Bayesian Inference from Non-Ignorable Network Sampling Designs

Simón Lunagómez, Edoardo M Airoldi
Department of Statistics, Harvard University
1 Oxford Street, Cambridge, MA 02138, USA
Corresponding author: Simón Lunagómez, e-mail: lunagomez@fas.harvard.edu

Abstract

Consider a population where subjects are susceptible to a disease (e.g. AIDS). The objective is to perform inferences on a population quantity (like the incidence of HIV on a high-risk subpopulation, e.g. intra-venous drug abusers) via sampling mechanisms based on a social network (link-tracing designs, respondent-driven sampling). We phrase this problem in terms of the framework proposed by Rubin for making inferences on a population quantity and, within this context, prove that respondent-driven sampling is non-ignorable. By non-ignorable it is meant that the uncertainty of the sampling mechanism must be modeled in the likelihood in order to get valid inferences. We develop a general framework for making Bayesian inference on the population quantity that: models the uncertainty in the underlying social network, incorporates dependence among the individual responses according to the network, and deals with the non-ignorability of the sampling design. The proposed framework is general in the sense that it allows a wide range of different specifications for the components of the model we just mentioned. Our model is compared with state of the art methods in simulation studies and it is applied to real data.

1 Introduction

Usually not including explicitly the sampling mechanism in the likelihood function does not have an impact in the inference (either likelihood-based or Bayesian) of a population quantity. All that is needed is the vector of indicators that encode which individuals have been included in the sample. Rubin [12] and Heitjan and Rubin [9] have developed a rigorous approach for tackling the question of when it is valid to ignore the functional form of the sampling mechanism for performing inferences. A key notion for this approach is the one of *ignorability*, which establishes when the probability distribution of the sampling design is relevant for modeling the distribution of random quantities corresponding to individuals not included in the sample. Under Rubin's framework ignorability is equivalent to saying that the posterior of the population quantity can be computed without conditioning on the functional form of the sampling design. A sampling design is called *non-ignorable* if its functional form has to be expressed explicitly in the model in order to perform likelihood-based or Bayesian inference.

We consider a situation where non-ignorability arises because the sampling design is driven by a network, which is progressively discovered through sampling. This could happen in at least two ways:

1. The probability distribution of the sampling design depends on features corresponding to the portion of the graph that was not sampled. An obvious implication of this is that changes in the underlying network will affect the likelihood. An equally important, but greatly overlooked aspect of this is that we could have different realizations of the sampling mechanism, leading to the same set of sampled individuals, but conveying different information about the network structure, therefore producing different inferences.
2. The network induces a dependence structure on the responses, in such case the responses of not sampled individuals need to be taken into account for computing the likelihood.

The first point is illustrated in Figure 1. It shows the case when the likelihood of the sample is affected by the underlying network and the case where the same set of sampled individuals can arise from different realizations of the design implying different values for the likelihood.

From a statistical perspective, the issues of inferring a population quantity using a non-ignorable sampling design on a social network include: Modeling the unobserved part of the social network

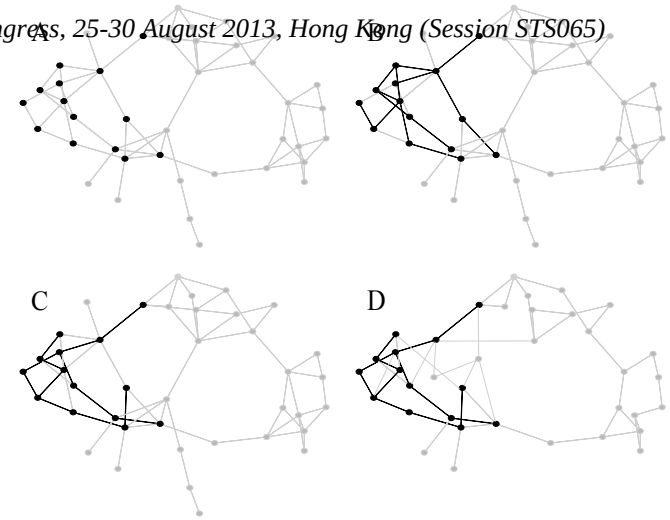

Figure 1: An illustration of the sources of non-ignorability in link-tracing designs. *Panel A:* A population graph and a random sample. *Panel B:* A realization of a link-tracing design on that produces the sample in panel A. *Panel C:* A different realization of a link-tracing design on that produces the same sample in panel A. *Panel D:* The same realization of a link-tracing design in panel C, but on a different population graph; it has a different likelihood.

