86	—	—	—
		23	49	70	100	73	82	88
		24	21	69	93	69	77	84
		25	11	67	98	67	70	83
		26	1	75	75	—	—	—
		27	8	75	101	—	—	—
		29	1	85	85	—	—	—
		33	1	87	87	—	—	—
	12	24	13	82	106	82	84	91
		25	94	67	108	76	88	93
		26	23	86	105	86	93	97
		27	120	77	109	83	90	94
		28	55	79	108	82	91	97

Table 3. (Continued)

n	Partition level	T_k	Number of occurrences	A_k Percentage points: Observed minimum A_k	Observed maximum A_k	.05	.25	.50
		29	67	83	107	86	93	98
		30	12	95	106	95	99	102
		31	66	83	110	84	94	99
		32	7	85	108	—	—	—
		33	14	91	107	91	98	100
		34	2	93	103	—	—	—
		35	13	91	106	91	92	98
		36	8	93	104	—	—	—
		37	1	102	102	—	—	—
		41	1	95	95	—	—	—
		42	1	97	97	—	—	—
		43	2	104	104	—	—	—
		48	1	105	105	—	—	—
	13	35	54	89	114	93	102	106
		36	54	91	114	97	103	106
		37	95	95	114	98	104	108
		39	65	97	114	98	105	107
		40	38	94	115	100	107	110
		41	52	100	115	101	106	109
		43	15	97	114	97	104	106
		44	24	100	114	100	104	108
		45	51	98	115	100	107	110
		47	21	100	114	100	105	109
		51	8	107	111	—	—	—
		52	15	108	115	108	110	111
		55	1	113	113	—	—	—
		59	7	108	115	—	—	—
	14	56	60	109	118	109	113	115
		57	121	110	118	111	114	116
		60	113	106	118	111	114	116
		65	78	109	118	112	114	116
		72	71	110	118	113	116	117
		81	46	114	118	115	116	117
		92	11	114	118	114	116	117

of the pairs can be selected sequentially until the Level 2 partition is finally constructed. In particular, we obtain

$$P(T_2 = 2; A_2 = 2) = 1/5$$
$$P(T_2 = 2; A_2 = 3) = 4/5 \cdot 1/4 = 1/5$$
$$P(T_2 = 2; A_2 = 4) = 4/5 \cdot 2/4 \cdot 1/3 = 2/15$$
$$P(T_2 = 3; A_2 = 3) = 4/5 \cdot 1/4 = 1/5$$
$$P(T_2 = 3; A_2 = 4) = 4/5 \cdot 2/4 \cdot 2/3 = 4/15 \; .$$

Obviously, the same procedure can be carried out at partition Level 2 for any value of n, and the following closed form expressions obtained: for $n \geq 4$,

$$P(T_2 = 2; A_2 = i + 2) = (2^{i-1}(n-2)(n-3) \cdot [(n-2)!])/(N \ldots (N-i)[(n-2-i)!]) \; , \quad (4.2)$$

where $N = \binom{n}{2} - 1$, $0 \leq i \leq n - 2$;

$$P(T_2 = 3; A_2 = i' + 3) = ((i' + 1)2^{i'+1}[(n-2)!])/(N \cdots (N - (i' + 1)) \cdot [(n - 2 - (i' + 1))!]) \; , \quad (4.3)$$

where $0 \leq i' \leq n - 3$.

Although it is theoretically possible to continue an exact analysis for all levels of a partition hierarchy, the bookkeeping burden is immense even for the possible partitions at Level 3. However, to gain some approximate notion of what these distributions will be, a random sample of assignments of the $n(n-1)/2$ proximity ranks can be made, and the complete-link hierarchy obtained along with the calculated values of A_k and T_k at each partition level k. Because of the size of the possible tables, Table 3 presents summary empirical information of this type only for n's of 8, 12, and 16 using 500 random allocations for each n. In particular, for each value of T_k within a partition level, Table 3 lists the number of occurrences, lower percentage points of .05, .25, and .50 for samples greater than ten, and the minimum and maximum values of A_k obtained within each of the subsamples.[4]

For Level 2 partitions, the complete conditional probability distribution can be found exactly using the closed form expressions given in (4.2) and (4.3). As an illustration, when $n = 8$ we have

$$P(A_2 = 2 \mid T_2 = 2) = .579 \qquad P(A_2 = 3 \mid T_2 = 3) = .422$$
$$P(A_2 = 3 \mid T_2 = 2) = .267 \qquad P(A_2 = 4 \mid T_2 = 3) = .338$$
$$P(A_2 = 4 \mid T_2 = 2) = .107 \qquad P(A_2 = 5 \mid T_2 = 3) = .167$$
$$P(A_2 = 5 \mid T_2 = 2) = .035 \qquad P(A_2 = 6 \mid T_2 = 3) = .059$$
$$P(A_2 = 6 \mid T_2 = 2) = .009 \qquad P(A_2 = 7 \mid T_2 = 3) = .012$$
$$P(A_2 = 7 \mid T_2 = 2) = .002 \qquad P(A_2 = 8 \mid T_2 = 3) = .002.$$
$$P(A_2 = 8 \mid T_2 = 2) = .000$$

At least in this trivial case, the sample percentage points (at .05, .25, and .50) given in Table 3 for $n = 8$ and partition Level 2 are also the "true" values; this observation may be verified immediately from the two conditional distributions for $T_2 = 2$ and $T_2 = 3$ just given.

