ANALYSIS OF SOCIOMETRIC STRUCTURE: A METHOD OF SUCCESSIVE GROUPING[1]

Jack Sawyer and Terrance A. Nosanchuk

University of Chicago

The method described here was developed for defining the structure of a social group, although it can be applied to grouping problems in many other areas. Essentially, it considers a given group of individuals and assigns each to one of a mutually exclusive and exhaustive set of sub-groups.

This problem of defining sub-groups has been one of general interest since the first work of Moreno (1934) on the two-dimensional graphical "sociogram" representation. The conception of the problem was broadened by Forsyth and Katz (1946) who introduced matrices in which the ij entry is the relational measure between individuals i and j--a one for "chosen", a zero for "indifferent," and a minus one for "rejected." They proposed rearranging the rows and columns of this matrix such that the plus ones were near the diagonal and the minus ones away from the diagonal; lines could then be drawn at appropriate places to separate sub-groups. Katz (1947) suggested, as a more definite criterion, minimizing $\Sigma\Sigma e_{ij}(i-j)^2$, where e_{ij} is the element in the ith row and the jth column. Beum and Brundage (1950) provided an iterative, and sometimes lengthy, procedure for achieving the minimization. Bock and Husain (1950) also employed a matrix in grouping individuals into sub-groups, but did so on the basis of Holzinger's B coefficient.

Luce and Perry (1949) utilized the matrix formulation in a different way--to obtain, by matrix multiplication, a special kind of sub-group composed of individuals all of whom choose each other. They extended this concept to include individuals connected to each other not through direct choice, but through a third person, as in the connection of persons A and C represented by A choosing B and B choosing C. Luce subsequently (1950) further generalized the concept of the clique and provided procedures for clique identification. Harary and Ross (1957) provided a general solution for obtaining all the cliques in a given matrix.

Bock and Husain (1952) factored a matrix of choices, a technique which groups persons having similar choice patterns. A "direct" factor analysis of the score matrix (rather than of the correlations) was used by MacRae (1959) to obtain two sets of factors, one grouping persons by whom they chose, the other grouping on the basis of whom they were chosen by. Another approach has been that of mathematical graph theory, the study of directed line segments, explored by Harary and Norman (1953). A recent summary of much of the work briefly mentioned here is available in Glanzer and Glaser (1959).

Applications in Other Areas

Although the grouping problem is particularly significant in the area of sociometrics, it also exists in many other areas. Rao (1952) employed a grouping method suggested by K. D. Tocher (1948) to determine how 12 Indian castes and tribes group together. In some cases groups are known and a multiple discriminant function may be used to classify members, as Rao illustrates (1952) in identifying the Highdown skull as Iron Age rather than Bronze Age.

A number of methods have been devised and utilized in connection with the grouping of psychological test scores, where *a priori* groups are not known. Holzinger and Harmon (1941) employ their method of B-coefficients to group 24 psychological tests on the basis of their inter-correlations. Cattell (1944) has summarized a number of methods of grouping psychological tests, all of which require, however, setting an arbitrary value above which two variables are considered to be related. More recently, McQuitty (1957, 1960) has proposed a number of "linkage analysis" methods.

The Grouping Methods

Any method of grouping takes as given some measure of relation among each of the pairs of individuals in the set being grouped. The present method works with any relational measure, such as strength of preference or frequency of interaction, although reference will be made to the "distance" between two points, perhaps a more general view. Such a distance measure might be determined, for example, by the method of multidimensional scaling, or by distance in a space spanned by orthogonal personality components.

Taking the matrix of relational measures as given, this method provides a completely determined, easily applied procedure for defining group structure, for any number of groups, 1, 2, ..., n-1,n, where n is the total number of points. The method proceeds by first regarding the n points as n one-point groups, and

[1] This research was supported in part by the National Science Foundation.

forming the best set of n-1 groups by combining those two groups whose combination minimizes some criterion. These two groups now form a new, single group, and the procedure is repeated on the resultant set of n-1 groups. Thus the best set of n-2 groups is found, and so on; at each stage, two of the previously existing groups are combined, until finally, the last two groups are combined to form a single group of n points.

