Evaluating Respondent-driven Sampling on an Gay Men Web Community

X. Lu[1], L. Bengtsson[2], T. Britton[3], M. Camitz[4], B. J. Kim[5], A. Thorson[2], F. Liljeros[1]

[1]Department of Sociology, Stockholm University, Stockholm, Sweden

[2]Department of Public Health Sciences, Karolinska Institute, Stockholm, Sweden

[3]Department of Sociology, Stockholm University, Stockholm, Sweden

[4]Department of Medical Biostatistics and Epidemiology, Karolinska Institute, Stockholm, Sweden

[5]Physics Research Division and Department of Energy Science, Sungkyunkwan University, Korea

Objective

To analyze the behaviors of respondent-driven sampling by simulating RDS on an empirically extracted network of a ”hidden” population.

Background

Respondent Driven Sampling (RDS) is a recently introduced, network-based sampling method for hidden populations that are generally difficult to access with standard probability-based sampling methods because of the lack of a well defined sampling frame. It is a version of snowball sampling that estimates and compensates for the systematic bias in the sampling procedure. There has been a rapid increase in RDS research during recent years, with more than 120 empirical studies in 30 countries (Fig. 1) targeting a wide range of hidden populations, such as injection drug users, men who have sex with men, sex workers and their sexual partners.

However, the estimators for RDS are based on several assumptions that may not be valid in real life. Particularly, the networks are assumed to be bipartite, respondents to recruit peers randomly and the sampling procedure to be with replacement. As the population is generally not known, biases introduced with these assumptions are hard to determine.

Methods

Extract a network of gay men from the Nordic’s largest web community for homosexual, bisexual, transgender or queer people. Simulate RDS on the network for testing the effect of violations against the basic RDS assumptions.

Network Description

The network contains 31 909 gay men, with an average of 11.4 outgoing links (out-degree) for each member. The maximum out-degree is much larger than the maximum in-degree and the in-degree distribution is more approximate to a “power-law”, while the out-degree has a fat tail (Fig. 2). There are a certain amount of members who have no ingoing link, nor outgoing link, implying that these members can not be recruited by others, nor recruit other respondents, which may result in bias for RDS. Cumulative degree distributions show that the network is relatively sparse, with more than 60% members having no more than 10 outgoing (ingoing) links.

Fig. 1 RDS studies around the world

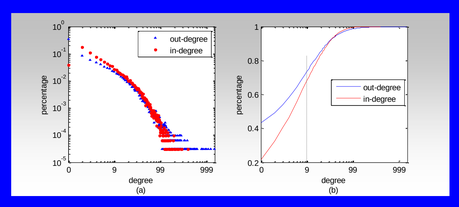

Fig. 2 (a) degree distribution (b) cumulative degree distribution

Results

1, RDS on undirected network

Asymptotically unbiased — The average RDS estimates of 10 000 times simulations for all the four groups approaches the true population value very fast (Fig. 3), while the unadjusted sample compositions always have large bias. Take age for example, with only 30 recruits in the sample, the *RDS Ⅱ* is expected to have only 0.5% deviation from the true proportion.

Resistant to seeds — The average mean absolute error (MAE) of sample size from sample size of 501 to 1000 changes little with the number and selection type of seeds (Fig. 4). *(In Fig. 4, t1 stands for selecting seeds uniformly from the network, and t2 stands for selecting with probability proportional to their degree)*

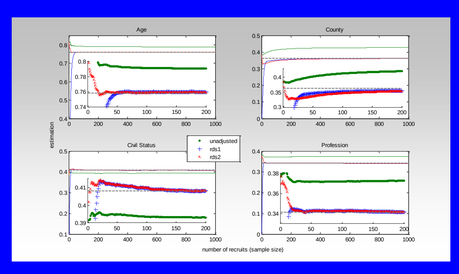

Fig. 3 Average estimations on undirected network

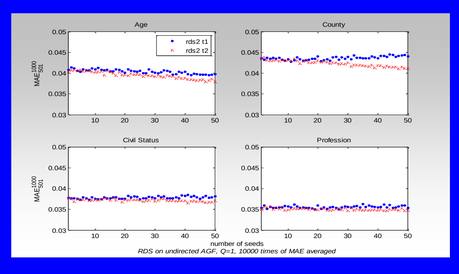

Fig. 4 Seeds effect on undirected network

2, RDS on directed network

Biased — When we take the direction of links into account and run RDS on the directed network, the average estimations of *RDS Ⅰ* and *RDS Ⅱ* are much biased from the true population value, and they never converge even when the sample comprises one third of the population, e.g., RDS estimations for county are expected nearly 10% larger than the true value, the average RDS estimations for age even have larger bias than the unadjusted sample composition (Fig. 5).

Sensitive to seeds — the number and selection type of seeds have significant effects on directed networks, that is, the more seeds with which we started the RDS, the less error would we get from the RDS estimations (Fig. 6).

