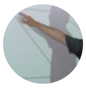
An Algorithm for Clustering Relational Data with Applications to Social Network Analysis and Comparison with Multidimensional Scaling

Article *in* Journal of Mathematical Psychology · August 1975

Impact Factor: 2.61 · DOI: 10.1016/0022-2496(75)90028-0

3 authors, including:



Ronald Breiger
The University of Arizona

AN ALGORITHM FOR BLOCKING RELATIONAL DATA, WITH APPLICATIONS TO SOCIAL NETWORK ANALYSIS AND COMPARISON WITH MULTIDIMENSIONAL SCALING

by

Ronald L. Breiger, Scott A. Boorman and Phipps Arabie

TECHNICAL REPORT NO. 244

December 13, 1974

PSYCHOLOGY AND EDUCATION SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

ACKNOWLEDGMENTS

We are indebted to Harrison White for detailed and productive comments, and also to Paul Holland, Ingram Olkin, Joel Levine, Paul Levitt, Frederick Mosteller, and Joseph Schwartz. Paul Levitt generously gave assistance in surmounting difficulties with a packaged APL version of the Hungarian method for optimal assignment. Together with Breiger, Schwartz co-discovered the mathematical convergence fact on which the CONCOR algorithm rests; he also has contributed many valuable ideas and comments, in particular the approach to simultaneous treatment of multiple relations (p. 16 below). We also thank S. F. Sampson for allowing us to cite data and interpretation from his unpublished study, Crisis in a Cloister. Support for the present research was obtained through NSF Grant GS-2689 (Principal Investigator: Harrison White) and NIMH Grant MH 21747 (Principal Investigator: Richard C. Atkinson).

Abstract

A method of hierarchical clustering for relational data is presented, which begins by forming a new square matrix of product-moment correlations between the columns (or rows) of the original data (represented as an n x m matrix). Iterative application of this simple procedure will in general converge to a matrix which may be permuted into the blocked form $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$. This convergence property may be used as the basis of an algorithm (CONCOR) for hierarchical clustering. The CONCOR procedure is applied to several illustrative sets of social network data and is found to give results which are highly compatible with analyses and interpretations of the same data using the blockmodel approach of White (in press). The results using CONCOR are then compared with results obtained using alternative methods of clustering and scaling (MDSCAL, INDSCAL, HICLUS, ADCLUS) on the same data sets.

AN ALGORITHM FOR BLOCKING RELATIONAL DATA, WITH APPLICATIONS TO SOCIAL NETWORK ANALYSIS AND COMPARISON WITH MULTIDIMENSIONAL SCALING

Ronald L. Breiger
Harvard University, Cambridge, Massachusetts 02138

Scott A. Boorman
Harvard University, Cambridge, Massachusetts 02138

Phipps Arabie
Stanford University, Stanford, California 94305

The first part of this paper describes an efficient algorithm for simultaneous clustering of one or more matrices and develops applications to sociometric and other social structural data.[1] Although the approach was originally motivated by applications to strictly binary network data, the algorithm can also be applied to matrices reporting data in integer or continuous form (e.g., application to Sampson's monastery data, pp. 32-38 below). The procedure hence represents a technique of considerable generality and gives promise of unifying a wide range of data analyses. From a formal point of view, the output of the algorithm may be represented as a hierarchical clustering (see Figs. 4 and 9 below). Unlike standard hierarchical clustering methods, however, the input to the present algorithm is not necessarily a proximity or a distance matrix, but rather one or more matrices representing arbitrary kinds of relationship (see pp. 12 ff. below).

The second part of the paper compares the results of the main algorithm to those of multidimensional scaling algorithms applied to some of the same data (the MDSCAL algorithm of Shepard [1962a, b] and Kruskal [1964a, b] and the INDSCAL algorithm of Carroll and Chang [1970]). This second part also reports exploratory sociometric applications of a recent nonhierarchical clustering algorithm of Arabie and Shepard (1973).

PART I. DESCRIPTION AND APPLICATION OF THE ALGORITHM

When several different disciplines encounter similar problems in research, it often happens that investigators in different areas make parallel and independent discoveries, with considerable duplication of effort entailed. Developments in hierarchical clustering constitute a prime example of such an occurrence. For instance, the most commonly used methods of hierarchical clustering in psychological research are those of Johnson (1967). However, as he pointed out, both of his methods had already been independently discovered. Specifically, the connectedness method (see p. 31 for details) had been described by Sneath (1957), and the diameter method by Sorenson (1948). Yet until Johnson's (1967) paper appeared, hierarchical clustering was virtually unheard of by psychologists doing research in areas other than test theory.

As in the case of Johnson's methods, the algorithm we present here represents an independent discovery of a method published earlier (McQuitty, 1967; McQuitty and Clark, 1968). McQuitty's work, although directed toward psychometricians, has received little attention from psychologists and none at all from sociologists, probably because most of his illustrative applications have used artificial data having limited interest. Since we have found that this method of hierarchical clustering can yield very meaningful results when applied to data with which sociologists and social psychologists are already familiar, we are presenting the present algorithm in a context that is quite different from McQuitty's.

We begin by describing the basic algorithm (acronym: CONCOR) as it applies to partitioning the vertices of a graph into similarity classes ("blocks"). There are direct generalizations to handling the simultaneous blocking of multigraph data, i.e., data which report more than one distinct kind of relation on the same population (pp. 16-17 below). As will later become apparent from the Bank Wiring Group and other applications, the ability to handle such multiple tie data is important for many analyses of concrete social structures.

Expressed in sociometric language, the basis of the procedure consists of systematically grouping together actors in a network who occupy similar positions with respect to ties sent, ties received, or both. The method is a rapidly convergent algorithm which (aside from exceptional cases of largely mathematical interest) will produce a bipartition of the set of actors (i.e., a partition into exactly two equivalence classes). As described below (Section 2D) this algorithm can be applied repeatedly at the discretion of the investigator to produce a partition of the actors with any desired degree of fineness.

The algorithm is applied to a number of illustrative data sets. These include: (1) social network data of sociometric or observed-reported type; (2) participation data on women in a Southern city; and (3) data on directorship interlocks among seventy large corporations and fourteen banks (originally studied by Levine [1972] from a multidimensional scaling standpoint). All data sets studied involve comparatively small populations (< 100), though there are no basic theoretical reasons for presence of this limitation. Emphasis will be placed on the interpretability of the obtained partitions in the light of the original relational data, as well as on connections between the present method and other methods of analysis applied by previous investigators of the same data. (See also Part II, where specific comparisons with multidimensional scaling are developed at length.)

1. Structural Equivalence and Blockmodels

This section provides substantive background which will motivate development of the algorithm and its applications. The reader who wishes immediate exposure to details may turn directly to Section 2. However, the ideas discussed below, in particular the zeroblock concept, have direct bearing on later data applications and will there be quoted freely.

Motivated by ideas of classical theorists such as Nadel (1957), White and collaborators have undertaken the development of formal theories which place great emphasis on the concept of "structural equivalence" in the description of concrete social structures (White, 1963, 1969; Lorrain and White, 1971; Bernard, 1971; Breiger, 1974). The structural equivalence concept is grounded in the network metaphors of theorists such as Simmel (1955):

> ... as the development of society progresses, each individual establishes for himself contacts with persons who stand outside [his] original group-affiliation, but who are

'related' to him by virtue of an actual similarity of talents, inclinations, activities, and so on. The association of persons because of external coexistence is more and more superseded by association in accordance with internal relationships. ...practical considerations bind together like individuals, who are otherwise affiliated with quite alien and unrelated groups.

(Compare also the work of von Wiese, who was strongly influenced by Simmel; e.g., von Wiese [1941: 29-30]).

Structural equivalence in White's work is seen as a unifying concept cross-cutting theories of roles, kinship, sociometry, and organization, where it repeatedly appears in many different guises and on various different levels of analysis. Although the concept of structural equivalence has been used in a number of distinct ways (Lorrain and White, 1971; Fararo, 1973), all these cited developments have adhered to a highly algebraic--and correspondingly rigid--concept of what structural equivalence should formally mean. Specifically, in any network (possibly involving multiple binary relations), the formal definition of structural equivalence is as follows:

<u>Definition 1</u>. Let S be a set and let $\left\{R_i\right\}_{i=1}^{m}$ be a set of binary relations on S, i.e., a set of subsets of SxS. Then individuals a, $b \in S$ are <u>structurally equivalent</u> with respect to the (multiple) network defined by $\left\{R_i\right\}_{i=1}^{m}$ if and only if the following criterion is satisfied: for any $c \in S$ and any relation R_i,

(1) $aR_i c \Longleftrightarrow bR_i c$

(2) $cR_i a \Longleftrightarrow cR_i b$.

This is a direct transcription of an equivalence ("indiscernibility") concept familiar in model theory within mathematical logic (e.g., Robinson, 1965; Schoenfield, 1967). However, it is immediately clear that if the above definition is applied directly to raw data, irregularities in real social structures of any size will allow very few instances of structural equivalence to be present. Hence, without some crucial weakening or idealization, the equivalence concept as given is essentially vacuous.

The route followed in the original work of White (cited above) centered around performing homomorphisms on algebraic structures (e.g., semigroups) generated by raw data matrices, and then employing such homomorphisms to induce various equivalence patterns. Specifically, the aim is to achieve structural equivalence in a reduced network, i.e., in the image network obtained from raw data when this data is subjected to a "functorial mapping" (essentially a generalized homomorphism). There is no need to describe here the detailed mechanics of performing such a homomorphism (see Lorrain and White, 1971; see also Fararo, 1973), but the crucial point is that the image network under homomorphism will typically be a much fatter network than the original one, i.e., a network with a much higher density of ties. For this reason alone it is not surprising that structural equivalence among actors will eventually emerge as a chain of successive homomorphic reductions is applied.

Taken in conjunction with a more broadly based rationale for the homomorphism concept (developed at length in White, 1969), this homomorphism strategy has been effective for giving insight into the "skeletal" structure of some varieties of complex social networks (see examples developed at length in Lorrain, in press). However, the approach has the crucial feature that it comes in a package: in order to achieve structural equivalence at the level of individual actors, one must make a long detour through complicated algebraic procedures involving powerful and highly restrictive assumptions on the treatment of compound ties (indirect social relationships).

In response to these limitations, White has subsequently developed a second, and distinct, approach to modeling social network data and finding structural equivalence patterns. This second approach is the point of departure for the development of the present algorithm and we will therefore discuss it in some detail. In later sections, the relation between White's own analyses and the results of the present algorithm will frequently be cited.

This second line of attack (White, 1973, 1974a, b; White and Breiger, 1974) centers around the concept of a "blockmodel." This is a very simple and natural combinatorial idea; unlike the homomorphism analyses it involves only minimal formal developments. For illustrative purposes, consider first the (imaginary) data of Fig. 1a. For simplicity, this example involves only one kind of reported tie; generalizations to multiple types of ties will be deferred until later, in the context of real data (e.g., Fig. 6).

The (i,j)th entry in the Fig. 1a matrix reports the presence ("1") or absence ("0") of a network tie from individual i to individual j. Both rows and columns are hence to be thought of as indexing the same population of individuals in some given order which is the same for both rows and columns. Otherwise, however, the row (respectively, column)

Fig. 1. Imaginary data illustrating blockmodels, lean fit, and zeroblocks.

(a)

$$
\begin{array}{r|cccccccccc}
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
2 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
3 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
4 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
5 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
6 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
7 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
8 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\
9 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
10 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
$$

(b)

$$
\begin{array}{r ccc|cccc|ccc}
2 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
7 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
8 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
\hline
3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
6 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
10 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
\hline
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
9 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{array}
$$

(c)

$$
\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}
$$

ordering as presented is quite arbitrary. In this obvious comment lies the source of blockmodel developments. By imposing the same permutation on both rows and columns, one may be able to discover a new way of presenting the data which is more interpretable (see Fig. 1b, which reports a permutation of the Fig. 1a matrix). The aim of this rearrangement may be made more definite by specifying that what is being sought is a permutation which reveals substantial submatrices all of whose entries are zero (White's term for such matrices is <u>zeroblocks</u>; under the superimposed division, Fig. 1b contains three zeroblocks). Finally, then, one may give a summary description of the data by means of a <u>blockmodel</u> (Fig. 1c), where a "0" in the model corresponds to a zeroblock in the data matrix while a "1" in the model corresponds to a block in the data matrix which contains at least some 1's.

What is being developed here is a new kind of generalization of the concept of structural equivalence, where now one is treating individuals in the same block as equivalent. [A minor terminological ambiguity arises here, since the term "block" will be used to denote both a set of actors and also a submatrix of a blocked matrix (see again Fig. 1b). Context should always make the intent transparent.] There is no longer any question of "massaging" the original network data, as in the case of the algebraic developments using homomorphisms. <u>The same data are retained and instead it is the equivalence concept itself which is weakened.</u>

The formal idea may be made more precise by starting from a given blockmodel (e.g., that in Fig. 1c) and making the following definition:

<u>Definition 2</u>. A blockmodel is a <u>lean fit</u> to a given data matrix M if

and only if there exists a permutation of M, leading to a permuted matrix M*, together with a subdivision of M*, such that:

(1) Zeroblocks in the permuted matrix correspond to 0's in the blockmodel;

(2) Blocks containing <u>some</u> 1's in the permuted matrix correspond to 1's in the blockmodel (compare again Figs. 1b and 1c).

There is a basic asymmetry here: zeroblocks are expected to contain <u>only</u> 0's, whereas 1-blocks merely need to contain <u>some</u> 1's. This is the sense in which the fit is said to be "lean" instead of "fat." Note that if the fit were indeed fat, so that all 1-blocks were completely filled with 1's, then individuals in the same block would be structurally equivalent in the original algebraic sense (p. 5 above).

The particular weakening of Definition 1 which the lean fit concept represents is a highly natural one in a wide variety of social network applications. Presence of an active tie often requires a clear effort on the part of one or both individuals concerned, whereas absence of a tie does not in general require work. In a tightly knit social structure it may be much easier to avoid maintaining a tie than to preserve an active "maverick" tie. Moreover, any kind of data collection procedure where reporting a tie depends on some kind of threshold cutoff criterion (as in forced-choice sociometric procedures) may also act to create gaps in 1-blocks. In contrast, such cutoff effects will not produce the opposite kind of error, that of reporting existence of an active tie where none in fact exists. On these and similar grounds (discussed more extensively in White, 1973, 1974b) it is unlikely that 1-blocks will be fat. From a purely formal standpoint, quite aside from substantive issues, one would expect the lean-fit criterion to be relevant as a criterion for clustering many varieties of sparse matrices.[2]

It is clear that blocks (in the lean-fit blockmodel sense just defined) need not be cliques in the standard graph-theoretic sense or any of its many sociometric generalizations (Luce, 1950; Hubbell, 1965; Alba, 1973, etc.) There is no implication that the members of a block cooperate or coordinate with one another. In fact, the individuals in a block need not be connected at all to one another (see again the third block in Fig. 1b), and in fact this absence of connection would not be at all surprising if the members of a block were "hangers-on" to some "leading crowd" and the relation being coded was something like "deference" (see also the interpretation of the Bank Wiring data, p. 27 below). This point stresses very forcefully that the criterion for lumping individuals into the same block is a consistency idea, not a connectivity idea: blocks are defined on the criterion that their members should relate consistently to other blocks, in the specific sense made precise by the lean-fit concept. In the present context, the emphasis on consistency implies in particular that in principle the <u>whole</u> social structure must be simultaneously taken into account in order to test <u>any</u> nontrivial blockmodel description for lean fit.

Practical development of blockmodel analyses now centers around the following problems. (1) Given a blockmodel (as in Fig. 1c), together with raw data (as in Fig. 1a), there is the problem of enumerating all (if any) concrete blockings of the data (e.g., Fig. 1b) which fit the blockmodel in the lean-fit sense. (2) Given only raw data, there is the problem of finding some lean-fit blockmodel for the data which involves a reasonably small and interpretable set of blocks. Next, since what one is interested in is a summary description of a complex structure, one may also (3) weaken the strict lean-fit criterion of Definition 2 and proceed otherwise along the lines of (2). Finally, given any particular blockmodel and data arranged to fit this model, there is the problem (4) of assessing how convincing is the obtained fit, presumably in a statistical significance sense relative to some null hypothesis. Other things being equal, it is clear that a lean-fit of a seven-block model on a population of fifteen people is less likely to be impressive than is a fit of a three-block model to the same population. Taking the extreme case, each population of size n is a blockmodel of itself, with n blocks, and here blockmodels clearly add nothing.

In this outline of the problems to be solved, the algorithm we will describe below should be placed under (3). Specifically, the algorithm is a way of directly starting from raw data and obtaining a partitioning into clusters (actually, a hierarchical clustering). These obtained clusters typically do not bring out strict zeroblock structure in the data, as for example does the blocking in Fig. 1b. Nevertheless, extensive tests on data indicate that the results of the present algorithm are usually close to the most informative lean-fit blockmodels which have been found through trial-and-error methods (thus see below, p. 26). From the standpoint of White's work, therefore, the present algorithm may be interpreted as a search procedure for lean-fit blockmodels as characterized in Definition 2. The relation between the present algorithm and lean-fit models will be pursued at greater length in later discussion of data applications.

To avoid terminological confusion, we will observe the following conventions. When a blockmodel is spoken of as "fitting" given data, the default interpretation is that the fit is close to a perfect lean fit in the Definition 2 sense, but with some imperfections allowed (impure zeroblocks). We will always speak explicitly of the _strict_ lean fit criterion if the lean fit is perfect. Contrary to _a priori_ intuitions about the likelihood of imperfections in real social structures, it is surprising how often true zeroblocks are actually found (see, e.g., Fig. 6).

2. The Convergence-of-Iterated-Correlations (CONCOR) Algorithm

Consider an $n \times m$ real matrix M_0. An example could be a sociomatrix representing network ties; other examples will be encountered later. Treating the columns (alternatively, the rows) as separate vectors $\underset{\sim}{v}_i$, $i=1,2,\ldots,m$ (respectively $i=1,2,\ldots,n$), form the $m \times m$ (respectively $n \times n$) matrix M_1 whose (i,j)th entry is the standard product moment correlation coefficient between $\underset{\sim}{v}_i$ and $\underset{\sim}{v}_j$.[3] (M_1 will henceforth be referred to as the _first-correlation matrix._) Now apply the same procedure to M_1 and iterate, obtaining successively matrices M_2, M_3, etc., all of which will be square matrices of the same size as M_1.

Then the following mathematical statements appear to hold generally, aside from exceptional cases of "knife-edge" character: (1) $M_\infty = \lim_i M_i$ always exists; and (2) M_∞ is a matrix which may be blocked in the following bipartite form:

$$\begin{bmatrix} \boxed{1} & \boxed{-1} \\ \boxed{-1} & \boxed{1} \end{bmatrix} \quad \text{(2-BLOCK)}$$

These two assertions will not be investigated matematically here;[4] for the present, it is sufficient that both (1) and (2) have been empirically verified to hold on more than one hundred applications to sets and subsets of network-related data ranging in size up to 70 x 70 (for the initial matrix M_0). No exception to statement (2) has been found in any application to data. In the simplest case where M is a single binary matrix, representing a sociogram, the fact that the limit matrix M_∞ can be blocked as indicated above may be restated to say that the iteration procedure is an algorithm which splits into two parts the set of actors in the network. More general situations will also be encountered in the later applications, but the use of the algorithm is always to produce a bipartition of a concrete population. The algorithm will be designated CONCOR ("convergence-of-iterated-correlations").

Concerning the application of the CONCOR algorithm, the following initial observations should be made.

