Sampling a hidden population without a sampling frame: A practical application of Network Sampling with Memory*

Ted Mouw[1]
Ashton Verdery[1]
M. Giovanna Merli[2]
Jing Li[2]
Jennifer Shen[2]

Abstract

Mouw and Verdery (2012) propose a new method for sampling hidden populations, "Network Sampling with Memory" (NSM), which collects information on network members from the survey instrument to uncover the sampling frame for the target population. They show that NSM yields statistical estimates that are on average 98.5% more efficient than other popular approaches. Here, we present a practical application of NSM that reduces the cost of data collection by collecting contact information on up to three referrals from the current respondent, which eliminates the need to re-contact prior respondents to ask for referrals. We test this modification using simulated sampling on 215 school and university social networks. In addition, we report results from a pilot study using NSM, the 2013 Chinese African Health Study (CAHS) which sampled Chinese immigrants living in Tanzania, and we provide a step-by-step description of how to conduct an NSM-based survey in the field.

Acknowledgements

* We thank Mason Porter for providing access to the Facebook data set we use. This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

Affiliations

1. Department of Sociology, University of North Carolina at Chapel Hill
2. School of Public Policy, Duke University

Sampling a hidden population without a sampling frame: A practical application of Network Sampling with Memory

Although sampling methods that use a list of members of the target population, such as simple random sampling and multi-stage cluster sampling, are the backbone of modern survey techniques, sometimes a researcher wants to sample from a hidden or rare population where a list of eligible respondents does not exist and the use of screening questions to identify eligible members of the target population is not practical (Kalton and Anderson 1986; Sudman and Kalton 1986). The popularity of the Respondent Driven Sampling (RDS) method, which is a modified version of snowball sampling where respondents themselves recruit the next wave of respondents, attests to how much demand there is for survey techniques that can reach hidden populations in situations where no adequate sampling frame exists. Since its creation in 1997 (cf. Heckathorn 1997), for example, the National Institutes of Health has awarded a total of $150 million to 374 different research projects that list variants of "Respondent Driven Sampling" in the keywords (NIH 2013). Because it places the burden of identifying and recruiting additional sample participants on the people expected to know them best – other members of the population – RDS has been shown able to generate large, diverse, and time- and cost-efficient samples (cf. Abdul-Qadar 2006). In addition to its sampling properties, RDS claims a compelling statistical promise: it can provide asymptotically unbiased mean estimates (Volz and Heckathorn 2007). This inferential claim stems from the properties of random walks on graphs (e.g., a social network), where the probability of sampling a particular node (e.g., a person) is a function of the number of ties which that person has in the network (Lovasz 1993). These properties provide a set of general conditions under which a random walk based sampling method may yield unbiased estimates when cases are reweighted according to the inverse number of ties they have to other members of the target population.

Despite the popularity of RDS, a critical question concerns the precision of its estimates, even when all of its assumptions are met. An easy way to operationalize the precision of a sampling method is to calculate its design effect ("DE"): the ratio of the method's sampling variance to the theoretical sampling variance of simple random samples of the same size in the same population (Goel and Salganik 2009). Goel and Salganik (2010) and Mouw and Verdery (2012) use simulated sampling on observed network data to show that RDS often displays very high DEs, even on networks that are highly connected and fully meet the assumptions required for RDS. Mouw and Verdery (2012:234), for example, find that the average DE of estimates of the racial composition of students in 115 schools from Add Health data for RDS samples of 500 drawn with replacement is 47.8, which means that while the estimated proportion may be unbiased on average, the results from any particular sample are likely to be highly inaccurate.

