Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Social Networks 25 (2003) 141–160
SOCIAL NETWORKS
www.elsevier.com/locate/socnet

Two interpretations of reports of knowledge of subpopulation sizes

Peter D. Killworth [a,*], Christopher McCarty [b], H. Russell Bernard [b], Eugene C. Johnsen [c], John Domini [b], Gene A. Shelley [d]

[a] *James Rennell Division, Southampton Oceanography Centre, European Way, Southampton, Hants SO14 3ZH, UK*
[b] *University of Florida, Gainesville, FL, USA*
[c] *University of California, Santa Barbara, CA, USA*
[d] *Georgia State University, Atlanta, GA, USA*

Abstract

We asked respondents how many people they knew in many subpopulations. These numbers, averaged over large representative samples, should vary proportionally to the size of the subpopulations. In fact, they do not. We give two different interpretations of this finding. The first interpretation notes that the responses are linear in subpopulation size for small subpopulations, but with a non-zero offset, and become noisier for larger subpopulations. Our explanation assumes that respondents both invent and forget members of their networks in the subpopulations, in addition to guessing when the number concerned becomes large. The second interpretation notes that the responses are well described by a power law response, in which the mean number of subpopulation members reported known varies as the square root of the subpopulation size. Despite the apparent implausibility of this, we suggest a psychological mechanism and a model which is able to reproduce the behaviour. Other recall data are shown to have similar properties, thus widening the relevance of the findings.

Keywords: Reporting; Estimation; Accuracy; Power law

1. Introduction

Our recent research has been aimed at estimating the size of hard-to-count subpopulations by the use of social network techniques. The method involves asking a nationally representative group of respondents how many people they know in many subpopulations (whose size we already know) together with how many they know in some subpopulations whose size we wish to estimate. The pattern of responses by a single respondent concerning subpopulations of known size enables us to make a maximum likelihood estimate of

* Corresponding author. Tel.: +44-23-8059-6202; fax: +44-23-8059-6204.
E-mail address: p.killworth@soc.soton.ac.uk (P.D. Killworth).

0378-8733/03/$ – see front matter
PII: S0378-8733(02)00040-0

the number of people in that respondent's network. Then, the pattern of responses about subpopulations of unknown size, across respondents, together with the estimated sizes of their networks, enables a second maximum likelihood estimate to be made for the size of the subpopulation (Killworth et al., 1998b).

Somewhat surprisingly, the maximum likelihood estimates all involve linear and unweighted averages. As a result, it turns out that essentially most deductions from our model are linear. As an example, the number known in a subpopulation, averaged nationally, should be proportional to the size of the subpopulation, other things being equal. Given this latter (and at this stage deliberately vague) proviso, then, this proportionality should serve as a useful test of not just our specific model but also of how accurately respondents can estimate the size of quantities which are somewhat difficult to estimate (the reader may wish to consider, for example, how many people called Michael he or she knows).

This paper is concerned with what we found when we examined whether this simple proportionality actually held. While even with nationally representative surveys we might still expect a little scatter, we expected the proportionality to hold well. It did not. Furthermore, depending on how the data were plotted (whether on linear or log–log plots), two different interpretations of the same data sets could be made. Both required modification of our original model in the way respondents carry out the process of 'dredging' those members of their network who lie in certain subpopulations, especially in how the process of recall affects this process.

There is, of course, a large literature on how responses to stimuli are produced (and can lead to non-linear behaviour, particularly power laws), and we shall briefly discuss this below. However, as far as we are aware, the majority of research discussing such responses has been aimed at how the stimulus response occurs, what are the reaction times, etc. rather than at the process of recall of information. Yet data from recall are the mainstay of many important areas of social science research, and the possibility of a consistent distortion of actual values by a reporting process, together with the better known examples of simple random noise, omission and deletion, etc. needs investigating. The problem of 'informant inaccuracy' has long been studied in many fields which rely on their data by asking people for information (cf. the review article by Bernard et al., 1984). Studies of informant accuracy are mainly in necessarily simplistic terms (an informant said such-and-such; was this true?) rather than more generally seeking to explain and/or predict the variety and magnitude of the errors obtained in the data gathering.

This paper, then, carries out two parallel tasks, after a discussion of our data sources and expectations (Sections 2 and 3). First, we note that the same data sets are capable of generating at least two interpretations (Section 4). Second, we argue that either interpretation requires a modification of our original model, which assumes respondents essentially respond completely and accurately (Sections 5 and 6). We then discuss these two viewpoints (Section 7).

2. Data sources

Three national telephone surveys were conducted, which we term here "survey 1", "survey 2" and "clergy", with 796, 594 and 131 respondents, respectively. Surveys 1 and 2 were

Table 1

Subpopulation
First name “Michael”
First name “Christina”
First name “Christopher”
First name “Jacqueline”
First name “James”
First name “Jennifer”
First name “Anthony”
First name “Kimberly”
First name “Robert”
First name “Stephanie”
First name “David”
First name “Nicole”
American Indian
Woman given birth last 12 months
Woman adopted a child in past 12 months
Man or woman widowed and under age 65
Person on kidney dialysis
Postal worker
Commercial pilot
Member of JAYCEES
Person with diabetes
Opened a business in past 12 months
Have twin brother or sister
Licensed gun dealer
Came down with AIDS
Male incarceree in state or federal prison
Homicide victims in past 12 months
Committed suicide in past 12 months
Died in motor accident in past 12 months

based on nationally representative samples of respondents; clergy was based on a representative sample of clergy across the US, from a range of religions (e.g. Jewish, Baptist, etc.). Details of the survey sources and methodology are given in McCarty et al. (2001); note in particular that the clergy sample, purchased through a nationally known sampling service, represented the wide variety of clergy roughly proportional to their representation in the population.

Each respondent was asked, inter alia, how many people they knew[1] in each of the 29 subpopulations (i.e. clearly defined subsets of the US, such as diabetics, people named Michael, etc.; a list of these is in Table 1).

