The Network Scale-Up Method: Background and Theory

H. Russell Bernard and Christopher McCarty
University of Florida
February, 2009

Suggested citation: Bernard, H. R. and C. McCarty 2009. The Network scale-up method: Background and theory.
+URL where found.

The network scale-up team

- H. Russell Bernard (Univ. of Florida)
- Peter D. Killworth (Southampton Oceanography Centre)†
- Christopher McCarty (Univ. of Florida)
- Eugene Johnsen (UC-Santa Barbara)
- Gene A. Shelley (Georgia State Univ.)

- The network scale-up method was developed by a team of researchers under grants from the U. S. National Science Foundation to H. Russell Bernard and Christopher McCarty at the University of Florida. The method can be applied now to estimating the size of hard-to-count (or impossible-to-count) populations but the method is a work in progress. Each new application provides data for improving the validity and accuracy of the estimates. As with the development of the model, these improvements require the efforts of survey researchers, mathematicians, and ethnographers.

Finding the distribution of the number of people whom people know

- Our objective is to understand the basic components of social structure. One quantity that seems important to us is the number of people whom people know.
  - We call this c

- The network scale-up method was developed in conjunction with our team's research on the rules governing who people know and how they know them. The particular list of people who people come to know in a lifetime may appear random, but the rules governing who we come to know are surely not random. One basic component of social structure is the number of people whom people know. We call this number c.

- This number has a distribution, of course, and it probably changes across societies.

- 1984 Peter D. Killworth, H. R. Bernard, and C. McCarty. Measuring Patterns of Acquaintanceship. *Current Anthropology* 25:381–97.
- 1988 H. Russell Bernard, Peter D. Killworth, Michael J. Evans, Christopher McCarty, and Gene A. Shelley. Studying Social Relations Cross Culturally. *Ethnology* 27:155–79
- 1990 H. Russell Bernard, Peter D. Killworth, Christopher McCarty, Gene A. Shelley, and Scott Robinson. Comparing Four Different Methods for Measuring Personal Social Networks. *Social Networks* 12:179-215

A primitive model

- We can derive this number from an assumption.
  - Let t be the size of a population (e.g. the U.S.), and let e be the size of some subpopulation within it.
  - We assume that the fractional size $p = e/t$ of that subpopulation also applies to any individual's network, other things being equal.
  - That is, everyone's network in a society reflects the distribution of subpopulations in that society.

Some history

- The original network scale-up model was a four-part equation: (1) the event population (called e); (2) the total population (called t) within which e is embedded; (3) the probability, p, that anyone in t knows someone in e; and (4) the number of people whom people know, c. Some history: Bernard was in Mexico City, soon after the earthquake there in the fall of 1985. No one knew how many people had died in that earthquake, but one person told Bernard that "there must be thousands dead, because everyone knows someone who died." We did a random, representative street-intercept survey and found the percentage of people who reported knowing someone who died in the quake. That gave us two parts of the equation. We knew t (Mexico City had around 18 million people at the time) and we knew p. We reasoned that if we knew c, then we could solve for e. This set up our research program on finding, not just the average c in a population, but its distribution. For work on the early development of the model, see:
- 1989 H. Russell Bernard, E. Johnsen, P. Killworth, and S. Robinson. Estimating the Size of an Average Personal Network and of an Event Population. In: *The Small World* , ed. by M. Kochen, 159–75. Norwood, NJ: Ablex Publishing.
- 1990 Peter D. Killworth, Eugene C. Johnsen, H. Russell Bernard, Gene A. Shelley, and Christopher McCarty. Estimating the Size of Personal Networks. *Social Networks* 23:289–312.
- 1991 H. Russell Bernard, P. D. Killworth, E. C. Johnsen, and S. Robinson. Estimating the Size of an Average Personal Network and of an Event Subpopulation: Some Empirical Results. *Social Science Research* 20:109–21.

The scale-up method to estimate *c*

- To test this, we ask a representative sample of people to tell us how many people they know in many sub-populations whose sizes are known:
  - e.g., diabetics, gun dealers, postal workers, women named Nicole, men named Michael

- 1998 Killworth, P.D., E.C. Johnsen, C. McCarty, G.A. Shelley, and H.R. Bernard. A Social Network Approach to Estimating Seroprevalence in the United States. *Social Networks* 20:23-50.
- 1998 P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of Seroprevalence, Rape and Homelessness in the U.S. Using a Social Network Approach. *Evaluation Review* 22:289–308.

- To find c and its distribution, we ask a representative sample of people in t to tell us how many people they know in many subpopulations whose sizes are tracked reliably in public statistics. If our model is working, then we should be able to estimate accurately the size of those same populations. To the extent that we can do that, we have more confidence in our estimates of subpopulations whose sizes are unknown.

Do people answer accurately?

