Index: An Informant-Defined Experiment in Social Structure

H. Russell Bernard; Peter D. Killworth; Christopher McCarty

Social Forces, Vol. 61, No. 1 (Sep., 1982), 99-133.

Stable URL:
http://links.jstor.org/sici?sici=0037-7732%28198209%2961%3A1%3C99%3AIAIEIS%3E2.0.CO%3B2-2

Social Forces is currently published by University of North Carolina Press.

INDEX: An Informant-Defined Experiment in Social Structure*

H. RUSSELL BERNARD, *University of Florida*
PETER D. KILLWORTH, *Cambridge University*
CHRISTOPHER MC CARTY, *University of Florida*

ABSTRACT

This paper describes an informant-defined experiment, designed to answer the questions "whom does any informant know, and why?" Each of 50 informants was allowed to ask an unlimited number of questions about each of 50 target persons (all mythical, each with a created life history). When informants felt they had enough information, they told us which of their acquaintances was most likely to know the target person, or, more precisely, could serve as the first step in a chain of acquaintances.

The data show that four basic questions (location, occupation, age, and sex) account for more than 50 percent of questions asked, and a basic collection of six or seven questions would suffice for most circumstances. Less often used questions tend to be employed only when the basic set produces no useful information for that informant. The (verbal) reasons given for a choice could be succinctly defined using no more than four concepts. Analysis of these reasons shows that the basic pattern occurs throughout: location is important when the target is near, in a big town, and has a low occupation rating; occupation is important when the reverse holds.

In this paper we report the results of an experimental investigation of social structure. For our purposes, understanding social structure requires two essentially different kinds of information. First, we need to know, on average, how many people are known to each person in a group (such as in the U.S.), and who they are. This would provide a *description* of social structure. Second, we need to know *how* people think they are related to the people they know. This would provide an *explanation* of the description (at the cognitive level, at least).

*This work was supported under Office of Naval Research Contract #N00014-75-C-04410-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. We are deeply indebted to Bruce Mayhew for stimulating discussions about social structure.

 0037-7732/82/010099-33$03.50

This led us (Killworth and Bernard, a) to investigate experimentally how and why people think they know each other. In our experiment we presented 58 informants, or starters, with a long list of fictional people, or targets. (We have borrowed the terminology from the small-world literature which has influenced our thinking and our experiments. For a review of the literature, see Bernard and Killworth, and particularly, Pool and Kochen, Milgram, and Lin et al.). For each target, we provided some basic information: name, race/ethnicity, location, and occupation. Starters provided the name of a choice who they felt would be most likely to know the target, or to know someone who *might* know the target, and so on. In other words, informants provided the name of someone in their own network who would serve as the first link in a possible chain of acquaintances to the target (hence our debt to the small-world literature). Informants also provided some information about their choices. They told us their relationship with the choice (relative, friend, or acquaintance), and they checked a list of reasons for selection of a choice. The reasons were location, occupation, race/ethnicity, and "other." For example, if an informant said a choice was picked on the basis of location, then something about the location of the target and/or the choice were somehow connected in the informant's mind.

Overwhelmingly, location and occupation were the important reasons for choice. Characteristics of informants (except for their sex) had little effect on the type of, or reasons for, choices. However, characteristics of targets were highly correlated with both type of, and reasons for, choices. For example, the most likely reason for choosing an intermediary for a given target could be predicted 81 percent of the time, based on the target's occupation and distance from Morgantown, West Virginia, where the experiment took place.

This study yielded a lot of valuable information, but had two serious shortcomings. First, we had very little information about the choices. We knew their names (and thus, in most instances, their gender), and whether they were relatives or friends of the informant. Second, the instrument was closed-ended; it provided only a few selected pieces of information about each target, and forced informants to choose intermediaries and provide reasons for their choices, based solely on these pieces of information.

Informants' comments about the experiment revealed both an occasional need for more information, and that a connection between a choice and target could be very indirect. For example, some informants asked about the religion of the targets; many informants wanted to know the sex of targets whose exotic or foreign names concealed this piece of information.

Quite often, choices selected on the basis of location did not live (and *never* lived) anywhere near the target. Nonetheless, on these occasions, informants insisted that the choice was *associated* with the target's location. The choice's children, for example, might have gone to college in the target's home state.

The comments by our informants regarding the indirectness of such associations were convincing. We attempted to build both direct and indirect associations into a model of the process by which informants made their choices (Killworth and Bernard, b). In order to test this model, we assumed that each link in a chain of acquaintances belongs to one of a discrete set of classes or states in a Markov process. (We do *not* assume that the decision-making process is Markovian, only that the mechanics whereby the next choice is made are independent of the history of the small-world chain.) Many of the transition probabilities had to be guessed, lacking data about them, or even confirmation that all the states in our model existed. It could be argued, therefore, that the good fit between the model's predictions and known facts generated by small-world experiments was fortuitous or self-tuned.

In order to improve the credibility of such a model, we need to know what, if any, information informants need about a target (aside from location, occupation, sex, and race) to make their *best* choice. And we need to know how informants actually make their choice, once they are armed with a collection of facts about a target. In order to study this, an open-ended experiment was conducted. We consider this a member of a genre we call INDEX, or "Informant-Defined Experiment." The idea is to study social structure experimentally, but to allow the subjects of the study to define the information which is collected.

We turn now to a description of the experiment, followed by a discussion of the coding of the data. The analysis of the data deals with questions informants asked, choices they made, reasons they gave for their choices, and personal information about the informants and the targets. These groups of data are analyzed first singly, and then in combination with one another, where appropriate, in order to find out the relations between variables. In each section of analysis, we address one or both of the central questions we are asking about social structure: *Who do people know? How do they know them?*

I. The Informant-Defined Experiment

A list of 50 mythical targets was constructed. (There is now no direct connection with small-world experiments as such, but we retain the terminology for consistency.) Each target was given a name (and, therefore, gender), an occupation, a location, and a racial identity, as in traditional small-world experiments. Occupations were selected from the Duncan Scale to represent a cross-section of life in the U.S. Three housewives were included, and were assigned husbands with occupations;[1] three students were included, and were assigned both fields of study and part-time jobs;

three retired persons were included, but were assigned occupations prior to their retirement.

Location was rather more complicated. We divided the U.S. into six categories of location: (1) near-urban (i.e., Morgantown, WV); (2) near rural (i.e., the surrounding county); (3) "medium" urban (i.e., cities[2] within a 250 mile radius of Morgantown); (4) medium rural; (5) "far" urban (i.e., cities further than 250 miles from Morgantown; and (6) far-rural. The first two categories were assigned five targets each; the other four were assigned 10 targets each. Five black targets and 45 white targets were defined. Twenty-five males and 25 females were included.

In addition to these usual identifiers in social science experiments, we assigned some additional information to the targets, based on informants' comments during the Killworth and Bernard (a) study. Targets were assigned ages, ranging from 20–70 years; a religion; an education level, ranging from grade school to graduate degree (in six gradations); and a marital status. Table 1 summarizes the characteristics initially assigned to the 50 targets.

Six pretest subjects were asked to select a choice for each target. However, they were given no information whatever about any target—not even a target's name. Informants were told that they could ask for any information they felt they needed about any target in order to make a choice of intermediary. This pretesting revealed that targets' organizational affiliations and hobbies were frequently requested by informants. The instrument was modified and targets were assigned a maximum of five hobbies and five organizations each. Several pretest informants asked whether targets were active in religion; and whether targets had children, how many, and/or what ages. This information was added to the personal history of the targets.

Fifty informants provided us with the data reported in this paper. Informants were solicited by advertising and were offered $20 each for

Table 1. SOME DATA ABOUT THE 50 TARGETS; INCOME HAD NOT BEEN PROVIDED

	Mean	Standard Deviation	Median	Mode	Range
Age (years)	44.8	14.4	45.2	45	50
Occupation level (Duncan scale)	45.5	28.8	44.5	19	93
Income ($1,000/year)	19.9	11.8	16.1	14	65
Distance from Morgantown (km)	575	668	231	0	2,470
Number of hobbies	2.6	0.8	2.5	2	3
Number of organizations	2.5	0.8	2.2	3	4
Number of children	3.2	1.4	3.2	4	6

their participation. Interviews lasted, on average, 2.5 hours. Table 2 summarizes some characteristics of our informants.

We explained to our informants that we had complete life histories of 50 people from around the U.S., but that targets' names and characteristics had been shuffled in order to protect anonymity. Targets were presented to informants in random order. (We shall use the term, sequence, to mean when, from one to fifty, a target was presented to an informant.) Informants were to ask us questions about each target, until they felt able to make a choice. A choice was defined as someone who might know the target, or know someone who knew the target, or . . . etc. Informants were asked initially to explore any avenue they felt might be helpful, and to eliminate questions they found of no help as they went along. Many informants had difficulty in comprehending the task of matching a network member to a target. These people had to be taught how to play the game. We used non-explicit examples for illustration, in order to avoid biasing the informant: "After you have asked questions you will have a set of information about the target. Try to think of associations (whatever you think an association is) between the target and a friend or a relative, or an acquaintance of yours. You will probably want to pick the person you know who is somehow 'associated,' by your definition, with the target." With a few informants it was even necessary to illustrate the concept graphically. This was simply not an easy experiment to conduct; in collecting these data we may have channeled our informants' behavior in subtle ways which we have no way to control for.[3]

Figure 1 shows how informants rapidly adjusted to the experimental procedure, usually in the first seven targets. By about the eighth target,

Table 2. SOME DATA ABOUT THE 50 INFORMANTS

	Mean	Standard Deviation	Median	Mode	Range
Age (years)	41.9	15.1	42.5	30	59
Occupation level (Duncan scale)	42.9	24.5	43.5	19	75
Other occupation level*	34.0	24.6	19.2	19	61
Spouse's occupation level	49.8	26.4	51.0	19	94
Previous occupation level (first of several)	43.9	23.0	44.3	39	82
Income ($1,000/year)	17.4	8.6	16.2	13	25
Number of hobbies	2.8	1.0	2.8	3	4
Number of organizations	2.4	1.3	2.4	3	4
Number of children	2.6	1.6	2.2	2	7

*"Other occupation level" refers to any additional occupation the informant may possess (for whatever reason).