The value of the criterion measure (on the basis of which the groupings are made) after any grouping may be compared with the corresponding measure following each of the preceding groupings. A sharp jump might suggest not making that grouping, and reverting to the previous stage. There has been, however, no investigation of the stochastic properties of this measure to determine, for example, how large an increase must be to attain statistical significance.

The above method of regarding two groups, once joined, as permanently bonded, permits practical solution of the grouping problem; without this restriction there exist an overwhelming number of combinations--more than half a million different sets of groupings for only ten individuals, for example. Regarding two groups, once joined, as a single group, means that at a stage at which there are r groups, only r(r-1)/2 combinations need be considered.

In addition to a method for grouping, one needs a criterion on the basis of which to combine groups. The obvious general criterion is within-group homogeneity: groups should in some sense be internally homogeneous relative to other groups. Within this general conception, however, two distinctions can be made in the criteria employed. There is, first, the distinction whether in computing the "distance" between two sets of individuals (or among the individuals in a single group) one takes the sum of the individual inter-point distances, or their average. The second distinction concerns for which groups the average or total distance measure is computed.

The second distinction creates three cases; the first case minimizes the total (or average) distance between points which are in the same group. This measure starts at zero, of course, for the original n one-point groups, and grows to the sum (or average) of the distances between all the points when all the groups are finally combined into one. At each stage, one combines those two groups whose combination adds least to the intra-group distance.

In the second case, one considers each pair of groups by itself and combines those two which are "closest" in terms of the total (or average) distance between the points in the first group and the points in the second group. This case may, of course, give a different grouping from the preceding. If, for example, the two closest groups were very large, their combination would inflate the average much more than two smaller groups, even though the latter were more distant from each other. If the total distance, rather than the average distance, is employed, however, combining the two closest groups adds least to the total within-group distance over all groups, and this case becomes identical with the first.

The third case is somewhat interim, in that it considers the distance between two groups in relation to their distance from all other groups. The measure is thus a ratio, in which the numerator is the measure of the previous case, the total (or average) distance between two groups. The denominator is the total (or average) distance from the points in these two groups to all other groups. Two groups are combined, then, not simply on the basis of their absolute closeness, but in terms of their closeness relative to how far they are from all other groups. Thus two groups which were a moderate distance from each other, but exceedingly far from all the rest, might be combined prior to other pairs, which were nearer to each other but also nearer to the rest of the groups. The B-coefficient (Holzinger and Harman, 1941) is similar to this measure, although they differ in the numerator.

There thus exist five methods, as follows:

1. Total distance within groups
2. Average distance within groups
3. Average distance between two groups
4. Ratio of average distance between two groups to their average distance to points in all other groups
5. Ratio of total distance between two groups to their total distance to points in all other groups

For illustrative purposes, each of these five methods was applied to the same set of sociometric data. The data consisted of inter-person distances based upon the friendliness of each pair of 15 undergraduate fraternity members, obtained by the multidimensional scaling model (Morton, 1959). These data are presented in Table 1, columns A-O.

Computational Procedure

The balance of Table 1 illustrates the computational procedure for the first method, which minimizes the total distance between two points in the same group. The procedure is as follows.

1. Search the original nxn matrix for the smallest non-diagonal entry. Let it be d_{ij}. Record this amount, which is the increment added at this stage to the

Table 1: Total Distance Between Two Groups

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	FM	GK	CN	EL	IO	AJ	BD	CNH	AJGK	BDIO	ELFM	AJGK ·CNH	BDIO· ELFM
A	317	159	357	215	209	131	204	305	126^{6}	124	290	220	136	302	429	255	295	505	607								
B		270	217^{7}	262	417	376	296	218	302	357	320	415	242	302	832	733	512	582	520	619							
C			327	243	280	217	167	273	164	152	296	209	81^{3}	289	489	369											
D				323	451	401	351	347	350	388	388	437	255	355	888	789	582	711	702	707							
E					285	224	234	248	161	246	97^{4}	271	234	333	556	470	477										
F						111	252	422	241	117	363	70^{1}	298	396													
G							224	366	111	71^{2}	278	113	228	338	224												
H								195	220	184	342	238	149	171	490	408	316^{8}	576	366	424	647						
I									297	361	345	409	304	122^{5}	831	727	577	593									
J										117	207	215	202	329	456	228	266	368	626								
K											303	146	209	359	263												
L												349	289	386	712	581	585										
M													228	408													
N														305	526	437											
O															804	697	603	719									
FM																487	1015	1268^{11}	1635	885	1720	1505	1372	3355			
GK																	806	1051	1424	483^{9}	1522	1214					
CN																		1062	1171	661	1094						
EL																			1312	873	1293	1638	1924	2605			
IO																				1233	1222^{10}	1537	2657				
AJ																					1326	1085					
BD																						1741	2848				
CN•H																							2299^{12}	3278	3143		
AJ•GK																								5505	3296		
BD•IO																									5960^{13}	8783	
EL•FM																										6439	
AJGK•CNH																											15222

sum of the intra-group distances.

