Informant Accuracy in Social Network Data IV: A Comparison of Clique-Level Structure in Behavioral and Cognitive Network Data

H. Russell Bernard*
University of Florida
Peter D. Killworth
Cambridge University
Lee Sailer
University of Pittsburgh

This paper examines whether clique-structure in cognitive data (i.e. recall of who one talks to) may be used as a proxy for clique-structure in behavioral data (i.e. who one actually talks to). The answer to this question is crucial to much of sociometric and social net-theoretic studies of social structure.

We analysed the clique structures of the communication patterns of four naturally occurring groups of sizes 34 to 58, whose actual communications could easily be monitored, together with the groups' perceptions of their communications. The groups used were: radio hams, a college fraternity, a group of office workers, and an academic department. The analysis used clique-finding, block-modelling, and factor-analytic techniques, all employed in such a way as to maximize the accuracy of the cognitive data.

After defining a way to compare clique structures between behavioral and cognitive data, we found that there was no useful relationship between the two, and furthermore there was no significant difference in performance between any of the structure-finding algorithms.

We conclude that cognitive data may not be used for drawing any conclusions about behavioral social structure.

1. Introduction

Sociometric data are typically collected by asking people to state their preference for others in a group, or to recall their interactions with others. In either case, such data are collected in order to investigate the 'social structure' of a particular group. Social structure is assumed to be built up out of the interaction of the members of the group. Then, the plausible leap is made whereby the answers to the sociometric question reflect the pattern of interactions.

*Address all correspondence to Professor H. R. Bernard, Chair, Department of Anthropology, University of Florida, Gainesville, Fla. 32611, U.S.A. This work was supported under Office of Naval Research Contract N000014-75-C-0441-P00001, Code 452. The opinions expressed in this paper are those of the authors and do not necessarily reflect the position of the supporting agency.
In spite of a vast literature showing that what people say is quite often an unreliable proxy for what they do (see Deutscher, 1972 for a review), the response in sociometry or social network studies has been remarkably unsatisfactory. Clearly, faced with an untrustworthy instrument for gathering data one should either change or improve the instrument, or one should find its error bounds. Unfortunately, there do not seem to be any alternatives to asking people about their relationships; gathering reliable, long-term behavioral data is not practical in most situations. We can expect, therefore, that network data will continue to be collected by interviewing respondents. (Scientific citation and interlocking directorate studies, etc., are exceptions.) To be sure, there are moves towards collection of ‘cleaner’ data in sociometry (e.g. Hallinan, 1974), but these still involve asking people things. There is as yet no evidence that such methods do produce more accurate data. Another approach, pioneered by Holland and Leinhardt (1975), has been to assume that sociometric data are noisy, and to search for signals in the data by statistical methods. These approaches are attempts to improve the data-gathering or data-processing instrument.

We have approached the problem by attempting to find the accuracy (i.e. determine the error bounds) of the instruments by which network data are acquired. At its simplest level, network data are ‘accurate’ if, when i says he talked to j by some amount, then he did. We examined this possibility, with varying degrees of stringency, and found (not surprisingly) that he didn’t. We tried a variety of tests over a variety of data sets for which we had data on both behavior and reports of behavior (cognition) and never found more than a mild correspondence between them. (Overall, people report their dyadic communications inaccurately more than half the time.) In other words, one can not impute any level of interaction between i and j from what it says about his interaction with j. (See Killworth and Bernard, 1976, and Bernard and Killworth, 1977, for details.)

Of course, social network data are almost never analysed purely on the dyadic level. One is, after all, interested in structure, something not easily elucidated by counting dyads. When respondents report their interactions with others in a group, they may have in mind a far more complex and subtle framework than mere dyadic relations. This could have led us to count such relationships as ‘inaccurate’, instead of being slightly fuzzy representations of the behavioral structure.

A thorough examination of triadic level structure, however, showed that a) both behavioral and cognitive data are highly structured; but b) the structures are produced by totally different sets of triads. In particular, any given triad is inaccurately reported 76% of the time (Killworth and Bernard, 1978).

We are now forced to turn to the next higher level of group structure, the ‘clique’. Cliques are typically obtained by applying some algorithm to cognitive data. The assumption is that the cliques found in the cognitive data are those which would be found if one had corresponding behavioral data. It is quite possible that i states that he talked to j and k, when in fact he talked to l and m. This would produce great inaccuracy on both the dyadic and triadic
levels of structure. But, if i, j, k, l, and m form a clique, then i's report is a reflection of his interaction with that clique, though not its members. Thus a good clique-finding algorithm would be one which puts i, j, k, l, and m into a clique when applied either to cognitive or behavioral data. In other words, a good clique-finding device should reduce the noise which shows up as informant inaccuracy at the dyadic or triadic levels.

In this paper we compare the clique structures of matched pairs of behavioral and cognitive data about who people talk to in a variety of groups. (A discussion of the data is given in Section 2.) Since there is no agreement about what a clique is, we have used three essentially different and popular approaches: 1) factor analysis (MacRae, 1960); 2) an iterative correlational block modelling technique (CONCOR, see Breiger, Boorman, and Arabie, 1976); and 3) a graph-theoretic approach based on overlap of maximally complete subgraphs (COMPLT, see Alba, 1973).

We feel that the three algorithms we have chosen are representative of the three major trends among clique-finders. (We understand the difference between the term 'block', as commonly used when discussing CONCOR, and the classical term 'clique'. For our purposes here, the distinction is irrelevant.) As will be clear from succeeding sections of this paper, our selection of these three algorithms does not constitute our endorsement of them. Nonetheless, it is important to ascertain which (if any) algorithm produces cognitive cliques most like the behavioral cliques produced by the same algorithm. (In Section 4 we define what we mean by 'most like'.) Of course, we have no idea which, if any, cliques produced by any algorithm are the right ones. Hence, comparisons can only be made within an algorithm.

