Social Networks 12 (1990) 289–312
North-Holland

ESTIMATING THE SIZE OF PERSONAL NETWORKS *

Peter D. KILLWORTH **
Hooke Institute for Atmospheric Research

Eugene C. JOHNSEN
University of California, Santa Barbara

H. Russell BERNARD, Gene Ann SHELLEY
and Christopher McCARTY
University of Florida, Gainesville

Some methods for estimating the total size of personal communication networks are presented. All involve the scaling-up of a reported network size by a factor proportional to the number of people whom informants can recall when they are presented with a representative list of last names from a telephone directory. Estimates from Jacksonville, Florida give network sizes of 1700 ± 400; reevaluations of an estimate made for Orange County give 2025; and estimates from Mexico City give network sizes of about 600. The difficulties, and sources of error, in these estimates, are discussed. The estimates are compared with independent estimates based on the likelihood of informants knowing members of a small, countable subpopulation, which suggests for U.S. informants a network size of 1526. Thus consistent numbers are beginning to emerge, at least for U.S. informants.

1. Introduction

In a world increasingly connected by communications media and travel, it is perhaps surprising that we know so little about the actual process of communications themselves, or, indeed, who knows whom

* The research for this paper was funded by NSF grant SES 8803577. Without the work of many people, it would never have occurred. We especially wish to thank Alaina Michaelson for devotedly counting the 1984 Orange County phone book; Scott Robinson, Yolanda Hernandez, and Rosario Mata Castrejon in Mexico City for obtaining the data there; and Noah Friedkin for useful discussion.
** Hooke Institute for Atmospheric Research, Department of Atmospheric, Oceanic and Planetary Physics, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, England.

and for what reason. Remarkably, even such a simple statistic as the average number of people known to an individual on this planet remains far from guessed at, let alone measured. If the field of social network theory were at an advanced stage, such statistics as the mean number of friends, or acquaintances, or relatives, known to an individual would be easy to compute from the theory. A researcher would then be far more concerned with refining the details of the theory than calculating the statistics. However, the study of social networks is very young. Researchers, faced with the difficulties involved with obtaining reliable data from many informants about all details of their networks, are forced to study more manageable questions. Estimating the number of people an informant knows is one of those questions.

The problem is deceptively straightforward. Why not simply ask 1000 people to write down all the people they know, and then ask them about those people? We would learn a lot about personal networks, yet such an exercise has not been undertaken. Informants, after all, are human and can spare social science investigators only so much of their time – even assuming the task to be feasible. As a result, researchers have searched for *proxies* for the size of a person's network, rather than measuring the size of the network itself.

These proxies have been ingenious. They have involved personal diaries (Pool and Kochen 1978; Gurevich 1961); counts of names recognized in telephone books (Pool and Kochen 1978; Freeman and Thompson 1989); social surveys involving lists of only close friends, or members of support networks (Burt 1982; Fischer 1982; Wellman 1979); estimates based on the likelihood of informants knowing members of a small subpopulation of known size (Bernard *et al.* 1987; Bernard *et al.* 1989a); and more artificial (i.e. instrumental) studies such as the reverse small world (RSW) by Bernard, Killworth and others (Killworth and Bernard 1978; Killworth *et al.* 1984; Bernard *et al.* 1988).

The proxies proved somewhat frustrating. Estimates differed violently between the methods used. Insufficient estimates had been made on most methods to ensure a statistically reliable answer even for studies restricted to the U.S., let alone cross-cultural comparisons of network size. Pool and Kochen's (one-person) study of phone book recollection gave answers of 3100 and 4250 acquaintances depending on whether the Chicago or Manhattan phone book was used. Gurevich's 18 informants yielded estimates varying from 122 to 5053, with a mean

of 2130. The General Social Survey of 1985 asked informants to list up to five close friends, which precluded an estimate of network size. Support network studies (Fischer 1982) have suggested figures of 20 to 30. Data from Bernard *et al.*'s (1987) Mexican informants, using a method involving their knowledge of members of small subpopulations whose size is known, gave estimates varying between 173 and 810. The RSW studies, performed cross-culturally, gave lower results. Based on the number of persons triggered by a list of 500 target names as potential intermediaries in Milgram's (1967) small world technique, a uniform figure of around 130 was found for several studies in the U.S., and rather higher figures for Paiute Indians (256) and Ponape islanders (313). The number of persons generated can only rise as the number of target names presented increases, and crude estimates for the asymptotic number of people produced by this technique are around 250 for U.S. informants.

There have been several attempts to explain the differences between estimates. Freeman and Thompson (1989) – hereafter FT – note, for example, that the RSW method presents informants with a hypothetical situation. They argue that this makes the retrieval of persons in a network dependent upon the informant's strategy in answering the questions presented. Furthermore, informants can not produce more than 500 names in response to a 500-target RSW – although the number is usually much smaller. So even if the RSW figures are a measure of network size, it remains obscure how that measure could be scaled up to give an accurate estimate. Diaries of personal contacts can omit close friends who do not live near the informant; and so on. There is also the distressing possibility that informants whose data yielded a smaller estimate of network size simply had smaller networks than informants in other experiments.