2. Create a new group, called IJ, composed of the former groups (points) I and J.

3. Determine the total distance, $d_{ij.k}$, of the new group IJ from any other group K, by adding together d_{ik} and d_{jk}. Form of these $d_{ij.k}$'s a new column IJ., Strike out rows I and J and columns I and J.

4. Repeat steps 1-3 on the resulting (n-1)x(n-1) matrix.

Thus at each stage, the corresponding entries of two groups are combined into a new column, and the rows and columns of the former groups are eliminated. In Table 1, entries corresponding to combined groups are circled and the order of combination is indicated by a superscript. The succession of particular groupings obtained by this method is further illustrated by the tree diagram of Figure 1. The total computation consists, for a group of size n, of (n-1)(n-2)/2 additions, each composed of a single pair of numbers. While the first method is by far the simplest computationally, it illustrates the general procedure of combination for all methods.

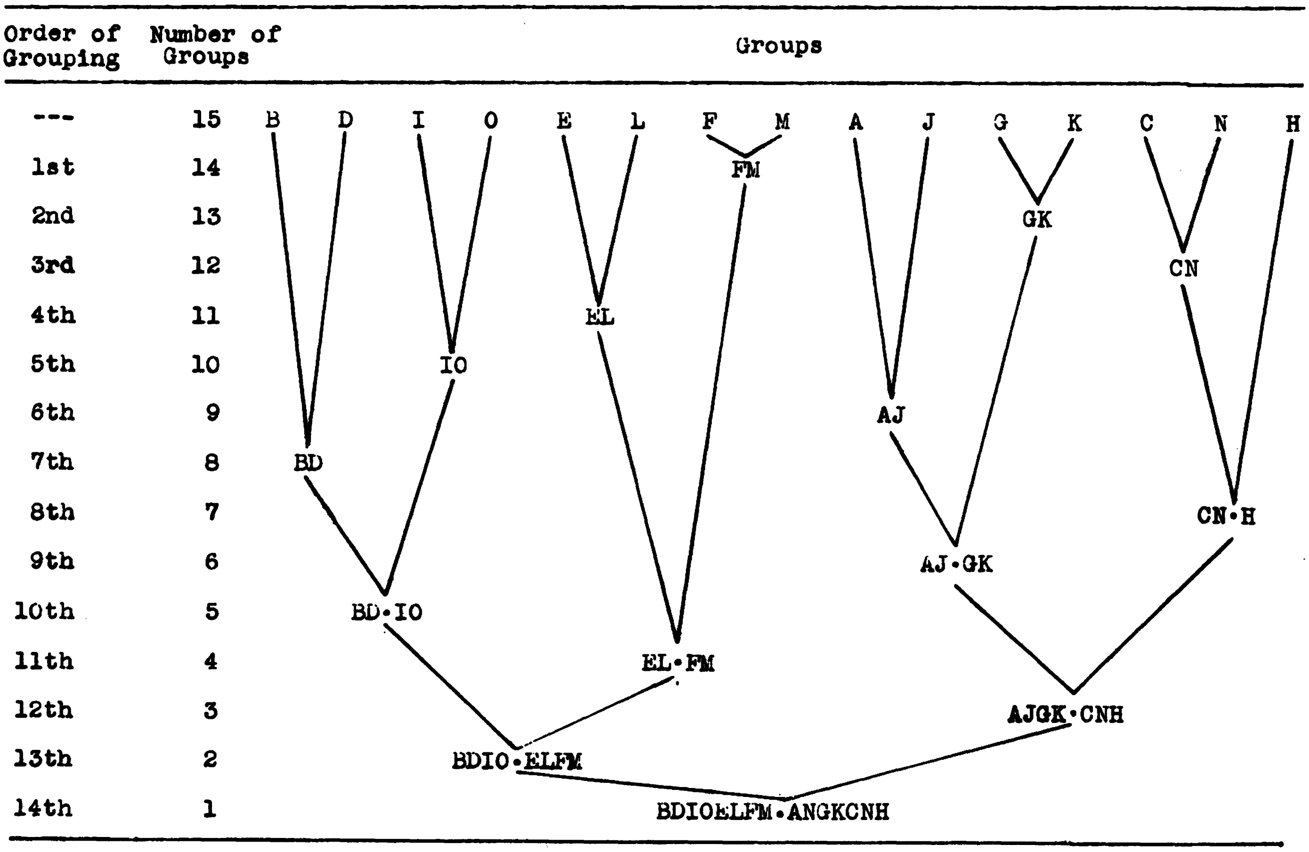
Comparison of the Five Methods

The results of applying each of the five methods are shown in Table 2. The rows indicate which two groups were combined at each stage, and the value of the criterion measure following that combination. As mentioned previously, sharp rises in the criterion may be employed as a stopping point. Thus, using the first method, one might stop after combining AJ (with nine groups) after AJ.GK (with six groups), or after AJGK.CNH (with three groups).

Comparing the methods, it is seen that FM and GK are first and second, respectively, in all five methods, and that the third and fourth groupings are always CN and EL. The agreement continues at a considerable level through the fifth and sixth groupings: IO is fifth in all but one method, where it is sixth; the combination AJ appears by the sixth grouping for three methods, and in the seventh for the other two methods. Beyond that point, there is less agreement, although in all, the 65 groupings result in only 27 different combinations (even though the more than half million

Figure 1

Successive Groupings; Total Distance Criterion



possible combinations would indicate little *a priori* chance of duplication). Some effects of the different methods can be observed. Use of the total distance influences toward groups of equal size, since in that way, the *number* of distances (and hence, likely, the total of distances) within groups is a minimum. Use of the average distance tends toward monolithic structure, one group being very large compared to the other. Other effects may be noted in further empirical evaluation. It is planned, for example, to obtain measures of interaction at successive time intervals in a newly formed group, to determine what model corresponds most closely to the process of the emerging group structure.

