The authors have developed and tested scale-up methods, based on a simple social network theory, to estimate the size of hard-to-count subpopulations. The authors asked a nationally representative sample of respondents how many people they knew in a list of 32 subpopulations, including 29 subpopulations of known size and 3 of unknown size. Using these responses, the authors produced an effectively unbiased maximum likelihood estimate of the number of people each respondent knows. These estimates were then used to back-estimate the size of the three populations of unknown size. Maximum likelihood values and 95% confidence intervals are found for seroprevalence, 800,000 ±43,000; for homeless, 526,000 ±35,000; and for women raped in the last 12 months, 194,000 ±21,000. The estimate for seroprevalence agrees strikingly with medical estimates, the homeless estimate is well within the published estimates, and the authors' estimate lies in the middle of the published range for rape victims.

ESTIMATION OF SEROPREVALENCE, RAPE, AND HOMELESSNESS IN THE UNITED STATES USING A SOCIAL NETWORK APPROACH

PETER D. KILLWORTH
Southampton Oceanography Centre

CHRISTOPHER McCARTY
H. RUSSELL BERNARD
University of Florida

GENE ANN SHELLEY
Georgia State University

EUGENE C. JOHNSEN
University of California, Santa Barbara

In this article, we describe improvements to a method (Killworth et al. 1998) for estimating the size of hard-to-count populations. The need to make such estimates is compelling. All efforts to redress social problems begin with the question: How big is the problem? How much money should be allocated

AUTHORS' NOTE: *The work was supported by NSF grant # SBR-9213615. The telephone survey was conducted by the Bureau of Economic and Business Research at the University of Florida.*

EVALUATION REVIEW, Vol. 22 No. 2, April 1998 289-308

to the problem of rape, for example? Of homelessness? Of helping people live with being HIV seropositive?

To determine the number of people who have ever been divorced in a population, we might survey a representative sample of the population and ask respondents directly if they have ever been divorced. A few people might find the question offensive, but most people would answer the question honestly. Then, if *n%* of the sample reported having been divorced, we would report that *n%* of the total population, plus or minus some error, have, in fact, been divorced.

We cannot, however, ask people if they have been raped or if they are HIV-positive and expect to get valid answers. We cannot interrogate populations that have perished (in wars, ethnic cleansings, or natural disasters). To estimate the size of such populations has heretofore required indirect methods—methods that are costly and time consuming and that produce inherently imperfect results. Moreover, the methods for indirectly estimating each hard-to-count population are ad hoc. New methods must be developed to take account of each population's special circumstances.

We have developed what we call "network scale-up methods" for estimating the size of all populations. These methods are based on a developing theory of social structure, so that improvements to the methods expand their usefulness for counting a greater variety of populations. The simplest network model equates two quantities. The first is the proportion of all those known by a person who are members of a subpopulation; this is equal to m / c, where each person knows c people, m of whom are in the subpopulation. The second is the proportion of a total population of size t, say, occupied by the subpopulation, of size e. Thus, $m / c = e / t$. Such a simple model works only under three strong assumptions. (a) Everyone has an equal chance of knowing someone in the subpopulation, (b) c (network size) is a constant, and (c) everyone has perfect knowledge about the members of their network (Bernard et al. 1989, 1991).

Clearly, some people in our society have a greater probability than others of knowing someone who is, for example, HIV-positive. Clearly, some people have bigger networks (know more people), whereas others have smaller networks. And clearly, our knowledge about the details of the lives of our network members is uneven—we might know if some of our network members are HIV-positive, but we would not know that about others.

Correspondence and requests for reprints should be sent to Dr. Peter D. Killworth, Southampton Oceanography Centre, Empress Dock, Southampton SO14 3ZH, England; e-mail: P.Killworth @soc.soton.ac.uk.

A theory of social networks would handle these problems. From quantities known about an individual (their gender, their income, their ethnicity, and so on), a network theory would let us assign to them a value for their network size c, a probability that they know someone who is HIV-positive, and a probability that if they know a person and if that person is HIV-positive, then they are aware that the person is HIV-positive (Bernard et al. 1991; Johnsen et al. 1995; Killworth et al. 1990, 1998).

To approach this problem, we ask respondents how many people they know in a large set of subgroups, some of known size, some of unknown size. We use this information to estimate c for each respondent (the size of his or her social network), and we then back-estimate the size of populations of unknown size.

METHODS

Data were collected from a nationally representative sample (in terms of gender, age, race, education, and income) of 1,554 members of the United States in a telephone survey in October and November 1994. The 1,554 compose 51.2% of the 3,035 originally contacted. (This response rate is better than the typical rate for random digit dial surveys, where the sample is representative of the United States.) Based on a recent survey of 1,920 adults across the United States involving eight reputable firms, the Council for Marketing and Opinions Research (1997) concluded that a refusal rate of 63% (the opposite of a cooperation rate) was normal given disclosure of survey length and no incentive, such as a token payment. This conclusion assumes a very strict method of calculating refusal rates, namely (initial refusals plus breakoffs) divided by (completes plus refusals plus eligible plus not available plus language barrier). Using this formula, our refusal rate was 43%, some 20% below the cited study.[1] Nonetheless, a potential nonresponse bias, whose size is unknown, remains.