- This works only if people can and do answer our questions accurately and we recognized early on that this was a problem.
- We expect that continued research on this problem will improve the estimates of hard-to-count populations.

- 2006 Killworth, P. D., C. McCarty, E. C. Johnsen, H. R. Bernard, and G. A. Shelley. Investigating the variation of personal network size under unknown error conditions. *Sociological Methods and Research* 35:84-112.

- We had earlier studied the problem of informant accuracy in network data – that is, the extent to which people could report accurately with whom they interacted over various lengths of time. We knew from this research that informant accuracy would threaten the validity of the estimates from the network scale-up model. Just as with c, we expect that research on informant accuracy will result in incremental improvements in the estimates of e.

A maximum likelihood estimate of an individual's network size:

$$c_i = t \bullet \frac{\sum_{j=1}^{L} m_{ij}}{\sum_{j=1}^{L} e_j}$$

where there are L known subpopulations. (Here i is the individual, who knows m_{ij} in subpopulation j.) Network size is (the sum of all the people you say you know in some subpopulations of known size, divided by the total size of those subpopulations) times the population within which the subpopulations are embedded.

Killworth, P. D., C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of Seroprevalence, Rape and Homelessness in the U.S. Using a Social Network Approach. *Evaluation Review* 22:289–308.

- To move beyond the basic model, we need to estimate the size of the network of each respondent in our scale-up surveys. We use a maximum-likelihood method to estimate the size of an individual respondent's network. This method for estimating c is due to Peter Killworth.
- Killworth, P. D., C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen 1998. Estimation of Seroprevalence, Rape and Homelessness in the U.S. Using a Social Network Approach. Evaluation Review 22:289-308.

Estimates of *c* are reliable

- This doesn't deal with the problem of informant accuracy, but the estimates of *c* for the U.S. are very stable.
- Across seven surveys, we consistently find an average network size of 290 (sd 232, median 231).
- And 290 is not an average of averages. It's a repeated finding.

Is 290 is an artifact of the method?

- We test this in three ways.
- (1) Make the estimates using a different method.
- (2) Experiment with parameters and see if the outcome varies in expected ways.
- (3) Compare values of *c* across populations of known relative sizes.

- We were surprised to find that the number 290 was so stable and we tried three different ways to disprove it. (1) We estimated the number using a method of counting that was different from the one we developed initially; (2) We introduced error into the data to see how it affected the outcome; and (3) We compared our results for c across populations of known relative sizes.

Reliability I

- (1) In one survey, we estimated *c* by asking people how many people they know in each of 17 relation categories – people who are in their immediate family, people who are co-workers, people who provide a service – and summing.
- This summation method once again produced a mean for *c* of 290.

- McCarty, C., P. D. Killworth, H. R. Bernard, E. Johnsen, and G. A. Shelley. Comparing Two Methods for Estimating Network Size. *Human Organization* 60:38–39

- In this test, we estimated c by asking respondents to tell us how many people they knew in each of 17 categories: people in their immediate family, people they know from work, people who provide a service, and so on. Once again, our estimate of the average c for the U.S. was 290.
- The summation method is due to McCarty.

- McCarty, C., P. D. Killworth, H. R. Bernard, E. Johnsen, and G. A. Shelley 2001. Comparing Two Methods for Estimating Network Size. *Human Organization* 60:38–39.

Reliability II: Change the data

- (2) In the second reliability check, we introduced errors into the data. We tested whether the changes in our estimates of c conformed to the changes we introduced to the data. We changed reported values at or above 5 to a value of 5 precisely. The mean dropped to 206, a change of 29%.
- We set values of at least 5 to a uniformly distributed random value between 5 and 15. We repeated the random change (5 – 15), but only for large subpopulations (with >1 million).
- The mean increased to 402, a change of 38% -- in the opposite direction.

