

In: *The Small World*, ed. by M. Kochen
159-75. NJ: Ablex Press

CHAPTER 9

Estimating the Size of an Average Personal Network and of an Event Subpopulation*

H. Russell Bernard

Department of Anthropology
University of Florida
Gainesville, FL

Eugene C. Johnsen

Department of Mathematics
University of California
Santa Barbara, CA

Peter D. Killworth

Hooke Institute for Atmospheric Research
Department of Atmospheric Physics
Clarendon Laboratory
Parks Road, Oxford, England

Scott Robinson

Department of Anthropology
Universidad Autónoma
Metropolitana-Iztapalapa
Mexico City, D.F., Mexico

INTRODUCTION

Soon after the devastating earthquake of September 1985, one of us happened to be in Mexico City. Everyone in the city seemed to know someone who had died. Was this true? It seemed easy enough to find out. Why not take a sample of the city and ask them?¹

¹We are not ignoring the tragic dimensions of what was one of the most disastrous earthquakes of this century. The impact of the earthquake and the relief effort will be discussed in another paper.

*Research reported here was supported by a grant from the National Science Foundation BNS-8318132 to Bernard and Killworth for the Cross Cultural Study of Social Networks. Additional support was provided by the Division of Sponsored Research of the University of Florida. Data were collected in Mexico by Yolanda Hernández Franco, María del Carmen Costa, Patricio Meade, Miguel Angel Riva-Palacio, and Alejandro Casteneira, students of Anthropology at the Universidad Autónoma Metropolitana—Iztapalapa.

We laid out a map of Mexico City, divided it into a grid of 2500 squares, and selected 20 of the grid squares at random. Interviewers went to each of the selected squares and asked 20 people in each area a few questions. In particular, did they know anyone, personally, who had died in the quake? Where did that person live? What was his or her occupation and level of education? How did the respondent know the deceased?

For our immediate purposes here, the important quantity from the data is this: 91 out of 400 informants told us that they had, indeed, personally known someone who had died in the quake. If our sample were representative, then about 23% of the residents of Mexico City lost someone in their personal network.

There is another important question that we wanted to answer. How many people had actually perished in the earthquake? The official government figure was 7000. Opposition newspapers reported 12,000, 15,000, and even 22,000. How could we determine the true figure? We could ask people how many people they thought had died (that question was asked for another purpose), but that really wouldn't help very much.

We reasoned that if we knew how many people are known, on average, by a person in Mexico City, that is, if we knew the average personal network size c in that metropolis, then we ought to be able to calculate, to within a certain probable error, how many people c had actually died, given a sample large enough to represent the average person.

Since we are just now attempting to estimate c , we are not yet in a position to estimate e , given the data collected. However, this inquiry led us to consider two further questions.

1. What is the *distribution*, across a population, of the number of people individuals know?
2. How can the size of *any* event subpopulation be estimated?

These two questions are related.

Consider the first question, which has been a vexing one for many years in social network analysis. When de Sola Pool and Kochen (1978) wrote their landmark paper on the small world problem, they needed the distribution of network sizes, but this wasn't known. Recall this problem. If we take any two people at random from a population of known size, what is the probability that they know each other? If they don't know each other, then what is the probability that they know someone in common? Or that people whom they know, know each other? And so on. The answer, of course, is that it depends on how many people each person knows and on the distribution of that number.

At the time that de Sola Pool and Kochen were writing their paper, the only value related to average personal network size was obtained by Gurevich (1961), who asked people to carry around diaries for 100 days and to write down all the people they encountered. This resulted in a value of around 500. The problem of

estimating how many people had died in the Mexico City quake suggests a method for estimating how many people each person knows, on average. Furthermore, if we know the distribution of personal network sizes we can estimate how many people had actually died.

ESTIMATING AVERAGE PERSONAL NETWORK SIZE AND EVENT SUBPOPULATION SIZE

Our data yield an estimate for the probability p that an individual in the population knows someone in the event subpopulation of $p \approx 91/400 = 0.2275$. One might proceed heuristically as follows (derivations of the other mathematical results which follow appear in the Appendix).