Respondents were asked, inter alia, about people they "knew"; for the survey, "knowing" someone meant that a respondent knew that person and vice versa, by sight or by name, that the person could be contacted, that contact (either in person, by telephone, or by mail) had occurred during the last 2 years, and that the person lived within the United States. Respondents stated how many people they knew (henceforth called their "active network," composed of all their alters) who were in each of the 32 subgroups (Table 1). Of these, 29 subgroups are of known sizes. The remaining 3, whose size we

wish to estimate, are HIV-seropositives, the homeless, and victims of rape. Because definitions vary, our wording of the questions was:

- How many people do you know who have tested positive for HIV, the AIDS virus? Just to be sure, are you certain that this person/these people actually received an HIV test that came back positive, or are you unsure?
- How many people do you know who are currently homeless? These are people who live in shelters, transient hotels (costing less that $12 per night), abandoned buildings or in open areas.
- How many women do you know who have been raped in the past 12 months? This *does include* women who have not reported the rape to the authorities.

The telephone survey cost $6.50 per respondent and took on average 10 minutes.

The estimation method used extends the simple model above. Suppose again that the population, here the United States, is size t and contains a subpopulation of size e. If a member of the population reports knowing m people in the subpopulation, and he or she knows an average of c people, then

$$\frac{m}{c} = \frac{e}{t}. \tag{1}$$

There is, of course, scatter here, which the methods below deal with.

In normal circumstances, c is the only unknown in Equation (1) and so can be estimated. If c is the considered known, Equation (1) can be used with observed m values to estimate unknown values of e. The utility of Equation (1) thus rests with the degree of accuracy with which c can be estimated. Here, we use maximum likelihood estimates based on many simultaneous values of m and e for different subpopulations. The idea is to find, for a given respondent, what value for the number of people he or she knows best fits the observed values of m; that is, to take the *pattern* of the ms and construct a best estimate of c.

Specifically, suppose there are L subpopulations, of size e_j, $j = 1, 2, \ldots$ and respondent i reports knowing m_{ij} members of subpopulation j. Then, if size of i's active network is c_i, and the probability of knowing any member of one subpopulation is independent of the others, we have

$$Prob\ (i \text{ knows } m_{ij}, j = 1,2,\ldots, L) = \prod_{j=1}^{L} {}_{c_i}C_{m_{ij}}\, p_j^{m_{ij}} q_j^{c_i - m_{ij}}, \tag{2}$$

by the binomial theorem. Here, p_j is the proportion of subpopulation j in the population, that is, $p_j = e_j / t$, and $q_j = 1 - p_j$. There is a unique value of c_i that

TABLE 1: The Subgroups Reported on By Respondents, the Mean Number in Each Subgroup Reported in Respondents' Active Networks, and the Size of Each Group in the United States

Subpopulation	*Mean Number Known*	*Subpopulation Size (millions)*
First name "Michael"	4.288	3.187
First name "Christina"	1.180	0.351
First name "Christopher"	1.649	1.255
First name "Jacqueline"	0.774	0.291
First name "James"	3.576	4.023
First name "Jennifer"	2.150	1.118
First name "Anthony"	1.536	0.874
First name "Kimberly"	1.449	0.642
First name "Robert"	4.049	3.811
First name "Stephanie"	1.192	0.510
First name "David"	3.128	2.865
First name "Nicole"	1.033	0.358
American Indian	2.677	2.0
Woman given birth last 12 months	3.720	4.0
Woman adopted a child in past 12 months	0.306	0.118
Man or woman widowed and under age	3.164	3.335
Person on kidney dialysis	0.420	0.17
Postal worker	2.310	0.81
Commercial pilot	0.649	0.534
Member of JAYCEES	1.504	0.19
Person with diabetes	3.401	6.5
Opened a business in past 12 months	1.125	0.63
Have twin brother or sister	1.999	5.3
Licensed gun dealer	0.545	0.240
HIV estimate	0.910	Our estimate
Came down with AIDS	0.469	0.207
Homeless	0.599	Our estimate
Women raped in past 12 months	0.220	Our estimate
Male incarceree in state or federal prison	1.342	0.892
Homicide victims in past 12 months	0.241	0.0274
Committed suicide in past 12 months	0.208	0.0302
Died in motor accident in past 12 months	0.480	0.0452

maximizes the probability (2) and thus provides a maximum likelihood estimate for c_i (Killworth et al. 1998). It can be shown that this is given by

$$c_i = t \cdot \frac{\sum_{j=1}^{L} m_{ij}}{\sum_{j=1}^{L} e_j}, \tag{3}$$

provided the subgroup sizes are small compared with t and that the m_{ij} are small compared with c_i. The estimate for c_i is unbiased (Killworth et al. 1998), with a standard error $\sqrt{(tc_i / \Sigma_{j=1}^{L} e_j)}$, so that accuracy improves with more subgroups. Also, estimates for $1 / c$ and back-estimates for e values are essentially unbiased when there are more than 20 subpopulations used for predictive purposes; the 29 used here are more than sufficient for the purpose.

For Equation (1) to hold, certain assumptions about statistical independence must hold.