A. <u>Counterexamples to the limit (2-BLOCK)</u>. There are a number of obvious counterexamples to the statements (1) and (2), i.e., cases where M_∞ does not exist or cannot be blocked in a bipartite form. However, if M_0 is perturbed slightly away from such a degenerate case, then convergence to the bipartite limit (2-BLOCK) will in general be restored. This indicates that the exceptions to the statements (1) and (2) above form a class of purely mathematical interest.[5] However, one particular case of degeneracy should be noted to arise if (in the case of iterated column correlations) some column of the original matrix M_0 has only 1's or only 0's (and dually for rows in the case of iterated row correlations). Then the column vector in question has zero variance, and hence the product moment correlations involving this column will be all undefined. In sociometric terms, this difficulty will occur when some individual is either chosen by everyone or chosen by no one. The former difficulty may be avoided by the technical expedient of imposing a zero diagonal on M_0 (no one is considered to "choose" himself). The second problem, that of an individual who is chosen by no one, is reminiscent of the degeneracies arising in multidimensional scaling when one or more points in the input distance matrix is very far from all other points (Shepard, 1962b; Arabie and Boorman, 1973; Table V).[6]

B. <u>Speed of convergence</u>. Approach of M_i to M_∞ is typically rapid. Define a cutoff criterion to be a parameter $c < 1$ such that the algorithm is terminated as soon as a matrix M_n is reached each of whose entries has an absolute value $\geqslant c$. The examples quoted in the applications below were for the most part constructed with a cutoff value of $c = 0.999$. In all examples of section 4 except the last one, Example E (the Levine data, where the population of corporations was of size 70), the 0.999 cutoff was reached in eleven or fewer iterations. In the case of the Levine corporate interlock data, a cutoff of $c = 0.9$ was reached after 12 iterations.

C. <u>Blocking on rows versus columns of a sociomatrix</u>. If the original matrix M_0 describes a network formed by forced-choice sociometric procedure (Bjerstedt, 1956), then the nature of the data collection procedure introduces an <u>a priori</u> asymmetry into the status of rows and columns. Specifically, as Holland and Leinhardt have stressed in a number of papers (e.g., 1969, 1970), forced-choice procedures have the effect of constraining row marginals and this constraint may have the effect of masking existing network structure. Holland and Leinhardt deal only with triad counts, but their concern also applies to the present situation and suggests that in such specific cases one should give preference to blockings based on column correlations rather than those based on row correlations. All sociometric and observer-reported applications of CONCOR presented in this paper (pp. 22 ff. below) are based on column, rather than on row, correlations. In many cases, row correlations have additionally been run. The results are typically close, though not in general identical, to those obtained using column correlations; the results of comparing row and column approaches will be reported elsewhere.

It is also possible to mix row- and column-correlation approaches in the same limiting process, as when successive iterations are alternately based on row and column correlations (an alternation procedure reminiscent of Mosteller row-column marginals equalization; see also p. 40 below).

D. Repetitions of the algorithm on successive subpopulations. The procedure just described may be separately repeated on each of the two obtained blocks. Specifically, one may repeat the procedure on each of the two submatrices formed by taking the columns of M_0 corresponding to each of the two blocks delineated by the previous bipartition. A new M_1 is then formed from each submatrix and the limit M_∞ is obtained. This repetition will lead to a new bipartite split of each of the original blocks in turn, leading to a finer overall partition with four blocks in all. Notice that although M_1 is a proximity matrix, the information contained in M_1 alone is insufficient for computing finer blockings; one must return to M_0 in each case. Repeating the algorithm on each of these finer blocks, we may obtain blockings to any desired degree of fineness, and thus the CONCOR algorithm leads to an algorithm for hierarchical clustering.

E. Multiple types of relation. Instead of data consisting of a single network, assume next that one is given a network where a number of distinct kinds of relations are reported. Specifically, assume that one starts with k $n \times n$ data-matrices, each reporting the incidence of a particular type of tie on an underlying population of size n (e.g., "Liking," "Helping," "Antagonism," etc.). The k matrices may be compounded into a single new matrix with nk rows and n columns, in which the individual data matrices are "stacked" one above the other in an arbitrary order but preserving the same column ordering for each matrix. (Alternatively, a $2nk \times n$ array including each matrix and its transpose may be formed.) An $n \times n$ first-correlation matrix M_1 may then be formed as usual and the CONCOR algorithm again applied as before. Note that the procedure as described implicitly gives equal weight to each component type of tie, and in particular makes no attempt to weight ties differentially according to the frequency of their incidence or other measures of comparative importance. Various natural refinements may be developed which respond to these difficulties by incorporating differential tie weights (compare the use of weighted Hamming metrics by Kruskal and Hart [1966]). However, only the simple unweighted procedure just sketched will be used in the exploratory applications below.

The ease with which the CONCOR method may be extended to handle multiple types of tie is a very important feature of the approach, and makes it a natural clustering method for many types of social network and other social structural data. In fact, there are few substantive contexts where it can be convincingly argued that only one kind of social relation is present, rather than multiple networks simultaneously existing in a population. Many characteristic aspects of concrete social structures in fact arise from the presence of multiple types of differentiated tie (see White, 1963: Chapter 1 for examples drawn from kinship and formal organizations). In many studies, empirical data collection procedures eliminate all but one type of tie, or use ad hoc aggregation procedures to reduce several distinct types of tie into a single type prior to the main analysis. The existence of CONCOR as a simple method which is able to handle a large number of types of tie as easily as one type may encourage empirical investigators to collect and report data on multiple distinct kinds of social networks.[7]

3. <u>Relation of the CONCOR algorithm to traditional aspects of clustering and scaling.</u>

Since the method of clustering introduced here is quite different from most methods encountered in the behavioral and biological literature, it is useful to relate CONCOR to the established framework of cluster analysis. In describing CONCOR as a hierarchical clustering algorithm, we should first emphasize that the phrase "hierarchical clustering" is here being carried over from the tradition of data analysis in psychology. There is no implication that CONCOR is a procedure specifically designed to extract status orderings or other social hierarchies from social network data, nor that such hierarchies will in fact be obtained in the applications below (contrast, for example approaches in Bernard [1973, 1974], where hierarchical structure in sociometric data is specifically sought and analysed).

A. Invariance properties. It is clear that the output from CONCOR is not in general invariant under arbitrary monotone transformations of M_0, considered as a matrix of real numbers. In the standard clustering literature, this absence of invariance is consistent with the metric approach of Ward (1963) rather than with the nonmetric approach of Johnson (1967). However, in the context of the present algorithm, the question of ordinal invariance does not have the same significance as is in the case of other methods, since in dealing with sociomatrices we are not viewing the input data M_0 as a distance or similarity matrix (cf. p.12 above, and compare Needham [1965:118] and Hartigan [1972:124-127]). The fact that M_0 need not be a distance matrix allows us to deal directly with binary matrices which cannot serve as direct input to commonly employed methods of clustering based on distance concepts.

However, from a formal standpoint, it is worth noting that the CONCOR algorithm does give results invariant under any transformation of $M_0=[m_{ij}]$ which takes m_{ij} to $a\ m_{ij}+b$.

B. The position of CONCOR in taxonomies of data and data analysis. In terms of Shepard's (1972:27-28) taxonomy for types of data and methods of analysis, we are of course dealing with profile data as soon as M_1, the first-correlation matrix, is computed. However, the fact that CONCOR is in many ways omnivorous with respect to M_0 (an n x m matrix) allows the algorithm to fall under several traditional headings simultaneously.

For example, the fact that M_0 need not be a square matrix allows the rows to correspond to entities completely different from the columns. Thus, in particular, we can deal with data which are appropriate to analysis by multidimensional unfolding (see, for example, the Levine data in Section 4E below). The possibility of clustering both the rows and the columns of a non-square M_0 makes this particular use of CONCOR quite similar in emphasis to Hartigan's (1972) method of "direct clustering" (see also MacRae, 1960).

In the useful terminology of Carroll and Chang (1970), the application of CONCOR to multiple types of relation constitutes 2-way scaling, since the result of forming M_1 on the stacked raw matrices is to study "subjects by subjects." We began with a 3-way data structure (the k distinct relations constituting the third level), but by stacking we reduced the problem to a 2-way analysis. This reduction in the complexity of the design is similar in intent to many applications of the more familiar 2-way procedures where one sums over conditions to obtain a group matrix (or sums squares, if one thinks that the raw data are actually distances [Horan, 1969]). An example of this standard approach is given by Shepard's (1972) reduction of the Miller-Nicely (1955) 3-way data on confusions between consonant phonemes, in order to convert the data into a form where they may be entered as an input to MDSCAL, which is inherently a 2-way procedure.

C. Relation to alternative methods of hierarchical clustering. We will not attempt in this paper to review or classify the many clustering algorithms presently available; the interested reader should consult Lance and Williams (1967a, b) and Jardine and Sibson (1971). However, we do wish to comment on the position of CONCOR with respect to some of the more well-known aspects of clustering procedures.

To begin with, the present algorithm is obviously "divisive," in contrast to the more commonly used "agglomerative" procedures (terminology of Lance and Williams, 1967a) which begin forming clusters by joining together single stimuli and then later merging clusters to obtain a tree structure.

Reflecting a commonly adopted standpoint, Jardine and Sibson (1971) suggest a basis for classifying clustering procedures, which would distinguish among procedures according to where they fall on a continuum whose extremes are respectively the connectedness and diameter methods of Johnson (1967) (see below, p. 31). The question naturally arises: where does CONCOR fall along such an axis?

We investigate this question in an Appendix. Specifically, the analysis there given employs one of the Boorman-Olivier tree metrics to quantify the similarity between CONCOR and Johnson's HICLUS solutions for two of the concrete data sets analyzed in Section 4 (the Bank Wiring Group data and the Sampson monastery data). The evidence derived from this analysis suggests no preferred position for CONCOR, and the Jardine-Sibson classification hence appears essentially irrelevant to the present approach.

Turning to a different set of problems, a common feature of many otherwise disparate clustering procedures is that they perform inadequately or unsatisfactorily when confronted with certain practical problems arising frequently in data analysis. Two such situations arise most frequently. These situations concern: (1) treatment of ties and (2) presence of an excessive number of levels for interpretation in the (output) hierarchical structure.

The presence of ties constitutes a real problem for clustering procedures which are based on a sequential pattern of merging/splitting. As Hubert (1973:48) observes, it is usually assumed that ties will not occur. If they do occur, some arbitrary decision must be made. In sharp contrast, ties in the raw data matrix M_0 do not in any way constitute a distinctive case for the CONCOR algorithm, which appears to deal very effectively with binary matrices--a rather extreme case of tie-bound data (see examples below in Section 4). The obvious reason is the fact that CONCOR passes immediately to the first correlation matrix M_1, and ties in M_0 will not in general be inherited as ties in M_1.

The experienced user of hierarchical clustering methods is well aware of the differences between the computer output from such methods and the published figures that subsequently appear. The chief discrepancy arises from the fact that most hierarchical methods yield n levels (where n is the number of stimuli) in the tree structure--far too many for either interpretability or ease of graphic presentation. The user is hence confronted with the task of collapsing over certain levels. The decision as to which levels are to be ignored is usually a rather subjective one, as there are no well defined criteria available for most hierarchical clustering methods. Of course, for situations in which a fine level of partitioning is ultimately or locally required, CONCOR is no different from the other hierarchical methods with respect to this particular problem: the user can continue applying CONCOR on a given matrix to reach any desired level of fineness.

D. What the CONCOR algorithm maximizes. Unlike some other hierarchical clustering schemes (e.g., Ward, 1963; Edwards and Cavalli-Sforza, 1963, 1965, Hubert, 1973), the CONCOR algorithm is not cast in the form of a solution of some maximum or minimum problem. However, there is numerical evidence that the performance is close to that of an algorithm designed to take the first-correlation matrix M_1 and to split the underlying population into two groups so as to maximize mean within-group correlations. For example, when M_1 for the Sampson data (Fig. 8) is rearranged in accordance with the two-block CONCOR model, $M_1 = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$, the mean correlation with submatrices A and D is .232 (excluding the diagonal entries of M_1, which are 1 by definition) and the mean correlation in submatrices B and C is -.098. This contrast is marginally sharper than for White's (in press) two-block trial-and-error model on the same data (which leads to the analogous correlations .185 and -.087 respectively).

4. Applications to the Analysis of Social Networks

We discuss five applications to sociometric, observer-reported, participation, and interlocking-directorate data.

A. Newcomb's Fraternity

Theodore Newcomb (1961; see also Nordlie, 1958) created a fraternity composed of seventeen mutually unacquainted undergraduate transfer students. In return for free room and board, each student supplied data over a four-month period, including a full sociometric rank-ordering each week, listing the sixteen other students according to his "favorableness of feeling" toward each. The experiment was repeated with different subjects in two successive years.

A small part of Newcomb's data (rankings for Year 2, Week 15) will serve as a first illustration of a two-block model produced by the CONCOR algorithm. Week 15 is the final week of the Year 2 experiment, and from looking at the Year 2 data as a whole it is clear that the preference rankings have reached what is roughly an equilibrium configuration by about Week 4 or 5 and have remained there since (see also Part II, which reanalyzes the full Year 2 data using INDSCAL.)

Specifically, form two binary matrices from the original rank-ordered data for the given week. The first matrix l ("most favorable feeling") is taken to contain a "1" for each of the top two choices of each student, with 0's elsewhere; the second matrix a ("least favorable feeling") is taken to contain a "1" for the bottom three choices of each student, with 0's elsewhere. In a simple way, these two matrices extract two extremes of sentiment out of the raw rank-orderings. The particular decision to take the top two and bottom three choices follows White (1974b); from exploring numerous alternatives it can be asserted that the blocking outcome will be robust over alternative ways of converting the data to binary form. In particular, the same analysis has been run taking top three and bottom three choices, with no essential difference in results.

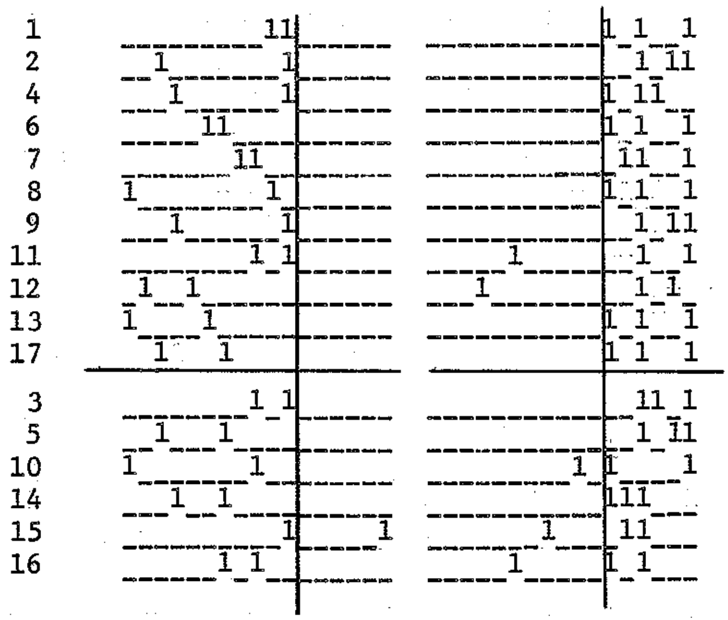
Given the binary matrices l and a, a 34 x 17 matrix $M_0 = \left(\frac{l}{a}\right)$ was now formed by stacking l over a. The 17 x 17 first-correlation matrix M_1 was now computed from the columns of M_0, and the CONCOR algorithm was applied to obtain M_∞. The bipartite blocking implied by M_∞ led to the blocks (1,2,4,6,7,8,9,11,12,13,17) and (3,5,10,14,15,16) (following Nordlie's [1958] numbering of subjects).

This blocking is identical to that obtained by White through trial-and-error and following (but not adhering strictly to) the lean-fit criterion (White, 1974b).[8]

Figure 2 now illustrates the obtained blocking on the present top- and bottom-choice matrices Z and α. It is clear that the pattern is close to a lean fit to the two-block two-relation blockmodel $H = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$, $T = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$,[9] though perfect lean fit is ruled out by a scattering of 1's in the low-density blocks. As a first way of developing a quantitative approach to blockmodel fit, beneath each blocked matrix in Fig. 2 is a table of the densities in each of the four blocks (i.e., the number of ties divided by the number of entries in the block and excluding cells which fall on the main diagonal). Note that there is a clear bimodality in density as between the low-density blocks (densities = 0, .02, .03, .05) and the high-density blocks (densities = .17, .20, .47, .50).

The blockmodel structure thus revealed is interpretable in a very simple way. One of the blocks contains persons (1) none of whom send top choices outside the block, and (2) who receive virtually all the top choices of the second block, and (3) who send virtually all their bottom choices to the second-block individuals. At the same time, the second block not only receives virtually all bottom choices from the other block, but also absorbs virtually all the bottom choices of its <u>own</u> members. This structure suggests a situation where there is a single, dominant central clique and a second population of "hangers-on."

Fig. 2. Two-block model for Newcomb fraternity data. Year 2, Week 15, rank-order data converted to binary form by taking top two and bottom three choices (see text).



.20	0
.17	.03

.02	.47
.05	.50

B. The Bank Wiring Room

The second application of the algorithm will concern an example of Homans (1950). This example is drawn from a classic study (Roethlisberger and Dickson, 1939) of a Western Electric production team transferred to a special room which an observer shared for six months. Rather than asking the men themselves for a statement of their relationships (as in the sociometric studies reviewed here), the original researchers inferred the incidence of six types of tie among the fourteen men (see Homans [1950:64-72] for a detailed description of each type of tie). The ties have no time referent and are thought of as stable.

The specific types of tie reported are as follows (see also Fig. 6 for the incidence of all ties except the Trading one): "Liking;" "Playing games with" (described as "Games" in Fig. 6; see Homans [1950: 68]); "Antagonism;" "Helping;" "Arguments about Windows with" (see Homans [1950:71]); and "Trading Jobs with." For the most part, these relational descriptions should be self-explanatory. "Liking," "Antagonism," and "Arguments about Windows" were all coded as symmetric ties. "Playing games with" was generically a positive sentiment tie, while "Arguments" was a particular kind of negative sentiment tie (Roethlisberger and Dickson, 1939:502-504). Each type of tie may be represented by a 14 x 14 matrix reporting its incidence on the fourteen-man population.

Our analysis excludes the highly specialized (and low-incidence) type of tie, "Trading Jobs," as we wish to achieve comparability of our results with White's seven-block model[10] formed on the remaining five relations. A 70 x 14 matrix M_0 was formed by vertically "stacking" the remaining five 14 x 14 matrices, taking care to preserve the ordering of columns. On the first iteration, a 14 x 14 column-correlation matrix M_1 was then formed (Fig. 3). Applying CONCOR, M_∞ yielded the bipartition: (W1,W2,W3,S1,W4,W5,I1,I3), (W6,S2,W7,W8,W9,S4). (This notation follows Homans' convention of numbering men within their job classification: W for wiremen, S for soldermen, I for inspectors.)

In order to obtain a finer blocking, the above process was repeated for each of these blocks in turn. (E.g., the next step was to form a 70 x 8 matrix composed of the columns corresponding to W1,W2,W3, S1,W4,W5,I1, and I3 of M_0 and to apply CONCOR with this new submatrix as M_0.) Eventually, nine blocks were found in this manner. In accordance with one standard way of representing hierarchical clusterings, a natural way of displaying the results of this repeated process is by a binary tree (Fig. 4). Each node in this tree represents a cluster (block) containing all men positioned below it.

Figure 5 indicates the similarity between Homans' analysis (which agrees in essentials with that of Roethlisberger and Dickson), the seven-block model in White (1974b), and our own findings using the present algorithm. Our two-block model essentially identifies Homans' two cliques, though also mixing in individuals whom Homans considers as outsiders. Our four-block model very nicely distinguishes the Homans cliques (Blocks 1 and 4) from their marginal members and outsiders (Blocks 2 and 3). This four-block model and White's seven-block model are compatible, i.e., the latter is a partition which is a refinement of the former.

Now return to the five data-matrices and impose our four-block

Fig. 3. First-correlation matrix M_1 formed on the Bank Wiring data by correlating columns of M_0 (described in text).