It should be remembered, however, that we are seeking to explain *why* networks are the way they are. The concentration on *mean* network size in the literature may well omit the fact that the *variation* in network size between informants (even in the same culture) needs explanation as well. Perhaps we should seek the distribution and variation of both means and standard deviations of network size between cultures.

This paper reports on attempts to define more closely the sizes of networks of informants in Jacksonville, Florida and Mexico City, by the simultaneous use of several network elicitors on the same group of

informants. The main results of this study are given elsewhere (Bernard *et al.* 1989b). They permit us to ask whether informants who have a large support network, for example, also have a large RSW network, etc., and permit the examination of overlap between various networks. Here we restrict attention simply to the size of personal networks (with only the crudest attempts at standard deviation estimates). Sections 2 and 3 give brief details of the methods used, and of the informants. Section 4 discusses the rationale for two estimates of network size: a modified FT estimate, and an estimate based on overlap between network elicitors. It is shown that the original FT estimate can be biassed, and in the case presented by Freeman and Thompson overestimated network size strongly. Section 5 gives the results for Jacksonville and Mexico City, and recomputes estimates for the Orange County data of FT. Comparisons are made with other estimates, including those of Johnsen *et al.* (1989) based on data by Laumann *et al.* (1989) on homicide victims. In Section 6 we consider proxies for these studies which could considerably speed up (and hence cheapen) the data gathering procedure. Section 7 discusses the problem of names which are nearly identical, and its effects on the estimates. Section 8 briefly estimates the size of friend, rather than acquaintance, networks. The paper concludes with a critique, and recommendations for future studies.

2. The network suite

The data gathering instrument, known as the network suite, is described fully in Bernard *et al.* (1989b). It comprises four network elicitors, known as *modules*. These are, in order of presentation to the informant:

(1) An approximation to the General Social Survey question (with whom can you discuss important matters?) was asked of each informant. Note that there was no limit put on informants as to how many people they could name.

(2) The support network. Eleven questions were asked of each informant:

(a) Who would take care of your house if you went out of town?

(b) If you work outside your home, who would you talk to about work decisions?

(c) Who, if anyone, has helped with household tasks in the last three months?
(d) With whom have you engaged in social activities in the last three months (such as going to movies, had over for dinner, etc.)?
(e) Who do you talk to about hobbies?
(f) Who is your "best friend"?
(g) Who do you talk to about personal worries?
(h) Who do you get advice from when making important decisions?
(i) If you needed a large sum of money, who could you borrow it from?
(j) Who are the adult members of your household, excluding you?
(k) Who do you feel especially "close to"?

Again there was no restriction on the number of people which could be mentioned; nor was there any restriction on repeating people between the eleven questions.

(3) The RSW instrument. This presented 500 targets to the informant. Each target possessed a name, location and occupation. The list contains 400 names around the world which are maintained constant between cultures – thus the names here have been presented to Paiutes as well as to Jacksonville and Mexico City informants. The remaining 100 names are local (i.e. in the same country) for the informant, but their distribution of occupation status, sex, etc. do not differ between cultures. Each target elicits a person whom the informant believes is more likely to know the target person than the informant him/herself. Again, persons generated by this procedure will almost certainly repeat between targets. The informant also indicated whether the occupation or the location of the target was responsible for making the choice of intermediary person.

(4) A modified FT telephone book instrument. 305 last names were generated randomly from the telephone book by the method FT describe. (Each different name was given an equal chance of occurring in the list, so that Smith and Abramowicz would both be equally likely to occur.) Informants were presented with each name, and asked to list any and all of the people they knew with that last name. Note that FT asked informants if they had *ever* known anyone of that name, whereas we merely asked if informants knew someone by that name. The Jacksonville informants certainly rejected people they knew to be deceased, but often included people with whom they had lost contact. It is not known how the Mexico City informants interpreted this question.

This suite of methods was administered entirely by computer, which did not, however, lessen the data problems we encountered (cf. Bernard *et al.*, 1989b). In all cases, informants provided the sex and relationship (friend, relative or acquaintance) of each of the persons they generated during the suite. However, the sex and relationships of the U.S. informants' choices in module 4 (the FT phone book method) were lost, save for those which could be filled in by reference to earlier modules. Informants also provided the usual data about themselves.

3. The data sources

The U.S. informants numbered 98 residents of Jacksonville, Florida (the majority of whom resided in Orange Park, within the Jacksonville urban area). They were obtained by advertisements in the local paper, and personal contacts of one of the authors. Informants provided data at the house of one of the authors. The Mexican informants numbered 99, all of whom lived in Mexico City. They were obtained through the network contacts of three data collectors; data were taken at the houses of the informants. Informants were paid for their participation.

The telephone books for Jacksonville and Mexico City were obtained, and the FT selection method implemented. This method, which involves choosing lines randomly in the book and rejecting those names which are not the first of their list of entries, was not practical for the Mexico City book due to its size. A less random method (actually biassed towards less frequent names) was employed. As we shall show below, this makes no difference statistically if the correct estimate for network size is used.

