PSYCHOMETRIKA—VOL. 32, NO. 3
SEPTEMBER, 1967

HIERARCHICAL CLUSTERING SCHEMES*

STEPHEN C. JOHNSON

BELL TELEPHONE LABORATORIES,
MURRAY HILL, NEW JERSEY

Techniques for partitioning objects into optimally homogeneous groups on the basis of empirical measures of similarity among those objects have received increasing attention in several different fields. This paper develops a useful correspondence between any hierarchical system of such clusters, and a particular type of distance measure. The correspondence gives rise to two methods of clustering that are computationally rapid and invariant under monotonic transformations of the data. In an explicitly defined sense, one method forms clusters that are optimally "connected," while the other forms clusters that are optimally "compact."

Introduction

In many empirical fields there is an increasing interest in identifying those groupings or clusterings of the "objects" under study that best represent certain empirically measured relations of similarity. For example, often large arrays of data are collected, but strong theoretical structures (which might otherwise guide the analysis) are lacking; the problem is then one of discovering whether there is any structure (i.e., natural arrangement of the objects into homogeneous groups) inherent in the data themselves. Recent work along these lines in the biological sciences has gone under the name "numerical taxonomy" [Sokal, 1963].

Although the techniques to be described here may find useful application in biology, medicine and other fields as well, we shall use psychology as an illustrative field of application. In that field, the "objects" under study might, for example, be individual human or animal subjects, or various visual or acoustic stimuli presented to such subjects. We might want to use measures that we have obtained on the similarities (or psychological "proximities") among the "objects" to classify the objects into optimally homogeneous groups; that is, similar objects are assigned to different groups.

Suitable data on the similarities among the objects (from which such a natural grouping might be derived) may be obtained directly or indirectly. For example, sometimes one obtains for every pair of objects a subjective

*I am indebted to R. N. Shepard and J. D. Carroll for many stimulating discussions about this work, and for aid in preparing this paper.

rating of similarity, or, (what is often very closely related) a measure of the confusion or "interchangeability" of the objects. Less directly, we may measure a number of attributes of the objects (often termed a profile of measures) and combine them to form a single measure of similarity. Various kinds of measures of profile similarity can be used for this purpose (e.g., product-moment-correlation, covariance, or the sum of squared or absolute differences between corresponding components of the profiles).

The problem of course, is that if the number of objects is large, the resulting array of similarity measures (containing, as it does, one value for each *pair* of objects) can be so enormous that the underlying pattern or structure is not evident from inspection alone. This paper discusses procedures which, when applied to such an array of similarity measures, constructs a hierarchical system of clustering representations, ranging from one in which each of the n objects is represented as a separate cluster to one in which all n objects are grouped together as a single cluster.

An algorithm for finding such a clustering representation was sought that would have the following features:

1. The input should consist solely of the $n(n-1)/2$ similarity measures among the n objects under study. This is in contrast to some previous methods which additionally require that each object be initially represented as a point in Euclidean space. (In many applications the restriction to a representation of the grouping in the concrete, spatial sense of an Euclidean metric seems unnecessarily and undesirably severe).
2. There should be a clear, explicit, and intuitive description of the clustering; i.e., the clusters should mean something. Some of the published clustering methods have nice algorithms, but when they have been carried out it is difficult to see exactly what problem has been solved.
3. The clustering procedure should be essentially invariant under monotone transformations of the similarity data. Often in psychology we have confidence in our data only up to rank-order; the absolute numbers obtained from the experiments may lie along virtually any scale. The method of Ward [1963], which inspired much of this current study, is indeed so general as to permit monotone invariant methods, but they are not explicitly treated.

The notion of a hierarchical clustering scheme, the central idea of this paper, was abstracted from examples given by Ward [1963]. We first consider such schemes, and develop a correspondence between hierarchical clustering schemes and a certain type of metric. Two recursive methods are then given for obtaining hierarchical clustering schemes from a given similarity matrix, and finally the significance of these two methods is discussed and illustrated by application to real data.