	W1	W2	W3	S1	W4	W5	W6	S2	W7	W8	W9	S4	I1	I3
W1	1.0													
W2	.30	1.0												
W3	.58	.18	1.0											
S1	.34	.05	.35	1.0										
W4	.46	.17	.38	.56	1.0									
W5	.07	.46	-.04	.01	.03	1.0								
W6	-.12	-.12	-.20	.09	.03	.11	1.0							
S2	-.05	-.05	-.06	.21	.22	-.07	.22	1.0						
W7	-.08	-.26	-.10	-.08	-.03	-.04	.33	.19	1.0					
W8	-.23	-.23	-.22	.07	.09	.01	.33	.21	.45	1.0				
W9	-.24	-.24	-.15	.05	-.09	.07	.38	.20	.50	.58	1.0			
S4	-.19	-.19	-.24	-.08	-.07	.11	.38	-.55	.30	.36	.43	1.0		
I1	.41	.27	.17	.37	.27	.27	-.07	-.04	.00	.03	.02	-.03	1.0	
I3	-.14	.41	-.18	-.08	-.07	.27	.04	.36	.00	-.08	-.09	-.15	-.11	1.0

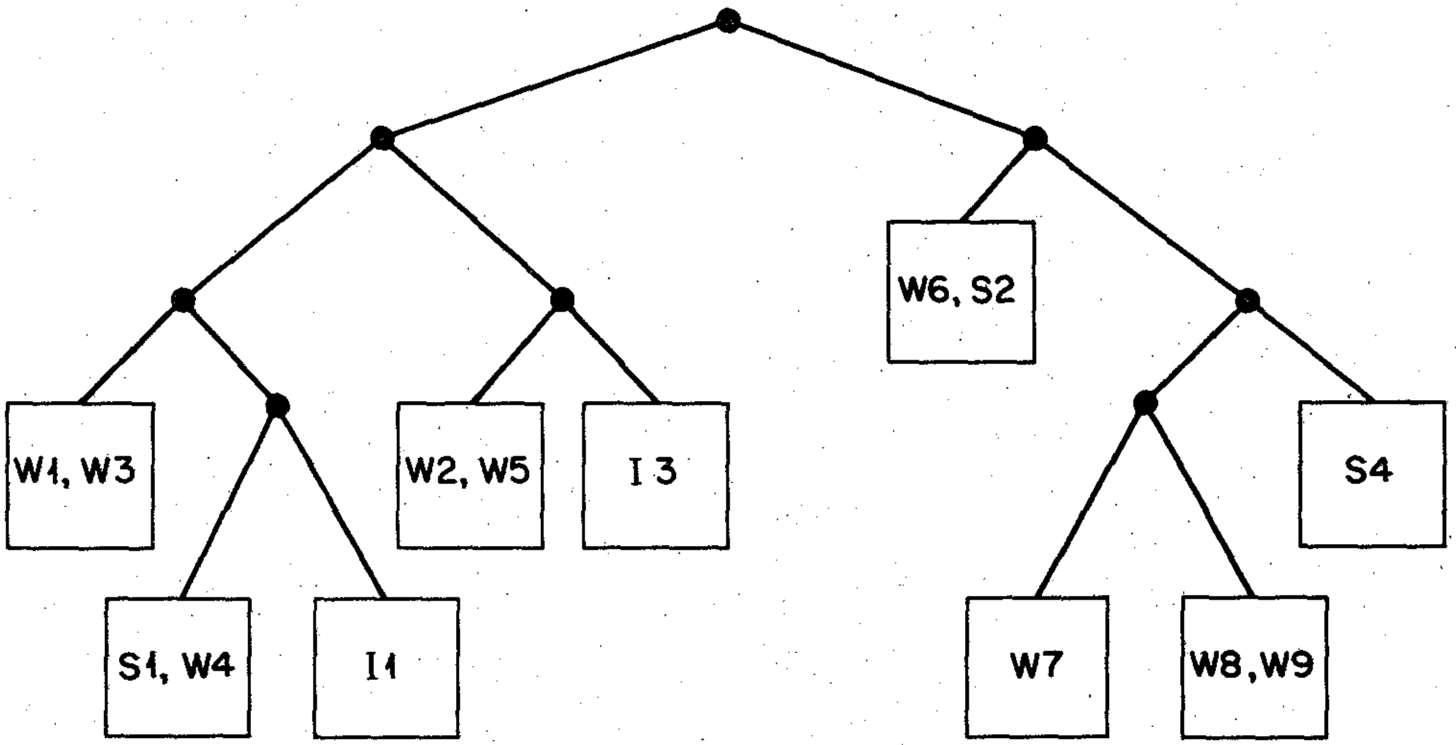


Fig. 4. Hierarchical clustering representation of the repeated application of the CONCOR algorithm on the Bank Wiring data.

Fig. 5. Comparison of the CONCOR results reported in Fig. 4 with the trial-and-error blockmodel analysis of White (1974b) and the discussion in Homans (1950).

Individual's identification	Homans' assignment+	(CONCOR algorithm) 2-block model	4-block model	9-block model	White's 7-block model (White, 1974b)
W1	A	1	1	1	2
W3	A	1	1	1	1
W4	A	1	1	2	1
S1	A	1	1	2	1
I1	A	1	1	3	2
W2	*	1	2	4	3
W7	B	2	4	7	5
W8	B	2	4	8	4
W9	B	2	4	8	4
S4	B	2	4	9	5
W6	**	2	3	6	6
W5	***	1	2	4	3
S2	***	2	3	6	6
I3	***	1	2	5	7

Key: Blocks are named by letter (Homans) or number (others).

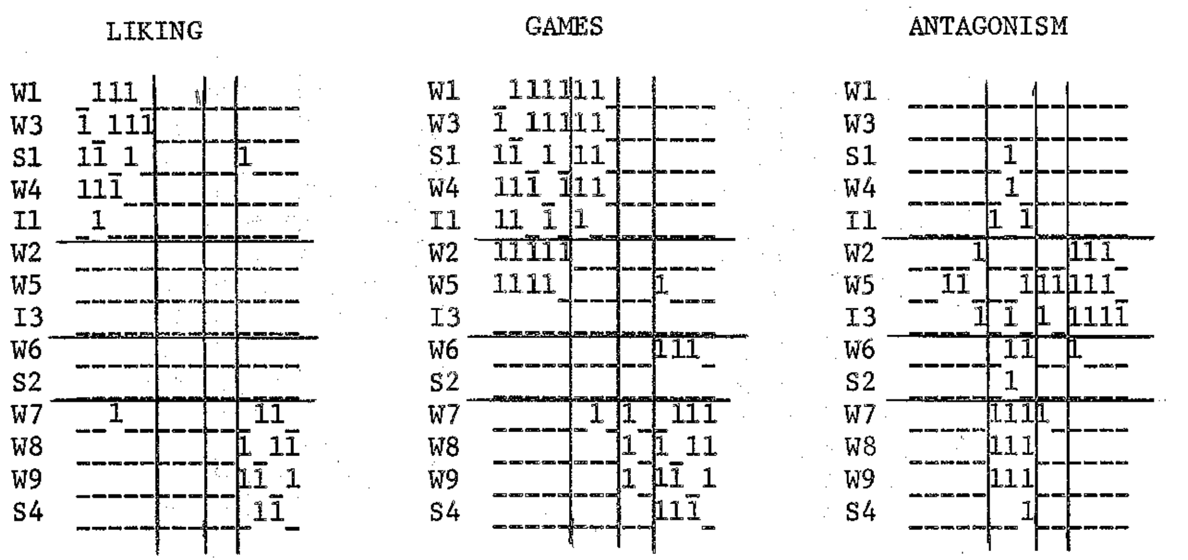
+ Based on Roethlisberger and Dickson (1939), pp. 508-510.
* Man W2 was oriented to but outside of Clique A and "had little to do with it; he entered little into conversation" (Homans, 1950: 70).
** Man W6 was oriented to Clique B but "in many ways was an outsider even in [this] group" (Homans, 1950: 71).
*** In Homans' judgment, men W5, S2, and I3 were not members of either clique.

model (Fig. 6). Below each data-matrix is placed a 4 x 4 matrix indicating the density of ties in the corresponding submatrices of data. As in the Newcomb case, this is a first approach to quantitative treatment of fatness of fit. Note the high frequency of zeroblocks (summing across relations, there are 5 x 16 = 80 blocks and almost half of these blocks [37] are zeroblocks). This occurrence supports the general observation at the end of Section 1, that even without explicitly trying to isolate zeroblocks CONCOR often has this effect when used on networks of basically low tie density.

The blocked "Liking" and "Games" matrices clearly delineate two cliques within which there is positive sentiment. (As mentioned above, Blocks 1 and 4 are identical to the central membership of Homans' cliques A and B, respectively.) The "Liking" matrix would yield a three-block blockmodel $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ according to the strict lean-fit criterion were it not for the presence of a single "discrepant" tie (S1 and W7 choose each other). As White (1974b) states, "this one tie, which abrogates the possibility of [an algebraic] role model based on Homans' cliques, is no accident; it is a significant part of the social structure, a tie between two leaders." (Compare the discussion of "bridges" in Granovetter [1973].)

The "Games" relation (see again Fig. 6) further suggests a status ordering as between central and marginal members of each clique: only central members of a clique play games together, while the marginal members of a clique ("hangers-on") play games only with the central members, <u>not</u> with each other. The appropriate submatrix blockmodel (taking either the first two or the last two blocks) is then of the form

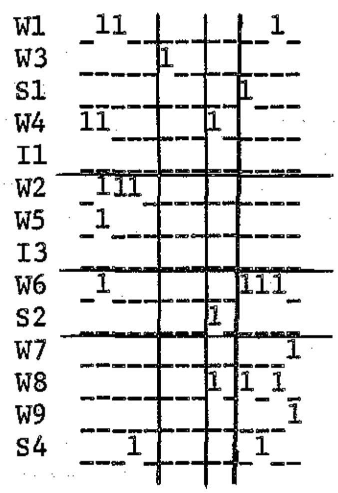
Fig. 6. Five Bank Wiring Group relations blocked into four blocks under CONCOR algorithm. Tie densities for blocks reported beneath each matrix.



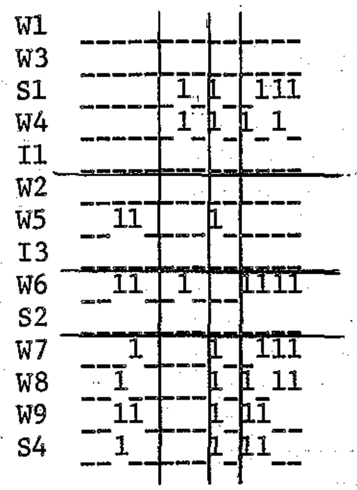
LIKING				GAMES				ANTAGONISM			
0.70	0	0	0.05	0.90	0.60	0	0	0	0.27	0	0
0	0	0	0	0.60	0	0	0.08	0.27	0.33	0.50	0.83
0	0	0	0	0	0	0	0.37	0	0.50	0	0.12
0.05	0	0	0.83	0	0.08	0.37	1.00	0	0.83	0.12	0

Fig. 6. (Cont.)

HELPING



WINDOWS



HELPING

0.20	0.07	0.10	0.10
0.27	0	0	0
0.10	0	0.50	0.37
0.05	0	0.12	0.42

WINDOWS

0	0.13	0.20	0.25
0.13	0	0.17	0
0.20	0.17	0	0.50
0.25	0	0.50	0.83

$E = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, where the "0" indicates absence of ties among the hangers-on population. (See also White and Breiger, 1974 for more extensive discussion in the context of other two-block models.) There is only one case of game-playing <u>between</u> cliques, and this involves a marginal member of one of the cliques.

The "Antagonism" relation is particularly revealing, and provides a substantial amount of additional information supporting a status line of interpretation. No central member of either clique is antagonistic toward either his fellow clique members <u>or</u> toward his opposite numbers (i.e., the four corner blocks are zeroblocks). The complete absence of antagonism between the two central cliques is very much in contrast to the naïve predictions of classical balance theory (Abelson and Rosenberg, 1958) or any of a number of substantially modified and weakened versions of this theory (e.g., Flament, 1963; Newcomb, 1968). <u>The central clique members are antagonistic only toward marginal members.</u>

Note also that there is more antagonism between the two hangers-on groups (3 symmetric ties) than within either of the hangers-on groups (1 symmetric tie). However, the hangers-on groups are both quite small and this last point is correspondingly weak.

Still considering the Antagonism matrix, one next observes that there is a strong asymmetry as between the two central cliques: the members of clique A direct antagonism only toward their own hangers-on, ignoring the hangers-on of clique B, whereas the members of clique B likewise direct antagonism toward the hangers-on of clique A and almost completely ignore their own hangers-on (there is only one exception, in the antagonism between W6 and W7; on this particular relation, see Homans

[1950:77]). Moreover, there is a substantially higher incidence of antagonism between the central members of clique B and the hangers-on of clique A than between these hangers-on and the central members of clique A (contrast the [1, 2] and [4, 2] cells of the blocked antagonism matrix in Fig. 6).

Summarizing this evidence, it is possible to interpret the observed asymmetries between the two cliques as evidence of the "dominant" position of clique A. This dominance is clear from the observer reports, and Homans in particular comments as follows (1950:71): "Each clique had its own games and activities, noticeably different from those of the other group, and clique A felt that its activities were superior to those of clique B." (See also Roethlisberger and Dickson, 1939:510). In developing this differential status interpretation, it is unfortunate that the reported antagonism relation is symmetric, since it is consequently impossible to differentiate negative sentiment ties as between sender and receiver.

In this connection, the "Helping on the Job" relation assumes a potentially important place, since it is the only relation in the data which is not fully symmetric.[11] Again, some status effects are indicated. The hangers-on to clique A did not help each other but helped the central members of clique A to a substantial extent which was not reciprocated. A similar asymmetry appears with respect to the marginal members of clique B. Observe that there are also instances where central members of one clique help central members of the opposite clique. However, these instances are too few and the density of the Helping matrix is too low to draw inferences about the relative status position of the two central cliques.

Finally, there is the Windows matrix, which describes the incidence of controversies about windows in the work room--specifically, whether they should remain open or shut. It is apparent that this was an activity which tended to center primarily around clique B. Homans (1950: 71) also describes several other activities which tend to be clique-specific. The present case admits a very simple interpretation if it is realized that the work room had assigned places for each of the men, and most of the members of clique B were located closer to the windows (see Fig. 2 in Homans [1950:57]).

The detailed analysis just concluded makes clear that the central importance of blockmodels is the way in which these models may be used to clarify relational structure from raw network data. This relational structure goes very much beyond mere partitioning or hierarchical clustering of the underlying population, such as is produced by CONCOR or any other hierarchical clustering procedure. However, it is obviously of interest to assess the performance of the CONCOR algorithm in producing blockings which may subsequently be used as a basis for detailed relational analysis. To this end, we now give a detailed comparative discussion of the relative performance of CONCOR and Johnson's well-known (1967) HICLUS procedures on the Bank Wiring data.

The HICLUS output, Fig. 7 shows the results of analyzing the first-correlation matrix M_1 in Fig. 3 by both Johnson's connectedness and diameter methods.[12] Recall that the diameter method substitutes the maximum distance into the original (proximity) matrix when a new cluster (i,j) is formed, i.e.,

Fig. 7. Hierarchical clustering of first-correlation matrix M_1, derived from Bank Wiring room group data, using HICLUS methods of Johnson (1967). The clusterings are reported in standard HICLUS format. There is no parallel in CONCOR to the cluster values α produced by the HICLUS procedures.

(a) Connectedness method

```
Similarity    W W W W S W W S W I S W W I
  value       6 7 8 9 4 1 3 1 4 1 2 2 5 3

  0.583       . . . . . XXX . . . . . . .
  0.581       . . XXX . XXX . . . . . . .
  0.564       . . XXX . XXX XXX . . . . .
  0.496       . XXXXX . XXX XXX . . . . .
  0.458       . XXXXX . XXXXXXX . . XXX .
  0.426       . XXXXXXX XXXXXXX . . XXX .
  0.408       . XXXXXXX XXXXXXXXX . XXXXX
  0.384       XXXXXXXXX XXXXXXXXX . XXXXX
  0.361       XXXXXXXXX XXXXXXXXX XXXXXXX
  0.300       XXXXXXXXX XXXXXXXXXXXXXXXXX
  0.221       XXXXXXXXXXXXXXXXXXXXXXXXXXX
```

(b) Diameter method

```
Similarity    W W S W W W I W W W W S S I
  value       1 3 1 4 2 5 1 7 8 9 6 4 2 3

  0.583       XXX . . . . . . . . . . . .
  0.581       XXX . . . . . . XXX . . . .
  0.564       XXX XXX . . . . XXX . . . .
  0.458       XXX XXX XXX . . XXX . . . .
  0.453       XXX XXX XXX . XXXXX . . . .
  0.384       XXX XXX XXX . XXXXX XXX . .
  0.361       XXX XXX XXX . XXXXX XXX XXX
  0.340       XXXXXXX XXX . XXXXX XXX XXX
  0.300       XXXXXXX XXX . XXXXXXXXX XXX
  0.272       XXXXXXX XXXXX XXXXXXXXX XXX
 -0.036       XXXXXXXXXXXXX XXXXXXXXX XXX
 -0.152       XXXXXXXXXXXXX XXXXXXXXXXXXX
 -0.258       XXXXXXXXXXXXXXXXXXXXXXXXXXX
```

$d[i,j],k) = \max[d(i,k),d(j,k)],$

whereas the connectedness method substitutes the minimum distance,

$d[i,j],k) = \min[d(i,k),d(j,k)].$

At the coarsest (two cluster) level, the Johnson connectedness method produces the two clusters (W6,W7,W8,W9,S4), (W1,W3,S1,W4,I1,S2,W2,W5,I3). The next splitting of the first cluster leads to (W6,W7), (W8,W9,S4), and one obtains similarly (W1,W3,S1,W4,I1), (S2,W2,W5,I3) for the second cluster (order of individuals follows output in Fig. 7 (a)). The two-cluster split is similar to the two-block CONCOR output, except that S2 in the CONCOR output is placed with (W6,W7,W8,W9,S4) rather than with the other cluster (W1,W3,S1, ..., I3). This difference does not clash in any major way with the substantive judgment of Homans that man S2 was not a member of either clique. At the four-block level, a more significant difference is that the Johnson method places W6 with W7, hence cutting across the boundary of the central clique B membership (W7 is assigned by Homans to clique B, whereas W6 is not).

Similarly the Johnson diameter method leads to the two-cluster split (W1,W3,S1,W4,W2,W5,I1), (W7,W8,W9,W6,S4,S2,I3), which may then be broken into four clusters (W1,W3,S1,W4), (W2,W5,I1), (W7,W8,W9,W6,S4), (S2,I3). The two-cluster diameter solution is again similar to the two-block CONCOR results, although the inspector I3 is now placed with (W7,W8,W9,W6,S4,S2). The four cluster solution deviates in an important way from the CONCOR results by placing man I1 in a different cluster from (W1,W3,S1,W4), hence breaking up the central clique A whereas CONCOR does not. In this respect, the performance of the diameter method is clearly inferior to CONCOR. As in the case of the connectedness method, the diameter method also places W6 with (W7,W8,W9,S4) at the four cluster level, hence again imperfectly discriminating clique B at this level.

In summary, although the performance of the three algorithms is quite similar, CONCOR is the only one of the algorithms to recover the Roethlisberger-Dickson-Homans cliques in a perfect way. The Appendix develops a more detailed quantitative comparison among the three methods, using the tree metrics approach of Boorman and Olivier (1973).

C. <u>Sampson's Monastery</u>

S. F. Sampson (1969) has provided a meticulous account of social relations in an isolated contemporary American monastery. Turbulence was emerging inside American Catholicism in the late 1960's, and there was a major conflict in this particular monastery toward the end of Sampson's twelve-month study. The upshot of this conflict was a mass departure of the members, with the result that Sampson's data is of special interest for what light it may shed on the structure of a social group about to disintegrate for internal reasons.

The wide variety of observational, interview, and experimental information which Sampson developed on the monastery's social structure included the formulation of sociometric questions on four specific classes of relation: Affect, Esteem, Influence, and Sanctioning. Respondents were to give their first, second, and third choices, first on the positive side (e.g., "List in order those three brothers whom you most esteemed"), then on the negative side (e.g., "List in order those three brothers whom you esteemed least"). Responses for eighteen members (not including senior monks) are presented for five time periods;

it should, however, be stressed that the data were obtained after the breakup had occurred, and hence are subject to the kinds of errors which make recall data often unreliable. The present analysis is confined to Sampson's fourth period, just before the major conflict and after a new cohort had initially settled in.