A natural question concerns heterogeneity: We have assumed the size of a subgroup is directly related to the probability that a respondent knows a member of the subgroup. This assumption has been examined in previous work. In Killworth et al. (1998), we tested this assumption with the subpopulation of people with particular first names, one of the subgroups we used in this research. In that study, we found that of the 14 names we used, the mean number known to respondents was correlated at .79 with our estimate of the proportion of the population with those names. Similarly, of the other subpopulations we used that were not first names, the mean number known was correlated at .83 with known subpopulation sizes. A similarly high correspondence between the estimate by respondents and the size of a subpopulation, which is arguably more accurately known (homicide victims within the last 12 months), was shown by Johnsen et al. (1995) in their estimate of AIDS victims using a similar method. Nonetheless, the assumption remains both a concern and a research question.

Another statistical requirement is that subgroups should be independent from respondents. This implies that (inter alia) the correlation between the numbers known in any two subpopulations should be positive; this is the case; only three of the ${}_{32}C_2$ combinations had a negative correlation (–.01 in each case). Additionally, common sense showed that knowledge of most subgroups was not affected by measurable indicators of the respondents, although occasional correlations remain likely to exist (respondents' income and knowledge of swimming pool ownership, for example). There were no significant differences between estimates of subpopulation size using the first or second half of the data (the "split-half" technique).

IMPROVEMENTS TO ESTIMATES OF UNKNOWN SUBPOPULATIONS

Although a maximum likelihood estimate has been previously used for estimating c_i, no such estimate was used for back-estimating unknown (or,

indeed, known) subpopulations. Instead, a value for each unknown e_J ($J > L$) was determined per respondent, and these values were averaged over respondents. Here, we improve the procedure by asking, Given a best estimate of the c_i, and the pattern m_{ij} of how many are known by i in subgroup j, what is the most likely value of e_J? Assuming independence between respondents, we have for subpopulation J,

$$Prob(c_i, m_{iJ}, e_J) = \prod_i {}_{c_i}C_{m_{iJ}} p_J^{m_{iJ}} (1-p_J)^{c_i - m_{iJ}}, \text{ where } p_J = \frac{e_J}{t}, \tag{4}$$

and this probability is to be maximized by varying e_J. Because the c_i and m_{iJ} are fixed, this is the same as maximizing $\Pi_i p^{m_{iJ}} (1-p)^{c_i - m_{iJ}}$, which in turn involves maximizing, over p,

$$p^{\Sigma_i m_{iJ}} (1-p)^{\Sigma_i c_i - \Sigma_i m_{iJ}}.$$

Because the maximum of $p^a(1-p)^{n-a}$ occurs when $p = a/n$, this implies that the maximum likelihood estimate is given by

$$\hat{e}_J = t \frac{\Sigma_i m_{iJ}}{\Sigma_i c_i} = \sum_i m_{iJ} \frac{\Sigma_j e_j}{\Sigma_j m_{ij}}, \tag{5}$$

from (3), which will be used henceforth. The statistics of this estimate are straightforward, assuming the c_i are estimated accurately. The m_{ij} are drawn from a binomial distribution of the size c_i and probability $p = e_J / t$, which is assumed small. Algebra then gives

$$E(\hat{e}_J) = e_J \text{ (so the estimate is unbiased)}, \tag{6}$$

$$\text{standard error of } \hat{e}_J = \sqrt{\frac{te_J}{\Sigma_i c_i}}. \tag{7}$$

This estimate becomes more accurate as the total size of the networks of all respondents increases, as would be expected.

This error estimate implicitly assumes that the estimated c_i are correct; the sum of these enters the denominator in (5). Each estimated $\hat{c}_i$ has its own error, of course. The effect of these errors can be estimated either by analysis or by direct Monte Carlo simulation (which confirms the analysis). We have that var $(\Sigma_i \hat{c}_i) = \Sigma_i$ var $(\hat{c}_i)$, in that the respondents are independent. From results already cited, the variance is just

$$\text{var}\left(\sum_i \hat{c}_i\right) = \frac{t\,\Sigma_i c_i}{\Sigma_j e_j},$$

so that

$$\frac{\text{s.error}\,(\Sigma_i \hat{c}_i)}{E\,(\Sigma_i \hat{c}_i)} = \frac{t^{1/2}}{(\Sigma_j e_j)^{1/2}(\Sigma_i c_i)^{1/2}} = \frac{1}{(\Sigma_{ij}\, m_{ij})^{1/2}}, \tag{8}$$

after use of definition of c_i. Thus, provided respondents altogether list an appreciable number of people ($\Sigma_{ij} m_{ij}$) in known populations, the denominator in (5) has negligible errors (in the cases reported here, 1.96 standard errors are less than 1% of the mean of the denominator) so that the main source of error is the numerator, as estimated.[2]

RESULTS

Applying our methods to the telephone survey data, the mean number of alters in the active networks of 1,554 nationally representative respondents is 286, SD = 291, with a predicted standard error of 40. This value for the mean c is significantly ($p < 1\%$) larger than the 108 found in our survey of 1,524 Florida residents (Killworth et al. 1998). The average number of alters among just the Florida residents in the national survey, however, is almost identical to the number found for all respondents. In our Florida study, however, we asked each respondent about only 8 subgroups, compared to 32 subgroups in the national study. We believe that the lower estimated value for c in our study of Florida residents is due, in part, to the far smaller number of subgroups asked about in that study.