I. *Clusterings and Metrics*

Figure 1 gives the typical results of a hierarchical clustering method, such as those discussed by Ward [1963] and others.

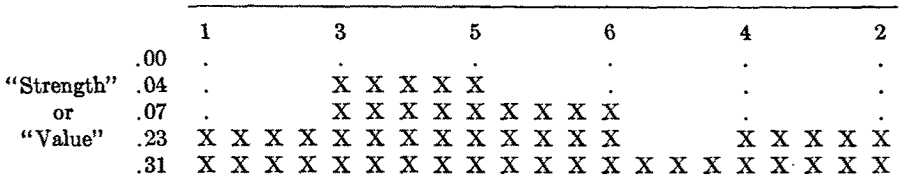

FIGURE 1
A Hierarchical Clustering Scheme

Notice the main features of such a result. The first clustering (top row) is the "weak" clustering—each object is a cluster, so with six objects we have six clusters. This is given the "value" or "rating" .00. Next we have a clustering with five clusters; the set [3, 5] is one cluster, and the remaining four objects are themselves clusters. This is given the value .04. At level .07 we have a clustering with four clusters [1], [4], [2], and [3, 5, 6]. At level .23 we have the two clusters [1, 3, 5, 6] and [2, 4], and finally at level .31 we have the "strong" clustering, with all objects in the same cluster.

We examine the following relevant features of this model. First, the "values" start at 0 and increase strictly as we read down the table. Second, and more important, the clusterings "increase" also, hierarchically; each clustering (except, evidently, the first) is obtained by the merging of clusters at the previous level. For example, if level .23 had had clusters [1, 3], [5, 6, 4], and [2] we would have not had a hierarchical clustering; the cluster [1, 3] cannot be obtained by merging any of the .07 level clusters. Finally we see that the first clustering is the weak clustering and the last is the strong clustering.

We now abstract from this simple example to the general notion of a hierarchical clustering scheme. We assume we have n objects, represented by the integers 1 through n. We have also a sequence of $m + 1$ clusterings, $C_0 , C_1 , \cdots , C_m$, and with each clustering C_i we have a number α_i , its *value*. We require that C_0 be the weak clustering of the n objects, with $\alpha_0 = 0$, and that C_m be the strong clustering. We require also that the numbers α_i increase; $\alpha_{i-1} \leq \alpha_i$, for $j = 1, 2, \cdots , m$, and the clusters "increase" also, where again $C_{i-1} < C_i$ means that every cluster in C_i is the merging (or *union*) of clusters in C_{i-1} . This general arrangement will be referred to as a *hierarchical clustering scheme*, or HCS for short.

This section will demonstrate that every HCS gives rise to a particular kind of distance, or *metric*, between the objects 1, 2, $\cdots$, n, and, conversely,

that given such a metric we may recover the HCS from it. This reduces the study of HCS's to the study of these metrics.

First, we shall assume that we are given an HCS, a sequence of clusterings $C_0 , \cdots , C_m$ with values $\alpha_0 , \cdots , \alpha_m$. For each pair x, y of objects, we shall define $d(x, y)$, a number, and prove that d is a metric.

We define d as follows: given the two objects x and y, we notice that in C_m (the strong clustering) x and y are in the same cluster. Let j be the *least* integer in the set $[0, 1, \cdots , m]$ such that, in the clustering C_i , x and y are in the same cluster. We define

$$d(x, y) = \alpha_j.$$

For example, in Fig. 1 we have $d(d, 5) = .04$ (since 3 and 5 are clustered at level .04 but not at level .0), $d(1, 4) = .31$ (since 1 and 4 are clustered at level .31 but not at level .23), $d(1, 6) = .23$, $d(5, 5) = .00$, $d(4, 2) = .23$, and so on—the complete distance matrix is given in Table 1

TABLE 1
Distance Matrix Corresponding to Figure 1

d	1	2	3	4	5	6
1	0	.31	.23	.31	.23	.23
2	.31	0	.31	.23	.31	.31
3	.23	.31	0	.31	.04	.07
4	.31	.23	.31	0	.31	.31
5	.23	.31	.04	.31	0	.07
6	.23	.31	.07	.31	.07	0

A few things are immediate from the definition: for example x and x are in the same cluster (evidently!) for all C_i ; 0 is the smallest j, so by definition

$$d(x, x) = \alpha_0 = 0.$$

Conversely, if $d(x, y) = 0$ for some x and y, it implies that x and y are in the same cluster in C_0—but, C_0 being the weak clustering, the only element in the same cluster with x is x itself—that is, $d(x, y) = 0$ implies $x = y$. Thus $d(x, y) = 0$ if and only if $x = y$.