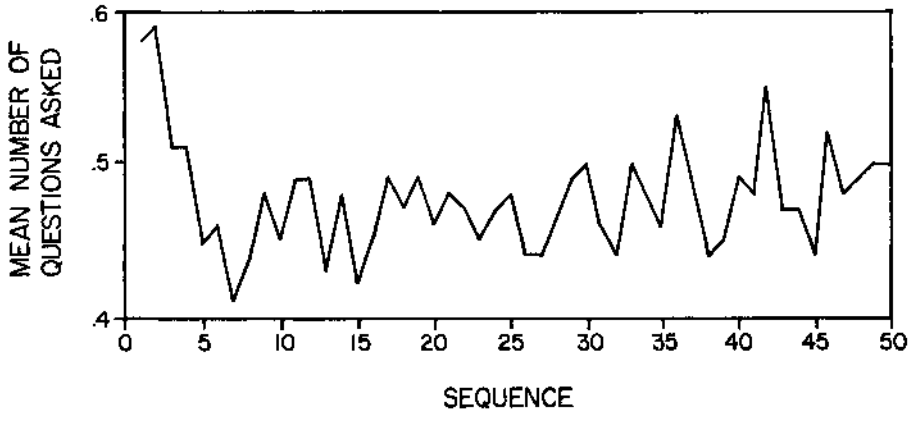

Figure 1. Mean number of questions asked about the target presented *i*th in sequence, for varying *i*.

informants settled down to asking a reasonably steady number (though not type) of questions. The number of questions asked per target did rise slowly towards the end of the sequence, as Figure 1 shows. However, in spite of our exhortation to eliminate unhelpful questions, many persons continued to ask questions throughout the experiment which (by their own claim) were never helpful. (Periodically, we asked informants why they kept asking questions which they rated as unhelpful. The typical response was that they were developing a "feel for the target," which allowed them to exclude many choices from their consideration, thus making the task of choosing more efficient.)

Of course, we did not actually have complete life histories for each target. The life histories were built up as the experiment progressed. Whenever an informant requested information about a particular target which was not already contained in the target's dossier, either the informant was told that the information was not obtainable, or the information was added to the target's dossier for potential use by later informants. This resulted in some inconsistent target characteristics. (For example, one target wound up with a father who had two distinct birthplaces.) If a piece of information was added to a target on the tenth informant, this meant that the previous nine informants had not requested the information. An example of a target's dossier, after 50 informants, is shown in Table 3.

Pretesting revealed that informants might ask some very narrow questions. This forced us, on some occasions, to interpret the informants response. For example, one informant asked if one target played the guitar. We interpreted this as a request for information about the target's hobbies, and we told the informant so. Some informants asked questions at the beginning of the interview which suggested that they did not understand their task. Such questions as "What color is the target's hair" or "What

Table 3. A TYPICAL TARGET DOSSIER, AFTER ALL 50 INFORMANTS HAVE ASKED QUESTIONS

*NAME:	Benjamin S. Clay
*LOCATION:	Charleston, West Virginia (South Charleston)
BIRTHPLACE:	Columbus, Ohio (July 12)
*OCCUPATION:	Papermill laborer
*PLACE OF EMPLOYMENT:	Wilhelm Paper Co. (worked there four years)
INCOME:	$15,000
*AGE:	32
*RACE:	White
*RELIGION:	Catholic (not an active member)
*EDUCATION:	Graduated from Richwood High School
*HOBBIES:	Pistol-shooting (competition), hunting
*ORGANIZATIONS:	National Rifle Association, Committee for Handgun Control, United Paperworkers International Union
SERVICE RECORD:	Served in the Navy for four years; stationed in San Diego, California
*MARITAL STATUS:	Divorced
*SPOUSE'S OCCUPATION:	Was a housewife
*CHILDREN:	Two sons, ages 10 and 12 (both live with spouse in Frostburg, Maryland)
FAMILY:	Came from small family
TRAVEL:	Kentucky, North Carolina (Smokey Mountains)
PREVIOUS OCCUPATION:	Gas station attendant in Columbus, Ohio

*Denotes information allocated at the beginning of the experiment.

kind of car does the target drive" were handled by giving the informant an answer and letting him or her judge the information's usefulness. In most cases, informants needed further explanations and stopped asking such questions. Subsequently, the questions were not recorded. If an informant insisted that a piece of information was useful, however, it was recorded as a question. An example of this was an informant who asked the exact birth date of a target. When presented with a birth date, she proceeded to make a choice on the basis of astrological signs. This question was then recorded as "used."

Each question was assigned a unique identifying number. As new questions were asked in the actual experiment, each was assigned a number. Table 4 presents a list of the twenty questions most frequently asked during pretest and test phases.

For each target the procedure was as follows: as each questions was

Table 4. MOST FREQUENT 20 QUESTIONS ASKED BY INFORMANTS (WITH A MNEMONIC OF FOUR LETTERS)*

		Most Often Asked	Usefulness of Question		
			Most	Somewhat	Not Useful
What is the target's occupation?	(OCCN)	1	1	2	1
Where does the target live?	(LOCN)	2	2	1	2
What is the target's age?	(AGE)	3	11	4	3
Is the target male or female?	(SEX)	4	12	3	4
What is the target's marital status?	(MARI)	5	20	7	5
What is the target's hobbies?	(HOBS)	6	3	5	7
What is the target's name?	(NAME)	7	--	14	6
Where does the target work (i.e., name of company)?	(WORK)	8	6	6	9
What organizations does the target belong to?	(ORGS)	9	4	9	8
Where is the target's location near (i.e., the closest well-known urban area)?	(NEAR)	10	5	8	15
What is the occupation of the target's spouse?	(OCCS)	11	9	11	10
What is the target's religion?	(RELI)	12	7	10	11
Does the target have children?	(CHIL)	13	--	12	12
How far has the target gone in school?	(EDUC)	14	17	13	14
Where did/does target go to school?	(SCHO)	15	8	15	13
How many children does the target have?	(NCHI)	16	--	16	20
What is/was target's field of study in school?	(STUD)	17	10	--	--
Does the target travel?	(TRAV)	18	--	18	18
What are the ages of target's children?	(AGEC)	19	--	--	--
Where was the target born and raised?	(BACK)	20	18	--	17

*"Most often asked" refers to frequency of appearance among top 20 questions. Usefulness of the question was estimated by informants.

asked, its code number was recorded, preserving the sequence in which it was asked. When informants had asked enough questions, they stated their choice, defined to be someone who would act as an intermediary in the small-world process. Then they provided a few sentences which explained why they had selected a particular intermediary (for example, "because he's a real estate agent," or "because his girlfriend's father is a pharmacist," or "because she was a graduate student at _____ and because he (target) would have been there two years ago when she (choice) left.") Next, informants ranked the questions they asked by the degree to which the answer had helped them make their choice. Informants were required to select a first-ranked question for each target, and were given the opportunity to rank other questions (if they had *asked* more than one) second, third, fourth, or fifth, stopping as they felt appropriate. We reminded informants of all questions they had asked (but had not ranked), and inquired whether each had been helpful or unhelpful. Thus each question asked for each target was accompanied by a code indicating its degree of usefulness. The relationship (friend or relative) of each choice to the informant was also recorded.

After completing the test, each informant answered a questionnaire. This consisted of basic socioeconomic data, and a personal response to any question ever asked by the informant about any target. For example, if the informant ever asked where a target's spouse went to school, then (unless the informant were single) he or she provided equivalent information about his or her own spouse.

II. Coding

The experiment yielded four different sets of data: (1) information (which we created) about 50 targets; (2) information about informants; (3) information concerning the questions asked by informants; and (4) information about the informants' choices and those choices' various connections to the targets.

The target data were coded first, since they contain more information than the equivalent informant information. The informant data contain less information because informants were not asked to provide data about themselves on questions they never asked. By the end of the experiment, the known answers to each of the questions ever asked about any target (Table 4) were coded in a format which left room for every possible answer. Informant data were then coded using the same format as for targets.

As noted above, questions were coded in the sequence asked by informants, followed by a ranking of each question's usefulness, which was also provided by the informant.

Finally, we developed a scheme to code the information about

choices' relations to targets. This information was contained in the short (usually one- or two-sentence) explanations given by informants on why they made a particular choice. Originally, two concepts were introduced, the "direct hit," and the "associated hit." If an explanation revealed that a characteristic of a choice matched exactly to a characteristic of the relevant target, this was a direct hit. For example, if a target lives in Los Angeles and the choice for that target also lives in Los Angeles, then this counts as a direct hit. If, on the other hand, a target lives in Los Angeles and the choice lives in San Francisco, then if, and only if, the informant said she selected the choice on the basis of location, this counts as an "associated hit." Associated hits can occur for various reasons. If an informant says she chose a pharmacist in order to get to a physician because "they are both in the medical field," then this is an associated hit. Similarly, a farmer and a tractor salesman may be associated by occupation; a student choice may be associated with a college administrator; a choice who plays jazz trumpet as a hobby may be associated with a target who collects jazz records, and so on. The concept of "associated location" and "associated occupation" were introduced in our earlier model of the decision process (Killworth and Bernard, b). Our experience in this experiment has broadened the concept to include associations such as hobbies, organizations, religions, etc.

In fact, our experience with these data has shown that simple associations are not enough to describe all the connections which informants claim exist between their choices and the targets. This led to the "associated via" and "intervening choice" categories. Consider the case of a choice who is a coal miner linked, by an informant, to a target who lives in Kentucky. The coal miner may, in fact, live in Ohio. But if the informant says "I chose him because he is a coal miner and he could contact people in Kentucky where there are lots of coal miners," then we believe this is best described as "associated with target's location via choice's occupation." Some other examples include the following: "I chose her because she belongs to the Sierra Club and the target works for the Environmental Protection Agency," then this counts as "associated with target's occupation via choice's organizational affiliation." "I chose him because he does cross-country skiing and the target lives in Vermont" is coded as "associated with target's location via choice's hobby." "I chose him because he collects rocks and the target is a geology student" is coded as "associated with target's field of study via choice's hobby."

Finally, many of our informants were apparently thinking two steps into the problem when they said such things as "I chose him because his girlfriend worked at Kroger's grocery and the target owns a grocery store." This counts as "associated with target's occupation via intervening choice's occupation." The *choice* was not associated with the target by any characteristics of his own; but his girlfriend (whom the informant may not have known well enough to name as his choice) is associated with the target's

occupation. For simplicity, we code the fact that the girlfriend is an intermediary choice, and that she is somehow associated with the target's occupation. Another example is the following: "I chose her because her father used to be a professional hustler. He could contact the target who likes to play pool." This was coded as "associated with target's hobby via intervening choice's occupation."