Given a population of size t persons, and an event subpopulation such as those who died in the earthquake of size e persons, there is a chance e/t that any individual is in the event subpopulation. If each individual knows c people (for simplicity c is assumed here to be a constant) then

$$(2.1) \quad c \cdot (e/t) \approx p$$

will yield a rough estimate of c , namely

$$(2.2) \quad c \approx (t/e) \cdot p.$$

If we take as representative values $t = 18,000,000$ (the current population of Mexico City) and $e = 7000$, (2.2) gives

$$(2.3) \quad c \approx 585.$$

This answer is flawed, unfortunately, for at least two reasons. First, an assumption is made that p is fairly small, for this to be valid. When this is not the case (as here), (2.2) should be replaced by

$$(2.4) \quad c \approx \ln(1 - p)/\ln(1 - e/t),$$

where the maximum personal network size is assumed to be very small relative to t and \ln denotes the natural logarithm. This yields

$$(2.5) \quad c \approx 664.$$

The second, and more fundamental, problem is that personal network size is not constant. Our interpretation of the above value must therefore only be in the

nature of an approximation to the average personal network size c . After some algebra, it can be shown that

1. the value obtained from (2.4) lies within the range of network sizes of the population;
2. this value is a *lower bound* for the average personal network size c ;
3. a nontrivial upper bound for the average personal network size cannot in general be obtained; and
4. if more information about the distribution of personal network size is known, the estimates on the average personal network size c may be sharpened. For example, the value in (2.4) becomes the average personal network size when network size is distributed binomially and e is also very small compared to t .

The value obtained in (2.4) depends on the relative sizes of t and e . For the Mexico City data, with $t = 18,000,000$, this lower bound estimate for c is shown in Table 1 for the various proposed values of e .

Turning the problem around, we now suppose for a population T of known size t that we have an accurate estimate for the average personal network size c . Such an estimate may have been obtained as a common estimate from a collection of previously investigated event subpopulations E_1, E_2, \dots, E_s with accurately known sizes e_1, e_2, \dots, e_s and accurately obtained corresponding probabilities p_1, p_2, \dots, p_s of a person in T (more precisely $T - E_1, T - E_2, \dots, T - E_s$) knowing someone in E_1, E_2, \dots, E_s , respectively.

If E_x is a new event subpopulation of unknown size e_x and unknown relative size $\epsilon_x = e_x/t$ for which an accurate value of the probability p_x , of a person in T (more precisely $T - E_x$) knowing someone in E_x , is obtained, and if

$$(2.6) \quad p_{k-1} < p_x < p_k \text{ for some } k, 1 \leq k \leq s,$$

then a bounded estimate for ϵ_x is given by

$$(2.7) \quad \epsilon_{k-1} < \epsilon_x < \epsilon_k.$$

Table 1. Values of the Lower Bound Estimate of Average Personal Network Size c for the Mexico City Earthquake Data with Varying Death Rates e

e :	7000	12000	15000	22000
c :	664	387	310	211

Further, for the estimate

$$(2.8) \quad \bar{e}_x = 1 - (1 - p_x)^{1/c}$$

we have

$$(2.9) \quad \bar{e}_x \leq e_x,$$

where the closer c is to the value in (2.4) or the closer e_x is to 0 the better the estimate \bar{e}_x is to e_x .

As an example, for the Mexico City earthquake event subpopulation E of (assumed accurately known) size $e = 7000$ and probability $p = 0.2275$ we obtain, by Table 1, $c \approx 664$. Now suppose for a new event subpopulation E_x we have $p_x = 0.1986$. Then the size e_x is bounded by $e_x < 7000$ and underestimated by $\bar{e}_x = \bar{e}_x \cdot t = [1 - (0.8014)^{1/664}] \cdot (18,000,000) = 6000.67$, so $6000 < e_x < 7000$.

DISCUSSION

There are 18 million people in Mexico City. If 7000 had actually died, and 91/400 of the 18 million knew at least one of them, then, under the probability model and a binomial distribution of personal network size, the average personal network size is $c \approx 664$. As long as e and the maximum personal network size are very small compared to t , this lower bound for c can be found rather precisely. This suggests an interesting situation. Presumably, if, say, 100,000 people had died in the earthquake, then the chance that a *good friend* died in the event would be far greater than it would be if only 7000 had died. Good friends are a subset of everyone's network, as are people with whom one discusses important matters (the name generator for the network component of the 1985 General Social Survey), and as are people from whom one would borrow money.

