Sociological Methods
& Research
Volume 35 Number 1
August 2006 84-112
© 2006 Sage Publications
10.1177/0049124106289160
http://smr.sagepub.com
hosted at
http://online.sagepub.com

Investigating the Variation of Personal Network Size Under Unknown Error Conditions

Peter D. Killworth
National Oceanography Centre, Southampton, UK
Christopher McCarty
University of Florida, Gainesville
Eugene C. Johnsen
University of California, Santa Barbara
H. Russell Bernard
University of Florida, Gainesville
Gene A. Shelley
Georgia State University, Atlanta

This article estimates the variation in personal network size, using respondent data containing two systematic sources of error. The data are the proportion of respondents who, on average, claim to know zero, one, and two people in various subpopulations, such as "people who are widows under the age of 65" or "people who are diabetics." The two kinds of error—transmission error (respondents are unaware that someone in their network is in a subpopulation) and barrier error (something causes a respondent to know more or less than would be expected, in a subpopulation)—are hard to quantify. The authors show how to estimate the shape of the probability density function (pdf) of the number of people known to a random individual by assuming that respondents give what they assume to be accurate responses based on incorrect knowledge. It is then possible to estimate the relative effective sizes of subpopulations and produce an internally consistent theory. These effective sizes permit an evaluation of the shape of the pdf, which, remarkably, agrees with earlier estimates.

Keywords: *social networks; errors; probability density function*

Authors' Note: The Bureau of Economic and Business Research at the University of Florida provided the telephone survey services. The work was funded through National Science Foundation (NSF) grant SBR-9710353. Anonymous referees much improved the article.

The natural sciences have made steady progress in the acquisition of knowledge about observed phenomena and theories that explain these phenomena in such a way that they can be tested. Social scientists have yet to agree on the fundamental building blocks of their science, but the idea of a social physics is a venerable part of social science. The first edition of Adolphe Quételet's book, *Sur l'homme et le developpement de ses facultés*, in 1835, carried the audacious subtitle, *essai d'une physique sociale*. The idea of a social physics continues to have appeal. At the beginning of the twentieth century, Georg Simmel began a program of research based on the analysis of triads (relations between three people), which he saw as the fundamental building blocks of society. Indeed, later work (Cartwright and Harary 1956; Davis 1967, 1970; Davis and Leinhardt 1972; Holland and Leinhardt 1970, 1971; Johnsen 1985, 1986, 1989) shows that the specification of microstructure in terms of triads can lead to specific forms of macrostructure, and conversely.

In the 1950s, George Homans laid out rules for the analysis of social behavior, suggesting, as his contemporary B. F. Skinner did, that we focus not on internal mental states to explain behavior but on the context and outcomes. During that time, he also took aim at structural-functionalism, asserting that social phenomena are produced by individuals, not social systems, and analyzing processes in which individuals make exchanges with each other, presenting his version of social exchange based on operant psychology (Homans 1961; see also Thibaut and Kelley 1959). Peter Blau (1964) pursued this further, looking at the patterning of social exchange as a basis for the study of social phenomena. Through the later work of Emerson, Cook, Willer, and many others (see the authors and coauthors in the referenced works by Cook and Willer), social exchange has grown and branched out to became theoretically and mathematically developed and to have its principles and hypotheses empirically tested in field and laboratory settings (Emerson 1972a, 1972b; Cook 1987; Cook, Molm, and Yamagishi 1993; Willer and Anderson 1981; Willer 1999; Willer et al. 2002).

Other research, including our own, also takes a more microstructural view, after Radcliffe-Brown (1957), seeing people as nodes connected to other nodes in ways that lend themselves to measurement and prediction. In other words, by understanding the fundamental properties that connect humans, we believe we should be able to predict social phenomena and, it is hoped, solve problems that arise from those connections. In our own early research, we addressed such fundamental issues as the accuracy of respondents' reports about these connections (Bernard et al. 1984). We felt this to be important since inaccurate data are likely to generate inaccurate theory.

There remains a staggering lack of reliable data on the connections between ordinary humans since monitoring of human interactions (or, indeed, self-monitoring) is a time-consuming and expensive business. Experience sampling (Csikszentmihalyi and Larson 1987) holds promise, but it has not been used to collect social network data. Thus, social scientists are in the unusual position of still not knowing some basic quantitative facts concerning the linkages between people. This lack of knowledge is responsible for the continuing interest in the "small-world" phenomenon. This is formally the appearance of coincidence when two seemingly unrelated social connections come together and the mechanisms whereby this can occur. In many studies, it involves the passage of information along a chain of acquaintances from a collection of starters (randomly selected individuals) and a specified target person. This number is found empirically to be of order 5-6 (Milgram 1967; Travers and Milgram 1969), which is found by most people to be surprisingly small.

If we knew that any individual knew, say, exactly 3,000 people, we would hardly be surprised about this since our friends' friends' friends would, without overlap of personal networks, comfortably exceed the world population. Basic calculations on network size were made by de Sola Pool and Kochen (1978; also cf. Kochen 1989's review volume), although the problem dates back to 1936 (see Problem 38 in Mauldin 1981), and the problem itself gives rise to the Erdös number (the connectivity of joint-authorship scientists to Paul Erdös). Thus, our small-world surprise is due to the suspicion that we know, on average, significantly fewer than 3,000 people. The word *suspicion* is used advisedly: There is remarkably little research on the size of people's networks.

The majority of small-world research has had to proceed without direct knowledge of one key fact: the probability density function (pdf), or simply density, of the number c of people known by a random individual. This pdf is simply some function of c, which we shall denote by $\bar{P}(c)$. $\bar{P}(c)$ is just the probability that a random individual knows precisely c people (under some definition of *knowing*). Obtaining $\bar{P}(c)$ must be one of the grails of social network theory. As far as we are aware, there are few estimates of $\bar{P}(c)$ for ordinary individuals; those we have suggest an average of about 290 people known to a random individual (Killworth and Bernard 1978; Killworth, Johnsen, et al. 1998; Killworth, McCarty, et al. 1998) with a long-tailed distribution. Silverman (1986) discusses approaches to estimating pdfs from data.

Mathematical models of the small-world problem have perforce been based on random graph models in which only simple linkage structures have been assumed, partly, of course, for mathematical tractability (cf. the biased net theory of Skvoretz, Fararo, and Agneessens 2004). The models have almost exclusively omitted any estimates for the connectivity of the general public (cf. Watts 1999, 2003; Newman, Watts, and Strogatz 2002), although accurate data have been employed for specialized subpopulations (Newman 2002). A popular assumption is that of a scale-free network (cf. Watts 2004). It is clear that progress cannot be made on modeling the small-world problem in nonspecialized settings unless the pdf for c is known.

Empirical approaches themselves contain several difficulties. Informants attempting to pass a message, or information, toward a target person have to make a choice, based on their own imperfect data, as to who is the "best" person to pass on the information to. Since there are many things people do not know about their acquaintances, the resulting data are likely to themselves be highly imperfect. Indeed, White (1970) showed that the only reason the typical path in a small-world experiment is of length 5-6 is that people make mistakes in choosing the next intermediary in a small-world chain. To date, our best estimates of how choices are made remain empirical: Compare Killworth and Bernard (1978) and the extensive decision trees in Bernard, Killworth, and McCarty (1982), based on the information respondents requested about the targets in a hypothetical small-world experiment. Dodds, Muhamad, and Watts (2003) used the World Wide Web to carry out a large "small-world" experiment involving chains of acquaintances toward a target person. Unfortunately, the experiment had a completion rate of only 0.6 percent, compared with 22 percent in the original Travers and Milgram (1969) study, and so provided rather less information on how small-world choices are made than might have been hoped.

Thus, two key items emerge if we wish to study the linkages between individuals. The first is the pdf $\bar{P}(c)$; the second is how to estimate it given imperfect data, obtained from asking informants questions to which they themselves only possess imperfect knowledge.

This article, then, is specifically about estimating this pdf, in the presence of data that are, traditionally, noisy and full of human error.

Sources of Data Error

Methods to measure c and its distribution empirically have been developed through proxy approaches (Freeman and Thompson 1989; Killworth et al.