Some caveats about our data sources and methods: (1) interviewing people by phone meant that respondents had only, on average, about 30 s to recall or estimate a number of

[1] We told respondents that knowing someone meant: “. . . you know the person and they know you by sight or by name; you can contact them in person, by telephone or by mail; and you have had contact with the person in the past two years”.

people they knew in each subpopulation; (2) the important feature of all 29 subpopulations is that their total size is known in each case; (3) we included the clergy sample because members of this group are widely thought to have larger-than-average social networks (i.e. the sample was *not* nationally representative); (4) survey 2 was in all respects similar to survey 1, and so will be treated here mainly as confirmation that our results were not merely produced from random noise; and (5) the data were not gathered with this study in mind, and our discoveries about the failure of simple proportionality models were not made until after the end of data acquisition. Thus, surveys including various confirmatory questions suggested in this work have not yet been made.

3. Expectations

We shall be concerned here with how the mean number of people known in a subpopulation depends on the size of that subpopulation (denoted by e; the entire population is of size t). Our model for this, discussed in previous work (e.g. Killworth et al., 1998a,b) is simple, and has a very simple and intuitive outcome. Suppose that the probability that a respondent knows c people is $P(c)$; the work cited gives our current estimates for the distribution of $P(c)$. We assume that each member of the subpopulation is equally likely to be a member of the respondent's network, with probability $p = e/t$ (the cited papers, and McCarty et al., 2001, discuss the potential shortcomings of this assumption, most of which are overcome with a sufficiently representative sample). Then the mean number of people in a subpopulation known to a respondent is given by

$$\bar{m} = \int P(c)\bar{n}\,\mathrm{d}c = \int P(c)\left\{\sum_{n=0}^{\infty} n \cdot \mathrm{prob}(n|c,\, p)\right\}\,\mathrm{d}c \tag{3.1}$$

where $\bar{n}$ is the mean number of people known in the subpopulation, and $\mathrm{prob}(n|c,\, p)$ is the probability that a respondent knowing c people will know exactly n from a subpopulation of size e. Whether we use a binomial or a Poisson distribution for this probability, the mean number known for a respondent knowing c people altogether is $\bar{n} = cp = ce/t$, so that

$$m = \frac{e}{t}\int cP(c)\,\mathrm{d}c \equiv \frac{e}{t}\bar{c} \tag{3.2}$$

where $\bar{c}$ is the average number of people known by a respondent overall. Thus, the mean number known in a subpopulation should be proportional to the size of the subpopulation. A logical consequence of this intuitive feature of our model is that, for example, in combining two populations, the mean number known in the combined set should simply be the sum of the mean numbers known in each separately. Is this intuitive result correct? We now examine the same data in two different ways, which will suggest two completely different departures from our simple model.

4. Results

4.1. Linear graphing of the data

Fig. 1 shows the mean number of people known to respondents in each of the 29 subpopulations, as a function of the (fractional) size of the subpopulation, for surveys 1 and clergy (recall that results for survey 2 are highly similar to survey 1, and are not shown for clarity). Qualitatively, the figure shows little unusual: the larger the subpopulation, the larger the number which respondents report knowing in that subpopulation. However, several aspects are unusual.

First, anomalously high reports are consistently found for small subpopulations (of size around 30,000, e.g. suicides in the last 12 months, victims of homicides, etc.). We deem these high since respondents claim on average to know about 0.2–0.3 such people. Yet these subpopulations occupy at most 1/10,000 of the US, so that to reproduce these figures, on average respondents would have to know, on average, 0.2–0.3 times 10,000, or 2000–3000 individuals. This is unlikely, given that our earlier work, using a variety of methods (Killworth and Bernard, 1977, 1978; Bernard et al., 1989, 1991; Killworth et al., 1998a,b), produced consistent estimates of $\bar{c}$ of just under 300 for the same definition of knowing.

Second, the data are roughly linear with subpopulation size for small subpopulations, but then—with some scatter—this linearity disappears for larger subpopulation, though less so for the clergy sample (in addition, the number known in each of the subpopulations by clergy are around twice those for surveys 1 and 2, confirming our belief that clergy are likely to know more people than the national average).

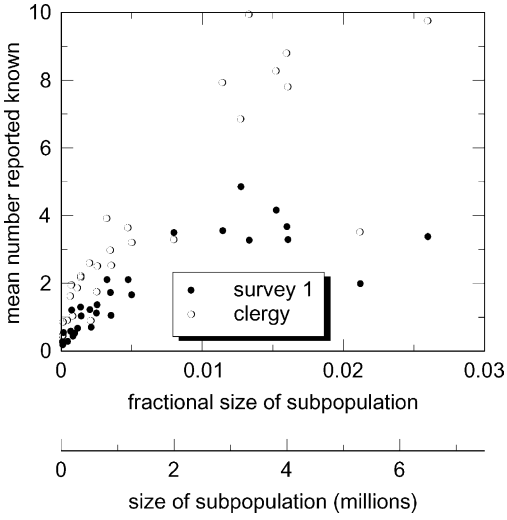

Fig. 1. Mean reported number known in each of the 29 subpopulations, against the size of that subpopulation within the US (shown as absolute and as a fraction of 250 million). Survey 1 is shown with filled circles, clergy with open circles. Survey 2 is very similar to survey 1 and is not shown for clarity.

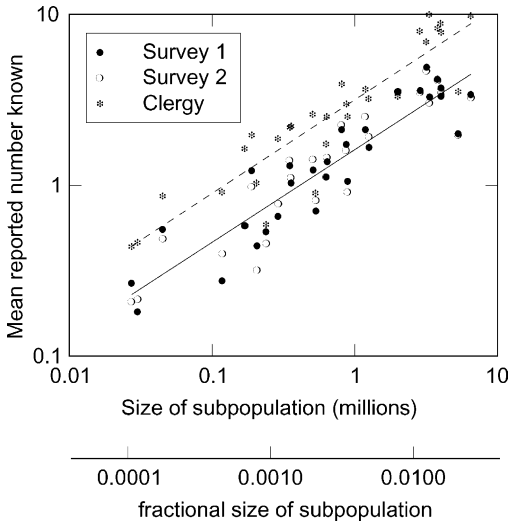

Fig. 2. Log–log plot of mean reported number known in each of 29 subpopulations, against the size of that subpopulation within the US (shown as absolute and as a fraction of 250 million). Also shown are best fit (power) lines for these data. Surveys 1 and 2 are indistinguishable, and are indicated by the firm line; clergy by the dashed line. The best fits correspond to reported numbers varying as the square root of subpopulation size.