Aside from empirical test, certain rational advantages of the foregoing type of method may be cited, the foremost of which is that it is completely determined. Given a set of data, these methods specify precisely the groupings, for any number of groups, without use of an arbitrary level to decide when a relation is to be considered, or when an individual is to be added to a group. Thus it is amenable to computer programming, and this step is anticipated.[2] These methods have a further advantage of being able to utilize data having any values, rather than being restricted to a dichotomy or trichotomy. Thus any added sensitivity in the measurement of distance between two individuals can be employed to advantage. Computationally the first method is extremely simple, although the others are somewhat more complex.

Some open questions remain. In particular there is no provision for error in the distance measure, and the effect of this on the grouping. Further, one would like a more satisfactory way of deciding just what number of groups was most reasonable. Finally, and related to the previous questions, it is often of interest to compare the structure of two groups, or of the same group at different times, and provision is needed for this.

[2] A similar procedure, developed by David L. Wallace and Benjamin Wright, has been programmed for UNIVAC at the University of Chicago.

Table 2

Successive Groupings and Criterion Measure for Five Methods

Order of	Total Distance		Average Distance		Average Distance		Ratio of		Ratio of	
							Average Distances:		Total Distances:	
Grouping	Within Groups		Within Groups		Between Two Groups		Two to all Groups		Two to all Groups	
1st	FM	70	FM	70	FM	70	FM	.2418	FM	.0093
2nd	GK	141	GK	71	GK	71	GK	.2990	GK	.0115
3rd	CN	222	CN	74	CN	81	EL	.3380	EL	.0130
4th	EL	319	EL	80	EL	97	CN	.3432	CN	.0132
5th	IO	441	IO	88	GK.J	114	IO	.3796	IO	.0146
6th	AJ	567	AJ	95	IO	122	FM.GK	.4212	AJ	.0214
7th	BD	784	AJ.GK	105	A.GKJ	127	AJ	.5564	BD	.0246
8th	CN.H	1100	CN.H	114	CN.H	158	AJ.FMGK	.5824	H.IO	.0337
9th	AJ.GK	1583	BD	122	AGKJ.FM	172	H.IO	.6058	FM.GK	.0383
10th	BD.IO	2805	AJGK.FM	142	AGKJFM.CNH	211	BD	.6396	AJ.CN	.0618
11th	EL.FM	4073	CNH.IO	166	BD	217	AJFMGK.CN	.6812	BD.EL	.0946
12th	AJGK.CNH	6372	BD.EL	180	AGKJFMCNH.EL	268	AJFMGKCN.EL	.8346	AJCN.FMGK	.1872
13th	BDIO.ELFM	12332	AJGKFM.CNHIO	224	BD.IO	306	AJFMGKCNEL.HIO	.9438	BDEL.HIO	.2208
14th	All	27454	All	262	All	335	All	∞	All	∞