Reliability III: Survey clergy

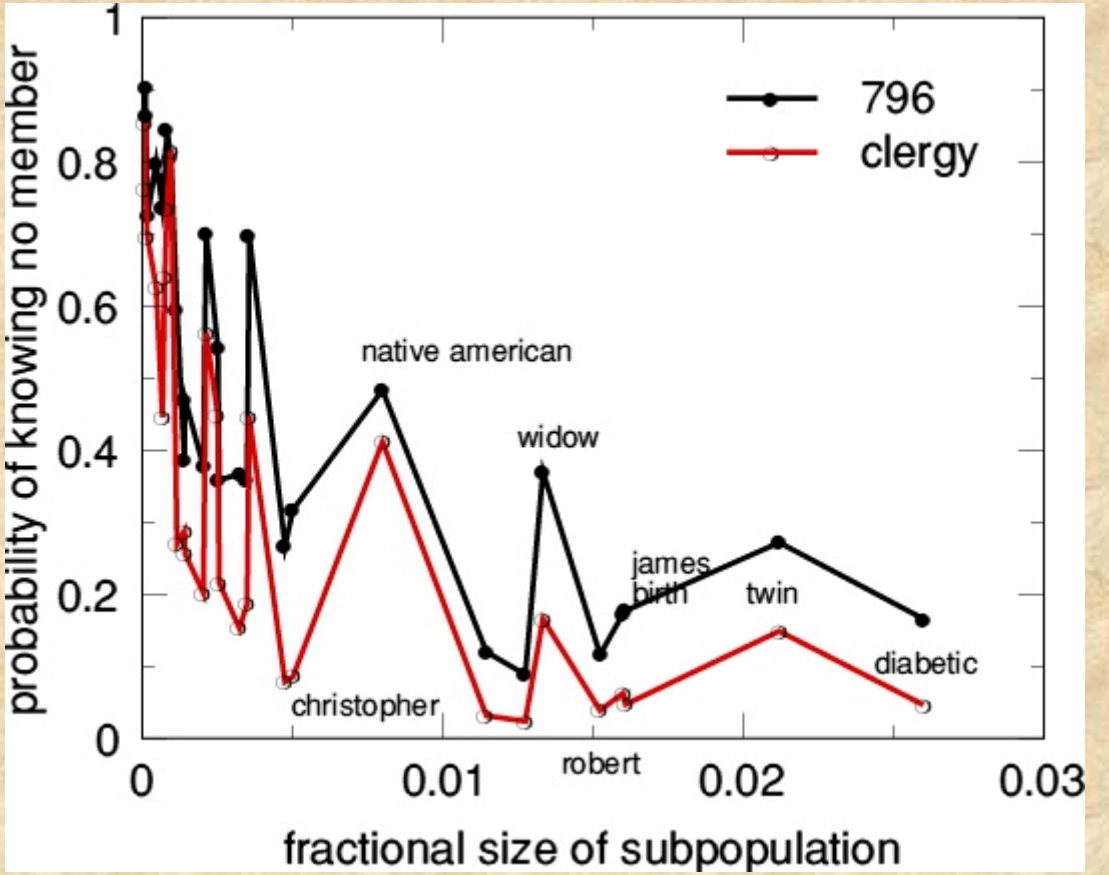
- (3) And in our third test of reliability, we surveyed a sample of 159 clergy – people who are widely thought to have large networks -- and estimated the size of their networks. We used our original scale-up method and the summation method. Here, the two methods produced quite different estimates of c, but both estimates are, as expected, larger than the 290 for the general population.
- Mean *c* = 598 for the scale-up method
- Mean *c* = 948 for the summation method

So, 290 is not a coincidence

- In each case, then, the result of our reliability tests were expected, giving us some confidence in the estimate of c for the U.S.
- 1. Two different methods of counting produce the same result.
- 2. Changing the data produces large changes in the results, and in the expected directions.
- 3. People who are widely thought to have large networks do have large networks.

