**Title**

Discussion of “Adaptive and Network Sampling for Inference and Interventions in Changing Populations” by Steven K. Thompson

**Permalink**

https://escholarship.org/uc/item/2zh5b0gr

**Journal**

Journal of Survey Statistics and Methodology, 5(1)

**ISSN**

2325-0984

**Author**

Handcock, Mark S

**Publication Date**

2017-03-01

**DOI**

10.1093/jssam/smw040

Peer reviewed

# Discussion of
# "Adaptive and Network Sampling in Changing Populations"

# 2015 Morris Hansen Lecture
# given by Steve K. Thompson

Mark S. Handcock*

January 27, 2017

The use of adaptive sampling designs has increased the range and efficiency[1] of conventional designs. Steve Thompson has been a leader in the development of adaptive sampling designs, and his Morris Hansen Lecture introduces more novel designs and methods for analyzing them. Over the years, Steve has made many contributions to survey sampling, including his books (**??**) and many research papers. All are distinguished by his exemplary writing and clarity of ideas. One stimulus for this is his background in environmental sampling and especially ecology. In these fields, the underlying mobile, interacting and structured populations prompted innovations in design to encompass them. In particular, Steve brought a spatial perspective to sampling that recognized the importance of geography and position as organizing principles in the populations. The designs themselves came to reflect and exploit this spatial dimension to go beyond what was possible using standard approaches. The move from spatial to networked populations is a natural one, as the latter are also characterized by complex interacting structure.

Network sampling methods have been very successful in improving our understanding of hard-to-survey or otherwise "hidden" populations. These populations are characterized by the difficulty in survey sampling from them

*Professor of Statistics, Department of Statistics, University of California, Los Angeles, CA 90095-1554 (E-mail: *handcock@ucla.edu*).

[1] This work was supported by the National Science Foundation (NSF, grant numbers MMS-0851555, SES-1357619, IIS-1546259), the National Institute of Child Health and Human Development of the National Institutes of Health (NICHD, grant numbers R21HD063000, R21HD075714 and R24-HD041022) and the Office of Naval Research (grant number N00014-08-1-1015).

using standard probability methods. Typically, a sampling frame for the target population is not available, and its members are rare or stigmatized in the larger population so that it is prohibitively expensive to contact them through the available frames. Examples of such populations in a behavioral and social setting include injection drug users, men who have sex with men, and female sex workers. Examples in non-epidemiological setting include recent immigrants, unregulated workers and the self-employed. Hard-to-survey populations are under-served by current sampling methodologies mainly due to the lack of practical alternatives to address these methodological difficulties.

The use of network information in studies has a long history. The earliest systematic work appears to date to the 1940s from the Columbia Bureau of Applied Social Research, lead by Paul Lazarsfeld. The Bureau became interested in the empirical study of personal influence via media and this lead to the study of opinion leaders and followers (**?**). Standard sampling of individuals was regarded as ineffective as such pairs were seldom selected in the sample (**?**, pp. 49-50). To increase the efficiency of the design, Robert Merton constructed pairs by asking already sampled individuals to name the people who influenced them. From these a second wave of influential people were interviewed as a "snowball sample" (**?**). **?** provides a fascinating history of the Bureau that is relevant to our reemerging interest in the study of influence via social media.

Following this, a graduate student working at the Bureau, Martin Trow, studied in his dissertation the support for anti-democratic popular movements in the US. To do this he conducted a demographic study of the men in Bennington, Vermont in 1954 with particular focus on their support for the right-wing demagogue Senator Joseph McCarthy. Trow conducted a snowball sample over the friendship networks of the men starting from "arbitrarily chosen lists of employees and occupational groups." (**?**, p. 297). He is very clear that this does not produce a representative sample, and goes on to provide a discussion of the issues with network sampling that is still relevant today (**?**, pp. 290-295). He surmises: "The resulting sample, while not meant to be representative of any specific population, nevertheless includes representatives of all the important occupational groups, ..." See also the insightful and still relevant **?**.

The adaptive link-tracing network sampling methods advocated by Professor Thompson in the lecture build on this tradition. They offer the same fundamental advantages of the cruder snowball sampling methods: exploiting the social links between individuals to collect data on a hidden population. In this sense, it is using the additional in the social network to improve the design. In addition it can adjust for discovered features in the population. All these lead to increased efficiency of sampling.