Our selection of algorithms needs a word of explanation. Initially, we wanted to compare as many different algorithms as possible. The best, and fairest, way to do this would be to provide each author of an algorithm with our data and have him run the data at his own installation. In fact, we offered to do just this. We placed an announcement in Footnotes (Newsletter of the American Sociological Association) in December 1976, in which we asked interested parties to participate directly in the research reported here. Several authors did respond, but eventually declined on various grounds. We selected three basically different kinds of algorithms which could be run on our installation.

One approach we have omitted is the 'input-output' approach (Hubbell, 1965), which assumes that interpersonal links are channels for the transmission of influence. Another is a generalization of factor analysis called 'ADCLUS' (Arabie, 1976). Neither of these has enjoyed widespread use. Of course, use or lack of use of an algorithm has nothing whatever to do with whether or not that algorithm is accurate, in any sense of the word.

Section 2 discusses the data and Section 3 outlines what we did to the data in order to provide the required input for each algorithm. Section 4 discusses our rationale (and some alternatives) for quantitative comparison, and Section 5 treats the results of such a comparison. Section 6 discusses other operations which could be performed on the data. Section 7 discusses the
results in terms of 'informant accuracy' and draws the firm conclusion that
clique analysis performed on cognitive data does not produce a reasonable
proxy for the clique structure of the corresponding behavioral data.

2. The data

Four different groups have been used in this study, each providing cognitive
and behavioral data on frequency of communication. Three of these data
sets have been described in detail elsewhere (Bernard and Killworth, 1977);
one is new to this study. The four sets of data are called Office, Tech, Hams,
and Frat. A brief description of each follows:

1) Office. These data are from a small social science research firm with
45 employees. This group is composed of several research project teams, each
having senior staff, lower-level assistants, clerks, and typists.

Cognitive data were collected from 40 persons; behavioral data were col-
lected from 44 persons. At Time 1, an observer walked through the office on
four consecutive workdays, covering the same ground every 15 minutes for
five hours each working day. He noted every dyadic contact, including those
 contained in $n$-tuple conversations. At Time 2, seven weeks later, the same
observational procedure was followed.

This was relatively unobtrusive observation; the observers were seen by
the participants, but not talked to. Bernard and Killworth (1977) showed
that this relatively unobtrusive observation had no noticeable effect on in-
formant accuracy as compared with totally unobtrusive observation.

Between Times 1 and 2, each participant was given the familiar deck of
cards containing the names of the other participants. They arranged (i.e.,
ranked) the cards from most to least on how often they talked to others in
the office during a normal working day. The question of frequency, amount,
and importance of contact was raised often by the participants (they are,
after all, social science researchers!), but this was deliberately left vague.
They were told to make up their own minds. Because their judgments were
explicitly based on a 'normal working day' the behavioral data from Time 1
and 2 were aggregated here. They do differ significantly, but whether this is
due to day-to-day fluctuation (which we do not define) or to a systematic
time variation in the group cannot be answered easily. The data produced
are called 'Office frequency' and 'Office ranks'.

2) Tech. These data are from a graduate program in technology education
at West Virginia University. The program's faculty, graduate students, and
secretaries are located in three buildings - two converted houses at the bot-
tom of a hill, and a suite of offices on the hill in the main education building
at the university. There are 37 people in the program; three of these are on
full-time field assignment over 100 miles from the university.

For one week a team of observers walked through the office spaces of the
Tech program. They covered the same ground every half hour, and noted all
occurrences of persons in verbal contact. Any two persons in contact were
scored. \( N \)-tuples were scored by counting each dyad. The same comments on obtrusiveness apply as for the Office data.

After a week of observation, each of the 34 persons on the main campus was handed a deck of cards containing the names of all other members of the group, and asked to rank the deck from most to least communication that week. The question was purposely left rather vague; amount, frequency, or importance of communication was not specified. The data produced are called ‘Tech frequency’ and ‘Tech ranks’.

3) Hams. Our third set of data comes from a group of amateur radio operators, commonly called ‘hams’, living in West Virginia, western Pennsylvania, and eastern Ohio. The hams belong to the Monongalia Wireless Association (MWA), which owns and maintains WR8ABM, a 2-meter, FM repeater station. Virtually all the 2-meter communication of the members of the MWA passes through the repeater station (perched atop a convenient hill) which receives local signals and sends them out over a wide area. WR8ABM is used by hams in a 30,000 mile\(^2\) area.

With the cooperation of the MWA, all conversations on WR8ABM were monitored around the clock for 27 days, using a voice-operated relay between a receiver and a tape recorder. By law and by convention, hams identify themselves with their ‘call’, the letters and numbers combination issued by the FCC. Thus, all communicants could be monitored, and the length of their conversations (in minutes) could be recorded. For current purposes, only the frequency of communication was used, for similarity with the rest of the data sets.

A ‘repeater’ allows groups of persons to participate in conversation, so long as only one person speaks at a time. Thus, this data set includes \( n \)-tuple interactions. All dyads were listed separately for the analysis.

At the end of the 27 day monitoring period, a list of 54 users was drawn up. (Eventually, we found a total of 107 users; by the end of the monitoring period, however, we had recorded 54 users who accounted for all but a small fraction of the repeater’s air time. The other 53 calls were mostly casual or transient users.) Each person was mailed a sheet with all 54 calls, and asked to scale them from 0 (no communication) to 9 (a great deal of communication). A total of 44 usable responses were obtained.

Under the rules of the FCC, hams must keep logs of all contacts made on the long-range bands. Many contacts on 2 meters are made from automobiles (hams use the 2-meter band as a substitute for the crowded citizens band), and are mostly local rather than long distance. A few years ago, the FCC ruled that 2-meter contacts need not be logged. Some of the old-timers in our survey, however, continue to keep logs of all their communication. Those who actually used their logs in responding to the scaling device were removed from the data for obvious reasons. The data produced are called ‘Ham frequency’ and ‘Ham scales’.