It was also necessary, for the estimates of network size, to determine the total length of each telephone book, and the number of different names in the book. It is of course straightforward to count the number of pages, columns per page, and lines per column in each book. However, many of the lines do not contain names as such (admonitions from the phone company, indented advertising, large size text, company names, etc.). It was necessary to take repeated random page samples from each book, and count the number of inadmissible entries on each page. Averaged, this provided an estimate of the fraction of each page which contained acceptable names, and thus an estimate of the total effective length of the book. The same page samples were also

used to estimate how many different names occurred on each page (sometimes zero!). This gave an average figure which, too, could be scaled up by the number of pages.

4. The rationale for the estimates of network size

All the estimates involve the people triggered by the list of phone book names; some involve the network elicitors. We make several assumptions for the estimates to follow. It is far from clear that all the assumptions are correct; however, they are the best that can be made at our current state of knowledge.

(i) A phone book provides a list of *entries*. Each entry is a one-line statement giving access to (on average) f people, where f is unknown. The parameter f includes the effects of families, multiple phones, children's phones, those without phones, etc. The size of f is unknown, but clearly one would expect that $f > 1$. (This "scale-up" factor f will appear with various subscripts as appropriate.) The entry is tagged by a last *name*.

(ii) There are 1 or more entries for a given name. Thus some names are more common than others.

(iii) All the people known to an informant have names which are contained in the phone book.

(iv) The distribution of names in a phone book is a proxy for the distribution of the names of the people in any informant's network. In other words, the ratio of Smith's to Abramowicz's in an average informant's network will be the same as it is in the phone book: and there are no forces acting to make certain names more or less likely in an informant's network *other than their natural abundance in the universe around the informant*.

(v) The network elicitors produce the same kind of network for each informant.

To proceed, we need a little terminology. We suppose that a list of L names, here 305 (but see below), have been selected from the phone book, which contains M entries in total. Of the G different names in the book, the list comprises $n_1, n_2, \ldots, n_L$. We then count the elicited list of persons known to the informant by those names. Call the size of

this list F. We then present a collection of network elicitors (our modules 1, 2, and 3) to the informant, yielding a list of persons in that type of network. The total of this list is R_i for list i ($i = 1$, 2, or 3 here). The number of *overlaps* O_i between the list R_i and the list F is counted. We assume that an informant knows N people; N is to be found.

For discussion, we also define the number of entries for name g (not necessarily one of the L presented) to be P_g, so that the total number of entries M is given by

$$\sum_{g=1}^{G} P_g = M$$

and the number of people in the world (or town) is fM. For convenience, we define the number of entries e_k for name n_k to be

$$e_k = P_{n_k}$$

in what follows. Then the total of entries for the list L is defined as

$$E = \sum_{k=1}^{L} e_k .$$

4.1. The modified Freeman–Thompson estimate N_Q

The estimate, which modifies the thinking in FT, is given by

$$N_Q = \gamma \frac{FM}{E},$$

where

$$\gamma = \frac{f}{f_E}$$

will usually be taken to be unity. Here f_E is the average number of people represented by each of the E entries of the list L. The formula scales up the reported triggered names F by (the total entries in the phone book M divided by the entries taken up by the list L). The rationale is as follows.

The probability that any member of a network has name n_k ($k = 1, 2, \ldots$) is simply the fractional number of people in the world with name n_k, namely $f_k e_k / fM$, by assumption (iv). (Here f_k is the average number of people represented by each of the entries e_k. We shall cease to define the f factors as they appear henceforth: their meaning will be obvious.) We define $q_k = f_k e_k / fM$, for convenience. Then the probability that an informant with network size N knows exactly r people with name n_k is given by the binomial theorem as ${}_N C_r q_k^r (1 - q_k)^{N-r}$. Thus the expected number of people in the network with name n_k is simply the mean of the binomial distribution, N_{q_k}. This finally yields

the expected number of people named

$$= \sum_{k=1}^{L} (Nq_k) = N \sum_{k=1}^{L} q_k = \frac{N}{fM} \sum_{k=1}^{L} f_k e_k = \frac{N f_E E}{fM} = F$$

by supposition. Rearranging this gives the estimate for N as

$$N_Q = \frac{f}{f_E} \frac{FM}{E} = \gamma \frac{FM}{E}$$

as above. The parameter γ is formally unknown (and enforced by our study of *names*, rather than *people*.) However, providing the names are reasonably randomly chosen (so that E and M are similarly representative of the "universe"), assumption (iv) implies that

$$f \approx f_E,$$

so that

$$\gamma \approx 1.$$

We shall test the hypotheses both of randomness of the names, and of unit γ, below.

