ESTIMATING THE SIZE OF AN AVERAGE PERSONAL NETWORK AND OF AN EVENT SUBPOPULATION: SOME EMPIRICAL RESULTS

H. Russell Bernard, University of Florida; Eugene C. Johnsen, University of California;
Peter D. Killworth, Hooke Institute, Oxford; Scott Robinson, Universidad Autònoma Metropolitana
Eugene C. Johnsen, Department of Mathematics, UCSB, Santa Barbara, CA 93106

Abstract:

An interesting problem in social network analysis is to determine the number of people whom an average person knows, i.e., his/her personal network size. An important related problem is to estimate the size of a given event subpopulation. In an attempt to estimate these quantities realistically, the authors have developed a probabilistic model and have applied it to a small first sample of data from the Federal District of Mexico City in order to relate various proposed sizes of the population of victims who died in the 1985 Mexico City earthquake to average personal network size in the Federal District.

The method involves asking members of a sufficiently large random sample of a population of size t if they know anyone in a fixed event subpopulation of size e. This produces an estimate of the probability p that anyone in the population (usually the population minus the event subpopulation) knows someone in the event subpopulation. Using an equal likelihood probability model, this leads to a lower bound estimate for c, the average number of people that a person in the population knows. When personal network size has a binomial distribution over the population this value is, in fact, an estimate for c itself. Except for pathological distributions, such as an extreme form of the two-point distribution, this appears to be approximately true for other distributions as well.

Here we test this method on further data from Mexico City, where a random sample of residents was asked whether they know personally anyone in each of several different event subpopulations of known and unknown sizes. Although the model does not fit the data for the individual subpopulations of known size well, we are able to develop procedures for obtaining various bounds and estimates for c and to determine some of the respondents' attributes on which variation in probability of knowing someone in an event subpopulation and variation in personal network size seem to depend. We apply these bounds and estimates to the estimation of the value of e for the population of rape victims in Mexico City. According to the model e increases monotonically with increasing p, and if we use an acceptable value of c supported by previous analysis we can obtain a bounded estimate for this unknown e. We also apply the method to data on AIDS victims in the United States and find a range of values for c which is consistent with that for Mexico City.

1. Introduction.

An important but vexing problem in social network analysis has been to determine in a population the number of people a person knows, i.e., his/her personal network size, and the mean, range, and distribution of this variable in the population as a whole (cf. [1]). These data have heretofore defied successful investigation. In an earlier paper [2], we presented a probabilistic method for estimating the average size of a personal network and the size of an event subpopulation in a given total population. We applied it to a first small random data sample of size 400 from the population of the Federal District of Mexico City in order to relate various proposed sizes of the subpopulation of victims of the 1985 Mexico City earthquake to the average personal network size in the Federal District. We give here and in the next two sections a brief summary of this method and these first results. We then apply the method to the data from a second larger random sample from the population of Mexico City proper in order to obtain estimates of the size of the average personal network from various known event subpopulation sizes, and then use this information to estimate the size of an unknown event subpopulation and to compare against results for a known event subpopulation in the U.S..

Consider a population T, of size $t >> 1$, 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 let k(u) denote the number of people in T that u "knows". Here "u knows v" means that u knows v personally, in that u knows v by name, knows where v lives, knows v's occupation, and that v knows the same about u. 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 give some results on the general case, the case where the distribution of k(u) is a single point n_0 (where $n = 0$) and 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) are 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 some of those discussed in this paper, e.g., the Mexico City earthquake) version B seems plausible. In the case of the earthquake, 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.

2. Average Personal Network Size.

For the general case, where the distribution of k(u) for u in T - E is unspecified, we have the following results (Bernard et al. [2]). We let p denote the proportion of the members of T - E who know someone in E, ln denote the natural logarithm, and $\epsilon = e / t$.

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

(1) $\quad \alpha \equiv \ln(1-p) / \ln[1 - e/(t-g)] \approx \ln(1-p) / \ln(1-\epsilon)$,

determined by the values of e, p *and* t *and the distribution of k(u), must lie within the range of values* $[n_0, n_0+n]$. *The right hand numerical approximation in (1) 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 (1).

Thus, with an empirical estimate r for p and under either of the distribution assumptions A or B we can estimate α, and frequently c, by (1).

We also have the following result.