Sampson presents his Time Four data in four tables, one for each class of relations, in which negative choices are represented by negative integers according to the choice level. (Thus, for example, a choice of "like most strongly" appears as +3 in the Affect table, while a choice of "most strongly dislike" appears as -3 in the same table.)

White (1974b) formulates blockmodels on choices which are made binary by using the top two and bottom two choices for each man. This leads to eight binary matrices in all, which are then blocked. We have instead applied the CONCOR algorithm directly to Sampson's reported data involving weighted choices. A 72 x 18 matrix M_0 was formed by vertically "stacking" the Affect, Esteem, Influence, and Sanctioning matrices, taking care as usual to preserve the ordering of columns. Starting with the first-correlation matrix M_1 shown in Fig. 8, CONCOR then produced a two-block partitioning (see Fig. 9) in which one block includes all individuals whom Sampson identifies as the "Loyal Opposition" faction (persons numbered 4, 6, 11, 5, and 9 in Fig. 8) and, in addition, three members whom Sampson terms "interstitial"--i.e., brothers not clearly belonging to any group (persons numbered 8, 10, and 13).

The CONCOR procedure was then repeated on the submatrix formed by taking columns of M_0 corresponding to the remaining block (i.e., columns

Fig. 8. First-correlation matrix M_1 formed on the Sampson monastery data (details in text).

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1.0																	
.23	1.0																
.02	-.07	1.0															
-.33	-.34	-.06	1.0														
-.29	-.48	-.10	.15	1.0													
-.08	-.17	-.23	.41	-.07	1.0												
.12	.24	-.14	-.42	-.01	.13	1.0											
-.04	-.28	-.37	.25	.24	.35	-.09	1.0										
-.19	-.21	-.15	.44	.26	.15	-.40	.02	1.0									
-.15	-.34	-.19	.05	.00	.18	-.02	.21	.00	1.0								
-.35	-.48	.06	.45	.18	.18	-.17	-.01	.10	.43	1.0							
.13	.19	-.26	-.25	-.19	.04	.00	.04	-.17	-.17	-.25	1.0						
-.06	-.33	.15	.02	.09	-.23	-.09	-.05	.04	.00	.04	-.24	1.0					
.10	.31	-.17	-.17	-.06	-.13	-.03	.02	-.04	-.33	-.39	.19	-.21	1.0				
.26	.38	-.16	-.41	-.17	.02	.23	-.12	-.14	.00	-.33	.17	-.26	-.01	1.0			
-.12	.31	-.18	-.24	-.09	-.28	-.02	-.16	-.26	.08	-.18	.17	.10	-.03	.20	1.0		
.11	-.14	.31	-.43	-.04	-.24	.12	-.26	-.17	-.15	-.22	.05	.19	.11	-.18	-.10	1.0	
.07	-.15	.25	-.37	-.05	-.56	.04	-.27	-.07	.12	-.09	-.11	.20	.08	-.06	-.01	.56	1.0

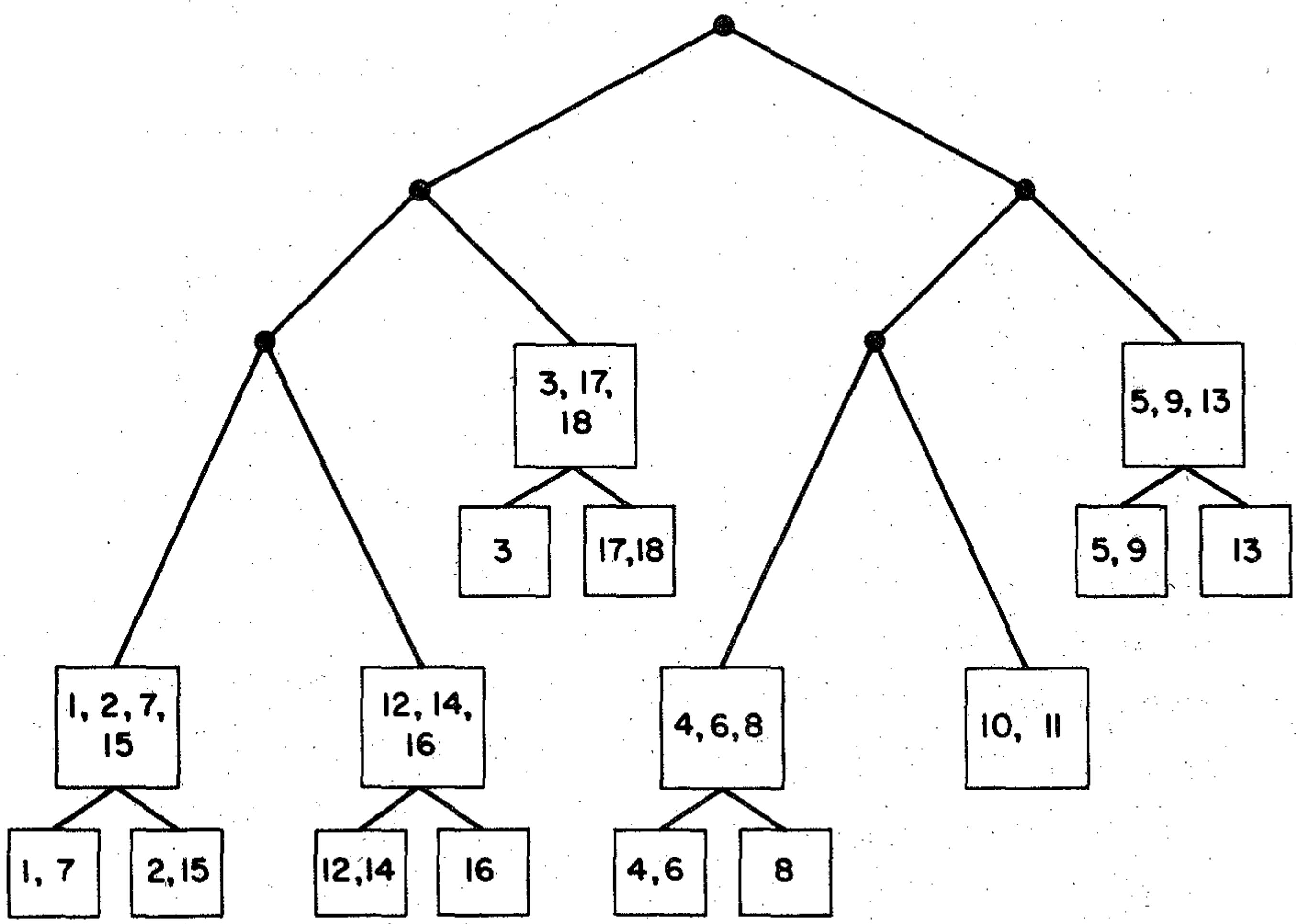


Fig. 9. Hierarchical clustering representation of repeated application of the CONCOR algorithm on Sampson's data.

1,2,3,7,12,14,15,16,17,18). Convergence of this 72 x 10 matrix resulted in the further partitioning (1,2,7,12,14,15,16) (3,17,18). The first group just enumerated corresponds identically with the "Young Turk" faction which Sampson identifies through a combination of many analytical techniques. The second group (3,17,18) coincides with the "Outcast" group which Sampson also identified. Together with the individual numbered 2, one of the leaders of the Young Turk faction, the Outcast group was the group whose expulsion from the monastery triggered a mass resignation which soon followed.

On the basis of an intuitive search for lean fit blockmodels, White (1974b, Table 10) has formulated a five-block model of the monastery's social structure, as well as a coarser three-block version formed from these five blocks. White's three-block model may be formally obtained by applying CONCOR to the stacked version of the eight raw Sampson matrices distinguishing "most" from "least," rather than the collapsed version of four stacked matrices on which the present analysis is based. White's three-block version and ours (just described) are identical with the exception of the individual numbered 13: White places him among the "Outcasts" and we place him with the "Loyal Opposition." Significantly, Sampson labels the individual in question as one of the three "interstitial" members of the monastery, implying that his structural position was ambiguous (see also p. 45 below).

The discussion thus far suggests excellent comparability of our results both with Sampson's own analysis and with White's three-block model. In order to explore the results further, we now return, as in the Bank Wiring analysis, to the original relational data.

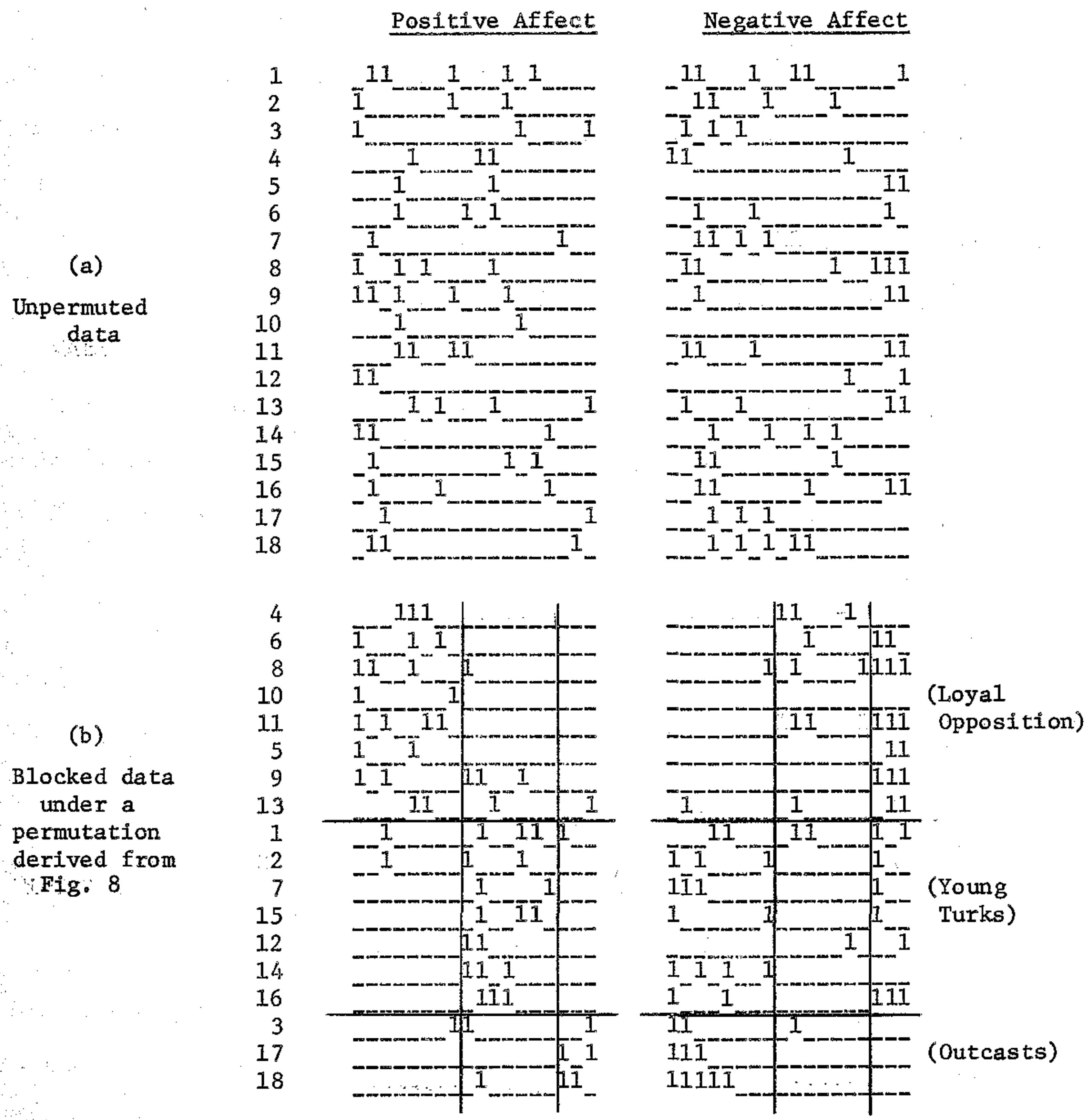
In Fig. 10 we present a summary description of the two kinds of affect relation with which Sampson deals. The matrices on the left of Fig. 10 consist of the Boolean union of Sampson's (positive) Affect, Esteem, Influence, and Sanctioning relations. The matrices on the right of Fig. 10 consist of the Boolean union of Sampson's Dislike, Disesteem, Negative Influence, and Negative Sanction relations. In obtaining the Boolean matrices from which these unions are formed, the top two and bottom two choices (respectively) are utilized.

The first row of matrices in Fig. 10 displays the Positive and Negative Affect relations in their unpermuted row-column order. The second row of Fig. 10 displays these same matrices permuted into a form compatible with the three-block model obtained above: using Sampson's labels, these blocks correspond to the Loyal Opposition + Waverers (persons numbered 4,6,8,10,11,5,9,13), the Young Turks (1,2,7,15,12, 14,16), and the Outcasts (3,17,18). The third row of Fig. 10 indicates densities of entries within the blocks of these last two matrices.

Examination of the third column of the blocked matrices in the second row of Fig. 10 strongly suggests why the Outcasts were so named: they receive a disproportionate share of the <u>negative</u> ties from individuals in other blocks, and virtually no <u>positive</u> ties.

Seen as a whole, the pattern evinced by Fig. 10 may be interpreted as an approximation to the sociometric "clustering" phenomenon discussed by Davis (1968), i.e., presence of "two or more subsets such that each positive line joins two points of the same subset and each negative line joins points from different subsets." Specifically, examination of the tie densities in the blocked Sampson data shows that most of the

Figure 10. Summary description of the Sampson data, showing unpermuted and blocked forms, and also block densities.



(c)

0.375	0.0893	0.0417
0.0357	0.429	0.0476
0.0417	0.0952	0.833

0.0357	0.161	0.625
0.286	0.0714	0.429
0.417	0.0476	0

positive affect ties are concentrated within blocks and most of the negative affect ties occur between blocks. It should be emphasized that this clusterability pattern is specific to the present data and does not necessarily generalize: just as blocks need not be cliques (see Section 1 above), so also blocks may--but need not--form clusters or approximate clusters in the above sense of Davis. As an illustration, turn back to the "Games" matrix in the Bank Wiring data (Fig. 6 above). Here the presence of numerous ties between the obtained blocks violates the Davis condition if blocks are to be understood as clusters in his sense. At the same time, however, the between-block positive ties are clearly interpretable in this case: they indicate the bonds between "hangers-on" and central clique membership.

Note that even though the Fig. 10 blocked matrices contain only one zeroblock, there is a clearly defined set of blocks which are close to being zeroblocks because of very low tie density. This judgment is borne out by a clear bimodality in the frequency histogram of block densities (Fig. 11).

Examining the pattern of block densities in more detail, it appears that the highest within-block density on positive sentiment is achieved within the Outcasts (.833, as opposed to .375 and .429 for Loyal Opposition and Young Turks respectively). Of the three groups, the small Outcast group hence approaches most nearly to the definition of a clique in classical sociometry. Note also that with respect to positive sentiment the Young Turks fall into two clear groups, (1,2,7,15) and (12,14,16), with the (12,14,16) subset distinguished by the absence of direct positive sentiment ties among its members (zeroblock on main

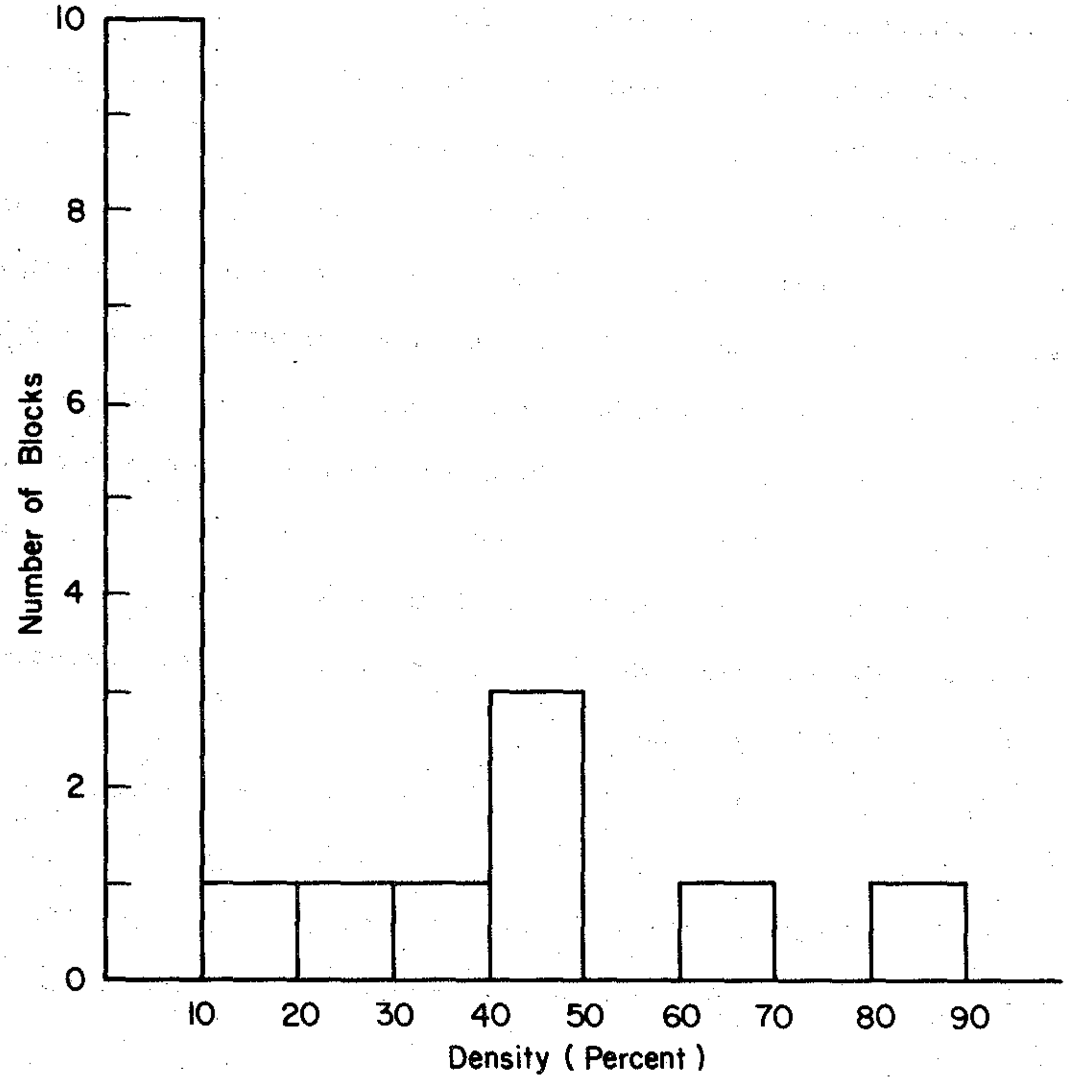


Fig. 11. Frequency histogram of within-block densities in the three-block Sampson model of Fig. 10 (see Fig. 10c).

diagonal in Fig. 10(b) positive affect matrix. This further division is reproduced by CONCOR (see again Fig. 9). The blocked negative sentiment matrix in Fig. 11(b) again reveals the Outcasts as a cohesive group, receiving a high incidence of negative sentiment from the other two groups (the [1,3] and [2,3] cells in the blocked negative matrix have densities .625 and .429 respectively, which are the two highest density cells in this matrix. Note that there is a virtual absence of negative sentiment directed from the Outcasts to the Young Turks (only one entry), which is in contrast to the quite high incidence of negative sentiment directed from Outcasts to the Loyal Opposition. This observation is consistent with the prevailing factional politics, since the Outcasts were among those later expelled whereas the Loyal Opposition formed the core of those remaining through all the subsequent resignations. Finally, note that there is a considerably higher incidence of negative sentiment ties directed by the Young Turks to the Loyal Opposition than vice versa([2,1] cell has density .286, while [1,2] cell has density only .161).

Finally, Fig. 12 shows the output of the Johnson connectedness and diameter methods on the M_1 Sampson matrix of Fig. 8. Both methods basically recover the three-way split into Loyal Opposition, Young Turks, and Outcasts, but both differ from CONCOR in Fig. 9 in placing the interstitial man 13 among the Outcasts. The diameter method also reveals the partition of the Young Turks earlier indicated, which splits them into the two subsets (1,7,2,15) and (12,14,16); the connectedness method does not reproduce this precise split.