Figure 1 shows a histogram of the c distribution. We believe this to be the first detailed histogram of network size. Its shape is similar to crude estimates found with other methods. Killworth and Bernard (1978) conducted a series of studies where respondents provided the name of someone they knew who would be the best first link in an imaginary chain to a mythical target person. As a consequence of providing links to 500 target people where duplicates could be used, respondents generated a list of all the people they knew who would be good initial contacts. The distribution of network size was similar to the one generated from this study.

Freeman and Thompson (1989) estimated network size among respondents by asking them if they knew anyone with a particular last name. The last names were randomly selected from the Orange County, California phone

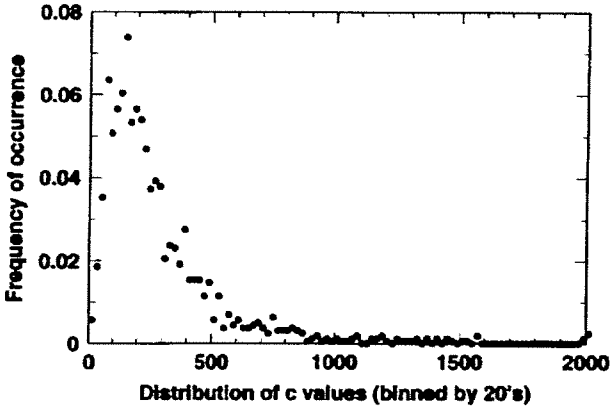

Figure 1: Histogram of the Frequencies of Occurrence of Active Network Size *c* Among Respondents With Bins of Size 20.
NOTE: Instances of *c* more than 2,000 are not shown directly.

book. Again, the distribution of network sizes for respondents generated by this method was similar to ours, although the actual network sizes were an order of magnitude different.

Although both methods generated actual network sizes that were different from what we found here, the *distributions* are quite similar. We are less concerned about actual sizes as we know this to be quite sensitive to the definition of a tie given to respondents and the cueing mechanism used to elicit names. We are encouraged that the distributions are similar, because this is something that should be the same even if actual levels are different. The fact that three very different methods generated similar distributions is telling.

Subpopulation sizes can be back-estimated using Equation (5), as well as by simple averaging. As a check on the accuracy of the method, each of the 29 known subpopulations was back-estimated using an estimate of c created using the other 28. The results (Figure 2) are most encouraging. The correlation between our estimates and the actual subpopulation sizes for these 29 groups is .79 (63% of variance), and the points cluster closely about the line of equality, save for a few outliers.

However, consistent with previous findings (Killworth et al. 1998), small subpopulations ($e < 2$ million, say) are overestimated, and large subpopulations underestimated. The two largest subpopulations, twins and diabetics, are highly underestimated. Neither being a twin nor being a diabetic carries

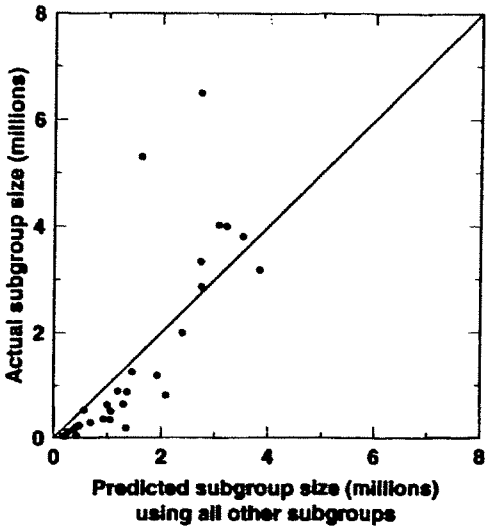

Figure 2: Predicted Versus Actual Subgroup Size for the 29 Subgroups of Known Size. NOTE: The calculation for a given subgroup uses the data from the other 28 known subgroups to estimate the active network size c, and then back-estimates using the maximum likelihood estimate Equation (5). The line of equality is shown.

any stigma (which is why some facts about people are not transmitted to all the alters in networks), but neither are they "newsworthy." We have anecdotal evidence that a decade or more can pass without a respondent finding out that close coworkers or business associates have a twin or are diabetic. Without these two subgroups, the correlation in Figure 2 rises to .94 (88% of variance).

Comparison with our older method of simply averaging e_j estimates over respondents (Killworth et al. 1998) is enlightening. The new maximum likelihood method is more accurate than the old method for 22 of the 29 known subpopulations, and it considerably reduces the overestimate of small subpopulations. The correlation between estimated and actual subpopulation sizes is marginally better for the old method (.80) although less so without the two outliers (.93). The two sets of estimates are very highly correlated.

Estimates for the unknown subpopulations are, using (5), with 95% confidence limits:

seroprevalence: 800,000 ± 43,000;
homeless: 526,000 ± 35,000;
women raped in the last 12 months: 194,000 ± 21,000.

Using our older methods, these figures are, respectively, 753,000, 772,000, and 303,000, which we will see are probably less accurate. Accordingly, we believe that this increase in accuracy demonstrates measurable progress toward the goal of a validated method to estimate subpopulation sizes.

DISCUSSION

The difficulties in producing accurate counts of these populations are well known. There is a second difficulty, in that we would wish to compare our estimates with other estimates relevant to the same time period. As much as we desire this, such other estimates are simply not available. This difficulty is one of the main reasons for our work.

We discuss estimates for each subpopulation in turn.

