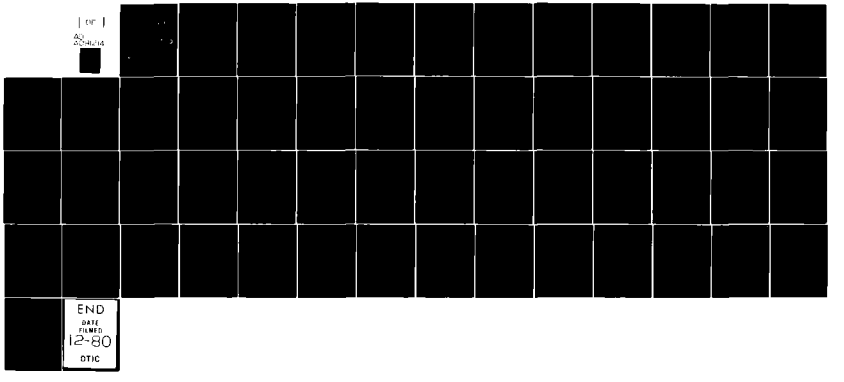
AD-A091 214 WEST VIRGINIA UNIV MORGANTOWN F/G 5/10
INFORMANT ACCURACY IN SOCIAL NETWORK DATA V: AN EXPERIMENTAL AT--ETC(U)
SEP 80 H R BERNARD, P D KILLWORTH, L SAILER N00014-75-C-0441
UNCLASSIFIED BK-120-80 NL



END
DATE FILMED
12-80
DTIC

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: (14) BK-120-80	2. GOVT ACCESSION NO.: AD-A091214
3. RECIPIENT'S CATALOG NUMBER: 224 BS (12)	
4. TITLE (and Subtitle): (6) Informant Accuracy in Social Network Data V: An experimental attempt to predict communication from recall Data	5. TYPE OF REPORT & PERIOD COVERED: (9) Interim Rept's
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): (10) H. Russell Bernard, Peter D. Killworth, Lee Sailer	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0442 P000 1
9. PERFORMING ORGANIZATION NAME AND ADDRESS: H.R. Bernard, Dept. of Anthropology, University of Florida, Gainesville, FL 3261	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (11) NR 177-772
11. CONTROLLING OFFICE NAME AND ADDRESS: ONR-Code 452 Arlington, VA 22217 (13)	12. REPORT DATE: September 1, 1980
	13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): West Virginia University Morgantown, WV 26505 (15) N00014-75-C-0442, NSF-IST78-22802	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Network analysis
Recall
Communications
Behavior & Cognition

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper seeks to discover whether the known inaccuracy of informant recall about their communication behavior can be accounted for by experimentally varying the time period over which recall takes place. The experiment took advantage of a new communications medium (computer conferencing) which enabled us to monitor automatically all the interactions involving a subset of the computer network. The experiment itself was administered entirely by the computer, which interviewed informants and recorded their responses.

(continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

374 100

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

80 10 6 110

Variations in time period failed to account for much of the inaccuracy, which continues, as in previous experiments at an unacceptably high level. One positive finding did emerge: although people do not know with whom they communicate, people en masse seem to know certain broad facts about the communication pattern. All other findings were negative. For example, it is impossible to predict the people an informant claimed to communicate with but did not; and it is impossible to predict who the five people are that an informant forgot to mention that she or he had had communication with.

Thus, despite their presumed good intentions, what people say about their communications bears no resemblance to their behavior. This immediately makes suspect all forms of data gathering, based on questions which require that informants recall their behavior.

ABSTRACT

This paper seeks to discover whether the known inaccuracy of informant recall about their communication behavior can be accounted for by experimentally varying the time period over which recall takes place. The experiment took advantage of a new communications medium (computer conferencing) which enabled us to monitor automatically all the interactions involving a subset of the computer network. The experiment itself was administered entirely by the computer, which interviewed informants and recorded their responses.

Variations in time period failed to account for much of the inaccuracy, which continues, as in previous experiments at an unacceptably high level. One positive finding did emerge: although people do not know with whom they communicate, people en masse seem to know certain broad facts about the communication pattern. All other findings were negative. For example, it is impossible to predict the people an informant claimed to communicate with but did not; and it is impossible to predict who the five people are that an informant forgot to mention that she or he had had communication with.

Thus, despite their presumed good intentions, what people say about their communications bears no resemblance to their behavior. This immediately makes suspect all forms of data gathering, based on questions which require that informants recall their behavior.

Informant Accuracy in Social Network Data V: An Experimental Attempt to Predict Actual Communication From Recall Data

by

H. Russell Bernard
Dept. of Anthropology
University of Florida
Gainesville, FL 32601

Peter D. Killworth
Dept. of Applied Mathematics
& Theoretical Physics
University of Cambridge
Silver Street
Cambridge CB3 9EW
England

Lee Sailer
Dept. of Anthropology
University of Pittsburgh
Pittsburgh, PA

This work was supported by a grant from the National Science Foundation (IST-7812801 A 01), and by a contract from the Office of Naval Research (N00C014-75-C-0441-P00001, Code 452). The opinions expressed in this paper are those of the authors and do not necessarily reflect the positions of the supporting agencies.

1. Introduction

Much of social science is conducted by asking informants to describe their behavior. This is true of studies of such disparate things as organizational communications, food consumption, child rearing practices, sex role behavior, and so on. Studies of naturally-occurring behavior fall into two groups, for our purposes: those in which it is possible to check directly the accuracy of informants' reports, and those in which it is not possible to do so. Social network data are typically of the latter kind; it is simply too unwieldy to check the accuracy of informants' responses to questions such as "who do you talk to?" Besides, if one could easily check the responses, then why ask informants questions in the first place?

Now it is obviously very important in any field to collect accurate data. Otherwise, theoretical deductions made from data (e.g. about social strructure) will be at best, suspect. The validity of data about human behavior has long been a source of vexation; La Pierre (1934) appears to have been among the first researchers to approach the problem experimentally. In a classic study, he toured the United States with a Chinese couple, staying at hotels and eating in restaurants along the way. They were served in 251 establishments, and were refused service in only one.

Six months after the trip was over, La Pierre obtained questionnaire responses from 128 of the establishments. Ninety two percent claimed that they would not "accept members of the Chinese race as guests." Since then a great deal of research has shown that attitudes just do not predict behavior in most cases. Deutscher (1972) has reviewed much of the literature up to 1970; and McGuire (1975) has wondered in print why researchers remain preoccupied with attitudes at all.

If the problem were simply of correspondence between attitudes and behavior, then it could be circumvented by asking people what they do rather than how they feel about certain things. Imagine, for example, what might have happened had La Pierre asked his respondents if they ever had given service to a Chinese person. One might assume that asking people what they do is a better proxy for what they do than asking them how they feel.

Unfortunately, as far as we are aware, this turns out not to be the case. For example, since at least 1951 (Meredith, et al.) researchers of human nutrition have known that people do not recall with any accuracy what they eat, even in "the past 24 hours." Researchers have been attempting to deal with the problem continuously since then, and especially in the last ten years (see, for example, Beaton et al., 1979, and Greger and Etnyre, 1978).

In human communications research, it appears that researchers have been rather more trusting of their data. We are unaware of any research before 1969 which addresses the problem in any way -- except for isolated calls that data accuracy should be checked (Tagiuri, Blake and Bruner, 1953). In 1969, however, Hammer, Polgar and Salzinger, as part of a study of speech predictability, were forced to conclude that informants' cognition "does not constitute an adequate substitute for observation [of behavior]." This pessimistic conclusion appears to have been universally ignored by students of social networks (we ourselves were unaware of it until recently).

In 1975, we began a series of papers (Killworth & Bernard, 1976; Bernard & Killworth, 1977; Killworth & Bernard, 1979a; Bernard, Killworth and Sailer, 1980 -- hereafter referred to as A [Accuracy] I-IV) to examine the accuracy of informants' cognition about one form of their behavior,

specifically the response to the question "who do you talk to?" This involved studying many naturally-occurring groups whose behavior was either automatically, or at least fairly unobtrusively, monitored. We compared the answers to the question "who do you talk to?" (recall data) with the actual communications of the informants (behavioral data).

Our main conclusion was that informants can not recall with acceptable accuracy whom they communicate with in a group over a period of time. For example, informants claim they talk to people they never actually talk to; they claim they never talk to people they do talk to; and they are unable to rank or scale their communications accurately even when referring to the people with whom they have communicated the most.

We considered the possibility that individual differences among informants (on socioeconomic indicators, or on how accurate they felt they were, for example) might help to account for variation in their accuracy (AII). We have found nothing that accounts for substantial parts of variation in informant accuracy. We also considered the possibility that different structures of groups of communicants might be related to accuracy of communication recall. We tested many different triadic structures, and again found nothing to account for variation in informant accuracy, though we did find that both recall data and actual communication data possess significantly high or low amounts of structure on every structural indicator we could think of. Unfortunately, the structures in any particular set of recall data were never produced by the same triads as those in the matched set of behavior data (AIII).

Finally, we considered the possibility that informant accuracy is a function of sub-group organization (AIV). Perhaps modern clique-finding algorithms might uncover an essential, underlying agreement between recall

and behavior data? Again, this turned out not be the case. The three clique-finders we used (chosen because they represent three major traditions in the literature) failed to produce similar cliques in our matched sets of recall and behavior data (or with each other).

Of course, it is possible that informant characteristics really are responsible for variations in accuracy of communications recall data (or any behavioral recall data). It may be that we have simply not made the correct comparisons. Similarly, there may be triadic structures which would give better answers than those we have tested; and there are certainly many clique-finders which we have not examined.[1] But the search would be endless. Clearly, another approach is needed.

In this paper we examine the possiblity that the inaccuracy we have found is a function of time period over which informants are asked to recall their behavior. All our previous data sets have been based on informant reports of their behavior during one of three "windows": the previous five days; the previous month; and the forthcoming month. Any period of time, or window, can be characterized by two quantities, which we call "lag" and "width." Width is the amount of time over which informants are asked to recall their behavior. Lag is the amount of time that has elapsed since the end of the window. Thus, the five-day windows in some of our previous experiments have a width of five days, and a lag of, at most, one day.

The majority of questions asked by students of social networks have a lag of less than one day, with widths that range from a few days to the life time of the informant. It seems plausible that very recent time windows should tend to be more accurate than windows far in the past. "Who did you talk to one minute ago?" should yield more accurate data

than "who did you talk to for a minute at this time last month?" Similar variations in accuracy could be caused by different widths: "who did you talk to during a period of a week, a month ago?" Is there a combination of lag and width which yields the most accurate social network data?