Theorem 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*

(2) $\quad n_0 \le \alpha \le c \le n_0 + n$.

For the one-point distribution, where n = 0, *all three inequalities in* (2) *are equalities. For a distribution with at least two points, where* n > 0, *all three inequalities are strict inequalities* <.

Thus, if $\alpha_1, \alpha_2, \cdots, \alpha_s$ are anchor values corresponding to different event subpopulations $E_1, E_2, \cdots, E_s$ then we have

(3) $\quad c \ge \max_{1 \le i \le s} \alpha_i$.

Our first Mexico City earthquake sample of 400 random respondents did not meet our statistical requirements; however, it was tantalizing to use the data from this sample to find α as a lower bound estimate for c. With r = 91/400 = 0.2275, we computed α for the different proposed death rates e to the nearest integer in Table 1 (taking g = 0). The right hand approximation in (1) is correct to within 0.1% here, assuming $n_0+n \le 10000$.

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

(a) Binomial Distribution

For a binomial distribution of k(u) over $[n_0, n_0+n]$, where n may be 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 a fixed probability of resulting in u knowing v (i.e., resulting in "success") we have

(4) $\quad c \approx \alpha$.

Thus, the values of c for the Mexico City earthquake 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 fixed trial success probability. More generally, we have the following result.

Theorem 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.

(b) Uniform Distribution

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

Table 1.
Values of the Lower Bound Estimate α of Average Personal Network Size c
for the Mexico City Earthquake Data with Varying Death Rates e (t = 18,000,000)

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

obtain the results shown in Table 2 to the nearest integer for the first sample of Mexico City earthquake data.

We note that in Table 2, as in Theorem 2 and its proof (Bernard et al. [2]), the value of n_0 does not 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 the columns of Table 2 are already the corresponding lower bound values shown in Table 1.

(c) Two-point Distribution

Finally, we examine the case where k(u) has a two-point distribution with $P(k(u) = n_0) = 1-\beta$ and $P(k(u) = n_0+n) = \beta$ for $0 < \beta < 1$. Here we can derive (cf. [2])

$$(5)\quad c = n_0 + \beta \left\{ \ln\left[(1-p)/[1-e/(t-g)]^{n_0}-(1-\beta)\right]-\ln\beta \right\} / \ln[1-e/(t-g)].$$

Now, the right side of (5) is only defined for

$$(6)\quad \beta > 1 - (1-p)/[1-e/(t-g)]^{n_0} \equiv \beta_\infty > 0,$$

where the second inequality in (6) is true since $\alpha > n_0$ for a two-point probability distribution. Then, as $\beta \to \beta_\infty^+$ the right side of (5) increases without bound, whence $c \to \infty$. 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 first sample of 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 $n_0 + n \le 10{,}000$ (used in all error analysis), $c = n\beta_n \le (10000)(0.2323) = 2323$, 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 (5), however, that as $\beta \to \beta_\infty^+$, $c \to \infty$ logarithmically, a relatively slow rate of unbounded growth, as seen in Table 3. Hence, no upper bound may be placed on c for this distribution (and thus for general distributions) unless more is known about the shape of the distribution.

3. Event Subpopulation Size.

For the probability distribution of k(u) with $\hat{p} \equiv 1 - p$, $\hat{\epsilon} \equiv 1 - \epsilon$, and $q_m = P(k(u) = n_0+m)$, $m = 0, 1, \cdots, n$, we obtained (cf. [2]) that

$$(7)\quad \hat{p} = \sum_{m=0}^{n} q_m \hat{\epsilon}^{\,n_0+m}.$$

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

$$(8)\quad \epsilon_0 = 0 < \epsilon_1 < \epsilon_2 < \cdots < \epsilon_s,$$

then we also have for their corresponding p_i values

$$(9)\quad p_0 = 0 < p_1 < p_2 < \cdots < p_s.$$

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	-	-	-

where $\hat{\epsilon}_0 = 1 - \epsilon_0$ and $\hat{p}_0 = 1 - p_0$ also satisfy (7).

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 (9)

(10) $\quad p_{k-1} < p_x < p_k$, for some k, $1 \le k \le s$.

Then, from (8) we have

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

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

(12) $\quad e_{k-1} < e_x < e_k$, for some k, $1 \le k \le s$.