4) Frat. The final set of data was obtained from a college fraternity in Morgantown, West Virginia. At the time the data were collected, the 58 occupants had been living together for at least three months. Senior stu-
dents had been living in the fraternity for up to three years. The pattern of data collection was similar to Office and Tech: an observer walked through the fraternity every 15 minutes, 21 hours a day, for five days, noting every group in conversation. The aggregate of these data provided behavioral frequency counts for each dyad in the group. At the end of five days, each member of the fraternity gave a rating, from 1 (no communication) to 5 (a great deal of communication) concerning his interaction with each of the other members. The data produced are called ‘Frat frequency’ and ‘Frat scales’.

Some characteristics of the data are given in Table 1. Note that in no case is content of communication part of the data. Obviously, content of communication may be an important factor in determining both interaction and people’s perception of interaction. However, as far as we are aware, content of communication is a topic universally avoided by researchers in social network theory.

3. Data treatment

The three clique-finding algorithms to be used (factor analysis, CONCOR and COMPLT) differ in their data requirements. Most social network data are binary (i.e. ones or zeros, depending on whether the relationship being examined does or does not occur, for each dyad in the group). Indeed, COMPLT is only capable of handling binary data, and always converts continuous data to binary data before analysis. However, both factor analysis (henceforth FACTOR) and CONCOR, while frequently used to handle binary data, are also able to accept valued data. By and large, binary sociometric data are collected by asking people whether or not they talk to a list of specified others, or by asking them to name their ‘n best friends’ or the people with whom they communicate the most. Our data, collected by rankings and scalings, are not of this form. In this paper we shall assume that the people whom i ranks most highly (or those to whom I assigns the highest scale rating) are the people with whom i believes he communicates the most. Therefore, we can convert ranked or scaled data into binary, for use by COMPLT, by choosing a cutoff level, K.

For ranking data this means that those i ranked 1st, 2nd ...Kth are allocated ones, and all others are allocated zeros. For scaled data, this means that those i gave a scale value > K are allocated ones, and all others are allocated zeros. For comparative purposes, it may be necessary to convert behavioral data to binary. This is done in a precisely similar fashion, with a cutoff, L (not necessarily the same as K). Other ways of producing binary data will be discussed in Section 6.

For a given set of behavioral and cognitive data, how should cutoff values be allocated? As in previous papers, these will be allocated in such a way as to maximize the accuracy of the informants. We must stress that the results presented here are the most accurate that a researcher armed only with cog-
Table 1. The data sets used in this study

<table>
<thead>
<tr>
<th></th>
<th>Office</th>
<th>Tech</th>
<th>Hams</th>
<th>Frat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking data</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Scaled data</td>
<td>no</td>
<td>no</td>
<td>yes (1 - 9)</td>
<td>yes (1 - 5)</td>
</tr>
<tr>
<td>Frequency of</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>communication</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>recorded</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group closed</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>naturally</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>simultaneous contacts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size of group</td>
<td>40</td>
<td>37</td>
<td>44</td>
<td>58</td>
</tr>
<tr>
<td>Group type</td>
<td>work</td>
<td>work</td>
<td>recreation</td>
<td>living</td>
</tr>
<tr>
<td>Communication mode</td>
<td>face-to-face</td>
<td>face-to-face</td>
<td>radio</td>
<td>face-to-face</td>
</tr>
<tr>
<td>Method of observation</td>
<td>observer</td>
<td>observer</td>
<td>continuous</td>
<td>observer</td>
</tr>
<tr>
<td>(15 mins)</td>
<td>(30 mins)</td>
<td></td>
<td>(15 mins)</td>
<td></td>
</tr>
<tr>
<td>Content of communication recorded</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

Cognitive data could produce. We are aware that different people interpret scales in different ways. Since a researcher collecting only cognitive data cannot know what these variations are, we shall use uniform cutoffs (one for cognition, and one for behavior) for all informants in a group in the main body of this paper.

These cutoffs are obtained by examining all possible combinations of cognitive and behavioral cutoffs. Then we choose the two cutoffs which minimize the difference between the total amounts of behavioral and cognitive interaction in the resulting binary matrices. This means that the density of links in behavior and cognition is as favorably distributed (i.e. as nearly equal) as possible.

For ranked data, cutoff problems are more difficult, since any ranking cutoff $K$ can be approximated exactly (in terms of linkage density) by a similar cutoff in the behavioral data (i.e. retaining the $K$ people i communicated with most). We have chosen to use identical cutoffs for behavior and cognition, with values chosen from the calculations of Bernard and Killworth (1977) which yielded the most accuracy amongst informants. (N.B. The cutoff for behavior is adjusted to allow for ties, an occurrence which cannot occur in the cognitive rankings. This involves reducing the number of links in the behavioral data.) The cutoffs used in this paper are shown in Table 2.
Table 2. Details of data treatment for COMPLT, FACTOR and CONCOR use the entire data. Note how different the linkage densities are, both before and after removal of asymmetric links, and – despite our efforts – between behavior and cognition.

<table>
<thead>
<tr>
<th></th>
<th>Office</th>
<th>Tech</th>
<th>Hams</th>
<th>Frat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavior ranked cutoff</td>
<td>8</td>
<td>8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cognition ranked cutoff</td>
<td>8</td>
<td>8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Behavior frequency cutoff</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Cognition scale cutoff</td>
<td>-</td>
<td>-</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Behavior linkage density</td>
<td>0.11</td>
<td>0.16</td>
<td>0.022</td>
<td>0.12</td>
</tr>
<tr>
<td>before removal of asymmetric links</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognition linkage density</td>
<td>0.21</td>
<td>0.24</td>
<td>0.023</td>
<td>0.13</td>
</tr>
<tr>
<td>before removal of asymmetric links</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavioral linkage density</td>
<td>0.035</td>
<td>0.13</td>
<td>0.022</td>
<td>0.12</td>
</tr>
<tr>
<td>after removal of asymmetric links</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognition linkage density</td>
<td>0.13</td>
<td>0.11</td>
<td>0.0074</td>
<td>0.064</td>
</tr>
<tr>
<td>after removal of asymmetric links</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximal frequency of behavioral communication</td>
<td>16</td>
<td>10</td>
<td>85</td>
<td>51</td>
</tr>
</tbody>
</table>

4. Similarity between behavioral and cognitive cliques

Suppose now that an algorithm has been used on a matched set of behavioral and cognitive data. This produces two sets of cliques. The issue is to define how similar these two sets of cliques are. Clearly, two linked problems are involved: 1) How similar are two cliques, and 2) How similar are two sets of cliques? We treat these in turn.