Mouw and Verdery (2012) show that it is possible to increase the precision of sampling from a hidden population by collecting network information as part of the survey. They build on recent

work in computer science about sampling the internet efficiently and propose a new method, "Network Sampling with Memory" (NSM) that uses information on network members from the survey instrument to uncover the sampling frame of the target population. The basic idea of NSM is intuitive: the network questions in the survey contain partial identifying information on respondents' contacts such as basic demographics, the first name and the last four digits of a cell phone number (cf. Dombrowski et al. [2012] for a similar data collection technique). Mouw and Verdery's innovation is to use this information to generate network rosters from the completed surveys, which can be used as a list of revealed members of the target population. When a respondent nominates a new, previously un-nominated, member of the population, they are added to the list of potential respondents. Gradually, as the survey continues, the list of nominated individuals begins to approach the underlying sampling frame of the population that is being studied.

NSM uses two sampling modes: Search and List. The Search mode seeks to discover the network as quickly as possible and thus uses the network data to identify "bridge ties" to unexplored clusters of the network. It does this by looking at the number of times each respondent's contacts have been nominated by other respondents. While a respondent in a heavily sampled cluster will have contacts who have been nominated by many other respondents, a respondent at the unexplored end of a bridge between clusters will have contacts who have not been nominated by anyone else. By increasing the probability of sampling bridge ties, the Search mode pushes the sampling algorithm to explore the frontier of the current network. The List mode, in contrast, samples with replacement from the accumulated list of nominated individuals. As the number of nominated individuals approaches the true size of the population, the sampling frame is revealed, and the List mode behaves similar to simple random sampling (SRS). NSM uses the Search mode to explore the network at the beginning of the survey, and the List mode to sample evenly from the population once the network has been explored. Design based weights are used when the sample is complete to account for the likelihoods of being sampled in either of the two modes. Mouw and Verdery (2012) report that simulated sampling with NSM on 215 social networks from Add Health and Facebook results in DEs of about 1.16, or 98.5% lower than the corresponding DEs that they found with RDS on the same networks.

Despite the theoretical promise of NSM, it is still not a "field ready" method. In this paper we consider three issues that may make NSM more viable as a practical means of sampling from a population connected via a social network. First, we test a variant of NSM that collects contact information (in addition to the partial identifying information) on up to 3 referrals for each respondent. This avoids having to re-contact previous respondents in order to get referrals to individuals whom they nominated. We call this "forward NSM" (fNSM) because the interview team never has to go back to earlier respondents, although the sample will jump around to different parts of the network. Using simulated sampling on 215 networks from Add Health and

Facebook, we show that fNSM results in DE of about 2.4, which, although it is worse than NSM, is still dramatically better than RDS on these networks.

The second innovation we test in this paper is to offer diagnostic guidelines that will help researchers know when they have collected a large enough sample. To do this, we test the effects of different stopping rules. One of the advantages of NSM is that the size of the target population can be estimated from the network data using the capture-recapture method (Dombrowski et al. 2012). Using simulated sampling, we show that the capture-recapture method provides an accurate estimate of the true population size after about 100 interviews in all 215 networks we analyze. Then we test the impact, on the DE, of stopping the survey before all of the nodes in the network have been nominated. We argue that one of the advantages is that the sampled network data can be used to provide feedback to the researcher on the accuracy of resulting sample.

Finally, we report on a pilot study using NSM in the field, the 2013 Chinese African Health Study (CAHS), which sampled immigrant Chinese living in Tanzania. As part of our test of the viability of NSM in the field, we provide a step-by-step description of carrying out an NSM-based survey in practice. It is our hope that these field reports will be useful for researchers thinking of implementing an NSM.

The CAHS survey was a pilot study with two parallel goals. First, it seeks to provide a preliminary characterization of Chinese migration to sub-Saharan Africa, the social network ties that exist in such migrant communities, and the implications of such migration patterns for migrants' health outcomes. Second, it was designed to test the viability of implementing NSM in the field. The CAHS survey consisted of 147 interviews of Chinese immigrants, and the combined network rosters resulted in a total of 898 uniquely identified individuals and a total of 1,211 nominations. There were a total of 96 refusals for a response rate of 60.4%. Figure 1 shows a graph of the social network of this sample. The nodes are color coded by the individual's home province, which provides a visual depiction of the degree of clustering by origin region.