Thus, the simple model of constant proportionality to the subpopulation size does not hold. The sample sizes are large enough that noise from small samples is unlikely to be generating the discrepancies. Thus, responses appear to fit some more non-linear law, with a region of linearity between—apparently—regions of over- and under-reporting.

4.2. Log–log graphing of the data

However, the data can be shown in a different form. Fig. 2 shows the identical data but on a log–log scale. The data uniformly suggest a gradient of 1/2 in log–log terms; in other words, the mean number reported varies as the square root of the size of the subpopulation. This is confirmed both by linear fits in log–log terms as well as least-squares fits to power laws using the actual data. The two fits are visually indistinguishable. Thus, the mean number reported apparently varies as the square root of subpopulation size. Only two best fit lines are shown, since the fits for surveys 1 and 2 are indistinguishable, but in all other respects show the same square root behaviour.

The power law persists even when we test for variation by sex and age. On average, men show a weak tendency for larger responses, but Fig. 3 shows that men and women exhibit very similar response behaviour. Fig. 4 shows that age does not affect response behaviour, either.

From the log–log perspective, then, we are led to the conclusion that the mean number reported known within a subpopulation varies as the square root of the subpopulation size. This is clearly counterintuitive, and the behaviour should not be expected. We can see this from a simple thought experiment. Suppose that respondents are asked about some

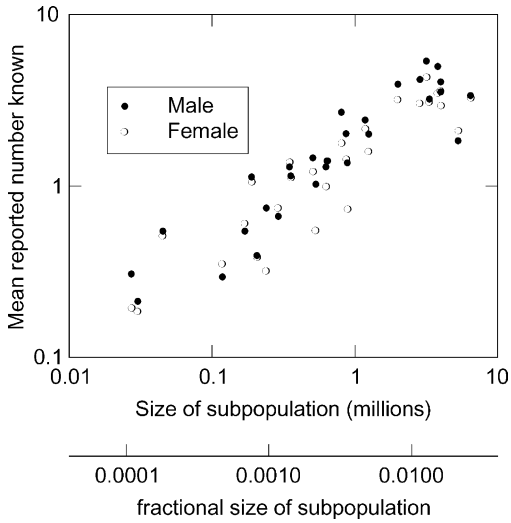

Fig. 3. As Fig. 2, but using the combined surveys 1 and 2, separating the data into male and female respondents.

subpopulation of size e, and report a mean value known of $\bar{m}$. Now suppose that we split this subpopulation into two equally sized subsets—of size $e/2$—e.g. those born in the first and second halves of the year, and ask respondents to estimate the numbers they know in each of the two subsets. They report $\bar{m}_1$ and $\bar{m}_2$, say. Then clearly we should expect $\bar{m}_1 + \bar{m}_2 = \bar{m}$. On average, of course, unless the breakdown of the original subgroup was chosen pathologically, we must have $\bar{m}_1 = \bar{m}_2 = \bar{m}/2$.

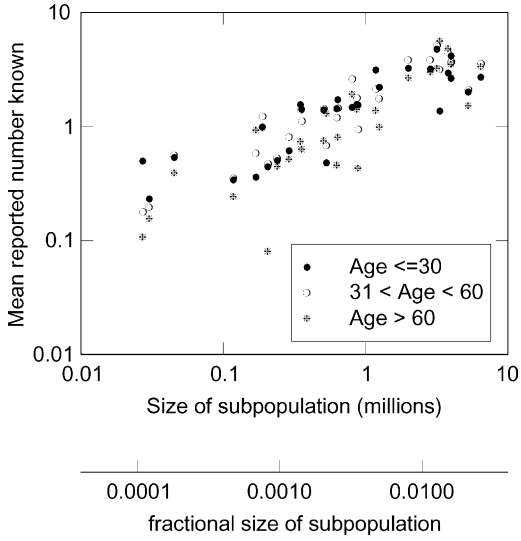

Fig. 4. As Fig. 2, but separated into the age categories shown.

Care must be taken with such arguments, since the presentation of portmanteau subpopulations in this manner was never made; the subpopulations were essentially disjoint (diabetics; people called Michael; etc.). Thus, the belief that the reported sizes of the unions of subpopulations should be additive may well be correct, but remains untested in our data. Presumably a respondent presented with the questions in the previous paragraph would indeed provide an additive response. However, we have not yet produced any rationale for the square root behaviour, so that we must return for the moment to the simple and intuitive model with which we began, that the mean number reported in a subgroup should vary proportionally to the size of the subgroup.

4.3. The problem

We have examined the same data in two ways, and two different interpretations have appeared, neither agreeing with our simple model. We now examine how the model might be altered to fit either of these interpretations.

5. Modifying the model: omission, commission and guessing

We first modify the linear response model. Our modifications are in two parts. First, we permit errors in reporting the number known in a subpopulation, including both errors of omission and of commission. Second, when the number known is high, we permit guessing to occur.

To include reporting errors, we again assume that the respondent knows c people, of whom n are members of a subpopulation of fractional size p. The probability of this is simply

$$ {}^{c}C_{n}(1-p)^{c-n}p^{n} $$

using a binomial distribution, where ${}^{c}C_{n}$ is the number of combinations of c items taken n at a time. However, the respondent does not report the number correctly. Specifically, there is a probability

- α that the respondent may forget any of the n members (i.e. an error of omission); and
- β that the respondent may falsely believe that any of the $c-n$ who are not in the subpopulation are in that subpopulation (i.e. an error of commission).