We stress that the four categories introduced above were sufficient to code *all* the 2,500 (50×50) reasons given by informants. We have not attempted to measure any inter-rater reliability for these codings (which were overwhelmingly handled by one investigator). As a measure of reliability, less than 2 percent of the codings required any discussion between the investigators. So bias, which must exist, should at least be consistent.

III. Questions (How Do People Know Each Other?)

A total of 82 different questions were asked by informants. (This does not include six questions which were asked only once, each by one informant.) Obviously, some questions were asked more frequently than others. Table 4 shows the most frequently asked twenty questions, together with a mnemonic for use in this paper (so that "where does the target live?" will be referred to as LOCN in what follows). Many of the less frequent questions were highly specific: "How long ago was the target divorced?" or "Is the target's child/children enrolled in a day care center?"

Figure 2 shows the probability that the frequent, and some less frequent, questions are ever asked. Note the dominance of OCCN (asked 92 percent of all occasions) and LOCN (90 percent). Other questions were asked much less frequently. The most commonly asked of these are AGE (asked 42 percent of the time), SEX (36 percent), MARI (24 percent), and HOBS (21 percent) (Note that this list is not consistent with standard SES variables employed by social scientists to define groups of people in our culture.). These probabilities can also be interpreted as a fractional contribution to the total number of questions ever asked, with OCCN and LOCN at 19 percent each, contributing 38 percent of all questions ever asked.

The final curve in Figure 2 shows the cumulative effects of the contributions. Four questions (OCCN, LOCN, AGE, and SEX) account for more than 50 percent of all questions ever asked. Ten questions account for more than 75 percent of all questions ever asked; eighteen questions account for 90 percent; twenty-five questions account for 95 percent.

Figure 3 shows a similar curve, restricting attention to the case when questions were declared by informants to have been most helpful to them in making a choice. Figure 3 shows that two questions (OCCN and LOCN) account for 64 percent of all "most helpful" (i.e., top ranked) questions. When HOBS and ORGS are added, over 75 percent are accounted for.

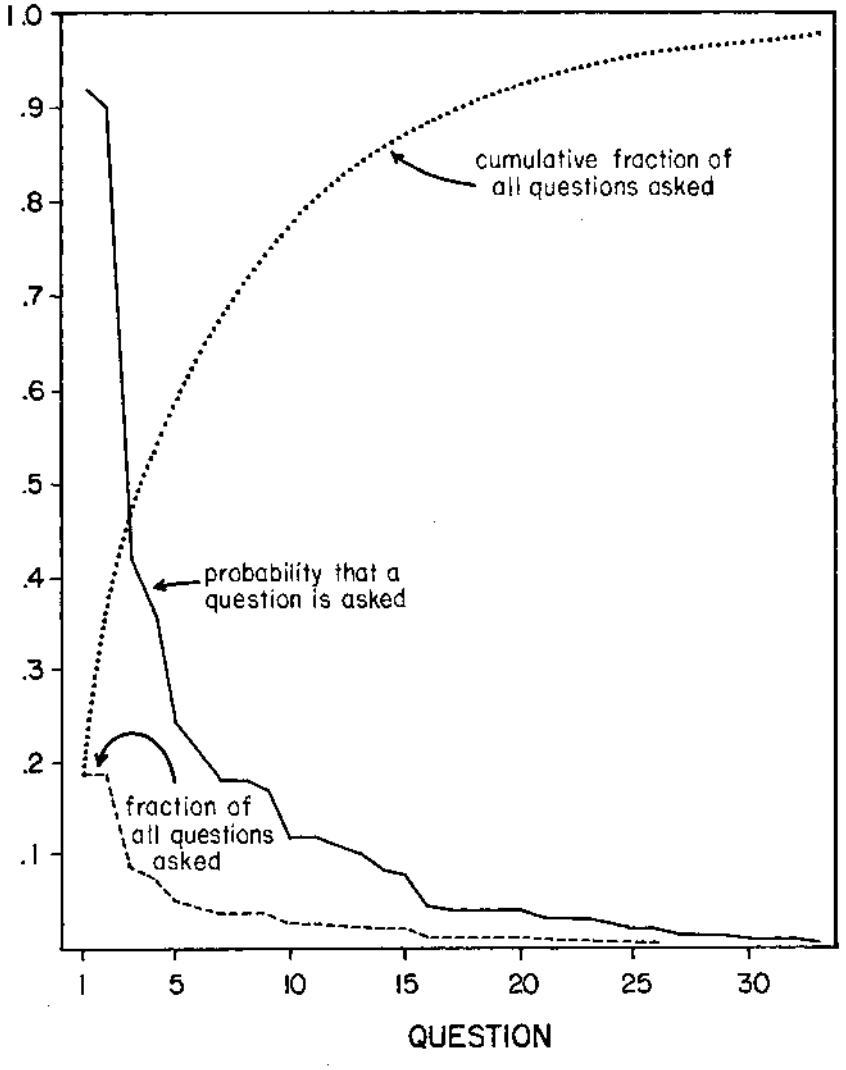

Figure 2. Question usage, by questions. The numbering of questions is that in column (a) of Table 4 (i.e. the most frequent two questions asked are OCCN and LOCN).

Another 5 percent is accounted for by NEAR; WORK accounts for 4 percent; and then the curve drops off.

The picture is changed subtly when we consider questions which are graded as at all helpful (not necessarily first-ranked) by informants. Again, LOCN and OCCN dominate (in that order, both more than 22 percent of the time). However, eight questions are required in order to

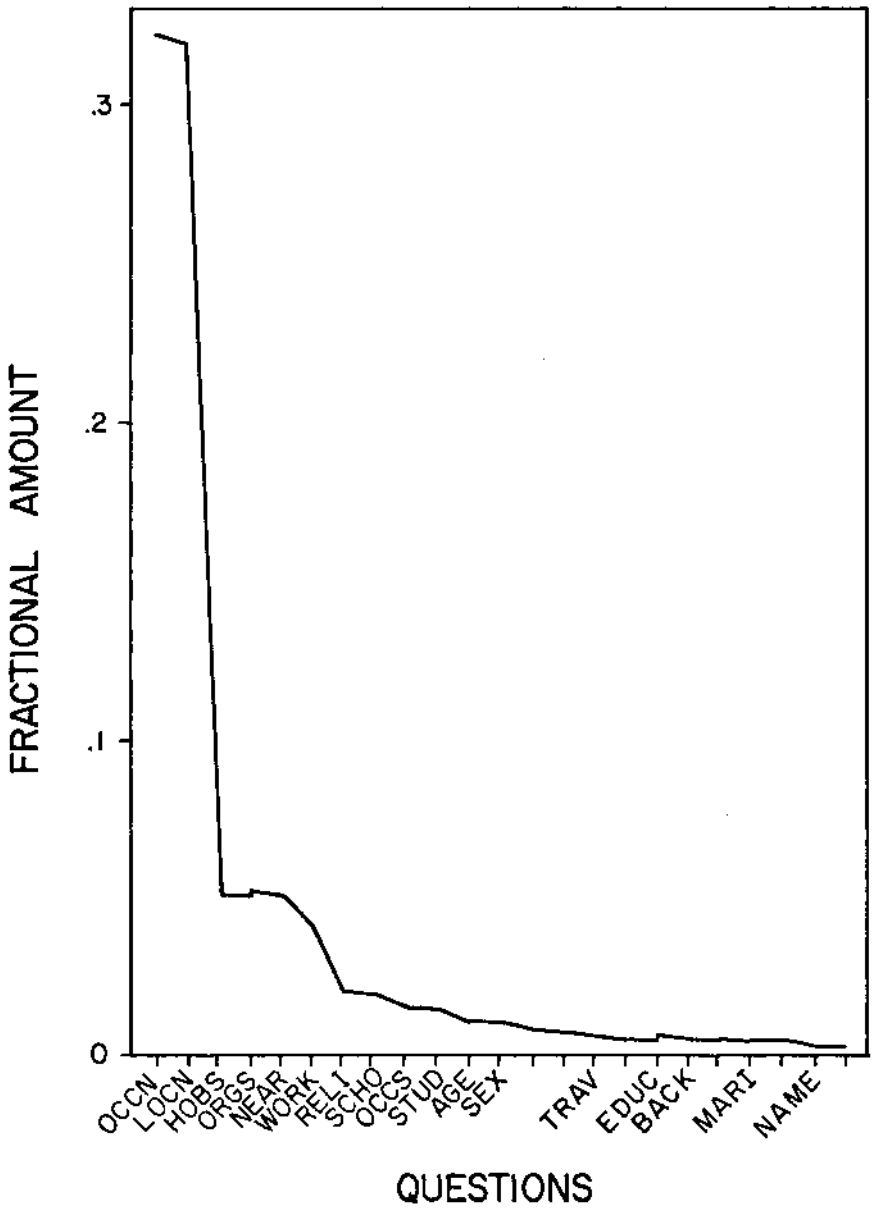

Figure 3. Fractional amount of questions deemed "most useful" by informants.

account for 75 percent of "at all helpful questions." The relative order of questions is also given in Table 4. It is perhaps surprising that the distribution of questions graded as "unhelpful" is largely the same as for those graded "helpful" (Table 4), with again OCCN and LOCN far above other questions (but only accounting for 15 percent each). This suggests that people tend to ask the same questions about all targets.

The number of questions asked by informants varied greatly, as shown in Figure 4. The mean number of questions in a string was 4.8, s.d. 2.7; but note that one informant once asked a string of 21 questions before making a choice. The mean number of questions asked by informants differs significantly** between informants, from 1.4 to 9.6. (Henceforth, single asterisks denote 5 percent significance levels; double asterisks denote 1 percent or better.) Similarly, some targets required significantly** more questions than others, from a minimum (average) of 3.4 for a target in Youngstown, Ohio, to a maximum of 5.5 for a target in Morgantown.