1990; Killworth, McCarty, et al. 1998; McCarty et al. 2001). These methods involve asking respondents how many people they know in certain categories or subpopulations. This instrument possesses (at least) two types of error, which we have termed *transmission errors* and *barrier effects*. A transmission error usually occurs when an individual is simply unaware that a member of his or her personal network lies within a subpopulation (e.g., is a diabetic), although false positives could also occur. A barrier effect occurs when the network position of a respondent relative to the network position of the subpopulation causes the respondent to know significantly more or fewer members of the subpopulation than would be expected.

The first of these errors (transmission) causes respondents to give incorrect answers about the membership of their network. The second of these errors (barrier) causes the responses about a subpopulation—although correct—to be uncharacteristic of the size of that subpopulation.

Although common sense tells us that barrier effects and transmission errors do exist, we are not certain of the extent of their impact for our empirical estimates of the size of subpopulations. The results of analyses of existing data from our various surveys briefly detailed here demonstrate the potential for these errors to confound many methods. It is worth, however, discussing the differences before we proceed.

First we consider barrier effects, which are a form of bias; some aspects of the barrier effect are dealt with in homophily studies and the "social distance" literature (e.g., McPherson, Smith-Lovin, and Cook's 2001 review), but many are not. For example, doctors are more likely to know doctors (homophily), but they are also more likely to know diabetics, which is not homophily. There are many possible barriers affecting exposure to certain subpopulations, with an obvious one being geographical, and others needing elucidation. We have begun an investigation of barrier effects by examining how respondents' reports of the number of people they know in subpopulations vary with properties of the respondents (though, as noted, there will also be a dependence on properties of the subpopulation under some circumstances). As an example, for a representative survey of 796 people in the United States, the mean number of Native Americans known to respondents in each of the 50 U.S. states and the fractional number of Native Americans in each state are highly correlated ($r = .58$, $p = .0001$). Thus, those living close to Native Americans know proportionally more than those living further away, with obvious effects on theories based on proxy information.

However, the difficulty this could present is mitigated if a survey is truly representative across the country or if many (and different) subpopulations are used. Surveys confined to small geographical areas may be subject to geographical barrier effects. For example, consider a survey limited to Florida (Killworth, Johnsen, et al. 1998), which predicted a much higher number for the HIV-positive subpopulation than the national survey reported here. This (positive) barrier effect occurred due to the larger proportion of HIV-positive people in Florida than nationally but disappeared when the national survey (Killworth, McCarty, et al. 1998) was undertaken.

If the distribution of the personal network size for a subpopulation differs significantly from that for the entire population, then a different form of barrier effect occurs. For example, if the subpopulation is AIDS victims, who are known to limit the size of their networks (Johnsen et al. 1995; Shelley et al. 1995), then the members of that subpopulation will be proportionally underreported compared with the actual size of the subpopulation. In our sample below, people who are imprisoned and, possibly, some other subpopulations may suffer from this effect. Conversely, subpopulations whose activities require them to know more people than normal (e.g., clergy, an example of which is given below) would be overreported by respondents.

It may be more difficult to remove barrier effects if they are less easy to anticipate. Analyses of variance across a combined national data set of almost 3,000 respondents showed significant differences in the number known in a subpopulation with almost any sociodemographic variable analyzed. This holds true even for apparently innocuous subpopulations such as "people named Michael." In other words, the results of proxy methods are subject to a variety of barrier effects; the simple assumption—that by using a representative sample, the effects can be overcome—may be erroneous.

Given that barrier effects are a function of the respondent and of the subpopulation, we can study their presence by looking for relationships between the attributes of the respondent and subpopulation and the respondent's report of the number known in the subpopulation. Transmission errors, on the other hand, are not necessarily a function of respondent characteristics but of the network members the respondent is trying to report about. Any empirical study of transmission errors must start with the members of the subpopulations.

We examined the reasons people do or do not tell people they know about their membership within four subpopulations: diabetics, Native Americans, twins, and widows younger than age 65. We chose these subpopulations because they were among the largest in our surveys and those we thought might be most susceptible to transmission error. For each of 24 subjects (6 from each of the four subpopulations), we elicited a network alter from each of five relation categories: family, nonsocial coworker, social

Table 1
Percentage of Respondents in a Given Population Who Told a Network Member From a Specified Category That They Are in the Population

	Population			
Network Member	Diabetic	Twin	Native American	Widow or Widower Younger Than Age 65
Family	70	0	13	83
Nonsocial coworker	78	71	37	20
Social coworker	75	86	37	40
Neighbor	60	57	37	17
Organization affiliation	67	86	13	17

coworker, neighbor, and organizational affiliation. We chose these relation categories for their potential variability in the transmission of membership information in the subpopulation.

Some of the results in Table 1 are due to the phrasing of the question. For example, we should not be surprised that none of our respondents who were twins told their family members about their status as members of these subpopulations—obviously, the family members knew already.[1] With the exception of family members, the table demonstrates wide variability across the rows and down the columns. In other words, respondents who are members of these subpopulations tell people about them based on whom they are talking to (family, coworker, neighbor, etc.), and the probability of telling varies by subpopulation (twins, diabetics, etc.). When further questioned about their reasons for not telling, respondents gave a variety of reasons. For example, diabetics sometimes did not tell because "it never came up." Others did not tell because they feared discrimination in the workplace.

We are aware that the true test is to examine whether the alters actually did know the information specified. However, under modern conditions, it is far from easy to obtain permission to contact alters, unlike, say, Laumann's (1969) study of the accuracy of respondents as to information known by their alters.

This study helped us verify that transmission error does exist; however, it provides us with no obvious way to correct it. This is due to the fact that transmission error probably varies by subpopulation in response to issues such as stigma, triviality, and so on. This error may be subpopulation specific or may depend only on classes of subpopulations (except that the correction should generally be positive).

Henceforth, we assume the existence of transmission error and barrier effects and subsume them under a single "error" in respondent reports; it is this error that we seek to mitigate.

Posing the Problem

We define the set of quantities P_n, where $n = 0, 1, 2, \ldots$. Here, $P_n(p)$ is the probability that a random respondent knows precisely n members of a subpopulation of fractional size p. We assume that $P_n(p)$ is independent of the respondent and that it depends solely on the size of the subpopulation. The quantities $P_n(p)$ are simple to compute—at least empirically—from survey data: One asks a large number of nationally representative respondents how many people they know in the population in each of several subpopulations (we use 29 below) of known fractional size within the total population and counts those reporting zero, one, two, and so forth in each subpopulation. We concentrate here on P_0, P_1, and P_2 for the most part (and P_0, P_1 for much of that) since there is distinct evidence of *heaping*, or reporting round numbers, when reported numbers get above 5 (Huttenlocher, Hedges, and Bradburn 1990; Baker 1992; Roberts and Brewer 2001). We are thus uncertain of the accuracy of reports of higher numbers while suspecting that reports of knowing precisely zero or one subpopulation member may be more accurately reported (while still possessing error).

Our model is simple (we are concerned here, after all, with the effects of error on its results). We assume that the probability that any member of a respondent's network is in a subpopulation of fractional size p is simply p. In other words, the only aspect of a particular subpopulation with any bearing on the probability of being known to an average respondent is the subpopulation *size*.[2] While the transmission and barrier effects discussed above will cause this assumption of independence to be in error—which we examine here—the model reduces to a simple binomial since all the choices are independent by our assumption.

If the pdf $\bar{P}(c)$ is known, P_n can be calculated as follows:

$$\begin{aligned} P_n &= \sum_c \bar{P}(c) \bullet \text{prob (someone knowing } c \text{ people knows} \\ &\qquad \text{precisely } n \text{ in the subpopulation)} \qquad (1) \\ &= \sum_c \bar{P}(c)\,{}_cC_n\, p^n (1-p)^{c-n}, \end{aligned}$$

since the probability of knowing precisely n is by assumption a binomial; here, ${}_aC_b$ is the number of combinations of b from a. As yet, we do not know the distribution $\bar{P}(c)$, but we can make deductions from the above without knowing $\bar{P}(c)$.