In order to test this, we conducted a totally automated experiment using a computer-based communication system known as EIES. Both behavioral and recall data were gathered by the computer. In section 2, we describe the communications medium, and in section 3, we provide details about the experiment itself and the data acquisition.

2. EIES: A Computer-Based Communications Medium

Prior to this experiment, all our work had been on single time windows. In order to study the accuracy of recall over multiple time windows, either of two things is required: a) many experiments, on many different groups, over many windows; or b) a single experiment on a group engaged in continual conversation over a long period of time.

An ideal example of the latter case is the Electronic Information Exchange System (EIES) at the New Jersey Institute of Technology. The system was developed and funded by the National Science Foundation as a means of improving communication among scientists. The idea was to enable scientists to communicate via computer rather than on a face-to-face basis, and to improve their scholarly productivity.

A complete description of EIES, including its technology and design philosophy may be found in Hiltz and Turoff (1978). Briefly, EIES allows an individual to exchange messages with others on the system by leaving the message in a central computer for pick-up during the next time the "receiver" logs on. Messages may be addressed to single individuals, with or without copies to other individuals. Messages may also be sent to

"groups." A typical group on EIES consists of between 10 and 100 people who have common interests and who are working on a common problem. Many groups on EIES are composed of scientists who hold ongoing "conferences" for periods up to two years since the introduction of EIES. Members of a group are free to enter into small or large conferences with subsets of their own groups, or of other groups.

"Conference comments" are a kind of public message submitted by a conferee for all members of a conference to read. Conference topics range from broad, theoretical discussions of, for example, general systems theory to very specific work-group discussions of, for example, data manipulation techniques. One EIES group planned and executed the experiment reported in this paper.

"Private messages" are communications between individuals; only the sender or the addressees of a private message are privileged to access that message. Private messages include side remarks about conferences; personal letters between friends, enemies and colleagues; and chit-chat between casual EIES acquaintances. Every EIES participant can be identified and addressed by name, nickname, or number (e.g., H. RUSSELL BERNARD, RUSS, or 357).

In other words, conferences function like the formal organizations of a business or university department. The private messages replace what might be called the "day-to-day communication network," where people talk about work and more casual social relations. Many studies of social networks in such environments have been conducted; the advantage of EIES for our purposes is that every non-formal communication (i.e., private message) can be permanently recorded. The privacy of the content of those messages is zealously guarded. We do not treat the content of messages

in the experiment, only what is known as "who-to-whom traffic," or who communicated with whom, and for how many lines of type.

At first glance, EIES may appear to be a rather "exotic" communications medium for a naturalistic study. After all, the overwhelming majority of scientists, much less the rest of the world, do not (yet) communicate via computer. Some of the data used in AI-IV (teletype messages between deaf people, voice activated tape recordings of ham radio operators) might also appear esoteric. There are at least two reasons why EIES <u>is</u> a legitimate medium for the experimental study of communications recall, and <u>is not</u> exotic.

1) The group is simply not exotic for what we are studying. It occurs naturally and involves a subset of the population we wish to study. Some subsets are indeed larger than others; there are more than three hundred thousand ham radio operators in the United States along, and there are more deaf teletype users than there are computer conferencers. But they are all human beings, of the same general cultural background, whose accuracy of recall we are interested in testing. Clearly, we can not generalize about the structures found in such groups to the world at large. But we can (and do) generalize about our informants' ability to recall their communications.

2) It is true that teletypes, radios, and computers are relatively rare media of communication. However, it turns out that the accuracy of informants who use these media is just as poor as that of informants who don't. We have studied several face-to-face groups: two offices (All) and a fraternity (Killworth and Bernard 1979b). All of the previous work, then, indicates that one should not expect EIES to be a "special case." Indeed, it turns out not to be.

Given that we are interested in comparing human recall of communication with actual behavior, EIES is an ideal experimental medium.

3. The Experiment

Between December, 1978 and April, 1979, 57 paid volunteer EIES users participated in our experiment. They ranged in age from 18 to 64, and included students and scientists from many different fields. An invitation to participate in the experiment was sent to over 150 EIES members via a personal message from Bernard.[2] Depending on the rate of their EIES use, each informant took up to 37 interviews, each for a specific lag and width. When an informant logged in to EIES, the computer selected a window and administered an interview. The informant was asked to list the people with whom he or she communicated during that window. Next, informants were given an opportunity to add or to delete names from the list, and were asked to estimate the number of messages and the number of lines sent to and received from each communicant recalled. Finally, they were asked to rate their confidence, in several different ways, on a scale from 1-7, about the information provided.

At the end of each interview, informants were given the opportunity to send the experimenters a message containing any observations or suggestions they wished to make. Twenty-seven windows were established according to the pattern shown in Table 1. Windows were selected for informants in random order. The window selection was modified by computer throughout the experiment to ensure even coverage of all the windows in the experiment. The remaining 10 windows we call "last-ons;" for these windows people were asked to recall their communications during the last time they were on EIES. This ranged from several weeks to several minutes

in lag. Twenty-three informants completed all 37 windows and both interviews, and, out of 57 informants, no regular window was taken fewer than 32 times or more than 38 times. Twenty-two informants took all 10 last-on windows, and 37 people took at least one.

On EIES there is a phenomenon called "deleted" messages -- messages sent, and possibly received, but then purged from EIES before our data collection routines could collect them. Eight percent of the 1211 interviews are contaminated by deleted messages, but never by more than one message per interview.

Two questionnaires were also administered by the computer. The first interview collected data on all our informants' age, sex, self-reported EIES use, and seven self-reported estimates of memory (e.g. "how well, on a scale from 1-7, do you remember birthdays?", "how well names?", etc.) The second interview was taken by the 22 informants who completed all 27 of the basic window interviews. It again asked for information on EIES use, and also asked informants about the 20 people with whom they had actually communicated most. For each of those 20, informants were asked to rate (on a scale of 1-7) the importance of the communication, how satisfying it was, how desirable communication was with that person, and how interesting it was.

Data collection in this experiment was, in a sense, scheduled at the leisure of the informant, and performed by the central computer itself. Thus, it was possible to allow our respondents some control over the progress of interviews. An informant could withdraw from the experiment (permanently or temporarily) at any time. Informants could check on their own accuracy for the previously completed interviews by using a routine called "feedback." They could also check on their general progress by using a routine called "windows."

Two other routines were introduced which we felt might illuminate the causes of variation in informant accuracy. These were called "raincheck" and the "harassment limit." The interviews were administered randomly at the very beginning of an EIES session at a rate sufficient to keep all the subjects at the same pace. For any given interview, a respondent was allowed to take a raincheck of from 1-7 days. (This was changed to 1-3 days later in the experiment, since we felt things were going too slowly.) After taking a raincheck, there was no way a respondent could avoid an interview the next time he or she logged onto EIES.

The harassment limit was the maximum amount of bother that an informant was willing to put up with in one session. After each interview, which averaged about 6-8 minutes, if sufficient time was left in the harassment limit, a last-on window was administered. Most informants selected 20 minutes as their harassment limit.

All the software for the experiment was written by Peter and Trudy Johnson-Lenz. This included all the routines which kept track of the behavioral data, as well as those which administered the interviews and which allowed participants to enter or withdraw from the experiment, to check on their progress, and so forth. David Harvey and the EIES technical staff at the Computerized Conferencing and Communications Center at the New Jersey Institute of Technology wrote the data from disk to tape. The success of this experiment is due entirely to the hard work of these individuals.

IV. Measuring Accuracy

There are various ways one might want to measure accuracy; each way is a function of what a researcher might want to do with the recall data at his or her disposal. For example, if the data were gathered in the form "who are the three people you communicate with the most?" then the researcher would only require that the three persons named by an informant were indeed the three most frequently communicated with persons in the informant's network. Furthermore, the ordering of the three would clearly be irrelevant. Another researcher might want to know the entire network of each person; he or she would then require that all and only those people spoken to by each informant be named. Yet another researcher might be analyzing the frequency (number of contacts or messages) or amount (number of lines, or words, or minutes) of communication. He or she would have far more stringent requirements on accuracy than the first researcher, who needed only three names. Clearly, different research goals invoke different definitions of "accuracy."

For our purposes, we concocted 48 different measures of accuracy, most of which were used previously in this series of papers. They fall into broad classes which make them easy to describe.

Each measure is computed separately for messages the informant recalls sending to people, those from people, and those both to and from, combined, shown in Table 2 as T,F,B.

The first six classes use only the names of those recalled and those actually communicated with. (Measures that use "number of messages," and "number of lines" as indicators of intensity of messaging follow.) T1, T1P, and T2P are straightforward. T12A counts the number of mistakes (T1 + T2) as meaningful in relation to the total number of people actually

communicated with. T12AR counts the number of mistakes as a percentage of the total number of possible mistakes (NA+NR), given the number of people recalled and the number of people actually communicated with for that informant and window.

The second and third classes of inaccuracy measures use either "number of messages" or "number of lines" as indicators of intensity of communication, noted in the table as M or L. This allows us to rank the recalled and actual communicants, and to see, for instance, whether people can recall with accuracy those people with whom they communicate most.

TOP5, TOP3, and TOP1 measure the percentage of errors people make about those they report as their most frequent communicants. WIN2 suggests that people might be able to recall those people most frequently communicated with, but that the exact ranks might be off by 2 or so, and still be counted as correct. WIN20 should indicate when a person recalls actually communicated with in the correct order, but does not penalize the informant for leaving people out randomly.

So, for example, T1PF is the percentage of messages from others recalled by the informant which in fact did not exist. And TOP5TL is the percentage of people reported to be in the top 5 most frequently communicated with (measured by estimated number of lines) not actually in top 5 (measured by actual number of lines). Virtually all of the percentages in this study are what Tukey (1977) calls "started." For example, instead of T1P = T1/NR, we actually use T1P = (T1 + 1/6)/(NR + 1/3) except, of course, when NR is zero, when T1P is undefined.[3] The specific purpose is to make a small adjustment to all of the ratios which will

permit later transformation by logs, inverses, ratios, etc., where values of zero cause problems.

All of these measures take a value close to zero when the recall is accurate, and increase with inaccuracy. Most measures tend to a maximum of 1 when the recall is totally inaccurate, the exception being T1 and T2 (which are straight counts) and T12A (which can, and frequently does, exceed unity.)

In the descriptions which follow (and indeed throughout this paper) we shall refer to the "windows" section of the data only (that is, leaving out "last-ons"), unless otherwise specified.