The length of a given question string, of course, depends on the difficulty the informant had in making a choice. In fact, there is strong evidence that if neither LOCN nor OCCN are very useful in a given case, the informants ask many more questions in an attempt to find some basis

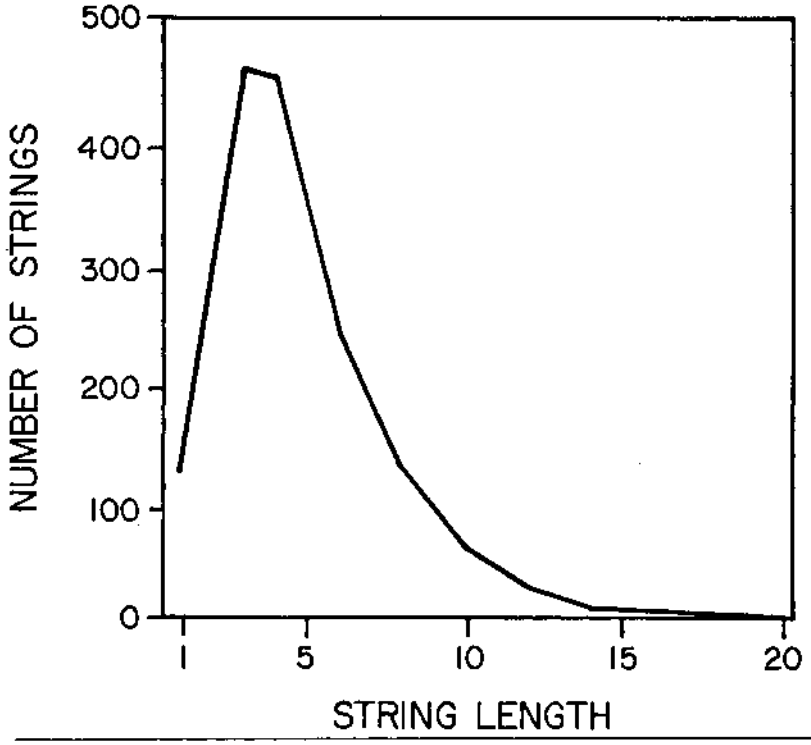

Figure 4. Distribution of lengths of question strings.

for making a choice. On the other hand, question strings in which LOCN or OCCN were ranked first or second most useful, on average, are significantly** (about one question) shorter than strings where this was not the case.

When informants had difficulty making a choice, they tended, not surprisingly, to stop when they reached the most useful question. Question strings ending with the most useful question were significantly** shorter, by 1.2 questions, than strings in which the most useful question was asked before the end. A different analysis gives similar conclusions: long strings (those containing six or more questions) end with the most useful question significantly** more often than would be expected by chance.

There is evidence that informants become set in their habits of asking questions, even when their own results suggest they should change. (This is an experimental phenomenon which obviously biases our results. Hence the randomizing of the order in which targets were presented.) The mean number of *different* questions generated by an informant, over all 50 targets, was 19.7, s.d. 8.9, although one informant asked only four different questions, and one asked 40. During interviews several informants were bothered by their inability to think of questions to ask. This led to the development of a basic set of questions which these informants used over the 50 targets. They often felt no reason to change their set of questions for a different target, as a new set of probabilities for a match was presented each time. (Note that the mean number of different questions generated by all informants per target was 28.3, s.d. 3.1. This much narrower variation per target than per informant suggests that the total amount of information needed by any informant for any target is remarkably uniform.)

Similarly, the correlation between the point in the sequence when a given question was first asked by an informant, and the percentage of time it was asked thereafter, is significantly** negative. In other words, questions asked about the first few targets tend to be used for most targets; questions which were asked for the first time for, say, the thirtieth target are *not* frequently used after that.

IV. Sequence of Questions (How Do People Know Each Other?)

The position of a given question in a string depended heavily on the particular question. LOCN is highly likely to be asked first or second (probabilities 0.35, 0.12 respectively), but much less likely thereafter. OCCN is almost equally likely to be asked first or second or third (about 0.25) but hardly ever after that. Questions such as HOBS or ORGS are almost never asked first, or second, but frequently occur further down the strings.

It is straightforward to define the most likely question string. Suppose we consider only strings beginning with LOCN. This may be useful

or non-useful for the informant. In each case a "most likely" second question will be selected; whether this is useful determines the third question, and so on. Figure 5A shows the sequences, and associated probabilities, for such strings. Sequences beginning with OCCN are almost identical, but with LOCN and OCCN interchanged.

A clear pattern emerges, with NEAR and WORK (Where is that near? Where does *T* work?—location questions) asked if LOCN is not useful but OCCN is. Questions like HOBS and ORGS are asked when one or another of LOCN or OCCN is unhelpful. These strings are usually quite

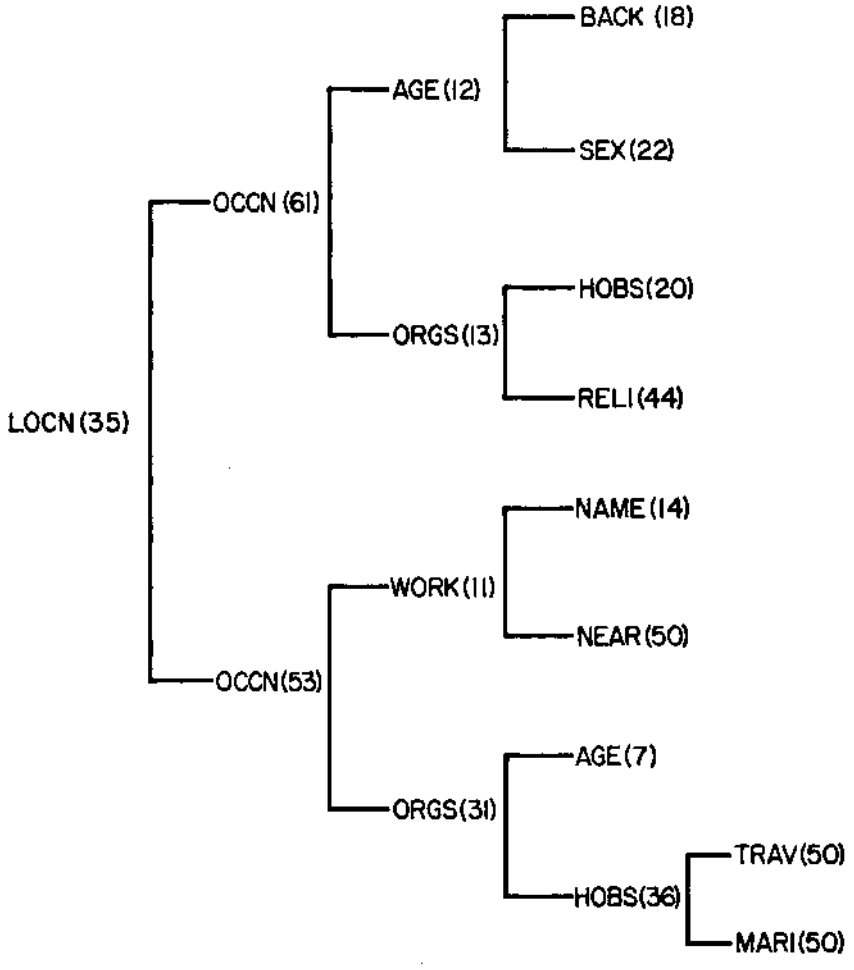

Figure 5a. Most likely sequence of questions for strings beginning with LOCN. Figures in parentheses are percentage probabilities of questions being asked, given that their predecessors were asked and were either useful or non-useful (a useful response branches upwards; a non-useful downwards). Sequences end when all potential strings are exhausted, or after 6 questions.

short, as indicated both in Figure 5A and in the previous section. Figure 5B shows a similar sequence for strings beginning with SEX. After asking AGE, the informant normally proceeds to OCCN and LOCN, before moving into sequences similar to Figure 5A.

There is, of course, a strong causal link between certain questions and those immediately following: "Where does the target travel?" is almost always preceded by "Does the target travel?" and can only be asked if the latter question received an affirmative answer.

Some of the causal links were not as strong as might have been anticipated. Examination of the data showed that this is caused by infor-

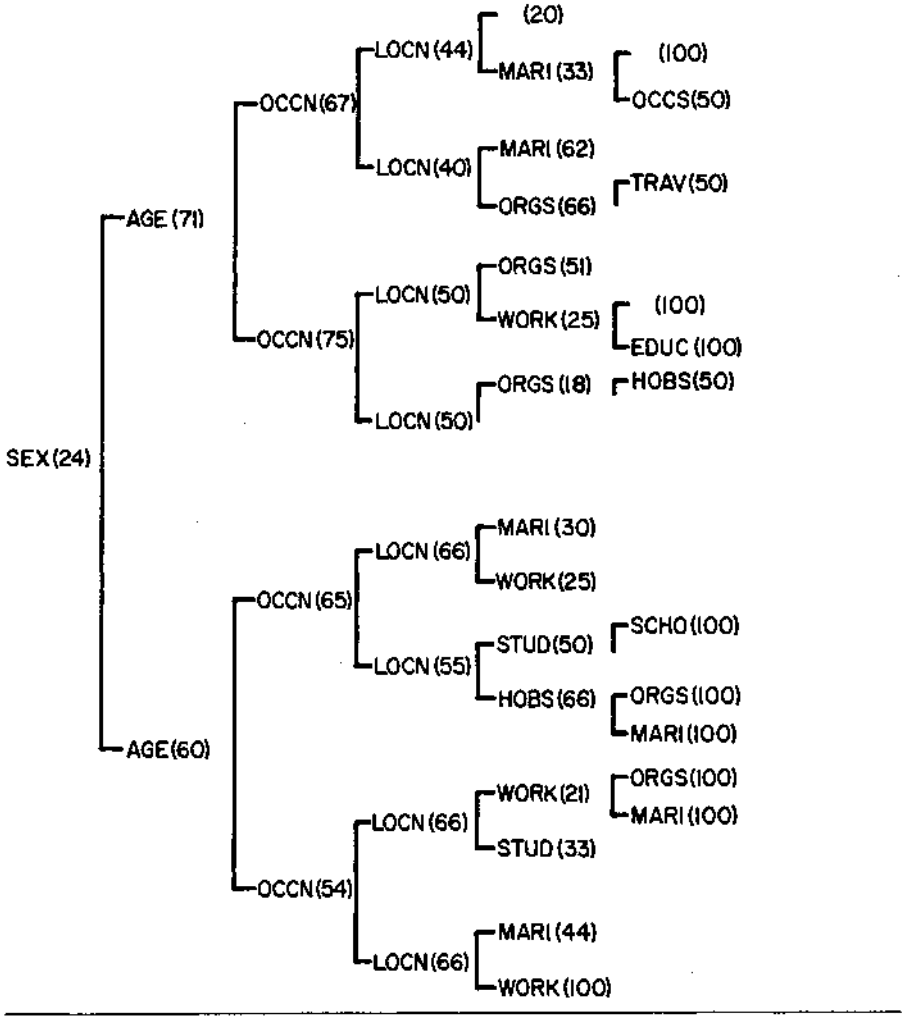

Figure 5b. As for Figure 5a, but for strings beginning with SEX.

mants shuffling the orders of some of their questions. This may be due partially to informants' attempts at breaking the monotonous routine of the experiment. Some felt that part of the game was to skip needless questions ("Why ask if a target has children when I can just ask the ages of children and get more information?"). Some informants even intentionally varied the order to their questions in order to avoid "getting in a rut." Examination of what questions preceded (anywhere in a string) a given question showed that there were 51 separate questions which could only occur after one or more questions of the set (NAME, AGE, OCCN, LOCN or MARI) had been asked. To some extent, of course, this reflects the strong probabilities of LOCN and OCCN being asked near the beginning of a question string.