This assumes a lot, however. First, the sample has to be representative and large enough to give a reliable estimate of p in each case. Given the budget for this study (\$400), our sample of 400 leaves much to be desired. Just four more people (one more percentage point) saying that they knew someone who had died changes our lower bound estimate of c in (2.4) by over 30. Thus, the random sample needs to be much larger (at most about 2400) to deliver an estimate within two percentage points 95% of the time of the percentage of $t - e$ who know someone in the event subpopulation. While we do not consider this intimidating, we make no strong claims about the actual situation in Mexico City, except to say that our estimation is a tantalizing first step.

Second, people have to tell the truth when asked whether they know someone in an event subpopulation. We are fairly confident that our informants were telling the truth for several reasons. We asked them whether they *knew* someone, not about their past behavior. The operational definition of whether people know someone is whether they say they do. Also, we asked people to choose between knowing versus not knowing. They did not have to scan a large collection of items and dredge up a list, as when informants are asked to name what they ate yesterday, or with whom they talked. When asked to tell us the occupation and address of the deceased, in nine cases (10%) respondents said they did not know; the deceased was either a casual acquaintance or was in the personal network of someone from whom the respondent had heard about the deceased. In 22 cases (18%) respondents said that they did not know where the deceased lived when the deceased was only someone they had heard about from someone in their own network.

Third, we assume that people *know* whether they know someone in the event subpopulation—in this case, that they have lost someone from their network as a result of the earthquake. This may not be the case. We often find out about such losses a long time after they occur, if the person involved is someone on the periphery of our network. This problem leads to the interesting possibility of defining first and second order network zones operationally on the basis of a cutoff time for finding out about what happens to people in those zones.

We have considered various network distributions which allow for variation in personal network size in the total population. We have shown that the value in (2.4) must occur within the range of values of network size and that this value is a lower bound for the average personal network size. Thus, $\ln(1 - p)/\ln(1 - e/t)$ is an important value associated with a population T and a given event subpopulation E , when the maximum personal network size is very small relative to t . Furthermore, in our investigation of the two-point distribution we found that there is no natural upper bound for the average personal network size other than one imposed by the maximum personal network size. As an example, there are informants from a previous study (Killworth & Bernard, 1978) who knew at least 1100 people.

There is an assumption that seems as if it might cause a problem, but does not. That is that the "world" being tested needs to be closed. While there are 18 million people in Mexico City, there are 80 million people in Mexico as a whole. More than half of the population of Mexico City has come to the capital in the last 20 years from elsewhere in the country, and most of them have extensive personal networks outside the capital. This is not a problem, however, because we are not trying to estimate average personal network size with respect to the entire country (or to the entire world), but only with respect to the population of Mexico City.

While some of the assumptions noted above are strong, and while we have

violated some of them in the study reported here, we have reason to be optimistic. The data from this one study are quite rich. In our sample there is strong evidence that people who *think* they know more people are indeed more *likely* to know more people. Respondents who claimed to know 800 or more, for example, were twice as likely to know someone (or know of someone) who had died in the quake, as were respondents who claimed to have networks of 100 or fewer persons. Women were more likely to know someone who died. Respondents who were born in Mexico City were likely to claim that they know many more people than do respondents who were born outside the capital. Respondents who are more aware of current events (we asked people if they had heard of the new IBM plant going in, a major topic of political conversation at the time) were 2-to-1 more likely to have known someone who died in the quake.

It appears that what is required is a series of studies that will produce a collection of lower bounds for the average sizes of general personal networks and for the average sizes of particular subnetworks (based on a variety of criteria), and correlates of personal network size (socioeconomic-demographic factors). Suppose we ask 20 independent samples of Americans, through the NORC, the ISR, or the Gallup poll, for example, whether they knew personally anyone who had been killed in the Vietnam War, or who had been killed by a drunk driver during the past 12 months, or who was a member of some other subpopulation of *known size*. By doing this for many representative samples, we should close in on the distribution of personal network sizes for a variety of classes of Americans: old people, young people, men, women, Whites, Blacks, Republicans, Democrats, etc. This could be extended to other societies as well, so that we could look for the global average network, and the factors that account for variation around that average. We believe that industrial, urban, college educated people have larger personal networks than uneducated farmers at the peasant level in developing nations. We could test that presumption. Indeed, recent data suggest the opposite (Bernard, Killworth, Evans, McCarty & Shelley, 1988).