We see also that $d(x, y) = d(y, x)$ for all objects x and y. To show that d is a good metric it remains only to show the triangle inequality. Let x, y, and z be any three objects, and let

$$d(x, y) = \alpha_i$$

$$d(y, z) = \alpha_k$$

Thus x and y are in the same cluster in C_j, and y and z are in the same cluster in C_k. Because the clusterings are hierarchical, one of these clusters includes the other; in fact, that cluster corresponding to the larger of j and k. Let this integer be ℓ, then in C_ℓ, x, y, and z are all in the same cluster. From the definition of d, we see thus that

$$d(x, y) \leq \alpha_\ell .$$

But $l = \max [j, k]$, and the α's increase as their subscripts do, so

$$\alpha_\ell = \max [\alpha_j , \alpha_k],$$

or, finally,

$$d(x, z) \leq \max [d(x, y), d(y, z)].$$

This is called the *ultrametric inequality*; we have shown that d satisfies it. It is plainly stronger than the triangle inequality, which would merely require

$$d(x, z) \leq d(x, y) + d(y, z),$$

for it is evident that

$$\max [d(x, y), d(y, z)] \leq d(x, y) + d(y, z),$$

so

$$d(x, z) \leq d(x, y) + d(y, z).$$

Thus we have taken a HCS and obtained a metric d on the objects which satisfies the ultrametric inequality. We now do the converse—given a distance matrix (such as Table 1) representing some metric d which satisfies the ultrametric inequality, we will construct a HCS (such as Fig. 7) from it. At level 0, we have the weak clustering—six clusters, each with but one object in them. The smallest element of the distance matrix, aside from the 0's, is the .04 entry that appears between objects 3 and 5. Accordingly, we create a clustering with value .04 with 3 and 4 in the same cluster, and the other objects constituting clusters by themselves. Now, we notice one very nice property of Table 1: 3 and 5 are exactly the same distance from any other object—that is, if x denotes object 1, 2, 4, or 6— then

$$d(3, x) = d(5, x), \quad \text{all} \quad x.$$

Thus, in fact it makes sense to talk about the distance from x to the cluster [3, 5].

We indicate this distance in Table 2. In effect we have a new object, [3, 5], which replaces 3 and 5 in the matrix. But now we get our next clustering by using the matrix in Table 2 and applying the same process, i.e., taking

TABLE 2
Distance Matrix for Table 1 After First Clustering

	1	2	[3, 5]	4	6
1	0	.31	.23	.31	.23
2	.31	0	.31	.23	.31
[3, 5]	.23	.31	0	.31	.07
4	.31	.23	.31	0	.31
6	.23	.31	.07	.31	0

the smallest nonzero entry (.07, between {3, 5} and 6) and clustering these together to obtain a clustering, at level .07, containing a cluster [3, 5, 6] and individual clusters [1], [2], and [4]. Once again, we define the distance from [3, 5, 6] to 1, 2, or 4 in a unique manner, construct another distance matrix, and so on. Eventually we end up clustering all objects together to get the strong clustering, and we find that we have completely reconstructed Fig. 1.

The key to the above process is being able to replace two (or more) objects by a cluster, and still being able to define the distance between such clusters and other objects or clusters. This property in turn depends on two essential facts: that d datisfies the ultrametric inequality, and that, at each stage, we cluster the minimum distances.

We now generalize this method, to enable us to get a HCS, given n objects and a metric d on them which satisfies the ultrametric inequality.