Questions tended to be asked in groups about a specific subject. For example, the 2,500 (50×50) question strings were treated as lists of 82 integers. The jth integer in a string was one or zero, depending on whether question j was asked in that string. Factoring these strings produced the questions which tend to occur together. Nineteen factors were found; several of these had only one question with a high (varimax) loading on that factor. Define a group to be those questions with a factor loading of 0.2 or more on one factor (the "typical" loading is 0.02). The groups found can be described simply as: children; socioeconomic status; travel; family; spouse; schooling; more socioeconomic questions; social life; previous location; occupation; spare time; and three small sets of very precise details.

It is interesting to note that OCCN and LOCN do not occur in these groups, although LOCN has a high loading on factor 2 but with the opposite sign (i.e., when LOCN is asked, group 2 tends not to be, and vice versa). A similar factoring but on strings with -1 (question asked and useful) yielded very similar groups. Clearly, the factoring produced detail about questions hardly ever asked, and suppressed the patterns, if any, of the frequent questions. To avoid this, we examined the seven useful questions (OCCN, LOCN, AGE, SEX, MARI, HOBS, and ORGS). In any question string, each of these questions may or may not occur. There are thus $2^7 = 128$ possible combinations of questions (56 of which never occurred). The combination "LOCN and OCCN asked, others not" accounted for 29 percent of all strings with, on average, only one more question needing to be asked to find a choice. The six most used combinations accounted for 59 percent of all question strings; LOCN is asked in all six, and OCCN in five.

In summary, the strong patterning of questions which we have seen throughout this section shows that *a basic set of seven questions gives sufficient information about a target to enable an informant to make a choice. This list is LOCN, OCCN, AGE, SEX, HOBS, ORGS and MARI* (with reservations about this last).

Although many English speaking social scientists might consider the list obvious, this is the first time such a list has been produced experi-

mentally—that is, by asking informants to generate the list. It might seem that we could learn the seven pieces of information needed by Morgantowners simply by asking them. Sensitive ethnography would be easier and less costly. On the other hand, it is not clear to us that it would be easy for an American ethnographer, using open-ended interviewing, to isolate the list of reasons that people know one another in, say, highland Bolivia, or in Nairobi. INDEX, however, *is* an ethnographic technique, in the tradition of ethnoscience and ethnomethodology. It is fairly open ended, but appears to exhaust quickly informants' cognition about how they know people. It also requires ethnographic sensitivity to a particular culture in order to take INDEX data and produce a closed-ended instrument. Most ethnographic techniques, however, except INDEX, do *not* allow us to tell whether, say, occupation or location is the more important locator variable; or under what correlative conditions (gender, age, etc.) one is more important than the other.

Typically, researchers (including us) decide in advance what the important variables might be, and then test to see if they guessed correctly. In Killworth and Bernard (a), for example, we provided the race of the targets as cues for our informants. It would seem rather obvious that race is a determining factor in calculating social distance in the U.S. Yet, in our informant-defined experiment, neither race nor religion played any part in defining social distance. Our informants simply did not need the information. Thus, the list is not as obvious as it might appear to be at first glance. Of course, other tasks asked of informants (e.g., estimating social distance) could be expected to require race or ethnicity for their completion.

V. Accounting for Questions (How Do People Know Each Other?)

The previous sections described the questions asked by informants. We now seek to explain the variation in questions by referring to differences among informants and between informant and target characteristics. At the simplest level, there are 2,500 question strings, each asked by an informant (with known background data) about a target (also with background data). It is logical, therefore, to attempt to account both for the number of questions asked, and for whether a given question was asked, on the basis of these background data. If this can be done, this will give clues as to *why* questions are asked.

Multiple regression of the number of questions in a string with both informant and target data accounted for only 16 percent of the variance. Although this amount is significant**, many correlations later in the paper account for far larger variance. To save space, we henceforth choose to ignore any regression accounting for less than 40 percent of the variance.[4] Similarly, the number of different questions per informant or per target

was not well-predicted by personal characteristics. Discriminant analyses were conducted on frequently asked questions, in an attempt to predict for which informants and targets any given question would be asked. Target data are of little use in this matter: only the background data of the informants have much bearing on whether they ask a question. For three questions, AGE, LOCN, and SEX, the discriminant function correctly predicts whether the question was asked 67, 93, and 73 percent of the time respectively. These are to be compared with the (null hypothesis) prediction of 58 percent not asked, 90 percent asked, and 64 percent not asked (i.e., the proportions of the 2,500 cases when they were asked). In all three cases the structure of the discriminant function is similar: the higher the informants' occupation and education levels, and the more organizations they belong to, and the more hobbies they have; then the more likely AGE and SEX are to be asked, and the less likely is LOCN to be asked.

In general, however, we found that little variation in questions could be accounted for on a string-by-string basis. Instead, question strings were analyzed, first averaged over all targets (i.e., retaining only informant data with which to explain the variation) and then averaged over all informants (retaining target data).

There are a great many questions which could be asked of either of these data sets. In searching for signals in the data we chose to examine whether differences in dichotomous variables produced significant differences in question usage. For example, do male informants ask a specific question more than females? Then we attempted to account for differences in question usage by regressing question usage against characteristics of informants or targets.

There were 13 examples of significant** differences in question usage between informants, split into two subgroups by various criteria. Of these, many referred to infrequently asked questions. However, several features showed up clearly. NAME was asked more** often (11 times) by informants with children than those without (2 times), with a similar division (9.2 times to 1.5 times) when it was found not useful. Where a target traveled was asked (or found most useful, or found at all useful, or found not useful) significantly more** often by informants who reported that they did not travel. Presumably those who do travel are likely to have connections with the target's location, and therefore have no need to enquire further. There is, lastly, a clear difference in usage of LOCN between the sexes. Male, more than female informants find LOCN to be most useful when asked *first* in a string of questions. Conversely, females, more than males find LOCN to be most useful, when asked *last* in a string.

Multiple correlation of question usage by informant characteristics yielded very little additional information. No informant characteristics accounted for: the mean number of questions asked; the number of different questions asked; the probability of a given question being asked;[5] the

probability of a given question being most useful; or the probability of a given question being not useful. Only three multiple correlations did produce acceptable results (i.e., more than 40 percent of the variance accounted for over 50 cases; nearly all the 125 multiple correlations were *statistically* significant).

Target characteristics, however, apparently play a much larger part in accounting for question usage, although there are again only 13 significant** differences in question usage for different targets, as split into two subgroups by various criteria. There were several interesting features. SEX is asked more often** for female targets (do informants somehow get clues to a target's sex from other questions and then seek to confirm their feeling?). OCCS (spouse's occupation) was asked more often** about female targets, and conversely OCCN is asked more often** about males (do informants perceive a male's occupation as more likely to yield a choice?). The split of targets into those in urban and rural areas confirms the findings in the Killworth and Bernard (a) study namely: there was a significant** tendency to find OCCN the most useful question for rural targets and LOCN for urban targets. This, as we shall see, is also because of the characteristics of the choices made.

Multiple regressions showed that target characteristics accounted for most of the frequently asked questions. (In fact, multiple regressions of any question usage or question-related topic on target characteristics almost invariably yield significant* amounts of variance accounted for.) Table 5 shows the details of these fits (we present these details so that readers can examine the structures in these data). Several clear signals, again, confirming our earlier (a) findings, occur in these fits. For example, the probability that the frequent questions (other than OCCN and LOCN) are asked, increases with the distance of the target from Morgantown. OCCN tends to be most useful for targets far from Morgantown, in a rural location, with high occupation level. Conversely, LOCN tends to be the most useful reason when targets live in urban locations; however, high occupation level and distance also tend to increase the probability that LOCN will be most useful. Note also that NAME is most useful, given that it was asked, only for targets near Morgantown, as might be expected. For such targets, the likelihood of NAME being most useful increases as the target's socioeconomic status decreases. (NAME was used in one of three ways: primarily as an identifier for the next person in the (nonexistent) chain; second, if a choice had the same name as the target; third, if the target's name implied ethnicity which could be used as a criterion for making a choice.) HOBS and ORGS also yielded plausible fits; the more hobbies or organizations a target has, the more likely is the relevant question to be useful; the likelihood is also increased for targets living further from Morgantown.