It is clear that the values of A_k given in Table 3 are generally open to large sampling error, especially those percentage points based on a small number of occurrences in the sample. Consequently, sample information of this type approximates only in a very crude way the actual values that could be derived theoretically, and any evaluation of a complete-link hierarchical clustering based on the data in Table 3 should be considered a very loose heuristic. In fact, because of the limited number of random allocations used (500), certain possible values of T_k are missing from the table entirely since none of the random assignments produced the appropriate partitions. Nevertheless, some idea of the accuracy of the Table 3 entries may be obtained by comparing the exact analysis for Level 2 given previously and the simulation results reported in more summarized form in Table 3. For instance, when $n = 8$, the obtained and expected values (rounded to the nearest integer) for Level 2 using the sample size of 500 are shown in Table 4.

[4] Additional tables that include n's from 6 to 16 (500 assignments), $n = 20$ (200 assignments), $n = 25$ (150 assignments), and $n = 30$ (100 assignments) may be obtained from the authors.

4. Comparison of Observed and Expected Frequencies for n = 8; Level 2; Sample Size of 500

A_2	$T_2 = 2$		$T_2 = 3$	
	Observed	Expected	Observed	Expected
2	268	278	—	—
3	135	128	7	9
4	55	51	11	7
5	15	17	7	3
6	1	4	0	1
7	1	1	0	0
8	0	0	0	0

Several other comparisons of a similar nature may also be given. For instance, it is easily verified that when $1 \leq k \leq n/2$.

$$P(T_k = k; A_k = k) = \prod_{i=0}^{k-1} (n - 2i)(n - 2i - 1)/[n(n - 1) - 2i] \; .$$

Thus, if n is, say, 16, the comparison between observed and expected frequencies given in Table 5 is immediate.

5. Comparison of Observed and Expected Frequencies for ($T_k = k$; $A_k = k$); n = 16; Sample Size of 500

k	Observed	Expected
2	380	382
3	233	214
4	86	82
5	21	20
6	3	3
7	0	0
8	0	0

In general, comparisons of this type give fairly close correspondences between the empirical and theoretical proportions, i.e., the empirical proportions appear well within the bounds that would be expected due to sampling, e.g., by chi-square tests for Table 4. As one final caution in the use of the Monte Carlo results, however, it should be pointed out that the values in Table 3 are subject to severe dependencies due to the sampling scheme employed. In other words, even minor distortions that appear early in the hierarchy will "chain" to the higher levels and produce correlated discrepancies throughout.

With the limitations of the simulation in mind, inspection of Table 3 still reveals several interesting empirical facts that deserve emphasis. At low numbered partition levels and low values of T_k, the minimum observed value of A_k is at or near the value of T_k, suggesting that the beginning partitions of a hierarchy produced by the complete-link strategy must "fit" the data. Furthermore, as both the index of the partition level and the value of T_k increase, the difference between T_k and the 50 percent point of the approximate distribution of A_k becomes quite large. This implies that at high partition levels a complete-link partition defined by the minimum number of edges is quite unlikely.

From a practical point of view, to use Table 3 to evaluate heuristically the goodness-of-fit of any partition within the complete-link hierarchy, the following steps would be taken:

1. the number of objects in each cluster would be used in conjunction with (4.1) to calculate the value of T_k;
2. the rank order of the proximity value A_k at which the partition was formed would be found;
3. using the values of A_k and T_k from Steps (1) and (2), the value of A_k would be compared to the approximate percentage points in Table 3. If the value of A_k was at a suitably small percentage point, the researcher has some assurance that the results are not comparable to that generated by a random process, and consequently, may reflect real characteristics in his data.

To illustrate the use of this heuristic strategy with the Holzinger and Harman data of Table 1, the complete-link partition hierarchy given in Table 2 includes the necessary values of A_k and T_k. Up to partition Level 4, the value of A_k equals the minimum value T_k, and consequently, no lack of fit is indicated. At Level 5, however, A_k is 17 and T_k is 8, and thus, A_k exceeds the approximate 25th percentage point of the distribution for A_k; in other words, when a random assignment of proximity values yields a partition with a minimum number of edges equal to 8, the complete-link partition generated a value of A_5 smaller than that obtained with the Holzinger and Harman results more than 25 percent of the time.