4.2. The original Freeman–Thompson estimate N_{FT}

The original estimate, made by FT, is subtly different. This is given by

$$N_{FT} = \frac{FG}{L}$$

and scales up the reported triggered names F by (the number of different names in the phone book G divided by the number of names in the list L).This estimate was created by FT, and was used by them to predict network sizes of 5520 for the Orange County, California area. However, the estimate relies on the sample of names L being truly random and representative. To see this, take the estimate N_Q already given. If we average over many samples of L names, the expression

$$\sum_{k=1}^{L} e_k$$

has an expected value of

$$\sum_{k=1}^{L}\left\{\sum_{g=1}^{G} f_g P_g.\text{prob}(g=n_k)\right\} = \sum_{k=1}^{L}\sum_{g=1}^{G} f_g P_g \frac{1}{G} = \sum_{k=1}^{L}\frac{1}{G}\sum_{g=1}^{G} f_g P_g$$

$$= \frac{1}{G}\sum_{k=1}^{L} Mf = \frac{L}{G}Mf$$

by the assumption of random names made by FT. Hence substitution into the N_Q formula above gives their estimate (the f's cancel here)

$$N_{\text{FT}} = \frac{FG}{L}$$

as required. Now assuming for the moment that $\gamma = 1$, the second estimate N_{FT} can only equal the estimate N_Q if

$$\frac{L}{G} = \frac{\text{number of names presented}}{\text{total number of different names}} = \frac{E}{M}$$

$$= \frac{\text{total entries for the list of names}}{\text{total entries in phone book}}$$

or, equivalently, that the fractional space taken up by the names in the phone book is precisely the fraction given by the number of names in the sample divided by the total number of names. (Immediately we see that a long list of names is necessary for the original FT estimate to

function correctly, since the ratio L/G can vary strongly between short but random lists.) In other words, N_{FT} *is likely to be a valid estimate for network size only if the sample of names chosen is sufficiently large that the random method used to produce it has indeed yielded a sample which occupies the same fraction in the space of different names as the entries do in the phone book*. We shall see that this is not the case for the Orange County data presented by FT.

4.3. The overlap estimate N_i for network elicitor i

This estimate is given by

$$N_i = \frac{FR_i}{O_i}.$$

It scales up the reported triggered names F by (the number of people in network i divided by the number of overlaps between network i and the F names). To obtain this, let us apply the modified FT estimate, N_Q and call the estimate N_i. Consider one of the N_i people in the network, chosen at random. The chance that it is one of the R_i people in the ith elicited network is clearly R_i/N_i. Now the O_i persons generated by the average informant are a random selection of the N_i, by assumption (iii) and the R_i people, however produced, have the same role between informants, by assumption (v). Hence

the expected number of overlaps between the F people

and the R_i people is

$$O_i = \frac{R_i}{N_i}F = \frac{R_i}{\gamma FM}EF = \frac{R_iE}{\gamma M},$$

which is, remarkably, independent of F. Now this equation can be solved for M, giving

$$M = \frac{R_iE}{\gamma O_i},$$

which in turn gives the expression N_i as

$$N_i = \frac{\gamma FM}{E} = \frac{FR_i}{O_i}$$

as above. Note that we have also obtained, *en passant*, an estimate of the unknown factor γ, namely

$$\gamma_i = \frac{R_i E}{O_i M},$$

and this can be used to test the claim of unit γ.

5. Estimates of network size for Jacksonville, Mexico City, and Orange County

We now present the estimates for the two data sets we obtained, together with a reevaluation of the FT estimate for Orange County. The initials J, MC, and OC will often be used to save space. The data and results are shown in Tables 1 and 2.

5.1. The Jacksonville results

The necessary figures are provided in Table 1. Their construction is as follows. The Jacksonville phone book possessed 445 pages, each of 4 columns of 106 lines, giving an initial estimate of total entries as (445)(4)(106) = 188,680. A 35 page random sample found 28 unaccep-

Table 1
The data

Locale	J	M	OC
L	305	271	297
F	11.5 ± 20%	4.2 ± 36%	15 ± 10%
M	176220	700280	450288
E	1457	6850	3335
G	34710	27775	87792
R_1	6.9 ± 14%	2.9 ± 18%	n.a.
O_1	0.041 ± 120%	0.061 ± 168%	n.a.
R_2	21.8 ± 15%	10.1 ± 13%	n.a.
O_2	0.184 ± 61%	0.081 ± 127%	n.a.
R_3	129 ± 11%	76.5 ± 21%	n.a.
O_3	0.70 ± 27%	0.28 ± 87%	n.a.

The percentages indicate ± 2 standard errors of the mean (i.e. 95% error bars). Note that some overlap figures are not significantly different from zero.

The estimates of N from overlaps are statistically dubious, because of the small numbers reported. These are 200, 524, and 1135, with corresponding γ's of 0.47, 1.22, and 2.64. Again, no γ estimate differs significantly from unity, although their average is about 1.4, with appallingly high error bars.

The best estimate for the number of people currently known to Mexico City informants is then 570 ± 460. This is close to the estimate of 664 by Bernard *et al.* (1989a) using informants' knowledge of earthquake victims. It is, however, about twice as large as the estimates of Bernard *et al.* (1987), based on informants' knowledge of doctors, mailmen, bus drivers and (later) knowledge of earthquake victims.

5.3. A reevaluation of the Orange County results of Freeman and Thompson (1989)

The difference between the Jacksonville result and the Orange County result (apart from the fact that FT asked about people ever known, while we asked about people actually known) led us to reexamine the FT data, and caused the production of the new N_Q estimate.