Clearly, if (10) is true but not (11) and (12) then either the data values e_i, p_i, $1 \le i \le s$, are poor or the original probability model is not completely correct. Thus, if the data are believed to be good we have a negative criterion for the full validity of the underlying probability model.

Assuming in this model that $\hat{p}$ is a differentiable function of $\hat{\epsilon}$, we have from (7) that

$$(13) \quad d\hat{p}/d\hat{\epsilon} \,\Big|_{\hat{\epsilon}=1} = \sum_{m=0}^{n} (n_0+m)\, q_m\, \hat{\epsilon}^{\,n_0+m-1} \,\Big|_{\hat{\epsilon}=1} = \sum_{m=0}^{n} (n_0+m)\, q_m = c .$$

Thus, for a fairly large value for c (at least 211 by Table 1) and for $\hat{\epsilon}$ less than but very close to 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 (10), the corresponding size of the bound for the approximation of ϵ_x in (11) will be considerably 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 (1) we derive the relation

(14) $\quad \hat{p} = \hat{\epsilon}^{\,\alpha}$,

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 (2), we have for $0 < \hat{\epsilon} < 1$ that

(15) $\quad \hat{\epsilon}^{\,\alpha} \ge \hat{\epsilon}^{\,c}$

or, by (14),

(16) $\quad \epsilon \ge 1 - (1 - p)^{1/c} \equiv \tilde{\epsilon}$.

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

(17) $\quad \tilde{\epsilon}_x = 1 - (1 - p_x)^{1/c} \le \epsilon_x$,

which yields a lower bound approximation $\tilde{\epsilon}_x$ to the true value ϵ_x. Here, the closer c is to α_x, or ϵ_x is to 0, the better the approximation $\tilde{\epsilon}_x$ is to ϵ_x. The latter implication corresponds

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	e:	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

to the fact that the closer $\hat{\epsilon}$ is to 1 the better the approximation of $\hat{\epsilon}^{\alpha}$ by $\hat{\epsilon}^{c}$.

As an example, from the first sample data for the Mexico City earthquake with event subpopulation E of assumed size e = 7000 and probability p = 0.2275, suppose we determined that $c \approx 664$. Now suppose for a new event subpopulation E_x we obtain $p_x = 0.1986$. Then the size e_x is bounded by $e_x < 7000$ and underestimated by $\tilde{e}_x = \tilde{\epsilon}_x \cdot t = [1 - (0.8014)^{1/664}] \cdot (18{,}000{,}000) = 6000.67$, so $6001 \le e_x < 7000$. Now, by (7), the graph of $\hat{p}$ as a function of $\hat{\epsilon}$ is increasing and concave upwards; hence, we can do a linear interpolation between the points $(\hat{\epsilon}_0, \hat{p}_0) = (1,1)$ and $(\hat{\epsilon}_1, \hat{p}_1) = (0.99961111\cdots, 0.7725)$ to obtain the tighter upper bound $e_x \le 6110$, whence $6001 \le e_x \le 6110$.

4. Application to Sample Data from Mexico City.

In a later second survey we obtained a larger random sample of 2260 from Mexico City proper (t = 10,700,000) in the hope of establishing a set of reference value pairs (p, ϵ) against which to compare new value pairs (p', ϵ') according to (10) and (11) above. The data for the six reference event subpopulations, with the 95% confidence ranges for p and corresponding ranges for α, are given in Table 4.

It is clear from this table that the monotonicity property of the model given by (8) and (9) does not hold, which indicates that either the data are inaccurate, the data are somewhat accurate but not very precise, or the model is not valid in its simple form for different subpopulations (or possibly a combination of either the first or second and the third).

By the nature of the survey and the data obtained, the first alternative (including its combination with the third) appears untenable. But, before ruling out the model in its simple form, we assume that the data are somewhat accurate but not very precise. This suggests analyzing the data at the aggregate level to gain precision and, it is hoped, detect a "signal" amidst the "noise". Since the simple model (1) with the assumption of an approximately binomial distribution of k(u) for u in T (or T - E) produces an approximately constant $c \approx \alpha = \ln \hat{p} / \ln \hat{\epsilon}$, which says that $\ln \hat{p}$ and $\ln \hat{\epsilon}$ vary linearly with respect to each other, we attempt to determine c (via α) by least squares linear regression of each variable against the other. Now, there are four ways to do this, namely