A simple comparison of the number of people recalled and the actual number of people communicated with demonstrates the unacceptable level of error in the data. On average, 2.5 (SD 4.2) people were recalled as being communicated with; this number ranged from 0-48 in the data. However, 6.0 (SD 10.9) people were actually communicated with. This number ranged from 0-111. Thus, the gross underestimation of communication found in AI,II continues to be present in these data.

The average values, standard deviations, minima and maxima of the 48 accuracy measures are given in Table 3. There are several things which are immediately apparent. For example, the levels of inaccuracy are indistinguishable among the "to," "from," and "both" values within any given measure, and the same is true for "messages" and "lines." Although the number of cases involved runs from almost 250 to 950 (one cannot define TOP1, for example, if no contacts were recalled), the only significant differences between T,F and B, or M and L, is in the simple count measures T1 and T2, which one would expect. This is a little surprising. We might have expected informants to better recall "to" messages, which they initiate, then "from" messages, which are

initiated by others. This is simply not the case, as Table 3 demonstrates. So unless otherwise specified, we will refer to measures without detailing to, from, both, messages or lines.

Only a small number of people were recalled who were not actually communicated with (T1) in a given window: 0.63 (SD 2.1). On one memorable occasion, however, 48 people were recalled -- the maximum number ever recalled, in fact -- but none were spoken to. Although 0.63 is an apparently small error, as a <u>percentage</u> of the number of people recalled (T1P), the error is 30% (SD 32%). Thus of those recalled, about one-third were not communicated with.

The figures are worse if one examines how many people were not recalled but should have been (T2). On average, 5.1 (SD 9.3) people were forgotten, with an awesome maximum of 93. This is also a high percentage of the number of people actually communicated with (T2P), namely 66% (SD 78%). In other words, two-thirds of the people an informant received messages from were forgotten.

Counting each occurrence of these two mistakes as an error, we can count how many errors each informant makes. If the informant says he or she talked to A,B, and C but really talked to A,B, and D, the informant made two errors: of commission for C and omission for D. Judged as a percentage of the number of people the informant really communicated with (here 3), this would give an error of two-thirds, or 67%. The real figure is rather higher, unfortunately: 79% (SD 46%). So, roughly, four-fifths of what an informant says is wrong in some way.

Now many sociometric studies concentrate on only the <u>main</u> communicants for each informant (neglecting infrequently communicated-with people, which are, it is hoped, the main sources of the above error). As we found in AI-II, however, it turns out that informants know their

most-frequent communicants no better than they know their other communicants. Whether one examines number of messages or number of lines, or to, from, or both, one finds:

(a) more than 52% of the time, informants choose the wrong most-frequent communicants (TOP1);

(b) more than 40% of the top three ranked communicants should not belong in the top three, (TOP3);

(c) more than 33% of the top 5 ranked communicants should not belong in the top 5 (TOP5);

(d) if one ranks the people recalled in order of the recalled communication, more than 45% have ranks differing by more than 2 from their position in the actual communication list (WIN2);

(e) in (d) above, more than 58% of those recalled have relative positions in the ranked list more than 10% removed (either way) from their relative positions in the actual communications list (WIN20).

In other words, we can not rely on the people an informant recalls, or the number of messages, or the number of lines, or the people an informant claims to speak to most, with any reliability. As a rough guide, we have the consistent result (see also AI,II) that at least half of what an informant says about his or her communication with others is incorrect.

It is clearly cumbersome to refer continually to 48 separate measures of accuracy, especially when, as we have seen, they are very similar. To reduce the level of complexity, the results of a factor analysis on those accuracy measures which lay between 0 and 1 was used to combine them into general indices. (We shall return to such measures as T1 later).

Five factors were created, each with a recognizable set of measures comprising the main factor loading. Since each set is, furthermore, a plausible subset of "similar" measures, we created five new overall inaccuracy measures as follows:

ACCT = average of (T1PT, TOP3TM, TOP5TM, WIN2TM, TOP3TL, TOP5TL, WIN2TL)

ACCF = average of (T1PF, TOP3FM, TOP5FM, WIN2FM, TOP3FL, TOP5FL, WIN2FL)

ACC2 = average of (T2PT, T2PF, T2PB, T12ART, T12ARF, T12ARB)

ACCTOP1 = average of (TOP1TM, TOP1TL, TOP1FM, TOP1FL, TOP1BM, TOP1BL)

ACC20 = average of (WIN20TM, WIN20TL, WIN20FM, WIN20FL, WIN20BM, WIN20BL)

where "average" above is defined as follows:

if two or more of the measures in a definition have non-undefined values, the "average" is a simple average of the non-undefined values; if only one or zero of the measures in a definition is defined, the "average" is undefined (i.e., missing).

The pattern of these five measures should be evident. ACCT is a compilation of "to" measures in errors of commission, roughly speaking; ACCF is the identical compilation of "from" measures. ACC2 involves a composite of T2P and T12AR, and roughly measures errors of omission. ACCTOP1 is a simple average of all TOP1 measures, and ACC20 a simple average of all WIN20 measures.

The values of the five new inaccuracy measures reflect the values of the 48 original variables well. ACCT has a mean of 0.46 (SD 0.31); ACCF 0.44(SD 0.29); ACC2 0.65 (SD 0.27); ACCTOP2 0.55 (SD 0.29); and ACC20 0.59 (SD 0.24). These means are based on a minimum of 460 valid cases.

WINDOW	WIDTH	LAG	TIME AGO	INTERVIEWS COMPLETED
1	30	31	60	36
2	30	1	30	36
3	14	47	60	35
4	14	17	30	36
5	14	1	14	35
6	7	54	60	32
7	7	24	30	34
8	7	8	14	35
9	7	1	7	34
10	3	58	60	34
11	3	28	30	36
12	3	12	14	37
13	3	5	7	35
14	3	1	3	34
15	2	59	60	36
16	2	29	30	36
17	2	13	14	35
18	2	6	7	34
19	2	2	3	35
20	2	1	2	34
21	1	60	60	37
22	1	30	30	38
23	1	14	14	33
24	1	7	7	37
25	1	3	3	33
26	1	2	2	34
27	1	1	1	37
28	LAST ON	LAST ON	LAST ON	37
29	LAST ON	LAST ON	LAST ON	34
30	LAST ON	LAST ON	LAST ON	29
31	LAST ON	LAST ON	LAST ON	25
32	LAST ON	LAST ON	LAST ON	24
33	LAST ON	LAST ON	LAST ON	24
34	LAST ON	LAST ON	LAST ON	23
35	LAST ON	LAST ON	LAST ON	23
36	LAST ON	LAST ON	LAST ON	22
37	LAST ON	LAST ON	LAST ON	22

TABLE 1

WINDOW LISTINGS

Width and lag are defined in the text; time ago is the time between the interview date and the start of the window. All times given in days.

T1 {T, F, B} -- The number of people recalled who were not actualiy communicated with.

T1P {T, F, B} -- T1/NR, where NR is the number of people recalled.

T2 {T, F, B} -- The number of people not recalled who were actually communicated with.

T2P {T, F, B} -- T2/NA, where NA is the number of people actually communicated with.

T12A {T, F, B} -- (T1 + T2)/NA

T12AR {T, F, B} -- (T1 + T2)/(NR + NA). This represents the percentage of the total possible number of mistakes made by the informant.

TOPn -- Let n be an integer (in fact n = 1,3 or 5), and define a "hit" to occur whenever a person is in both the top n most intense recalled and the top n most intense actually. Then

$$\text{TOPn} = 1 - \frac{\text{number of hits}}{n}$$

hence we may define

TOP1 {T, F, B}{M, L} ; TOP3 {T, F, B}{M, L} ; TOP5 {T, F, B}{M, L}

WIN2 {T, F, B}{M, L} -- Let a "hit" mean that the rank of a person on the recalled list is within 2 of his or her rank on the actual list. Then

$$\text{WIN2} = 1 - \frac{\text{number of hits}}{\text{number of recalled}}.$$

WIN20 {T, F, B}{M, L} -- Let a "hit" mean that the percentile rank of a person on the recalled list is within 10 of his or her rank on the actual list, so that

$$\text{WIN20} = 1 - \frac{\text{number of hits}}{\text{number recalled}}.$$

TABLE 2

INACCURACY MEASURES

T,F,B refer to 'to', 'from' and 'both to and from' respectively. M and L refer to number of messages and lines respectively. All measures are zero for accurate recall and increase with inaccuracy.

	Mean	S.D.	Min.	Max.		Mean	S.D.	Min.	Max.
T	0.70	2.0	0	45	T	0.40	0.30	0.04	0.97
T1 F	0.61	1.8	0	42	TOP3 F	0.43	0.32	0.04	0.96
B	0.63	2.1	0	48	B	0.40	0.30	0.04	0.97
T	0.37	0.34	0.01	1.0	T	0.42	0.30	0.04	0.96
T1P F	0.35	0.32	0.01	1.0	TOP3 F	0.44	0.31	0.05	0.96
B	0.30	0.32	0.01	1.0	B	0.42	0.30	0.05	0.96
T	3.3	7.7	0	85	T	0.35	0.28	0.02	0.97
T2 F	3.5	5.7	0	48	TOP5 F M	0.37	0.29	0.03	0.97
B	5.1	9.3	0	93	B	0.33	0.27	0.03	0.96
T	0.59	0.31	0.01	0.99	T	0.38	0.29	0.03	0.97
T2P F	0.67	0.27	0.03	0.99	TOP5 F L	0.37	0.29	0.03	0.97
B	0.66	0.28	0.01	0.99	B	0.36	0.27	0.03	0.97
T	0.81	0.60	0.02	6.9	T	0.49	0.33	0.01	1
T12A F	0.82	0.45	0.03	4.6	WIN2 F M	0.48	0.32	0.01	1
B	0.79	0.46	0.01	4.6	B	0.45	0.32	0.01	1
T	0.44	0.26	0.01	0.99	T	0.52	0.33	0.02	1
T12AR F	0.49	0.25	0.02	0.99	WIN2 F L	0.49	0.32	0.03	1
B	0.48	0.25	0.01	0.99	B	0.47	0.32	0.02	1
T	0.52	0.38	0.12	0.87	T	0.58	0.32	0.01	1
TOP1 F M	0.54	0.37	0.12	0.87	WIN20 F M	0.58	0.31	0.01	1
B	0.52	0.38	0.12	0.87	B	0.58	0.30	0.01	1
T	0.54	0.37	0.12	0.87	T	0.62	0.31	0.01	1
TOP1 F L	0.54	0.37	0.12	0.87	WIN20 F L	0.62	0.29	0.03	1
B	0.54	0.37	0.12	0.87	B	0.60	0.29	0.02	1

TABLE 3

VALUES OF INACCURACY MEASURES

(Measures are defined in Table 2)

V. The Effects of Lag and Width on Accuracy of Recall

The levels of inaccuracy found in the previous section are, as hypothesized, not uniformly distributed, at least over the 27 windows considered here. Figures 1-5 show contours of the five overall inaccuracy measures, as functions of lag and width. (All values for a given lag and width have been averaged, and those averages contoured. There is a wide variation between informants.) There is a strong, but not systematic, variation with lag and width for all five measures. Multiple correlations of the measures on lag and width account for at best 8% of the variance in the data (for ACC2); inclusion of quadratic terms is of little help, yielding only 14% at best (also for ACC2).