in probabilistic fashion, understanding the sampling mechanism as a probability model and including it in the likelihood, and finally, modeling the dependence structure of the responses given the network.

A sample design based on a social network structure that is widely used for inferring a population quantity is Respondent-Driven Sampling (RDS). This design was proposed by Heckathorn [7]. Later, Volz and Heckathorn [15] proposed an estimation procedure tailored for RDS based on the assumption that the relationship between the probability of inclusion of a given individual and the degree of the corresponding node is linear. Gile [4] improved this methodology by estimating the relationship between inclusion probability and degree distribution via an iterative procedure. None of this approaches is model-based, therefore, they are all vulnerable to the issues mentioned before. What is more relevant to the discussion is that these approaches assume that RDS is an ignorable design; we prove this is not true. A very interesting work that involves the concept of ignorability is Hancock and Gile [6]. Their focus is on estimating the parameters of the social network model, not in estimating a population quantity while allowing uncertainty for the network structure.

The purpose of this paper is to propose methodology such that: Allows us to perform inferences in cases where the sampling mechanism is non-ignorable. This has to be done while taking into account all relevant sources of uncertainty, *i.e.*, uncertainty regarding the underlying social network, the sampling mechanism, and the dependence structure of the responses.

2 Theory and Definitions

2.1 Rubin's Framework

The objective is to perform Bayesian inference of a population quantity Q. It is assumed that $Q = Q(X, W)$, here:

- W denotes the *response vector*, *i.e.* those variables that are the primary interest of the investigator (in the sense that the scientific question is related directly to them). To observe these variables it is a necessary condition that the corresponding individual is included in the sample.
- X denotes a vector of covariates that is available for all individuals in the population, regardless if they are included in the sample or not.

These assumptions and setting are established in [13]. We denote the sampling mechanism by I. Here I plays two roles: it represents the sampling mechanism as a probability model (how the sample was obtained) and the indicator function for a specific sample ($I = 1$ for individuals included in the sample, $I = 0$ otherwise).

Ignorability [13] is a property of the sampling mechanism I with respect to a model $p(W, X, I)$. A sampling mechanism is called *ignorable* with respect to a model $p(W, X, I)$ if:

$$\Pr(I \mid W, X) = \Pr(I \mid W_{INC}, X).$$

I is called *non-ignorable* otherwise.

For the purpose of this paper, it is reasonable to set $W = (Y, \mathcal{G})$, where $\mathcal{G}$ is the graph or social network and Y is the vector of univariate responses associated to each node of $\mathcal{G}$. This means that a sampling mechanism I is non-ignorable if its distribution is not constant with respect to unobserved data (unobserved entries of Y or features of the unobserved portion of the network).

2.2 Respondent-Driven Sampling

In order to perform inferences for a population quantity, where the population is regarded as hard-to-reach, sampling mechanisms that take advantage of the social network structure have been proposed. A first example of this is the snowball sampling proposed by Goodman [5]. Later, Respondent-Driven Sampling (RDS) was proposed by Heckathorn [7]; it constitutes an improvement over snowball sampling in the sense that it guarantees balance with respect to unobserved covariates regardless of the starting points (also known as seeds).

