Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241–254 (3755 citations in Google Scholar as of 4/1/2016).

One of the ten most-cited *Psychometrika* articles is by Stephen Johnson, who was a young computer scientist working at Bell Labs at the time the article was written. Interestingly, this piece is Johnson's sole published work in the area of statistics and/or psychometrics. It was obviously influenced heavily by other Bell Labs personnel at the time, including several future Psychometric Society presidents. The introductory footnote to the article reads: "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."

The Johnson paper in *Psychometrika* is nicely written and accessible, and provides for the first time in the literature the key concept of an ultrametric for characterizing all hierarchical clustering schemes. Johnson's inspiration for developing the correspondence between ultrametrics and hierarchical clustering schemes was from the work of Joe Ward, a personnel research psychologist working in the 1950s and 60s on occupational analyses for the United States Air Force. To quote Johnson (p. 242):

> 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.

To characterize more formally the basic problem posed 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, p_{ij}, is available (that might possibly be constructed from a more basic object by variable data matrix). It is assumed that the proximity measure has a dissimilarity interpretation so larger proximity values correspond to more dissimilar objects. These proximity values are collected into an $n \times n$ proximity matrix, $\mathbf{P} = \{p_{ij}\}_{n \times n}$. Any hierarchical clustering strategy produces (using Johnson's terminology) a hierarchical clustering scheme; the latter is a sequence or hierarchy of partitions of S, denoted $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_{n-1}$, from the information present in $\mathbf{P}$. In particular, the (disjoint) partition, $\mathcal{P}_0$, contains all objects in separate classes (Johnson's "weak" clustering); $\mathcal{P}_{n-1}$ (the conjoint partition) consists of one all-inclusive object class (Johnson's "strong" clustering); and $\mathcal{P}_{k+1}$ is defined from $\mathcal{P}_k$ by uniting a single pair of subsets in $\mathcal{P}_k$.

Generally, the two subsets chosen to unite in defining $\mathcal{P}_{k+1}$ from $\mathcal{P}_k$ are those that are "closest," with the characterization of this latter term specifying the particular hierarchical clustering method being used. We mention three of the most common options for this notion of closeness:

(a) complete-link: the maximum proximity value attained for pairs of objects within the union of two sets (thus, the maximum link [or the subset "diameter"] is minimized);

(b) single-link: the minimum proximity value attained for pairs of objects where the two objects from the pair belong to the separate classes (thus, we minimize the minimum link);

(c) average-link: the average proximity over pairs of objects defined across the separate classes (thus, the average link is minimized).

From the time of Johnson's 1967 paper (which mainly emphasized the single-link and complete-link criteria), it has been generally accepted that the complete-link criterion should be the default selection for the task of hierarchical clustering when done in the traditional agglomerative way that starts from $\mathcal{P}_0$ and proceeds step-by-step to $\mathcal{P}_{n-1}$. A reliance on the single-link criterion tends to produce "straggly" clusters that are not very internally homogenous or substantively interpretable. To quote Johnson (p. 252):

> ... the results of the Maximum Method [that is, complete-link] have appeared to be the more meaningful or interpretable. That is, the search for compact clusters (of small over-all "diameter") have proved more useful than the search for internally "connected" but potentially long chain-like clusters.

As noted by many users over the years, the average-link choice seems to produce results that are the same as or very similar to the complete-link criterion but relies on more information from the given proximities; the complete-link criterion (or for that matter, the single-link criterion) depends only on the rank order of the proximities. Given the Bell Labs context in which the Johnson paper was produced (with Kruskal and Shepard as senior colleagues to Johnson and the recent development of *nonmetric* multidimensional scaling) it shouldn't be surprising that Johnson emphasized only single-link and complete-link clustering. To quote (p. 253):

> Sokal and Sneath [1963, p. 190] have pointed out that, in methods like our Minimum or Maximum Methods [that is, single- or complete-link], 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 transformation of the similarity values.* [emphasis added]

Johnson notes that the fortran program he wrote to carry out both single-link and complete-link clustering (called `hiclust.f`, and still available at `www.netlib.org`) could be easily modified to include other clustering criteria such as average-link. Again, we give a relevant quote (p. 254):

> 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.

As noted earlier, the seminal contribution of Johnson's paper and the reason for its continued popularity is the characterization of a hierarchical clustering scheme in terms of what is called an ultrametric. We now turn to a more formal definition.

Given the partition hierarchies from any of the three criteria mentioned (complete-, single-, or average-link), suppose the values determined when the new subsets were formed (that is, the maximum, minimum, or average proximity between the united subsets) are placed into an $n \times n$ matrix, $\mathbf{U}$. In general, there are $n-1$ distinct nonzero values that define the levels at which the $n-1$ new subsets are formed in the hierarchy; thus, there are typically $n-1$ distinct nonzero values present in an appropriately row and column reordered matrix $\mathbf{U}$ that characterizes the identical blocks of matrix entries between subsets united in forming the hierarchy.

Based on a matrix such as $\mathbf{U}$, the partition hierarchy can be retrieved immediately along with the levels at which the new subsets were formed. In fact, any (strictly) monotone (that is, order preserving) transformation of the $n-1$ distinct values in such a matrix $\mathbf{U}$ would serve the same retrieval purposes. Thus, as one illustration, the $n-1$ distinct values in $\mathbf{U}$ could be replaced by the simple integers, $(1, 2, 3, \ldots, n-1)$, and the partitions of the hierarchy could still be reconstructed easily. Generally, a matrix $\mathbf{U}$ that can be used to retrieve a partition hierarchy in this way is called an ultrametric (matrix):

A matrix $\mathbf{U}$ is called an ultrametric (matrix) if for every triple of subscripts, $i, j,$ and k, $u_{ij} \leq \max(u_{ik}, u_{kj})$; or equivalently (and much more understandably), among the three terms, $u_{ij}, u_{ik},$ and u_{kj}, the largest two values are equal.

The last paragraph of Johnson's paper is very prescient about where his introduction of the ultrametric concept will lead. This last paragraph begins (p. 254):

> 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. ... To the author's knowledge, this problem is unsolved.[1]

So, as Johnson foreshadowed, it is now common in the literature to reformulate this hierarchical clustering task as that of locating a best-fitting ultrametric matrix, say, $\mathbf{U}^* = \{u^*_{ij}\}$, to the given proximity matrix, $\mathbf{P}$, such that the least squares criterion

$$\sum_{i<j} (p_{ij} - u^*_{ij})^2 \ ,$$

is minimized. Nevertheless, Johnson's paper is a true classic, as is evidenced by the fact that about one-fifth of its citations were received in the most recent four years since its appearance almost fifty years ago.

References

[1] Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy.* San Francisco: W. H. Freeman.

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

[1] This comment about being "unsolved" is somewhat of an understatement. The task of finding the "closest" ultrametric falls within the class of optimization problems now known as NP-hard, which includes all the old combinatorial chestnuts, such as the traveling salesman problem. Although Johnson was a computer scientist, he cannot be faulted for not proving or knowing this in 1967. The notion of a problem being NP-hard was not even introduced into the computer science literature until the early 1970s.