For this paper, we focus on the second goal of CAHS. As a pilot study, the 2013 CAHS sought to show that it is possible to use NSM in the field, not to complete a full sample of the network. Nevertheless, a visual inspection of Figure 1 illustrates several potential benefits of NSM in sampling from a hidden population. First, after 147 interviews, the accumulated roster of 898 nominated individuals provides a list of potential respondents drawn from the target population. Second, unlike snowball-sampling based techniques like RDS, which draw the next wave of respondents from those recruited by current respondents, NSM jumps around in the network as it has been revealed up to that point. This allows NSM to sample the network more efficiently. In Figure 1, the region at the center of the network represents a cluster of nodes that have been

relatively oversampled after 147 interviews. The NSM "Search" mode, which uses the current network data to look for bridge ties to unexplored parts of the network, would place a higher sampling probability on the nodes on the edges of the network in Figure 1, which increases the chance of discovering previously unknown clusters of individuals. Finally, the social network data gathered by NSM provides researchers the opportunity to test novel hypotheses about social network structure and effects. In the case of the substantive area of interest to CAHS – migration – it allows us to examine the structure and function of migrant social networks.

NSM is not intended as a replacement for conventional methods of survey sampling when a sampling frame exists. Nonetheless, there are situations with important substantive and policy interest where conventional sampling frames do not exist. In this paper we argue that if network data is collected as part of the survey then the use of a network-based sampling strategy, such as NSM, can result in both gains in efficiency and precision of the resulting estimates, as well as provide information to the researcher about the complexity of the underlying network and what an adequate stopping point for the survey might be. By considering how to make NSM more practical for use in the field, this paper will help demographic researchers understand new populations and questions.

References

Abdul-Quader, Abu S., Douglas D. Heckathorn, Keith Sabin, and Tobi Saidel. "Implementation and Analysis of Respondent Driven Sampling: Lessons Learned from the Field." Journal of Urban Health 83, no. 1 (November 1, 2006): 1–5. doi:10.1007/s11524-006-9108-8.

Dombrowski, Kirk, Bilal Khan, Travis Wendel, Katherine McLean, Evan Misshula, and Ric Curtis. "Estimating the Size of the Methamphetamine-Using Population in New York City Using Network Sampling Techniques." Advances in Applied Sociology 2, no. 4 (2012): 245–252.

Goel, Sharad, and Matthew J Salganik. "Assessing Respondent-driven Sampling." Proceedings of the National Academy of Sciences 107, no. 15 (2010): 6743–6747.

Goel, Sharad, and Matthew J. Salganik. "Respondent-driven Sampling as Markov Chain Monte Carlo." Statistics in Medicine 28, no. 17 (2009): 2202–2229. doi:10.1002/sim.3613.

Heckathorn, Douglas D. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." Social Problems 44, no. 2 (May 1, 1997): 174–199. doi:10.2307/3096941.

Kalton, Graham, and Dallas W. Anderson. "Sampling Rare Populations." Journal of the Royal Statistical Society. Series A (General) 149, no. 1 (January 1, 1986): 65–82. doi:10.2307/2981886.

Kalton, Graham, and Dallas W. Anderson. "Sampling Rare Populations." Journal of the Royal Statistical Society. Series A (General) 149, no. 1 (January 1, 1986): 65–82. doi:10.2307/2981886.