FT reported that the OC phone book possessed 1416 pages, with 4 columns of 92 lines, and concluded that the book contained $(1416)(4)(92) = 512{,}088$ entries. It is clear that their estimate had already removed some unacceptable lines which occurred on each page. A reexamination of the book found 1415 pages of 4 columns with 122 lines per column, giving an initial estimate for M of $(1416)(4)(122) = 691{,}008$. A random 35 page sample found 170 unacceptable entries per page, or a rate of 35%. This lowered the value of M to 450,288 cited in Table 1. This estimate is comfortably similar to the FT estimate of 512,088. However, our estimates of total names (although not needed for the N_Q estimate) differed rather more. FT counted names in 100 columns, giving 112,147 distinct names. Our count of 35 random pages yielded 62 (s.d. 39) different names per page, or $G = 87{,}792$. The difference between our estimate and FTs estimate is statistically significant.

The list of 305 names used by FT turned out to have 8 names which did not exist in the OC phone book, probably due to typing errors at some stage. This reduced L to 297. The total entries of the 297 names were 3335.

Immediately we see that the names chosen were not very random. The fraction G/L was 296, while M/E was 135, a factor of 2.2 lower. Put another way, FT's list of names occupied 2.2 times the amount of space in the phone book than it would have, had it been truly random (which, given the care with which they prepared the list of names, demonstrates the difficulty of generating random samples).

The effect of all this is to generate an estimated N_Q of only 2025, compared with FT's original estimate of 5520. Note that this drastic reduction is less evident if one used the biassed estimate N_{FT}, which has the revised value 4434, which is only 20% smaller than they estimated. However, as we have seen, this estimate is not in general correct.

Since the OC study was performed five years ago, we can also roughly estimate temporal variability in the estimates (though not in actual network size, which would involve replicating the study). We used the 1988/9 OC phone book, and applied the same methods. We found: $M = 425{,}228$ (a reduction since 1984); $G = 97{,}160$ (an increase since 1984); L was reduced to 240, since there were 65 names in the FT list which did not now occur, and their entries totalled $E = 3531$. (The sample remained biassed: $G/L = 404$, while $M/E = 120$.) The estimated N_Q was now 1806 ($\pm 10\%$), compared with 2025 five years earlier. These estimates are just significantly different.

5.4. Comparison

Apart from the drastic reduction in the FT estimate, one thing is clear: Mexico City informants have a *much smaller* acquaintance network, from our data, than do U.S. informants. Orange County informants are predicted to have significantly larger networks than Jacksonville informants, which is to be expected given that the network of people an informant has ever known must contain the network of people an informant currently knows. However, the spread of the Jacksonville estimates including the overlap calculations is such as to include the single Orange County value, so that we cannot necessarily conclude that Jacksonville and Orange County informants differ in their network sizes despite the different question asked.

It is difficult to see why Mexico City informants should have networks so much smaller than their U.S. counterparts. Since personal network size typically increases with age, followed by a decrease in

Table 2
The results

Locale	J	M	OC
N_Q	1391 ± 20%	429 ± 36%	2025 ± 10%
N_1	1935 ± 154%	200 ± 222%	
N_2	1362 ± 96%	524 ± 176%	
N_3	2119 ± 58%	1135 ± 144%	
γ_1	1.39 ± 134%	0.47 ± 186%	
γ_2	0.98 ± 76%	1.22 ± 140%	
γ_3	1.52 ± 38%	2.64 ± 108%	

The percentages indicate a simple sum of all the 2 standard error percentages entering the relevant calculations. When the denominator is significantly nonzero, the error is slightly overestimated by this method.

table lines per page (businesses are listed separately, hence the low figure). This reduces M by 6.6% to the value given. The same sample yielded 78 (s.d. 46) different names per page, so that the original N_{FT} estimate could be computed. This gave the G estimate listed, with quite high error bars. The 305 names occupied 1457 entries (E).

The modified FT estimate, N_Q, was 1391 (±20%), assuming that $\gamma = 1$. This is very close to the N_{FT} estimate, 1308 (±20%), because it turned out that the Jacksonville name sample was fairly random, on two grounds. First, the ratios G/L and M/E, used as scaling factors for F, are very similar: $G/L = 114 \pm 33\%$; $M/E = 121$. Second, the correlation between the number of entries for each of the 305 names in the phone book, and the number of people triggered by each name was 0.82, so that the name distribution in the book and that in informants' networks are equivalent.

The estimates involving overlaps with other networks are: N1 = 1935; $N_2 = 1362$; and $N_3 = 2119$, with corresponding estimates for γ of: $\gamma_1 = 1.39$, $\gamma_2 = 0.98$, and $\gamma_3 = 1.52$. It is clear from the low rate of overlaps that estimates other than N_3 are statistically dubious; the numbers of both informants and names from the phone book need increasing to remove this difficulty. Can we estimate γ from these figures? We have argued that random name sampling, or use of assumption (iv), would imply that $\gamma = 1$. However, our list of names was but a single sample of such lists, and might have been biassed.

The two-standard-deviation spread of the estimates shows that on no occasion can the null hypothesis that $\gamma = 1$ be rejected. However, the

average of the three estimates would give a best guess of $\gamma = 1.3$. Using such a value would raise the N_{FT} estimate to 1804 (which is automatically the mean of the N_i estimates). It is possible that on average each of our 305 names accounted for 30% less people than did the average entry in the Jacksonville phone book, but not probable. (There is obviously *some* effect of last name on family size: families with (e.g.) Irish last names may well be larger than some others, so that the f factors must vary from name to name. Our data did not permit us to examine any groupings of network members into families.) The problem is frustrating, since we are trying to count people, rather than names, which are merely artefacts.