The maximum values in all cases are for two- or four-week lags (usually two) and widths of one day. As hypothesized, asking people about "one day a long time ago" does, indeed, produce highly inaccurate answers (at least 74% incorrect on any of the five measures). Curiously, a lag of two months and width of one day is systematically more accurate than two-week or one-month lags with the same width, although the differences are not statistically significant.[4] This suggests that for such windows, informants tend to report those whom they believe they "usually talk to." In fact, this explanation was offered by several users of EIES in comments which they made to us on the system about the experiment. Although our informants' technique for handling these awkward windows (one day, sixty days ago) yields more accurate data, their data for such windows remain at least 70% inaccurate.

Increasing the width of the window, as might be expected, increases the accuracy, although the trends in any measure are by no means uniform. We had anticipated that a lag and width of one day (i.e., yesterday) would

uniformly produce the most accurate data. On only two of the measures (ACCT and ACC2) was this the case. ACC20 was the most extreme, with greatest accuracy involving a week-long window, ending the day prior to the interview.

Let us consider each inaccuracy measure in turn. ACCT (Figure 1) measures people's inability to recall who they sent messages to. This inability tends to increase as either lag or width increase. ACCF measures inability to recall who people received messages from. The effects of width are mainly confined to 1-3 days. For larger widths the inaccuracy depends only weakly on width. ACC2 measures the ability to invent communicants they didn't really communicate with. Here, accuracy is best for lags of 1-2 days. For longer lags, inaccuracy increases with lag and decreases with width. ACCTOP1 measures people's inability to recall their most "used" communicant (in terms of either frequency or amount of communication). For widths above three days, the measure is insensitive to both lag and width. ACC20 measures the inaccuracy of what an informant recalls, with little penalty for omitting communicants. Increasing lag or decreasing width both increase the error here, although for small lags (i.e., less than two days) the effects of width are weak.

All the cases examined so far allow the possibility of intervening communication on EIES between the end of a window and the time of an interview. It seems likely that this could be a major source of inaccuracy for informants. That is, the intervening communication might be confused by an informant with communication during a particular window.

The last-on windows were included in order to test for this hypothesis. In other words, we believed that informants might be more accurate in reporting their communications with others the last time they used EIES than they would be in reporting their communications during any of the 27

windows. Indeed, this is the case. The five inaccuracy measures, computed for *last-on* interviews only (with a minimum of 97 cases), have the following mean values: ACCT 0.38, SD 0.35; ACCF 0.31, SD 0.32; ACC2 0.48, SD 0.34; ACCTOP1 0.37, SD 0.30; ACC20 0.43, SD 0.32. In each case, these values are more accurate than the corresponding value for the 27 windows.

It is not clear how to decide whether these values are significantly better, due to the many contributory factors involved (not the least of which is the persistent strong differences in accuracy between informants). A naive t-test between pairs of means shows significant** differences in every case. (Henceforth, single asterisks denote significance at the 5% level or better; double asterisks denote significance at the 1% level, or better). Now, 80% of all last-on interviews involve lags of at most two days, whereas the average windowed lag is 20 days. Thus, the last-on inaccuracies would be expected to be less than regular window inaccuracies, due to this fact alone. Restricting attention to windows and last-ons possessing identical lags and widths, the results continue to be significantly** more accurate for last-ons.

Is last-on accuracy affected by lag? Multiple regression of the inaccuracy measures for last-ons with lag (and order of presentation, to illuminate a possible learning effect), accounts for, at most, 15% of the variance (in this case for ACCF). So, informants are not systematically more accurate for shorter lags, even for last-on communication. In fairness, the 15% of variance accounted for in ACCF is significant**, but the scatter implied by this low figure is sufficiently great to invalidate the use of very short lags in order to obtain accurate results.

Although last-on inaccuracy is less than window inaccuracy, it is clearly still too large for reliable use of recall data in network

studies. In order to improve accuracy still further, we examined the 93 last-on windows which had a lag of zero days. In other words, informants for each of these 93 interviews had used EIES earlier the same day as their interview. In fact, they had logged off EIES no more than 20 minutes ago. One would assume that informants would be highly accurate, given that they were being asked to recall their communications such a short time ago. The results were quite mixed. Some people, as usual, are very accurate, while others are not. For example, of the 35 cases in which ACCTOP1 could be computed, 20 were correct. However, the mean inaccuracies remain unacceptably high: ACCT has a mean of 0.30, SD 0.33; ACCF 0.21, SD 0.26; ACC2 0.42, SD 0.35; ACCTOP1 0.34, SD 0.29; ACC20 0.38, SD 0.29. Surprisingly, only ACCF is significantly* less inaccurate for same-day last-ons than for last-ons with a lag of one or more days.

...Given the very short lags for same-day last-ons (i.e., no more than 20 minutes) we can examine how inaccuracy varies in very short time intervals. The scatter still remains too high to account for variations in inaccuracy. Multiple regression of the five measures in lag and width now measured in minutes) still only accounts for, at best, 18% of the variance (in this case, for ACC2). In no case is a significant amount of variance accounted for. As an indication of the scatter involved, note that of 5 interviews conducted just one minute since the informant had last been on EIES, on two of these occasions ACCT had values larger than 0.87, and on three occasions values less than 0.2. The predominant factor determining accuracy is simply wide variation amongst informants. Some people are fairly accurate, while others are grossly inaccurate. We will examine these differences further in Section VI.

Recapitulating, a researcher asking for communications data could expect the most accurate results from data on very recent time windows. However, there is no way to know a priori what width the window should be for greatest accuracy. It is highly plausible that more recent events should be recalled more accurately than less recent events. But, while hardly surprising, these results are not trivial. Consider that data on a lag of two days and a width of one day are distinctly less accurate than data on a lag and width of one day. Hence, the exact positioning of the window in time has an extreme effect on the accuracy of the data acquired: even tiny alterations in the lag or width of the window produce large alterations in the accuracy.

Nor are these results very comforting. The most accurate value, for any non-last-on window, of each of the five measures, still yields 36% inaccuracy, on average. Arguably, this could be counted as 64% accurate data; however, (a) there is no way to know which data are accurate and (b) recall that all cases when either of NR or NA — the number of people recalled and actually communicated with — is zero have been excluded from consideration; these are also highly inaccurate. (Including values of zero for NR or NA, would yield infinite values of inaccuracy. Removing those values, however, only serves to raise artificially the level of accuracy. Section VIII discusses this case in detail.)

Still, some researchers might choose to interpret this finding as an encouraging sign that asking people who they talk to (and/or how much they talk to others) can yield data which are sufficiently accurate for further manipulation. We would consider such an interpretation unproductive for the following reasons. First, consider that the minimum value of ACCTOP1, over any window, is 0.32. This simply means that, for the most accurate window (in this case lag one, width two), on

32% of the occasions informants could not name correctly the person with whom they communicated with most frequently. Second, to repeat, there is no way for a researcher taking data to choose the "most accurate window" for any given study. Even if this were possible, the researcher would have to settle for less inaccuracy of one kind at the cost of getting higher inaccuracy of other kinds. Finally, the most accurate source of data is on windows with a lag of a few minutes. But researchers collecting data in the field would themselves have been present during these "more accurate" windows. Thus, at best, they would have been able to observe communication directly (in which case, why ask for data from informants?); and, at worst, their presence will have modified the communications being measured.

VI. What Else Acccounts For Inaccuracy?

We have seen that the dependence of accuracy on the lag and width of time windows is not strong. Clearly, other variables are contributing to informant inaccuracy. Some of these variables are presumably functions of the personal history and qualities of each informant. Some informants have better memories than others, for example; some use EIES more frequently than others; and so on. Some variables may be a function of the particular window under consideration. Perhaps the window involved a lot (or very little) message traffic; perhaps the informant was in a hurry when being interviewed; or perhaps the informant's first few interviews were less accurate than later ones.

During the background interview, we asked each informant, how well, on a scale of 1-7, he or she could remember each of the following: zip codes, phone numbers, names, faces, dates, lyrics, and birthdays.

Perhaps an informant's self-evaluation of memory is related to his or her accuracy in recalling communication. At the end of each window interview, informants also provided estimates of their confidence, on a scale of 1-7, about their recall of the following: list of communicants, number of messages sent, number of messages received, number of lines sent, and number of lines received. Both the memory and the confidence measures averaged around 4, as might be expected. Since these variables are too highly intercorrelated to use separately in regressions, we factored each set. This produced three memory variables: the average of names and faces; birthdays; and phone numbers. A similar factoring on confidence measures reduced them to two: confidence in the list of communicants; and the average of the other four.

Surprisingly, the memory variables were almost uncorrelated with the five inaccuracy measures; however the two confidence measures were reasonably correlated ($r = -0.2$ to -0.3) with inaccuracy. Of course, the lack of correlation of memory and inaccuracy could be produced by other, more subtle cross-correlations. Accordingly, a large number of variables was entered in a multiple correlational search to find the predictors of accuracy. In the search, at various levels of inclusion, were: sex and age of informant; number of people recalled ("to," "from," and "both"); time to take the window; total time ever spent on EIES by the informant; lag, width; number of people communicated with (again for the three categories); the three memory variables; the two confidence variables; the number of times "feedback" had been used by an informant to check previous accuracy; and the order of presentation of the window.

Little variance was accounted for, even by such a list of variables. Eighteen percent of the variance of ACCT was accounted for, mainly by

number of communicants "to" (recalled and actual), and both confidence measures. Only 15% was accounted for ACCF, by number "from" (recalled and actual), lag, and confidence in messages and lines. ACC2 was best accounted for (37%), by number of recalled communicants and confidence in that list. ACCTOP1 had 16% accounted for, by total time ever spent on EIES and confidence in list of communicants; ACC20 had 22% accounted for, by number of actual communicants and confidence in messages and lines.