To begin with, consider $P_0(p)$, the probability of knowing nobody in a subpopulation. There are two fixed points. Apart from errors of commission (inventing a member of a subpopulation; cf. Killworth et al. 2003), a respondent cannot know anyone in a subpopulation of size zero. Thus, $P_0(0) = 1$. Similarly, if a subpopulation comprises the entire population, it is impossible to know zero members of it, so $P_0(1) = 0$. We also know that $P_1(0) = P_2(0) = \ldots = 0$ since one cannot know one, two, and so forth members of a subpopulation of size zero.

Consider

$$\frac{dP_n}{dp} = \sum_c \bar{P}(c)_c C_n \{np^{n-1}(1-p)^{c-n} - (c-n)p^n(1-p)^{c-n-1}\}.$$

Now

$$P_{n+1} = \sum_c \bar{P}(c)_c C_{n+1} p^{n+1}(1-p)^{c-n-1},$$

and

$${}_cC_{n+1} = {}_cC_n\left(\frac{c-n}{n+1}\right),$$

so that

$$\begin{aligned}\frac{dP_n}{dp} &= \frac{n}{p}\sum_c \bar{P}(c)_c C_n p^c(1-p)^{c-n} \\ &\quad - \left(\frac{n+1}{p}\right)\sum_c \bar{P}(c)_c C_{n+1} p^{n+1}(1-p)^{c-n-1} \\ &= \frac{n}{p}P_n - \left(\frac{n+1}{p}\right)P_{n+1}.\end{aligned}$$

Thus,

$$P_{n+1} = \frac{1}{(n+1)}\left\{nP_n - p\frac{dP_n}{dp}\right\}.$$

Note the special case

$$P_1 = -pdP_0/dp. \tag{2}$$

It is easy to show that

$$P_n(p) = \frac{(-1)^n}{n!} p^n \frac{d^n P_0}{dp^n},$$

so that knowledge of the distribution of the probability that nobody is known in a subpopulation (i.e., $P_0(p)$) means that the distribution of the probability of knowing any required number in a subpopulation is known, provided that $P_0(p)$ can be differentiated sufficiently accurately.

This is a serious requirement when actual data are confronted. We here use data from two surveys: a "standard" nationally representative survey of 796 respondents, as well as a survey of 131 clergy of different faiths, chosen on the (correct) assumption that clergy know more people than nonclergy know. In each survey, we asked respondents how many they knew in each of 29 subpopulations of known size and estimated $P_0(p)$, $P_1(p)$, and $P_2(p)$, where p is the known fractional size of each subpopulation. Details of these surveys and the subpopulations are given in McCarty et al. (2001).

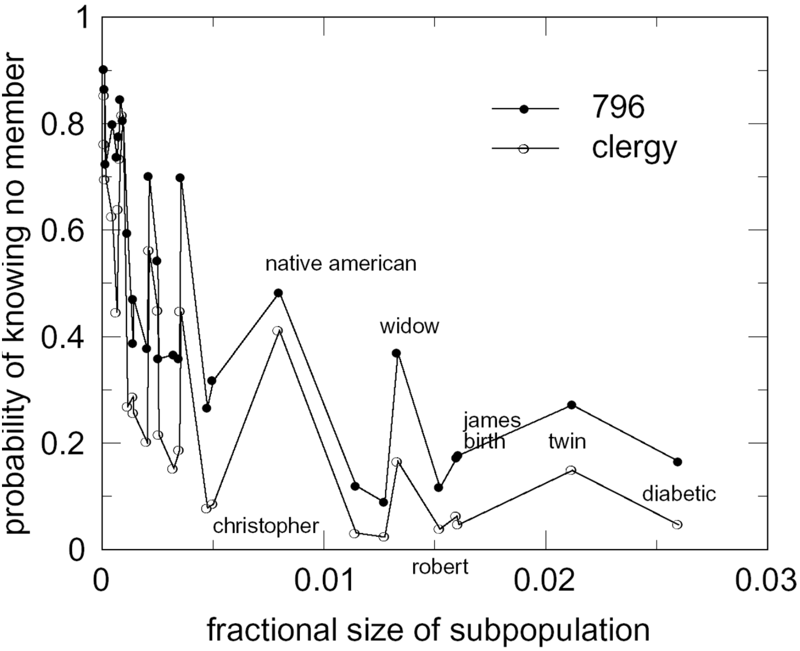
Figure 1 shows $P_0(p)$ for the two surveys. The two curves are well correlated, indicating that our surveys are replicable (McCarty et al. 2001). However, the two curves are hardly smooth. Computing their gradients to generate a potential $P_1(p)$ would be impossible. The same jaggedness extends to P_1, P_2 (Figure 2), although there are hints that this is lessened for higher numbers known (henceforth, data are only shown for the 796 survey respondents for clarity).

Some jaggedness would be expected from simple experimental error (in that we use a finite sample to estimate a probability). However, that random variability could not be so strongly correlated between samples and is far in excess of what would be expected from standard sampling theory. Whatever is happening is then *not* random but a signal of a process in respondents' minds. How to cope with this forms the body of this article.

Exploring a Possible Solution

Any approach has to accept that the existence of errors causes respondents to provide data consistent with a fractional population size something other than the actual fraction. Statistical fitting techniques exist and could be employed, but our purpose is to see if error effects can be included sensibly in calculations by a direct, if simple, model of the effect of the error

Figure 1
The Probability $P_0(p)$ of Knowing No Member of a Subpopulation of Fractional Size p



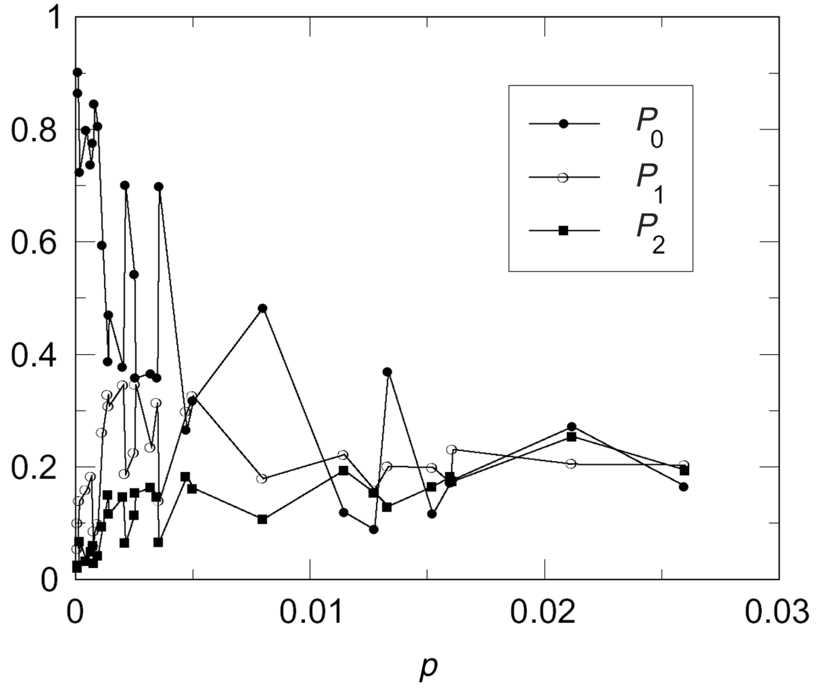
Note: Two sets of data are plotted: Full circles show the standard data used elsewhere in this article, a nationally representative survey of 796 respondents; open circles show the same for a survey of clergy of different religions. Some subpopulations are identified.

process. Two possible models suggest themselves: (a) assuming errors lie in the P_0 values and (b) assuming errors lie in the p values.

Errors Occur in P_0 Values

One can consider the possibility of adjusting each observed value of P_0 in Figure 1 up or down to make the P_0 curve smooth in some way. Initially, this appears attractive since it maintains the actual values of p (which are certainly the most accurately known of our data). Examination of either of the two curves suggests that there may be two (or more) distinct

Figure 2
The Probabilities $P_0(p), P_1(p)$, and $P_2(p)$ as Functions of the Fractional Size of the Subpopulation p



Note: The first curve is a copy of that in Figure 1.

sets of values in existence, depending on the type of subpopulation. For instance, it is clearly more difficult to know that a member of one's network is a widow than it is to know that a member is called Robert, although the size of both these subpopulations is similar.