In summary, *characteristics of targets control most of the questions which*

Table 5. TARGET CHARACTERISTICS RELATED TO THE FREQUENTLY ASKED QUESTIONS (SIGN IS DIRECTION OF RELATIONSHIP)

	Ordinal Variables								Dichotomous Variables (Opposite sign if Dichotomy Negated)				
Topic	Income	Occupational Level	Age	No. of Hobbies	No. of Orgs.	Education Level	Distance	Population Size	Male	Urban	Children	Active Religion	% of Variance Accounted For
No. of questions asked		-	+	+	+	-	+		-	+	-		44
No. of times RELI asked		-	+	-	-	+	+	-	-	+		-	46
" " " HOBS "		-	+	+	+	+	+	-	-	+		+	62
" " " SCHO "		-	-	-	+	-	+	+	+	-		-	52
" " " ORGS "		-	+	-	+	+	+	+		-		+	58
" " " TRAV "			+	-	+	+	+	+		+		+	56
No. times OCCN most useful		+	-	-	+	+	+		+	-			55
" " OCCS " "		-	+	+	-	-	-		-	-	+	+	51
" " NEAR " "		-	+	+	+	+	+		-	-	+	+	43
Probability NAME was most useful (given that it was asked)	-	-	+		-	-	-	+		+		+	62
" OCCN "	-	+	-	-	-		+		+	-		-	67
" LOCN "		+	-	+	+	+	+	-	-	+		+	48
" HOBS "	+	-	+	+	+	-	+	+	+	-			58
" ORGS "	+		+	-	+	+	+	-	-	+		+	52
" SEX "	-	+	-	+	+	+	+	-	-	+		-	46
Probability LOCN was second most useful		+	-		+	-	-		+	+	-	-	71
" NEAR "		+	+	+	+	-			-	-	+	+	51
" SCHO "		+	-	-	+	+	+		+	+	+	+	46
Probability OCCN not useful	-	-	+	+	+	-	-	+	-			+	52
" LOCN "	+	-	+	+	-	-	+		+	-		-	72
" BACK "	+	-	+	-		+	+	-					46
" HOBS "		-	+	-	-	+	+	-	-	+		+	54
" EDUC "	+	-	+	+	-	+	+	-	-	+		+	59
" SCHO "	-		-	-	+	-	+	+	+	-		-	58
Probability ORGS not useful	-	-	+		-	+	+	+	+	-		+	50
" MARI "	-	+	+	+	+	+	+	+	-	+		+	41
" TRAV "	+	-	+	-	+	+	+	-	-	+		+	53

informants ask; and characteristics of informants do not appear to have much influence on which questions were asked. Of course, on a one informant–one target basis, this is untrue (witness the discriminant functions). The signals only emerge on averaging over all informants or targets.

VI. Choices (Whom Do People Know?)

On average, informants used 40.7 different choices for the 50 targets (s.d. 4.9). This number is significantly** higher than the 34.7 different choices for the first 50 targets in our earlier study. The difference is of course due to the inclusion of ten very local targets in the current experiment. In fact, on average, 9.2 different choices (s.d. 0.9) were used for the ten local targets, suggesting that each informant has a large number of choices for local targets, as expected from intuition. Only two of the remaining 40 non-local targets, on average, had one of the local choices used for them. If locality did not matter to informants, this low number would occur by chance less than one in 10^{20} times. Hence local choices are only used for local targets.

Informants made male choices 67 percent of the time (which is significantly** higher than the 60 percent found in Killworth and Bernard (a), but reflects the same tendency to choose males). Informants made choices just because the choices "knew a lot of people" only 7 percent of the time (s.d. 15 percent). The distribution of these choices across *targets* is significantly** less scattered, suggesting that the decision to use someone who "knows many people" is a function solely of informants, and not of targets. Similarly, the number of intervening choices used per target varies significantly less** than per informant, so that the decision to use an intervening choice depends only on the informant.

Friends and acquaintances account for 80 percent (s.d. 11 percent) of all choices made, with family members accounting, of course, for the other 20 percent. These figures are almost identical to the 82 percent (s.d. 10 percent) found in Killworth and Bernard (a).

Again, qualities of informants and targets account for some of the variation in the above values. As in our (a) study, gender is the factor easiest to account for. Male informants make significantly** more male choices (37.9) than female informants (31.1); male choices are made significantly** more often for male targets (38.6) than for female targets (27.9); female informants used relatives as their choice significantly** more often than did males (11.6 to 7 respectively). We noted in the earlier study that females used family more than males did, but we had no way to know whether this would still hold when the target population included 20 percent local targets (10 out of 50).

The number of different choices made can be thought of as a crude indicator of the size of an informant's network. This number can be fitted

well (43 percent) by a linear combination of informant characteristics. It increases with informants' occupation and education levels, number of organizations and hobbies, and income; it decreases with age, and if the informant is active in religion and/or has children. The number of *male* choices made, decreases with the informant's age, occupation and education levels, and number of organizations; it is higher if the informant is male or has children (46 percent of variance accounted for).

The number of male choices per target may also be accounted for by target characteristics (53 percent of variance). The number increases with target's occupation and education levels, number of hobbies and distance from Morgantown; it also increases for male targets with children. Number of male choices per target decreases with occupation and education levels and number of target's organizations; and increases if the target lives in a rural area or is female (44 percent of variance).

This brief discussion shows that the types of people used for choices are well predicted if we know details about informants and targets. There is a clear differentiation in usage of choices on three dichotomous variables: local versus non-local targets; sex of choice; and the use of friends or family; statistical fits allow an excellent prediction of this usage.

VII. Reasons for Choices (How Do People Know Each Other?)

It is difficult to separate reasons totally from questions or choices, so that the degree of usefulness of a question, for example, has frequently been a feature of the previous sections. However, we can now extend the concept to include features of the choices discussed in the coding section: namely the direct hits, associated hits, vias, and intervening choices.

In the (a) study we found that, for any target, the most popular reason for choice was always location and occupation (but only "ethnicity" and "other" were the other possible reasons). The current data permit testing of this finding. Over the 50 targets, LOCN was the most popular reason for choice 23 times, and OCCN 25 times. Only twice were there any other most popular reasons: once OCCS, once WORK.

The finding is repeated if we consider the most popular reason for choice per informant. Twenty-one informants used OCCN most often, twenty-six used LOCN most often, and three informants used one each of AGE, HOBS, and ORGS most often. Hence the dominant role of location and occupation as overriding reasons for choice is confirmed.

The probability of a direct hit is surprisingly high; on average, there were 0.9 (s.d. 1.1) direct hits per informant–target combination. Certain questions are the most likely to be direct hits, e.g., LOCN (0.19 direct hits) and OCCN (0.15 direct hits), which therefore account for over one-third of all direct hits. In fact, on the 21 percent of occasions when two or more

direct hits occurred, LOCN or OCCN occurred 59 percent of the time (but only 9 percent of the time together).

Of course, Morgantown targets attracted very many (36) direct LOCN hits over the 50 informants. But little target information accounted for the large or small numbers of direct hits on any given question (indeed, the distributions of the numbers of direct hits over the 50 targets was usually random, on a chi-square test). Only the number of LOCN direct hits per target could be accounted for (75 percent of the variance**) over our 40 percent threshold: the more urban, and nearer, the target, the more likely for a LOCN direct hit. In fact, rural targets living far from Morgantown *never* had a LOCN direct hit; even rural targets living a medium distance away only had such a hit less than 1/2 percent of the time.

There was a similar number of associated hits (0.95, s.d. 0.82) per informant–target combination, and associated vias (0.88). Again, occupation and location dominate: location was an associated hit 0.31 times, and a via 0.26 times; occupation was an associated hit 0.28 times, and a via 0.28 times. This pair thus accounted for over 60 percent of all associated questions. Again, little target or informant information accounts for variations in the number of questions. However, OCCN occurs as an associated hit significantly** more for male targets than for female targets. LOCN occurs both as an associated hit and as an associated via more** for targets who live in a rural area than for targets in an urban area (i.e., informants find a choice who lives in a city near the target). These findings agree both with other findings in this paper and those in our (a) study.

Finally, there were only 0.1 (s.d. 0.4) intervening reasons (and a similar number of intervening vias), so that these are *rare* events. Again, LOCN and OCCN dominate the usage (0.03 intervening reasons, in both cases or 59 percent of all occasions, and a similar number of intervening vias).

Hence, as usual, LOCN and OCCN are seen, not only as frequent questions asked, but also as dominant connections between choices and targets, as the *reasons* for choices being made. But little variation in these reasons is explained by informant or target information. In other words, we can predict (Section VI) the type of choice made quite well, but not the details of that choice which led to informant selection.

VIII. The "Tag" Concept and its Use in Predicting Choices (Whom Do People Know? How Do They Know Each Other?)

This paper began by seeking simple answers to the questions "whom do people know, and how and why do they know them?" So far we have been concerned solely with a description of, and some statistical derivations from, the data. Implicit in any form of data gathering must be either a

theory or a desire to create a theory. The need for even a preliminary theory of who knows whom is obvious. At present, we can account for little of the details about the people chosen by informants in INDEX. But even if we know *all* the people known by an informant (his or her total network, in other words), could we select from that network the person the informant would choose for a given target? This is important: if we are to understand how an individual makes a choice when presented with limited information about a target, we need to model the decision process in some way that allows testing. The model we present here is very simple, and surprisingly successful in predicting the choices made by each informant.

We shall assume that an informant selects a choice for a given target because, in some sense, the informant perceives the choice to be similar to the target. Furthermore, we assume that if there are several choices who resemble a target, the informant chooses the most similar such choice (however this may be evaluated by the informant). In other words, the dissonance between the choice and what is known about the target is minimized.

How is similarity between choice and target to be measured? Clearly, the actual decisions involve highly complex cognitive processes about which we understand little. As a simplification, therefore, we assumed that a choice and a target are perceived as similar if and when some facet of the choice (e.g., where the choice went to school) and some facet of the target (e.g., where one of the target's children attends school) are either connected or, at best, identical. In some cases, of course, we had to suggest this concept to informants; this inevitably must weaken the following case slightly.

We shall term each facet of a target's personal history a "tag." Although targets began the experiment with very few tags (see Section I), as the experiment progressed and more data were invented for each target, the number of tags grew. Of course, the nature of the tags differed widely, as did their coding within the data. In order to count and catalogue the target tags, we again simplified the problem. We treated all tags *coded in location style* as location tags (i.e., target's location, previous location, family location, where the target travels, etc. are all location tags). Overlaps were removed (so that if a target still lives where he was born, only one tag is created). Similar tags were counted for occupations, hobbies, organizations, age, sex, and religion (the latter three, for targets, consisted of one tag apiece, of course).

Targets developed many tags during the experiment (from 11 to 23 tags), with a mean of 15.7, s.d. 2.7. The numbers of OCCN, HOBS and ORGS tags were all similarly distributed, with means of 2.5 to 3 and s.d.s about 1, while on average each target possessed 4.7 LOCN tags, s.d. 2.2. Now, on average about 1–2 hobbies and organizations, together with one

occupation and one location tag had been created for each target before the experiment. Hence about 2 more OCCN and 4 more LOCN tags were created by questioning, again demonstrating the need informants felt for information on these topics.