$$(18) \quad \ln \hat{p} = \alpha \cdot \ln \hat{\epsilon},$$

$$(19) \quad \ln \hat{p} = \alpha \cdot \ln \hat{\epsilon} + \ln \beta,$$

$$(20) \quad \ln \hat{\epsilon} = \alpha^{-1} \cdot \ln \hat{p},$$

Table 4.
Known sizes of, Probabilities of Knowing Someone in, and Corresponding Values of α for Six Reference Subpopulations of Mexico City Proper (t = 10,700,000)

subpopulation	e	range of p	α	range of α
Doctors	30426	0.3889 ± 0.0201	173	162 - 185
Mailmen	14728	0.1473 ± 0.0146	116	103 - 128
Bus Drivers	11696	0.2571 ± 0.0180	272	250 - 294
Quake Victims	10000	0.2668 ± 0.0182	332	306 - 359
TV Repairmen	4013	0.2619 ± 0.0181	810	745 - 876
Priests	1595	0.2854 ± 0.0186	2254	2082 - 2431

Table 5.
Probabilities of Knowing Someone in the Event Subpopulation According to the Partition of the Sample by Age in Years

subpopulation	≤ 20	20-34	35-49	50-65	≥65
Doctors	0.3171	0.3756	0.4484	0.4921	0.4000
Mailmen	0.1111	0.1503	0.1659	0.1746	0.1143
Bus Drivers	0.2685	0.2599	0.2646	0.2063	0.2000
Quake Victims	0.2824	0.2789	0.2399	0.2275	0.2286
TV Repairmen	0.2847	0.2798	0.2399	0.1852	0.0857
Priests	0.2269	0.2720	0.3363	0.3598	0.4000

(21) $\ln \hat{\epsilon} = \alpha^{-1} \cdot \ln \hat{p} + \ln \gamma$,

where $\ln\beta$ and $\ln\gamma$ are included in the unconstrained regressions. We obtain the following results, with α to the nearest integer, for the data in Table 4 :

(18.r) $\alpha = 196$,

(19.r) $\alpha = 56$, $\beta = 0.7761$,

(20.r) $\alpha = 274$,

(21.r) $\alpha = 221$, $\gamma = 1.0003$.

The values of α in (18.r) and (20.r) (for the simple model) yield the range 235 ± 39, which includes the α values for Bus Drivers and line (21). The data point for Bus Drivers almost lies on all three lines (19), (20) and (21). Except for the data point for Quake Victims, the other data points do not lie very close to any of these lines. Line (19) suggests some problem with the data or the model, whereas line (21) does not. We may also determine the best fit lines (18) and (19) in the sense of least squares distance of the data points from the lines (lines (20) and (21) are then, respectively, equivalent to (18) and (19)). For these we obtain the following results from the data in Table 4:

(18.s) $\alpha = 274$,

(19.s) $\alpha = 221$, $\beta = 0.9356$.

These best fit lines, which treat the variables symmetrically, are virtually the same as (20) and (21), respectively, further supporting the corresponding α values 274 and 221.

For each of the six reference subpopulations the sample (and hence also the total population) was partitioned into subclasses according to (i) zone of survey interview, whether socioeconomically lower, middle or upper class, (ii) age of respondent, whether < 20, 20-34, 35-49, 50-65, and > 65 years, (iii) highest education level attained by respondent, whether < 4, 4-6, 7-12, 13-16, and > 16 years, (iv) socioeconomic class of respondent, whether upper, middle, and lower class, (v) occupation of respondent, whether working at home (home), out of home (oohm), retired (retd), unemployed (unem), students (stud), and a residual class (rsdl), and (vi) respondent's reporting of how many people he/she believes to be in his/her personal network, whether ≤ 100, 100-500, 500-1000, 1000-1500, > 1500, and no answer. We present the p values for the subsamples determined by age, education, socioeconomic class, occupation, and number believed known in Tables 5, 6, 7, 8 and 9. Since the full sample was not a quota sample for the different subclasses of T the subsample sizes in these subclasses varied considerably,

Table 6.
Probabilities of Knowing Someone in the Event Subpopulation
According to the Partition of the Sample by Education in Years

subpopulation	≤4	4-6	7-12	13-16	≥17
Doctors	0.2539	0.2769	0.3735	0.5146	0.7333
Mailmen	0.0933	0.1076	0.1549	0.1683	0.2519
Bus Drivers	0.3161	0.2749	0.2833	0.1756	0.1556
Quake Victims	0.1554	0.2032	0.2853	0.3244	0.3481
TV Repairmen	0.1399	0.1096	0.2814	0.3976	0.4444
Priests	0.2176	0.2470	0.2892	0.3146	0.4074