To deal with this would indeed involve some ad hoc method to raise or lower one set of points against another to make the curve smooth. This seems to us to be unsatisfactory on several grounds. First, the amount that any point needs to be shifted depends on the type of subpopulation (i.e., on the degree of error) so that two points with almost identical p values might need to be adjusted by different amounts. Second, the effects on P_0, P_1, and so on would not be identical (since (1) shows that p enters these quantities in a nonlinear fashion) and so would be impossible to predict. We invested considerable time and effort in an attempt to develop weights to adjust for the transmission error we knew to exist in each of the populations that we asked respondents to estimate. These efforts included two studies: (a) the development of a decision model to explain to which members of a population the members of a target group tell their membership in the target group and (b) an ambitious "alter-chasing" study where we identified "starter" members of several populations, elicited a set of network alters, and called those alters to ask whether they knew about the starter's membership in the population. These studies demonstrated that any real attempt to estimate this variability empirically would be very difficult and expensive and could potentially introduce as much noise, via the weights, as we are attempting to remove.

We also sought some smooth functional form for $P_0(p)$ and performed fitting exercises. The scatter in Figure 2 demonstrates that almost any decreasing function of p would fit the data equally well, and this approach has also to be discarded.

Errors Occur in *p* Values

We therefore proceed with the second approach—namely, that the combined effect of transmission errors and barrier effects has simply misplaced each point on the p-axis. In other words, respondents react as if the above equations hold but with modified values of p, which reflects the fraction of their network that they are aware lies within a subpopulation. More formally, we assume that transmission errors and barrier effects conspire so that subpopulation i, of fractional size p_i, is actually reported as if it were of size p'_i, where

$$p'_i = \lambda_i p_i,$$

and λ_i is a factor accounting for transmission errors and barrier effects whose value may be more or less than unity depending on the form taken by the errors. Recall that this factor is independent of the respondent since we deal exclusively with averages across all respondents. A value of $\lambda_i < 1$ might indicate transmission error or barrier effects, making it hard to know members of a subpopulation (e.g., those who limit their networks). Values of $\lambda_i > 1$ might involve subpopulations whose networks are larger than usual (e.g., clergy or members of some political groups).

In effect, the x-abscissas of Figures 1 and 2 are misplotted (and may be in the wrong order) because they should be plotted using the (unknown)

p_i' as values, so our data essentially refer to $P_0(p')$, not $P_0(p)$. Put another way, we assume that respondents are giving data based on a subset of a subset of their network: the part of the subset of all those in a subpopulation of which the respondent is aware (for transmission error) or can be aware (for barrier effect). For "visible" subpopulations—those with small errors—this will be the majority of the subpopulation that is known. For less visible subpopulations—those with large errors—it will be only some fraction of the subpopulation. (This, so far somewhat arbitrary, process will be both quantified and validated later.)

Evidence for this belief is found by reordering the data in Figure 2 in descending order of P_0. This satisfies the requirement from (2) that P_0 should decrease with p. This order of subpopulations is unique. Replotting, with a completely arbitrary x value representing the position in the descending order, gives Figure 3, where five passes of a smoothing operator were carried out between adjacent points.[3]

We now regard the modified (or "effective") fractional subpopulation sizes p_i' as the unknown values and solve for them by requiring (2) to hold.[4] We rewrite (2) as

$$P_1 + p dP_0/dp = 0.$$

To evaluate this numerically, we first define $Q_n(j)$ to be the jth value of P_n in the ordered list: In other words, $Q_n(j)$ are the data points in Figure 3. Then (2) can be evaluated as

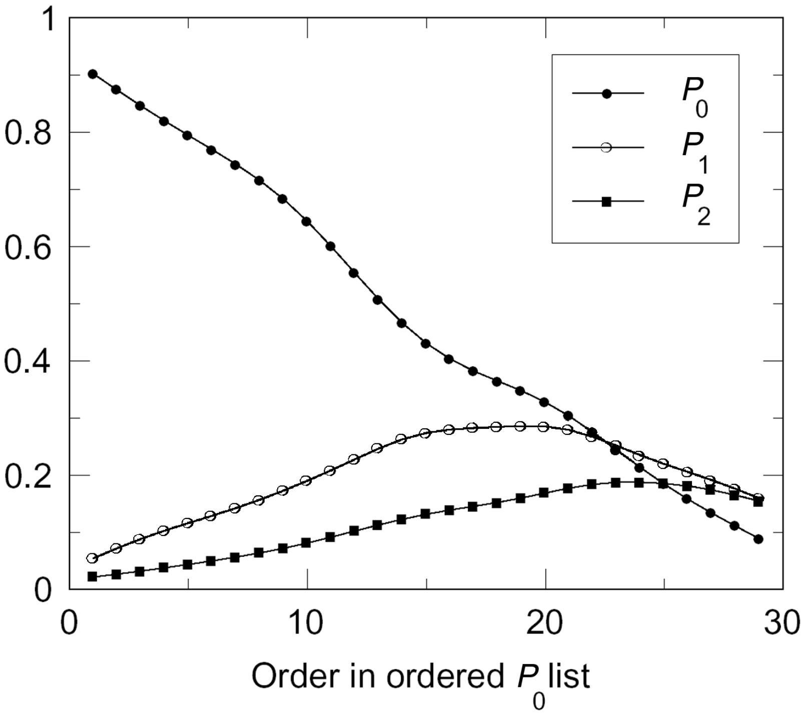
$$Q_1(j)p_j'\left[\frac{Q_0(j+1) - Q_0(j-1)}{p_{j+1}' - p_{j-1}'}\right] = 0, j = 1, 2, \ldots, N,$$

where the derivative dP_0/dp has been evaluated as a simple centered difference of the P_0 values. Multiplying by $p_{j+1}' - p_{j-1}'$ gives a matrix problem for the p_j',

$$-Q_1(j) + p_{j-1}' + \{Q_0(j+1) - Q_0(j-1)\}p_j' \\ + Q_1(j)p_{j+1}' = 0, j = 1, 2, \ldots, N.$$

These are N equations for the N unknowns $p_1', p_2', \ldots, p_N'$. However, some care is needed at the endpoints of the range where values for $j = 0$ and $N + 1$ are required. Now the actual p_j values reach as low as 4×10^{-4}, which is almost certainly close enough to zero that we can safely add an extra "known" value $j = 0$ to the list, where $p_0' = 0$, $Q_0(0) = 1$, without any loss of accuracy. At the high end of the probabilities, however,

Figure 3
The Data in Figure 2 but Ordered in Decreasing Order of $P_0(p)$, With an Arbitrary x-Axis



Note: The smoothed version shown is used for the following calculations.

the actual p_j values never exceed 0.026. Thus, we cannot add the obvious $p'_{N+1} = 1$, $Q_0(N + 1) = 0$ (although this would close the problem numerically) since because 0.026 is not close to unity, this would involve evaluation of dP_0/dp using values of p' over the wide range (0.026 and 1), which is so widely spaced that the centered-difference approximation would yield completely inaccurate results. (Reposing the problem using $\ln(p)$ as the unknown does not alleviate the problem.) Indeed, it can be shown that since $P_n(1)$ will be expected to be small for small n, P_0 must tend to zero with a very flat behavior as 1 is neared, confirming that a centered difference is ineffective.

It is of course unrealistic to seek P_0 values for p near unity; even to estimate the vanishingly small fraction of respondents knowing zero or one males, for example, is impossible. Instead, we choose here to use a one-sided difference for the gradient at the right-most point—namely,

$$Q_1(N) + p'_N \left[\frac{Q_0(N) - Q_0(N-1)}{p'_N - p'_{N-1}} \right] = 0, \text{or}$$
$$Q_1(N)\{p'_N - p'_{N-1}\} + \{Q_0(N) - Q_0(N-1)\} p'_N = 0.$$

This is a much more accurate representation of (2) for the last value of j, but it suffers from the disadvantage that the system is now homogeneous in the p'_j. In other words, if there is a solution p'_j, then Ap'_j is another solution for any constant A, so that the solution is unique only up to an arbitrary multiplicative scaling.