SEROPREVALENT

Official seroprevalence estimates are discussed more fully elsewhere (Killworth et al. 1998). The method of back-calculation based on the number of HIV-infected individuals necessary to account for the incidence of AIDS gives estimates between 800,000 and 1 million (Gail and Brookmeyer 1988; Rosenberg et al. 1991). Other methods are potentially error prone, for example, the survey of child bearing women, which uses sentinel hospital reports and adjustments, first to all women and second to all men; the errors in this procedure are unknown. Other estimates range between 300,000 and 1.5 million (CDC 1990). Our own estimate based on Florida resident responses to an earlier survey was 1.6 million (Killworth et al. 1998). The current estimate is thus thoroughly in line with official estimates, but is half our previous estimate.

This difference relates to at least four effects.[3] (a) We have improved the estimating equation since conducting the Florida survey. Applying Equation (5) to the Florida data reduces our earlier estimate from 1.6 million to 1.2 million. (b) Asking about fewer subgroups (8 compared to 32 in the present study) lowered our estimate of c by half. (c) People in Florida report knowing 0.63 HIV-positive persons compared with 0.91 reported by the national survey. Recall that recalculation using only responses from Florida respondents in the current survey gave almost identical results to the full survey, so that the higher incidence of AIDS in Florida is not accounting for the difference in estimates. (d) We have corroborated the fact that people who

test positive for HIV withdraw from interaction with people who may be judgmental of them, effectively limiting the size of their networks (Shelley et al. 1995).

HOMELESS

Extrapolating from the self-reports of 1,507 respondents (interviewed by telephone in a national, representative, survey) Link et al. (1994) concluded that (a) 5.7 million people were "literally homeless" (sleeping in shelters, abandoned buildings, bus and train stations, etc.) during the 5-year period 1985 through 1990, and that (b) an astonishing 13.5 million people in the United States have been literally homeless at some time during their lifetimes. Although these estimates help us evaluate the overall problem, they do not tell us anything about the number of people who are currently homeless.

This so-called point estimate has proven very difficult, as Jencks's (1994) review makes clear. The Census Bureau attempted a brute force count during the month prior to the 1990 census (Wright and Devine 1992). The attempt involved some 15,000 workers who visited urban shelters, bus and train stations, and other places where they might find homeless people. This effort produced a count of 228,821, but was held to be very low, even by the Census Bureau (Society 1994). The Urban Institute's 1987 study of 20 cities produced an estimate of about 508,000 homeless adults in the United States with from 567,000 to 600,000, including children (Burt and Cohen 1989; Rossi et al. 1987).

We asked our respondents how many people they knew who were currently homeless, defining the term in the literal sense above. We made it clear to respondents that we wanted to know only about currently homeless persons, but respondents may still have reported network alters who had been homeless in the recent or not-so-recent past. It is also possible that the homeless, like HIV-positive persons, restrict their contact with others and thus have networks of lower-than-average size. These factors would influence our estimate, but our point estimate of 526,000 ± 135,000 agrees well with the estimates of Burt and Cohen (1989) and Rossi et al. (1987).

RAPED

Perhaps the most difficult of the three populations to estimate is the annual number of rape victims. Respondents are reluctant to provide information. Statistics for rape victims are made either by compilations from FBI reports or by interview techniques (Bureau of Justice Statistics 1994; Kilpatrick et

al. 1992). Interview reports from the National Crime Victimization Survey (NCVS) yield about 130,000 rapes per year between 1987 and 1991, of which approximately half are stated to have been reported to the police (Bachman 1994). Now police reports, when other acts of violence are committed with rape, list only the other act involved. Even taking this into account, FBI reports (of 102,560 in 1990) are believed to underestimate interview reports by 20% to 30% (Reiss and Roth 1993), although there are suggestions (Koss 1992) that NCVS figures themselves are underestimates. Another example of underestimation is given by Kilpatrick et al. (1992), who find only 12% of the National Women's Study rape victims reported the rape to police within 24 hours, with 4% more later.

The number of rapes projected by any survey, however, is most strongly influenced by how the question is asked.

The NCVS has been conducted annually since 1973, to estimate the actual number of crimes that occur each year in the United States. In 1992, the NCVS was based on some 59,000 households comprising 110,000 inhabitants older than 11. Until 1991, the projected count of rapes from the NCVS was based on volunteered responses to a series of questions about attacks and threats of attacks. "When a woman indicated" on the NCVS "that she had been the victim of an attempted or completed rape, she was not asked to explain what happened further. Her personal classification of the incident as a completed or attempted rape was accepted and recorded" (Bachman 1993). Note that respondents were *not* asked about rape or any form of sexual assault in this survey—the question concerned violence in general. The implication is that reports of rape are again likely to be underestimates.

Beginning in 1992, respondents were asked directly, "Has anyone attacked or threatened you . . . with rape, attempted rape, or other type of sexual assault?" and "Have you been forced or coerced to engage in unwanted sexual activity by (a) someone you didn't know before, (b) a casual acquaintance, or (c) someone you know well?" (Bachman and Saltzman 1995, 8). If respondents indicate confusion about the meaning of the question, then the interviewer reads the following definition:

> Rape is forced sexual intercourse and includes both psychological coercion as well as physical force. Forced sexual intercourse means vaginal, anal, or oral penetrations by the offender(s). This category also includes incidents where the penetration is from a foreign object such as a bottle. (Bachman and Taylor 1994, 508; cited in Koss 1996, 60)

The NCVS projected approximately 172,000 completed rapes in 1992 and 1993 (Bachman and Saltzman 1995).