Additional numerical comparison of the three methods on the present

Fig. 12. Application of Johnson (1967) HICLUS methods to first-correlation matrix M_1 for Sampson data.

(a) Connectedness method

```
Similarity   0 0 0 1 0 0 1 1 0 0 1 0 1 1 1 0 1 1
  value      5 8 6 0 9 4 1 2 7 1 4 2 5 6 3 3 7 8

  0.556      . . . . . . . . . . . . . . . . XXX
  0.452      . . . . . XXX . . . . . . . . . XXX
  0.440      . . . . XXXXX . . . . . . . . . XXX
  0.432      . . . XXXXXXX . . . . . . . . . XXX
  0.411      . . XXXXXXXXX . . . . . . . . . XXX
  0.383      . . XXXXXXXXX . . . . XXX . . . XXX
  0.352      . XXXXXXXXXXX . . . . XXX . . . XXX
  0.313      . XXXXXXXXXXX . . . . XXXXX . . XXX
  0.308      . XXXXXXXXXXX . . . XXXXXXX . . XXX
  0.307      . XXXXXXXXXXX . . . XXXXXXX . XXXXX
  0.264      XXXXXXXXXXXXX . . XXXXXXXXX . XXXXX
  0.236      XXXXXXXXXXXXX . XXXXXXXXXXX . XXXXX
  0.204      XXXXXXXXXXXXX . XXXXXXXXXXX XXXXXXX
  0.193      XXXXXXXXXXXXX XXXXXXXXXXXXX XXXXXXX
  0.120      XXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXX
  0.117      XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

(b) Diameter method

```
Similarity   0 0 1 0 0 0 1 0 0 0 1 1 1 1 1 0 1 1
  value      6 8 0 5 9 4 1 1 7 2 5 2 4 6 3 3 7 8

  0.556      . . . . . . . . . . . . . . . . XXX
  0.452      . . . . . XXX . . . . . . . . . XXX
  0.383      . . . . . XXX . . XXX . . . . . XXX
  0.352      XXX . . . XXX . . XXX . . . . . XXX
  0.264      XXX . XXX XXX . . XXX . . . . . XXX
  0.247      XXX . XXX XXX . . XXX . . . . XXXXX
  0.234      XXX . XXX XXX . XXXXX . . . . XXXXX
  0.185      XXX . XXX XXX . XXXXX XXX . . XXXXX
  0.178      XXXXX XXX XXX . XXXXX XXX . . XXXXX
  0.154      XXXXX XXX XXX . XXXXX XXX . XXXXXXX
  0.125      XXXXX XXX XXX XXXXXXX XXX . XXXXXXX
  0.103      XXXXX XXXXXXX XXXXXXX XXX . XXXXXXX
 -0.031      XXXXX XXXXXXX XXXXXXX XXXXX XXXXXXX
 -0.072      XXXXXXXXXXXXX XXXXXXX XXXXX XXXXXXX
 -0.118      XXXXXXXXXXXXX XXXXXXXXXXXXX XXXXXXX
 -0.328      XXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXX
 -0.562      XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

data is contained in the Appendix.

D. Social Participation in "Old City"

As part of their classic Deep South study, Davis, Gardner and Gardner (1941:146-151) present research on the social participation of eighteen women at fourteen social events (such as a card party, a church supper, and so on) held during the course of a year. Their goal was to determine cliques present in this small population. This example was subsequently used by Homans (1950:82-86) in his section on the "Definition of the Group." Breiger (1974) has employed an *ad hoc* clique detection procedure to this data which emphasizes the "duality" of persons and groups.

The unpermuted data matrix, whose (i,j)th entry signifies the presence ("1") or absence ("0") of woman *i* at event *j*, is shown in Fig. 13a. Columns are arranged chronologically and rows are ordered arbitrarily.

The present algorithm was applied to the (single) original matrix. Blockings into two blocks were obtained separately for columns (events) and for rows (women). Then these distinct partitionings were imposed (respectively) onto columns and rows of the original data (see Fig. 13b). In the reordered matrix, one may directly observe a strong association of the first cluster of women with the second cluster of events. The presence of this association is corroborated by a Yule's Q of -.941 on the 2 x 2 table formed by taking within-block sums of the Fig. 11b matrix.

The two-block partition of women thus obtained is (Eleanor, Ruth, Charlotte, Brenda, Laura, Evelyn, Theresa, Frances) and (Dorothy,

Fig. 13. Participation data on women in a Southern city, illustrating the use of the CONCOR algorithm to block membership data. In Fig. 13 (a), women (rows) are ordered arbitrarily and events (columns) chronologically (adapted from Homans [1950:83]). Fig. 13 (b) displays this same matrix after applying CONCOR separately to rows and columns.

(a)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1.	Eleanor		1		1				1				1		
2.	Brenda		1		1			1	1		1		1	1	
3.	Dorothy						1						1		
4.	Verne				1	1	1						1		
5.	Flora	1					1								
6.	Olivia	1					1								
7.	Laura		1	1	1			1	1		1		1		
8.	Evelyn		1	1			1	1	1		1		1	1	
9.	Pearl						1		1				1		
10.	Ruth		1		1		1						1		
11.	Sylvia				1	1	1			1		1	1		1
12.	Katherine					1	1			1		1	1		1
13.	Myrna					1	1			1			1		
14.	Theresa		1	1	1		1	1	1				1	1	
15.	Charlotte		1		1			1						1	
16.	Frances		1					1	1				1		
17.	Helen	1			1	1				1			1		
18.	Nora	1			1	1	1		1	1		1			1

(b)

	1	5	6	9	11	14	2	3	4	7	8	10	12	13
Eleanor							1		1		1		1	
Brenda							1		1	1	1	1	1	1
Laura							1	1	1	1	1	1	1	
Evelyn			1				1	1		1	1	1	1	1
Ruth			1				1		1				1	
Theresa			1				1	1	1	1	1		1	1
Charlotte							1		1	1				1
Frances							1			1	1		1	
Dorothy			1										1	
Verne		1	1						1				1	
Flora	1		1											
Olivia	1		1											
Pearl			1								1		1	
Sylvia		1	1	1	1	1			1				1	
Katherine		1	1	1	1	1							1	
Myrna		1	1	1									1	
Helen	1	1		1					1				1	
Nora	1	1	1	1	1	1			1		1			

Flora, Olivia, Pearl, Verne, Sylvia, Katherine, Myra, Helen, Nora). The first block contains the seven women whom Homans (1950:84) identifies as members of one clique, while the second block contains the five women whom Homans terms members of the other clique. In Homans' evaluation, the remaining six women were marginal to one or both cliques.

This application illustrates the usefulness of CONCOR as a method for analyzing individual-by-committee membership data.

E. Levine's "Sphere of Influence"

Levine (1972) has studied a set of interlocked directorates of the boards of major American banks and corporations. Specifically, this study starts with a 70 x 14 matrix whose (i, j)th entry is the number of directors shared by corporation _i_ and bank _j_. His "study of network representation" employs an unfolding variant of Guttman-Lingoes smallest space analysis to produce a gnomonic map of the "sphere of influence." We have applied the CONCOR algorithm separately to the rows and columns of Levine's original 70 x 14 matrix in our own effort to identify clusters of corporations and of banks which are highly interrelated. Figures 14 and 15 show respectively the results of column (banks) and row (corporations) applications.

With respect to columns (banks) of the 70 x 14 matrix, the first bipartition (Fig. 14) separates the five Chicago banks from the others. Repeating the CONCOR algorithm with respect to the non-Chicago banks, these latter are separated at the next step into New York banks and Pittsburgh banks. The one exception is that Chemical Bank of New York is placed with the Pittsburgh group. Levine's three-dimensional joint space also recovers the regional bank groupings.

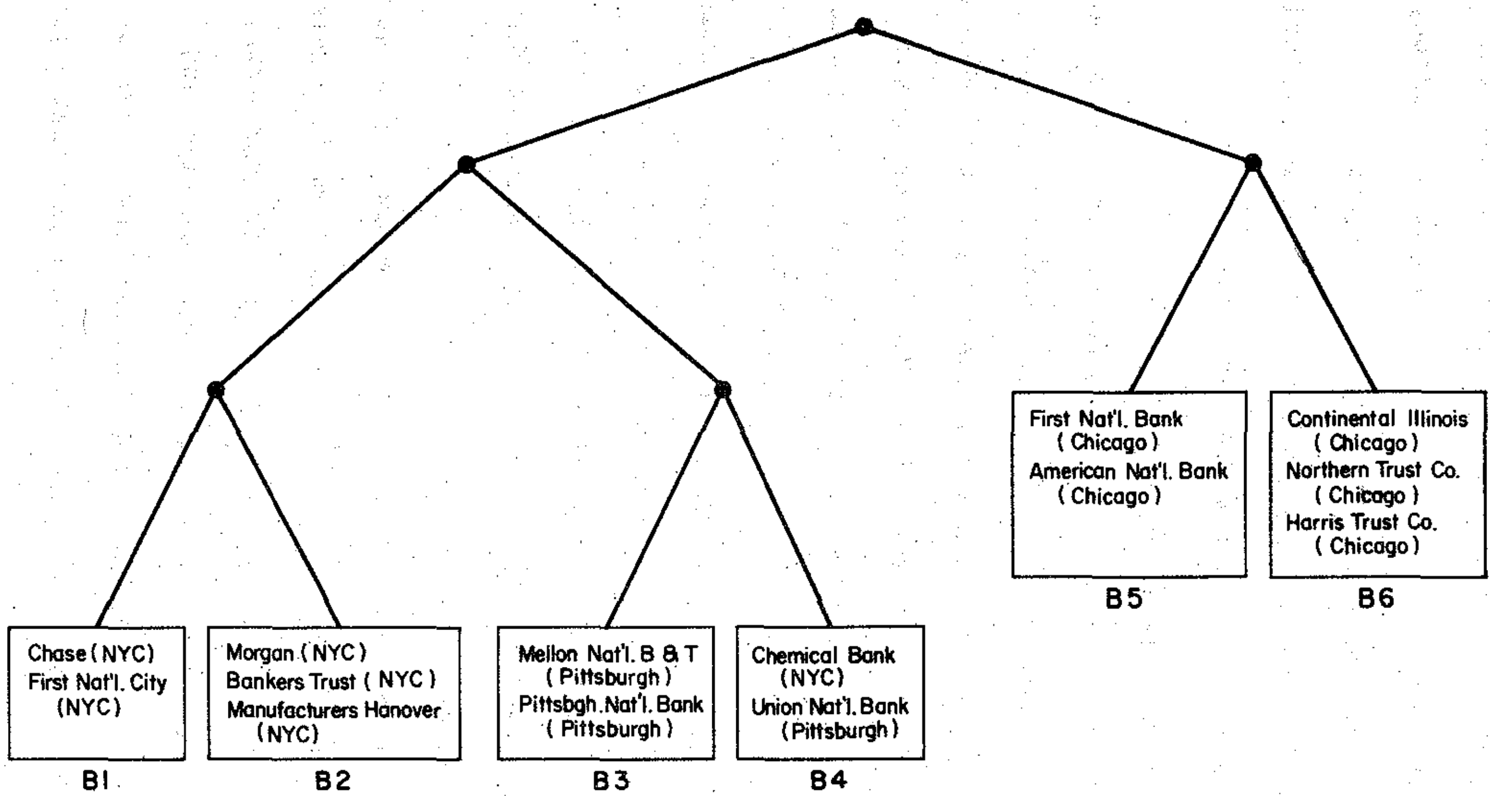


Fig. 14. Hierarchical clustering representation of repeated CONCOR application on the columns (banks) in the Levine (1972) data.

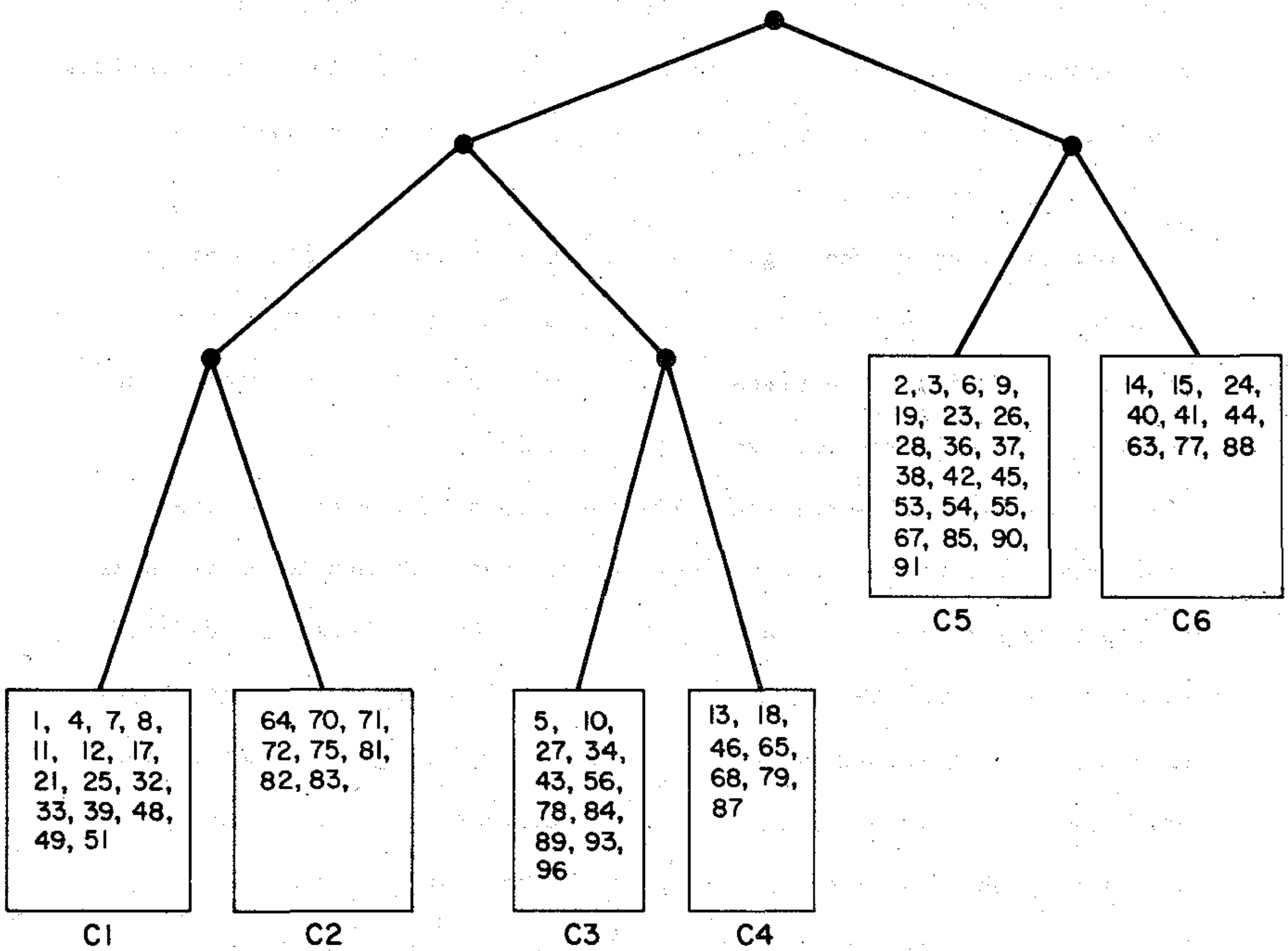


Fig. 15. Hierarchical clustering representation of repeated CONCOR application on the rows (corporations) in the Levine (1972) data. Numbering follows Levine.(*)

(*) There is an error in labeling TRW in Fig. 10 of Levine (1972:25), which reports TRW as Corporation 92 instead of 93 as in his Fig. 5 (1972:19). We follow Levine's Fig. 5 for the present numbering.

Turning next to the rows (corporations) of Levine's matrix, we formed the 70 x 70 first-correlation matrix M_1. Blocking this matrix through repeated applications of CONCOR leads to the six-block partitioning of the seventy corporations shown in Fig. 15. (Corporations are numbered consecutively in the ordering in which they appear in Fig. 5 of Levine, 1972.)

One may compare the Fig. 15 structure to Levine's "sphere of influence" obtained by Guttman-Lingoes scaling (specifically, to his Fig. 10 [1972:25]). The present results are generally consistent with clusters in the Levine smallest space solution.

One may also consider the self-consistency of the present dual procedure for blocking on both rows and columns. Figure 16a shows sums within blocks of the original 70 x 14 matrix, where blocks are defined by cross-tabulating the separate bank and corporation partitions. Rows of Fig. 16a index blocks of corporations (the ordering of blocks is their ordering from left to right in Fig. 15); columns of Fig. 16a index blocks of banks (as ordered in Fig. 14). Utilizing a method of Mosteller (1968; see also Romney, 1971) one may also correct for the effects of unequal row and column marginals by simultaneously normalizing row and column sums in Fig. 16a. The resulting matrix (Fig. 16a) has the property that the largest entry (i,j) in any row _i_ is also the largest entry in column _j_. This may be taken as an indication of the mutual tendency of particular groups of banks and corporations to share directors.

Fig. 16. (a) Number of director interlocks between each of the six sets of corporations obtained in Fig. 15 and the six sets of banks in Fig. 14.

(b) The result of normalizing the previous matrix to have both row and column marginals = 1 (i.e., doubly stochastic form).

(a)

	B1	B2	B3	B4	B5	B6
C1	4	25	3	17	0	7
C2	1	8	3	2	0	0
C3	1	12	18	0	1	1
C4	0	0	0	0	5	21
C5	39	12	4	3	2	1
C6	6	3	0	0	13	3

(b)

	B1	B2	B3	B4	B5	B6
C1	0.0594	0.224	0.048	0.562	0	0.107
C2	0.074	0.358	0.239	0.329	0	0
C3	0.0331	0.24	0.643	0	0.0498	0.034
C4	0	0	0	0	0.259	0.741
C5	0.636	0.118	0.0703	0.109	0.0491	0.0168
C6	0.197	0.0596	0	0	0.642	0.101

PART II. APPLICATIONS OF MULTIDIMENSIONAL SCALING TO THE SOCIAL STRUCTURE DATA OF PART I

Three applications will be developed, dealing respectively with the Bank Wiring Room data, and the Sampson monastery data, and the Newcomb-Nordlie fraternity data. The scaling procedures used are the MDSCAL program of Kruskal (1964a,b) and the INDSCAL algorithm of Carroll and Chang (1970). In addition to these scaling procedures, certain aspects of the MDSCAL solution in the Bank Wiring group example have also been interpreted through use of a recent non-hierarchical clustering algorithm of Arabie and Shepard (1973) (acronym: ADCLUS). This last means of representation is of special interest because it explicitly makes allowance for the possibility of overlapping clusters. This raises the possibility of isolating ways in which the CONCOR algorithm, and blockmodels more generally, may distort or oversimplify overlapping membership properties inherent in social structures to which they are applied.

In MDSCAL and ADCLUS applications, the algorithms are applied to first-correlation matrices derived from raw data as in Part I (e.g., Figs. 3 and 8). It should be noted that similar matrices describing correlations among sociometric positions have been studied using factor analysis in a number of earlier investigations on other data (e.g., MacRae, 1960; see also Katz, 1947, Glanzer, and Glaser, 1959).[13]

The data applications are now presented in the following order: MDSCAL and ADCLUS on the Bank Wiring Room group; MDSCAL on the Sampson monastery group; INDSCAL on the Newcomb Year 2 data.

1. MDSCAL and Non-hierarchical Clustering Analysis of the Bank Wiring Group

As a first scaling application, Kruskal's nonmetric multidimensional scaling program, MDSCAL (Kruskal, 1964a,b) is applied to the first correlation matrix reported for the Bank Wiring data in Fig. 3. The MDSCAL algorithm is well-known and its details will not be resummarized here. The result of this application is shown in Fig. 17, which displays the obtained two-dimensional MDSCAL 5M solution giving the best stress (.126, formula 1) of 20 alternative random initial configurations.[14]

Notice that this approach to network scaling is quite distinct from that employed by Laumann and co-workers in studies of the social structure of a German community elite (Laumann, 1973; Laumann and Pappi, 1973; Laumann, Verbrugge, and Pappi, 1974). Specifically, Laumann and Pappi start by defining a distance matrix in terms of the least-path distance between individuals in a given network (all relational ties presumed symmetric). There is no formation of a correlation matrix such as M_1 in Fig. 3, and the Laumann approach measures connectivity rather than similarity of structural position.

