Learning about networks using sampling*

Andrew Gelman

15 Jan 2016

In survey research we are often interested in hidden, hard to reach, and marginalized populations. Even when potential respondents are sitting out in plain sight, it can be difficult to gather a representative sample, to persuade sampled people to respond to a survey, and to get accurate responses. We are seeing this now in early polls for the upcoming election campaign: even setting aside the difficulties or working with nonrandom, biased samples, it's not always clear how to interpret presidential preferences, months before the first primaries and nearly a year in advance of the general election.

Often what we're interested in is not just hard-to-reach groups but how they relate to the larger society, as Steve Thompson discussed the study of the transmission of disease.

As political scientists, we are interested in groups not merely themselves but in what we call their *penumbra*, the number of family members, friends, and acquaintances of people in the group; see Figure 1. The size and the shape of the penumbra can relate to the political salience and influence of a social group.

We studied penumbras using two surveys administered by YouGov on a panel 12 months apart, with about 3,000 respondents in wave 1 and 2,106 re-interviewed in wave 2. We asked about penumbra membership in 14 social groups and 8 names, and attitude questions on 12 related policies (Margalit and Gelman, 2016). We have put similar questions on the General Social Survey (DiPrete et al., 2011).

Penumbras are typically much larger than the group; for example, less than 1% of American adults are in the active military but nearly half of respondents know someone in the service. This represents some combination of social networks and uncertainty about classification; for example, you might be counting a friend who is no longer in active service, and this would count in the penumbra but not in the group size. The gay/lesbian penumbra is particularly large, with nearly three-quarters of respondents reporting that they know someone among this group which is generally estimated to comprise about 3% of the population.

Recent immigrants and gays/lesbians have about the same numbers in the United States, but the penumbra of gays and lesbians is much larger, which could have political repercussions, as suggested by the rapid gain in acceptance of same-sex marriage in recent years.

At the other extreme, very few people report knowing someone who had an abortion in the past five years, despite there being millions of women who fall into this category. This can arrive from a classification or transmission error in that women who have had abortions do not always reveal this fact to their acquaintances (Cowan, 2013), and it can also be considered as part of the definition of penumbra in that if you do not realize that a friend falls into a particular group, that affects how the group is perceived.

This is a common issue in survey research, that we are measuring some underlying reality, but perceptions are also important in themselves. Economists care about the volume of business transactions and also consumer confidence; criminologist measure victimization rates and also perceptions of public safety; and so on.

Now I want to speak more generally about network sampling. Sometimes we do network sampling because we want to, other times because we have to. We can have network structure in the sampling, or network structure in the population. Steve Thompson talked a lot about network structure in the

*Discussion of the 2015 Morris Hansen lecture by Steve Thompson. To appear in the *Journal of Survey Statistics and Methodology*.

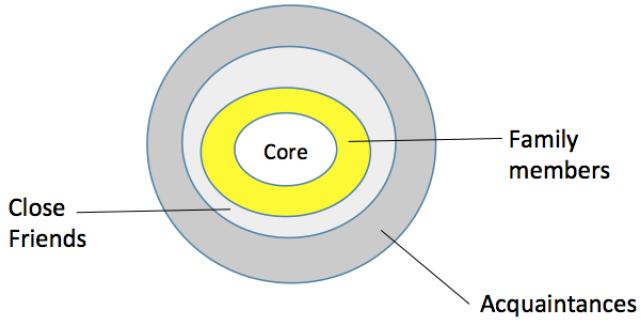

Figure 1: *Sketch of the social penumbra of a group in the population.*

population, and Mark Handcock in his discussion talked about network structure in the sampling, about survey methods that sampled along paths in the social network.

When you use network sampling to learn about the general population, often you'd like to do network sampling but you can't: that's the usual paradigm in respondent-driven sampling (Heckathorn, 1997). There is a general model-based solution to these problems, which is to poststratify: you take your sample and adjust it to match the population. You want to adjust for variables that matter in the sample and also matter in the population.

If you do a network sample, two key variables are *closeness to the intake point* and *gregariousness* or network size. Closeness to the intake is important because a sample that goes along a network will tend, by its nature, to overrepresent people near the seeding point of the sample, and gregariousness matters because, if a potential respondent knows more people, he or she will be more likely to be referred. Personal network size plays a role similar to "number of telephone lines" in a traditional phone survey.