Step 1. Clustering C_0 , with value 0, is the weak clustering.

Step 2. Assume we are given the clustering C_{i-1} with the distance matrix between each cluster or object and every other. Let α_i be the smallest nonzero entry in the matrix. Merge the pair of points and/or clusters with distance α_i , to create C_i , of value α_i .

Step 3. We may create a new distance matrix, treating the new clusters as objects, in an unambiguous manner.

That is, if x and y are two objects (possibly clusters) at level C_{i-1} , and if $d(x, y) = \alpha_i$ (so that x and y become clustered in C_i), and if z is any other object or cluster at level C_{i-1} , then $d(x, z) = d(y, z)$. The proof of this is easily sketched—if $d(x, z) \neq d(y, z)$ one must be larger—say $d(x, z) > d(y, z)$. Then, however, the ultrametric inequality demands

$$d(x, z) \leq \max (d(x, y), d(y, z))$$

$$\leq \max (\alpha_i , d(y, z))$$

By assumption,

$$d(y, z) < d(x, z).$$

Thus,

$$d(x, z) \leq \alpha_i \ .$$

But α_i was chosen to be the *least nonzero distance* in the matrix; thus

$$d(x, z) = \alpha_i \ .$$

This then in turn requires that

$$d(y, z) < d(x, z) = \alpha_i \ ,$$

a contradiction, since no nonzero distance can be strictly smaller than α_i . Thus the hypothesis that $d(x, z) \neq d(y, z)$ leads to a contradiction; the two distances must be equal, and we can construct our reduced matrix.

Step 4. We now repeat Steps 2 and 3 until we finally obtain the strong clustering—we are then finished.

This procedure evidently produces a hierarchical clustering scheme, since each clustering is a merging of clusters from the previous clustering and the α_i increase. There is thus a complete correspondence between HCS's on the one hand, and metrics satisfying the ultrametric inequality on the other.

II. *The Two Methods*

In the first section we developed a natural way of going from a metric d, satisfying the ultrametric inequality, to an HCS. In general, however, (if only because of noisy data) the similarity matrix does not satisfy the ultrametric inequality—we will thus try to modify our method to give us reasonable clusterings in this case.

When we went from a metric to an HCS in Sect. I, we required the ultrametric inequality only in Step 3; we assumed that we have clusters and/or objects x and y from C_{i-1} which clustered in C_i (i.e., $d(x, y) = \alpha_i$). We then took any third cluster or object z and attempted to define $d([x, y], z)$. The ultrametric inequality told us that $d(x, z) = d(y, z)$, and thus led to a natural definition:

$$d([x, y], z) = d(x, z) = d(y, z).$$

In general we may not expect $d(x, z) = d(y, z)$, but we still may formally define $d([x, y], z)$ as some function f of $d(x, z)$ and $d(y, z)$,

$$d([x, y], z) = f(d(x, z), d(y, z)),$$

and then proceed as in Sect. I, above. It would be natural to require that if $d(x, z) = d(y, z)$, then

$$f(d(x, z), d(y, z)) = d(x, z) = d(y, z).$$

Then if d satisfies the ultrametric inequality, the process will give the same HCS as the "natural" one described in Sect. I. This still leaves us with a large number of choices for f—geometric means, various weighted averages, and so on. We evidently need stronger conditions on the function f.

This work was strongly influenced by the work of Shepard [1962a & b and Kruskal [1964] on multidimensional scaling, in which the results are invariant under montone transformations of the similarity matrix. Since much data of psychological interest is of this type, it seemed worthwhile to try to develop a clustering program with this feature. Immediately we ruled out most of the common functions for f, since the operations of addition, multiplication, square root, and so on are not monotone invariant. The functions *max* and *min*, however, give rise to monotone invariant clustering methods; the corresponding methods may be summarized as follows:

Minimum Method: Given a similarity function d on n objects, we build an HCS as follows:

Step 1. Clustering C_0 , with value 0, is the weak clustering.