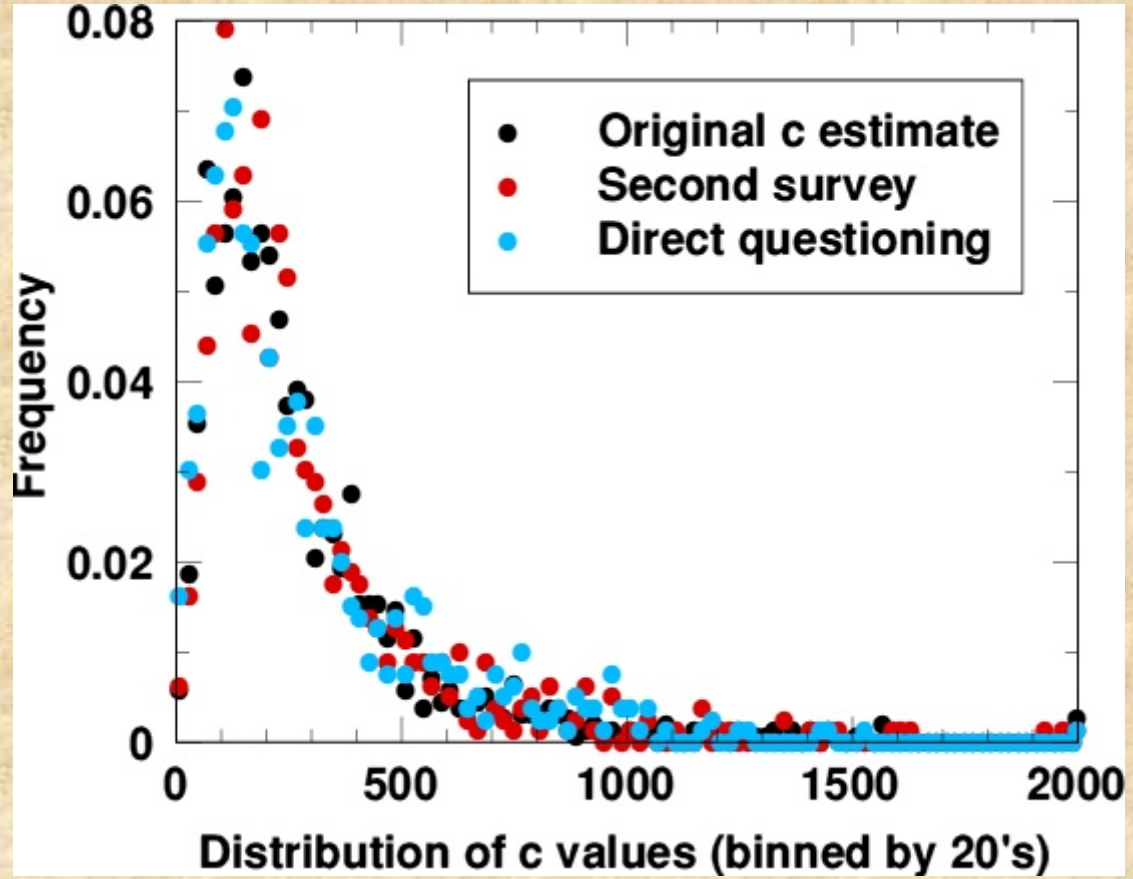
Something is going on

- This next slide shows the probability, for two of our surveys, of knowing no one in each of 29 populations of known size, by the actual size of those populations.
- The two distributions track, except for the expected offset.


probability of knowing no member
fractional size of subpopulation
796
clergy
native american
widow
james
birth
twin
diabetic
christopher
robert
0
0.2
0.4
0.6
0.8
1
0
0.01
0.02
0.03

The distribution of *c*

- Here is the graph of the distribution of network size:


0.08
0.06
0.04
0.02
0
Frequency
0
500
1000
1500
2000
Distribution of c values (binned by 20's)
Original c estimate
Second survey
Direct questioning

Reliability vs. validity

- We are measuring something, and we are doing so reliably, but if our model works, then we ought to be able to use it to estimate the populations whose sizes are <u>not</u> known.
- We can create a maximum likelihood estimate for the size of an unknown subpopulation based on what all respondents told us and our estimates of their network sizes – roughly speaking, inverting the previous formula.

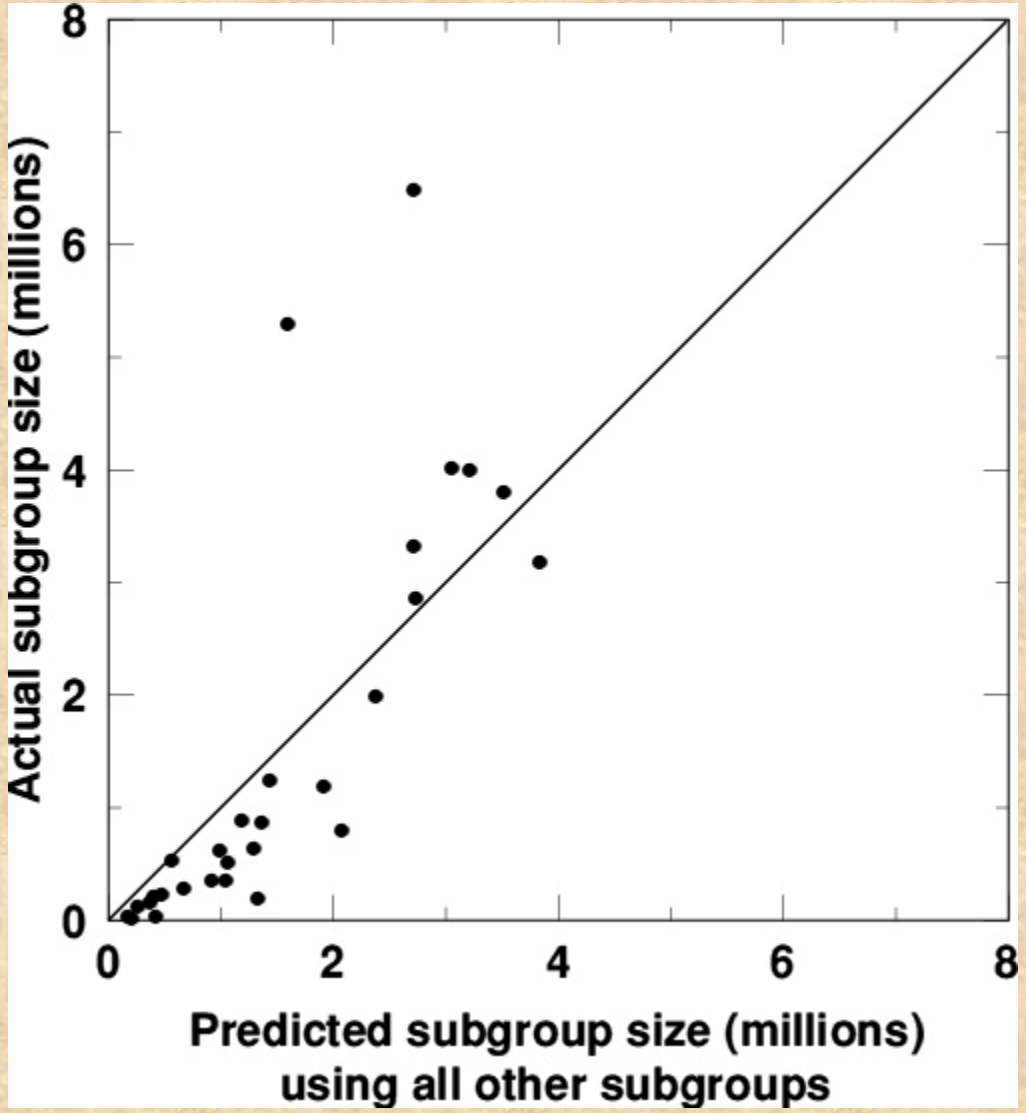
Can we predict what we know?