Then the expected number reported will be

$$ (c-n)\beta + n(1-\alpha) = \beta c + n(1-\alpha-\beta). \tag{5.1} $$

To obtain the expected number reported across a representative sample, we average this over the distribution of c and over all n, namely

$$ \begin{aligned} \bar{m} &= \int P(c) \sum_{n} {}^{c}C_{n}(1-p)^{c-n}p^{n}\{\beta c + n(1-\alpha-\beta)\} \\ &= \beta\bar{c} + (1-\alpha-\beta)\int P(c)\sum_{n} {}^{c}C_{n}(1-p)^{c-n}p^{n}n \end{aligned} \tag{5.2} $$

where we have used the identity

$$\sum_n^c C_n (1-p)^{c-n} p^n = 1.$$

To find the second term, we note that

$$\sum_n^c C_n (1-p)^{c-n} p^n n = \text{expected value of } n = cp.$$

Thus, the mean number reported is simply

$$m = \beta\bar{c} + (1-\alpha-\beta)\int P(c)cp = \beta\bar{c} + (1-\alpha-\beta)p\bar{c}. \tag{5.3}$$

This is linearly proportional to $\bar{c}$, as should be expected, and also varies linearly with respect to p. However, there is an offset when $p = 0$ caused by commission errors; the respondent will believe that some network members are in a subpopulation of zero size.

The best fit from survey 1 to this expression for subpopulations under 0.01 in fractional size (where Fig. 1 shows that a good linear fit applies) is

$$\bar{m} = 0.33 + 365p \tag{5.4}$$

(with a correlation of 0.92) so that under this model

$$\begin{aligned} \beta\bar{c} &= 0.33 \\ (1-\alpha-\beta)\bar{c} &= 365. \end{aligned} \tag{5.5}$$

This is insufficient to solve for the triplet $\bar{c}$, α, and β. However, since we expect $\bar{c}$ to be several hundred, β is clearly small. Then $(1-\alpha)\bar{c} \approx 365$. If α is also small, then

$$\begin{aligned} \bar{c} &\approx 365 \\ \beta &\approx 0.0009 \end{aligned} \tag{5.6}$$

and α is not well determined. This simple model thus predicts that the mean number known rises from about 290 (for our original model, cf. Killworth et al., 1998b) to 365, and there is a chance of about 0.1% of a respondent believing that any network member is in any given subpopulation when in fact the member is not.

This model predicts a linear variation with p; Fig. 1 shows that this is not the case for larger values of p. Where the linear variation fails clearly depends on the class of respondents, since the clergy show a linear relationship for p distinctly above those where the relationship has failed for survey 1 (and survey 2, not shown). We amend the model a second time. When a respondent is asked to report a number which is in fact above some cut-off value (which may vary between respondents; we have no way to tell at this time), instead of examining their network members mentally and counting, they *estimate*. We have evidence of this happening. McCarty et al. (2001) show that respondents provide m values which are rounded, frequently being multiples of 5, when the numbers reported are above 10. We assume here that estimating includes simple guessing ("Oh, I must know at least 30 Michaels").

The result of the guessing is increasing inaccuracy for the simple model here as p increases, since more and more respondents are placed in the guessing rather than counting

regimes. Thus, the linear response to increasing p becomes a scatter, often with considerable underestimation. The model does not as yet provide a rationale for why under-, rather than overestimation is preferred. Such biases are known to appear systematically in other fields, e.g. part–whole proportion judgements, and models for the occurrence of estimation bias are beginning to appear (e.g. Hollands and Dyre, 2000).

6. Modifying the model: why should a power law occur?

6.1. Power laws

Power laws are, of course, ubiquitous in science. Over a century ago, Fechner, interested in understanding the relationship between the strength of a stimulus and the strength of the sensation it elicited in the mind, delineated a logarithmic functional relationship between the mind and body. Empirical research on power law phenomena in human perception has focused mostly on sensory modalities, which are easily manipulated under controlled, laboratory conditions. Stevens (1971) posited a quantifiable relationship between the physical and perceived intensity of a stimulus. This relationship can be expressed as the following equation, known as Stevens' power law:

$$\Psi = \kappa\phi^{\beta} \tag{9}$$

where sensory perception (Ψ) is the product of a constant κ (which is dependent upon the sensory system in question) and the physical intensity of the stimulus ϕ raised to some power, β (which is also characteristic of the sensory system being examined). Stevens (1975) gives a full discussion. The conclusion from this formula is that there exists within the processes of human cognition, an algorithmic method of magnitude estimation for sensory perception. Cognitive psychologists have extended research on power laws to memory and cognition. Anderson and Tweeney (1997), for example, demonstrate that recall and cognition decline as a power function of time. Laming (1973) in a discussion of Weber's Law notes that discriminators report the *square* of the incremental signal.

However, by far the majority of psychological research has been built around the concepts of recency, primacy and salience, with detailed treatments of both response to stimuli (e.g. estimation) and response times when recalling category contents from memory (a classic example of the latter being Indow and Togano, 1970). We have been unable to find examples in any literature which discuss how *many* items respondents can recall from a list of a given size, or their ability to judge the *size* of the list (without explicit enumeration being required on the respondent's part). The category recall experiments consistently assume that given enough time, respondents recall the entire list. To make our interpretations more difficult here, we cannot know the size of the list the respondent is attempting to reproduce, since this will depend on many factors, not least the size of the respondent's total network.

Anderson and Tweeney (1997) also raise the possibility that some power curves might be artifacts (see also Myung et al., 2000). Individual curves as a functional description of forgetting are exponential, but Bakan (1954) and Estes (1956) argued long ago that the averaging of exponential functions could result in a function that is no longer exponential. Of course, if power functions in perceptual response data are, in fact, artifacts, doubt would

be cast on theories that explain mental processes in the frame of power law learning and forgetting.

Another component of artifaction deals with the analysis of noisy data. When looking at the level of individual responses, much of the data possesses random error. Whether or not aggregation of data followed by their linearisation under a log–log transformation would result in an artifactual power curve has yet to be determined.

6.2. Processes leading to power law response

We have identified at least six possibilities which may lead to a non-linear response (though not necessarily the square root behaviour as observed). None can be rejected. Our data are insufficient to reject any of these six, though some are more likely than others. We shall briefly discuss each in turn:

1. respondents are reporting accurately, but their knowledge is inaccurate;
2. the effects of salience cannot be neglected;
3. the power law is an artifact of the averaging process;
4. respondents are misreporting accurate knowledge;
5. during a telephone survey, respondents have a finite time to respond to each question, which results in misreporting as a power law behaviour;
6. the process of 'dredging' from a list of n can produce a response varying as $n^{1/2}$.