1) How similar are two cliques? To identify a specific cognitive clique with one of a set of behavioral cliques (or vice versa), we quantify the degree of similarity between two cliques.

In doing this, we have set several ground rules.

a) The behavioral clique is, by definition, correct, and the cognitive clique is to be judged against it. (Henceforth, cognitive and behavioral cliques will be referred to as c-cliques and b-cliques respectively).

b) Researchers draw conclusions about group structure from cognitive data. Therefore, a statistical test of how likely it is that a given b-clique will be represented as a particular c-clique is irrelevant. The only thing that matters is how well the b-clique is represented in the cognitive data. For example, a four-person b-clique (1-2-3-4) misrepresented as a four-person c-clique
(1-2-3-5) is just as wrong in a ten-person group as it is in a group of a hundred persons. This holds in spite of the fact that such misrepresentation is far more likely in a small group than in a large group.

c) The only important ingredients in judging the similarity of cliques are the members of both cliques, and their intersection.

By rule c, we can count the number of people who are in, or not in, any given pair of cliques being compared. These counts can be arranged in a $2 \times 2$ contingency table, as in Fig. 1. What is needed is a suitable combination of $a$, $b$, $c$ and $d$ which represents the similarity (or lack thereof) between the cliques. If the two cliques are identical, then $a = m$ (the number of people in the clique), $b = c = 0$, and $d = n - m$. This corresponds to perfect similarity.

Figure 1. Contingency table for comparison of cliques.

At first glance, $\chi^2$ (or an equivalent measure of association, like $\Phi$) would appear suitable. However, $\chi^2$ suffers from two defects. First, large entries for $b$ and $c$ (the minor diagonal) produce just as significant values of $\chi^2$ (i.e. for inaccuracy) as do large entries for $a$ and $d$ (the major diagonal) for accurate representations. Second, the group size $n$ becomes important in the measure – essentially, misrepresenting (1-2) by (1-3) becomes a less serious error for large $n$ than for small $n$. We feel that a measure which includes how many people in the group were not in a given clique is too favorable in its measure of accuracy for small cliques vis-a-vis large ones. By rule b, this measure is rejected.

Katz and Powell (1953) recognized this problem, and devised a measure of association $\Gamma$ as

$$\Gamma = \frac{na - (a+b)(a+c)}{(a+c)(b+d)}$$

which takes the value 0 when the two cliques are independent, and 1 when they are identical. Assuming that $d \gg a, b, c$ (as is usually the case for a large group), simplifies to

$$\Gamma \approx \frac{a}{a+c}$$
which has the advantage (by rule b) of being independent of group size. However, the formula above has a deep intuitive flaw. Consider these two possibilities. First, suppose a b-clique (1-2-3-4-5) is misrepresented as (1-2) in cognition. The Katz and Powell measure yields \( \Gamma \sim 2/5 \). Second, suppose that (1-2-3-4-5) is misrepresented as (1-2-6-7-8); then \( \Gamma \sim 2/5 \) again.

Now the omission of 3, 4, and 5 from the clique is a serious error, and \( \Gamma \) demonstrates this. But \( \Gamma \) also fails to penalize the addition of 6, 7, and 8 to the clique. Since (1-2-6-7-8) is by any measure a worse misrepresentation than (1-2), we must reject \( \Gamma \), and its other forms suggested by Katz and Powell (1953).

The above example leads straightforwardly to a measure of agreement. It is clear that to reach (1-2-6-7-8) from (1-2-3-4-5), six operations must be undertaken: the omission of 3, 4, and 5, and the inclusion of 6, 7, 8. This suggests that an operation count is a valid measure of dissimilarity, and we shall define a dissimilarity measure \( Q \) by

\[
Q = \frac{b + c}{a + c}
\]

as the fractional number of operations needed to convert the b-clique to the c-clique. We choose the fractional number so that misrepresenting (1-2-3) by (1-2) gives the same dissimilarity \( (Q = 1/3) \) as misrepresenting (1-2-3-4-5-6) by (1-2-3-4). Some examples of \( Q \) values are shown in Table 3.

Note that the last two examples in Table 3 overcome the objection to the Katz and Powell measure \( \Gamma \), in that (1-2-6-7-8) is 'twice as wrong' as (1-2). (It is, of course, not obvious whether omission or commission is the worse error; we have chosen to penalize them equally).

Of course, \( Q \) can be related to \( \chi^2 \) just as can \( \Gamma \); Katz and Powell's (1953) argument really makes the obvious statement that

\[
\chi^2 = \Gamma(\chi^2/\Gamma);
\]

the argument trivially holds for any association measure. However, as we have seen, \( \chi^2 \) is not a relevant statistic for drawing conclusions from sociometric data about structure.

<table>
<thead>
<tr>
<th>Behavior clique</th>
<th>Cognitive clique</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2-3-4-5</td>
<td>1-2-3-4-5</td>
<td>0</td>
</tr>
<tr>
<td>1-2-3-4-5</td>
<td>1-2-3-4</td>
<td>1/5</td>
</tr>
<tr>
<td>1-2-3-4-5</td>
<td>1-2-3-4-6</td>
<td>2/5</td>
</tr>
<tr>
<td>1-2-3-4-5</td>
<td>6-7-8-9-10</td>
<td>2</td>
</tr>
<tr>
<td>1-2-3-4-5</td>
<td>1-2</td>
<td>3/5</td>
</tr>
<tr>
<td>1-2-3-4-5</td>
<td>1-2-6-7-8</td>
<td>6/5</td>
</tr>
</tbody>
</table>
Table 4. The clique comparison matrix for invented data. Double underlines show identified cliques; single underlines show multiple representations.