- We can test this by seeing how well we do on the 29 populations of known size.
- The next slide presents the sequence for applying the method. Details on how to conduct a network scale-up study are in the accompanying presentation.

- The sequence for applying the method is as follows: (1) Ask a representative sample of people how many people they know is a list of populations whose size is tracked in public records. (2) Apply the maximum likelihood measure (in slide 8) to estimate c (the size of the network) of each person in the survey. (3) Estimate the size of each subpopulation. The word “estimate” here means checking the accuracy of the method by comparing what the method predicts against the known-sizes of the populations. The original scale-up method used 29 populations of known size and 3 populations of unknown size. See slide 50 for the results of this test. With 20 known-size populations, there should be sufficient data to check the accuracy of the method for any given test. (4) If the estimates of the known-size populations are accurate, then the estimates of the unknown-size populations may be reasonable. Of course, if alternative estimates are available for the unknown-size populations, then these should be compared to the estimates from the network scale-up method.

- As the next slide shows, the overall result is encouraging, but we don't estimate some known-size populations well.
- Note the two outliers (way above the diagonal) in the next slide. These are twins and diabetics. When these two data points are removed, the correlation rises from 0.79 to 0.94. The problem, of course, is that the outliers represent something that needs to be accounted for in order to improve the accuracy of the method. Just removing the outliers, therefore, improves the correlation but not the method itself.
- Killworth, P.D., C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen 1998. Estimation of Seroprevalence, Rape and Homelessness in the U.S. Using a Social Network Approach. *Evaluation Review* 22:289–308.

r =.79 … rises to .94 without the outliers



- From the previous slide, we see that people tend to overestimate small populations (<2 million -- people named Nicole, people who are undergoing kidney dialysis) and underestimate large ones (>3 million -- people who have a twin sibling, diabetics)
- In other words, when we ask people “how many people do you know who have a twin brother or sister), the answers tend to be overestimates. And conversely for small populations: people tend to underestimate the number of people they know in small populations. A valuable piece of information comes from out of this: In building a network scale-up survey, researchers should shoot for known-size populations that are in range of sizes.

Stigma vs. not newsworthy

- What causes these tendencies to over- and under-estimation? Is it just the size of the subpopulation? Or is it the fact that some things (like being HIV-positive) are stigmatizing.
- In that case, we may know people who are HIV-positive but not know that we know them because they haven't told us about their HIV status.
- But being a twin or a diabetic is neither stigmatizing, nor (as in violent crime) newsworthy. The fact that someone has a twin may simply never come up in conversation, even after decades of knowing someone. From ethnographic evidence, personal information about close co-workers or business associates can take a decade or more to be transmitted ... and in the case of being a twin or a diabetic, may never be transmitted.