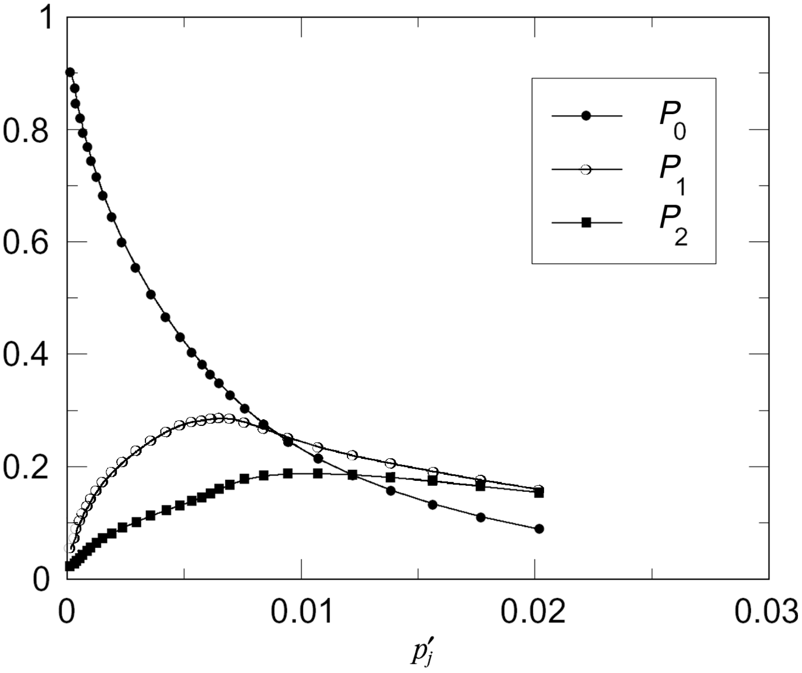
We have experimented with other approaches but found no way to avoid this numerical difficulty (a completely different approach is sketched in the appendix, but this also proves hard to set absolute values on the p'_j). We choose to close the problem (which is a tridiagonal system and so trivial to invert) by choosing the scaling to make the sum of squared differences

$$\sum_{j=1}^{N} (p'_j - p_j)^2$$

a minimum (i.e., we select a solution that is closest in Euclidean distance to the original values). By this scaling, we expect very "visible" subpopulations (e.g., those who died in a car wreck in the past 12 months) to show an increased p'_j compared with p_j and less visible subpopulations to show a decrease; recall that barrier effects can have either sign. Changing this scaling merely rescales everything, mutatis mutandis. Note that the solution set $\{p'_j\}$ would be found—subject only to the final scaling chosen—for *any* set $\{p'_j\}$ since the values are determined solely by the values of the $Q_0(j)$ and $Q_1(j)$. This is logical since we are arguing here that the values of P_0 and P_1 determine the p'_j values uniquely up to the scaling factor.

Solving for the p'_j gives the distributions of P_0, P_1, and P_2, shown in Figure 4. The values and orders of the P_n are unchanged, but now they have an x-abscissa, which is the predicted probabilities. The largest probability, by this choice of scaling, is somewhat smaller (just over 0.02), corresponding to the subpopulation with the smallest P_0 value. A glance

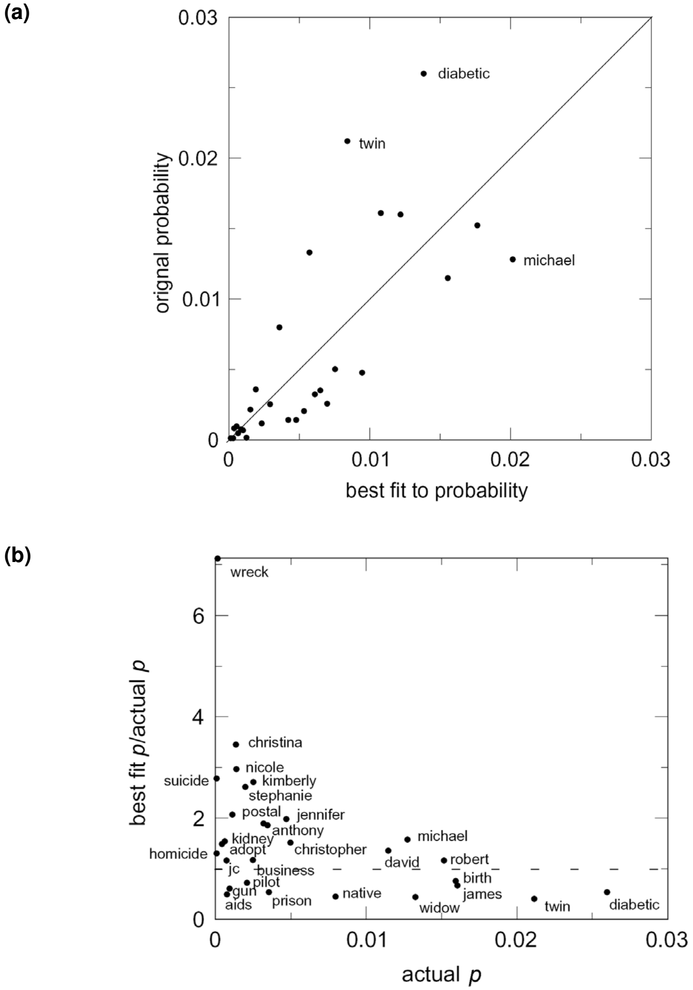
Figure 4
$P_0(p), P_1(p)$ and $P_2(p)$ (Smoothed), Plotted as Functions of the Computed p' Rather Than the Original p



at the right-hand end of Figure 4 will show that evaluation of the slope of P_0 there by extending the curves to $p' = 1, P_0 = 0$ would be grossly in error, validating our choice above of the one-sided derivative.

Figure 5a shows how the p'_j compare with the original p_j values. The root mean square (RMS) difference between the p'_j and the p_j is 0.005. The correlation between the two sets is excellent (0.75), although the relative scatter among small p values cannot be shown clearly by such a plot. Instead, Figure 5b plots the ratios λ_j against p_j, with the points labeled with the subpopulation name. The results are often, but not always, intuitive. The subpopulations that are "visible" include car wreck victims (by this scaling, an increase in p'_j/p_j of over 7), many female names, and suicide victims. Many male names, homicide victims, and postal workers

Figure 5
(a) Scatterplot of the Best Fits to Subpopulation Fractional Size p'_j Compared With the Original Values p_j and (b) Estimates of the Amount by Which p_j Must Be Scaled Down to Obtain p'_j (i.e., the Ratio p'_j/p_j) as Functions of the Actual p_j



Note: In (a), a few subpopulations are labeled. In (b), the points are labeled showing which subpopulation is referred to. A dotted line is added at a ratio of 1.

occupy a middle range of ratio values. Those we have identified as hard to know (i.e., transmission errors) such as recent incarcerees, recent widows, twins and diabetics, and those for which there are barrier effects (e.g., those with AIDS) all have very small ratios.

(The difference in behavior between male and female names is interesting. The ratios p'_j/p_j are significantly smaller for male names than for female names. The size of even the most popular female name subpopulation is small compared with popular male names, which may make female names more "visible" : Respondents may be more aware that they know someone with a fairly rare name. However, since we are dealing almost exclusively with the chance of knowing zero members or one member of a subpopulation, it is hard to see why this would mitigate against popular male names since while one may be uncertain how many Michaels one knows, one can be sure if one knows at least one.)