Once the basic quantities were known for the U.S., for example, we could begin to solve our model equations and inequalities for unknown values of e . For example, if we knew sufficient information about the distribution of peoples' personal network sizes in the U.S., we could then ask them whether they know anyone who has ever been raped or who has AIDS, etc. There are important social, economic, moral, and political reasons for wanting to know such data and others that have eluded realistic estimation. Many such studies will provide values of unknown e 's (e.g., for rape victims, child abuse victims). The reliability of these estimates depends on how much we know about the distribution of personal network size for various populations. All of this also depends on the essential validity of the probability model which is the framework for this analysis.

APPENDIX

Preliminaries

We consider a population T , of size $t \gg 1$, having a subpopulation E , of size $e > 0$, which is the subgroup of T associated with some attribute or event. For each member u of $T - E$ we let $k(u)$ denote the number of people in T that u "knows." For our purposes here "u knows v" means that u knows v personally, in the sense that u knows v by name, knows where v lives, and knows v 's occupation. The people in T whom u knows will be called the *personal network* of u , denoted by $K(u)$.

We allow $k(u)$ to vary with u over $T - E$ and to take its values on a finite interval of nonnegative integers $[n_0, n_0 + n]$ where $n \geq 0$. Regarding average personal network size, we first examine the general case and the case where the distribution of $k(u)$ is a single point n_0 (where $n = 0$) and then the special cases where $k(u)$ has either a binomial, uniform, or two-point distribution. We then address the question of estimating event subpopulation size.

Now, we need to make a fundamental assumption, either about the distribution of the members in the various personal networks $K(u)$ or about the distribution of the members of E , as follows:

- A. For a random member u of $T - E$, all subsets of $T - \{u\}$ of size $k(u)$ were equally likely to have been the subset $K(u)$ known by u .
- B. All subsets of T of size e were equally likely to have been the subpopulation E .

In some situations (but possibly not the Mexico City earthquake) version B seems plausible. If all of the downtown buildings in a city were similar in level of earthquake survivability and all socioeconomic strata of the population were randomly represented in the downtown population when an earthquake occurred centered downtown, etc., then this assumption may not be a bad one. Version A implies the assumption that for a random u in $T - E$ the probability any particular member of $K(u)$ is in E is just the relative size of E in T , e/t , when e is very small compared to t .

We first consider the general case where the distribution of $k(u)$ for u in $T - E$ is unspecified. We let $P(X)$ denote the probability that the event X occurs and $P(X|Y)$ the conditional probability that the event X occurs given that event Y has occurred. For a random member u of $T - E$ we let N denote the event that u knows no one in E and let $P(k) \equiv P(k(u) = k)$ and $P(N|k) \equiv P(N|k(u) = k)$. Then, in general, we have

$$(A.1) \quad P(N) = \sum_{m=0}^n P(N | n_0 + m)P(n_0 + m).$$

Now, for random u in $T - E$, we have by version A above that

$$(A.2A) \quad P(N|k) = C(t - e - 1, k)/C(t - 1, k),$$

where $C(n, m)$ denotes the binomial coefficient $n!/m!(n - m)!$, and by version B that

$$(A.2B) \quad P(N|k) = C(t - k - 1, e)/C(t - 1, e).$$

It is easily verified that the right sides of (A.2A) and (A.2B) are algebraically the same, which means that either one may be used in (A.1) for $k = n_0 + m$. Their common expression can be written as

$$(A.3) \quad P(N|k) = [1 - e/(t - 1)][1 - e/(t - 2)] \dots [1 - e/(t - k)] \\ = [1 - e/(t - g_k)]^k,$$

where, by continuity, such real numbers g_k exist with $1 \leq g_k \leq k$ for $k > 0$.

Substituting (A.3) into (A.1) and again invoking continuity, we obtain

$$(A.4) \quad P(N) = \sum_{m=0}^n [1 - e/(t - g)]^{n_0 + m} P(n_0 + m),$$

where such real number g exists with $1 \leq g \leq n_0 + n$. Now, without loss of generality, we may assume that the distribution of the values $k(u)$ on the integers in $[n_0, n_0 + n]$ is such that $P(n_0) > 0$ and $P(n_0 + n) > 0$. Then, since $P(n_0 + m) \geq 0$ for $0 \leq m \leq n$,

$$(A.5) \quad \sum_{m=0}^n P(n_0 + m) = 1,$$

and $0 < 1 - e/(t - g) < 1$, we have from (A.4) that

$$(A.6) \quad [1 - e/(t - g)]^{n_0 + n} \leq P(N) \leq [1 - e/(t - g)]^{n_0},$$

with strict inequality, $<$, in both inequalities of (A.6) when $n > 0$.