<table>
<thead>
<tr>
<th></th>
<th>Behavior cliques</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>5/2</td>
<td>1/2</td>
<td>3/4</td>
<td>3/2</td>
<td>3/2</td>
</tr>
<tr>
<td>3</td>
<td>1/2</td>
<td>3/2</td>
<td>3/4</td>
<td>3/2</td>
<td>1/2</td>
</tr>
<tr>
<td>Cognitive</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1/2</td>
<td>1</td>
</tr>
<tr>
<td>cliques</td>
<td>5</td>
<td>3/2</td>
<td>1/2</td>
<td>5/4</td>
<td>3/2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>3/2</td>
<td>3/2</td>
<td>3/4</td>
<td>3/2</td>
</tr>
<tr>
<td>7</td>
<td>3/2</td>
<td>1/2</td>
<td>3/4</td>
<td>3/2</td>
<td>3/2</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1/2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

2) How similar are two sets of cliques? Comparison of sets of cliques presents some quite different problems from those encountered in a simple clique-by-clique analysis. For example, the type of cliques generated by one algorithm may be radically different from those of another algorithm. One algorithm may allocate an individual to a unique clique; while another algorithm may permit individuals to hold membership in several cliques. It is even possible for some algorithms to produce two identical copies of a particular clique.

These problems do not disappear when one considers the accuracy of a single algorithm in representing b-cliques by an appropriate set of c-cliques. There may be different numbers of b- and c-cliques; a c-clique may resemble several different b-cliques very strongly; or several c-cliques may all resemble a single b-clique.

The appropriate tool for analysing the differences between a set of b-cliques and a matched set of c-cliques is the clique comparison matrix. This matrix displays the dissimilarity Q between every pair of b- and c-cliques. (The concept of this matrix parallels the triad comparison matrix used by Killworth and Bernard, 1978). An example of a clique comparison matrix is shown in Table 4 for some invented data. The six b-cliques and eight c-cliques have been chosen to include most of the problems encountered in
analysing real data. Note the overlap in clique membership in both sets of
data, and the duplication of c-cliques 4 and 8. Only two b-cliques (5 and 6)
have been identified exactly by two c-cliques (1 and 3, respectively), as
shown by the zero Q scores.

We need somehow to obtain a mean dissimilarity D from the clique com-
parison matrix that has an immediate interpretation. For example, if D were
as low as 0.25, then one would like to think that, on average, a b-clique
(1-2-3-4) would be misrepresented as (1-2-3) or (1-2-3-4-5).

To obtain D we will define a set of rules for identifying b- and c-cliques,
followed by an immediate operationalization of them. The first is obvious:

a) A c-clique must be identified (i.e. matched) with each b-clique if at all
possible. This identification must be made with the ‘best’ c-clique, i.e. that
c-clique whose dissimilarity coefficient Q from the given b-clique is as small
as possible.

Application of this rule to the invented data in Table 4 is equivalent to
seeking column minima in the clique comparison matrix. Unequivocal best
fits are made to b-cliques 1, 5, and 6 (by c-cliques 3, 1, and 6 respectively,
the latter two being totally accurate). Problems arise as to the choice of col-
umn minima for b-cliques 2, 3, and 4, handled by a second, equally trivial,
rule:

b) When choices arise, the choice must be made which minimizes D (i.e.
reduces the measure of inaccuracy).

Application of this rule to Table 4 turns out to have no effect on the out-
come, so that without loss of generality c-cliques 2, 8 and 4 are identified
with b-cliques 2, 3 and 4 respectively. These identifications are denoted by
double underlines in the table.

Note that two things have happened during clique identification. First,
c-clique 3 has been identified with both b-cliques 1 and 6. Second, c-cliques
5, 6 and 7 are not identified with any b-clique. These are symptoms of more
general problems which can occur, namely ‘inadequate representation’ and
‘multiple representation’.

Inadequate representation occurs when a single c-clique is identified with
several b-cliques. This can occur either because there are fewer c-cliques than
b-cliques (so that one c-clique must be identified with several b-cliques) or
because, for some reason, that c-clique is most like several b-cliques. The
converse, multiple representation, occurs when a c-clique is not identified
with any b-clique (the reason for the terminology will become apparent in a
moment). This can occur when there are more c-cliques than b-cliques (when
some c-cliques are missed by the above identification procedure) or because
some inadequate representation on a b-clique caused several c-cliques to be
overlooked.

Both inadequate and multiple representation represent errors in the cog-
nitive structure. Inadequate representation corresponds to the omission of
b-cliques in the cognitive data, whereas multiple representation corresponds
to the commission of c-cliques which have no counterpart in the behavioral
data.
Before a measure $D$ can be defined, it is necessary to deal with the unidentified c-cliques (i.e. multiple representation). This involves our final rule:

c) Any c-cliques that have not been identified by rules a or b are now identified with the b-clique most like them. Operationally, this means that any rows which do not have any double underlines in them (i.e. to have any identifications) are allocated multiple identifications on the basis of their row minima. Thus c-cliques 5 and 7 are identified with b-clique 2, and c-clique 6 with b-clique 5 in Table 4, as indicated by the single underline. The term 'multiple representation', then, refers to the fact that b-cliques 2 and 5 are now identified with more than one c-clique.