Step 2. Assume we are given the clustering C_{i-1} with the similarity function d, defined for all objects or clusters in C_{i-1} . Let α_i be a minimal nonzero entry in the matrix. Merge the pair of objects and/or clusters with distance α_i to create C_i , of value α_i .

Step 3. We create a new similarity function for C_i in the following manner: if x and y are clustered in C_i and not in C_{i-1} (i.e., $d(x, y) = \alpha_i$) we define the distance from the cluster $[x, y]$ to any third object or cluster, z, by

$$d([x, y], z) = \min [d(x, z), d(y, z)].$$

If x and y are objects and/or clusters in C_{i-1} not clustered in C_i , $d(x, y)$ remains the same. We obtain a new similarity function d for C_i in this way.

Step 4. We now repeat Steps 2 and 3 until we finally obtain the strong clustering—we are then finished.

Maximum Method. Same as the Minimum Method, except in Step 3, where we define

$$d([x, y], z) = \max [d(x, z), d(y, z)]$$

when x and y are two objects and/or clusters of C_{i-1} which cluster in C_i, and z is any third object or cluster of C_i.

NOTE: It is tacitly assumed in the discussion of the methods that the distances in the original matrix are all distinct except for 0. This is not important in the Minimum Method, but difficulties do arise when applying the Maximum Method to matrices with large numbers of identical entries. In practice this restriction rarely produces an ambiguous result.

The above two methods are clearly related to the method described in Sect. I. In particular, if d satisfies the ultrametric inequality, the two methods reduce to the method of Sect. I, as promised.

III. *Nature of the Solutions*

Monotone Invariance

One of the requirements that we have set for our methods is that the solutions be invariant under monotone transformations of the original data. Monotone invariant processes are those which are dependent only on the *rank order* of the data. In our methods, we use the matrix elements twice; first, we find the smallest nonzero matrix element, and second, we form the maximum or minimum of two matrix elements. Both these processes may be carried out knowing nothing of the data except the rank order. Thus the clusterings are unaffected by monotone transformations of the similarity matrix. The values assigned to the clusterings also are determined merely by rank order—thus a monotone transformation of the similarity matrix transforms the values of the clusterings, but leaves the clusterings invariant.

What the Clusterings Mean

Both methods depict basic attributes of the original similarity matrix. In particular, the values assigned to the clusterings have simple meanings in the two methods.

If we are given a clustering obtained by the Maximum Method, we may represent the value of the clustering as follows: for each cluster in the clustering, compute the *diameter* of the cluster (the largest intra-cluster distance). For a given Maximum Method clustering, the value of the clustering is the *maximum diameter of the clusters in the clustering.* At any stage, the distance from object/cluster x to object/cluster y is exactly the diameter of the set x union y. This gives us a simple means of visualizing the clusterings—the Maximum Method attempts at each stage to minimize the diameter of the clusters.

The analysis for the Minimum Method is slightly more involved, but equally basic. A *chain* from object x to object y is any sequence of objects $z_0, z_1, \cdots, z_\ell$ with $z_0 = x$ and $z_\ell = y$. The *size* of a chain is the largest link distance:

$$\text{size} = \max_{i=1,\cdots,\ell} [d(z_{i-1}, z_i)].$$

Given a clustering, we say that the *chain distance* d' from x to y is the minimal chain size of all chains from x to y;

$$d'(x, y) = \min_{\substack{\text{all chains} \\ \text{from } x \text{ to } y}} [\text{size of chain.}]$$

It turns out that d' satisfies the ultrametric inequality and is indeed associated with the HCS we obtain from the Minimum Method. The chain distance intuitively measures a kind of *connectedness* of x and y through intermediate points. We may thus describe the value of a Minimum Method clustering by

$$\text{Value of Clustering} = \max_{\substack{x \text{ and } y \\ \text{in same cluster}}} [d'(x, y)].$$

The above statements may be proved easily by induction; the proofs are omitted here as not being central to the argument.