The first two possibilities could be created by several circumstances which we are currently studying (cf. McCarty et al., 2001). Many effects can intervene to distort the simple linear relationship; the most frequent we term "barrier" and "transmission" effects. "Barrier" effects are where respondents cannot be aware of some subpopulation members for various reasons (e.g. the subpopulation lives on a remote island), and "transmission" effects are where knowledge of subpopulation membership is not transmitted to all the people in the subpopulation's combined network with equal probability (for example, because the information is uninteresting to some).

An example of a transmission effect is salience: diabetics are more likely to know diabetics than are most other respondents (but those whose name is Michael are *not* more likely to know other Michaels!). The averaging process, of course, removes—or at worst lessens—most transmission effects, since we use nationally representative samples. To see this, consider an unpublished example from the data of Shelley et al. (1995). They found that on average, seropositive respondents knew 52.5 people with AIDS, compared with 0.44 for the general US population, an increase by a factor of 120. Clearly, salience has a strong effect for respondents within this subpopulation. However, there are only about 800,000 seropositive individuals in the US (Killworth et al., 1998b) so that seropositive respondents would contribute to a number of people known who have AIDS by an amount (800,000/250 million) times 52.5, or 0.16, which is only one-third of the national average number of AIDS victims known anyway.

Transmission effects in general do play a strong role in reports. Consistently in our data, some subpopulations (e.g. diabetics) appear—by any graphic technique—to be underreported compared with those of similar sizes. We assume this is because it is simply hard for a respondent to know that a network member is a diabetic; other than when the member

was diagnosed diabetic, this fact is usually judged un-newsworthy by the network member concerned. Often, a random comment during a meal with the network member may be the only means whereby the respondent discovers the information.

The third possibility is that the power law may be an artifact of the averaging process (cf. the discussion by Anderson and Tweeney, 1997). This does appear unlikely, since the linearity of our averaging is transparent from the simple arguments above. A more complete analysis, using a binomial distribution to predict how the number known by a respondent, who knows c people, is distributed (which is part of the procedure we use in our approach to predict the sizes of subpopulations), also shows complete linearity when averaged across a national survey. Nonetheless, as we shall show, a square root power law also approximately survives national averaging, so that care must be taken in any interpretation here.

The next possibility (of simple misreporting) can clearly not be ruled out (and, indeed, the model of remembering presented below can be thought of as an example of this). We have argued that over-reporting of small subpopulations may well be occurring. Also, preferential selection of numbers estimated (as round, rather than precise, numbers) is clearly evident. McCarty et al. (2001) discuss this in detail, and show that these effects play a small role in the final average figures produced. However, one can create a pattern of misreporting which would turn a linear response into a square root response, but finding a rationale for such a pattern is not easy. For example, since all reported numbers are small (typically under 10), one would need a systematic under-reporting of larger numbers and over-reporting of small ones to achieve the observed behaviour.

This may well be happening, since we are aware from focus group studies we carried out (cf. McCarty et al., 2001, for details) that respondents choose to estimate rather than enumerate when the m they report passes some size. It is possible that this estimation produces under-reporting. Nonetheless, we have no model of this process at our disposal. Also, simple tests such as mentioned above (would the sum of respondents' estimates of, say, the male and female diabetics they know be equal to their reported estimate of the total diabetics they know?) were not carried out for our data because we did not ask respondents about the gender of each of their network members.

The fifth possibility—that respondents only have a finite time to respond, which may bias their responses—certainly occurs. In fact, respondents were not given a formal time in which to respond, but the structure of a telephone survey certainly must induce some feeling of finite time in which to respond, and we believe 30 s is a good estimate. Thus, when attempting to remember all those in some category, respondents produce a cumulative total over time. A model for this is well known (e.g. Bousfield and Sedgewick, 1944 and later papers). It predicts that if respondents know n members of a category, they report a cumulative total $n_c(t) = n[1 - \exp(-\lambda t)]$ in time t, for some unknown decay rate λ. If λ is independent of the category size, then for similar interview times, the same constant fraction of size is predicted, independent of the category size. This could not explain our findings. Indow and Togano (1970) extend the Bousfield and Sedgewick model, suggesting that λ varies inversely as n (en route, they suggest that the probability that respondents can find members of the category during some time interval at time t varies as $[1 - n_c(t)/n]$, a point we shall refer to below).

Their model would thus predict that the number recalled in some fixed time would be $m = n[1 - \exp(-K/n)]$ for some constant K. Since we assume that n is some constant

(but unknown) multiple of e, it follows that $m = \alpha e[1 - \exp(-L/e)]$, for unknown constants α, L. The best fit to our data (not shown) is reasonable, but lies beyond the 95% confidence limits for most of the data. On the other hand, the square root fit lies outside the limits for very few of the data points. Thus, we cannot reject the possibility that the finite respondent time yields a curve similar, but not identical, to a square root law—especially as Indow and Togano (1970) note that the observational evidence for their inverse n behaviour for the decay term is poor, though there is clear evidence of the decay decreasing with n. So a more general dependence of the decay term could well fit the data at least as well as the square root behaviour. Note, however, that the numbers recalled by our respondents are small (typically under 10) whereas typical category lists involve an order of magnitude larger size; the data reported by Indow and Togano would suggest that respondents would have no difficulty in recalling lists of 10 completely during our telephone surveys. The process required to reproduce the lists used by Indow and Togano, however, differs strongly from the enumeration required of our respondents, who must both extract a list of their network members and then judge who within that list lies within each subpopulation.

The sixth and final possibility does not appear to have been discussed in the literature, namely that respondents attempting to recall from a category of size n yield a response varying as $n^{1/2}$. In the next section we produce a simple model of how this might occur.

7. A simple model of knowledge recall

We here pursue a rationale whereby the respondent's knowledge of n members of a subgroup could be reported as less than n. To be specific: the mean number reported known is an average over respondents whose network size c varies as well as over a probability distribution for the number actually known in a subpopulation, modified by the process we seek here to explain—which is how this is transformed into the number reported. Although this *process* is non-linear, it turns out that the dependence on subpopulation size propagates through the averaging process as long as this dependence is a power law or close to one. A proof is given in Appendix A. Thus, the average of a square root process remains a square root as far as such a power law dependency is concerned.