Compatibility of Fig. 17 with blockmodel approaches using CONCOR is extremely good, to the point where one can infer most of the hierarchical clustering shown in Fig. 4 from examining convex clusters in the scaling solution of Fig. 17. The two central cliques A and B identified in Homans' analysis (p. 26 above) emerge as well-separated clusters in the scaling. The wiremen W2 and W6, who are both essentially classified by Homans as hangers-on, occur in positions close to, but somewhat removed from, their respective cliques. This summary statement

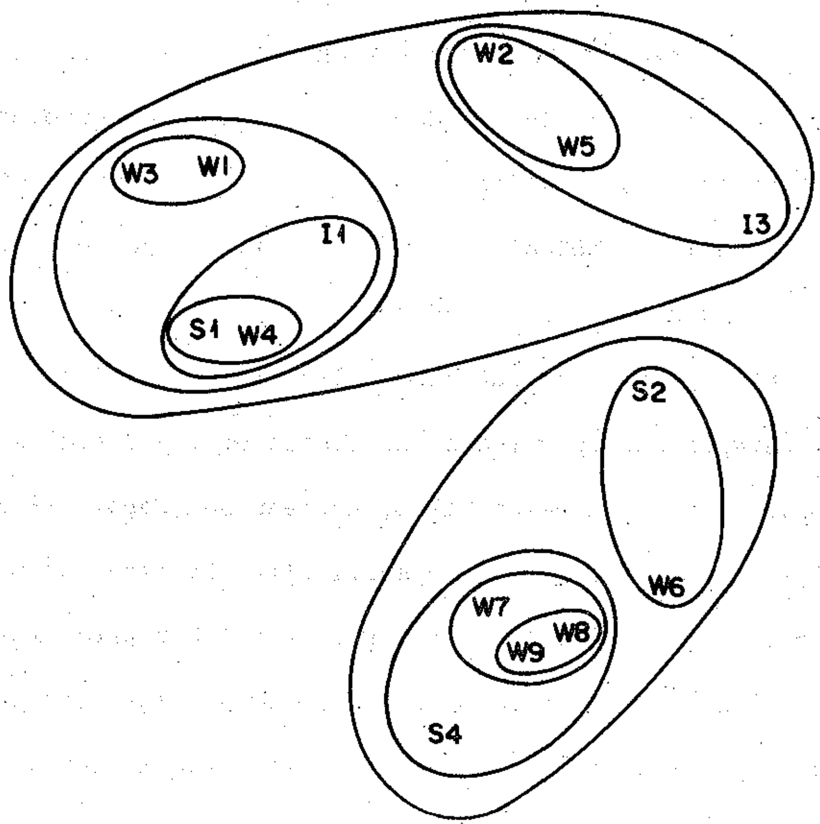


Fig. 17. Two-dimensional MDSCAL-5M solution for input proximity data given by Fig. 3 (first-correlation matrix for the Bank Wiring Group). Stress formula 1, stress = 12.6%. Superimposed clusters are obtained from the CONCOR results shown in Fig. 4.

is also true, to a somewhat lesser extent, of S2 and (W5, I3), which the 4-block CONCOR model places as hangers-on to Homans' cliques A and B, respectively. The further CONCOR applications reported in Fig. 4, which lead to still finer partitionings, are also clearly reflected in the scaling; thus clique A in the scaling breaks up into (W1,W3) and (S1,W4,I1), and this last cluster in turn splits into I1 and (S1, W4), again reflecting the CONCOR performance shown in Fig. 4.

Despite this very close agreement between the two algorithms, CONCOR and MDSCAL, there is also good reason to probe as hard as possible in the direction of non-hierarchical ways of describing the social structure. In order to explore this direction, application has been made of the recent ADCLUS algorithm of Arabie and Shepard (1973). Given a single proximity matrix $P = [P_{ij}]$ on n items, this algorithm is designed to select a family $\mathcal{S}$ of (possibly overlapping) clusters or subsets of these items and assign a positive numerical weight w_C to each cluster C, in such a way as to achieve a best fit to the additive membership model

$$P_{ij} \cong \sum_{C} \delta_{iC}\, \delta_{jC}\, w_C, \quad \delta_{iC} = \begin{cases} 1 & \text{if item } i \text{ is contained in cluster } C, \\ 0 & \text{otherwise,} \end{cases}$$

i.e., a model which predicts the similarity between two items to be the sum of the weights of clusters containing both.

Starting from the correlation matrix in Fig. 3, application of the ADCLUS algorithm led to the set of clusters and associated weights (which accounted for 91.2% of the variance) shown in Fig. 18. Many of the clusters are identical or close to those which are implied in the CONCOR tree (Fig. 4). It is worth noting that the ADCLUS algorithm also

Fig. 18. List of clusters and cluster weights obtained from the Fig. 3 Bank Wiring Group correlation matrix by the ADCLUS (non-hierarchical clustering) algorithm (Arabie and Shepard, 1973).

	Cluster (C)	Weight (w_C)	Present as Subtree in Fig. 4	Δ^{+}
1.	(W2,W5,I3)	.4888	Yes	0
2.	(S2,I3)	.4155	No*	.500
3.	(W6,W7,W8,W9,S4)	.3951	No*	.167
4.	(W2,I3)	.3358	No*	.500
5.	(W1,W2,W3,S1,W4,I1)	.2994	No*	.167
6.	(W1,W3,S1,W4)	.2742	No*	.200
7.	(S1,W4,W5,W6,S2,W7, W8,W9,S4,I1,I3)	.2303	No	.214
8.	(W1,W2,W5,I1)	.2284	No	.500
9.	(W1,W3)	.2181	Yes	0
10.	(W9,S4)	.2120	No*	.500
11.	(W5,W6,S4)**	.2012	No	.667
12.	(W6,S2,W7,W8,W9)	.1189	No*	.667
13.	(W6,S2,I3)	.1162	No*	.333
14.	(W1,W2,W3,S1,W4,W5,S2,I1)	.1041	No	.222
15.	(S1,W6,S2,W7,W8,W9)	.0808	No	.286
16.	(S1,W5,W8,W9,I1)	.0788	No	.600
17.	(W1,W3,S1,W4,W5,S2,W7,I1)	.0640	No	.400
18.	(W1,W2,W3,W4,I1)	.0635	No	.333
19.	(S1,W4,W6,S2,W7,W8)	.0587	No	.500

* Differs from some subtree Fig. 4 only by one man (either added or subtracted).

** This cluster is the only cluster in the high-weight group [Clusters 1-11] whose meaningfulness is clearly in doubt.

+ $\Delta = \Delta(C) \equiv \min\limits_{S \in T} \left(\frac{|S \Delta C|}{|S \cup C|} \right)$, where T is the Fig. 4 tree, $S \in T$ means that S is a cluster implied by T (in the terminology of Boorman and Olivier, 1973, S is the <u>node set</u> of a subtree of T), and Δ is the standard set-theoretic symmetric difference operation. $|\;|$ denotes the size of a set. $\Delta(C)$ has the properties of a distance measure (see Boorman, 1970).

assigns major weight to some clusters which are not directly implied by Fig. 4 and yet which have been given explicit interpretation in Homans' verbal description. Among such clusters are (W1,W2,W3,S1,W4,I1) ($w \cong .30$) and (W6,W7,W8,W9,S4) ($w \cong .40$). The second of these particular clusters, however, appears in both of the Johnson HICLUS solutions for the Bank Wiring data (see Fig. A1 in Appendix 1). Homans (1950:69) speaks specifically of these two clusters as the two groups of individuals who participated in games (cf. also Games matrix in Fig. 6). Neither of these clusters appears in the CONCOR solution of Fig. 4. It should also be observed that there is a clear elbow in the distribution of assigned weights of the ADCLUS clusters, with a large jump from the cluster (W5,W6,S4), with an assigned weight ~ .20, to the cluster (W6,S2,W7,W8,W9), with an assigned weight ~ .12.

There is, however, little question that the Bank Wiring Group data basically sustains the hierarchical subgroup organization shown in Fig. 4. It is possible that the presence of this hierarchical cluster structure may to some extent reflect the extent to which the Bank Wiring data reports an isolated group in equilibrium. Again, it should be stressed that hierarchical clustering structure has nothing in general to do with the presence of social hierarchy, and represents a totally distinct concept. Presence of such structure is further borne out by the fourth column of Fig. 18, which reports a measure of the discrepancy between each given ADCLUS cluster and the CONCOR tree in Fig. 4. Taking the product moment correlation between the weights w_C and the Δ column of Fig. 17, one obtains $r = -.37$. This indicates a <u>positive</u> relation between the magnitude of ADCLUS weights and the property of being close to some CONCOR block. In other words, the highest weight clusters in the ADCLUS solution also tend to be similar to clusters obtained in the Fig. 4 hierarchical clustering.

2. MDSCAL Analysis of the Sampson Monastery

The same scaling procedure as in the last section has also been applied to the 18-man Sampson monastery group. Figure 19 reports the two-dimensional MDSCAL solution starting with first-correlation matrix used in the CONCOR analysis of this same data (Fig. 8 above). Again, the scaling algorithm reproduces the basic blockmodel clusters. The Young Turks emerge as a distinct cluster, as also are the Loyal Opposition and the Outcasts (see Fig. 19, and notice the strong similarity to Fig. XVII [p. 370] in Sampson, 1969). The interstitial status of man 13 emerges very clearly from the scaling plot, and it is evident from this position why there might be some ambiguity as to his placement (Loyal Opposition or Outcasts, but it is not clear which). Men 8 and 10, whom Sampson also views as waverers, are clearly placed between the core Loyal Opposition and the Young Turks, although closer to the former cluster. This last placement is one respect in which the scaling solution gives information which CONCOR does not (see Fig. 9).

The further applications of CONCOR, leading to the Fig. 9 tree, are somewhat less consistent with the detailed structure of the scaling solution than in the Bank Wiring case. For example, (10,11) and (5,9,13) are both blocks obtained through CONCOR, but these blocks crosscut one another in the Fig. 19 scaling.

Viewed within the context of the MDSCAL solution, some of the more elongated clusters in Fig. 19 look suggestive of the "chaining" effects

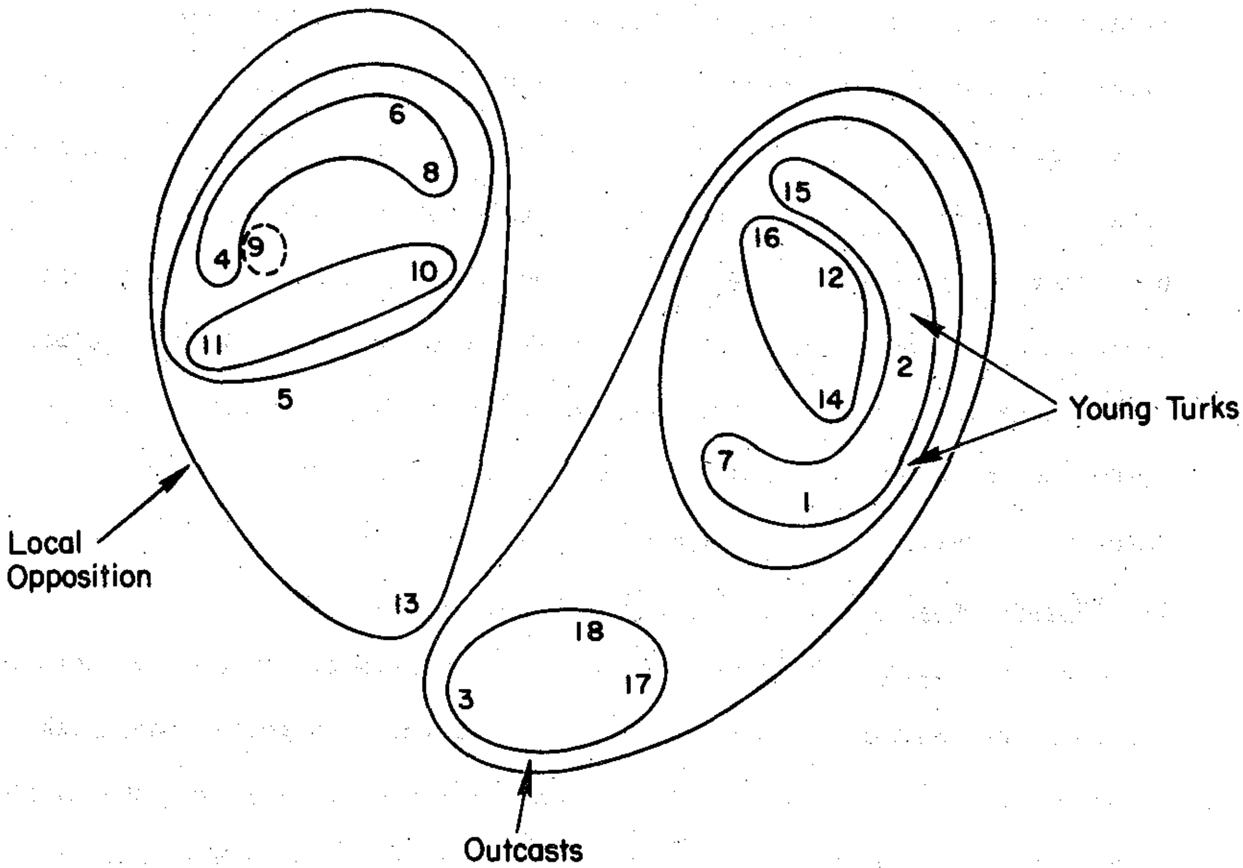


Fig. 19. Two-dimensional MDSCAL-5 solution for input proximity matrix data given by Fig. 7 (first-correlation matrix for Sampson's monastery data). Stress formula 1, stress = 18.6%. Superimposed clusters selected from the CONCOR results in Fig. 9.*

*There is one particular cluster (5, 9, 13) implied by Fig. 8 which for reasons of clarity is not indicated in the present Figure.

(Lance and Williams, 1967a; Jardine and Sibson, 1971) that often stigmatize the connectedness method in HICLUS. Specifically, "chaining" is a generic term for the tendency displayed by certain clustering methods to add new elements to pre-existing clusters as one moves up the hierarchy, rather than for elements to act as the nucleus of new groups (Lance and Williams, 1967a:374). However, more systematic investigation in the Appendix indicates that the overall mathematical behavior of CONCOR on the Sampson data is actually closer to the diameter method than either method is to the connectedness method.

3. <u>INDSCAL Analysis of the Newcomb Fraternity Data</u>

In its entirety, Nordlie-Newcomb data consists of complete preference profiles for fraternity groups in each of two years, reported each week for sixteen weeks (except for the absence of reported data from the ninth week of Year 2; see Nordlie, 1958). Henceforth, following Newcomb, we will enumerate Year 2 weeks with references to this missing week and starting with Week 0, thus 0,1,2,3,4,5,6,7,8,X,10,11,12,13,14,15. This depth of longitudinal information is exceptional in the published literature, and opens the possibility of systematically tracing the evolution of the social structure in each year (compare the use of MDSCAL in Arabie and Boorman [1973] to trace the over-time changes in the social structure of a vervet monkey troop, drawing on data of Struhsaker [1967] and partition metrics developed in Boorman [1970] and Boorman and Arabie [1972]). Specifically, even very crude examination of the Newcomb-Nordlie data suggests that the final situation in Week 15 of Year 2 was the equilibrium outcome of a process which starts in Week 1 and rapidly approaches the final structure by Week 4 or Week 5.

For instance, consider the specific two-block model obtained earlier (p. 24) and note that the number of errors associated with this blocking is 1 (in the l matrix, lower right) and 5 (in the a matrix, top and bottom left), giving 6 errors in all. Over the fifteen weeks the number of errors counted in this same way for each week lead to the 15-term sequence, starting from Week 0 (37,33,30,30,25,15,8,11,10,X,10,9,11,10,9,6) (the X reflects the data not recorded from Week 9). It is clear that initially in Week 0 there is a very large number of errors which indicates essentially no tendency toward the final blocking, that in Weeks 1-5 this number of errors decreases sharply, and that from Weeks 6-15 the number of errors is much lower and roughly constant, indicating that equilibrium block structure has been essentially reached, although some individual variability among weeks continues to be present.

We will now try to recapture this evolution in a way which does not explicitly read backward from a blockmodel analysis performed on data in the final week. The Carroll-Chang INDSCAL algorithm is a natural vehicle for making this attempt. Because use of INDSCAL has been almost exclusively restricted to the psychological and marketing literature (e.g., Wish and Carroll, 1973; Carroll, 1973 and references there), we first give a brief restatement of aim of the algorithm.

The basic idea is one of dual scaling. Initially, using the standard psychological interpretation, suppose that one has a group of m subjects who each give a judged proximity matrix among n items. It is desired to place the n items in a single ("stimulus") space reflecting some kind of group (or composite) judgment, and simultaneously to place the m subjects in a second ("subject") space reflecting individual differences among subjects. The very strong and specific hypothesis is now made that subjects differ from one another only through differential weights which they attach to the dimensions of a Euclidean stimulus space having a non-arbitrary orientation. Specifically, given m nxn proximity (similarity) matrices $P_1, P_2, \ldots, P_m$, the idea of INDSCAL is first to convert the matrices P_j into distance matrices D_j by means of a linear transformation and then to find n stimulus vectors

$\underset{\sim}{x}_1 = (x_{1i})_{i=1}^{k}$, $x_2 = (x_{2i})_{i=1}^{k}, \ldots, \underset{\sim}{x}_n = (x_{ni})_{i=1}^{k}$ and m subject vectors

$\underset{\sim}{w}_1 = (w_{1i})_{i=1}^{k}, \ldots, \underset{\sim}{w}_m = (w_{mi})_{i=1}^{k}$ such that in a k-dimensional "modified" Euclidean space, the distance between stimuli r and s, for subject j is:

$$D_j(r,s) = \sqrt{\sum_{i=1}^{k} w_{ji}(x_{ri}-x_{si})^2}.$$

(For a more detailed description giving the exact least-squares target function and nonlinear least-squares fitting procedures, see Carroll and Chang, 1970.) Thus, the obtained vectors $\underset{\sim}{x}_i$ constitute the stimulus space solution and the vectors $\underset{\sim}{w}_j$ constitute the subject space solution. It is to be emphasized that, unlike MDSCAL, this algorithm is a metric scaling procedure, i.e., will not give results invariant under monotone transformation of the input proximity data. The stimulus space solution also comes equipped with a set of preferred axes along with the weights w_{ji}, so that the obtained solution is also not rotation-invariant.

For the present application of the Newcomb data, the stimulus and subject spaces will be given the following nonstandard interpretations:

Standard interpretation	Newcomb data interpretation
Subjects	Weeks
Stimuli	Fraternity members.

No confusion should arise if it is explicitly emphasized that the fraternity members in the Newcomb data are not being treated as analogous to subjects in the INDSCAL input.

The procedure is now as follows. Starting with the raw preference rankings (as reported by Nordlie, 1958), the first step is to convert these data into a form suitable for INDSCAL input. A number of ways of doing this have been explored, but the simplest approach also turns out to give the best results. Specifically, convert each preference matrix for each week j into a matrix of distances among fraternity members r,s,... by setting

$$D_j(r,s) = \frac{1}{2}(P_r(s)+P_s(r)), \tag{AV}$$

where $P_r(s)$ is the preference position assigned to s by r and $P_s(r)$ is the analogous position assigned to r by s (thus both $P_r(s)$ and $P_s(r)$ can assume integral values from 1 to 16 inclusive). In the absence of the Week 9 data, there are then fifteen 17 x 17 matrices D_j thus defined. These are taken as distance matrices for INDSCAL input; the INDSCAL algorithm has been run on this data in each of dimensions k = 4,3, and 2, accounting, respectively, for 64,56, and 45% of the variance.