We now describe the modality of RDS to be discussed in this paper:

1. r individuals belonging to the population of interest are selected, and recruited to the study. These individuals are known as *seeds*;
2. for each seed, m neighbors in the social network are selected according to an uniform distribution over the neighbors of that node that have not been sampled previously. In the case that the number of available neighbors of the seed (*i.e.*, the ones that have not been sampled yet) is less or equal to m, all available neighbors (if any) are selected. The nodes sampled this way are known as the *first wave*;
3. for each individual belonging to the $i-$*th wave*, m neighbors in the social network are chosen according to an uniform distribution over the neighbors of that node that have not been sampled previously. Again, if the number of available neighbors of the node is less or equal to m, all available neighbors (if any) are selected. The nodes sampled this way are known as the $(i+1)-$*th wave*;
4. sampling is performed until a pre-specified sample size n is attained.

Response-Driven Sampling is non-ignorable. Let I denote RDS. The distribution of I depends on the subgraph of $\mathcal{G}$ that is included in the sample, but it also depends on information regarding the unobserved part $\mathcal{G}$. If a node is recruited and has d neighbors, where $d > m$, m being the number of coupons per individual, then the distribution of I depends on the number of neighbors of the node not included in the sample. Then,

$$\Pr(I \mid Y, \mathcal{G}) \neq \Pr(I \mid Y_{INC}, \mathcal{G}_{INC}).$$

It follows that I is non-ignorable.

3 Statistical Methodology

3.1 Bayesian Modeling When the Sampling Propagates Through a Social Network

We now describe the methodology we propose. Our method provides a probabilistic approach for modeling and inferring a population quantity Q when the sampling propagates though a social network $\mathcal{G}$. By social network we mean the graph that encodes the relationships between the individuals. In this Section, Y and I have the meaning given in Section 2.1.

The graph is incorporated in the modeling at two different levels: first, it is assumed that the sample is propagated using the social network; second, the graph imposes a dependence structure on Y. This second assumption makes this approach more realistic for real-life applications. As discussed before, mechanisms such as RDS are not ignorable, which means that a term regarding the distribution of I has to be included in the likelihood function in order to obtain valid inferences.

Based on these ideas, we proposed a model of the form:

$$p(Y, I, \mathcal{G}, \alpha, \gamma) = p(\alpha)p(\mathcal{G} \mid \alpha)p(I \mid \mathcal{G})p(\gamma)p(Y \mid \mathcal{G}, \gamma). \quad (1)$$

Here $p(\mathcal{G} \mid \alpha)$ is a random graph model for the social network; this factor is used mainly to model the unobserved portion of the network. The random graph has parameter α, with prior $p(\alpha)$. $p(I \mid \mathcal{G})$ models the distribution of the sampling mechanism given the graph, by adding this term we take care of the issues regarding non-ignorability. It is reasonable to assume that such distribution does not depend on α. The term $p(Y \mid \mathcal{G}, \gamma)$ is used to model the dependence structure of Y given the graph. In it was assumed a Markov Random Field to induce a dependence on Y given $\mathcal{G}$. To specify such distribution, extra parameters (denoted by γ) are needed. Here $p(\gamma)$ denotes the prior for γ.

We now explain how Bayesian inference can be performed by the model given in Expression 1. First we introduce some notation: Let Y_{INC} and $\mathcal{G}_{INC}$ denote, respectively, the observed responses (Y for which $I = 1$) and the observed subgraph of $\mathcal{G}$. Let Y_{EXC} and $\mathcal{G}_{EXC}$ denote,

respectively, the unobserved responses and the unobserved part of the network. To deal properly with the non-ignorability issues due to the missing data, and in order to compute the part of the likelihood corresponding to $p(Y \mid \mathcal{G}, \gamma)$, Y_{INC} and $\mathcal{G}_{INC}$ have to be augmented; we denote by Y_{AUG} and $\mathcal{G}_{AUG}$ the augmented portion for the response vector and the graph, respectively. Within this setting, inference on Q assuming a model of the form given by Expression 1 is performed via Bayesian model averaging (see [10] and [11], Section 7.4):

$$p(Q \mid Y_{INC}, \mathcal{G}_{INC}, I) = \sum_w p(\mathcal{M}_w) \int_{\Theta_w} p_w(Q \mid \theta_w) p(\theta_w \mid Y_{INC}, \mathcal{G}_{INC}, I) d\theta_w.$$