The model presented here involves active recall, and so would be valid for recall of a few members of a network (all of the subgroups in Fig. 1 have a mean number recalled of under 5). (Other approaches, relevant to the estimation process and hence to larger subpopulations, is discussed in Johnsen et al., submitted for publication.)

The model is devised only for recall of items which are hard to extract from memory in a finite time. If a respondent has five friends, he or she is presumably aware of that and will respond accordingly. If asked how many diabetics he or she knows, however, a more complex listing process must be made, and our model refers to how estimates from this list (i.e. 'dredging' of information) are made.

Let us suppose that the respondent is actively involved in the recall of n people, and has so far recalled i of them (n is here at least 1, else there is nothing to model). This process becomes steadily more difficult as i increases. The simplest model of this is to assume

that the probability that the respondent can recall another member is simply the fractional number of members remaining. So at the start of the process all members are available, and the respondent can recall at least one with probability unity (we assume the respondent is *aware* that at least one of his or her network is in the subpopulation, hence the initial probability of unity). If the respondent reaches halfway, the probability that another member can be recalled is one-half. If the respondent has recalled all members, then, sensibly, the probability of recalling another member must be zero.

To solve this simple problem for the expected number recalled, we write

$$p_i = \text{probability of recalling exactly } i \text{ members.}$$

Now the respondent can always recall one member at the start of his or her recall. The chance that this is the *only* member recalled is 1 minus the probability that he or she can recall another, which is $(n-1)/n$. Thus, $p_1 = 1-(n-1)/n = 1/n$. Continuing, the probability that the respondent has recalled at least up to i members is

$$1 - \sum_{j=1}^{i-1} p_j, \quad i \geq 2. \tag{7.1}$$

Our simple assumption above then means that the probability that, having recalled i, the respondent can recall one more (at least), is

$$\frac{n-i}{n}. \tag{7.2}$$

(This has similarities with, but is *not* identical to, the arguments of Indow and Togano (1970).)[2]

Equivalently, the probability of not recalling an additional member is

$$1 - \frac{n-i}{n} = \frac{i}{n}. \tag{7.3}$$

Thus, the probability that exactly i are recalled is the product of (a) the probability that the respondent has recalled up to i members (7.1 above) and (b) the probability that the respondent cannot recall another member (7.3 above), so that

$$p_i = \frac{i}{n}\left(1 - \sum_{j=1}^{i-1} p_j\right), \quad p_1 = \frac{1}{n}. \tag{7.4}$$

An analytical solution to this is given in Appendix B, but is not enlightening. We require the expected number recalled, which is

$$E(n) = \sum_{i=1}^{n} i p_i. \tag{7.5}$$

[2] In their paper, the expression (7.2) represented the chance that respondents could locate additional members of a list during a specified time interval. If they failed, they would simply continue during the next time interval until they had eventually retrieved the entire list. In our model, once a respondent fails to locate *one* additional member of the list, the respondent quits and reports no further members.

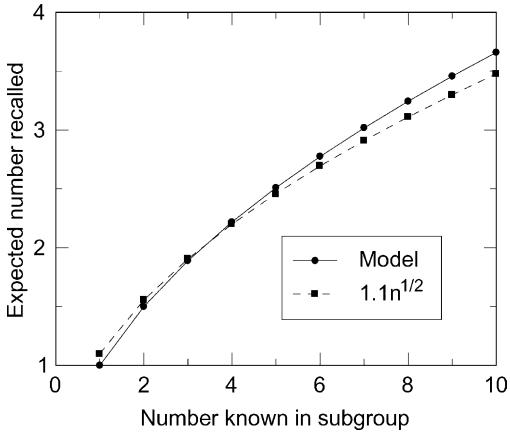

Fig. 5. A plot of the predicted number reported known by a respondent, against the actual number known, using the simple model for reported knowledge in this paper. The firm line shows the predicted number; the dashed line shows a square root power law $1.1n^{1/2}$. Over a much wider range of n (up to 100, the largest number reported by nationally representative respondents) a power law proportional to $n^{0.524}$ is visually indistinguishable from the model predictions.

Fig. 5 shows $E(n)$ against n over the range $1 \leq n \leq 10$, together with the curve $1.1n^{1/2}$, showing that the response of the model strongly resembles a square root power law. (Respondents in survey 1 reported n values up to 95; over this larger range the power law

$$E(n) \approx 1.078n^{0.524} \tag{7.6}$$

is visually indistinguishable from the accurate solution.) We argue that neither (7.5) nor (7.6) is specifically a power law; but their numerical values are indistinguishable from a power law, which is what matters.

Thus, our simple model is able to reproduce the approximate square root behaviour. This statement assumes that the averaging process over respondents with different network sizes does not strongly modify the square root behaviour. Given that for small network sizes, n can mainly be 0 or 1, there will be some quantisation effects.

8. Discussion

This paper began from the intuitively plausible (and modelled) belief that a nationally representative sample of respondents would, on average, report knowing a number of people in a subpopulation which is proportional to the size of that subpopulation. We have shown that this belief is not in accord with the data, and replicably so. Further, the way we perceive that reports deviate from simple proportionality can depend on the manner in which the data are plotted. In our case, there are at least two rival interpretations, which we have modelled as (a) low probability errors of omission and commission, together with randomised guessing when the number to be recalled becomes larger; and (b) a process whereby the number

recalled from a group of size n varies as $n^{1/2}$. Both model modifications, we believe, are plausible. How, then, can we choose?

We have not answered this question and cannot answer it with the data we have at hand. Our preference is, perhaps naturally, for the modified linear model (which does not suffer from the non-linear difficulties of interpretation discussed above), but we have not disproved either possibility, and there may be others. Our purpose here has been to lay out the problem and provide for a program of research that can address the problem. This may not be easy. Certainly, selected additional data would cast light on the issues. For larger subpopulations, where there is clearly scatter in the responses under any graphing technique, investigation of additional subpopulations would be useful.