However, current methods also inherit the challenges of their ancestors. For example, adaptive web sampling requires the following of a link from a previously sampled unit to a new unit. For stigmatized human populations this process raises privacy concerns, e.g, illegal immigrants, injection drug users. There is also the practical issues of the unique identification of new individuals. For example, **?** applies Adaptive Web Sampling to a population at high risk for HIV in Colorado Springs, CO. The population is interconnected via drug-using relationships between individuals. As Thompson notes, the population was hidden so that it is prohibitively expensive to contact them through the available sampling frames and following links to identify new members of the population is the only practical approach. However, this requires members of the high risk population to identify and disclose their peers to the researchers. While this was famously possible in Colorado Springs (**?**), this is rarely feasible for most hidden populations where these methods are most likely to be applied. In addition, the typical advice is to have the initial sample of people, referred to as “seeds”, to be at least 50% of the total sample. This large proportion of seeds increases the coverage (“low bias”) while the subsequent adaptive / link-tracing component increases the efficiency (“low variance”). In this sense, Adaptive Web Sampling acts more like a “turbo-charger” to the design. If this advice is followed, choosing the seeds via a known sampling design is difficult for many populations. This would require large conventional designs or other innovations like spatial cluster sampling to implement (**?**). The combination of the privacy issues and the not insignificant sampling of seeds make the use of many link-tracing designs infeasible for the majority of hidden stigmatized populations.

Attempts have been made to ameliorate the effects of these two issues. If there are no privacy concerns then one could chose a small number of seeds via a non-probability sample and use link-tracing to collect the majority of the sample. While such approaches are often effective at acquiring a sample, the degree to which it can be considered a probability sample is usually unclear.

One popular approach of this kind is Respondent-Driven Sampling (RDS, introduced by **?**). RDS presents two main innovations for this setting: a design for sampling from the target population and a corresponding strategy for estimating population properties based on the resulting sample. It is from the former that the method draws its name: the RDS design relies on the respondents at each wave to select or "drive" the next wave of sampling through their selection of other members of the target population. This is typically achieved through the distribution of coupons by respondents to their alters via the underlying unobserved social network. This strategy reduces the privacy concerns generally associated with sampling from stigmatized populations. However, the sampling mechanism is now unknown as it depends on the fickle choices of the respondents rather than being under the control of the researchers. There have been many attempts at estimating population characteristics from RDS data based on the estimation of individual inclusion probabilities for the sampled units. These rely on various approximations of the sampling mechanism (**????**). Each of these require strong assumptions about the unknown sampling process for their validity. **?** and **?** were the first works to systematically evaluate the statistical properties of current estimators based on RDS data. More recent approaches attempt to adjust for a convenience sample of seeds. For example, **?** extend the estimator of **?** to correct for the bias introduced by seed selection in the presence of homophily.

While these approaches are useful they also highlight the challenges of adaptive network sampling for hidden populations. Typically the sampling mechanisms is partially unknown and needs to be estimated. If design-based inference is used the uncertainty in the estimates of the inclusion probabilities needs to be incorporated into the inference. In addition, joint inclusion probabilities, $P(s|\mathbf{y}, \phi)$ are required to account for the sampling dependencies.

These challenges suggest that future advances will depend on two areas. The first is model-based inference and especially Bayesian approaches. **?** develop a hybrid approach where design-based estimates are used and the inclusion probabilities are estimated based on a model for the network and the sampling design. Models for the static or dynamic networks, as developed in this Hansen Lecture, can better leverage the information available and account for the various sources of uncertainty. However the development of realistic models for networked populations is difficult. **?**, **?**, and others describe continuous-time Markov models for evolution of social networks (See **?** for a review). One well developed model is the *actor-oriented* model of **?** and **?**, which can be viewed in terms of actors making decisions to form and dissolve ties to other actors. This model was then extended by **?** to jointly model actors' network-related choices ("selection") and the effects of neighboring actors on each other's attributes ("influence").

The second area is in the development of novel sampling designs that better collect information on the network while preserving the privacy of the networked population. One approach is privatized network sampling (**?**). This is an area in rapid development (e.g., **?**).