RESEARCH ARTICLE

One-step estimation of networked population size: Respondent-driven capture-recapture with anonymity

Bilal Khan[1]*, Hsuan-Wei Lee[1], Ian Fellows[2], Kirk Dombrowski[1]

1 Department of Sociology, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America,
2 Fellow Statistics, San Diego, California, United States of America

* bkhan2@unl.edu

OPEN ACCESS

Citation: Khan B, Lee H-W, Fellows I, Dombrowski K (2018) One-step estimation of networked population size: Respondent-driven capture-recapture with anonymity. PLoS ONE 13(4): e0195959. https://doi.org/10.1371/journal.pone.0195959

Editor: Ming Tang, East China Normal University, CHINA

Received: February 22, 2018

Accepted: March 31, 2018

Published: April 26, 2018

Data Availability Statement: All Java software developed and used is available at https://github.com/grouptheory/telefunken-support/tree/master/java. All R software developed and used is available at https://github.com/grouptheory/telefunkensupport/tree/master/R-v1. All data inputs and outputs are available at https://github.com/grouptheory/telefunken-support/tree/master/figures_and_data.

Funding: Research reported in this publication was supported by the National Institutes for Health,

Abstract

Size estimation is particularly important for populations whose members experience disproportionate health issues or pose elevated health risks to the ambient social structures in which they are embedded. Efforts to derive size estimates are often frustrated when the population is hidden or hard-to-reach in ways that preclude conventional survey strategies, as is the case when social stigma is associated with group membership or when group members are involved in illegal activities. This paper extends prior research on the problem of network population size estimation, building on established survey/sampling methodologies commonly used with hard-to-reach groups. Three novel one-step, network-based population size estimators are presented, for use in the context of uniform random sampling, respondent-driven sampling, and when networks exhibit significant clustering effects. We give provably sufficient conditions for the consistency of these estimators in large configuration networks. Simulation experiments across a wide range of synthetic network topologies validate the performance of the estimators, which also perform well on a real-world location-based social networking data set with significant clustering. Finally, the proposed schemes are extended to allow them to be used in settings where participant anonymity is required. Systematic experiments show favorable tradeoffs between anonymity guarantees and estimator performance. Taken together, we demonstrate that reasonable population size estimates are derived from anonymous respondent driven samples of 250-750 individuals, within ambient populations of 5,000-40,000. The method thus represents a novel and cost-effective means for health planners and those agencies concerned with health and disease surveillance to estimate the size of hidden populations. We discuss limitations and future work in the concluding section.

1 Introduction

Estimating the size of hidden and hard-to-reach populations is of critical importance to health officials seeking to mitigate the extent of health problems that may be concentrated within such populations [1], or when "reservoirs" of infection among a hidden population pose a

80. Bender EA, Canfield ER. The asymptotic number of labeled graphs with given degree sequences. Journal of Combinatorial Theory, Series A. 1978; 24(3):296–307. http://dx.doi.org/10.1016/0097-3165(78)90059-6.

81. Bollobás B. A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs. European Journal of Combinatorics. 1980; 1(4):311–316. http://dx.doi.org/10.1016/S0195-6698(80)80030-8.

82. Newman MEJ, Strogatz SH, Watts DJ. Random graphs with arbitrary degree distributions and their applications. Phys Rev E. 2001; 64:026118. https://doi.org/10.1103/PhysRevE.64.026118

83. Albert R, Barabási AL. Statistical mechanics of complex networks. Rev Mod Phys. 2002; 74:47–97. https://doi.org/10.1103/RevModPhys.74.47

84. Illenberger J, Flötteröd G. Estimating network properties from snowball sampled data. Social Networks. 2012; 34(4):701–711. https://doi.org/10.1016/j.socnet.2012.09.001

85. Handcock MS, Gile KJ, Mar CM. Estimating hidden population size using respondent-driven sampling data. Electronic journal of statistics. 2014; 8(1):1491. https://doi.org/10.1214/14-EJS923 PMID: 26180577

86. Coronado-García M, Thrash CR, Welch-Lazoritz M, Gauthier R, Reyes JC, Khan B, et al. Using Network Sampling and Recruitment Data to Understand Social Structures Related to Community Health in a Population of People Who Inject Drugs in Rural Puerto Rico. Puerto Rico Health Sciences Journal. 2017; 36(2):77–83. PMID: 28622403

87. Verdery AM, Fisher JC, Siripong N, Abdesselam K, Bauldry S. New Survey Questions and Estimators for Network Clustering with Respondent-driven Sampling Data. Sociological Methodology. 2017; p. 0081175017716489. https://doi.org/10.1177/0081175017716489

88. Verdery AM, Siripong N, Pence BW. Social Network Clustering and the Spread of HIV/AIDS Among Persons Who Inject Drugs in 2 Cities in the Philippines. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2017; 76(1):26–32. https://doi.org/10.1097/QAI.0000000000001485 PMID: 28650399

89. Carter JL, Wegman MN. Universal classes of hash functions. Journal of Computer and System Sciences. 1979; 18(2):143–154. http://dx.doi.org/10.1016/0022-0000(79)90044-8.

90. McCreesh N, Johnston LG, Copas A, Sonnenberg P, Seeley J, Hayes RJ, et al. Evaluation of the role of location and distance in recruitment in respondent-driven sampling. International journal of health geographics. 2011; 10(1):56. https://doi.org/10.1186/1476-072X-10-56 PMID: 22008416

91. Rocha LE, Thorson AE, Lambiotte R, Liljeros F. Respondent-driven sampling bias induced by community structure and response rates in social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2017; 180(1):99–118. https://doi.org/10.1111/rssa.12180

92. Sperandei S, Bastos LS, Ribeiro-Alves M, Bastos FI. Assessing respondent-driven sampling: A simulation study across different networks. Social Networks. 2017;.

93. Cho E, Myers SA, Leskovec J. Friendship and Mobility: User Movement in Location-based Social Networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'11. New York, NY, USA: ACM; 2011. p. 1082–1090. Available from: http://doi.acm.org/10.1145/2020408.2020579.