An extra attempt was made by inventing such variables as effort (time taken during window per communicant recalled), and activity during window (number recalled per day of width). Again, logical and empirical transformations of the data were made to improve the fit.

The conclusions of this section still hold. In short, everything we have measured seems to be related to inaccuracy in a reasonable way. The problem is that nothing *seems to* matter very much.

VII. The Special Case of No Communication

A special case of these calculations occurs when NR or NA are zero (i.e., when an informant claims he spoke to no-one or when she actually spoke to no-one). This case automatically removed many inaccuracy measures from previous consideration as they could not be defined.

On 29% of occasions, in fact, an informant had no actual communication during the window under consideration. And on 28% of occasions an informant recalled communicating with no-one. If these two sets of occasions completely overlapped, the informants would always be accurate when they claimed not to speak to anyone.

The overlap is, of course, imperfect. On 41% of those occasions when an informant recalled having no communication, he or she did in fact have communication; and on those occasions she or he communicated with 4.8 different people. Similarly, on 19% of occasions when informants

actually had no communications during a particular window, they claimed, on average, to have communicated with 2.1 different people. Consistently, in all our work, we have found that errors of omission are more severe than those of commission.

Most of these figures are well-predicted** by the width (but not the lag) of the window under consideration. Both the percentage of times a mistake occurs, and the number of omitted or committed communicants, increase strongly with width, with correlations of the order of 0.7 to 0.8. Only the mean number of commissions (given a commission occurred) is weakly described by width ($r = 0.27$**). Hence, the longer the time over which informants recall their interaction, the more errors of omission or commission are made by those informants.

VIII. What Is The Best We Can Do?

It is already clear, both from the preceding sections and from AI-IV, that data from informants about their communications, over any time period, are unreliable. Given this, are there any positive statements which could be made? This and the next two sections are attempts to find specific rules for treating the data so as to yield reliable results. This section examines whether one can predict the list of people communicated with, given only informants' recall.

The situation is difficult, as Table 4 demonstrates. One might arguably be able to find some rules to predict the 0.63 people not communicated with but recalled; but it is unclear how to predict who the 5.1 people are who are not recalled but were communicated with. (The entry in the lower right-hand corner depends on the size of population involved and is not easy to define; the number involved is obviously large, but defining the entries here to be "accurate" hardly helps the situation.)

Let us first seek to predict the numbers in Table 4. (The equivalent tables for "to" and "from" are equally predictable, and omitted here as are "last-on" cases, which are much more scattered.) We are given only NR (number recalled) plus the information detailed in Section VI. Now NA can be predicted to 64%** of its variance, overwhelmingly by a linear function of NR, whose coefficient is about 1.44; the underestimation is typical of all our data sets. Since

$$NR = a + b$$

is known, and

$$NA = a + c$$

is well predicted, only one more quantity needs to be predicted to define a, b and c. In fact T1 (i.e., b) and T2 (i.e., c) can also be predicted, the former to 36%**—again a linear function of NR—and the latter to 52%**, by NR, and total time ever spent on EIES. As a result, a, b and c are all predicted by linear functions of NR, with coefficients 0.68, 0.32 and 0.77 respectively.

Predictability of numbers of people in various categories, of course, is of little help to a researcher concerned with mapping the communication structure of a group. The recorder needs to know which people fall into the four categories. Is there some rule which would enable the researcher to obtain recall data from an informant and then to select some of those communicants and be sure they were in category (a), i.e., were actually communicated with? We are not here requiring a rule which specifies the entire of category (a); merely a reliable subset—no member of category (b) is to be allowed. Given the high level of inaccuracy involved, this is clearly the best one might hope for.

recall \ Behavior	communicated with	not communicated with
communicated with	2.14 (a)	0.63 (b)
not communicated with	5.10 (c)	??

TABLE 4

Accuracy contingency table

The entry in each box is the mean number of communicants for that box: e.g. 5.1 people were communicated with but not recalled. The lower right entry cannot easily be defined.

There are two ways this might be achieved. Obviously the rule must involve selecting those people an informant reported communicating with most frequently. The chances are slim at best that someone would be reported as spoken to only rarely and yet be consistently in category (a). The simplest rule, then, is to define some (small) integer n and specify that the people reported as spoken to first, second, ..., nth most often are actually spoken to. Recall that there may be other actual communicants; this rule would not seek to find them.

Let us define an inaccurate "score" which is rather similar to T2P. For a given n, the score is the ratio (undefined when both NR and NA are zero).

$$\text{score} = \frac{\text{number of those in category (b) predicted by the rule}}{\min(n, \text{number of reported communicants})}$$

The rule is accurate when the score is zero, and totally inaccurate when the score is unity. When n exceeds the number of reported communicants, all communicants are selected by the rule.

Somewhat surprisingly the score almost always decreased monotonically with n. A peak in inaccuracy usually occurred for very low n—suggesting that the frequent restriction by sociometricians to an informant's "top 3" choices may be dangerous. In fact the median value for the most inaccurate cutoff n for this rule turns out to be n = 2, where the score takes an average value of 79%. In other words, 79% of the people selected by "use the top 2 recalled communicants" are not spoken to!

Because of the improvement in accuracy by increasing n, the optimal rule involves selecting all recalled communicants as being actual communicants. However, this still yields 19% inaccuracy. Thus,

although this is the most accurate version of the rule, it is unreliable once in every five occasions, and clearly unacceptable.

The second possible method would be to modify the cutoff used. It might be argued that only those individuals perceived as "communicated with a great deal" should be included by the rule. In other words, the inclusion rule ceases to be relative ("take the top 5," etc.) and becomes absolute ("choose all those recalled as having more than x communication" for some x).

We chose to make the cutoff point be a function of informant. Each informant's total communication was scanned, and the maximum number of messages and lines was recorded over all windows and all communicants. The selection rule then became "choose a recalled communicant only if the amount of recalled communication (messages or lines) exceeds x% of that informant's maximum communication." What value should x take in order to achieve totally reliable data?

Unfortunately, x needs to be 100 percent (and the data are not reliable even then). Figure 6 shows histograms of the required cutoffs, over the informants. The largest peaks are in the 91-100 percent band, indicating that for at least twelve informants any rule of this type would be spurious. There is a cutoff of 10 percent or less for only 6 informants. In general, the scatter in Figure 6 is too great to produce a reliable rule.

Nor is the situation improved by considering the numerical values of the cutoffs rather than their percentage values. Eighty percent of these cutoffs lie in the lowest 10 percent of the message or lines traffic. For

example, the cutoff for 41 informants involved fewer than 10 messages for total reliability; for 8 informants (16 percent of those for whom the calculation could be performed) the cutoff was two messages or less for total reliability.

We are forced to conclude that there is no reliable way to select a subset of those recalled who are actually communicated with. If we select only those communicants with reported communication, more than 90 percent of the maximum ever achieved—a very stringent criterion—no less than 25 percent of the time the data are wrong.

IX. Global Statistics

Many of the results presented so far have been based on dyadic measures; that is, two people are involved: an informant, and a communicant. In our previous papers (AI, III, IV) we analyzed higher level data, including triads and n-tads or "cliques." The data were progressively more inaccurate as the level of structure became more complex. Because data in this paper were taken from a small subset of a closed group, repeating the analyses at the triadic or clique levels would be fruitless. However, this does not invalidate the less stringent task of searching for similarities in the global structures of recall and behavioral data. This section investigates "net popularity," and the structural equivalence of the two data sets.

a) Popularity

Interest in locating the most popular persons in a group goes back to the beginnings of sociometry. Most groups appear to have a small subset of their members who are communicated with significantly more often than others in the group. Although informants' recall is poor at the dyadic level, do they nonetheless "know" who the popular members <u>are</u> in the group? We tested this in two ways.

In the first method, we estimated the actual popularity of each member of EIES, by adding up all the messages/lines ever sent, by the informants in any of the windows to that member of EIES. For these purposes, there are 364 members of EIES. Due to temporal overlap of some of the windows, the results may be slightly, but unavoidably, biased. We ranked the top 20 of the 364 in order of communication, by both messages and lines. A similar procedure was carried out for recall data, and the two sets of ranks were compared. Here the results are rather encouraging. The person in EIES who is communicated with most (messages or lines) is the

fourth most popular person in the recall data. Nonetheless, the four most popular people in the behavioral data are the same as the four most popular people in the recall data, but in wrong order. (The consistent underestimation continues; both lines and messages are underestimated by about 50%.) Even the top 10 seem reasonable: only one in two of the behavioral top 10 are omitted in the recall data.

The same results held when we restricted our attention to a subset of the data. Instead of recording all messages from an informant to the entire population of EIES, we recorded only the communication (actual and reported) for each of our informants to the n persons on EIES with whom each informant communicated with first, second.....nth most often during a given window. Here n takes the values 1,3, or 5. Precisely similar results are found. In other words, informants may not know who they speak to the most; but they appear to know, in general, who is most spoken to.

In the second method, we examined the popularity of our informants rather than of EIES in general. This time we counted incoming messages from all persons on EIES to our informants (again, both messages/lines and behavior/recall data). We ranked the informants in order of popularity, and we obtained results similar to those obtained in the first method. The first three informants (ranked by messages) are the same for both behavior and recall, though in the wrong order. The most popular person (ranked by lines) was the same for both behavior and recall. Although the second most popular person in behavior was valued sixth in recall (again for lines) the top six were the same in both cases.

Similar results are found by restricting attention to the top 1,3 or 5 communicants, although the resulting most-popular person is never the same for recall and behavior.

b) Structural equivalence

Although informants are inaccurately recalling their communication at many levels, we showed above that they have an accurate "feel" for the popular members of the group. Do they in fact recall accurately the relative positions of themselves and others in the group? In other words, how equivalent are the structures present in behavioral and recall data?[5] (Again, the small subset of the group comprising the informants precludes other analyses such as centrality and the like.)

The strong inaccuracy at the dyadic level suggested that any comparison between behavior and recall at all but the simplest level would probably fail. Hence we simplified both behavioral and recall data to a (57x384) matrix m_{ik}, where

$$m_{ik} = \begin{cases} 1 & \text{if } i \text{ ever communicated with } k \\ 0 & \text{otherwise} \end{cases}$$

We then defined three (57x57) matrices on the subset of our informants. The first is a simple symmetric distance measure d_{ij}, where

$$d_{ij} = \begin{cases} \sum_k (m_{ik} - m_{jk})^2 & i \neq j \\ 0 & i = j \end{cases}$$

where the sum is taken over all k in the entire group, and the zero diagonal value is for later convenience. Thus d_{ij} is small when i and j are "similar" and large when i and j are "dissimilar."