We are currently investigating whether subpopulations can be cast into what might be termed 'visible' and 'invisible'—that is, difficult and easy to be aware of knowing. It is easy, for example, to be aware that one knows someone named Michael. For the population of Michaels, there may be low transmission error. By contrast, it is very difficult to be aware that one knows a diabetic (it just comes up in conversation) or someone who is HIV-positive (it is the sort of stigmatising information that people do not easily divulge). For these populations, there would be high transmission error. We are looking for a systematic way to allocate subpopulations to either of these categories (or, perhaps, onto a unidimensional scale of transmission error) so that the scatter for large subpopulations may be removed or alleviated.

For small subpopulations, investigation of additional subpopulations would also be useful. Even given the noise associated with the above transmission error biases, the data certainly suggest that a non-zero average number would be reported known for a subpopulation of vanishingly small size. Sufficiently many very small subpopulations (of fractional size under 0.1% of the total population, say) would enable us to distinguish between a non-zero offset (the modified linear theory) and a square root fall off (the power law).

The work presented here also serves as a potential warning. While all researchers examine their data visually as a matter of course, the increasing reliance on statistical packages to do this, together with their default layouts, can easily lead to only one view of data possessing several interpretations.

These findings, indeed, may not be confined to the narrow data sources described here. Indeed, any research design that calls for respondents to recall the number of occurrences of things or events, where the range of occurrences is large, may be subject to this phenomenon. Within the social network literature this kind of task is fairly common. Network studies are often based on asking people to estimate how frequently they interact with people they know. But the same problem is likely in asking people to recall the number of times last month that they ate certain foods, or the number of times last year that they visited the doctor. The consumption of foods that are eaten infrequently, for example, may be overestimated, while foods eaten frequently may be underestimated.

As a test, we returned to data of Killworth and Bernard (1978) concerning accuracy of reports by amateur radio operators of their interactions. During a month, all interactions were monitored (on public air waves), and the number of contacts between all pairs of respondents was recorded. The respondents were asked to provide an estimate (using a 0–9 scale, rather than a specific estimate) of the number of contacts they had made with each of

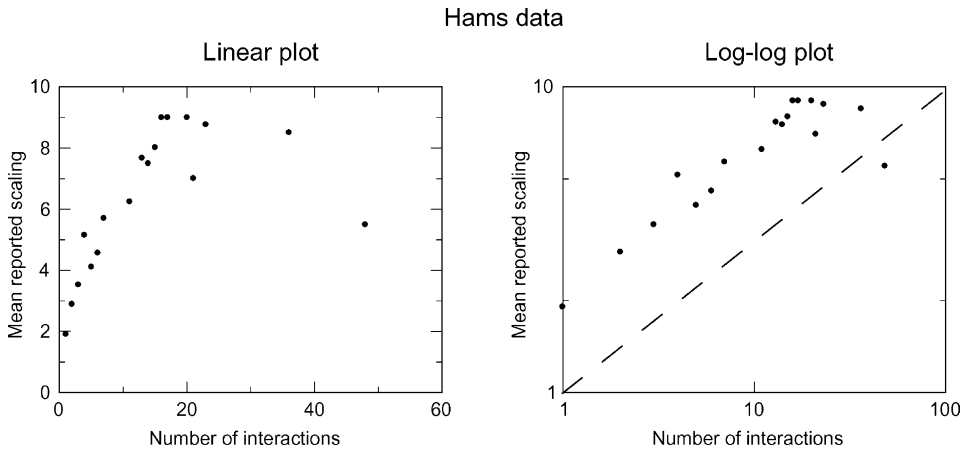

Fig. 6. Two plots of the amateur radio operators recall data. The left plot is linear–linear; the right log–log. Both show the same data. The *x*-axis is the number of interactions between two individuals during a month. The *y*-axis is the mean (binned over all pairs of individuals with that amount of interaction) of the estimated (0–9) frequency of interaction. The dashed line shows a square root dependence. Both diagrams display the same behaviour as for the subpopulation data.

the other operators. We computed the mean reported scaling for each possible amount of interaction. In this task, just as in the main body of this paper, informants were attempting to recollect a number of similar items—in this case, how many interactions they had had with someone. We show both the linear and log plots in Fig. 6.[3]

The diagrams are similar to Figs. 1 and 2 shown earlier. The linear plot displays a tendency for a non-zero estimate even for no interactions, a linear growth, followed by an apparently random behaviour for large actual numbers of interactions. The log–log plot shows a well-defined square root dependence of the mean reported scaling on the actual number of interaction. As before, either of these descriptions could be taken as 'truth', and there is no straightforward way to choose between them.

We therefore conclude that both the structures within our data, and the two interpretations of them, are probably applicable beyond the confines of our work.

Acknowledgements

We are indebted to Amber Yoder and David Kennedy for the help they provided with this survey. The Bureau of Economic and Business Research at the University of Florida provided the telephone survey services as well as a place for us to work during analysis and write-up. The work was funded through NSF grant SBR-9710353.

[3] Different respondents may interpret the 0–9 scale in differing ways, depending on their experience; thus an average confounds several possible interpretations by respondents. In addition, there are few occurrences of interactions occurring more than 10 times, so that the data are potentially noisy.

Appendix A. Convolution of a power law response through averaging processes

Respondents know differing numbers of people c, with some unknown probability distribution $P(c)$. Some estimates of this distribution are given in Killworth et al. (1998b). If a respondent recalls members of a subpopulation based on a power law, the survey-average response involves passing this power law through two averaging processes: (a) across the distribution of possible responses by respondents knowing c, and (b) across the range of c itself. It is not a priori clear that a power law can pass through both processes relatively unscathed. We demonstrate here for a square root law that this is in fact approximately the case.

The survey mean number recalled will be

$$\bar{m} = \int F(c, e)\, P(c)\, \mathrm{d}c \tag{A.1}$$

where $F(c,e)$ is the mean number recalled by respondents knowing c people altogether, for a subpopulation of size e. In turn,

$$F(c, e) = F(E) = e^{-E} \sum_{n=0}^{\infty} \alpha n^{1/2} \frac{E^n}{n!}, \quad \text{where } E = \frac{ce}{t}. \tag{A.2}$$

Here t is the size of the entire population, and we have assumed a Poisson distribution for the (rare) case of knowing someone in a small subpopulation. In such a case, the expected number known is E, the probability of knowing precisely n is $e^{-E}E^n/n!$, and when n are known, we assume a number $\alpha n^{1/2}$ is reported.