Then, letting p denote the proportion of the members of $T - E$ who know someone in E and \ln denote the natural logarithm, we obtain from (A.6)

$$(A.7) \quad n_0 + n \geq \ln(1 - p)/\ln[1 - e/(t - g)] \geq n_0,$$

which leads to the following general result, where $\epsilon = e/t$.

Lemma A.1. Under either of the assumptions A or B, the value

$$(A.8) \quad \alpha \equiv \ln(1 - p) / \ln[1 - c/(t - g)] \approx \ln(1 - p) / \ln(1 - \epsilon),$$

determined by the values of c , p and t and the distribution of $k(u)$, must lie within the range of values $[n_0, n_0 + n]$. The right hand approximation in (A.8) is excellent when $n_0 + n$ is very small compared to t .

The value α is an *anchor value* for the range of personal network sizes in $T - E$ and must be within this range for any frequency distribution of personal network size. In particular, if $k(u)$ has the one-point distribution, where $n = 0$, the average personal network size is $c = n_0 = \alpha$. When $n_0 + n$ is very small compared to t the error due to taking $g = 0$ is insignificant, the value of α is virtually independent of the frequency distribution of $k(u)$, and we obtain the right hand approximation in (A.8).

Thus, with an empirical estimate r for p and under either of the distribution assumptions A or B we can estimate α by (A.8).

Although the Mexico City earthquake sample does not meet our statistical requirements, it is tantalizing to use the data from this sample to estimate α . With $r = 91/400 = 0.2275$, we estimate α for the different death rates ϵ to the nearest integer in Table 1, taking $g = 0$ and replacing c by α . The right hand approximation in (A.8) is correct to within 0.1% here, assuming $n_0 + n \leq 10,000$.

Average Personal Network Size

From (A.4), setting $P(N) = 1 - p$ and defining $q_m \equiv P(n_0 + m)$ for $m = 0, 1, \dots, n$, we obtain

$$(A.9) \quad 1 - p = [1 - c/(t - g)]^{n_0} \sum_{m=0}^n [1 - c/(t - g)]^m q_m,$$

or

$$(A.10) \quad \alpha = n_0 + \ln \left\{ \sum_{m=0}^n [1 - c/(t - g)]^m q_m \right\} / \ln[1 - c/(t - g)] \geq n_0.$$

Letting c denote the average value of $k(u)$ in $T - E$, we have

$$(A.11) \quad c = \sum_{m=0}^n (n_0 + m)q_m = n_0 + \sum_{m=0}^n mq_m \leq n_0 + n.$$

Thus, we have $\alpha \leq c$ if and only if

$$(A.12) \quad \sum_{m=0}^n [1 - c/(t - g)]^{mq_m} \geq \prod_{m=0}^n [1 - c/(t - g)]^{mq_m},$$

which, since each $q_m \geq 0$ and $\sum_{m=0}^n q_m = 1$, is just the generalized arithmetic mean - geometric mean inequality for the positive quantities $[1 - c/(t - g)]^m$, with weights q_m , $m = 0, 1, \dots, n$ [cf. Hardy, Littlewood & Polya, 1952, p. 17]. Since all of these quantities $[1 - c/(t - g)]^m$ are different, we in fact have strict inequality, $>$, in (A.12) when at least two of the values q_m are positive, i.e., when the probability distribution for $k(u)$ has at least the two points n_0 and $n_0 + n$ for $n > 0$. Thus from all the above, we have the following result.

Theorem A.2. Under either assumption A or B and for any probability distribution of the values $k(u)$ on the integer interval $[n_0, n_0 + n]$, the anchor value α and the average value c of the personal network sizes $k(u)$ must satisfy the inequalities

$$(A.13) \quad n_0 \leq \alpha \leq c \leq n_0 + n.$$

For the one-point distribution, where $n = 0$, all three inequalities in (A.13) are equalities. For a distribution with at least two points, where $n > 0$, all three inequalities are strict inequalities $<$.

Of course, strict inequality in an inequality of (A.13) is only a strict numerical inequality, which may not be substantively significant between quantities which are almost equal. Note that if $\alpha_1, \alpha_2, \dots, \alpha_s$ are anchor values corresponding to different event subpopulations E_1, E_2, \dots, E_s then we have

$$(A.14) \quad c \geq \max \alpha_i, \\ 1 \leq i \leq s$$

We now examine some special cases where we assume a particular distribution for $k(u)$.