Table 7.
Probabilities of Knowing Someone in the Event Subpopulation
According to the Partition of the Sample by Socioeconomic Class

subpopulation	lower	middle	upper
Doctors	0.2696	0.4644	0.6323
Mailmen	0.1239	0.1743	0.1097
Bus Drivers	0.3211	0.2245	0.0710
Quake Victims	0.2200	0.3066	0.2903
TV Repairmen	0.1606	0.3203	0.5097
Priests	0.2250	0.3285	0.3742

from 16 in How Many: No Answer to 1158 in Age: 20 - 34.

We note that there is a great deal of monotonicity in this data, either increasing or decreasing values of p with increasing values of the table variable when the table variable has a natural ordering. For example, the data for Priests shows that p increases without exception with increasing age, education, and socioeconomic class and almost without exception with increasing believed personal network size. For Doctors, p increases without exception with increasing education, socioeconomic class, and believed personal network size and almost without exception with increasing age. A similar but weaker version of this occurs for Mailmen.

With regard to all six reference subpopulations, p increases with increasing believed personal network size without exception for Doctors and Quake Victims and almost without exception for Mailmen, Bus Drivers, TV Repairmen, and Priests. Thus, believed personal network size behaves like actual personal network size does under the assumption of the simple model. Except for believed personal network size, p decreases without exception for Bus Drivers with increasing socioeconomic class and almost without exception with increasing age and education. For Quake Victims and TV Repairmen the patterns of how p varies with increasing values of the four table variables are very similar. For socioeconomic class and education the patterns of how p varies for the six reference subpopulations are also very similar, indicating that functionally, for our purposes here, these are very similar attributes.

In order to see whether there is more information in this data, we plot the known value of 100ϵ against the various values of p in Tables 5 - 9 for each of the six reference subpopulations. For each reference subpopulation this yields a range of p values, from the minimum to the maximum, which we call the p-spread for that subpopulation. These are plotted as horizontal lines in Figure 1. On each p-spread is shown the full sample p value and corresponding α value from Table 4. Now, although there is no simple model curve of the form (14) which comes close to passing through all the points (p, ϵ) given by the data in Table 4, we try to obtain the next best thing, namely, a simple model curve which intersects a maximum number of these p-spreads. Unfortunately, there is no such curve which intersects all six or even five of the six p-spreads. However, there is small set of such curves which intersect the p-spreads for Doctors, Bus Drivers and Quake Victims, and comes close to intersecting the p-spreads for both Mailmen and TV Repairmen. The best estimating curve of this set, which just fails to intersect the p-spreads for Mailmen and TV Repairmen by about the same difference in p, has a nearest integer α value of 220. This value is virtually the same as that for line (21) and the best fit line found above and is easily

Table 8.
Probabilities of Knowing Someone in the Event Subpopulation
According to the Partition of the Sample by Occupation

subpopulation	home	oohm	retd	unem	stud	rsdl
Doctors	0.3986	0.3906	0.6957	0.1948	0.4196	0.3391
Mailmen	0.1026	0.1613	0.0870	0.1039	0.1473	0.1845
Bus Drivers	0.2243	0.2575	0.0870	0.1948	0.2455	0.3734
Quake Victims	0.1862	0.2849	0.2609	0.2078	0.2857	0.3133
TV Repairmen	0.1432	0.2651	0.3043	0.2078	0.3750	0.2575
Priests	0.3174	0.2868	0.4348	0.2468	0.2545	0.2790