Computed values of $F(E)$ are shown, with unit proportionality α, for $0 \leq E \leq 10$ in Fig. 7; also shown is $E^{1/2}$. The agreement is clearly within the error margins of the power law results, so that we may assume that a square root law passes through the varying responses for a given c without too much distortion.

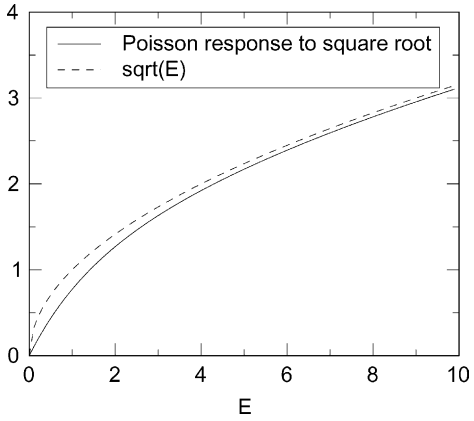

Fig. 7. A plot of how a square root response at the individual respondent level is averaged through a binomial or Poisson process, showing that the response is only slightly distorted by averaging and remains approximately square root in structure.

Then in (A.1), the role of e is merely a multiplicative factor of approximately $e^{1/2}$ (since F varies similarly), and so

$$\bar{m} \propto e^{1/2}$$

to a good degree of approximation. In other words, if there is a square root power law in individual responses, it survives approximately the averaging necessary to form a representative mean.

Appendix B. The solution to the simple model

We write

$$q_i = \sum_{j=1}^{i-1} p_j, \quad i > 1. \tag{B.1}$$

Then

$$p_i = \frac{i}{n}(1 - q_i) \tag{B.2}$$

is to be solved. We note that (B.1) implies

$$p_i = q_{i+1} - q_i$$

so that (B.2) becomes a recurrence relation

$$q_{i+1} = \frac{i}{n} + \frac{n-i}{n} q_i \tag{B.3}$$

with initial value $q_1 = 0$. Eq. (B.3) has a special case solution $q_i = B$, where B is some constant. Since this does not satisfy the initial condition, a complementary function must be added on, which is a solution of (B.3) without the first term on the right-hand side, namely

$$\frac{r_i}{n^i}, \quad r_i = \frac{(n-1)!}{(n-1)!} A$$

for some constant A. Substituting into the two conditions on q_1, q_2 we have $B = 1, A = -n$, so that

$$q_i = 1 - \frac{(n-1)!}{(n-i)! n^{i-1}}, \quad i > 1; \quad q_1 = 0$$

giving finally

$$p_i = \frac{i}{n^i} \frac{(n-1)!}{(n-i)!}.$$

References

Anderson, R.B., Tweeney, R.D., 1997. Artifactual power curves in forgetting. Memory and Cognition 25, 724–730.
Bakan, D., 1954. A generalization of Sidman's results on group and individual functions and a criterion. Psychological Bulletin 51, 63–64.
Bernard, H.R., Killworth, P.D., Kronenfeld, D., Sailer, L.D., 1984. The problem of informant accuracy: the validity of retrospective data. Annual Review of Anthropology 13, 495–517.
Bernard, H.R., Johnsen, E.C., Killworth, P.D., Robinson, S., 1989. Estimating the size of an average personal network and of an event subpopulation. In: Kochen, M. (Ed.), The Small World. Ablex, Norwood, NJ, pp. 159–175.
Bernard, H.R., Johnsen, E.C., Killworth, P.D., Robinson, S., 1991. Estimating the size of an average personal network and of an event subpopulation: some empirical results. Social Science Research 20, 109–121.
Bousfield, W.A., Sedgewick, H.W., 1944. An analysis of sequences of restricted associative responses. Journal of General Psychology 30, 149–165.
Estes, W.K., 1956. The problem of inference from curves based on group data. Psychological Bulletin 53, 134–140.
Hollands, J.G., Dyre, B.P., 2000. Bias in proportion judgments: the cyclical power model. Psychological Review 107, 500–524.
Indow, T., Togano, K., 1970. On retrieving sequence from long-term memory. Psychological Review 77, 317–331.
Johnsen, E.C., Killworth, P.D., Domini, J., Bernard, H.R., McCarty, C., Shelley, G.A., submitted for publication. Recalling numbers of alters in target subpopulations: an explanatory model for a power law.
Killworth, P.D., Bernard, H.R., 1977. Informant accuracy in social network data II. Human Communication Research 4, 3–18.
Killworth, P.D., Bernard, H.R., 1978. The reverse small-world experiment. Social Networks 1, 159–192.
Killworth, P.D., Johnsen, E.C., McCarty, C., Shelley, G.A., Bernard, H.R., 1998a. A social network approach to estimating seroprevalence in the United States. Social Networks 20, 23–50.
Killworth, P.D., McCarty, C., Bernard, H.R., Shelley, G.A., Johnsen, E.C., 1998b. Estimation of seroprevalence, rape and homelessness in the US using a social network approach. Evaluation Review 22, 289–308.
Laming, D., 1973. Mathematical Psychology. Academic Press, London.
McCarty, C., Killworth, P.D., Bernard, H.R., Johnsen, E.C., Shelley, G.A., 2001. Comparing two methods for estimating network size. Human Organization 60, 28–39.
Myung, I.J., Kim, C., Pitt, M.A., 2000. Toward an explanation of the power law artifact: insights from response surface analysis. Memory and Cognition 28, 832–840.
Shelley, G.A., Bernard, H.R., Killworth, P.D., Johnsen, E.C., McCarty, C., 1995. Who knows your HIV status? What HIV+ patients and their network members know about each other. Social Networks 17, 189–217.
Stevens, S.S., 1971. Neural events and the psychophysical law. Science 170, 1043–1050.
Stevens, S.S., 1975. Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects. Wiley, New York.