Where $\theta_w = (Y_{AUG}(w), \mathcal{G}_{AUG}(w), \alpha(w), \gamma(w))$. $p(\mathcal{M}_w)$ is the distribution that dictates how many extra nodes and edges will be added to $\mathcal{G}_{INC}$; It is constructed as follows:

1. We obtain D samples from $p(\mathcal{G}, \alpha) = p(\alpha)p(\mathcal{G} \mid \alpha)$.
2. For each sample, a realization of $(I \mid \mathcal{G})$ is obtained.
3. Therefore, we have D Monte Carlo versions of $(\mathcal{G}, \mathcal{G}_{INC})$.
4. From them, the number of nodes and edges to be augmented is computed.

The quantity of interest Q is not explicitly represented in the model. It is assumed that Q is a function of the response vector Y and the graph $\mathcal{G}$. To compute $p_k(Q \mid \theta_k)$ when Q is the population mean we proceed as follows: For each iteration of the MCMC we obtain:

$$\tilde{Y} = (Y_{INC}, Y_{AUG})$$

and let the Monte Carlo version of Q be the sample mean of $\tilde{Y}$. To compute $p(\theta_k \mid Y_{INC}, \mathcal{G}_{INC}, I)$ we implemented a Metropolis-Hastings algorithm with a mixture of kernels.

3.2 Model Specification

We now present the specific choices we made for these distributions: For $\mathcal{G}$ an Erdös-Rényi model [3] was assumed, with a single probability of inclusion $\alpha \in (0, 1)$. A Beta(ω_1, ω_2) was used as prior for α. Our specification for $p(I \mid \mathcal{G})$ is an RDS with m coupons per wave and sample size n. For this paper we will assume that y_i is binary, and

$$\Pr\{y_i = 1 \mid Y_{-i}, \mathcal{G}, \gamma\} = \Phi\left(\psi + \sum_{\{k|A(i,k)=1\}} \zeta y_k\right), \tag{2}$$

where A is the adjacency matrix for $\mathcal{G}$ and $\gamma = (\psi, \zeta)$. In other words $p(Y \mid \mathcal{G}, \gamma)$ is specified as a Markov Random Field (MRF) based on a probit model. We used a scaled Beta as prior for ζ, *i.e.*

$$p(\zeta \mid \eta_1, \eta_2, \delta) = \frac{1}{B(\eta_1, \eta_2)} \frac{\zeta^{\eta_1 - 1}(\delta - \zeta)^{\eta_2 - 1}}{\delta^{\eta_1 + \eta_2 - 1}} \times \mathbb{I}_{(0,\delta)}(\zeta), \tag{3}$$

and a density of the form:

$$p(\psi \mid \nu_1, \nu_2, \xi) = \frac{1}{B(\nu_1, \nu_2)} \frac{(\psi + \xi)^{\nu_1 - 1}(-\psi)^{\nu_2 - 1}}{\xi^{\nu_1 + \nu_2 - 1}} \times \mathbb{I}_{(-\xi,0)}(\psi) \tag{4}$$

as prior for ψ.

It is worth to emphasize that we need to make specific choices for this distributions for the sake of concreteness. None of the choices we made for these distributions are essential for applying our approach: other distributions could be used.

4 Results

We conducted a simulation study in order to gain better understanding on the performance of our method. We considered the following regimes:

- For the graph topology we used a Small World (SW) model on a circle. The degree on the initial state (the lattice, before re-wiring) was set as 8. Four probabilities for re-wiring were considered: 0.15, 0.35, 0.75, and 0.95.
- To understand the impact of the strength of the dependence among the responses, we varied the parameters of the MRF to represent *low* dependence ($\psi = -0.82$, $\zeta = 0.01$). and *high* dependence ($\psi = -1.1$, $\zeta = 0.15$).

The size of the underlying network was set as 100, the sample size was fixed in 35. RDS was run using a single seed and 3 coupons in all cases. The specification of the MRF that we used in combination with the random graph model imply $Q = 0.2$ in all scenarios.