With caveats for the low reliability of some of the N estimates, this gives our best estimate for the number of people currently known to Jacksonville informants as 1700 ± 400. The error estimate here merely represents the spread of the estimates, and does not include the individual errors present in each estimate.

5.2. *The Mexico City results*

The Mexico City phone book contained 2135 pages, each of 4 columns with 152 lines, giving an initial estimate of M as 1,298,080. A random 35 column sample found 70 unacceptable lines per column, or 280 per page. This becomes 46% unacceptable entries, lowering M to 700,280. A 50 page random survey found only 13 (s.d. 19) different names per page, giving G as 27,775 ($\pm 41\%$). That Mexico City should have fewer distinct names than Jacksonville, despite its much greater size, was quite surprising to us. We prepared a random list of 305 names from the book. More careful examination, after the experiment, revealed that 34 of these were unacceptable, since they were names of companies, and had no corresponding personal name in the book. These names occupied 6850 entries (E). It turned out that Mexican informants generated far fewer people from these 271 names, only 4.2 (s.d. 7.6).

The modified FT estimate, N_Q, was 429 ($\pm 36\%$). Coincidentally, the original N_{FT} estimate was 430, since (a) $G/L = 102.5$ ($\pm 41\%$), and $M/E = 102$ are the same, and (b) the correlation between number of names reported and length of name entry, although lower (0.64), was still highly significant, showing that the choice of names was reasonable.

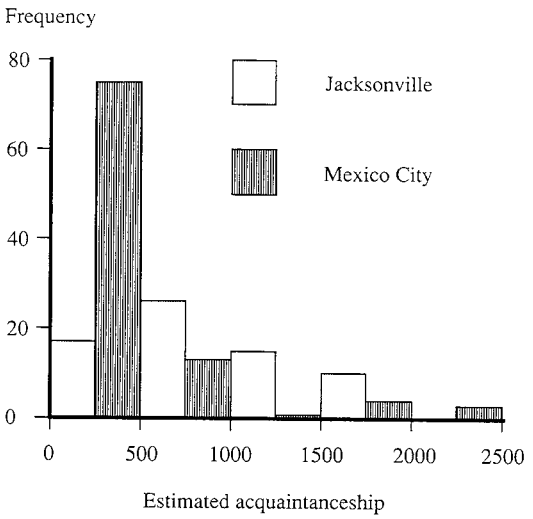

Fig. 1a. Histograms of the predicted distribution of current network size, using the N_Q method for each informant. Totals are: Jacksonville 98 informants; Mexico City 99. Each pair of entries fits into the interval shown.

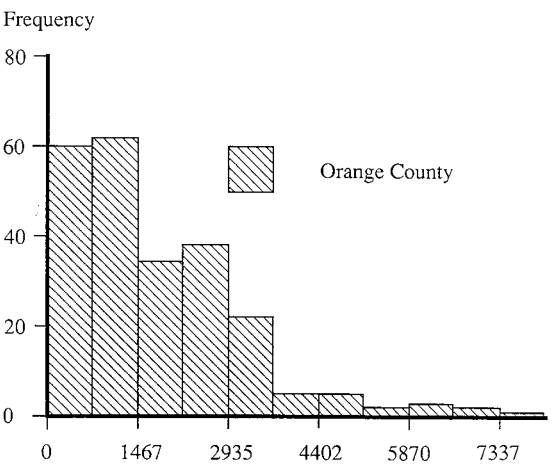

Fig. 1b. The redrawn histogram of the predicted (ever known) network size for the 247 Orange County informants.

later life, one might speculate that MC would have relatively more young people owing to its population explosion. However, the two sets of informants were not statistically different in age (J = 37, MC = 33); indeed, network sizes were uncorrelated with informants' ages.

The distribution of network sizes for the three data sets is given in Figure 1a and 1b in histogram form (the data for Orange County taken from FT's figure 1, rescaled). Notice that although the OC average is only 50% larger than the J average, the distributions are quite different, with proportionally many more OC informants having networks much larger than the mean. This may be caused by differences between the informant's total life network and his or her current network; it may be an artifact of the larger number of informants (i.e. the likelihood of finding informants with large networks increases with the number of informants); and the nonrandom sample used in OC (all students) may also play a part. The two networks, of course, may just be different.

We may note that the estimates of 1700 ± 400 for Jacksonville, and 2025 for Orange County, are of the same order as the independent estimate of Johnsen *et al.* (1989). They used national data on the likelihood of informants knowing one or more homicide victims to estimate the mean acquaintance size, and found a value of 1526. This too lies in the range for the Jacksonville informants, but is significantly smaller than the Orange County estimate. The estimates we make here also agree quite well with Gurevich's (1961) figure of 2130; the network size distribution is qualitatively similar in this case.