IV. *Illustrative Application*

In the course of exploring some alternative methods for analyzing existing data on the confusability of various speech sounds under different conditions of filtering and noise, R. N. Shepard recently applied the two methods described here to the data of Miller and Nicely [1955] on confusions among sixteen English consonants. As an illustration of the kind of meaningful results that can be obtained in this way we present the HCS's that were obtained for one of Miller and Nicely's sets of data; viz., their Table VII (based on the condition in which only the low audio frequencies from 200 to 300 cps were passed).

For the purpose of applying the present methods, a symmetric matrix was constructed giving, for each of the $n(n-1)/2$ pairs of consonants x and y, a measure of their similarity, $s(x, y)$, defined by

$$s(x, y) = \frac{f(x, y)}{f(x, x)} + \frac{f(y, x)}{f(y, y)},$$

where $f(x, y)$, for example, is the frequency with which the consonant x was heard as the consonant y according to Miller and Nicely's Table VII. In the analysis, then, the similarity estimate, $s(x, y)$, is treated as an approximately monotonically decreasing function of an assumed underlying distance $d(x, y)$.

The representations obtained by the Maximum and Minimum Methods are shown in Figs. 2 and 3, respectively. Across the top, in each table, are indicated the phonetic symbols for the sixteen consonants studied by Miller and Nicely. (In both cases, it turned out that the diagram for the HCS could be constructed with these consonants in the same order and, indeed, in exactly the order in which they originally were listed by Miller and Nicely.)

The numbers down the left-hand side of each table are the similarity values associated with each clustering in the hierarchical representation. (Notice that since $s(x, y)$ is inversely related to distance, these numbers begin at ∞ for the weak clustering and *decrease*.

Similarity Value	Consonants															
	p	*t*	*k*	*f*	θ	*s*	ʃ	*b*	*d*	*g*	*v*	ð	*z*	ʒ	*m*	*n*
∞	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2.635	.	.	.	.	.	.	.	.	XXX	.	.	.	.	.	.	.
2.234	XXX	.	.	.	.	.	.	.	XXX	.	.	.	.	.	.	.
2.230	XXX	.	.	.	.	XXX	.	.	XXX	.	.	.	.	.	.	.
2.185	XXX	.	.	.	XXXXX	.	.	.	XXX	.	.	.	.	.	.	.
2.123	XXX	.	.	.	XXXXX	.	.	.	XXX	.	.	.	XXX	.	.	.
2.108	XXXXX	.	.	.	XXXXX	.	.	.	XXX	.	.	.	XXX	.	.	.
1.870	XXXXX	.	.	.	XXXXX	.	.	XXXXX	.	.	.	.	XXX	.	.	.
1.683	XXXXX	.	.	XXXXXXX	.	.	.	XXXXX	.	.	.	.	XXX	.	.	.
1.604	XXXXX	.	.	XXXXXXX	.	.	.	XXXXX	.	.	.	.	XXX	.	XXX	.
1.577	XXXXX	.	.	XXXXXXX	.	.	.	XXXXX	.	.	.	XXXXX	.	.	XXX	.
1.567	XXXXX	.	.	XXXXXXX	.	.	.	XXXXX	.	.	XXXXXXX	.	.	.	XXX	.
1.065	XXXXXXXXXXXXX							XXXXX	.	.	XXXXXXX	.	.	.	XXX	.
1.009	XXXXXXXXXXXXX							XXXXXXXXXXXXX							XXX	.
0.425	XXXXXXXXXXXXXXXXXXXXXXXXXXX														XXX	.
0.270	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX															

FIGURE 2
The HCS Obtained on the basis of Miller and Nicely's Table VII by the Minimum Method