Now that a c-clique is identified with each b-clique, and vice versa, some measures of $D$ may be defined. The first, $B$, examines how well each b-clique was represented, and is defined by

$$B = \frac{\text{sum of all double-underlined } Qs}{\text{number of b-cliques}}$$

$B$ is a pure average of the dissimilarity between each b-clique and its best-fitted c-clique. It carries no penalty for additional, perhaps totally erroneous, c-cliques. Thus the error of commission is ignored. The error of omission, on the other hand, is partially represented, because one c-clique is forced to identify with several b-cliques. Thus, for the data in Table 4,

$$B = \frac{1}{6} \left( \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 1 + 0 + 0 \right) = 0.42$$

In other words, the average dissimilarity between each b-clique and the best-fitted c-clique was 0.42; but the fact that there were three unidentified c-cliques is ignored.

A second measure can be defined which partially allows for this problem. This is to examine how accurate each identified clique was. Thus, define the measure $C$ by

$$C = \text{average of all underlined } Qs \ (\text{single and double}).$$

This allows for both omission and commission, at least to some extent. For the invented data,

$$C = \frac{1}{9} \left( \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 1 + 0 + \frac{1}{2} + 0 \right) = 0.44$$

This value is higher than $B$, but only slightly; the penalty for adding c-cliques 5, 6 and 7 is only small by this measure. In fact, adding 20 approximate copies of a b-clique has very little effect on $C$, which is rather unsatisfactory.

After considerable experimentation, involving such operations as row minima before, rather than after, column minima (thus identifying a b-clique for each c-clique rather than vice versa) we have produced the definition of $D$ which most accurately fits our intuition about the values it should take. Additionally, omission and commission are equally penalized, at least for 'pure' data. Define $D$ by

$$D = \frac{\text{sum of all identified } Q \text{ scores (single and double underlined)}}{\text{number of b-cliques}}$$
This can be thought of as a simple column average, except that multiply-identified b-cliques increase the numerator but not the denominator. Thus, in the example,

$$D = \frac{1}{6} \left( \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 1 + 0 + \frac{1}{2} + 0 \right) = 0.67$$

We believe that this figure most accurately represents how dissimilar the c-cliques are to the b-cliques. However, opinions may differ on this matter, so that the B and C measures, which are pure averages, will also be quoted in the text.

No measure can be completely satisfactory; it is always possible to construct examples whereby deliberate mis-identification of cliques can reduce the average dissimilarity, though no such examples occur in our data as far as we are aware.

As a final comparison of the three measures, consider the following examples of ‘pure’ omission or commission. Suppose that the behavioral data consists of n disjoint cliques, each of size m. In the case of omission, we shall suppose that the cognitive data consist precisely of the first p (<n, by supposition) b-cliques, but no others. Then the measures give

$$B = C = D = \frac{2(n-p)}{n}$$

which are all identical. All score p zeros and (n-p) 2’s for dissimilarity. The measures are all proportional to (n-p), the number of omitted b-cliques.

In the case of commission, we suppose that the cognitive data consist of p (>n) disjoint cliques of size m, the first n of which are the b-cliques. The remaining (p-n) of them are total inventions and involve members which do not occur in the b-cliques. Then

$$B = 0; \quad C = \frac{2(p-n)}{p}; \quad D = \frac{2(p-n)}{n} > C$$

so that the B measure is most lenient, then C, and D is the most stringent. But note that of the three measures, only D gives the same value for omission of \(|p-n|\) cliques.

We have detailed the arguments which lead to the selection of D for three reasons. First, it was not at all obvious how a comparison of two structures should have been made, or what the best measure should be. Second, it is not obvious even after the above discussion that D is the ‘best’ available measure; since readers may disagree, we have given a subset of our reasoning to show why we rejected other measures. Last, we believe that the development of comparison measures is important for many branches of social science: not only for analysis of measurement error in network data, but also because social structure, at any level, is a dynamic process. At this time, all that can be said about two sets of data obtained at different times is that they differ; with some additional sophistication, a level of significance can be added. But the ability to demonstrate how much a structure changes is important, and urgently requires research.
5. Clique accuracy

The three algorithms (COMPLT, CONCOR, FACTOR) were applied to the four pairs of data (Office, Tech, Hams, and Frat). This produced twelve sets of comparisons between b-cliques and c-cliques, as shown in Table 5.

Before discussing Table 5 as a whole, four points need to be made. These points are numbered as footnotes in Table 5, but are discussed here for clarity.

1) The COMPLT algorithm produces cliques, together with a list of 'peripherals' for each clique. The values quoted here are for clique comparisons without peripherals. Including the peripherals in their respective cliques improves COMPLT's accuracy ($D = 3.71, 0.59, 0.54, 1.11$ for the four sets of data, left to right). However, the cliques with peripherals included contain nearly all the members of the group being analysed. We do not feel that this is a constructive way to improve accuracy.

2) COMPLT on the Hams data is the most accurate of all the comparisons. However, the accuracy is largely due to the amalgamation by COMPLT of smaller cliques. $D$ between the smaller cliques is 1.1, while amalgamation produces $D = 0.50$. As it turns out, the Hams cognitive data were the most
accurate at the dyadic level of the four data sets. (See Bernard and Killworth, 1977).

3) COMPLT amalgamated the seven Frat b-cliques into one large clique, but left the seven c-cliques unamalgamated. This would have given a $D > 7$. In the general spirit of this research, we analysed the b-cliques from an earlier stage of the algorithm in comparison with the final c-cliques, thus decreasing COMPLT’s $D$ score for these data.

4) Because CONCOR is a divisive algorithm, one must decide at what level to stop the splitting. Our choice of four cliques is common in the literature (but see Section 6).

With the above codicils, we now turn to a discussion of the results presented in Table 5. The first thing we notice is that the best $D$ in the entire table is 0.50 (for COMPLT on Hams). The reader will immediately appreciate what this means:

For three major clique-finders, run on four different sets of data, there is never more than a 50% concordance between the clique structure produced by people’s recall of their interaction, and that produced by their interaction.