Although we cannot make improved estimates for γ, we can examine estimates for f. A useful experimental analogue is that for those J informants who generated people in module 4, the mean number of people generated per last name was $1.38 \pm 7\%$. The equivalent calculation for MC of the number of people generated per last name gave an "f" of $1.6 \pm 17\%$, very similar to the J value. However, a different estimate can be obtained using the fact that the 1988 population of Jacksonville was 677,077, implying that the "true" value for f, the number of people accounted for by each entry in the book, would be $(677{,}077)/(176{,}220) = 3.8$, which is higher than either of the estimates.

6. Other proxies for network size

The instrument used to gather our data was costly both financially and in terms of informants' time. It is natural to wonder whether reliable estimates of network size could be obtained with less work by informants. We examined two possibilities within the Jacksonville data.

(a) do the names have to be randomly chosen?

The rationale for the N_Q estimate does not require that names be randomly selected, although it would be natural to do so. As an extreme test of the necessity for randomness, the equivalent figures were computed assuming that only a list of the seven most frequent names (Cochran, Little, Merritt, O'Brien, O'Connor, Ross and Sweat) had been presented to informants. The correlation between length of name entry and number of people triggered was 0.81 (significant at the 5% level), so that the sample still paralleled informants' networks. The length of entries E was 598, and 3.15 names from this set were generated on average. This led to an estimate for N_Q of $(3.15)(176{,}220)/(598) = 929 \pm 9\%$. This estimate is significantly lower than the original Jacksonville N_Q, but remains of the same order of magnitude. Thus many fewer names presented to informants can still yield network size estimates of about the same size.

(b) does the size of the telephone book matter?

Clearly no single phone book from one U.S. city can contain all the names of residents of the U.S., because of the locations of various ethnic groups. However, a smaller phone book is easier to manipulate. As an experiment, we used the Orange Park phone book to test a smaller book. Orange Park (the similarity of names with Orange County is unfortunate) is a suburb of Jacksonville, in which many of our informants lived. It has a separate phone book, whose entries are of course also contained in the Jacksonville book.

The correlation between length of name entry in Orange Park's directory and the main book was 0.91, which indicated that it could well serve as a proxy for the main book. (Of course, many names in the main book, including the majority of the rare names in our list of 305, were not present in the OP directory.) The relevant figures were

$M = 16{,}380$; $E = 135$. These gave, remarkably, M/E (Orange Park) = 121, precisely the same figure as for the main Jacksonville calculation; so the prediction of N_Q is the same as for Jacksonville. Thus we must conclude, for these data, that a small phone book does indeed serve as a proxy for a large one. Whether this is true in other samples is unknown.

7. The problem of "similar" names

It became obvious during the analysis that informants had not always responded precisely to the task given. A total of 47 responses to the name Higgenbotham were generated in Jacksonville by 98 informants. Given that only one entry in the phone book listed Higgenbotham, but that there were 116 entries for Higginbotham, informants were giving names which were similar in some sense to the names of people they knew. There were many other examples: our choice of Richerson produced many responses, with only two entries, while Richardson (the normal spelling) had 252 entries.

This is a potentially severe source of error in the method. If informants are really responding to a much larger list of names, then our estimate for N_Q (but not the overlap estimates) will be an overestimate, and will have to be scaled down somehow. We examined the list of names and responses, and selected nine names for which the problem appeared serious. (Our criterion was that in total, the 98 informants produced more people than there were entries in the phone book, and that the "nearby" entries were sufficiently sizeable to affect the calculation of N_Q.)

This yielded nine names: Burrow versus the more popular Burrows, Cheney vs. Chaney, Cornwall vs. Cornwell, Dees vs. Deese, Dixson vs. Dixon, Higgenbotham vs. Higginbotham, Kline vs. Klein, Marshell vs. Marshall, and Richerson vs. Richardson. This set lengthened E from 1457 to 2292, a 57 percent increase. This correspondingly reduces the estimate N_Q to 884, compared with the original estimate of 1391. Thus the original estimate (subject to the unknown γ factor) must be seen as an upper bound on the actual N_Q value in all cases. It may well, as in this case, be a severe overestimate.

8. Split by sex and relationship

Jacksonville informants generated significantly more male people than females (by percentage, 65% to 35%, s.d. 23%; by actual count, 7.11 to 4.45, or 62% males). These percentages are significantly higher than those reported by FT (56% males). The equivalent figures for Mexico City informants are: by percentage, 59% males and 41% females (s.d. 32%), and by actual count 1.88 males to 1.73 females (52% males). For both sets of informants, the percentage of male choices generated by the FT technique was larger than for the RSW module, which itself tends to generate more male choices than female.

A full discussion of the role and number of friends, relatives and acquaintances is given in Bernard *et al.* (1989b). However, we simply note here estimates of the equivalent calculations for friends and relatives; acquaintances then, of course, being obtained by subtraction. In Jacksonville, we had very limited knowledge of the proportion of friends in the FT module. The percentage of friends generated tends to decrease as the module number increases, so that module 1 has the largest proportion of friends and module 4 the smallest. As our only guide to the number of friends in the Jacksonville module 4, then, we were forced to use the fraction of friends mentioned in the RSW module, which was 37%. (This is probably an overestimate: Mexico City informants reported 45% friends in module 3 and only 30% in module 4.) Assuming for the moment that the module 3 figure holds also for the module, this yields estimates for friends of precisely 37% of those already given, namely: (515, 716, 504, and 784), with an average of 629 friends per informant. In Orange County, FT report that 22.2% of reported persons were friends. This gave a figure of 450 friends per informant; had a lower estimate of the percentage of friends been used, the agreement would have been better.