Similarity Value	Consonants															
	p	*t*	*k*	*f*	θ	*s*	ʃ	*b*	*d*	*g*	*v*	ð	*z*	ʒ	*m*	*n*
∞	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2.635	.	.	.	.	.	.	.	.	XXX	.	.	.	.	.	.	.
2.234	XXX	.	.	.	.	.	.	.	XXX	.	.	.	.	.	.	.
2.230	XXX	.	.	.	.	XXX	.	.	XXX	.	.	.	.	.	.	.
2.123	XXX	.	.	.	.	XXX	.	.	XXX	.	.	.	XXX	.	.	.
1.855	XXXXX	.	.	.	.	XXX	.	.	XXX	.	.	.	XXX	.	.	.
1.683	XXXXX	.	.	.	XXXXX	.	.	.	XXX	.	.	.	XXX	.	.	.
1.604	XXXXX	.	.	.	XXXXX	.	.	.	XXX	.	.	.	XXX	.	XXX	.
1.525	XXXXX	.	.	.	XXXXX	.	.	.	XXX	.	.	XXXXX	.	.	XXX	.
1.186	XXXXX	.	.	.	XXXXX	.	.	.	XXX	.	XXXXXXX	.	.	.	XXX	.
1.119	XXXXX	.	.	.	XXXXX	.	.	XXXXX	.	.	XXXXXXX	.	.	.	XXX	.
0.939	XXXXX	.	.	XXXXXXX	.	.	.	XXXXX	.	.	XXXXXXX	.	.	.	XXX	.
0.422	XXXXXXXXXXXXX							XXXXX	.	.	XXXXXXX	.	.	.	XXX	.
0.302	XXXXXXXXXXXXX							XXXXXXXXXXXXX							XXX	.
0.019	XXXXXXXXXXXXX							XXXXXXXXXXXXXXXXX								
0.000	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX															

FIGURE 3
The HCS obtained on the basis of Miller and Nicely's Table VII by the Maximum Method

For these data the Maximum and Minimum Methods yield very similar results. The principal difference between the two representations is confined to the order in which the last three clusters [*p*, *t*, *k*, *f*, θ, *s*, ʃ], [*b*, *d*, *g*, *v*, ð, *z*, ʒ], and [*m*, *n*] combine into two clusters. For the Maximum Method the last two of these combine with each other before joining the first, whereas for the Minimum Method the first two combine with each other before joining the last. Otherwise, although the precise numerical values associated with the clusterings differ somewhat between the two methods, the topological structures of the two representations are alike. That is, above the level of three clusters, we find that exactly the same subclusters appear in both representations and (consequently) that each such subcluster divides into exactly the same sub-subclusters. This close agreement suggests that these data do not seriously violate the assumed ultrametric structure.

Moreover, both of the obtained HCS's are meaningfully related to the distinctive features presumed [e.g., by Miller and Nicely, 1955] to govern the discrimination of consonant phonemes. At the level of five clusters, for example, the sixteen phonemes divide into the unvoiced stops [*p*, *t*, *k*], the corresponding voiced stops [*b*, *d*, *g*], the unvoiced fricatives [*f*, θ, *s*, ʃ], the corresponding voiced fricatives [*v*, ð, *z*, ʒ], and the (voiced) nasals [*m*, *n*]. Then, at the level of three clusters, the stops and fricatives coalesce for the voiced and unvoiced phonemes, separately, to yield just the nasals, the remaining voiced, and the corresponding unvoiced consonants.

Analyses of other of Miller and Nicely's matrices (which were obtained under different conditions of filtering) led to clusterings that, although highly consistent (across independent sets of data), departed systematically from the HCS's presented here for their Table VII. These divergent results will be covered in a forthcoming report by Shepard; their detailed discussion here would require too extensive a detour into the substantive problems of psychoacoustics. One further observation should perhaps be made here, though, regarding these further analyses. In this particular kind of application anyway, it has generally appeared that, to the extent that there is an appreciable departure between the HCS's obtained by the Maximum and Minimum Methods, the results of the Maximum Method have appeared to be the more meaningful or interpretable. That is, the search for compact clusters (of small over-all "diameter") has proved more useful than the search for internally "connected" but potentially long chain-like clusters. The reverse may of course prove to be true in other types of applications.