If you take a network sample, you can measure the degree or gregariousness of each respondent, and we can also record the distance that he or she is from the starting point—how many links it took to reach that person. Then you can use statistical methods to estimate the distribution of these variables in the population, and you can adjust using the basic poststratification formula,

$$\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j},$$

where θ is your population average of interest (for example, the proportion of people who would respond Yes to some survey question, if asked); the j's represent poststratification cells (in the case of network sampling adjustment, these cells would be determined by gregariousness, closeness to intake points, and probably some demographic variables such as age, sex, ethnicity, and so forth; the N_J's are the sizes of the cells in the population (which themselves would have to be estimated in some way, given that there is no national roster of people characterized by gregariousness); and the θ_j's are the population averages within each cell.

Implementing such corrections can take work. When we are doing poststratification we will have a large number of cells. The logic of survey adjustment is that as surveys become worse and worse we must adjust for more and more variables. Traditionally we adjust by raking, but there's a limit to how effective raking can be with granularity of data, which motivates Bayesian and multilevel models (Gelman, 2007).

Learning about network structure itself, though, can be a challenge. The fundamental difficulty of network sampling is that a small sample of a network doesn't look like a network itself.

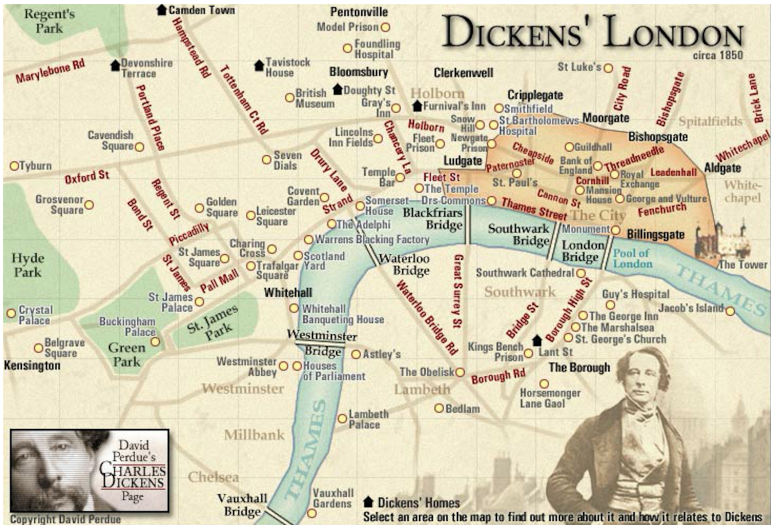

Figure 2: *London in the novels of Charles Dickens was a small, close-knit community. The statistics of sampling from networks can explain why the author felt it necessary to include so many coincidences in his story. Image from David Perdue's Charles Dickens Page,* `http://charlesdickenspage.com/`.

In traditional survey research we have been spoiled. If you work with atomistic data structures, a small sample looks like a little bit of the population. But a small sample of a network doesn't look like the whole. For example, if you take a network and randomly sample some nodes, and then look at the network of all the edges connecting these nodes, you'll get something much more sparse than the original. For example, suppose Alice knows Bob who knows Cassie who knows Damien, but Alice does not happen to know Damien directly. If only Alice and Damien are selected, they will appear to be disconnected because the missing links are not in the sample.

This brings us to a paradox of literature. Charles Dickens, like Tom Wolfe more recently, was celebrated for his novels that reconstructed an entire society, from high to low, in miniature. But Dickens is also notorious for his coincidences: his characters all seem very real but they're always running into each other on the street (as illustrated in Figure 2) or interacting with each other in strange ways, or it turns out that somebody is somebody else's uncle. How could this be, that Dickens's world was so lifelike in some ways but filled with these unnatural coincidences?

My contention is that Dickens was coming up with his best solution to an unsolvable problems, which is to reproduce a network given a small sample. What is a representative sample of a network? If London has a million people and I take a sample of 100, what will their network look like? It will look diffuse and atomized because of all those missing connections. The network of this sample of 100 doesn't look anything like the larger network of Londoners, any more than a disconnected set of human cells would look like a little person.

So to construct something with realistic network properties, Dickens had to artificially fill in the network, to create the structure that would represent the interactions in society. You can't make a flat map of the world that captures the shape of a globe; any projection makes compromises. Similarly you can't take a sample of people and capture all its network properties, even in expectation: if we want the network density to be correct, we need to add in links, "coincidences" as it were. The problem is, we're not used to thinking this way because with atomized analysis, we really can create samples that are basically representative of the population. With networks you can't.