Binomial distribution. We first consider a binomial distribution of $k(u)$ over $[n_0, n_0 + n]$, where n is viewed as the number of opportunities or encounters (the "trials" of the binomial distribution) that u has with other members v of T , over and above a fixed set of n_0 members whom u already knows, each of which has probability π of resulting in u knowing v (i.e., resulting in "success"). The average number of people in T that the members of $T - E$ know is then $c = n_0 + n\pi$. We then have

$$(A.15) \quad P(n_0 + m) = C(n,m)\pi^m\xi^{n-m}, \text{ where } \xi = 1 - \pi,$$

so (A.4) becomes

$$(A.16) \quad P(N) = [1 - e/(t - g)]^{n_0} \sum_{m=0}^n C(n,m) [\pi - \pi e/(t - g)]^m \xi^{n-m} \\ = [1 - e/(t - g)]^{n_0} [\pi + \xi - \pi e/(t - g)]^n$$

or

$$(A.17) \quad P(N) = [1 - e/(t - g)]^{n_0} [1 - \pi e/(t - g)]^{(c - n_0)/\pi}.$$

Setting $P(N)$ equal to $1 - p$ and taking natural logarithms of both sides of (A.17), we obtain

$$(A.18) \quad \ln(1 - p) = n_0 \cdot \ln[1 - e/(t - g)] + \{(c - n_0)/\pi\} \cdot \ln[1 - \pi e/(t - g)],$$

which, since $\ln[1 - e/(t - g)] \neq 0$, becomes by (A.8)

$$(A.19) \quad n_0 + \{(c - n_0)/\pi\} \cdot \ln[1 - \pi e/(t - g)] / \ln[1 - e/(t - g)] = \alpha.$$

Now, for e and $n_0 + n$ very small relative to t we have, to an excellent approximation, that

$$(A.20) \quad (1/\pi) \cdot \ln[1 - \pi e/(t - g)] \approx \ln[1 - e/(t - g)],$$

whence (A.19) becomes

$$(A.21) \quad c \approx \alpha.$$

Thus, the values of c for the Mexico city data when the values of $k(u)$ have a binomial distribution over the integer interval $[n_0, n_0 + n]$ are virtually the same as those given in Table 1, and are practically independent of the value of the trial success probability π . More generally, we have

Theorem A.3. Under either assumption A or B, if the personal network sizes of the members of T - E have a binomial distribution and e and $n_0 + n$ are very small relative to t then every anchor value α for the distribution range is an excellent approximation to the average personal network size c .

Uniform distribution. We next consider the case when the values of $k(u)$ have a uniform distribution over the integer interval $[n_0, n_0 + n]$. Here $P(n_0 + m)$

$= 1/(n + 1)$ for $m = 0, 1, 2, \dots, n$, so that $c = n_0 + n/2$ or $n = 2c - 2n_0$, and (A.4) becomes

$$(A.22) \quad P(N) = \{1/(n + 1)\} \cdot [1 - c/(t - g)]^{n_0} \sum_{m=0}^n [1 - c/(t - g)]^m.$$

Then for $P(N)$ set equal to $1 - p$ this becomes

$$(A.23) \quad 1 - p = \{(t - g)/(2c - 2n_0 + 1)e\} \cdot [1 - c/(t - g)]^{n_0} \cdot \{1 - [1 - e/(t - g)]^{2c - 2n_0 + 1}\}.$$

Numerically solving (A.23) for c for the Mexico City data ($r = 0.2275$) for $n_0 = 0, 100, 200, 300, 400, 500$ and 600 , using $g = 1$, we obtain the results in Table 2, to the nearest integer.

We note, by Theorem A.2, that the value of n_0 cannot exceed α , and as n_0 approaches α from below the value of c approaches α from above, which means that n approaches 0 and the distribution of $k(u)$ approaches the one-point distribution. In fact, the relationships among these values are sufficiently robust that the lowest values in each column of Table 2 are already the corresponding lower bound values shown in Table 1.