A test of whether the answer is reliable is to predict the values of P_2 using the finite-difference form

$$Q_2^{\text{predicted}}(j) = \frac{1}{2}\left\{ Q_1(j) - p'_j \frac{[Q_1(j+1) - Q_1(j-1)]}{p'_{j+1} - p'_{j-1}} \right\},$$

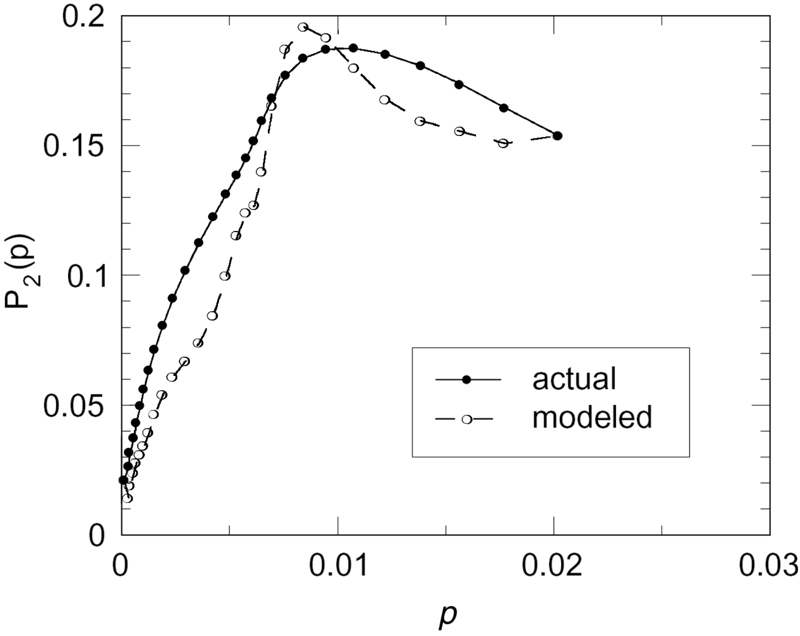
where the actual values are substituted at the endpoints $j = 1, N$. The result is shown in Figure 6. The degree of agreement is reasonable (given our expectation of larger inaccuracy in reports of size 2 and above), with a good correlation between the actual and predicted values, although the predictions are mostly slightly lower than the actual values.

Solving for the Distribution of $\bar{P}$(c)

We argued at the start of this article that researchers really needed the distribution, or pdf, $\bar{P}(c)$. We argued in the last section that we have good approximations to $P_0(p)$, $P_1(p)$, and so on. To complete the solution, the problem must be inverted for $\bar{P}(c)$. Because there are only 29 known values of $P_0(p)$ (and we cannot use P_1, etc. since they are automatically consistent and add no information), it would appear that either we cannot obtain more than 29 individual values of $\bar{P}(c)$ or that the problem is underdetermined. (One could, for example, add in a smoothness criterion and find a least squares solution.)

While there are inversion methods for various binomial sums (Riordan 1968), they are not well suited for the underdetermined nature of the problem.

Figure 6
Actual and Predicted (From the Theory, Equation (3)) Values of P_2, as Functions of p



However, we can replace the binomial with a Poisson distribution. This is because all the probabilities considered here are small; for large probabilities, P_0 is vanishingly small, and so the error remains tiny. Indeed, the algebraic results above hold identically for the Poisson distribution.

Then (1), for $n = 0$, becomes

$$P_o(p) = \sum_c \bar{P}(c) \cdot \exp(-pc).$$

Approximating the sum by an integral, we replace this by

$$P_o(p) \approx \int_o^\infty \bar{P}(c) \exp(-pc) dc,$$

so that $P_0(p)$ is merely the Laplace transform of $\bar{P}(c)$. (In practice, $\bar{P}(c)$ is normally summed over a bin of size about 20; for the purposes of inversion, this is ignored.) A fair approximation to $P_0(p)$ is, from a least squares fit (a fit minimizing absolute errors is similar),

$$P_o(p) \approx \left(\frac{\alpha}{p+\alpha}\right)^v,$$

where the power $v = 1.1794$ and $\alpha = 4.1866 \times 10^{-3}$. An overall less accurate fit, although with a much better fit to the decay at larger p values, is given by $v = 1.5, \alpha = 4.25 \times 10^{-3}$; this has relevance below. Other functional forms—for example, $ab/[(p+a)(p+b)]$—give indistinguishable answers. The resulting inversion for $\bar{P}(c)$ is

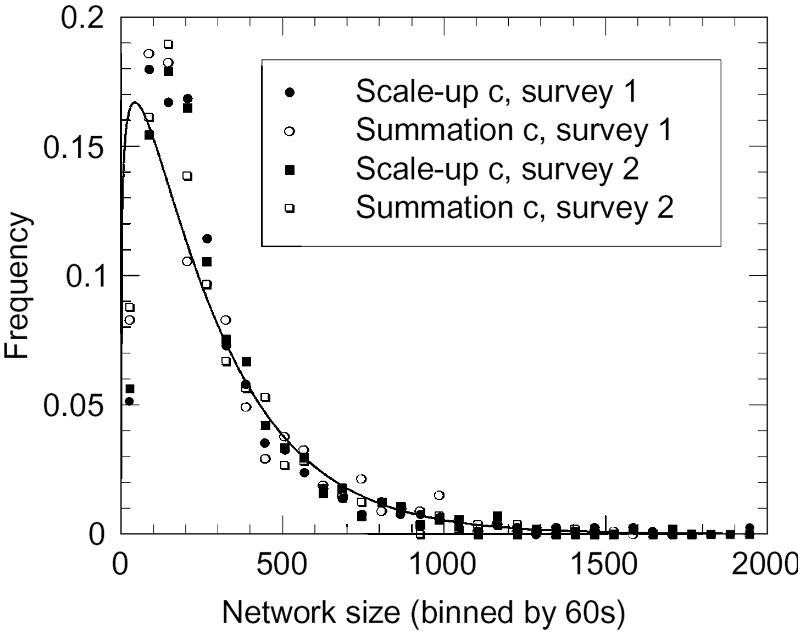
$$\frac{\alpha^v}{\Gamma(v)} c^{v-1} \exp(-ac),$$

where Γ represents the gamma function; this is indeed a probability density function.

The inversion is shown as the firm curve in Figure 7 and is long-tailed. The solution is a little sensitive to the power v chosen, in that the position of the modal value of c can be changed by up to a factor of 2. However, the best fit strongly resembles the discrete solutions found by other methods also shown in Figure 7 (Killworth, McCarty, et al. 1998; McCarty et al. 2001). The c value for the maximum (i.e., modal) value is given by $(v-1)/\alpha$, here 43, perhaps somewhat smaller than the value of about 100 indicated in the comparison data, although the broad binning (to smooth the empirical data) bands make the comparison less straightforward. The peak value of the (binned) probability is 0.167, again slightly smaller than the comparison data would indicate.

However, in all other respects, the agreement is excellent. The mean value of c across the population is $v/\alpha = 282$, with a standard deviation of $v^{1/2}/\alpha = 259$, which are extremely close to the estimated values of 290 and 260, respectively, in the best available estimates we are aware of (McCarty et al. 2001). We are unable to judge the agreement more quantitatively, as the earlier estimates are empirical from a few data sets (although consistent between them), and the new distribution here is obtained by inverting a set of values that possess unknown errors. We also lack other independent estimates of $\bar{P}(c)$ with which to make a comparison, and there is considerable overlap in the data sets used in this and prior estimates.

Figure 7
Empirical Distribution of $\bar{P}(c)$ From Two Previous Surveys and Two Different Approaches, Scale-Up and Summation (cf. McCarty et al. 2001), Shown by Symbols, Together With the Inverted Distribution $\bar{P}(c)$ Found From $P_0(p)$, Binned Into 60s, Shown by the Firm Curve



Nonetheless, this agreement is most encouraging (although we stress the sensitivity to the precise shape of the fit to P_0 and hence to the structure of P_0 itself). Although there is no connection between the scale-up methodology approach to $\bar{P}(c)$ and the approach used here, they agree in the form, shape, and size of $\bar{P}(c)$ and its first two moments.

Conclusions

This article has examined how to obtain an estimate of the probability density function of the number of people known, *c*, from proxy data that possess various types of systematic (as well as random) error. We have argued that respondents may be giving accurate responses based on incorrect knowledge (due to transmission errors and barrier effects) and that it is possible to estimate the relative effective proportional sizes of subpopulations so as to produce an internally consistent theory. Given these effective sizes, it is possible to invert the problem and obtain the shape of the pdf of c, which, remarkably, agrees better than qualitatively with earlier estimates as well as qualitatively with estimates based on recall of names in telephone books (Freeman and Thompson 1989; Killworth et al. 1990). Given that two entirely different methodologies were used to obtain estimates of $\bar{P}(c)$, this level of agreement is most encouraging.

The approach we have used is not one of statistical fitting, which would seek a signal among the P_0 values in the presence of random noise. Here the level of random noise—while clearly present—does not cause the difficulties, as the remarkable level of replicability between the two data sets in Figure 1 demonstrates. Instead, we have sought to model the process whereby incorrect knowledge can be used "as is" in a description of personal network size.