Table 9.
Probabilities of Knowing Someone in the Event Subpopulation
According to the Partition of the Sample by Believed Personal Network Size (in hundreds)

subpopulation	no ans	≤ 1	1 - 5	5 - 10	10 - 15	≥ 15
Doctors	0.3750	0.2634	0.3941	0.4552	0.5282	0.5680
Mailmen	0.0625	0.1156	0.1525	0.1253	0.1897	0.2524
Bus Drivers	0.0000	0.2245	0.2585	0.2353	0.2615	0.4272
Quake Victims	0.3125	0.1801	0.2754	0.3095	0.3385	0.3981
TV Repairmen	0.3125	0.1438	0.2472	0.3350	0.4359	0.4320
Priests	0.1875	0.2124	0.2895	0.2890	0.3436	0.4806

within the bounds 235 ± 39 given by lines (18) and (20). This lends further support to a simple model with $c \approx \alpha \approx 220$. We note that the p values where the curve for $\alpha = 220$ intersects the p-spreads for Doctors, Bus Drivers and Quake Victims are convex linear combinations of p values for certain of the subsamples given in Tables 5 - 9. Each such convex combination may thus be viewed as an indicator for the corresponding event subpopulation. The important question here is whether an indicator gives consistent results under repeated sampling.

From the number of respondents in each class with respect to *believed* personal network size (except No Answer), using the midpoint value as the mean for each of the first four size classes and 1800 for the mean of the last size class, we obtain the average *believed* personal network size of 516. This is consistent with the results for α from both Mexican surveys.

Qualitatively, the above results are consistent with or explainable by current social knowledge and theory, lending support to the simple model and data. However, for no value category of any table variable does the monotonicity relation between p and ϵ, as given by (8) and (9), occur even approximately, which tends to disconfirm the model in the absence of major error in the data. It appears that developing from this an accurate and precise method for estimating the size of an unknown subpopulation will require either elaboration of the model or further improvement of the data, or both.

5. Application to Another Event Subpopulation in Mexico City.

We next attempt to estimate the unknown size of the subpopulation of Rape Victims in Mexico City proper. In the second survey we obtained the estimated $p \approx 0.1491$ for t = 10,700,000. For the various α values of interest obtained earlier, we obtain the estimates of the subpopulation of Rape Victims given in Table 10.

From the curve for $\alpha = 220$ in Figure 1 we can estimate bounds for the number of Rape Victims in Mexico City proper according to min(p) = 0.0571 (for Age ≥ 66 yrs.) and max(p) = 0.3111 (for Education ≥ 17 yrs.). When the resulting p-spread for Rape Victims intersects this curve at minimum p, ϵ is minimum, and when it intersects it at maximum p, ϵ is maximum. As generally shown by the figure, we obtain $\min(\epsilon) = 0.000267$ and $\max(\epsilon) = 0.001692$, which translates to min(e) = 2859 and max(e) = 18110, so $2859 \leq e \leq 18110$.

Figure 1.

p - Spreads for Six Reference Subpopulations with Best Estimating Simple Model and Estimate of Relative Size Range for Unknown Raped Subpopulation in Mexico City Proper (t = 10,700,000)

