Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Minimum distance estimators of population size from snowball samples using conditional estimation and scaling of exponential random graph models

David A. Rolls [*], Garry Robins

Melbourne School of Psychological Sciences, University of Melbourne, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 20 May 2016
Received in revised form 14 April 2017
Accepted 18 July 2017
Available online xxxx

Keywords:
Network sampling
Hidden population sampling
Network size
Link-tracing designs
Social networks
Hard-to-reach populations

ABSTRACT

New distance-based estimators of population size for snowball sample network data using exponential random graph models (ERGMs) are presented. After ERGM parameters are obtained using conditional estimation it is possible to simulate networks from the ERGM across a range of hypothesized sizes and then estimate the population's size. This is done by creating simulated snowball samples from the simulated networks and then minimizing their distances from an observed network statistic across network sizes. The number of nodes in the snowball sample (snowball size) combined with a moment-based distance is shown to be an effective estimator. For ERGM conditional estimate parameters, the moment-based snowball size estimator can outperform a multivariate Mahalanobis estimator, where the latter would be a maximum likelihood estimator under the assumption the network statistics are multivariate Gaussian. "Extreme" ERGM scaling across network sizes, which prevents finding a minimum-distance estimate, is also discussed.

1. Background

For hard-to-reach or "hidden" populations, link tracing designs like snowball sampling and respondent-driven sampling are an efficient way to collect samples. However, a network sample leaves unclear the size of the population, which may be interesting on its own (e.g., in a public health context) or may be important for further research (e.g., for simulating networks to study the effectiveness of public health interventions). In this paper we consider an approach for estimating the size of populations using exponential random graph models (ERGMs) fit to snowball sampled network data.

There is previous literature on estimating the size of hidden populations, including for link-tracing samples (Frank and Snijders, 1994), respondent-driven sampling (Handcock et al., 2014; Crawford et al., in press) and ad hoc methods using ERGMs (Rolls et al., 2013). Methods for estimating the size of hidden populations using other sampling designs include capture/re-capture estimators (Bao et al., 2010; Berchenko and Frost, 2011; Paz-Bailey et al., 2011) and a technique combining link-tracing and cluster sampling (Félix-Medina and Thompson, 2004).

For link tracing designs, Frank and Snijders (1994) assume both that the seed set is an IID sample of the population and arcs in the adjacency matrix for the population network are IID Bernoulli. Handcock et al. (2014) consider data from respondent-driven sampling (RDS). In particular, they use "unit size", a generalization of node degree. They use a Bayesian approach and a four component Gibbs sampler to approximate the joint posterior distribution for the population size, the parameter of the unit size distribution, the unit sizes of the unobserved members of the population, and the sequence of

[*] Corresponding author.
E-mail address: drolls@unimelb.edu.au (D.A. Rolls).

http://dx.doi.org/10.1016/j.csda.2017.07.004
0167-9473/© 2017 Published by Elsevier B.V.

unit sizes through each step of the data collection. Another important paper is Handcock and Gile (2010), which presents a general framework for considering network sampling designs, considers when sampling designs are *amenable* to an ERGM model and discusses likelihood estimation. In particular they bring the sampling design into the joint estimation of the ERGM parameters and ψ, where ψ is a parameter of the sampling design (e.g., the probability a member of the population is included in the seed set.)

Starting from ERGM parameters obtained using conditional estimation (Pattison et al., 2013), Rolls et al. (2013) describe an ad hoc method to select an appropriate network size starting from snowball sample data. Koskinen et al. (2013) consider related problems in the context of inference for ERGM models with "covert" actors in a Bayesian framework. Shalizi and Rinaldo (2013) discuss the inappropriate use of the same ERGM parameters across a range of network sizes, but do not consider conditional estimation which is a technique to estimate ERGM parameters of the full population using a network sample. For dyad independent ERGMs and egonet sampling, (i.e., models without clustering), Krivitsky et al. (2011) and Krivitsky and Kolaczyk (2015) introduce an offset term into the ERGM specification so the same parameters can be used for a range of network sizes.