V. *Discussion*

Relation to Other, Similar Methods

Although the methods described here were developed independently, they were subsequently found to be closely related to some methods that

had been developed earlier for applications to biological taxonomy. In particular, what are here called the Minimum Method and the Maximum Method appear to be essentially like the earlier methods of Sneath [1957] and of Sørensen [1948], respectively, (which methods Sokal and Sneath [1963, pp. 180–181] have, in turn, called "clustering by single linkage" and "clustering by complete linkage," respectively). A more recent method of this same general type is that of "hierarchical linkage analysis" proposed by McQuitty [1960]. In view of the general upsurge of interest in clustering methods that is currently taking place in a number of different fields, it is likely that similar methods have been proposed by others as well.

Apart from what function it may serve to bring such methods to the attention of psychologists, the present report has the advantage of providing, for the first time, a unifying conceptual formulation; specifically, a formulation based upon the notion of the ultrametric. This ultrametric conceptualization, moreover, leads directly to readily mechanizable computing algorithms for both the Minimum and Maximum Methods, as well as certain intermediate methods (which are briefly mentioned below). More importantly, it allows one to specify—as had not previously been done—precisely what type of underlying structure is being assumed and, hence, precisely what problem is being solved.

A Computer Program

Another step that has been taken here is the construction of a computer program that will carry out both the Maximum and Minimum Methods on an arbitrary matrix of similarities or "proximities." (The program is written in FORTRAN and is suitable for IBM machines of the 709-7090 class.) The solutions displayed in the present Tables 11 and 12 were in fact computed and printed out (in the form shown) by this program. When necessary the program also determines an appropriate reordering of the "objects" so that such a table can be constructed. On an IBM 7094, the analysis is completed quite rapidly; in another application with 64 objects, solutions were obtained for both the Minimum and Maximum Methods in just 10.1 seconds.

Possible Extensions

Sokal and Sneath [1963, p. 190] have pointed out that, in methods like our Minimum or Maximum Methods, the merging of two clusters depends upon a single similarity value (viz., the least or greatest in the appropriate set). They suggest that, for greater robustness of the solution, it may sometimes be desirable to use some sort of average value instead. As we have already noted, to base such a procedure upon averages of the more obvious types is to lose the invariance, sought here, under monotone transformations of the similarity values. More importantly, the solutions would no longer

have the clear-cut meaning of the "connected" or "compact" solutions obtained by the conceptually simpler Min. and Max. Methods.

Nevertheless, when this seems desirable, the methods described here can be (and, indeed, have been) modified to yield solutions intermediate between those obtained by these two extreme methods. J. D. Carroll (personal communication) has suggested an average method based upon medians which, of course, do have the desired property of monotone invariance. The main problem, in the case of medians, is the choice of an appropriate procedure for dealing with the ambiguities that tend to arise when two or more of the initial similarity estimates are tied. Moreover, as in the case of other sorts of averages, the solution no longer lends itself to a simple characterization (e.g., in terms of "compactness" or "connectedness").

Finally, a different kind of possible extension will be briefly indicated here, in the form of a presently unsolved problem. In Section I we saw that the construction of an HCS is equivalent to finding a metric which satisfies the ultrametric inequality. Given a similarity measure d, we would in general like to find the closest metric D which satisfies the ultrametric inequality—various measures of closeness could be used. For example, we could use a rank-order correlation between $d(x, y)$ and $D(x, y)$ over all objects x and y. To the author's knowledge, this problem is unsolved.

REFERENCES

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964, **29**, 1–27.

McQuitty, L. L. Hierarchical linkage analysis for the isolation of types. *Educational and Psychological Measurement*, 1960, **20**, 55–67.

Miller, G. A. and Nicely, P. E. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 1955, **27**, 338–352.

Shepard, R. N. Analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 1962a, **27**, 125–140.

Shepard, R. N. Analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 1962b, **27**, 219–246.

Sneath, P. H. A. The application of computers to taxonomy. *Journal of General Microbiology*, 1957, **17**, 201–226.

Sokal, R. R. and Sneath, P. H. A. *Principles of Numerical Taxonomy*. San Francisco: W. H. Freeman, 1963.

Sørensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 1948, **5** (4), 1–34.

Ward, J. H., Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 1963, **58**, 236–244.

Manuscript received 5/9/66
Revised manuscript received 12/14/66