For each scenario we simulated 100 datasets. Each dataset was generated as follows : (1) A realization of the random graph model was simulated. (2) A vector of responses was simulated

from the MRF model given the graph. (3) RDS was performed. (4) Using the prior described in Section 3.1 (the density of the graph was centered at the true value), the mixing distribution of the BMA was calibrated. This was done using 250 samples from the prior. (5) The BMA procedure was executed with 4 samples from the mixing distribution, with 500 samples for each MCMC and a burn-in of $7,500$ iterations. Results are summarized in Table 1.

Estimator	Dependence	0.15	0.35	0.75	0.95
Bayes	high	0.087	0.045	0.021	0.017
VH	high	0.024	0.026	0.024	0.025
Bayes	low	0.094	0.051	0.023	0.019
VH	low	0.021	0.022	0.021	0.021

Table 1: Average bias of the estimator based on our methodology (Bayes) and Volz-Heckathorn (VH). The regimes are given by the strength of the dependence implied by the MRF model ($\psi = -1.1$, $\zeta = 0.15$ for high and $\psi = -0.82$, $\zeta = 0.01$ for low) and the probability of re-wiring in a SW graph. The size of the network was set at 100 and the sample size was 35. The average bias was computed over 100 simulations. For all regimes $Q = 0.2$.

As expected, our method outperforms the augmented VH when the re-wiring probability is high. Not surprisingly, our procedure tends to be more biased when the probability of re-wiring is low. This was expected since the distribution of $\mathcal{G}_{AUG}$ is a function of the prior $p(\mathcal{G} \mid \alpha)$. In other words: a drastic error in specifying the random graph prior will lead to bias results. The method we propose shows a better improvement over the augmented VH when there is high dependence in the response.

5 Real Data

We applied our methodology to the data derived from the study discussed in [2]. This was a large RDS study implemented in a single location, namely the community of Campinas in the state of Sao Paulo, Brazil. Since RDS was used, non-ignorability is an issue for likelihood-based inferences. The aim of the study was to infer the prevalence of HIV among gay men in Campinas, Brazil.

The study comprised 658 men who have sex with men. The inclusion criteria used for this study, were:

1. born male;
2. had anal or oral sex with another man or transvestite in the past six months;
3. 14 years of age or older;
4. reside in the Metropolitan area of Campinas.

RDS was implemented using 16 seeds and a maximum of 3 referrals per subject (*i.e.*, $m = 3$). Point estimates (sample proportion and Volz-Heckathorn) and Bootstrap confidence intervals are shown in Table 2. The results shown in this table are not model-based.

	Naive		Volz-Heckathorn	
$\hat{Q}$	0.0789	$(0.0577, 0.1001)$	0.0711	$(0.0466, 0.0955)$

Table 2: Point estimators (sample proportion, Volz-Heckathorn) and the corresponding 95 per cent Bootstrap confidence intervals.

We also applied our method to this data set. We used the distributions described in Section 3.2. Since no prior information for the social network $\mathcal{G}$ is available, a sensitivity analysis was conducted. For the Erdös-Rényi model, the density and the size of the graph were allowed to vary. Results from the sensitivity analysis are summarized in Table 3. Inferences tend to be quite stable with respect to the density and the network size assumed for the prior.

6 Discussion

In this paper we developed methodology for performing inference on non-ignorable designs on a network. The authors in [6] discuss the idea of *amenable* designs and work with sampling mechanisms that fulfill that condition. It would be interesting to understand the relationship between amenability and graph ignorability. One key difference between these two concepts is that amenability deals with situations where the objective of the inference is α, the parameter vector for the graph, in contrast, the definition of graph ignorability was formulated for performing inferences on a quantity $Q(Y, \mathcal{G})$. It is reasonable to think that the methodology proposed here can be used for dealing with designs that are not amenable. It is our understanding that all the available literature falls into one of two categories: Either they do not discuss ignorability, but assume