The general approach described here also considers a range of hypothesized network sizes. Unlike Koskinen et al. (2013) these sizes are used for simulating networks (using the conditional estimates) and not estimation of ERGM parameters themselves. Unlike Handcock et al. (2014) the data is assumed from a k-wave snowball sample, not an RDS design, and ERGM model parameters are not estimated with the population size. Consistent with Shalizi and Rinaldo (2013), we actually exploit the idea that fixed ERGM parameters across different network sizes lead to networks with different properties. In our approach, snowball samples with the same number of seeds and waves as the empirical sample are taken from simulated population networks. These are used to create a distribution for the graph statistics over snowball samples for each population size. A distance from the graph statistics of the empirical sample to the distribution is minimized to create the network size estimate. An advantage of this approach is it places the emphasis on capturing observed properties of the data (e.g. graph statistics) rather than on network parameters. This is particularly important since the sensitivity of observed features to small changes in ERGM parameters is unclear. In contrast to Rolls et al. (2013), we also provide a framework to use multiple graph statistics to create the population size estimate.

2. Methods

2.1. Notation

Exponential random graph models (ERGMs) (Frank and Strauss, 1986; Robins et al., 2007a) are a particular class of network models that have proven useful in modelling social networks. Under a homogeneity assumption whereby all structurally identical subgraphs are equally probable, these models have the form

$$\Pr(Y = y) = \exp\left(\eta \cdot z(y)\right) / \kappa, \tag{1}$$

where Y is an $N \times N$ binary matrix of variables for N nodes denoting whether a tie is present (1) or absent (0), y denotes a realization of Y, η is a vector of model parameters, $z(\mathbf{y})$ is vector of corresponding network statistics observed in y, and κ is a normalizing constant which ensures Eq. (1) describes a proper distribution. ERGMs provide a parsimonious method to model structural features of a network (e.g., edge density, clustering) and features related to nodal attributes (e.g., homophily). The network statistics $z(\mathbf{y})$ may include structural quantities like the number of edges or triangles or any of the geometrically-weighted alternating statistics proposed in the literature (Snijders et al., 2006; Hunter, 2007). (See Robins et al., 2009 for a concise summary of ERGM alternating statistics.) They may also include node related quantities like the number of nodes with a shared attribute, as a means to capture homophily. Computation of κ is extremely demanding for all but the simplest ERGMs. For unknown κ, estimation and simulation for ERGMs usually involves Markov Chain Monte Carlo maximum likelihood estimation (MCMCMLE) and MCMC methods, respectively. ERGMs provide a parsimonious method to model structural features of a network (e.g., edge density, clustering) and features related to nodal attributes (e.g., homophily).

We assume the observed data is a k-wave snowball sample. A snowball sample of a network Y is formed by starting with a collection S_0 of seed nodes (referred to as the "seed set" or "zone 0"), usually chosen as a random sample although other schemes like probability proportional to degree are possible. Let S_1 ("zone 1") be all the nodes not in S_0 sharing an edge with a node in S_0. Similarly, let S_i ("zone i") be all the nodes sharing an edge with a node in S_{i-1} that are not already in $\cup_{j=1}^{i-1} S_j$. Let s_i be an observation of S_i. A 1-wave snowball sample is the subgraph formed by the nodes in S_0 and S_1, the edges between nodes in S_0, and the edges between nodes in S_0 and nodes in S_1. A k-wave snowball sample is a $(k-1)$-wave snowball sample together with the nodes in S_k, the edges between nodes in S_{k-1}, and the edges between nodes in S_{k-1} and nodes in S_k.

Let y_{obs} be the observed sample, s_{obs} be the seed nodes of the observed sample, and $n_0 = |s_{obs}|$ be the size of the seed set in the observed sample. Unless otherwise mentioned, we assume $|S_0| = n_0$ arising from a sampling choice rather than a random process. Further, we assume seed sets are equally likely and formed by choosing n_0 nodes at random without replacement from the population. In the case of an ERGM with only structural parameters (i.e., no nodal attributes), specifying the conditions $S_0 = s_{obs}$ and $|S_0| = n_0$ are equivalent. For an ERGM with nodal attributes they are generally different because s_{obs} also includes information on the joint frequency of attributes within the seed set.

Handcock and Gile (2010) provide a notation to express the sampling design for snowball samples. Let $D_N(y, s_0)$ be the 0/1 $N \times N$ matrix that is 1 if the arc represented by (i, j) in graph y was sampled using seed set s_0. Notice this is fully determined

Please cite this article in press as: Rolls, D.A., Robins, G., Minimum distance estimators of population size from snowball samples using conditional estimation and scaling of exponential random graph models. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.07.004.