Two-point distribution. Finally, we examine the case where $k(u)$ has a two-point distribution. Here we have $P(n_0) = 1 - \beta$ and $P(n_0 + n) = \beta$ for $0 < \beta < 1$. The value of c is easily found to be $c = n_0(1 - \beta) + (n_0 + n)\beta = n_0 + n\beta$, whence $n = (c - n_0)/\beta$. In this case (A.4) becomes

$$(A.24) \quad P(N) = [1 - e/(t - g)]^{n_0}(1 - \beta) + [1 - e/(t - g)]^{n_0 + n\beta},$$

Table 2. Values of the Average Personal Network Size for the Mexico City Data with a Uniform Distribution ($t = 18,000,000$)

n_0	e:	7000	12000	15000	22000
0	c:	695	405	324	221
100	c:	686	396	316	214
200	c:	678	391	311	211
300	c:	673	387	310	—
400	c:	668	—	—	—
500	c:	665	—	—	—
600	c:	664	—	—	—

from which can be derived, after setting $P(N) = 1 - p$ and substituting for n ,

$$(A.25) \quad c = n_0 + \beta \{ \ln[(1-p) / (1 - c/(t-g))]^{n_0} - (1-\beta) \} - \ln \beta / \ln | 1 - c/(t-g) |.$$

Now, the right side of (A.25) is only defined for

$$(A.26) \quad \beta > 1 - (1-p) / (1 - c/(t-g))^{n_0} \equiv \beta_\infty > 0,$$

where the second inequality in (A.26) is true since $\alpha > n_0$ for a two-point probability distribution. Then, as $\beta \rightarrow \beta_\infty^+$ the right side of (A.25) increases without bound, whence $c \rightarrow \infty$. The latter, of course, is not substantively feasible; however, it represents the possibility that c can become very large. For the case $n_0 = 0$, where $\beta_\infty = p$, the values of c for various values of β approaching $\beta_\infty \approx r = 0.2275$ for the Mexico City data are given to their nearest integers in Table 3, using $g = 1$. Here, under the plausible bound on personal network size given by $n_0 + n \leq 10,000$ (used in all error analysis), $c = n\beta_n \leq (10000)(0.2323) = 2323$, which is already overstepped in the first two columns of Table 3.

Note that, for $\beta = 0.9999$, the two-point distribution with $n_0 = 0$ is very close to the one-point distribution at $n_0 + n = n$, and the values in Table 3 support this. Note from (A.25), however, that, as $\beta \rightarrow \beta_\infty^+$, $c \rightarrow \infty$ logarithmically, a relatively slow rate of unbounded growth, as seen in Table 3. Hence, no upper

Table 3. Values of the Average Personal Network Size for the Mexico City Data with a Two-Point Distribution and $n_0 = 0$ ($t = 18,000,000$)

β	c:	7000	12000	15000	22000
0.9999	c:	664	387	310	211
0.5000	c:	780	455	364	248
0.3000	c:	1095	639	511	348
0.2500	c:	1548	903	722	492
0.2300	c:	2674	1559	1247	850
0.2280	c:	3589	2093	1674	1141
0.2276	c:	4523	2638	2110	1439

bound may be placed on c for this distribution (and thus for general distributions) unless more is known about the shape of the distribution.

Event Subpopulation Size

From (A.4) with $P(N) = 1 - p \equiv \hat{p}$ and $c/(t - g) \approx c/t = \epsilon$, to an excellent approximation, we obtain with $1 - \epsilon = \hat{\epsilon}$ and $q_m = P(n_0 + m)$, $m = 0, 1, \dots, n$,

$$(A.27) \quad \hat{p} = \sum_{m=0}^n q_m \hat{\epsilon}^{n_0+m}.$$

Since $\hat{\epsilon} > 0$ and $q_m \geq 0$ for all $m = 0, 1, \dots, n$ we see that \hat{p} and all its derivatives with respect to $\hat{\epsilon}$ are positive, whence \hat{p} is an increasing function of $\hat{\epsilon}$. Thus, if there are event subpopulations E_1, E_2, \dots, E_s , ordered so their corresponding $\hat{\epsilon}_i$ values satisfy

$$(A.28) \quad \hat{\epsilon}_0 \equiv 1 > \hat{\epsilon}_1 > \hat{\epsilon}_2 > \dots > \hat{\epsilon}_s,$$

then we also have for their corresponding \hat{p}_i values

$$(A.29) \quad \hat{p}_0 \equiv 1 > \hat{p}_1 > \hat{p}_2 > \dots > \hat{p}_s,$$

where $\hat{\epsilon}_0$ and \hat{p}_0 also satisfy (A.27). Alternatively, these chains of inequalities are equivalent to