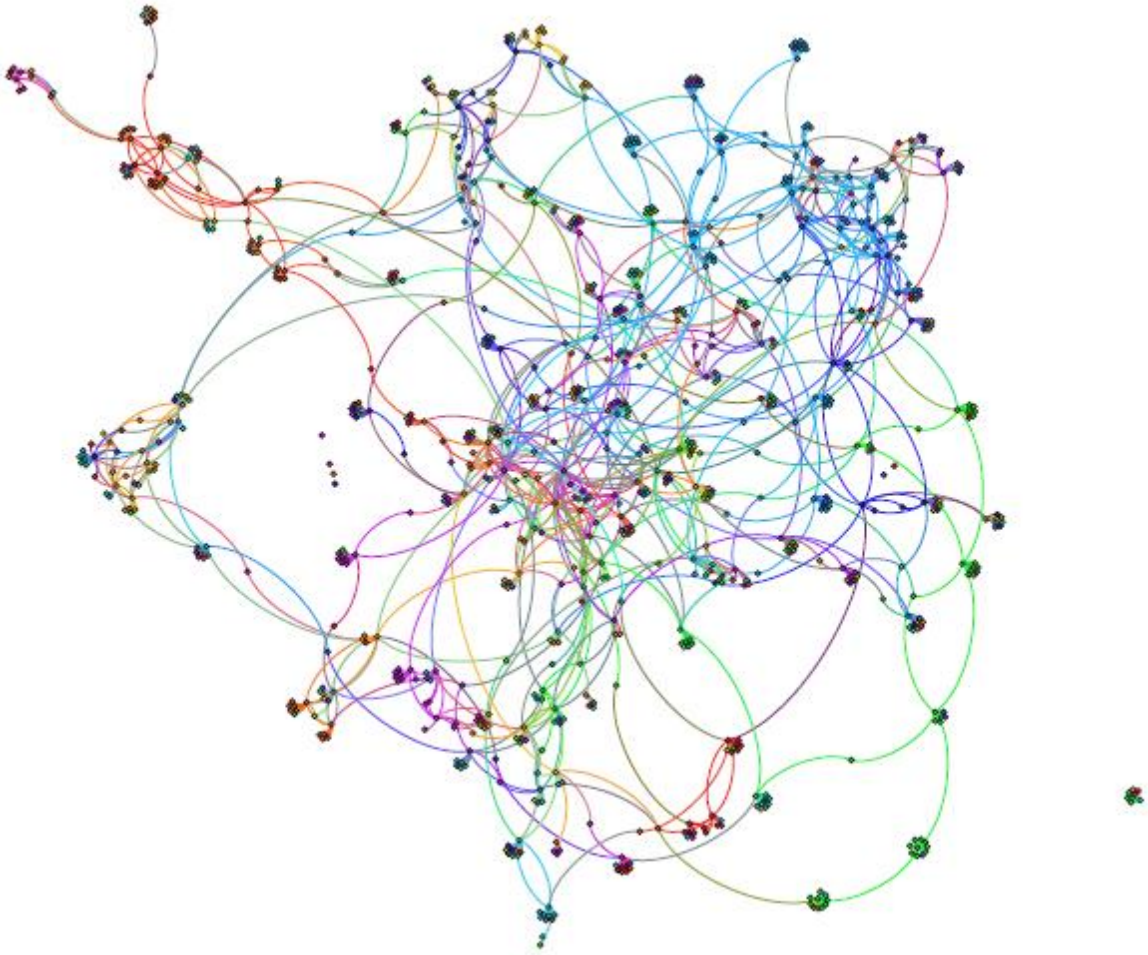
Lovász, László. "Random Walks on Graphs: A Survey." Combinatorics, Paul Erdos Is Eighty 2, no. 1 (1993): 1–46.

Mouw, Ted, and Ashton M. Verdery. "Network Sampling with Memory A Proposal for More Efficient Sampling from Social Networks." Sociological Methodology 42, no. 1 (August 1, 2012): 206–256. doi:10.1177/0081175012461248.

Sudman, Seymour, and Graham Kalton. "New Developments in the Sampling of Special Populations." Annual Review of Sociology 12 (January 1, 1986): 401–429. doi:10.2307/2083209.

Volz, Erik, and Douglas D. Heckathorn. "Probability Based Estimation Theory for Respondent Driven Sampling." Journal of Official Statistics 24, no. 1 (2008): 79.

Figure 1. 2013 CAHS network.



Notes: Node and edge colors show the origin province of Chinese immigrants in Tanzania.