In Mexico City, the figure is much lower. Informants reported 30% of friends, giving only 129 friends in an average network. (It should be noted that the much smaller values for the Mexican data for all quantities connected with network size holds up consistently; cf. Bernard *et al.* 1989b.)

Figures for relative usage are: for Jacksonville, module 3 estimates suggest 17%, and for Mexico City only 5% of persons recalled were relatives.

9. Discussion

The introduction noted that we have little idea of the mean size of an informant's network, let alone how it varies between informants and across cultures. However, for the U.S. at least, evidence presented here is beginning to give consistent estimates of network size from a variety of methods and sources of data. Estimates using phone books, and estimates using overlaps between phone book studies and network generators, are all yielding numbers around 1700, although with a spread of at least 400. A rather smaller, and less accurate, number of estimates for Mexico City informants is beginning to give estimates around 570.

We obviously need far more studies of network size, using as many methods as possible (ideally on the same informants to permit a direct intercomparison). However, such studies can only yield at best one number – a network size. Such a number would be a useful constraint on any theory of social networks. Lacking any such theory, we must seek more information while obtaining these estimates. If the phone book method, or something similar, is to be used to predict the scaling-up of more restricted networks such as support and RSW networks, then we need to know how many people the phone book really represents (the f factor in our calculations). How many names are really represented by one entry, given our findings about similar names? Could we acquire the equivalent of a world phone book? *Why* do informants know the people they do? What accounts for the differing proportions of friends, relatives and acquaintances in their networks? Why do informants in areas of low population density (e.g. Paiutes, Ponape) appear to have far more people in their networks than do informants who live in medium and high population density areas? Would another descriptor, instead of population density, prove both more enlightening and more relevant? Can the degree of overlap within friendship networks, discussed at length by Pool and Kochen (1978), be easily measured?

To answer these questions is difficult – both for investigators and for informants. It involves applying several different methods to the same set of informants. This takes time; time that many informants may not wish to give, even with payment for their services. Would the selection of those informants who will give data bias the results of such surveys? To get statistical reliability, many more informants are needed

than the few hundreds used so far. Again, time and expense are difficult issues.

An alternative strategy might be to use simpler network elicitors as proxies for the data one really wishes to collect. But as this paper shows, the correct way to scale up these proxies to the full network is not simple; and research is needed both on proxies and full networks to find out the correct scaling method. The questions remain both difficult and important, if we are ever to understand the glue that binds human beings together on this planet.

References

Bernard, H.R., E.C. Johnsen, P.D. Killworth, and S. Robinson
1987 "Estimating the size of an average personal network and of an event subpopulation: some empirical results." *American Statistical Association*, Proceedings of the section on Survey Research Methods.

Bernard, H.R., P.D. Killworth, M. Evans, C. McCarty, and G.A. Shelley
1988 "Studying social relations cross-culturally." *Ethnology 27*, 155–179.

Bernard, H.R., E.C. Johnsen, P.D. Killworth, and S. Robinson
1989 "Estimating the size of an average personal network and of an event subpopulation," in: M. Kochen (ed.) *The Small World*, pp. 159–175. Norwood, NJ: Ablex.

Bernard, H.R., E.C. Johnsen, P.D. Killworth, C. McCarty, S. Robinson, and G.A. Shelley
1989 "Comparing four different methods for measuring personal social networks." *Social Networks*, in press.

Burt, R.S.
1982 *Toward a Structural Theory of Action*. New York: Academic Press.

Fischer, C.
1982 *To dwell among Friends: Personal Networks in Town and City*. Chicago: University of Chicago Press.

Freeman, L.C. and C.R. Thompson
1989 "Estimating acquaintanceship volume", in: M. Kochen (ed.) *The Small World*, pp. 147–158. Norwood, NJ: Ablex.

Gurevich, M.
1961 *The social structure of acquaintanceship networks*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Johnsen, E.C., H.R. Bernard, P.D. Killworth, and G.A. Shelley
1989 Estimating the size of event populations: the number of AIDS and homicide victims in the U.S. To be submitted.

Killworth, P.D. and H.R. Bernard
1978 "The reverse small-world experiment." *Social Networks 1*: 159–192.

Killworth, P.D., H.R. Bernard, and C. McCarty
1984 "Measuring patterns of acquaintanceship." *Current Anthropology 23*: 381–397.

Laumann, E.O., J.H. Gagnon, S. Michaels, R.T. Michael, and J.S. Coleman
1989 "Monitoring the AIDS epidemic in the United States: a network approach." *Science 244*: 1186–1189.

Milgram, S.
1967 "The small world problem." *Psychology Today 1*: 60–67.
Pool, I. deS. and M. Kochen
1978 "Contacts and influence." *Social Networks 1*: 5–51.
Wellman, B.
1979 "The community question." *American Journal of Sociology 84*: 1201–1231.