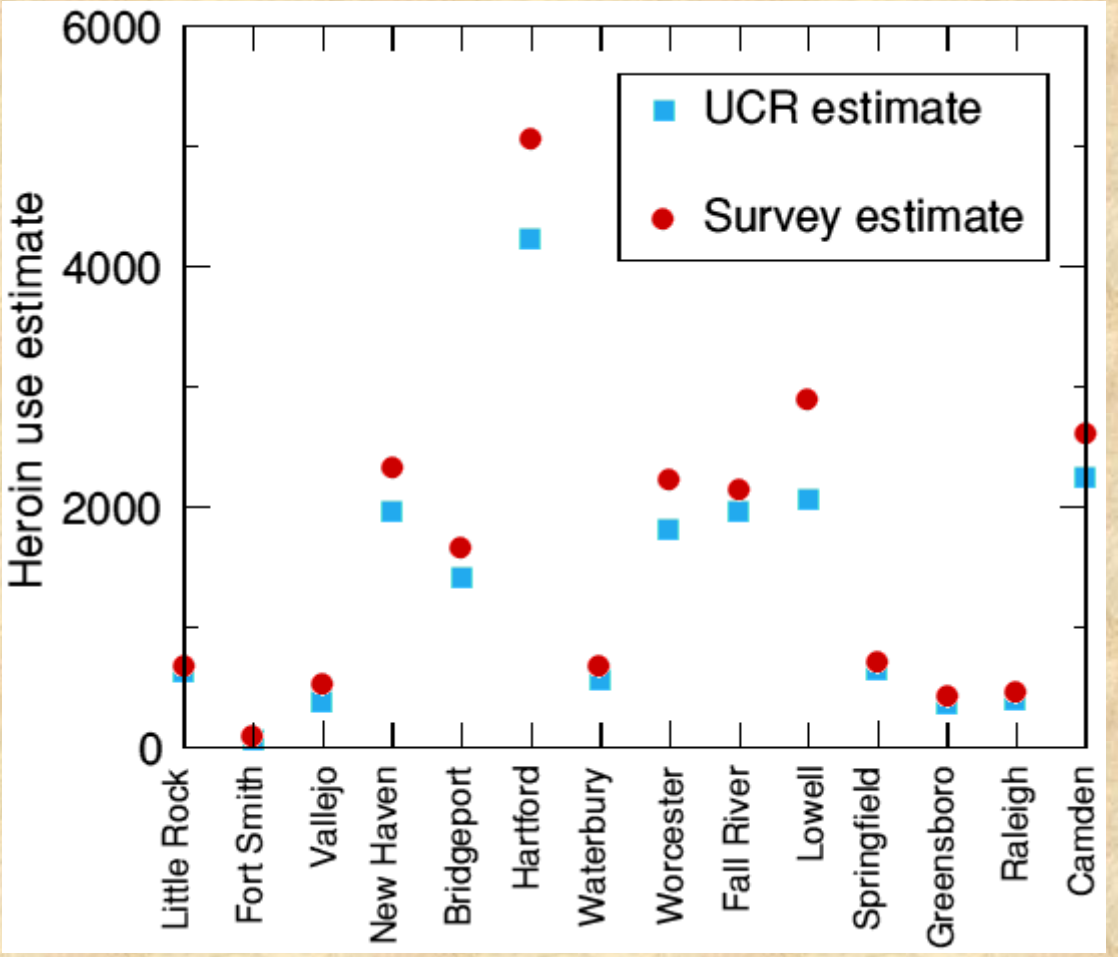
- 1990 Gene Anne Shelley, H. R. Bernard, and P.D. Killworth. Information Flow in Social Networks. *J. of Quantitative Anthropology* 2:201–25.
- 1995 Shelley, G.A., H. R. Bernard, P. D. Killworth, E. C. Johnsen, and C. McCarty. Who Knows Your HIV Status? What HIV+ Patients and Their Network Members Know About Each Other. *Social Networks* , 17, 189-217.
- 2006 Shelley, G. A., P. D. Killworth, H. R. Bernard, C. McCarty, E. C. Johnsen, and R. E. Rice. Who knows your HIV status II: Information propagation within social networks of seropositive people. *Human Organization* 65:430-444.

Another encouraging result

- Charles Kadushin ran a national survey to estimate the prevalence of crimes in 14 cities, large and small, in the U.S.
  - He asked 17,000 people to report the number of people they knew who had been victims of six kinds of crime and the number of people they knew who used heroin regularly.

- 2006 C. Kadushin, P. D. Killworth, H. Russell Bernard, and A. Beveridge. Scale-up methods as applied to estimates of heroin use. *Journal of Drug Issues* 36:417-440.

- Here are the estimates for the number of heroin users in each of the 14 cities, along with the estimates from the UCR.
  - The UCR is the Uniform Crime Report system in the U.S. In this slide, the UCR estimates for heroin use in the 14 cities are the blue squares. The estimates from the network scale-up survey in the same cities are the red dots. For the most part, these estimates are very similar.


Heroin use estimate
6000
4000
2000
0
UCR estimate
Survey estimate
Little Rock
Fort Smith
Vallejo
New Haven
Bridgeport
Hartford
Waterbury
Worcester
Fall River
Lowell
Springfield
Greensboro
Raleigh
Camden

It’s less expensive, but ...

- The fact that we track well with official estimates means only that we have a much, much less expensive way to get at these estimates – not that the estimates are correct.
- And estimates of other crimes in those 14 cities did not track so well.