The second and third matrices are "substitutability" measures s_{ij} and t_{ij}. Both measure how well i and j can substitute for each other in terms of their patterns of communication. The s_{ij} matrix is symmetric, by dividing the intersection of i's and j's communication by the union:

$$s_{ij} = \begin{cases} \dfrac{\sum_k m_{ik}m_{jk}}{\sum_k (m_{ik} + m_{jk} - m_{ik}m_{jk})} & \text{if denominator} \neq 0 \\ 1 & \text{if denominator} = 0 \end{cases}$$

where the 1 indicates perfect substitutability if i has no communication. The t_{ij} matrix is asymmetric, by normalizing by i's total communication:

$$t_{ij} = \begin{cases} \dfrac{\sum_k m_{ik}m_{jk}}{\sum_k m_{ik}} & \text{if denominator} \neq 0 \\ 1 & \text{if denominator} = 0 \end{cases}$$

These last two matrices increase with i's similarity to j; the first, d_{ij}, decreases with i's similarity to j. All have zero diagonal values.

We may now compare behavioral and recall versions of each matrix, by the Γ measure introduced by Katz and Powell (1953) and extended by Hubert and Baker (1978). Γ is no more than the correlation coefficient between the behavioral and recall entries of d_{ij}, s_{ij} or t_{ij}. Its significance can then be tested by Mantel's strategy (see Hubert and Baker). This examines whether relabeling the 57 informants in the recall matrix would produce a significantly better or worse fit to the un-relabelled behavioral matrix. Hubert and Baker provide an approximate Z-score for Γ [(mean-expected mean) ÷ standard deviation] together with a pessimistic estimate of significance level, Q. The Z-score of course yields an optimistic level; above 1.96 the results are significant. Monte Carlo simulations would be necessary if the results showed conflicting significance estimates.

The results for the three matrices are:

d_{ij} : $\Gamma = 0.64$; $Z = 5.1$; $Q = 3.7\%$

s_{ij} : $\Gamma = 0.30$; $Z = 9.1$; $Q = 1.2\%$

t_{ij} : $\Gamma = 0.39$; $Z = 4.1$; $Q = 5.6\%$

In all cases the degree of structural agreement between behavior and recall is at least significant*, with very high Z scores. So the behavioral and recall matrices possess similar signals. However, the detailed agreement is rather poor: the variance accounted for in the behavioral data by the recall data is 41%, 9%, and 15% respectively. In other words, one data set could not be used as a proxy for the other.

In summary, then, at a global level there is reasonable agreement between recall and behavior. Recall data yields a list of "popular" people which is very similar to the list produced by behavioral data. Similarity and dissimilarity measures between informants show considerable correlation between behavior and recall data, but recall accounts for insufficient variance in behavioral data for it to be used as any kind of predictor.

X. Can We Calibrate the Recall Data?

Implicit in most empirical studies is the concept of cost-effectiveness. How much will it cost to collect good data, and will it be worth it? The two specific extreme choices in our case are (a) use inexpensive measures of message traffic, such as recalled messages to (RMT) and recalled messages from (RMF) some person, and collect large amounts of data; or (b) use costly, direct observational measures of message traffic, in our case the actual number of messages to (AMT) and from (AMF) some person. This is only feasible on a small dataset. Typical research projects in network analysis use economical but inaccurate measures. In this section we suggest and demonstrate a technology that may help improve the accuracy of the cheap measure for a few extra dollars.

In the data sets we work with, we purposely record the expensive measures and the inexpensive measures for all the cases. (In fact, we chose our research population because the observational measures, usually so expensive, are cheap.) One simple and general measure of the accuracy of the cheap measure is the mean square error, in this case

$$MSE(RMT) = \frac{\sum_{i=1}^{N} (AMT_i - RMT_i)^2}{N}$$

To "improve" RMT, we adapt what in sampling theory is called Regression Estimation. Suppose that in a large data set some concept is measured inaccurately (the usual case). Regression estimation proceeds as follows:

1. Choose a small, simple random sample of cases from the data set.
2. Measure (again) each case in the sample using the expensive, accurate measure (AMT or AMF in our case).
3. Using any and all cheap measures and statistical tricks, develop a prediction equation for the accurate measures. (In our case, AMT

is a function of RMT, RMT^2, number of people recalled, effort, lag, width, perceived activity, experience using EIES, and several interactions of similar variables.) Since this is a simple random sample of the data set, the prediction equations should generalize to the data set.

4. The independent variables in the prediction equation are all cheap (by our design) and have been measured for all cases in the data set. Call the value of the predicted valued for each case in the data set "corrected RMT," or CRMT. In other words, RMT is corrected for bias, and various individual characteristics by using the relationship between AMT and RMT, effort, etc. in the sample.

Statistical theory that the connected RMT in the entire data set will be a better proxy for AMT than uncorrected RMT. In our data set we can assess this claim directly, since we know AMT. The accuracy of CRMT, is therefore

$$MSE(CRMT) = \frac{\sum_{i=1}^{N} (AMT_i - CRMT_i)^2}{N}$$

The relative accuracy of CRMT and RMT in measuring AMT is, for our data,

$$\frac{MSE(CRMT)}{MSE(RMT)} = 80\% \quad .$$

The same result for CRMF and RMF is

$$\frac{MSE(CRMF)}{MSE(RMF)} = 80\% \quad .$$

The corrected RMT and RMF are roughly 20% better than the raw measures. While this might encourage some, it is not really as good as it might be. Being 20% better than awful is not good; it is medium bad.

Still, if the project must go on, there are two alternatives. The

researcher must choose to a) measure N_1 cases at c_1 dollars per case or b) measure N_2 cases at c_1 dollars per case and n_2 cases at c_2 dollars per case, where c_2/c_1 is large and n_2/N_2 is small. N_1 and N_2 are about equal. For example, instead of 1000 cases at $1 per case, one could collect 750 cases at $1 per case and 50 cases at $5 per case. The total cost is the same. But if the 50 case sample can be used to improve the accuracy of the data set by a factor of more than $\sqrt{1000}/\sqrt{750} = 1.15$, then the final results from plan (b) should be much more accurate in the long run. Calibration of the recall data in this paper unfortunately yielded abysmal results, but this may be because we failed to put the right quantities into the regressions. We will have more to say about the implications of this in the conclusions.

XI. Conclusions

In an effort to determine how much lag and width of a time window affected communication recall, we designed a totally automated experiment. The experiment took advantage of a new communications medium (computer conferencing) which enabled us to monitor automatically all interactions involving a subset of the computer network. In previous experiments we had found little which accounted for the gross inaccuracy in human recall of communication. We believed that the concepts of lag and width might prove helpful.

Although lag and width account for some of the <u>variation</u> in accuracy (small lags and widths tended to be more accurate than large ones), the amount of <u>variance</u> accounted for is small (typically about 10%). Consideration of a wide variety of other variables still failed to account for most of the variation in accuracy (never more than 37%, and usually less than 20%).

Nor are people more accurate when they recalled communicating with nobody. On 41% of such occasions, communication had taken place, with 4.8 different people, on average.

Only one positive statement can be made about accuracy from our results. Although individual people do not know with whom they communicate, people en masse seem to know certain broad facts about the communication pattern. Specifically, if we examine the aggregate of what everybody says about their communications with everybody, the resulting "most-frequently-communicated-with members of the group turn out to be correct. That is, the list of the top six most "popular" people is the same for both recall and behavioral data.

All other findings were negative. It is impossible, for example, to produce an accurate list of those with whom an informant has communication, given his or her recalled list together with estimates of amount or frequency. It is impossible to predict who the (on average) five people are that an informant forgot to mention that she or he had had communication with. It is impossible to predict the people an informant claimed to communicate with but did not. And, finally, although the structure of recall and behavioral data are correlated, the scatter remains far too high to use one as a proxy for the other.

XII. Discussion

We began this series of papers in 1975 because we distrusted conclusions drawn by network researchers (including ourselves) about the structure of communications in human groups. We had no reason to distrust the motives of our (or anyone else's) informants. As far as we know, if a researcher inquires about an informant's communications, the data obtained are an accurate (i.e., honest) description of how the informant believes

he or she communicates. We continue to assume that the amount, frequency, and persons involved all accurately represent the informant's view of his or her network. However, one consistent and unavoidable conclusion has emerged from our studies of informant accuracy in network data: what people say, despite their presumed good intentions, bears no useful resemblance to their behavior.

This immediately makes suspect all forms of the instruments "what do you _____?" and "who do you _____?" It may very well be that peasant farmers can report accurately how many bushels of wheat they harvested last year, or it may not be. It appears that people's reports of their voting behavior are accurate, if the data are gathered immediately. (What proportion of the population today would claim to have voted for Richard Nixon in 1972?). On the other hand, asking people about their consumption of goods and services produces appalling results. As far as accuracy of recall about communication is concerned, the only thing people have ever recalled accurately in our experiments is who the most "popular" people are in their group. (By "popular" we mean who in the group is communicated with the most.) Even then, informants get the most popular individual wrong most of the time.

We feel that it is vital in any field to have accurate (not just reliable) data. It is virtually impossible to develop a theory for any process unless one can obtain accurate data about that process. This must be just as true for human communications (and interactions in general) as for black holes, DNA molecules, or the movement of tectonic plates. Still, it is obvious that in research on human beings in natural settings, acquiring full, accurate data on their behavior is nearly impossible. We have been able to achieve this only because we selected groups whose behavior could

be monitored, and not because of any interest we might have had in the groups themselves. Our interest has been exclusively methodological.

There are at least two ways to treat the dilemma of needing accurate data and not having any. Both ways are important and should be implemented. The first requires the collection of behavioral data in natural settings on child rearing practices, alcohol consumption, leisure activities, health care activities -- in short, on everything in which social scientists are interested, and for which they normally rely on recall data. It is not necessary (we hope) to collect full, matched sets of recall and behavioral data such as we have done in our program of methodological studies. It should be sufficient to obtain, for each behavior being studied, a sample from the population, in order to calibrate the data obtained from informants' recall. It logically follows that we should not pretend to study quantitatively things that can not be measured by direct observation, or at least by using accurate and calibrated (if indirect) instruments.

The second way is to seek other quantities, hitherto unmeasured, which may be accounting for inaccurate recall. Quantities which come to mind are motivation, content, importance, meaning, ecological conditions, population density, norms, detail of the interview procedure, and so on. These quantities need to be defined, then collected -- accurately! -- and finally checked to see if they are related to, or predict, the behavior which we are trying to study. We cannot simply "blame" inaccurate data on these quantities until and unless we have examined whether this is the case. So far, everything we have tested fails to account for inaccuracy. The unpleasant possibility is that nothing accounts for variations in accuracy, except individual (that is, random) differences . . .