Summary

This method takes as given a measure of relation (based, for example, upon frequency of interaction, strength of preference, or distance in a space spanned by orthogonal personality components) between all pairs of n individuals in a group. On the basis of these measures, from 1 to n groups may be formed, using mutually exclusive sub-sets of individuals. The method, at all stages, is completely determined. One proceeds by first regarding the n individual as n one-person groups, then joins those two groups whose combination minimizes some criterion. These two groups now form a new, single group, and the procedure is repeated on the resultant set of n-1 groups. As this successive grouping process is continued, the increment added to a measure of group homogeneity may be used as a criterion for stopping. Various models for behavior in the formation of groups are represented by different criteria for grouping, which vary in two respects: (1) whether the total or average relation between sets of individuals is employed as criterion, and (2) whether this criterion refers to only the two groups being joined, or to all groups. A comparison of the resultant cases is made upon friendship distances for a group of undergraduate men.

References

Beum, C. O., & Brundage, E. G. A method for analyzing the sociomatrix, *Sociometry*, 1950, 13, 141-145.

Bock, R. D., & Husain, Suraya Z. An adaptation of Holzinger's B-coefficients for the analysis of sociometric data. *Sociometry*, 1950, 13, 146-153.

Bock, R. D., & Husain, Suraya Z. Factors of the tele: a preliminary report. *Sociometry*, 1952, 15, 206-219.

Cattell, R. B. A note on correlation clusters and cluster search methods. *Psychometrika*, 1944, 9, 169-184.

Forsyth, Elaine, & Katz, L. A matrix approach to the analysis of sociometric data: preliminary report. *Sociometry*, 1946, 9, 340-347.

Glanzer, M., & Glaser, R. Techniques for the study of group structure and behavior: I. Analysis of structure. *Psychol. Bull.*, 1959, 56, 317-322.

Harary, F. & Norman, R. Z. *Graph theory as a mathematical model in social science*. Ann Arbor: Institute for Social Research, Univer. of Michigan, 1953.

Harary, F., & Ross, I. C. A procedure for clique detection using the group matrix. *Sociometry*, 1957, 20-205-215.

Holzinger, K. J., & Harmon, H. H. *Factor analysis: a synthesis of factorial methods*. Chicago: University of Chicago, 1941.

Katz, L. On the matrix analysis of sociometric data. *Sociometry*, 1947, 10, 233-241.

Luce, R. D. Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 1950, 15, 169-190.

Luce, R. D., & Perry, A. D. A method of matrix analysis of group structure. *Psychometrika*, 1949, 14, 95-116.

MacRae, D., Jr. Direct factor analysis of sociometric data. Paper presented at the American Association for the Advancement of Science, Chicago, December, 1959.

McQuitty, L. L. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. *Educ. psychol. Measmt.*, 1957, 17, 207-229.

McQuitty, L. L. Hierarchical linkage analysis for the isolation of types. *Educ. psychol. Measmt.*, 1960, 29, 55-67.

Moreno, J. L. *Who shall survive?* Washington: Nervous and Mental Disease Monograph, No. 58, 1934.

Morton, A. S. *Similarity as a determinant of friendship: A multidimensional study*. Princeton: Princeton University and Educational Testing Service, 1959.

Rao, C. R. *Advanced statistical methods in biometric research*. New York: Wiley, 1952.

Tocher, K. D. Discussion on Mr. Rao's paper. *J. roy. statist. Soc., Sec. B*, 1948, 10, 198-199.