While the methodology we have outlined should work with other data sets involving other subpopulations, a difficulty remains in that no clue has yet appeared as to how one might prejudge the degree of over- or underestimation λ_j for any new subpopulation. Such information would be vital in teasing out subpopulations that are "hidden" but not obviously so. For example, we know from much work that AIDS victims limit their networks. But are there less stigmatizing illnesses whose sufferers also limit their networks, of which researchers are unaware (and so use of social network counting methods might underestimate the size of the subpopulation)? In other words, while the "effective" subpopulation size is a convenient construct for our present purposes, it would be useful if one could deduce the degree of misestimation in data for use in clarifying further issues in social network theory. Policy makers in particular need to know the actual sizes of important populations, not the effective sizes. To move from effective to actual requires a theory both to explain why the ratios of effective to actual are what they are and to predict what the ratios should be for subpopulations whose size is initially unknown. Such a theory remains lacking.

The best fit for $\bar{P}(c)$, as well as our earlier empirical estimates (Killworth, Johnsen, et al. 1998; Killworth, McCarty, et al. 1998), is clearly fundamental for modeling in the small-world literature. It is of interest that its shape does not fall into any previously modeled category (it is not, for example, scale free), and it would be interesting to see whether theoretical predictions based on the "power law times exponential" pdf for c would fit observed path distributions in small-world experiments.

Finally, although this article is devoted to the refinement of estimates of c and a description of the pdf for c, this work fits into a larger research agenda. Personal network size involves several parameters that would comprise a social physics. Social scientists must agree on a set of fundamental parameters on which they can begin to build their science. These parameters can of course be refined and, in some cases, completely changed. We may ultimately determine that personal network size is not as basic a construct as we currently think it is. Until we can agree on the methods for establishing and measuring such parameters, the study of society will remain a moving target, rather than a science where one finding is built upon another.

Appendix
An Alternative Approach

The approach in the main article suffered from the difficulty that no absolute values for the revised probabilities could be set because of inaccuracies if one attempted to evaluate derivatives beyond the smallest observed P_0. Here we explore another approach, although this will also suffer from estimation difficulties.

We rewrite (2) as

$$\frac{dp}{dP_o} = -\frac{p}{P_1}, \tag{A1}$$

and assuming we know how P_1 varies with P_0, we can regard this as a differential equation to determine p as a function of P_0. Its solution is

$$p = \exp\left(-\int_o^{P_o} \frac{dP_o}{P_1} \right), \tag{A2}$$

where we use the fact that when $p = 1, P_0 = 0$. The behavior of (A2) is dominated by places where P_1 is small. These are at $P_0 = 1$, by definition, and at $P_0 = 0$, where the size of P_1 is expected to be small. In the former area, the manner in which P_1 tends to zero is important. It must do so in a way that makes the integral in (A2) infinite at $P_0 = 1$, so that $p = 0$ there.

But P_1 becoming zero linearly or superlinearly with P_0 gives very different solutions for p due to the exponential in (A2). Similarly, the behavior of P_1 near $P_0 = 0$ is crucial since if it tends to some small nonzero value, this value will dominate the way in which p tends to zero, again due to the exponential in (A2). It would appear to be difficult to acquire data for such values of P_0.

If we ignore these (nontrivial) difficulties, then for suitably chosen distributions $P_1(P_0)$, which can be guessed at by examining how P_1 and P_0 covary, (A2) can be inverted and the problem solved. In practice, it is usually simpler to examine the structure of $P_0(p)$ from curves such as Figure 4, add disposable parameters, compute $P_1(p)$ using (2), and eliminate p between P_0 and P_1 to obtain the $P_1 - P_0$ relationship directly.

One such family of relationships, which resembles the shape of Figure 4, is

$$P_o(p) = \left(\frac{1 - p^\alpha}{1 + p^\alpha}\right)^m. \tag{A3}$$

$$P_1(p) = \frac{\alpha m}{2} P_o^{(m-1)/m}\left(1 - P_o^{2/m}\right) \equiv 2\alpha m p^\alpha \frac{\left(1 - p^\alpha\right)^{m-1}}{\left(1 - p^\alpha\right)^{m+1}}. \tag{A4}$$

(Here, $a > 0, m > 0$.) Plots of (A4) are strongly independent of the parameter m once it becomes large (which is necessary to reduce values at $p = 1$). P_1 reaches a maximum as a function of P_0 at

$$P_o = \left(\frac{m-1}{m+1}\right)^{m/2},$$

which is a very weak function of m (changing from 0.33 for $m = 2$ to 0.36 for $m = 40$) and is well approximated by $1/e$ for m above 2. This value fits the P_1 maximum observed rather well. To obtain the correct magnitude for the curve requires $\alpha \approx 0.77$, again essentially independent of m. Thus, we possess a wide range of solutions, by varying m, all of which give a good fit (not shown) to the observed $P_1 - P_0$ distribution.

However, this wide choice of m again has very strong effects on P_0 as a function of p. It is clear from (A3) that P_0 decays increasingly rapidly as m increases. Thus, from the $P_1 - P_0$ relationship alone, we cannot with sufficient accuracy distinguish between the wide range of possible m values, yet the value of m determines the effective scaling for the revised probabilities. We have tried other functional relationships, with the same inability to distinguish between wide parameter changes, and suspect that other approaches (e.g., spline fitting with specified endpoints and again a smoothness requirement) might be more productive.

Notes

1. Family includes nonconsanguineal relatives, so that the figures for Native Americans and widow(er)s and family make sense.

2. The reader may be concerned that properties of the respondent could also affect the probability here, particularly the concept of "it takes one to know one," if the respondent is a member of the subpopulation. Recall, however, that the probability we define is averaged over the entire population of respondents and so is only permitted to depend on the subpopulation itself.

3. This smoothing took the form of replacing $Q_n(j)$ by $0.25Q_n(j-1)+0.5Q_n(j)+0.25Q_n(j+1)$, which conserves the mean but reduces the variance (and hence the noise) in the fields. Smoothing is *not* essential to what follows; indeed, the entire calculation has been carried out on the raw data. There are some small sign reversals, but qualitatively the same picture emerges.

4. The (continuous) problem is formally underdetermined, and it is possible to imagine a solution for P_0 that possesses oscillatory second derivatives and fits the data perfectly while being completely unrealistic. In such cases, the problem would be closed by the addition of some smoothness criterion. This has some degree of arbitrariness, and we prefer the approach here, wherein the centered-difference formulas themselves imply smoothness.

References

Baker, Michael. 1992. "Digit Preference in CPS Unemployment Data." *Economics Letters* 39:117-21.

Bernard, H. Russell, Peter D. Killworth, David Kronenfeld, and Lee D. Sailer. 1984. "The Problem of Informant Accuracy: The Validity of Retrospective Data." *Annual Review of Anthropology* 13:495-517.

Bernard, H. Russell, Peter D. Killworth, and Christopher McCarty. 1982. "INDEX: An Informant-Defined Experiment in Social Structure." *Social Forces* 61:99-133.

Blau, Peter M. 1964. *Exchange and Power in Social Life*. New York: John Wiley.

Cartwright, Dorwin and Frank Harary. 1956. "Structural Balance: A Generalization of Heider's Theory." *Psychological Review* 63: 277-93.

Cook, Karen S., ed. 1987. *Social Exchange Theory*. Newbury Park, CA: Sage.

Cook, Karen S., Linda D. Molm, and Toshio Yamagishi. 1993. "Exchange Relations and Exchange Networks: Recent Developments in Social Exchange Theory." Pp. 296-322 in *Theoretical Research Programs*, edited by Joseph Berger and Morris Zelditch Jr. Stanford, CA: Stanford University Press.

Csikszentmihalyi, Mihali and Reed W. Larson. 1987. "Validity and Reliability of the Experience-Sampling Method." *Journal of Nervous and Mental Disease* 175:526-36.

Davis, James A. 1967. "Clustering and Structural Balance in Graphs." *Human Relations* 20:181-7.

———. 1970. "Clustering and Hierarchy in Interpersonal Relations: Testing Two Graph Theoretical Models on 742 Sociomatrices." *American Sociological Review* 35:843-51.

Davis, James A. and Samuel Leinhardt. 1972. "The Structure of Interpersonal Relations in Small Groups." Pp. 218-51 in *Sociological Theories in Progress*, vol. 2, edited by Joseph Berger, B. Anderson, and M. Zelditch Jr. Boston: Houghton Mifflin.

de Sola Pool, Ithiel and Manfred Kochen. 1978. "Contacts and Influence." *Social Networks* 1:1-51.