Reliability, validity, and accuracy

- So, while definitely reliable and perhaps valid, our estimate of network size (and its distribution) is not sufficiently accurate.
- Which raises the question: How can we improve the accuracy of the method?
- There are at least three sources of inaccuracy: transmission effects, barrier effects, and informant reporting.

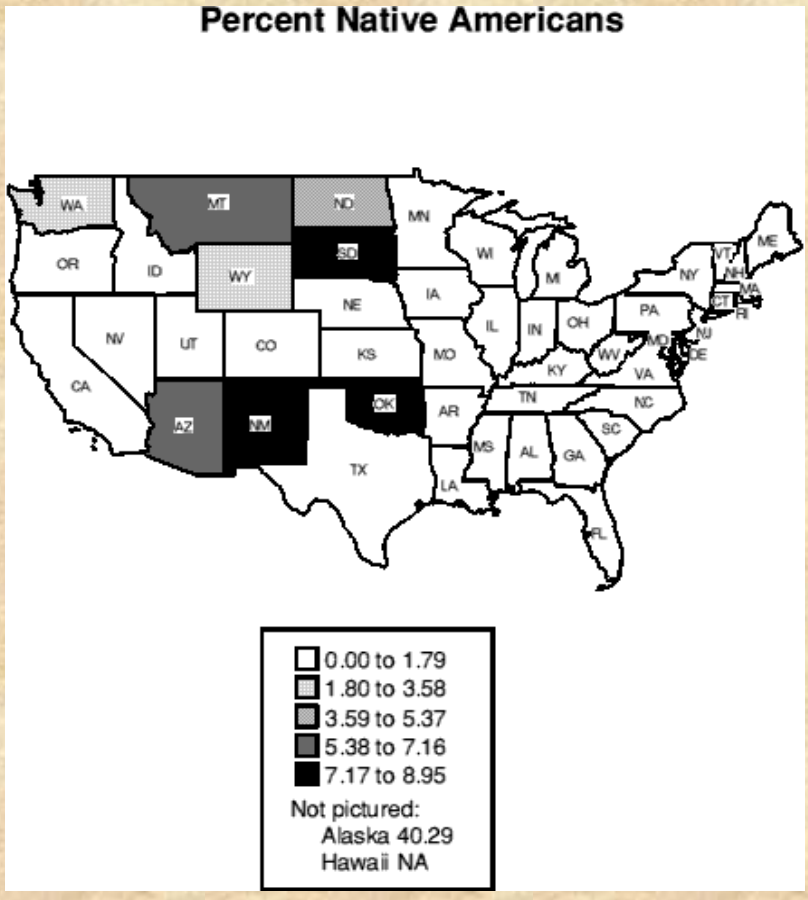
Compromising assumptions

- 1. Transmission effects: Everyone knows everything about everyone they know.
- 2. Barrier effects: Everyone in *t* has an equal chance of knowing someone in *e*.
- 3. Inaccurate recall. People don't recall accurately the number of people they know in the subpopulations we ask them about.
  - The accuracy problem is discussed earlier.

- Transmission effects show up when people do not know that they know something about their friends, family members, and acquaintances. For example, you may know someone you work with every day and not know that she is a member of a particular church or that she is suffering from some chronic illness, and so on. To learn these facts about someone requires that they tell you. Not knowing things about people you know because those people don't tell you is an information transmission problem.
- A lot of information about people, however, is blocked by social and physical barriers (more about this in the next slides). And even when people know things about their network members, they may not dredge up the information when a survey researcher asks about it. Thus, asking "How many people do you know named Michael?" may result in an underestimate because of poor recall or an overestimate because of rounding (that is, people saying to themselves, "well, I can count three Michaels who I know and I must know more than that, so I'll just say five in answer to the question."

Network physical barriers

- There are physical and social barriers to knowing people in various populations. Geography is physical barrier.
- There are more American Indians in Oklahoma than there are in Florida. We expect that people in Oklahoma will, therefore, know more American Indians, on average, than people in Florida do. The next graph shows this.



From our surveys, across the U.S., the correlation between the average number of Native Americans known and the percent of the population that is Native American is 0.58 (p=0.0001).

Network social barriers

- Besides physical barriers, there are social barriers to knowing people in various populations.
- Black people are more likely to be diabetic than are white people and so we expect that black people are more likely to know a diabetic than are white people.
- Men in the U.S. may know more gun dealers than women do.
- Even first names are subject to barrier effects. People in California are more likely to know someone named Carmen than are people in states that have a low fraction of people with Hispanic names.
- Over time, with many applications of the network scale-up method, we should continue to improve the accuracy of network scale-up method.
- We address the barrier effect by using a random, nationally representative sample of respondents.
- However, using the method on specific populations may still lead to incorrect estimates.