The fact that the agreement was not worse than 50% is because of the amalgamation of the cliques, noted in 2) above. In fact, COMPLT put all those who talked to each other (more than the cutoff used) into a clique; it put all those who said they talked to each other into a clique; and the agreement between those two cliques was about 50%.

A second observation on Table 5 is that the different algorithms produce widely varying answers on the same set of data. For example, FACTOR produces uniformly more cliques than the other algorithms. The accuracy of algorithms also varies widely. FACTOR is usually the least accurate (although the only two precisely correct c-cliques were both 1-person, found by FACTOR). CONCOR is the least inaccurate for Office ($D = 1.16$), Tech ($D = 0.78$) and Frat ($D = 1.25$), and COMPLT the least inaccurate for Hams ($D = 0.50$).

Thus the average ‘best’ $D$, over all four data sets, is 0.89 (for comparison, the mean $D$ over Table 5 was 1.6). The average $Ds$ for COMPLT, CONCOR and FACTOR were 2.18, 1.15, 1.48 respectively. Very similar results are found for $B$ and $C$. The variation between data sets is sufficiently large that no algorithm is significantly better than any other, on a one-way analysis of variance.

6. Additional results

Table 5 does not exhaust the ways that the three algorithms could be applied to the data. In this section we discuss briefly several alternative approaches which could have been used.

The choice, for CONCOR, of subdividing into four-cliques is traditional in the literature. We examined subdivisions into two- and eight-cliques on the
Hams data (for which CONCOR was most inaccurate). In both cases, $D$ rose to 3.07 for two-cliques and 1.76 for eight-cliques, with similar rises for $B$ and $C$. Thus the number of subdivisions has an important effect on the accuracy of CONCOR; unfortunately there is no information available to a researcher to enable him to choose what this number should be.

Users of CONCOR usually dichotomize their data (Breiger, Boorman and Arnie, 1975) by applying a cutoff as in Section 3. Therefore the binary version of the Frat data, used by COMPLT, was presented to the CONCOR algorithm to see if this improved its accuracy. Unfortunately, $D$ rose to 2.18, while $B$ and $C$ rose to 1.65 and 1.56 respectively. This suggests that removal of data (by dichotomizing) is not a good thing to do.

Finally, it might be argued that noisy data should be cleaned before applying an algorithm. As far as we are aware, our own CATIJ (Bernard and Killworth, 1973) is the only data treatment which has any claim to filtering randomly inserted noise. Of course, information inaccuracy is probably non-random. If it were random, then any cleaning method should be expected to produce more accurate results for any given clique-finding algorithm.

As an experiment, CATIJ was computed on the Tech ranks data, and a comparison computation was performed on Tech frequencies. Continuing the spirit of this paper, the Tech frequencies were converted to ranks in order of communication frequency, with ties resolved in favor of the lower-numbered members of the group (a rather biased method). Applying FACTOR to the CATIJ matrices gave $D = 1.32$, $B = 0.83$, $C = 0.99$, which are better than FACTOR on the raw data, but not significantly so. Analysing the CATIJ row-1 links with COMPLT (i.e. a binary matrix, with 1's when CATIJ showed a direct link) yielded $D = B = C = 1.10$ without peripherals, or 0.86 with peripherals (and also quite large cliques). Thus CATIJ yields some improvement, but not much; there is no indication that the additional labor involved produces worthwhile results in terms of greater accuracy.

7. Discussion
(a) The data

We have so far tried to avoid any value judgments in discussing the accuracy of various algorithms in producing behavioral and cognitive cliques. But now we must address the fundamental question: how accurately should an algorithm produce cognitive versions of behavioral cliques in order for the cliques to be of any use in understanding behavioral social structure? At the dyadic level, the 50% inaccuracy we found (Bernard and Killworth, 1977) was sufficient to make any deductions about dyadic communication, based on cognition, totally valueless. At the triadic level (Killworth and Bernard, 1978) the 76% inaccuracy was even worse. The hope, throughout both of these papers, was that at the clique level the inaccuracy would be reduced to some acceptable level.

Unfortunately, at the clique level, the average $D$ was 1.60, and the average 'best' $D$ was 0.89. Roughly speaking, the clique structure determined from a
set of cognitive data differs 160% from the behavioral clique structure it is supposed to represent. For example, for any algorithm, the b-clique (1-2-3-4-5-6) is typically represented by the c-clique (1-7-8-9-10); this is, of course, the c-clique that best represents the b-clique.

We expected, at the outset, that Ds of 0.2 or so might occur; indeed, even 40% inaccuracy would be better than that seen at dyadic and triadic levels. After all, representing (1-2-3-4-5) by (1-2-3-4-6) was not, we felt, too bad a misrepresentation. A useful by-product of this paper, we had hoped, would be to find the algorithm which most nearly fitted these reasonable demands on accuracy. But none did.

It seems obvious to us that accuracy of clique representation could be improved by tinkering with default parameters, choosing individual cutoffs for binary data production, and so on. But how can a researcher know a priori how to do this? We feel that our decisions in this paper are quite reasonable, and fair both to the data and to the algorithms. Furthermore, considering the additional information at our disposal (the behavioral data), our decisions are probably superior to those that would be made by a researcher working with only cognitive data.

We are now convinced that cognitive data about communication cannot be used as a proxy for the equivalent behavioral data. This one fundamental conclusion has occurred systematically in a variety of data sets, with a variety of treatments, all as kind to the data as possible. We must therefore recommend unreservedly that any conclusions drawn from data gathered by the question “Who do you talk to?” are of no use in understanding the social structure of communication.

(b) Beyond the data

The basic goal of all social network research must be to formulate the rules by which structures form and evolve in any group. These rules must be precise and modelable; that is, they must be quantifiable, so that they can be tested. The usual procedure in physical science is to formulate a set of rules (a theory) and then to collect data in order to test those rules. This is done, on rare occasions, in the social sciences. For example, structures at the micro-level (e.g. transitivity) have been studied by statistical techniques, such as those developed by Holland and Leinhardt (1975). By necessity, the more common procedure in the social sciences is to collect large amounts of data, in order to examine them for uniformities or signals which (hopefully) will indicate rules.