A similar situation occurs at Level 6 where the calculated value of A_6 is close to the median value constructed by random assignment. This last result is especially interesting since this particular decomposition corresponds to the standard factor analytic interpretation given by Holzinger and Harman in terms of a "spatial relations" group of tests and a "verbal" group of tests. In other words, the obtained two-group partition does not appear to be inconsistent with a hypothesis of randomness and apparently there are too many edges existing between the pair of four-object subsets for the partition to represent clearly an underlying relationship among the objects.

5. SUMMARY

The relation developed here between node colorability and complete-link clustering makes a graph-theoretic approach to goodness-of-fit possible. Although incompletely developed, a graph-theoretic paradigm based on the "extraneous edges" concept provides an alternative, at least in complete-link hierarchical clustering, to the usual correlation techniques for evaluating goodness-of-fit. The tables provided, although open to sampling error, may be used to evaluate whether the clusters observed at a particular level could conceivably have been obtained by a random process.

Information of this type is valuable to the user of the complete-link procedure in determining whether his results should be the basis for further substantive interpretation, or possibly, that some other clustering technique may be more appropriate. Hopefully, the orientation presented here will stimulate the search for complete analytic solutions or approximations that will obviate the need to generate a more extensive set of results by Monte Carlo approximation. Clearly, what is needed are extensive theoretical analyses comparable to what Ling [15] has done for the single-link strategy, and moreover, a general application of such work to the evaluation of a partition hierarchy obtained from the complete-link criterion.

[*Received April 1975. Revised November 1975.*]

REFERENCES

[1] Anderberg, Michael R., *Cluster Analysis for Applications*, New York: Academic Press, Inc., 1973.
[2] Anglin, Jeremy M., *The Growth of Word Meaning*, Cambridge, Mass.: The M.I.T. Press, 1970.
[3] Baker, Frank B., "Stability of Two Hierarchical Grouping Techniques, Case I: Sensitivity to Data Errors," *Journal of the American Statistical Association*, 69 (June 1974), 440–45.
[4] ——— and Hubert, Lawrence J., "Measuring the Power of Hierarchical Cluster Analysis," *Journal of the American Statistical Association*, 70 (March 1975), 31–8.
[5] Berge, Claude, *The Theory of Graphs and Its Applications*, New York: John Wiley & Sons, Inc., 1962.
[6] Cunningham, Kathrine M. and Ogilvie, John C., "Evaluation of Hierarchical Grouping Techniques: A Preliminary Study," *The Computer Journal*, 15 (August 1972), 209–13.
[7] Harary, Frank, *Graph Theory*, Reading, Mass.: Addison-Wesley Publishing Co., 1969.
[8] Hartigan, John A., *Clustering Algorithms*, New York: John Wiley & Sons, Inc., 1975.
[9] Holzinger, Karl J. and Harman, Harry H., *Factor Analysis*, Chicago: The University of Chicago Press, 1941.
[10] Hubert, Lawrence J., "Some Applications of Graph Theory to Clustering," *Psychometrika*, 39 (September 1974), 283–309.
[11] ———, "Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarchical Clustering Procedures," *Journal of the American Statistical Association*, 69 (September 1974), 698–704.
[12] Jardine, Nicholas and Sibson, Robin, *Mathematical Taxonomy*, New York: John Wiley & Sons, Inc., 1970.
[13] Johnson, Stephen C., "Hierarchical Clustering Schemes," *Psychometrika*, 32 (September 1967), 241–54.
[14] Ling, Robert F., "A Probability Theory of Cluster Analysis," *Journal of the American Statistical Association*, 68 (March 1973), 159–64.
[15] ———, "An Exact Probability Distribution of the Connectivity of Random Graphs," *Journal of Mathematical Psychology*, 12 (February 1975), 90–8.
[16] ——— and Killough, George G., "Probability Tables for Cluster Analysis Based on a Theory of Random Graphs," *Journal of the American Statistical Association*, 71 (June 1976), 293–300.
[17] Lingoes, James C. and Cooper, Terry, "Probability Evaluated Partitions-I," *Behavioral Science*, 16 (May 1971), 259–61.
[18] Matula, David W., Marble, George and Isaacson, Joel D., "Graph Coloring Algorithms," in Ronald C. Reid, ed., *Graph Theory and Computing*, New York: Academic Press, Inc., 1972, 109–22.
[19] Miller, George A., "A Psychological Method to Investigate Verbal Concepts," *Journal of Mathematical Psychology*, 6 (June 1969), 169–91.
[20] Ore, Oystein, *Theory of Graphs*, Providence, R.I.: American Mathematical Society, 1962.
[21] Overall, John E. and Klett, C. James., *Applied Multivariate Analysis*, New York: McGraw-Hill Book Co., 1972.
[22] Sneath, Peter H.A. and Sokal, Robert R., *Numerical Taxonomy*, San Francisco: W.H. Freeman & Co., 1973.