Figures 20 and 21 illustrate respectively two corresponding two-dimensional projections of the four-dimensional INDSCAL subject-space and stimulus-space solutions. Examining the subject-space solution first, there is clearly a coherent trend across weeks, with the later

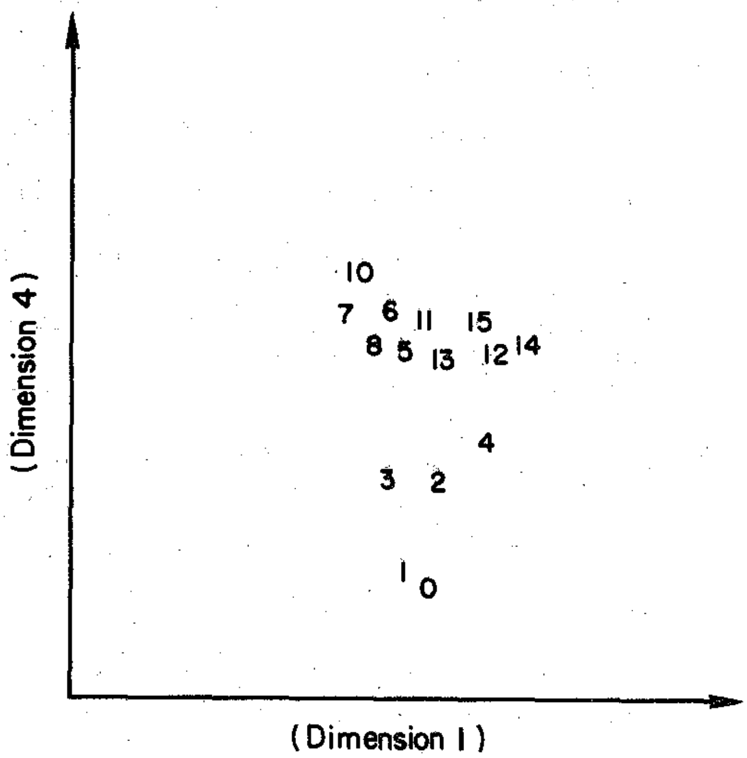


Fig. 20. Subject-space for two-dimensional INDSCAL solution on Newcomb-Nordlie data (Year 2), showing evolution of group structure over the fifteen reported weeks. Plot is obtained from k=4 INDSCAL solution, projecting onto dimensions 1 and 4.

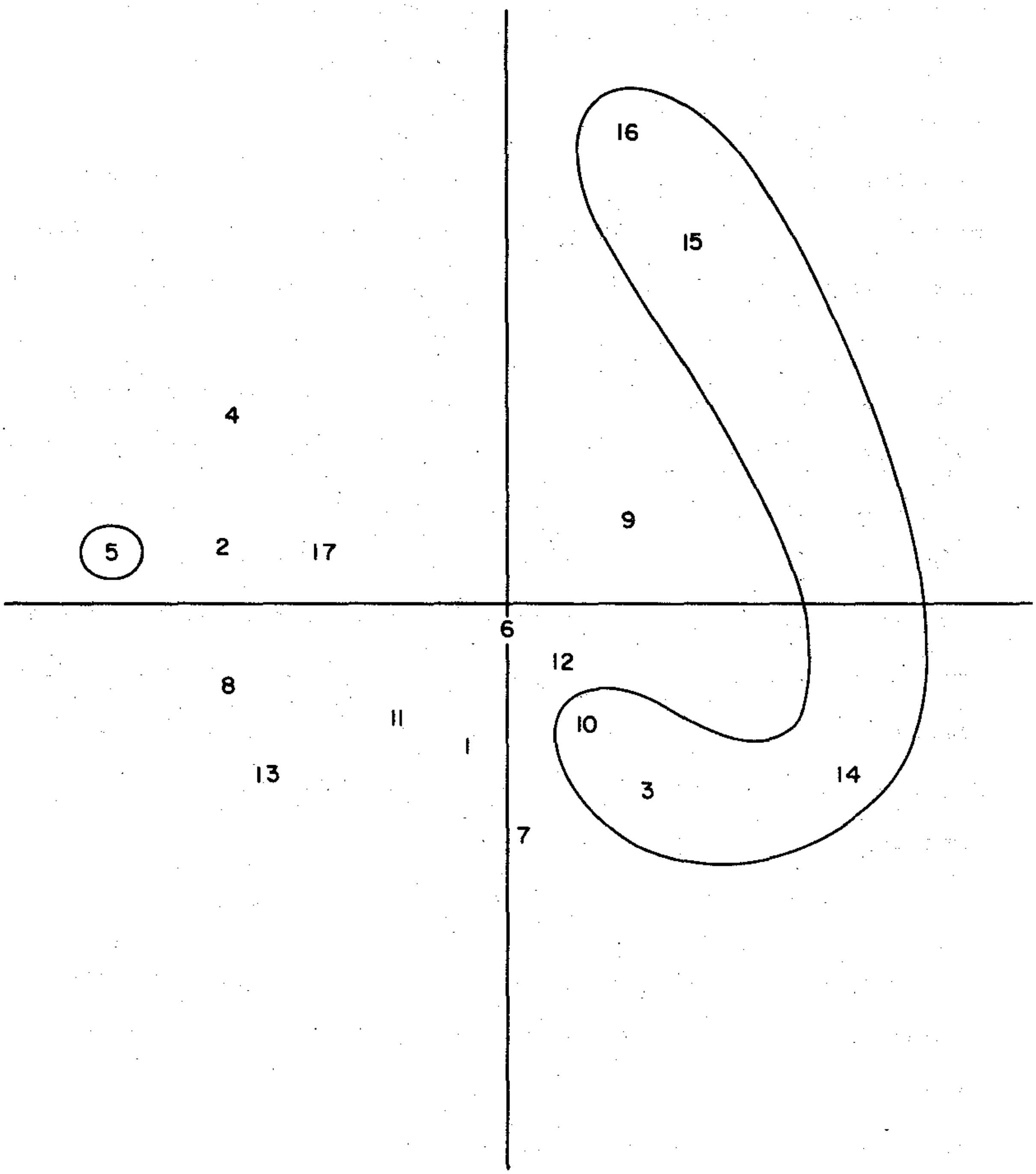


Fig. 21. Stimulus-space INDSCAL solution corresponding to Fig. 20, obtained axes superimposed. Circled points correspond to second (hangers-on) block in Fig. 2 CONCOR solution.

weeks (6-15) being clustered much more tightly than the early ones (0-5). The same separation is also clear for three of the six other two-dimensional projections implied by the four-dimensional subject-space solution; the k=2 INDSCAL subject-space solution shows an analogous pattern, though here the clustering of the later weeks becomes so tight as to make discrimination among these weeks difficult. These positive results are reinforced when one now turns to the stimulus space solution (Fig. 21). This second solution places individual fraternity members in a common two-dimensional Euclidean space. Superimposed on this space, we have indicated the earlier two-block CONCOR division shown in Fig. 2. It is clear that the members of the second CONCOR block (individuals numbered 3,5,10,14,15,16), whom we earlier characterized as hangers-on, are now placed mainly as outlying points in the INDSCAL solution. This placement is consistent with earlier hangers-on interpretations and suggests that INDSCAL is here recovering a kind of center-periphery dimension in polar coordinates.

DISCUSSION

There are two separate topics for summary comments. The first concerns the contribution of the CONCOR algorithm to the blockmodel approach and its relation to other blockmodel analyses. The second topic concerns the comparative merits of blockmodels versus multidimensional scaling approaches to social network data.

As far as the CONCOR algorithm specifically is concerned, the applications we have explored in the present paper show that this algorithm produces results which stand generally in close relation to trial-and-error blockmodels satisfying White's criterion of lean fit. Specifically, the partitionings produced by CONCOR are in general close to a strict lean-fit blockmodel if any such model exists (e.g., see the Fig. 5 comparison of CONCOR with White's analysis). This is true even though the CONCOR algorithm is not explicitly guided by a search for zeroblocks. The CONCOR algorithm hence emerges as a useful way of systematically searching for blockmodels on unexplored raw data. Of course, CONCOR is clearly not the only algorithm which could be used to find blockmodels, and other hierarchical clustering algorithms applied to a first-correlation matrix may in fact produce similar results. In specific comparisons with HICLUS on various data sets, there is evidence that CONCOR performs in a superior way at the four-block level. However, the actual utility of CONCOR cannot be assessed on so narrow a basis. Most importantly, unlike standard hierarchical clustering algorithms such as Johnson's HICLUS (Johnson, 1967), CONCOR admits full exploitation of row-column duality because of the possibility of blocking separately on both rows and columns of rectangular matrices. While we have not emphasized these alternatives for sociometric data (examples A-C in Part I), the nonsociometric examples D and E make heavy use of this dual blocking possibility. CONCOR therefore emerges as a natural way of unifying algorithmic approaches to the several distinct network-related kinds of social structural data, including committee membership data as well as sociometric data (Breiger, 1974).

In most data investigations, it is reasonable and desirable that both the strict zeroblock criterion and the CONCOR algorithm should be independently applied. The search for blockmodels which are strict lean fits to given data is greatly facilitated by an unpublished algorithm due to G. H. Heil. This algorithm takes as input a given blockmodel (e.g., Fig. 1c) and given data (e.g., Fig. 1a), and produces as output a list of all (if any) permutations of the original data which conform to the proposed blockmodel in the lean fit sense (e.g., Fig. 1b). This algorithm will be described in detail elsewhere (see Heil and White, 1974). One extremely valuable feature of the Heil algorithm, which is not shared by CONCOR, is the light which it is able to cast on nonuniqueness of blockmodel solutions. There is no question that many data sets possess some inherent ambiguity; we have already run across cases of such ambiguity in the presence of "interstitial" men in the Sampson data (p. 34 above). Bringing out this ambiguity is clearly not a task which can be accomplished by a single algorithm like CONCOR producing a unique solution. It is also very interesting that one may be able to obtain partitionings identical to CONCOR by directly applying the Heil algorithm to raw data under an appropriately chosen blockmodel "hypothesis." Developments along this last line are pursued in White and Breiger (1974).

Next, there is the problem of assessing the scaling analyses in Part II. The result of applying MDSCAL-5 M to the Homans and Sampson first-correlation matrices is impressive (and especially so in the light of the Homans and Sampson analyses) and is also in excellent agreement with the output of CONCOR on the same matrix. This suggests that MDSCAL of a first-correlation matrix is a valuable probe into a concretely presented social structure. This way of applying MDSCAL appears new and supplements the use of more classical techniques like factor analysis (e.g., MacRae, 1960).[15]

This consideration leads to a very important additional point. The most interesting substantive results of the present paper have been obtained when we have returned to the original raw data and imposed on this data the row and column permutations implied by a CONCOR blocking (e.g., Figs. 6 and 10). This feedback to underlying relational data is a distinctive feature of blockmodel analysis which is not shared by scaling procedures. The ultimate aim of blockmodel analysis is to analyze the network of relations among blocks; in fact, blocks are defined in the first place through reference to such a network. In this sense, it is actually misleading to speak of blockmodels in terms of structural equivalence of individuals alone: blockmodels imply equivalence of individuals in the same block, but at the same time also imply networks of relations among blocks.[16] By contrast, the aim of the scaling applications is to recreate as much as possible of a social structure in a Euclidean space (more generally, in a Minkowski r-space), dispensing with the original network structure and substituting a more familiar spatial one.[17]

Finally, we should again stress the complementarity between the two modes of analysis. Scalings obtained as in Figs. 17-21 explicitly lose track of network structure, but bring out the geometry of structural position in a much richer way than is possible through any clustering technique (e.g., by use of CONCOR). Blockmodel analyses are inherently restricted to clusterings, but make use of these clusterings to extract direct information out of raw network structure.

APPENDIX

Numerical Studies of the Similarity of CONCOR to Johnson's Connectedness and Diameter Methods in Two Data Cases

The present appendix gives a combinatorial approach to the problem of comparing CONCOR with the two methods of Johnson's HICLUS algorithm (connectedness and diameter methods). Specifically, we view the output of each hierarchical clustering method as a binary tree and we apply one of the tree distances $\beta(T_1, T_2)$ developed in Boorman and Olivier (1973).

One difference between CONCOR and HICLUS is the absence of any valuation of levels in CONCOR trees analogous to cluster values α in Johnson's procedure (see also Fig. 7). In the terminology of Boorman and Olivier (1973), the output of CONCOR is hence a bare tree, whereas the HICLUS methods lead to valued trees. In order to compare bare to valued trees, either of two strategies may be followed. On the one hand, there are various possible procedures for converting a bare tree into a valued tree, e.g., by assigning a value to each node which is the size of the corresponding subtree. Alternatively, it is possible to treat any valued tree as bare by simply disregarding the associated cluster values. We presently follow the latter approach as the less artificial strategy.

Given any bare binary tree (e.g., as represented in Figs. 3, 9, etc.) one may equivalently represent the tree as the collection of all its node sets, i.e., sets of items falling under some given node. Thus in the Bank Wiring tree (S1, W4, I1) and (W1, W3, S1, W4, I1) are node sets, whereas (S1, W4, W3) is not. Notice that the full structure of the original tree may be recovered from the collection of all its node sets, so that no information is lost in passing from the original tree structure to this set of sets.

Any binary tree on n items will then lead to a collection of n-2 subsets, where without loss of information one ignores both the trivial node set consisting of all items and the singleton sets formed by taking each item alone (i.e., the highest and lowest levels of hierarchical clustering).

Then one may define as follows a distance $\beta(T_1,T_2)$ between any two bare trees:

<u>Definition</u> (=Definition 1.1 in Boorman and Olivier [1973:29]).

$\beta(T_1,T_2) = \min_f \sum_{i=1}^{n-2} | W_i \Delta X_{f(i)} |$, where $\{W_i\}_{i=1}^{n-2}$ is the collection of node sets formed from T_1, $\{X_i\}_{i=1}^{n-2}$ is the analogous collection formed from T_2, and f is a permutation of the first n-2 integers. Here Δ represents the operation of forming the symmetric difference between two sets (i.e., the set of elements contained in one or the other, but not in both), and $|\ |$ denotes the size of a set.

The distance β may be shown to have various desirable properties, and in particular is a metric. The definition of β represents a special case of a very general principle which may be employed to define structural distances in many situations (Boorman, 1970). In general, the computation of $\beta(T_1,T_2)$ reduces to an optimal assignment problem (Ford and Fulkerson, 1962), but in simple cases the optimal assignment may be readily computed without recourse to a linear programming algorithm.

We now apply the metric $\beta(T_1,T_2)$ to the present problem. Figures A1 and A2 show respectively the node sets obtained through each of the three methods (CONCOR, diameter, connectedness) on the Bank Wiring data and the Sampson data respectively. Both figures are presented in such a way that identical clusters fall on the same line.

To compute $\beta(T_1,T_2)$ between any pair of methods in Figs. A1-A2, it is only necessary to find an optimum correspondence between clusters not produced by both methods. Figures A3-A4 show calculation of the optimum correspondences for the two data sets. Given the correspondence, calculation of $\beta(T_1,T_2)$ is then immediate; the results are also reported in Figs. A3-A4.

The result of these calculations shows that there is no simple relation among the three methods. In the Bank Wiring case, CONCOR is more similar to both HICLUS methods than either of these methods is to the other. Of the two methods, CONCOR is more similar to the connectedness method. Taken alone, this result is evidence for placing CONCOR in an intermediate position on a diameter-connectedness continuum, hence following the classificatory strategy of Jardine and Sibson (1971) and paralleling the intermediate position on such a continuum of various other clustering methods (e.g., Sokal and Michener [1958]; Hubert [1972]). On the other hand, this situation is reversed in the case of the Sampson data. Here the two HICLUS methods are actually closer to one another than CONCOR is to the connectedness method. In this second case, therefore, the relevancy of the diameter-connectedness continuum proposed by Jardine and Sibson quite clearly breaks down. Also, this result helps to alleviate suspicions that CONCOR may in general behave

Fig. A1. Clusters produced by three algorithms (CONCOR, diameter method, connectedness method) on the Bank Wiring data. Trivial clusters (consisting of single individuals or the entire population) are not recorded; since these clusters are produced by all methods, they do not affect computation of $\beta(T_1,T_2)$. Identical clusters are placed on the same line.

CONCOR	Diameter method	Connectedness method
C1: (W1,W3)	D1: (W1,W3)	K1: (W1,W3)
C2: (S1,W4)	D2: (S1,W4)	K2: (S1,W4)
C3: (S1,W4,I1)		
C4: (W1,W3,S1,W4,I1)		K3: (W1,W3,S1,W4,I1)
C5: (W2,W5)	D3: (W2,W5)	K4: (W2,W5)
C6: (W2,W5,I3)		K5: (W2,W5,I3)
C7: (W1,W3,S1,W4,I1, W2,W5,I3)		
C8: (W6,S2)		
C9: (W8,W9)	D4: (W8,W9)	K6: (W8,W9)
C10: (W7,W8,W9)	D5: (W7,W8,W9)	K7: (W7,W8,W9)
C11: (W7,W8,W9,S4)		K8: (W7,W8,W9,S4)
C12: (W6,S2,W7,W8,W9,S4)		
	D6: (W2,W5,I1)	
	D7: (W1,W3,S1,W4)	K9: (W1,W3,S1,W4)
	D8: (W1,W3,S1,W4,W2,W5,I1)	
	D9: (S2,I3)	
	D10: (W6,S4)	K10: (W7,W8,W9,S4,W6)
	D11: (W7,W8,W9,W6,S4)	
	D12: (W7,W8,W9,W6,S4,S2,I3)	K11: (W2,W5,I3,S2)
		K12: (I1,W1,W3,S1,W4, S2,I3,W2,W5)

Fig. A2. As Fig. A1, for the Sampson monastery data.

CONCOR	Diameter method	Connectedness method
C1: (1,7)		
C2: (2,15)	D1: (2,15)	K1: (2,15)
C3: (1,7,2,15)	D2: (1,7,2,15)	
C4: (12,14)	D3: (12,14)	
C5: (12,14,16)	D4: (12,14,16)	
C6: (1,7,2,15,12,14,16)	D5: (1,7,2,15,12,14,16)	K2: (1,7,2,15,12,14,16)
C7: (17,18)	D6: (17,18)	K3: (17,18)
C8: (3,17,18)	D7: (3,17,18)	K4: (3,17,18)
C9: (1,7,2,15,12,14,16,3,17,18)		
C10: (4,6)		
C11: (4,6,8)		
C12: (10,11)		
C13: (4,6,8,10,11)		
C14: (5,9)	D8: (5,9)	
C15: (5,9,13)		
C16: (4,6,8,10,11,5,9,13)		
	D9: (6,8)	
	D10: (10,6,8)	
	D11: (4,11)	K5: (4,11)
	D12: (5,9,4,11)	
	D13: (5,9,4,11,10,6,8)	K6: (5,9,4,11,10,6,8)
	D14: (7,2,15)	
	D15: (13,3,17,18)	K7: (13,3,17,18)
	D16: (1,7,2,15,16,12,14,13,3,17,18)	K8: (1,7,2,15,16,12,14,13,3,17,18)
		K9: (9,4,11)
		K10: (10,9,4,11)
		K11: (6,10,9,4,11)
		K12: (8,6,10,9,4,11)
		K13: (2,15,16)
		K14: (2,14,15,16)
		K15: (1,2,14,15,16)
		K16: (7,1,2,14,15,16)

Fig. A3. Computation of optimal assignment between distinct clusters produced by the different methods on the Bank Wiring data. Clusters referred to in notation of Fig. A1. An optimal assignment (not necessarily unique) pairs corresponding columns and rows, e.g. (in [a]) C3 to D9, C4 to D7, etc. $\beta(T_1,T_2)$ is hence given by the trace $T = \sum_i a_{ii}$ for each of the interger-valued matrices shown.