N	Density	Mean	SD	0.025	0.05	0.5	0.95	0.97
1316	0.1	0.113	0.007	0.098	0.101	0.113	0.123	0.125
1316	0.05	0.114	0.008	0.097	0.100	0.114	0.128	0.131
1316	0.01	0.109	0.017	0.076	0.077	0.112	0.136	0.140
1316	0.005	0.119	0.008	0.104	0.106	0.119	0.134	0.136
1316	0.001	0.116	0.009	0.099	0.101	0.116	0.133	0.136
1316	$\frac{1}{N}$	0.114	0.006	0.101	0.103	0.114	0.125	0.129
2632	0.1	0.133	0.007	0.119	0.122	0.133	0.144	0.147
2632	0.05	0.135	0.008	0.118	0.122	0.134	0.150	0.151
2632	0.01	0.138	0.014	0.108	0.110	0.140	0.158	0.161
2632	0.005	0.146	0.011	0.122	0.123	0.147	0.166	0.169
2632	0.001	0.155	0.009	0.138	0.140	0.155	0.169	0.173
2632	$\frac{1}{N}$	0.129	0.015	0.096	0.101	0.131	0.152	0.156

Table 3: Summaries of the posterior for Q. These include: mean, standard deviation, and quantiles. Posterior samples were obtained assuming different values of the density for the Erdös-Rényi model and size of the underlying network N.

implicitly that the sampling mechanism of their choice is ignorable (*e.g.*, [4]), or they discuss ignorability and then they restrict their discussion to what they regard as ignorable designs [6]. In either case, the problem of making inference on a population quantity using a non-ignorable design is not addressed.

An important feature of the methodology we propose is that it is highly modular. By this we mean that the term $p(I \mid \mathcal{G})$ does not have to correspond to RDS. All the arguments hold for any other design that is not graph-ignorable. The choices we made for $p(\alpha)$ and $p(\mathcal{G} \mid \alpha)$ were based on considerations such as simplicity and computational convenience. In principle, nothing prevents the reader from using a different random graph specification. Specifying the term $p(Y \mid \mathcal{G}, \gamma)$ is more delicate: Using a different MRF model would imply substantial changes in the MCMC procedure. Moving away from the MRF assumption would be even more challenging, and therefore an interesting line of research.

References

[1] Joseph Blitzstein and Sergiy Nesterko. Bias-variance and breath-depth tradeoffs in respondent-driven sampling. 2012.

[2] M. de Mello, A.A. Pinho, M. Chinaglia, W. Tun, A. Barbosa Junior, and M.C.F.J. Ilario. Assessment of risk factors for hiv infection among men who have sex with men in the metropolitan area of campinas city, brazil, using respondent-driven sampling. Technical report, Washington DC: Population Council, 2008.

[3] Paul Erdos and A. Renyi. The evolution of random graphs. *Magyar Tud. Akad. Mat. Kutato Int. Kolz*, 5:17–61, 1960.

[4] Krista J. Gile. Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association*, 106:135–146, 2011.

[5] L. A. Goodman. Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170, 1961.

[6] Mark S. Handcock and Krista J. Gile. Modeling social networks from sampled data. 2011.

[7] Douglas D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44:174–199, 1997.

[8] Douglas D. Heckathorn. Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential degree. *Sociological Methodology*, 37:151–207, 2007.

[9] D.F. Heitjan and Donald B. Rubin. Ignorability and coarse data. *Annals of Statistics*, 19:2244–2253, 1991.

[10] A. Raftery, D. Madigan, and C. Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). volume 5 of *Bayesian Statistics*, pages 323–349. Oxford University Press, 1996.

[11] Christian P. Robert. *The Bayesian Choice, Second Edition*. Springer-Verlag, 2001.

[12] Donald B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[13] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, 1987.

[14] Matthew J. Salganik and Douglas D. Heckathorn. Sampling and estimation in hidden populations using respondent driven sampling. *Sociological Methodology*, 34:193–239, 2004.

[15] Erik Volz and Douglas D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24:79–97, 2008.