$$(A.30) \quad \epsilon_0 \equiv 0 < \epsilon_1 < \epsilon_2 < \dots < \epsilon_s$$

and

$$(A.31) \quad p_0 \equiv 0 < p_1 < p_2 < \dots < p_s.$$

Now let E_x be a new event subpopulation of unknown size e_x and unknown relative size ϵ_x for which the probability p_x , that a random u in $T - E_x$ knows anyone in E_x , satisfies in (A.31)

$$(A.32) \quad p_{k-1} < p_x < p_k, \text{ for some } k, 1 \leq k \leq s.$$

Then, from (A.30) we have

$$(A.33) \quad \epsilon_{k-1} < \epsilon_x < \epsilon_k.$$

Thus, if our probability model is reasonably close to correct then, given a sufficiently broad range of ϵ_i, p_i pairs from previous event subpopulations, we should be able to bound the size of the new event subpopulation between successive values

$$(A.34) \quad c_{k-1} < \epsilon_x < c_k, \text{ for some } k, 1 \leq k \leq s.$$

Clearly, if (A.32) is true but not (A.33) and (A.34), then either the data values $\epsilon_i, p_i, 1 \leq i \leq s$, are poor or the original probability model is not correct. Thus, if the data are believed to be good we have a negative criterion for the validity of the underlying probability model. Since the earthquake data do not furnish more than one pair of values ϵ_i, p_i , we are not yet in a position to make this test.

Now, assuming in this model that \hat{p} is a differentiable function of $\hat{\epsilon}$, we have from (A.27) that

$$(A.35) \quad d\hat{p}/d\hat{\epsilon} \Big|_{\hat{\epsilon}=1} = \sum_{m=0}^n (n_0 + m)q_m \hat{\epsilon}^{n_0+m} \Big|_{\hat{\epsilon}=1} = \sum_{m=0}^n (n_0 + m)q_m = c.$$

Thus, for a fairly large value for c (at least 211 by Table 1) and for $\hat{\epsilon}$ less than 1 but in the vicinity of 1, we see that large changes in \hat{p} correspond to small changes in $\hat{\epsilon}$. This indicates that, whatever the size of the bound within which p_x sits in (A.32), the corresponding size of the bound for the approximation of ϵ_x in (A.33) will be much smaller.

Now suppose that, for a fixed population T (more precisely, $T - E$ for which e is very small relative to t), we know the average personal network size c . From (A.8) we derive the relation

$$(A.36) \quad \hat{p} = \hat{\epsilon}^\alpha,$$

where α is a function of $\hat{\epsilon}$ and \hat{p} and, hence, need not be constant over different pairs $\hat{\epsilon}, \hat{p}$. Now, by Theorem A.2, we have for $0 < \hat{\epsilon} < 1$

$$(A.37) \quad \hat{\epsilon}^\alpha \geq \hat{\epsilon}^c$$

or, by (A.36)

$$(A.38) \quad \epsilon \geq 1 - (1 - p)^{1/c} \equiv \bar{\epsilon}.$$

Thus, for E_x of unknown relative size ϵ_x , with accurately estimated probability p_x of a person in $T - E_x$ knowing someone in E_x , we have

$$(A.39) \quad \bar{\epsilon}_x = 1 - (1 - p_x)^{1/c} \leq \epsilon_x,$$

which yields a lower bound approximation $\bar{\epsilon}_x$ to the true value ϵ_x . Note that the closer c is to α , or ϵ_x is to 0, the better the approximation $\bar{\epsilon}_x$ is to ϵ_x . The latter implication corresponds to the fact that the closer $\hat{\epsilon}$ is to 1 the better the approximation of $\hat{\epsilon}^\alpha$ by $\hat{\epsilon}^c$.

REFERENCES

- Bernard, H.R., Killworth, P.D., Evans, M.J., McCarty, C., & Shelley, G.A. (1988). Studying social relations cross-culturally. *Ethnology*.
- de Sola Pool, I., & Kochen, M. (1978). Contacts and influence. *Social Networks*, 1, 5-51.
- Gurevich, M. (1961). *The social structure of acquaintanceship networks*. Unpublished doctoral dissertation, MIT.
- Hardy, G.H., Littlewood, S.E., & Pólya, G. (1952). *Inequalities*. Cambridge, England: Cambridge University Press.
- Killworth, P.D., & Bernard, H.R. (1978). The reverse small world experiment. *Social Networks*, 1, 159-192.