Making the question more direct is certainly responsible for some of the dramatic increase in the estimated number of completed rapes from 1990 to 1992. The problem is further confused by the adding of the phrase "psychological coercion." As Koss (1996, 60) points out, the addition of the phrase was "probably meant to refer to verbal threats of bodily harm or rape, which are crimes" but may suggest to respondents "situations involving false promises, threats to end the relationship, continual nagging and pressuring, and other verbal strategies to coerce sexual intercourse, which are undesirable but not crimes."

The National Women's Survey of 1992 was based on a representative sample of 4,008 women older than 18 in the United States who were contacted by random digit dialing (Koss 1996, 62). Women were asked, "Has a man or boy ever made you have sex by using force or threatening to harm you or someone close to you? Just so there is no mistake, by sex we mean putting a penis in your vagina." (The survey also asked about other forms of rape.)

Extrapolating from the results of the survey, the National Victim Center reported that 683,000 rapes (± 135,000) had probably taken place in 1990 (Kilpatrick et al. 1992; Koss 1996).

Our results for 1994, based on Equation (5) above, are in the middle of the wide range of estimates for the number of rapes in the United States and are almost identical to the 1992 and 1993 estimates of the NCVS. In future research, however, a more specific question on rape will be asked for better intercomparison, using the Bachman and Taylor wording. It should also be noted that our survey involved only those older than 18, whereas the NCVS data included females older than 12.

RELIABILITY OF THE ESTIMATES

Estimates depend on the number of subgroups used in the back-estimation. Figure 3 shows the range of this behavior, with three examples of populations of known size and the three populations whose size we want to know.

The estimated number of pilots is consistently accurate, irrespective of the number of subgroups in the survey. The number of "women who gave birth last year" is slightly underestimated with data from just four subgroups, but beginning with five subgroups it is well estimated. The estimate remains close to the true figure until 19 subgroups, but with 20 or more subgroups the number of women who gave birth is overestimated. The number of people who committed suicide in the last 12 months is consistently overestimated.

From these data, then, no firm guide can be given as to the optimal or minimum number of subgroups that need to be included in the network

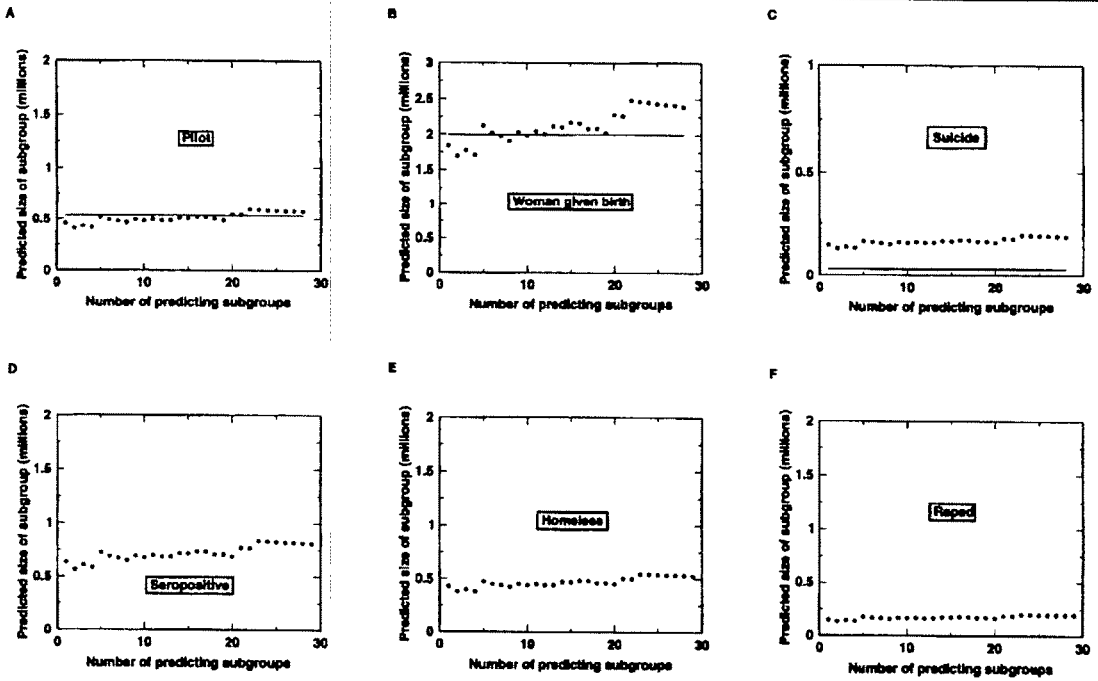

Figure 3: The Behavior of Estimates as the Number of Subgroups Used to Make the *c* Estimate Is Increased.
NOTE: The subgroups are chosen to illustrate the widest variety of behavior possible. The horizontal firm line shows the actual subgroup size. The subgroups were added in the order in Table 1. Shown are (A) pilots, (B) woman given birth, (C) suicides, (D) seropositive, (E) homeless, and (F) rape victims.

scale-up survey. However, in the three cases of populations whose size we do not know (seropositive people, homeless people, and rape victims), the estimates are independent of the number of subgroups. We are encouraged by this because the estimates for these populations are consistent with estimates made by other researchers, using entirely different methods.