It's not all bad news, though, because we can also use networks to learn about social structure

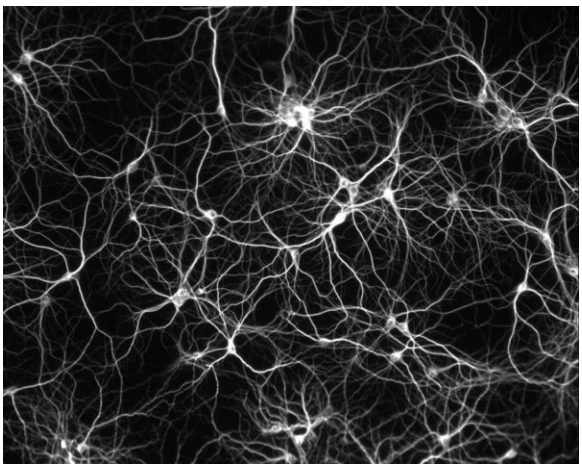

Figure 3: *Fractal structure. This is an image of neurons (from the Fractal Foundation,* `http://fractalfoundation.org/OFC/OFC-1-6.html`*), but the point is to illustrate the idea that a multistage fractal sample can include details at many different levels at once.*

from non-network samples. For example, certain people are hard to reach in a survey: people in prison, school, and other institutional settings. We can use networks to learn about such people: just survey the general population and ask, "How many prisoners do you know?" That simple question tells us a lot about the social network of prisoners—you can learn from the responses what are the demographics of people who know prisoners, what are their attitudes, and so on. And you can also ask them about the prisoners whom they know.

What we have to do, in the Federal statistical agencies and elsewhere, is to go beyond the individualistic approach to survey sampling. Traditionally you ask a person about themselves, maybe about their family, and that's it, as if each respondent or each household lives in a little bubble. There is such a thing as society, and we can ask people all sorts of things about who they know, what their friends do, and so forth. We can ask whatever you want! We can also use data from other sources, Facebook, cell phone records, whatever, but let's not forget the power of direct questions.

There's another idea which I call fractal sampling. When you do a survey, you want to learn at all levels. For example, if you're studying politics, you'll want to know what's happening nationally, you'll want a nationally representative sample. But you'll also want to know what's happening at the state level, the city level, and the neighborhood level. You can't expect to get good estimates for all the neighborhoods in the country or all the cities or even all the states, but you'll want *some* information at all these levels. That's what fractal sampling is all about.

Usually we take multistage cluster sampling because simple random sampling would be too cumbersome–in a face-to-face survey, you wouldn't want to have to parachute interviewers into 1500 randomly-selected locations around the United States—but here I'm arguing that, even if you could do this sort of idealized independent sampling, you'd still want a cluster sample so that you can learn something about social structures at different levels of aggregation.

You can't take a sample of 1000 or even 10,000 Americans and learn anything about local communities—remember Charles Dickens, sampled networks don't look like real networks, and all that? Instead we need to do a fractal sample: we sample states at random, in some states we sample some cities, in some cities we sample some neighborhoods, and so on. We want a multistage cluster sample to be able to make inferences at these different levels.

We can also think about fractal sampling in time. For example in a food consumption survey,

instead of asking people every day what they eat, or asking them every two weeks or every month, we ask people sporadically, at different intervals, sometimes asking several days in a row to learn about short-term eating patterns, and surveying other people far apart in time to get a sense of longer-scale variation.

Such fractal surveys might not be cheap, but on the other hand they take advantage of features such as multistage sampling and irregular sampling which are often considered as problems rather than opportunities in sampling.

In conclusion, I think we've been living in an artificial world which one might call atomistic sampling. I appreciate Steve Thompson's and Mark Handock's work in that they move us toward a closer match between statistics and survey research, on one hand, and the social structure we are studying, on the other. This research is challenging but it has real payoffs. We can use statistical modeling tools to learn from network samples and also about network structure.

References

Cowan, S. K. (2013). Secrets and social influence. Ph.D. thesis, Department of Sociology, University of California, Berkeley.

DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J., and Zheng, T. (2011). Segregation in social networks based on acquaintanceship and trust. *American Journal of Sociology* **116**, 1234–1283.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science* **22**, 153–164.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems* **44**, 174–199.

Margalit, Y., and Gelman, A., (2016). The political impact of social penumbras. In preparation.