FOOTNOTES

1. Burt and Bittner (1980) have pointed out that the clique-finders which we used do not necessarily produce statistically adequate subgroups. We support their call for testing the adequacy of subgroups, and we note that this has not been done until very recently with the advent of algorithms for doing so. We have a gnawing suspicion that this will only further invalidate much of sociometric and social network research.

2. A copy of the two page invitation letter, and full documentation of the experiment is contained in a technical report, available from the authors; the Office of Naval Research, Code 452, Arlington, Va., 22217; or NTIS. The report is called "An experiment on the degradation of accuracy in human recall of communications," (see Bernard, Killworth, and Sailer 1979) and contains a codebook for the publicly available tape of the data from the experiment. The tape is available from Bernard.

3. The removal of certain measures from consideration when, for example, NR or NA is zero, may appear to bias the averages which follow in the text. (We defer consideration of NA or NR with values of zero until Section VII.) The averages of various measures quoted in this paper are biased in a statistical sense, due to the starting involved. This results in shift toward 0.5 in all fraction-type measures; high inaccuracy is decreased by this, low inaccuracy is increased. The differences are numerically small except at extreme cases, near zero or unity, when they increase to about 10%. Monte Carlo simulations show that the mean of the unstarted fractions is unbiased but inefficient; the mean of the started fractions is biased but more efficient. Opinions were divided between the authors as to which is the better approach. In the end, it is probably a question of each researcher's background.

4. We realize that we do not have independent cases, normal distributions, etc. We use the word "significant" to mean sizable, or notable, or whatever. The probabilities are those produced by the statistical packages and are included for information rather than statements about some population.

5. We are indebted to Ronald Burt for discussions leading to this investigation.

BIBLIOGRAPHY

Beaton, G.H., J. Milner, P. Corey, V. McGuire, M. Cousins, E. Stewart. M. de Ramos, D. Hewitt, P.V. Grambsch, N. Kassim and J.A. Little (1979) Sources of variance in 24-hour dietary recall data: Implications for nutrition study design and interpretation. American Journal of Clinical Nutrition, 32, 2546-2559.

Bernard, H.R. and P.D. Killworth (1977) Informant accuracy in social network data II. Human Communication Research, 4, 3-18.

Bernard, H.R., P.D. Killworth and L. Sailer (1979) An experiment on the degradation of accuracy in human recall of communications. Technical Report BK-118-79, Office of Naval Research, Code 452.

Bernard, H.R., P.D. Killworth and L. Sailer (1980) Informant accuracy in social network data IV, or a comparison of clique-level structure in behavioral and cognitive data. Social Networks, 2, in press.

Burt, R.S. and W.M. Bittner (1980) A note on inferences regarding network subgroups. Social Networks, in press.

Deutscher, I. (1972) What we say/what we do. Scott, Foresman and Co., Glenview, Illinois, 370 pp.

Greger, J.L. and G.M. Etnyre (1978) Validity of 24-hour dietary recalls by adolescent females. American Journal of Public Health, 68, 70-72.

Hammer M., S. Polgar and K. Salzinger (1969) Speech predictability and social contact patterns in an informal group. Human Organization, 28, 235-242.

Hiltz, S.R. and M. Turoff (1978) The network nation. Reading, Mass: Addison-Wesley.

Hubert, L.J. and F.B. Baker (1978) Evaluating the conformity of sociometric measurements. Psychometrika, 43, 31-41.

Katz, L. and J.H. Powell (1953) A proposed index of the conformity of one sociometric measurement to another. Psychometrika, 18, 249-256.

Killworth, P.D. and H.R. Bernard (1976) Informant accuracy in social network data. Human Organization, 35, 269-286.

Killworth, P.D. and H.R. Bernard (1979a) Informant accuracy in social network data III, or a comparison of triadic structures in behavioral and cognitive data. Social Networks, 2, 19-46.

Killworth, P.D. and H.R. Bernard (1979b) Help?! An offer of data for analysis. Connections, 2, no. 1, 40-41.

La Pierre, R.T. (1934) Attitudes vs. actions. Social Forces, 13, 230-237.

McGuire, W.J. (1975) The concepts of attitudes and their relations to behaviors. In: Perspectives on attitude assemssment: Surveys and their alternatives. H.W. Sinaiko and L.A. Broedling, eds. Technical Report, Office of Naval Research.

Meredith, A., A. Matthews, M. Zichefoose, E. Weagley, M. Wayave and E.G. Broon (1951) How well do school children recall what they have eaten? Journal of the American Dietary Association 27, 749-751.

Tagiuri, R., R.R. Blake and J.S. Bruner (1953) Some determinants of the perception of positive and negative feelings in others. Journal of Abnormal and Social Psychology, 48, 585-592.

Tukey, J.W. (1977) Exploratory data analysis. Reading, Mass: Addison-Wesley.

CAPTIONS FOR FIGURES 1-6

Figure 1. Contours of the means of the ACCT (inaccuracy "to") measure as a function of lag and width. Both lag and width are expressed in days, on a log-log scale for clarity (i.e. "in the last n days" corresponds to the upright axis). Contours are every 0.05, labelled every 0.1. The heavy dots indicate the location of the 27 windows; the sparcity of data in the upper left quadrant means that the smoother contours there should be interpreted with caution. The minimum value is 0.30 (lag=width-1); maximum 0.74 (lag=14, width=1).

Figure 2. Contours of the means of the ACCF (inaccuracy "from") measure, displayed as in Figure 1. Minimum value 0.24 (lag=1, width=2); maximum 0.81 (lag=30, width=1).

Figure 3. Contours of the means of the ACC2 (inaccuracy "to and from") measure, displayed as in Figure 1. Minimum value 0.45 (lag=width=1); maximum 0.85 (lag=14, width=1).

Figure 4. Contours of the means of the ACCTOP1 (inaccuracy "top ranked person") measure, displayed as in Figure 1. Minimum value 0.32 (lag=1, width=2); maximum 0.83 (lag=14, width=1).

Figure 5. Contour of the means of the ACC20 (inaccuracy "error in ranking by ±10%") measure, displayed as in Figure 1. Minimum value 0.47 (lag=1, width=7); maximum 0.75 (lag=14, width=1).

Figure 6. Histograms of the minimum percentage of total message or line communication required for accuracy. If an informant reports communication with someone above this percentage cutoff, then that person is in fact communicated with. Below the cutoff, this may not be true. The solid bars show messages; the plain bars, lines.