(a) CONCOR-diameter method

	D9	D7	D6	D8	D10	D11	D12
C3	5	3	4	4	5	8	10
C4	7	1	6	2	7	10	12
C6	3	7	2	6	5	8	8
C7	8	4	5	1	10	13	13
C8	2	6	5	9	2	5	5
C11	6	8	7	11	4	1	3
C12	6	10	9	13	4	1	1

β(CONCOR, diameter) = 13

(b) CONCOR-connectedness method

	K9	K12	K11	K10
C3	3	6	7	8
C7	4	1	6	13
C8	6	9	4	5
C12	10	13	8	1

β(CONCOR, connectedness) = 9

(c) Diameter method-connectedness method

	K5	K12	K11	K3	K8
D6	2	6	3	6	7
D8	6	2	7	2	11
D9	3	7	2	7	6
D10	5	11	6	7	4
D12	8	12	7	12	3

β(diameter, connectedness) = 16

Fig. A4. As Fig. A3, for the Sampson monastery data. Notation for clusters follows Fig. A2.

(a) CONCOR-diameter method

	D14	D16	D11	D9	D12	D10	D15	D13
C1	3	8	4	4	6	5	6	9
C9	7	2	12	12	14	13	8	17
C10	5	12	2	2	4	3	6	5
C11	6	13	3	1	5	2	7	4
C12	5	12	2	4	4	3	6	5
C13	8	15	3	3	5	2	9	2
C15	6	11	5	5	3	6	5	6
C16	11	16	6	6	4	5	10	1

β(CONCOR, diameter) = 20

(b) CONCOR-connectedness method

	K15	K16	K14	K13	K8	K5	K11	K10	K12	K9	K7	K6
C1	5	4	6	5	9	4	7	6	8	5	6	9
C3	3	2	4	3	7	6	9	8	10	7	8	11
C4	5	6	4	5	9	4	7	6	8	5	6	9
C5	4	5	3	4	8	5	8	7	9	6	7	10
C9	5	4	6	7	1	12	15	14	16	13	8	17
C10	7	8	6	5	13	2	3	4	4	3	6	5
C11	8	9	7	6	14	3	4	5	3	4	7	4
C12	7	8	6	5	13	2	3	2	4	3	6	5
C13	10	11	9	8	16	3	2	3	1	4	9	2
C14	7	8	6	5	13	4	5	4	6	3	6	5
C15	8	9	7	6	12	5	6	5	7	4	5	6
C16	13	14	12	11	17	6	3	4	2	5	10	1

β(CONCOR, connectedness) = 34

(c) Diameter method-connectedness method

	K16	K14	K15	K9	K12	K11	K10	K13
D2	2	4	3	7	10	9	8	3
D3	6	4	5	5	8	7	6	5
D4	5	3	4	6	9	8	7	4
D8	8	6	7	3	6	5	4	5
D9	8	6	7	5	4	5	6	5
D10	9	7	8	6	3	4	5	6
D12	10	8	9	1	4	3	2	7
D14	3	3	4	6	9	8	7	2

β(diameter, connectedness) = 25

quite similarly to the connectedness method, and in particular that CONCOR may be prone to similar difficulties of a "chaining" type (see also above, pp. 45-46).

Of course, all results based on *a priori* metrics do not take account of substantive features of particular data sets, and hence have limitations for this reason. Also, there is as yet no developed distribution theory for the values of tree metrics, which would enable statements about levels of significance to be made. Ling (1971) presents results which constitute a start in this direction. Prior to development of such a theory, only ordinal comparisons among distances between clusterings may be made with any rigor.

REFERENCES

Abelson, R. P., and Rosenberg, M. J. Symbolic psycho-logic: A model of attitudinal cognition. Behavioral Science, 1958, 3, 1-13.

Alba, R. D. A graph-theoretic definition of a sociometric clique. The Journal of Mathematical Sociology, 1973, 3, 113-126.

Arabie, P. Concerning Monte Carlo evaluations of nonmetric scaling algorithms. Psychometrika, 1973, 38, 607-608.

Arabie, P., and Boorman, S. A. Multidimensional scaling of measures of distance between partitions. Journal of Mathematical Psychology, 1973, 10, 148-203.

Arabie, P., and Shepard, R. N. Representation of similarities as additive combinations of discrete, overlapping properties. Paper read at Mathematical Psychology Meetings, Montréal, August, 1973.

Bernard, P. A few sociological considerations on structural equivalence. Unpublished, Department of Sociology, Harvard University, 1971.

Bernard, P. Stratification sociométrique et réseaux sociaux. Sociologie et Sociétés, 1973, 5, 128-150.

Bernard, P. Association and hierarchy: The social structure of the adolescent society. Unpublished Ph.D. thesis, Harvard University, Department of Sociology, 1974.

Bjerstedt, A. Interpretations of sociometric choice status. Lund: C. W. K. Gleerup, 1956.

Boorman, S. A. Metric spaces of complex objects. Unpublished Senior Honors Thesis, Harvard College, Division of Engineering and Applied Physics, 1970.

Boorman, S. A., and Arabie, P. Structural measures and the method of sorting. In R. N. Shepard, A. K. Romney, and S. B. Nerlove (eds.), Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, Vol. 1, New York: Seminar Press, 1972, pp. 225-249.

Boorman, S. A., and Olivier, D. C. Metrics on spaces of finite trees. Journal of Mathematical Psychology, 1973, 10, 26-59.

Boyd, J. P. The algebra of group kinship. Journal of Mathematical Psychology, 1969, 6, 139-167.

Breiger, R. L. The duality of persons and groups. Social Forces, 1974 (in press).

Carroll, J. D. Models and algorithms for multidimensional scaling, conjoint measurement and related techniques. Appendix B of P. E. Green and Y. Wind, Multi-attribute Decisions in Marketing. New York: Holt, Rinehart and Winston, 1973.

Carroll, J. D., and Chang, J. J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckhart Young" decomposition. Psychometrika, 1970, 35, 283-319.

Clark, J. A., and McQuitty, L. L. Some problems and elaborations of iterative, intercolumnar correlational analysis. Educational and Psychological Measurement, 1970, 30, 773-784.

Davis, A., Gardner, B. B., and Gardner, M. R. Deep south: A social anthropological study of caste and class. Chicago: University of Chicago Press, 1941.

Davis, J. A. Clustering and structural balance in groups. Human Relations, 1968, 20, 181-187.

Edwards, A. W. F., and Cavalli-Sforza, L. L. A method for cluster analysis. Biometrics, 1965, 21, 362-375.

Fararo, T. C. Mathematical Sociology. New York: Wiley, 1973.

Flament, C. Applications of graph theory to group structure. Englewood Cliffs, N. J.: Prentice-Hall, 1963.

Ford, L. R., Jr., and Fulkerson, D. R. Flows in Networks. Princeton, N. J.: Princeton University Press, 1962.

Glanzer, M., and Glaser, R. Techniques for the study of group structure and behavior. I. Psychological Bulletin, 1959, 56, 317-332.

Granovetter, M. The strength of weak ties. American Journal of Sociology, 1973, 78, 1360-80.

Griffith, B. C., Maier, V. L., and Miller, A. J. Describing communications networks through the use of matrix-based measures. Unpublished mimeo., Graduate School of Library Science, Drexel University, 1973.

Hartigan, J. A. Direct clustering of a data matrix. Journal of the American Statistical Association, 1972, 67, 123-129.

Heil, G. H., and White, H. C. An algorithm for finding homomorphic correspondences between given multi-graphs. Unpublished, Department of Sociology, Harvard University, 1974.

Holland, P. W., and Leinhardt, S. Masking: The structural implications of measurement error in sociometric data. Unpublished, Department of Statistics, Harvard University, 1969.

Holland, P. W., and Leinhardt, S. A method for detecting structure in sociometric data. American Journal of Sociology, 1970, 76, 492-513.

Holman, E. W. The relation between hierarchical and Euclidean models for psychological distances. Psychometrika, 1972, 37, 417-423.

Homans, G. C. The human group. New York: Harcourt Brace, 1950.

Horan, C. B. Multidimensional scaling: Combining observations when individuals have different perceptual structures. Psychometrika, 1969, 34, 139-165.

Hubbell, C. H. An input-output approach to clique identification. Sociometry, 1965, 28, 377-399.

Hubert, L. Some extensions of Johnson's hierarchical clustering algorithms. Psychometrika, 1972, 37, 261-274.

Hubert, L. Monotone invariant clustering procedures. Psychometrika, 1973, 38, 47-62.

Jardine, N., and Sibson, R. Mathematical Taxonomy. New York: Wiley, 1971.

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

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

Klahr, D. A Monte Carlo investigation of the statistical significance of Kruskal's nonmetric scaling procedure. Psychometrika, 1969, 34, 319-330.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 1964a, 29, 1-28.

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. Psychometrika, 1964b, 29, 115-159.

Kruskal, J. B., and Hart, R. E. A geometric interpretation of diagnostic data for a digital machine: Based on a Morris, Illinois Electronic Central Office. Bell System Technical Journal, 1966, 45, 1299-1338.

Lance, G. N., and Williams, W. T. A general theory of classificatory sorting strategies. 1. Hierarchical systems. The Computer Journal, 1967a, 9, 373-380.

Lance, G. N., and Williams, W. T. A general theory of classificatory sorting strategies. 2. Clustering systems. The Computer Journal, 1967b, 10, 271-277.

Laumann, E. O. Bonds of pluralism. New York: Wiley Interscience, 1973.

Laumann, E. O., and Pappi, F. U. New directions in the study of community elites. American Sociological Review, 1973, 38, 212-230.

Laumann, E. O., Verbrugge, L. M., and Pappi, F. U. A causal modelling approach to the study of a community elite's influence structure. American Sociological Review, 1974, 39, 162-174.

Levine, J. H. The sphere of influence. American Sociological Review, 1972, 37, 14-27.

Ling, R. F. Cluster analysis. Unpublished Doctoral Dissertation, Department of Statistics, Yale University, 1971.

Lorrain, F. P. Réseaux sociaux et classifications sociales. Paris: Éditions du Hermann, (in press). English language version: Doctoral thesis, Harvard University, Department of Sociology, 1972.

Lorrain, F. P., and White, H. C. Structural equivalence of individuals in social networks. Journal of Mathematical Sociology, 1971, 1, 49-80.

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

MacRae, D. Direct factor analysis of sociometric data. Sociometry, 1960, 23, 360-371.

McQuitty, L. L. Multiple clusters, types, and dimensions from iterative intercolumnar correlation analysis. Multivariate Behavioral Research, 1968, 3, 465-477.

McQuitty, L. L., and Clark, J. A. Clusters from iterative, intercolumnar correlational analysis. Educational and Psychological Measurement, 1968, 28, 211-238.

McQuitty, L. L., and Clark, J. A. Some problems and elaborations of iterative, intercolumnar correlational analysis. Educational and Psychological Measurement, 1970, 30, 773-784.

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

Mosteller, F. Association and estimation in contingency tables. Journal of the American Statistical Association, 1968, 63, 1-28.

Nadel, S. F. The theory of social structure. London: Cohen and West, 1957.

Needham, R. M. Applications of the theory of clumps. Mechanical Translation, 1965, 8, 113-127.

Newcomb, T. M. The acquaintance process. New York: Holt, Rinehart and Winston, 1961.

Newcomb, T. M. Interpersonal balance. In R. P. Abelson _et al._ (eds.), _Theories of cognitive consistency: A sourcebook._ Chicago: Rand McNally, 1968, pp. 28-51.

Nordlie, P. G. A longitudinal study of interpersonal attraction in a natural group setting. Unpublished Ph.D. Thesis, University of Michigan, 1958.

Robinson, A. _Introduction to model theory and to the metamathematics of algebra._ Amsterdam: North-Holland, 1965.

Roethlisberger, F. J., and Dickson, W. J. _Management and the worker._ Cambridge, Mass.: Harvard University Press, 1939.

Romney, A. K. Measuring endogamy. In P. Kay (ed.), _Explorations in mathematical anthropology._ Cambridge, Mass.: MIT Press, 1971, pp. 191-213.

Sampson, S. F. Crisis in a cloister. Unpublished Ph.D. Thesis, Cornell University, 1969.

Schoenfield, J. R. _Introduction to mathematical logic._ Reading, Mass.: Addison-Wesley, 1967.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. _Psychometrika_, 1962a, _27_, 125-140.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. _Psychometrika_, 1962b, _27_, 219-246.

Shepard, R. N. Psychological representation of speech sounds. In E. E. David, Jr., and P. B. Denes (eds.), _Human Communication: A unified view._ New York: McGraw-Hill, 1972, pp. 67-113.

Simmel, G. (Trans. by K. H. Wolff and R. Bendix). Conflict and the web of group-affiliations. New York: Free Press, 1955.

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

Sokol, R. R., and Michener, C. D. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, 1958, 38, 1409-1438.

Sorenson, 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, 1-34.

Stenson, H. H., and Knoll, R. L. Goodness of fit for random rankings in Kruskal's nonmetric scaling procedure, Psychological Bulletin, 1969, 71, 122-126.

Struhsaker, T. T. Social structure among vervet monkeys (Cercopithecus aethiops). Behavior, 1967, 29, 83-121.

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

White, H. C. The anatomy of kinship. Englewood Cliffs, N. J.: Prentice-Hall, 1963a.

White, H. C. Notes on finding models of structural equivalence: Drawing on theories of roles, duality, sociometry and balance. Unpublished, Department of Sociology, Harvard University, 1969.

White, H. C. Chains of opportunity. Cambridge, Mass.: Harvard University Press, 1970.

White, H. C. Equations, patterns and chains in social structure. Unpublished, Department of Sociology, Harvard University, 1973.

White, H. C. Multiple networks in small populations. II. Compound relations and equations. Unpublished, Department of Sociology, Harvard University, 1974a.

White, H. C. Models for interrelated roles from multiple networks in small populations. In P. J. Knopp and G. H. Meyer (eds.), Proceedings of the Conference on the Application of Undergraduate Mathematics in the Engineering, Life, Managerial and Social Sciences. Atlanta: Georgia Institute of Technology Press, 1974b.

White, H. C., and Breiger, R. L. Multiple networks in small populations. I. Blockmodels. Unpublished, Department of Sociology, Harvard University, 1974.

Wiese, L. von (ed. and annotated by F. Mueller). Sociology. New York: Oskar Priest, 1941.

Wish, M., and Carroll, J. D. Applications of "INDSCAL" to studies of human perception and judgment. In E. C. Carterette, and M. P. Friedman (eds.). Handbook of perception, Vol. 2. New York: Academic Press, 1973.

FOOTNOTES TO TEXT

[1] The algorithm was initially suggested by the empirical discovery of convergence of iterated correlations (see below, p. 12) on network data reporting contacts among research scientists in an emerging biomedical specialty area (described in Griffith, Maier, and Miller [1973]). Subsequently, Dr. Tragg of the University of Surrey pointed out that work constituted an independent rediscovery of the "iterative, intercolumnar correlation analysis" proposed by McQuitty and his co-workers (McQuitty, 1968; McQuitty and Clark, 1968; Clark and McQuitty, 1970). See text for further discussion.

[2] A closely related view is expressed by Needham (1965:117):

> The moral of this is that we should not look for an "internal" definition of a cluster, that is, one depending on the resemblance of the members to each other, but rather for an "external" definition, that is, one depending on the non-resemblance of members and non-members.

Translating "resemblance" into "presence of network ties," it is clear that the idea here is very similar to the present conception.

[3] I.e., if $\underset{\sim}{x} = (x_i)_{i=1}^{n}$, $\underset{\sim}{y} = (y_i)_{i=1}^{n}$,

$$r(\underset{\sim}{x},\underset{\sim}{y}) = \frac{\underset{\sim}{x}' \circ \underset{\sim}{y}'}{\|\underset{\sim}{x}'\| \ \|\underset{\sim}{y}'\|}$$

where $\underset{\sim}{x}' = (x_i - \bar{x})_{i=1}^{n}$, $\bar{x} = \frac{1}{n} \Sigma x_i$, etc., and $\circ$ and $\| \ \|$ denote the Euclidean inner product and norm, respectively. If $\underset{\sim}{x}'$ or $\underset{\sim}{y}' = \underset{\sim}{0}$ then $r(\underset{\sim}{x},\underset{\sim}{y})$ is formally undefined, which gives rise to certain exceptions to the basic convergence fact (2-BLOCK).

[4]McQuitty and Clark (1968) attempt to give a formal proof of the convergence, but their argument does not appear to be rigorous and gives little information on the mathematical behavior of the algorithm.

[5]The knife-edge character of the exceptions was pointed out and investigated by Joseph Schwartz. Clark and McQuitty (1970) report certain exceptions to the convergence; additional classes of exceptions have also been communicated to us by Ingram Olkin of the Stanford Department of Statistics (personal communication).

[6]A second formal class of exceptions which should also be noted occurs when M_0 is taken to be of the form $M_0(i,j) = \frac{c_i d_j}{N}$, where $\sum_i c_i = \sum_j d_j = N$ (i.e., where M_0 corresponds to the standard null hypothesis of row-column independence in a contingency table). Then forming correlations either between rows or between columns, one obtains $M_1(i,j) = 1$ for all i and j and it is clear that statement (2) fails.

[7]In principle, the semigroup (White, 1969) and category-functor (Lorrain and White, 1971) approaches to the algebraic analysis of social networks also give an important place to simultaneous treatment of multiple types of tie. However, existing computational methods do not easily extend to handle more than two distinct relations simultaneously. As a result, for many applications it is necessary to aggregate quite substantially before applying the algebra.

[8]White (1974b) also reports a more refined five-block model of the same data. White and Breiger (1974) develop a three-block model which is a

refinement of the two-block model in the text, viz. (13,9,17,1,8,6,4), (7,11,12,2), (14,3,10,16,5,15). This three-block model is obtained by using the Heil enumeration algorithm (see p. 52 below) rather than CONCOR, and hence provides an interesting check on the CONCOR solution.

[9] Letter notation for 2 x 2 blockmodels follows conventions adopted by White (1974b, Table 1).

[10] White (1974b) actually presents two blockmodels for the Homans data. We discuss only his model which closely resembles our own. See White (in press) for a discussion of the substantive differences between his two models of the Bank Wiring group data.

[11] The "Trading Jobs" matrix is also not symmetric (in fact, it is asymmetric) but the tie density is very low (number of entries = 7; see Homans [1950:67]) and hence this relation is little help in clarifying status relations among groups.

[12] The two methods are also referred to in the literature by a wide variety of other terms. The diameter method is also referred to as the compactness or minimum method (Johnson, 1967), the furthest-neighbor method (Lance and Williams, 1967a), and the complete-link method (Jardine and Sibson, 1971). Similarly, the connectedness method is also referred to as the minimum method (Johnson, 1967), nearest-neighbor method (Lance and Williams, 1967a), and single-link method (Jardine and Sibson, 1971). The terminological jungle is a nuisance.

[13] There are some slight variants in procedure. For example, Katz (1947) proposes to leave out any mutual choices between two individuals i and j when correlating their positions in data given by a standard

positive-choice sociometric procedure. Obviously, this modification will make little effective difference if the group is of any size.

[14] It is noteworthy that the configuration (in Fig. 17) corresponding to the lowest stress value of .126 was by no means the first obtained in the series of 20 different initial configurations. (In fact, Fig. 17 was the thirteenth obtained solution; the twelfth solution had yielded a stress of .321.) The value of .126 for a two-dimensional solution with 14 stimuli is, of course, quite respectable according to Klahr's (1969) Monte Carlo study. However, arguments have been advanced elsewhere (Arabie, 1973) as to why the values in that Monte Carlo study (which, along with that of Stenson and Knoll [1969] gives the most useful data currently available) are inflated, owing to unfortunate properties of Kruskal's L-configuration.

[15] It is worth noting, however, that MDSCAL (as also INDSCAL) is an expensive technique by virtually any measure, especially in the light of the initial configuration problems discussed in the preceding footnote. One major practical side of CONCOR(shared, of course, with many other hierarchical clustering methods) is that it is cheap and extremely easy to implement.

[16] Note, however, that in introducing blockmodels one is explicitly decoupling structural equivalence from the idea of compounding or concatenating social relationships (contrast White, 1963; Lorrain and White, 1971; also White, 1970; Boyd, 1969). This is the major substantive break between blockmodels and the earlier algebraic approaches to social network analysis represented by work of White,

Lorrain, Boyd, and other investigators.

[17]For a derivation of the relation between Euclidean distance models (e.g., the MDSCAL solutions presented here) and hierarchical representations such as Johnson's methods, see Holman (1972).