The consistent overestimation of small subgroups and the underestimation of large subgroups remains to be explained. We assume that this must be due to patterns of misreporting by respondents and that this relates to the difficulty of knowing that someone in one's active network is, in fact, a member of a given subgroup. For example, information about the death of a close friend in a motor vehicle accident will propagate quickly through one's network; it may take more time to find out that a casual acquaintance has suffered the same fate. Some sensitive information (like seropositive status) may remain hidden forever from all but a few people in a network (cf. Shelley et al. 1990; Shelley et al. 1995). Were all subpopulations equally stigmatized, this would be less of a difficulty, but the contrast between a stigmatized and nonstigmatized group (e.g., those named Jacqueline) is likely to cause difficulties. This issue needs further work. In one potential area of bias (subpopulations based on first names) Brewer (1997) examines—and rejects—bias.

Thus, there appear to be transmission problems (rate of information propagation, and whether information is passed on) and barrier problems (affecting whether information can propagate at all to certain parts of networks). We tried to deduce the factors that account for overreporting and underreporting by respondents, but the results supported neither our interpretation of the problem nor any other obvious interpretation. (The method involved seeking numerical factors by which reports should be multiplied to give completely consistent answers. Initial guesses for each factor were taken as unity. An interactive procedure was followed, in which successive improvements to each factor in the 29 were made repeatedly until further changes were computed to be zero, so that the solution had been obtained.) We plan to examine these issues in future work.

The network scale-up method described here has been shown to generate estimates of subgroup size in line with other estimates made by other methods. At the same time, it generates some further questions.

Among these is the question of the utility of a statistically accurate prediction of subgroup size (recall the 93% correlation between estimates and actual sizes of the known subpopulations after removal of the two outliers) when its numerical error is a sizable fraction of the actual subgroup size. This is a frequent difficulty with statistical methods and is clearly relevant here.

We also need to know more about how information propagates in networks, that is, about barrier and transmission effects. This involves investigation of

complete networks, not just parts of networks (e.g., the study of sexual partners in seropositive studies; Service and Blower 1995). The study of limited networks, however, may be useful in examining heterogeneity (e.g., in age) within networks.

CONCLUSION

We have described developments in network scale-up methods to estimate the size of uncountable subpopulations, to demonstrate the robustness of the answers produced. Our method involves asking representative samples of the U.S. population how many people they know in many populations of known size and how many they know in a few populations whose size we want to estimate. We use each respondent's responses about populations of known size to make a maximum likelihood estimate of how many people that respondent knows. Then we use the pattern of all respondents' responses about populations of unknown size to estimate the size of those (unknown) subpopulations.

The estimate of the size of the seropositive community agrees strikingly with estimates made by the Centers for Disease Control and Prevention; the estimate for homelessness is well within the widely varying published estimates and resembles strongly the estimate that is most widely accepted by social scientists and public policy analysts; the estimate for rape victims is in the middle of the estimate range and agrees strongly with the estimate made by the National Center for Victim Studies of the Bureau of Justice.

We tested our method by estimating the size of 29 populations of known size. The network scale-up method estimates accurately the size of 20 out of those 29 populations. These estimates, plus the close approximation of our estimates to other estimates of populations of unknown size, strengthen belief in the network scale-up methods.

We have identified problems with both under- and overreporting of population knowledge by respondents. (We are aware of underreporting for small populations and overreporting of large populations, because respondents, respectively, forget or guess. We suspect the reverse can also occur, as hinted at in Figure 2.) In addition, underreporting occurs because of *barrier* and *transmission* effects. A barrier occurs when there are factors (such as distance) that prevent—or reduce the likelihood of—a respondent from knowing someone in a population. Transmission effects occur when a respondent is not aware that someone he or she knows is in a specific population. Overreporting occurs when respondents are forced to choose a specific

number for their knowledge of a subpopulation and to guess. Our future program of research will investigate these features.

NOTES

1. Although it is encouraging that the practice of listing cooperation or refusal rates in academic studies is becoming more common, the reader should note that there are several methods for calculating such rates and should treat the results with this in mind.

2. A final source of error remains, of course, model error (if our assumptions are false). Although we discuss this after the results, we can note that the accurate findings concerning the test subpopulations at least suggest that our assumptions are unlikely to be too far in error.

3. Note also that the national survey took place 1 year after the Florida survey, during which time the number of HIV-positive would have increased. We do not have access to estimates on this fine a time scale for comparison.

REFERENCES

Bachman, R. 1993. Predicting the reporting of rape victimizations. *Criminal Justice and Behavior* 20:254-70.

———. 1994. Violence against women: A National Crime Victimization Survey Report. NCJ-145325. Washington, DC: U.S. Department of Justice.

Bachman, R., and L. E. Saltzman. 1995. Violence against women: Estimates from redesigned survey. NCJ-154348. Washington, DC: U.S. Department of Justice.

Bachman, R., and R. M. Taylor. 1994. The measurement of family violence and rape by the redesigned National Crime Victimization Survey. *Justice Quarterly* 11:499-512.

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 *the small world*, edited by M. Kochen, 159-75. Norwood, NJ: Ablex.

———. 1991. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social Science Research* 20:109-21.

Brewer, D. D. 1997. No associative biases in the first name cued recall procedure for eliciting personal networks. *Social Networks* 19 (4): 345-53.