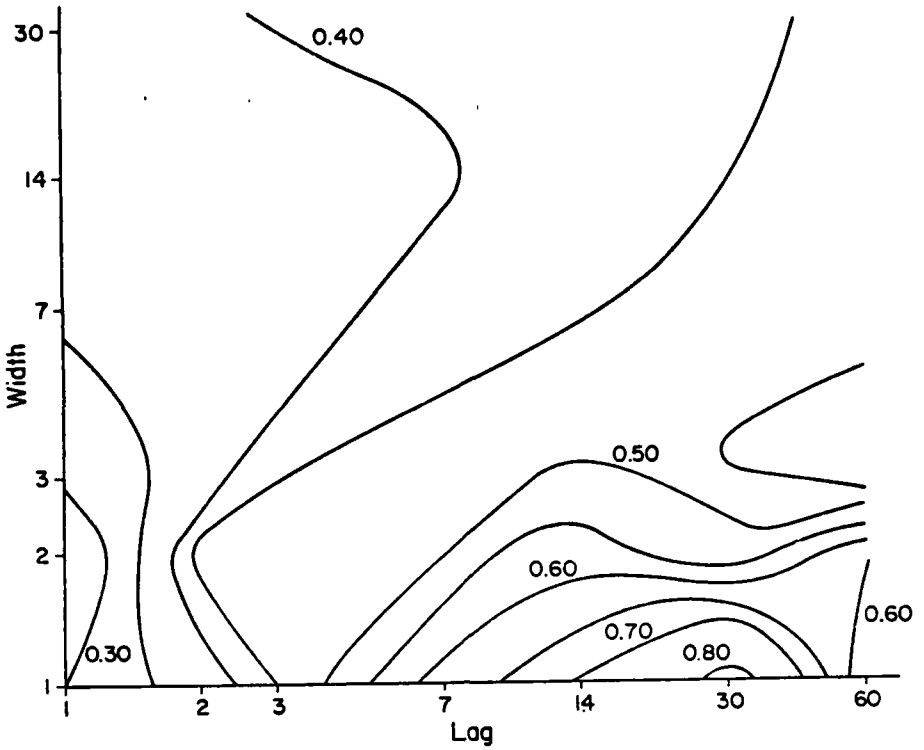

30
14
7
3
2
1
Width
0.40
0.50
0.60
0.70
0.80
0.30
0.60
1
2
3
7
14
30
60
Lag

FIG 2

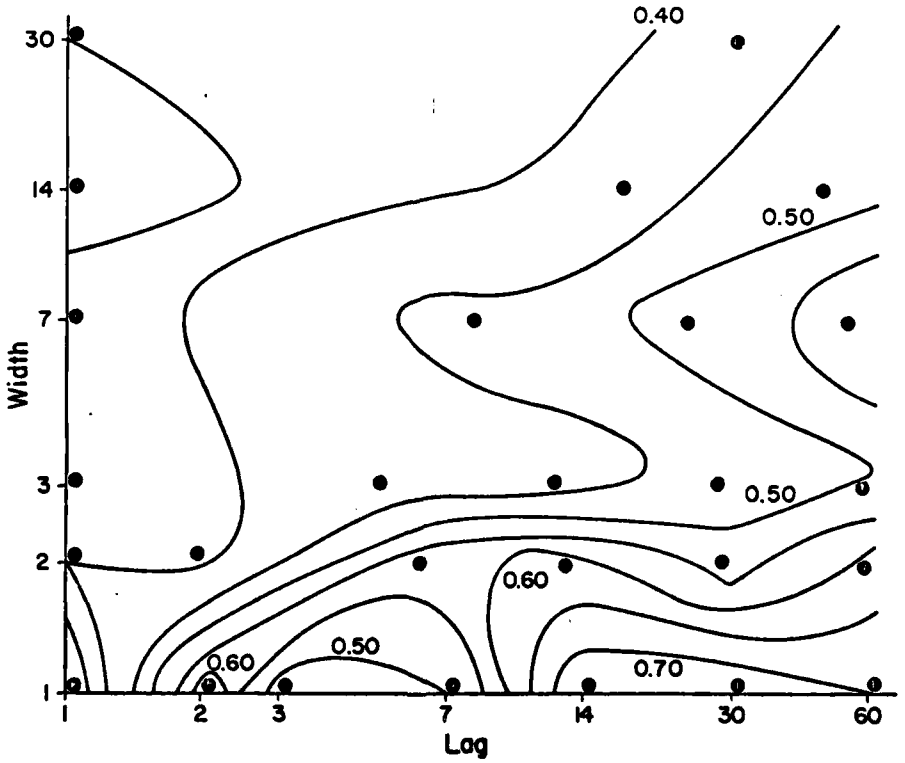
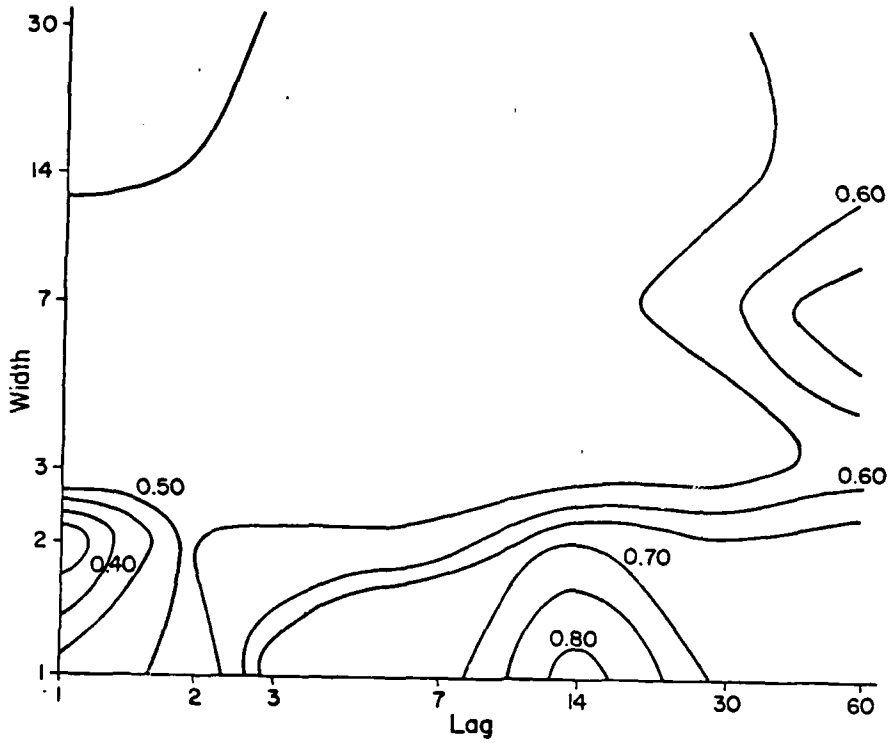
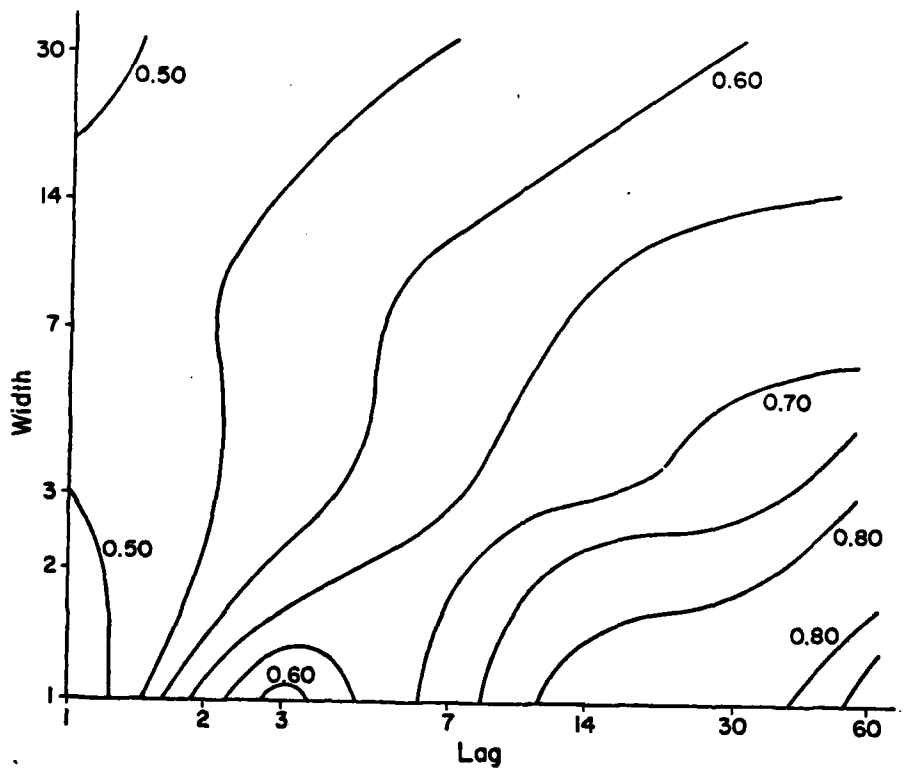
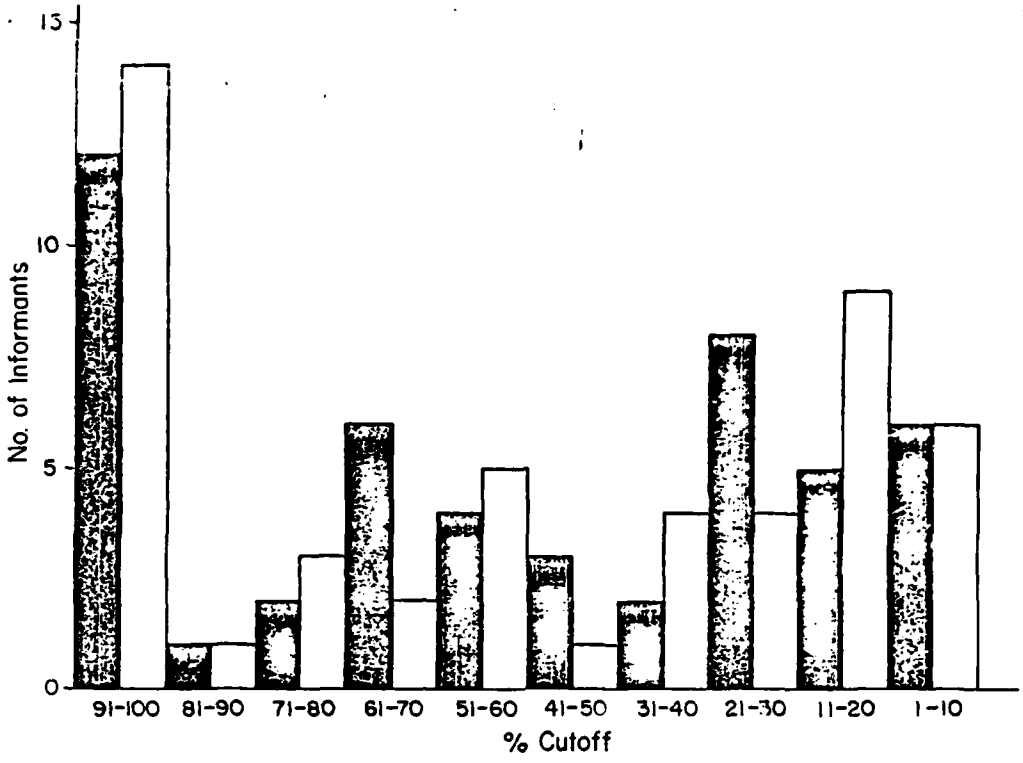

30
14
7
3
2
1
Width
0.40
0.50
0.50
0.60
0.60
0.50
0.70
1
2
3
7
14
30
60
Lag

FIG 1


30
14
7
3
2
1
Width
0.60
0.50
0.60
0.40
0.70
0.80
1
2
3
7
14
30
60
Lag


30
14
7
3
2
1
Width
0.50
0.60
0.70
0.50
0.80
0.60
0.80
1
2
3
7
14
30
60
Lag


13
10
5
0
No. of Informants
91-100
81-90
71-80
61-70
51-60
41-50
31-40
21-30
11-20
1-10
% Cutoff

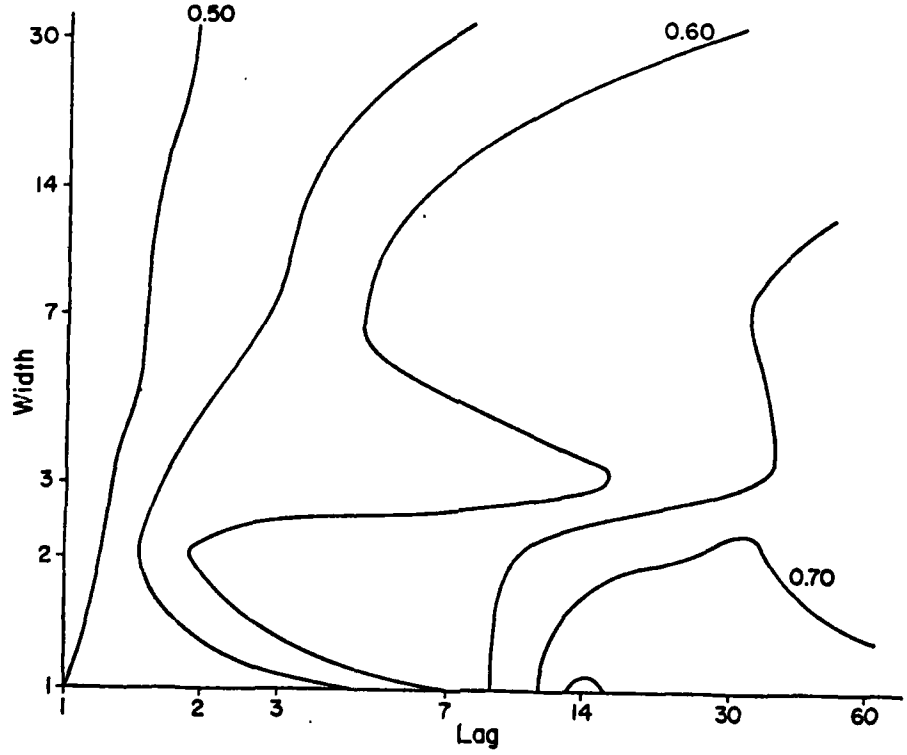

0.50
0.60
0.70
30
14
7
3
2
1
Width
1
2
3
7
14
30
60
Lag

FIG 5