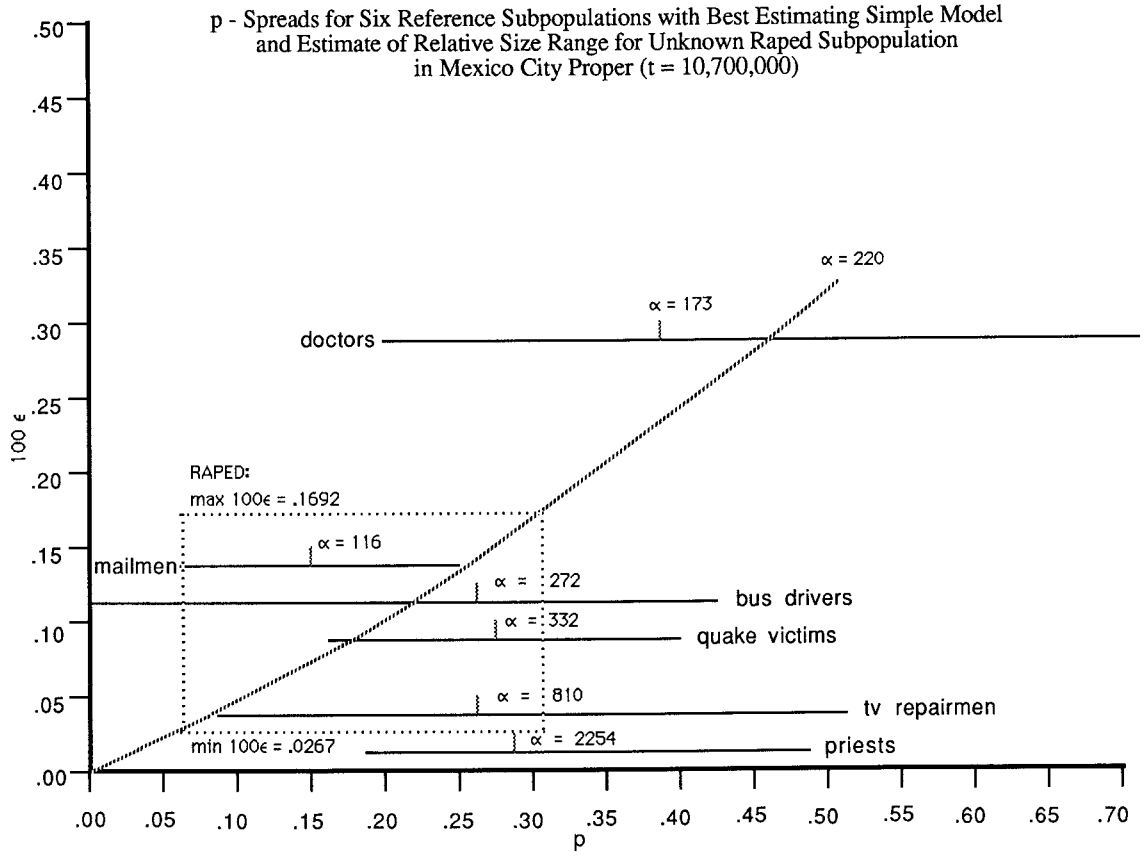

Table 10.
Estimates of the Number of Rape Victims in Mexico City
for Various Obtained Values of α

α:	116	196	220	235	274	332	810	2254
e:	14883	8811	7850	7349	6303	5202	2133	766

Note that all e values in Table 10 except those determined by α = 810 and 2254 lie in this interval. This suggests that the event population of priests and possibly also that of TV Repairmen are not good reference event subpopulations for purposes of estimating others. For the α values lying in the interval 235±39 or $196 \leq \alpha \leq 274$ we have $6303 \leq e \leq 8811$.

6. Application to a Subpopulation in the U.S.

In a Media General - Associated Press poll of 1304 randomly selected adults across the U.S. taken in April 1987, one of the questions asked was whether the respondent knew anyone with AIDS (cf. [3]). Seven percent of them said they did. Using this figure, an estimated May 1, 1987 U.S. adult population (over 17 years of age) of 179,955,000 based on U.S. Census data (cf. [4], [5]), and the diagnosed number of AIDS victims as of early May 1987 of 35219, we can apply the simple model to estimate the corresponding value $\alpha \approx 371$. For the given three percent margin of error (at an assumed confidence level of 95 percent), this gives a range of $208 \leq \alpha \leq 539$. This is consistent with other values and bounds for α which we have determined with this model for the Federal District of Mexico City and for Mexico City proper.

References and Notes

[1] de Sola Pool, I. and M. Kochen, "Contacts and Influence", *Social Networks* 1 (1978), 5 - 51.

[2] Bernard, H. R., E. C. Johnsen, P. D. Killworth, and S. Robinson, "Estimating the Size of an Average Personal Network and of an Event Subpopulation", in The Small World (M. Kochen, ed.), Ablex, N.J., 1988.

[3] Kilman, L., "AIDS Most Feared Of Diseases: Poll", *Santa Barbara News Press*, Santa Barbara, California, May 12, 1987.

[4] U.S. Bureau of the Census, "Estimates of the Population of the United States, by Age, Sex, and Race: 1980 to 1986", *Current Population Reports* , Population Estimates and Projections, Series P-25, No. 1000, Issued February 1987, Table A, p. 2.

[5] U.S. Bureau of the Census, "Estimates of the Population of the United States to May 1, 1987", *Current Population Reports* , Population Estimates and Projections, Series P-25, No. 1007, Issued July 1, 1987.

[6] Supported in part by NSF Grant BNS-8318132 and by a Faculty Research Grant from the Graduate School, University of Florida. We particularly wish to express our thanks to Dr. Donald Price, Vice President for Research, University of Florida, for generous support of this research, and to student researchers Maria del Carmen Costa, Alejandro Casteneira, Yolanda Hernandez Franco, Patricio Meade, and Miguel Angel Riva-Palacio for their conscientious efforts in the collection of the main data analyzed in this paper.