The transmission effect

- Transmission bias comes from the fact that some things more difficult than other to know about our acquaintances. Stigmatizing information (for example, a suicide in the family's history) is not shared with everyone one knows. Some things, like having a twin sibling may simply never come up in conversation.
- We study transmission bias by asking people why they do or do not tell their network members various things about themselves.
- We recruited 30 people who were members of one of the known populations used in the network scale-up method.

Interview egos and alters

- To test the transmission effect, we randomly selected male and female first names proportionate to their representation in the 1990 US census.
- We asked 30 people if they knew someone named Michael, someone named Nicole, and so on. We chose the names so we'd have a good distribution – some common names, some less common, and some uncommon. We kept asking people about first names until we got a hit – that is, the respondent said that he or she knew someone with that first name. We did that until we got 25 people in each respondent's network. Then we asked the respondent for some information about each of the 25 network alters.
- With 30 respondents and 25 alters for each respondent, there were 750 alters. We were able to contact 220 of the 750 and ask them things about themselves – things that we had asked the original respondent about them. The next slide shows a summary of results.

Population	% who knew	% who did not know	Respondents	# of alters
Am. Ind.	100	0	2	12
Diabetic	55	45	6	44
Birth in last 12 mos.	93	7	3	27
Gun dealer	92	8	1	12
Member of JC's	58	42	1	12
Dialysis	88	12	5	26
Business in last 12 mos.	75	25	4	16
Postal worker	100	0	1	10
Has twin	88	12	2	24
Widowed <65	97	3	4	38

Findings from the alter study

- From the previous slide, we see that it is much easier to know that someone is a kidney dialysis patient than it is to know that they are a diabetic.
- Diabetes is much less visible.

Some things are easy to get right

- 99% of the respondents in the study knew the marital status of their network members.
- People know how many children 89% of their alters have.
- 98% know the employment status of their alters.

Some things are harder to know

- When asked about the number of siblings a network member has, people say they don't know 52% of the time.
- People say they know the state in which 70% of their network members were born, but only 57% of the reports (ego's and alter's) agree on this.

Some people withdraw

- One source of transmission bias is the fact that people in stigmatized populations withdraw from interacting with people in their networks. This was found first in ethnographic research:
  - Gene Shelley found that people who are HIV+ withdraw from their network in order to limit the number of people who know their HIV status.
- And was later confirmed mathematically.
  - Eugene Johnsen confirmed this by showing that HIV+ people have, on average, networks that are one-third the global average.

- 1995 Johnsen, E. C., H. R. Bernard, P. D. Killworth, G. A. Shelley, and C. McCarty. A Social Network Approach to Corroborating the Number of AIDS/HIV+ Victims in the U.S. *Social Networks* 17:167–87.
- 1995 Shelley, G.A., H. R. Bernard, P. D. Killworth, E. C. Johnsen, and C. McCarty. Who Knows Your HIV Status? What HIV+ Patients and Their Network Members Know About Each Other. *Social Networks* , 17, 189-217.
- 2006 Shelley, G. A., P. D. Killworth, H. R. Bernard, C. McCarty, E. C. Johnsen, and R. E. Rice. Who knows your HIV status II: Information propagation within social networks of seropositive people. *Human Organization* 65:430-444.

Can we account for these errors?

- Can we use this kind of information to tweak the model?
- We tried to develop weightings for classes of characteristics about subpopulations ... classes like “things that carry a strong stigma” and “things that carry a moderate stigma” and “things that just don’t come up in conversation.”
  - While we found some signals like these, we don’t know how to know whether two populations require the same weighting.

- So far, we have not been able to improve the model using the information we've collected about transmission and barrier effects.
- We turned to modeling the errors, but with more empirical tests of the model we expect further improvements.

Returning to using this to scale up

- Finally, let's return to the application of the network scale-up model. In the mid-1990s, when we first tested the model, did random digit dialing surveys of 1554 adults in the U.S.
- We estimated the population of people who are HIV-positive, people who are homeless, and women in the U.S. who had been raped in the previous 12 months. The next slide shows our estimates for those populations.

- RDD telephone survey of 1554 adults in the U.S. in 1994.
- Seroprevalence: 800,000 ± 43,000;
- Homeless: 526,000 ± 35,000;
- Women raped in the last 12 months: 194,000 ± 21,000.
  - These are all close to other estimates made with various enumeration or surveillance methods. For details, see article below.

- 1998 P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of Seroprevalence, Rape and Homelessness in the U.S. Using a Social Network Approach. *Evaluation Review* 22:289–308.

Next steps

- We can’t claim that the network scale-up method produces the most accurate estimates of hard-to-count and uncountable populations.
- However, as our knowledge improves about transmission and barrier effects, the estimates improve – and by a known amount for the known-size populations.
- Our goal is incremental improvement in estimating the size of any population.
- As it has in the past, this will take continued collaborative effort among ethnographers, modelers, and survey researchers.