We lacked directly comparable data about choices; collection of such data would have presented enormous complexity and was not attempted.[6] Instead, we chose to *deduce* choice data by using the reasons informants gave for each choice. Whenever a choice achieved a direct, associated, or intervening match with the target it was chosen for (possibly several targets), that piece of target data was added to the list of tags for that choice. For example, if a choice was "connected with Los Angeles" at some time, that choice henceforth possessed a "Los Angeles" LOCN tag.

The number of choice tags (again, only distinct ones are counted) is much fewer than for targets. The number ranged from 0 to 12 over all informants and choices, but was strongly peaked about 1 and 2, giving a mean of 1.8, s.d. 1.2. Split into categories (Figure 6), the dominance of LOCN and OCCN tags is again clear (mean numbers of 0.75, 0.62, compared with at most 0.14 for all other tag types).

We can now test the simple hypothesis that, for any given target, an informant will choose the choice that has the largest number of matching tags with that target. (This procedure is of course biased by the way we obtained the choice tags: the correct choice, for a given target, almost invariably possesses some tags in common with that target. However, we will allow for this statistically below.) We define two non-location tags to

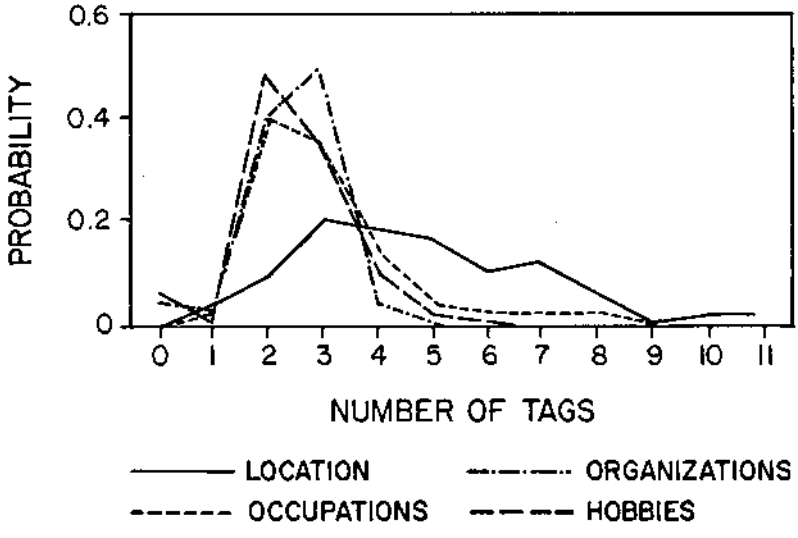

Figure 6. Probability of a choice possessing a given number of tags, distributed by different category of tag.

match only if they agree completely; in other words, an occupation tag of "symphony orchestra conductor" does not match with "symphony orchestra player." Location tags match if the choice and target location tags correspond to positions in the U.S. less than some cutoff distance apart. These cutoffs were taken to be 444 km, 222 km, 111 km, etc., down to 7 km, and a final cutoff of 0 km (corresponding to a location match only if the two locations are identical).

Thus, for each distance cutoff, and for each informant–target combination, we can nominate the "optimal" choice(s) as being the choice(s) which has (have) the most tags in common with the target, and then compare the optimal with the actual choice. We have defined two ways to measure the accuracy of this procedure, which we term the "easy and difficult scores." The easy score is defined as unity whenever the actual choice is among the optimal choices, and zero otherwise; the difficult score as 1/(no. of optimal choices) if the actual choice is among the optimal choices, and zero otherwise. In other words, the easy score counts how often the actual choice was correctly (but not necessarily uniquely) predicted; the difficult score counts how often we would be correct if we chose at random from among the optimal choices.

The results of this are shown in Table 6, averaged over all informant–target combinations. Two things are immediately obvious. The first is that maximal accuracy is obtained when the location matches exactly, although there is little degradation if two LOCN tags are up to 28 km apart. Henceforth we shall treat LOCN tags like all others, and require an exact match. The second is the high rate of accuracy: the actual choice is among the optimal choices 89 percent of the time, and predicted 60 percent of the time even if the choice is selected at random from among the (possibly several) optimal choices.

The high success rate, combined with the simplistic approach, suggests that we might improve the accuracy still further if we weighted the tags in some fashion before counting the matches. We examined eight different weighting combinations, shown in Table 7. Note that *no* weighting achieves the accuracy of the simplest counting scheme; also, the overall importance of LOCN and OCCN is confirmed by scheme 3's scores of 0.52, 0.82. We are forced to conclude that if the model has relevance to the decision process, then all tags should be counted, independent of their

Table 6. AVERAGE ACCURACY SCORES OBTAINED BY PREDICTING THAT INFORMANT WILL MAKE THE "OPTIMAL" CHOICE

Cutoff (in kms)	444	222	111	56	28	14	7	0
Difficult score	0.31	0.37	0.48	0.54	0.57	0.59	0.59	0.60
Easy score	0.53	0.63	0.77	0.85	0.87	0.89	0.89	0.89

Table 7. AVERAGE EASY AND DIFFICULT SCORES FOR VARIOUS WEIGHTINGS OF TAGS

Weighting	Difficult Score	Easy Score
1. Only LOCN tags counted	0.34	0.59
2. Only OCCN tags counted	0.33	0.49
3. Only LOCN and OCCN tags counted	0.52	0.82
4. Only LOCN tags counted; tags which are direct hits are given double weighting	0.32	0.46
5. As 4, but for OCCN	0.33	0.49
6. As 4, but for LOCN and OCCN	0.47	0.70
7. Only non-(LOCN or OCCN) tags counted	0.19	0.31
8. All tags counted; LOCN and OCCN weighted as in 4	0.55	0.77

directness or lack thereof. In linear programming terms, then, all tags have an equal *utility*.

Informants occasionally asked about schooling. In our coding, schools were not allocated a precise location (i.e., they were not given Cartesian coordinates) but were recorded by state and an identifying arbitrary code. Including schools as a separate tag might improve accuracy. However, if the school's *state* is used as a tag (so that 11 schools in the same state are identical for predictive purposes) this weakens the scores to 0.59, 0.87. Using the school's unique *code* improves the accuracy, but only by 1 percent, to 0.60, 0.90 for difficult and easy scores respectively. Hence inclusion of schools is of no real help in predicting choice.

The difficulty with interpreting these results stems almost entirely from the biased way the choice tags were obtained. It seems intuitively obvious that if one obtains some choice tags from the target for which that choice was made, then that choice is likely to be the one with the largest number of tags matching with that target. Clearly, we need to estimate how likely it is that we achieve the levels of accuracy observed in our data.

This calculation is given in the Appendix. It is shown that the easy score (89 percent) is significantly** higher than the 86 percent expected by chance; the difficult score (60 percent) is likewise significantly** higher than the 27 percent expected by chance. Hence, although the method is biased towards a high easy score, the tag model is giving excellent results. The 89 percent success rate of the straightforward counting procedure, although biased, obviously accounts for a large amount of the decision process. Nonetheless, the tag concept is a simplistic one. A detailed ethnographic study of how informants made a selection between choices of apparently equal utility would be very valuable.

The number of tags differs strongly between targets; and the total number of choice tags differs between informants. We examined whether characteristics of either targets or informants could account for this variation. Only the number of LOCN tags possessed by a target could be well accounted for by target variables: 45 percent of the variance could be accounted for by a linear combination of socioeconomic variables. The largest contributors are targets' age and occupation level: the higher the target's age, and the lower his or her occupation level, the more occupation tags that target possesses. This is plausible: too low an occupation level forces informants to search for occupations related to that target.

Similarly, informant data account for little variation in the total number of tags (not necessarily distinct) which their choices possessed. Each informant had a total of 75 tags on average, of which 29 were location and 25 occupation. The only significant result is that female informants have more* occupation tags than male informants, by 28 to 21. Thus, *little about targets or informants accounts for variation in tag density.*

Another, more subtle, bias in our use of tags is the degree of utility of each tag. Because of the manner in which choice tags were deduced, most of them have, in some undefined sense, a high degree of utility for that informant. Thus, with hindsight, equal tag weighting is likely to yield the most accurate results.

Clearly, not all tags are really of equal utility: it seems plausible that a choice currently living in Chicago is more likely to be chosen for a Chicago target than, say, a choice whose father travels to Chicago. Given the limitations of our experiment we could not test for this. However, this does suggest a variety of informant-defined experiments both to find out what weighting of tags and tag types is necessary to yield accurate prediction of choices and to find what other qualities of informants and/or targets are important in determining why some targets have stronger ties with some informants than with others.[7]

Conclusions and Discussion

We began this experiment with the intention of finding out two things: what informants need to know about a target to make a choice, and how they make that choice. To be sure, the experiment possessed two drawbacks. First, the task ("tell us who you know who is most likely to know this target") is artificial, as are most ways of gathering data in social science; an interview, as here, is at least more natural than a questionnaire. However, it is straightforward to modify the experiment to make the task less artificial. Instead of selecting choices on the basis of knowing a target, members of an informant's network could be selected for a specific task relevant to the particular culture: "Who would you talk to about bailing

your child out of prison?" or "Who would you talk to about taking out a loan?" etc. The tasks should ideally be defined by the members of the culture, and not predesignated by researchers.

Second, the list of seven facts appears to be devoid of meaning. On the other hand, the list of possible social structures, defined from functional networks, is endless, for all practical purposes. A systematic-cross-cultural study of any particular network might be a waste of time. How could we tell, *a priori*, which networks are the important ones? We believe that the INDEX technique addresses this problem. We assume that any given network relation (borrowing money, babysitting for, being kin . . .) requires that people know one another or *of* one another. INDEX, then, subsumes functional networks at a level of generality which makes systematic comparison possible. It produces a high enough level of generality so that it can be both exhaustive and rather short. (We predict that all such lists, irrespective of the variation in cultural content, will be around seven in industrialized states, and around three in hunting/gathering societies.)

The third drawback in the experiment is that the information produced is totally *cognitive*. There are no behavioral data being gathered. In other words, while we can realistically assume that if *A* chooses *B*, then *A* does (behaviorally) know *B*, we cannot assume that *B* is the best route from *A* to a target, however this might be defined. We do not think that this is a serious problem. The fact that *A believes B* to be the best route will generally imply that *A* would choose *B* were some behavior (e.g., folder-passing in small-world studies) to be involved. Best route or not, it will be used because it is perceived as best.