Dodds, Peter S., Roby Muhamad, and Duncan J. Watts. 2003. "An Experimental Study of Search in Global Social Networks." *Science* 301:827-9.

Emerson, Richard M. 1972a. "Exchange Theory, Part I: A Psychological Basis for Social Exchange." Pp. 38-57 in *Sociological Theories in Progress*, vol. 2, edited by Joseph Berger, B. Anderson, and M. Zelditch Jr. Boston: Houghton Mifflin.

———. 1972b. "Exchange Theory, Part II: Exchange Relations and Network Structures." Pp. 58-87 in *Sociological Theories in Progress*, vol. 2, edited by Joseph Berger, B. Anderson, and M. Zelditch Jr. Boston: Houghton Mifflin.

Freeman, Linton C. and Claire R. Thompson. 1989. "Estimating Acquaintanceship Volume." Pp. 147-58 in *The Small World*, edited by Manfred Kochen. Norwood, NJ: Ablex.

Holland, Paul W. and Samuel Leinhardt. 1970. "A Method for Detecting Structure in Sociometric Data." *American Journal of Sociology* 70:492-513.

———. 1971. "Transitivity in Structural Models of Small Groups." *Comparative Group Studies* 2:107-24.

Homans, George C. 1961. *Social Behavior: Its Elementary Forms*. New York: Harcourt, Brace and World.

Huttenlocher, Janellen, Larry V. Hedges, and Norman M. Bradburn. 1990. "Reports of Elapsed Time: Bounding and Rounding Processes in Estimation." *Journal of Experimental Psychology, Learning, Memory and Cognition* 16:196-213.

Johnsen, Eugene C. 1985. "Network Macrostructure Models for the Davis-Leinhardt Set of Empirical Sociomatrices." *Social Networks* 7:203-24.

———. 1986. "Structure and Process: Agreement Models for Friendship Formation." *Social Networks* 8:257-306.

———. 1989. "The Micro-Macro Connection: Exact Structure and Process." Pp. 169-201 in *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences*, edited by Fred Roberts. New York: Springer-Verlag.

Johnsen, Eugene C., H. Russell Bernard, Peter D. Killworth, Gene A. Shelley, and Christopher McCarty. 1995. "A Social Network Approach to Corroborating the Number of AIDS/HIV+ Victims in the U. S." *Social Networks* 17:167-87.

Killworth, Peter D. and H. Russell Bernard. 1978. "The Reverse Small-World Experiment." *Social Networks* 1:159-92.

Killworth, Peter D., Eugene C. Johnsen, H. Russell Bernard, Gene A. Shelley, and Christopher McCarty. 1990. "Estimating the Size of Personal Networks." *Social Networks* 12:289-312.

Killworth, Peter D., Eugene C. Johnsen, Christopher McCarty, Gene A. Shelley, and H. Russell Bernard. 1998. "A Social Network Approach to Estimating Seroprevalence in the United States." *Social Networks* 20:23-50.

Killworth, Peter D., Christopher McCarty, H. Russell Bernard, Eugene C. Johnsen, John Domini, and Gene A. Shelley. 2003. "Two Interpretations of Reports of Knowledge of Subpopulation Sizes." *Social Networks* 25:141-60.

Killworth, Peter D., Christopher McCarty, H. Russell Bernard, Gene A. Shelley, and Eugene C. Johnsen. 1998. "Estimation of Seroprevalence, Rape and Homelessness in the U.S. Using a Social Network Approach." *Evaluation Review* 22:289-308.

Kochen, Manfred, ed. 1989. *The Small World*. Norwood, NJ: Ablex.

Laumann, Edward. O. 1969. "Friends of Urban Men: An Assessment of Accuracy in Reporting Their Socioeconomic Attributes, Mutual Choice and Attitude Agreement." *Sociometry* 32:54-69.

Mauldin, R. Dan, ed. 1981. *The Scottish Book: Mathematics From the Scottish Café*. Boston: Birkhaüser. (cf. also a typescript transmitted by Stanley Ulam in 1958 from discussions at the Scottish Cafe; cf. http://www-gap.dcs.st-and.ac.uk/ ~ history/HistTopics/Scottish_Book.html)

McCarty, Christopher, Peter D. Killworth, H. Russell Bernard, Eugene C. Johnsen, and Gene A. Shelley. 2001. "Comparing Two Methods for Estimating Network Size." *Human Organization* 60:28-39.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Reviews of Sociology* 27:415-44.

Milgram, Stanley. 1967. "The Small World Problem." *Psychology Today* 22:60-7.

Newman, Mark E. J. 2002. "Ego-Centered Networks and the Ripple Effect." *Social Networks* 25:83-95.

Newman, Mark E. J., Duncan J. Watts, and Steven H. Strogatz. 2002. "Random Graph Models of Social Networks." *Proceedings of the National Academy of Science* 99 (suppl. 1): 2566-72.

Radcliffe-Brown, Alfred R. 1957. *A Natural Science of Society*. Glencoe, IL: Free Press.

Riordan, John. 1968. *Combinatorial Identities*. New York: John Wiley.

Roberts, John M. Jr. and Devon D. Brewer. 2001. "Measures and Tests of Heaping in Discrete Quantitative Distributions." *Journal of Applied Statistics* 28:887-96.

Shelley, Gene A., H. Russell Bernard, Peter D. Killworth, Eugene C. Johnsen, and Christopher McCarty. 1995. "Who Knows Your HIV Status? What HIV+ Patients and Their Network Members Know About Each Other." *Social Networks* 17:189-217.

Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Skvoretz, John, T. J. Fararo, and F. Agneessens. 2004. "Advances in Biased Net Theory: Definitions, Derivations, and Estimations." *Social Networks* 26:113-39.

Thibaut, John W. and Harold H. Kelley. 1959. *The Social Psychology of Groups*. New York: John Wiley.

Travers, Jeffrey and Stanley Milgram. 1969. "An Experimental Study of the Small World Problem." *Sociometry* 32:425-43.

Watts, Duncan J. 1999. "Networks, Dynamics, and the Small-World Phenomenon." *American Journal of Sociology* 105:493-527.

———. 2003. *Six Degrees: The Science of a Connected Age*. New York: Norton.

———. 2004. "The 'New' Science of Networks." *Annual Reviews of Sociology* 30:243-70.

White, Harrison. 1970. "Search Parameters for the Small World Problem." *Social Forces* 49:259-64.

Willer, David, ed. 1999. *Network Exchange Theory*. Westport, CT: Praeger.

Willer, David and Bo Anderson, eds. 1981. *Networks, Exchange and Coercion*. New York: Elsevier.

Willer, David, Henry A. Walker, Barry Markovsky, Robb Willer, Michael Lovaglia, Shane Thye, and Brent Simpson. 2002. "Network Exchange Theory." Pp. 109-44 in *New*

Directions in Contemporary Sociological Theory, edited by Joseph Berger and Morris Zelditch Jr. Lanham, MD: Rowman & Littlefield.

Peter D. Killworth is a professor and individual merit scientist at the National Oceanography Centre, Southampton, United Kingdom. His research interests include the propagation of ocean planetary waves, the dynamics of climate variability, and the physics that tie persons together.

Christopher McCarty is the survey director at the Bureau of Economic and Business Research at the University of Florida, Gainesville. His research interests include the analysis of the structural properties of personal networks, software development for personal networks, survey research methods, and economic indicators for developing countries.

Eugene C. Johnsen, professor emeritus of mathematics at the University of California, Santa Barbara, is a research and consulting mathematician whose interests are in applying mathematics to the study of important problems in the social sciences. He is currently investigating social influence processes in groups with Noah Friedkin, as well as social network theory and methods for large populations with his coauthors on this article.

H. Russell Bernard is a professor of anthropology at the University of Florida, Gainesville. His research, in Greece and Mexico, is on social networks, technology and social change, and indigenous literacy.

Gene A. Shelley is an adjunct associate professor in the Department of Anthropology and Geography at Georgia State University in Atlanta. Her research interests include intimate partner violence, child abuse, and HIV.