Since the social sciences rely so completely on data to inform potential theory, the data must satisfy two requirements: 1) they must be error free, or the error level of the data acquisition technique must be known; 2) they must be meaningful, that is, they must relate to something else.

The first requirement is obvious. Noisy data make noisy theory. The second requirement is more subtle. For example, data about the structure of defensive and offensive plays by a football team can be related directly to measurable indices of performance — yards gained per play. Altering the
'structure of work', or the 'task flow' in a factory can be related directly to measurable indices of productivity.

It is apparent that manipulating such structures is likely to produce changes in the indices of outcome. We say "it is apparent" because such relationships between structure and outcome have intuitive appeal to most coaches, managers, and social scientists in our society.

However, many 'apparently' important phenomena (which are so dominant to our senses that they seem to be 'important' to study) turn out to be irrelevant because they do not relate to anything else of interest. For instance, the first and most overwhelming impression of an ocean might be its color; yet color is totally unrelated to ocean temperature, salinity, or practically anything else that comes easily (i.e., intuitively) to mind. Ocean color is related to certain other observables, such as angle and wavelength of incident light; however, the instrumentation required to measure these phenomena is sufficiently complex that they remained non-observable for centuries. We cannot, therefore, assume that our intuition is a reliable guide to what is the meaningful quantity to study. Furthermore, once we have selected something to observe, we are obliged to check that it is, indeed, related to other quantities of interest.

Social networks are usually based on behavior: communication, economic exchange, etc. The obvious thing to observe when studying a network is the behavior, because behavior must be correlated with other important things about the network. For example, A cannot influence B directly, or receive information from B directly if A does not talk to B. Unfortunately, direct measurement of behavior such as communication is very difficult. In fact, in most naturally occurring situations, such measurement is impossible with present technology.

Traditionally, this problem is circumvented by shifting the object of study to something that can be observed, recorded, and measured: individuals' reports of their behavior, i.e., their cognition or recall. These reports are then taken as a proxy for behavior – both because it is convenient to do so, and because it is plausible that they are. The question then arises, is cognition about behavior related to anything of interest? In other words, is behavior recall a meaningful object of study in its own right?

In repeated experiments, we have been unable to show (at least for the instrument "Who do/did you talk to?") that cognition is related to behavior in any meaningful way whatever. We maintain that if cognition about a particular behavior does not relate to the behavior, it is unlikely to relate to anything else, either. For example, in our data there are many cases of people claiming very strongly (i.e., ranked first of second, or scaled high) that they talked to someone in a given time period when in fact they didn't. Conversely, there are many cases of people who talked to someone else a great deal but who claimed not to have done so at all. We cannot expect these behavioral reports to correlate with, say, innovation diffusion, when the very pathways upon which diffusion depends do not exist!
Of course, all of this does not prove that cognition about behavior is not related to anything else. As far as we are aware, however, in naturally occurring groups cognition about behavior is only correlated with other data on cognition, e.g. "Who do you like?" tends to correlate well with "Who do you talk to?".

Nor does our work prove that other instruments which tap recall of a behavior are not accurate proxies for that behavior. Many people have pointed out to us that people can remember some behaviors better than they can remember others. This may be true, but it requires testing. The claim or belief that one instrument is more accurate than another, no matter how plausible, is simply not good enough to allow the rules of social structure to be discovered.

Appendix A

The data in every set are represented by a square matrix (rows representing choosers and columns representing chosen in cognitive data). The behavioral data matrices are naturally symmetric, and the \((i,j)\) entry contains the observed amount or frequency of interaction of person \(i\) with person \(j\). There are two kinds of cognitive data, scales and ranks. Entry \((i,j)\) of a scale matrix contains a small integer, directly proportional to the intensity of interaction with person \(j\) reported by person \(i\). Entry \((i,j)\) of a rank matrix ranges from first \((1)\) to last \((n - 1, n = \text{group size})\).

For each raw data set, that set of people who both responded to the cognitive questionnaire and were present during observation was extracted to produce a square matrix. We realize that this means throwing away data, but we wanted the input to the different algorithms to be uniform, where possible.

COMPLT: The process by which behavioral and cognitive ordinal data were converted to dichotomous data is described in the text. COMPLT reports 'peripheral' members of each clique. For the clique comparison section, results with and without peripherals are contrasted.

COMPLT's default options were used throughout this analysis, with the exception of the RECIP option. Thus we ignore two-person cliques, retain only symmetric pairs, and combine subsets which display significant overlap.

Frat: The behavioral data produced one clique with 45 members, five peripherals, and eight non-members. We also include the results from the next earlier stage of the algorithm, which produces seven cliques. Cutoff for both data sets is 5.

Office and Tech: It is more difficult to match cutoffs for ranked data than for scaled data. Cutoffs of eight links were chosen in all cases, except that dichotomous behavioral data is sparser, due to the fact that sometimes a choice occurs between a cutoff value 5 (for example) which produces three links, and a cutoff value of 6, which produces 40 links. This often occurs. We chose in all cases the lower cutoff value.
References

Alba, R.

Arabie, P.

Bernard, H. R., and P. D. Killworth
1973  "On the social structure of an ocean-going research vessel and other important things. *Social Science Research* 2:145–84.

Breiger, R. L., S. A. Boorman and P. Arabie

Deutscher, I.
1972  *What We Say/What We Do*. Glenview, Ill.: Scott Foresman.

Hallinan, M. T.

Holland, P. and S. Leinhardt

Hubbell, C. H.

Katz, L. and J. H. Powell
1953  "A proposed index of the conformity of one sociometric measurement to another". *Psychometrika* 18:249–56.

Killworth, P. D. and H. R. Bernard

Killworth, P. D. and H. R. Bernard

MacRae, D., Jr.