Bureau of Justice Statistics. 1994. Criminal victimization in the United States, 1992. NCJ-145125. Washington, DC: U.S. Department of Justice.

Burt, M., and B. Cohen. 1989. *America's homeless: Numbers, characteristics, and programs that serve them*. Washington, DC: The Urban Institute Press.

Centers for Disease Control. 1990. HIV prevalence estimates and AIDS case projections for the United States: Report based upon a workshop. *MMWR* 39:7.

Council for Marketing and Opinion Research. 1997. CMOR refusal rates and industry image survey (summary of results). *Survey Research* 28:1-2.

Freeman, L. C., and C. R. Thompson. 1989. Estimating acquaintanceship volume. In *The small world*, edited by M. Kochen, 147-58. Norwood, NJ: Ablex.

Gail, M. H., and R. Brookmeyer. 1988. Methods for projecting course of acquired immune deficiency syndrome epidemic. *Journal of the National Cancer Institute* 80:900-11.

Jencks, C. 1994. *The homeless*. Boston: Harvard University Press.

Johnsen, E. C., H. R. Bernard, P. D. Killworth, G. A. Shelley, and C. McCarty. 1995. A social network approach to corroborating the number of AIDS/HIV+ victims in the U.S. *Social Networks* 17:167-87.

Killworth, P. D., and H. R. Bernard. 1978. The reverse small-world experiment. *Social Networks* 1:159-92.

Killworth, P. D., E. C. Johnsen, H. R Bernard, G. A. Shelley, and C. McCarty. 1990. Estimating the size of personal networks. *Social Networks* 12:289-312.

Killworth, P. D., E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. 1998. A social network approach to estimating seroprevalence in the United States. *Social Networks* 20:23-50.

Kilpatrick, D. G., C. N. Edmunds, and A. K. Seymour. 1992. *Rape in America: A report to the nation*. Arlington, VA: National Victim Center.

Koss, M. P. 1992. The underdetection of rape: Methodological choices influence incidence estimates. *Journal of Social Issues* 48:61-75.

Koss, M. P. 1996. The measurement of rape victimization in crime surveys. *Criminal Justice and Behavior* 23:55-69.

Link, B. G., E. Susser, A. Steuve, J. Phelan, R. E. Moore, and E. Streuning. 1994. Lifetime and five-year prevalence of homelessness in the United States. *American Journal of Public Health* 84:1907-12.

Reiss, A. J. Jr., and J. A. Roth, eds. 1993. *Understanding and preventing violence*. Washington, DC: National Academy Press.

Rosenberg, P. S., R. Biggar, J. J. Geodert, and M. H. Gail. 1991. Backcalculation of the number with human immunodeficiency virus infection in the United States. *American Journal of Epidemiology* 133:276-85.

Rossi, P. H., J. Wright, and G. Fisher. 1987. The urban homeless: Estimating composition and size. *Science* 235:1336-41.

Service, S. K., and S. M. Blower. 1995. HIV transmission in sexual networks: An empirical analysis. *Proceedings of the Royal Society, Series B* 260:237-44.

Shelley, G. A., H. R. Bernard, and P. D. Killworth. 1990. Information flow in social networks. *Journal of Quantitative Anthropology* 2:201-25.

Shelley, G. A., H. R. Bernard, P. D. Killworth, E. C. Johnsen, and C. McCarty. 1995. Who knows your HIV status? What HIV+ patients and their network members know about each other. *Social Networks* 17:189-217.

Society. 1994. Social science and the citizen. *Society* 33 (1): 2-3.

Wright, J., and J. A. Devine. 1992. Counting the homeless: The Census Bureau's "S-Night" in five cities. *Evaluation Review* 16:355-64.

Peter D. Killworth, individual merit scientist 3 (1), is a physical oceanographer at the Southampton Oceanography Centre, UK. His main research is on ocean process modeling, including the behavior of planetary waves, bottom flows, and the parameterization of baroclinic ocean eddies. He has also worked for 25 years on collaborative work in social networks, examining the rules that tie people together, with a recent emphasis on counting the uncountable.

Christopher McCarty is director of the Survey Program at the Bureau of Economic and Business Research at the University of Florida. McCarty received his Ph.D. in anthropology in 1992 and has been actively involved in studies relating to social networks, demography, and most recently democracy and governance. He is involved in developing methods to elicit samples of total personal networks that can be adapted to specific research areas, such as HIV/AIDS.

H. Russell Bernard is professor of anthropology at the University of Florida. Recently, he has been working with indigenous people to develop publishing outlets for works in previously nonwritten languages. He also does research on social networks, focusing recently on the development of a network model for estimating the size of uncountable populations.

Gene Ann Shelley is a behavioral scientist with the Division of Violence Prevention, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention. Her interests include using social network variables to identify hard-to-count or stigmatized populations and also the mechanism of how social networks may function to inhibit or encourage violent behavior, such as rape or partner violence.

Eugene C. Johnsen, professor of mathematics emeritus, is an applied research mathematician and director of Summer Sessions at the University of California, Santa Barbara. His research efforts lie at the intersection of mathematics and substantive social science and include individual work on developing and evaluating models of social structures and processes related to the formation of solidarity, collaborative work with Noah Friedkin on models for social influence in groups, and collaborative work with the present authors on estimating the size of personal networks and hard-to-count populations.