We were reasonably successful with our first aim, to find out what an informant needs to know about a target to make a choice. The set of questions LOCN, OCCN, AGE, SEX, HOBS, ORGS and MARI—with reservations about MARI—seem to contain virtually all an informant needs to know about a target. We reiterate that such a set is not trivial; repeating the experiment in other cultures will almost certainly produce a different set of questions. What, we wonder, would the union of all the sets of questions over many cultures contain? Would there be few or many questions? The main reasons for choice were LOCN and OCCN. Four categories enabled us to code each of the 2,500 verbal reasons in a uniform manner. (Again, would there be different categories in other cultures or for other tasks?) Characteristics of the target appear to control which questions were asked by informants, in the main. Finally, the simple "tag" model was highly successful in accounting for which choices were made for given targets.

We were less successful, notwithstanding, with our second aim, to understand how informants make choices. To be sure, if we know all the tag details of all an informant's choices, we can predict which choice the informant will make for each target. But we cannot account for the variation

in tags between informants. Similarly, we can usually predict accurately the likelihood of the basic set of questions being asked, or being useful, or being a direct hit, or whatever, knowing target details. But we cannot account for the differing number of questions between informants, or the differing number of choices. In other words, we cannot ask a simple list of socioeconomic questions of informants and be able to say anything about their immediate networks of choices; too much appears to depend on personal history of informants. So our goal of constructing a model of social structure, based on firm data about who knows whom, and why, remains tantalizingly elusive.

Notes

1. In Killworth and Bernard (a), 25 percent of the targets were housewives; the uniformity of response by informants to these targets led us to reduce the number of housewives for the present experiment. With hindsight, three housewives out of 50 targets are too few to produce reliable statistics.
2. "Cities" are defined (except for Morgantown) as places we felt informants anywhere in the U.S. would recognize.
3. As in all experiments, there is an element of unnaturalness in our procedures. In actual day-to-day activities, people go to others for information, for favors, to repay debts, etc. Compared to such naturally occurring behaviors, the requirement of small-world experiments that informants make a choice connected with a target may seem improbable. In fact, in both INDEX and our (a) study we remove the requirement that informants even make contact with their choice of intermediary. We argue that the choices made by our informants are no more or less hypothetical than those in almost any social science experiment.
4. This is rather unusual for the social sciences, perhaps. However, a scatter diagram of a regression accounting for 16 percent of the variance shows very little in the way of a signal; hence our suppression of low variance.
5. We examined only the 25 most frequently asked questions, which account for 95 percent of all questions ever asked.
6. Of course, we had a great deal of data about the choices in the one or two sentence explanations given by informants. However, these data were not collected with the idea of systematic comparison, and we simply did not know how to code the data for such comparison. It is obvious how to collect comparable data about choices; but this would have increased the time required for interviews by so much that we were forced to abandon this part of our original design.
7. To extend the investigation, we conducted a follow-up interview with informant 15. He provided two new sets of data. The first was a count of the number of connections or tags between each of his 33 choices and each of the 50 targets, now given *all* target information, rather than the limited information he requested during the original experiment. Then, armed with all the information, he told us which choice he would now make for each target.

This mini-experiment was slightly flawed for two reasons. First, in order to reduce the many hours of the follow-up, we had collated all locations and occupations relevant to each target. As a result, if the target lived in a small town but was attending college in a neighboring big city, *both* the big city ("the town is near X") and the colleges ("attends X college") were available as LOCN tags. This doubling-up of essentially the same information made interpretation of the data somewhat difficult. Second, the informant ignored the myriad of possible intervening choices, as we had requested. This automatically removed such reasons as "I choose Y, because she knows someone at Z oil company," which had been used in the original experiment. A more precise, informant-defined interview would be very valuable.

Informant 15 generated many tags between choices and targets. On average, between any choice and any target, he found 0.27 LOCN matches, 0.05 OCCN matches, 0.21 HOBS matches, and 0.06 ORGS matches, or 0.58 matches in all. Corresponding s.d.s are 0.3, 0.06,

0.21, 0.06, and 0.41, respectively. Note how much more likely it is that a LOCN or HOBS match occurs than an OCCN or ORGS match.

In the original experiment, the difficult and easy scores for the tag concept for informant 15 had been 0.50, 0.86 respectively. Repeating the calculation based on the more complete tag information reduces the accuracy noticeably to 0.44, 0.25 (although this is now unbiased, of course, so interpretation of the scores is somewhat altered). His final choices contained 17 alterations from the original set of choices; on three occasions he would now prefer to make a choice outside his initial 33. Rather surprisingly, the tag scores based on his *final* choices decreased slightly to 0.42, 0.24.

We examined the cases where simple tag-counting yielded the wrong choice. Almost invariably it was a matter of "how strong" a tag was: a choice living in the target's location (1 tag) being preferred, for example, to a choice who was born in that location and whose family still lives there (2 tags). Thus the original tag experiment had, as we suspected, automatically scaled the utility of most tags, and chosen the most useful ones. The follow-up interview, however, did not contain any utilities, and this probably accounts for the reduction in accuracy.

Appendix. Expected Tag Scores by Chance

To calculate this we need three sets of probabilities. The first, α_r, $r=0,1,2,\ldots,5$, are the probabilities that the actual choice has r tags matching the target. These probabilities are calculated from the data: the mean number of matching tags is 1.56, s.d. 0.84. The other sets are β_n, $n=11,12,\ldots,23$, the probabilities that a target has n tags and γ_m, $m=0,1,\ldots,12$, the probabilities that a choice has m tags; both are again known from the data.

We assume all tags are of a similar type (retaining the different categories would involve awkward partitions of integers, without adding significantly to the results). Let there be N tags in total (there are 451 different target tags in all: 126 location, 84 occupation, 80 hobbies, 109 organizations, 37 ages, 2 sexes, and 13 religion tags). We might take N as 451, or perhaps only (126 + 84), depending on our interpretation of the number of tags.

Now the probability that a random choice tag matches a random target tag is clearly $1/N$. If the target has n different tags, and the choice has m different tags, then the probability of exactly t matching tags is

$$q_t = {}_mC_t(n/N)^t(1 - n/N)^{m-t}$$

by the binomial theorem (the n/N factor derives from n chances of $1/N$, of course). Hence the probability of less than r matches, P_r is given by

$$P_r = \sum_{s=0}^{r-1} q_s$$

Thus, if the correct choice has r matches, the probability that another random choice achieves less than r matches, and is not optimal, is P_r, and the probability that a random choice achieves the same number of matches is q_r. The probability that another choice achieves *more* tag matches than the correct choice is $(1 - P_{r+1})$.

These probabilities, so far, are conditional upon the values of r, m, and n. Summing over r, m, and n, and multiplying by $\alpha_r\beta_n\gamma_m$, yields the probability P that a choice has fewer tag matches than the correct choice, and Q that a choice has the *same* number of tag matches (and, therefore, $R \equiv 1 - (P + Q)$ that a choice has *more* tag matches).

The expected value for the easy score E is then given, since there are on average 41 different choices, by

$$\mathrm{E}(E) = (P + Q)^{40} = \mathrm{E}(E^2)$$

which is simply the probability that all the other choices score less than the correct choice. The variance is then given by

$$\sigma_E^2 = (P + Q)^{40} - \left[(P + Q)^{40}\right]^2.$$

The difficult score is slightly more awkward to evaluate numerically. The probability that any other choice achieves the same score as the correct choice is

$$\mu_a \equiv {}_{40}C_a Q^a P^{40-a}$$

giving expectancies for the difficult score D as

$$E(D) = \sum_{a=1}^{40} \mu_a/a,\ E(D^2) = \sum_{a=1}^{40} \mu_a/a^2$$

and variance

$$\sigma_D^2 = E(D^2) - E(D)^2.$$

We can now compare the observed scores with those expected. With N=451 different tags,

$P = 0.922$, $Q = .074$
$E(E) = 0.859$, $\sigma_E = 0.348$
$E(D) - 0.271$, $\sigma_D = 0.204$

Over 2,500 cases, the observed mean easy score was 0.89. Dividing r_E by $(2{,}500)^{\frac{1}{2}} = 50$, we find that 0.89 is some 4 standard deviations above expected. Similarly, the observed difficult score of 0.60 is 80 standard deviations above expected. Thus both observed scores are significantly** higher than expected by chance, although it should be borne in mind that 0.86 is the expected easy score (i.e., the system is biased toward a high score).

Adjusting the effective number of tags only makes this conclusion firmer. Restricting attention to location and occupation, which, from Table 7, achieves difficult and easy scores of 0.52, 0.82 respectively, yields expected scores of 0.15 (s.d. 0.15), 0.65 (s.d. 0.48) respectively. Again, the observed scores are significantly** high.

It should be noted that we deliberately did *not* restrict the target's tags to what each informant knew about the target (i.e., we compared choice location tags with a target's places of travel, whether or not the informant had asked about the target's travel). This was to permit the other choices more chance of matching tags with the target. If we *do* restrict the target's tags to what each informant asked about, the mean difficult and easy scores rise still further to 0.78, 0.94 respectively. It is clear that this is too biased to be regarded as a fair test of the tag concept.

References

Bernard, H. Russell, and P. D. Killworth. 1979. "A Review of the Small-World Literature." *Sociological Symposium* 28(Fall):87–100.

Duncan, O. D. 1961. "Socioeconomic Index Scores." In A. J. Reiss, Jr. (ed.), *Occupations and Social Status*. New York: Free Press.

Killworth, P. D., and H. R. Bernard. a:1978. "The Reverse Small-World Experiment." *Social Networks* 1:159–92.

_____. b:1979. "A Pseudomodel of the Small World Problem." *Social Forces* 58(December):477–505.

Lin, N., P. Dayton, and P. Greenwald. 1978. "Analyzing the Instrumental Use of

Relations in the Context of Social Structure." *Sociological Methods and Research* 7(2):149–66.
Milgram, S. 1967. "The Small-World Problem." *Psychology Today* 1:61–67.
de Sola Pool, I., and M. Kochen. 1978. "Contacts and Influence." *Social Networks* 1:1–48.