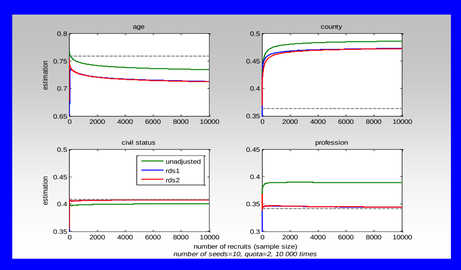

Fig. 5 Average estimations on directed network

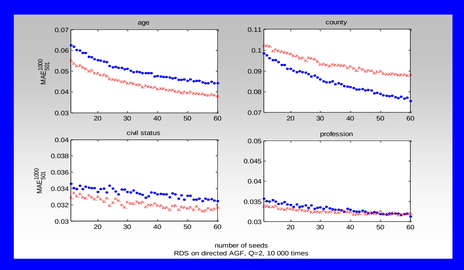

Fig. 6 Seeds effect on directed network

3, The recruit-with-mistake model

When recruiting, there would be two types of mistake respondents make: on the one hand, they may miss or cannot recognize some of their reciprocal links and thus report smaller degrees, we call this as “mistake 1”; on the other hand, they may take some irreciprocal links as reciprocal, we call this as “mistake 2”, which may inflate their reported degrees. we use $p_{re\to di}$ to represent the probability for respondents to make mistake on reciprocal links, and $p_{di\to re}$, to represent the probability for making mistake on irreciprocal links.

To test the sensitivity of RDS estimators, average MAE from 500 to 1000 recruits is calculated and the error surfaces are presented for different probabilities (Fig. 7). *(to avoid the effect of disconnected components when respondents only recruit on reciprocal links, we add a single link for each pair of disconnected components defined by reciprocal links)*

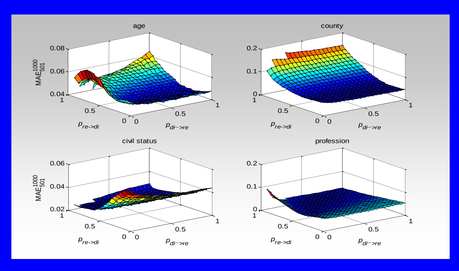

Fig. 7 Error surfaces of recruiting with mistake

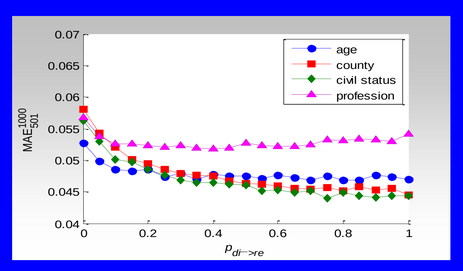

Fig. 8 Average MAE of making mistake on irreciprocal links

Robustness to mistakes on irreciprocal links — The average MAE varies little along with the change of mistakes made on irreciprocal links, while mistakes made on reciprocal links might result in uncertain error. Take county and civil status for example, given a certain level of mistakes made on reciprocal links, the average MAE almost constant no matter how many irreciprocal links are regarded as the potential links to be recruited.

2-D plots of average MAE when respondents only make mistake on irreciprocal links are depicted in Fig. 8. Obviously, the more irreciprocal links taken as potential links to be recruited, the less error would be for RDS estimations. Even when the respondents take all outgoing links as potential, the average MAE will only vary within 1.5%, indicating the robustness of RDS estimators for making mistake on irreciprocal links.

Conclusions

Different behaviors of RDS estimates under various conditions are analyzed in this paper. The RDS estimates may bring in large bias and misleading estimations when the network are directed, indicating that RDS methods should be carefully used when the studied population forms irreciprocal relationships.

There are lots of hints from this study: first, if the network fulfills all the assumptions, that is, reciprocal, well connected, etc., the sample size should be as large as possible to reduce the estimation error; second, seeds should be sufficient and diverse enough to generate a representative sample of desired size; third, when interviewing, respondents should be asked to report as many potential recruits as they know to avoid missing their reciprocal connections.

Acknowledgement

This work is funded in part by the Riskbankens jubileumsfond (dnr: P2008-0674) and FAS in Sweden (Swedish Council for Working Life and Social Research).

References

1. Heckathorn D D. Respondent-driven sampling: A new approach to the study of hidden populations [J]. Social Problems, 1997, 44 (2): 174-199
2. Volz E, Heckathorn DD. Probability-based estimation theory for respondent-driven sampling. Journal of Official Statistics 2008, 24(1):79-97
3. Malekinejad M, Johnston L, Kendall C, Kerr L, Rifkin M, Rutherford G. Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review [J]. AIDS and Behavior, 2008, 12 (0): 105-130
4. Salganik M J. Variance estimation, design effects, and sample size calculations for respondent-driven sampling [J]. Journal of Urban Health-Bulletin of the New York Academy of Medicine, 2006, 83 (6): I98-I112

Correspondence to: lu.xin@sociology.su.se