Bayesian Inference from Non-Ignorable Network Sampling Designs

Simón Lunáñez1, Edouard Airoldi2,3

1Harvard University Department of Statistics, 2Harvard FAS Center for Systems Biology

Introduction

Consider individuals interacting in a social network and a response that can be measured on each individual. We are interested in making inferences on a population quantity \( q \) that is a function of both the response and the social interactions. In this work, we rely on the notion of ignorability for coarse data proposed by Heitjan and Rubin. This notion is the key element for assessing when to include the function of the sampling mechanism in the likelihood. A design that has to be included in the likelihood in order to obtain legitimate inferences is called non-ignorable. The objective of this work is to propose a general framework for performing Bayesian inference when the sampling mechanism on the social network is non-ignorable.

On Ignorability

There are situations where computing the likelihood depends not only on who is included in the sample, but also on how the sample was collected:

- The probability distribution of the sampling design may depend on features corresponding the portion of the graph that was not sampled.
- Two realizations of the sampling mechanism can lead to the same observed subgraph, but the probabilities associated to the samples are different.

General Model & Assumptions

We assume a data generative model of the form:

\[ p(Y; G, n, \alpha, \beta) = p(Y|G) p(G|n, \alpha, \beta) \]

where we denote \( Y \) as the realization of a random graph with parameter vector \( \alpha \). The distribution of the sampling mechanism \( G \) is conditional on a given realization of \( \alpha \); it is not necessary to know how the graph was generated to sample \( \alpha \). We assume that \( \alpha \) induces a dependence structure on the response vector \( Y \). To fully specify such distribution, additional parameters (denoted by \( \beta \)) are needed, that explains the dependence among \( \alpha \) and \( Y \).

We included the factor \( p(Y; G) \) to deal with ignorability issues. Still, we need to model the partial observed data, that is the purpose of additional components: \( p(G|n, \alpha, \beta) \). We assume that the dependence structure modeled by \( p(Y; G) \) is a Markov Random Field (MRF).

Inference

- Let \( Y_{INC} \) and \( Y_{EXC} \) denote, respectively, the observed and partially observed portions of \( G \);
- \( Y_{INC} \) and \( Y_{EXC} \) are defined in an analogous manner for the response vector;
- \( Y_{INC} \) and \( Y_{EXC} \) have to be augmented to compute \( p(Y; G) \);
- There is uncertainty regarding how many nodes to augment;
- The observed data contains little information about how augment the graph.

Model Specification

For an Erdős-Rényi model was assumed \( 3, \) with a single probability of inclusion \( \alpha \). A Beta\((\alpha,\omega)\) was used as prior for \( \alpha \). Our specification for \( p(\alpha) \) is an RDS \([4]\) with \( m \) coupons per wave and sample size \( n \). For this paper we will assume that \( \alpha \) is binary, and

\[ p(\alpha = 1 | \text{INC}, \text{EXC}) = \alpha + \sum_{i \in \text{INC}} (1 - \alpha) \]

where \( \text{INC} \) is the adjacency matrix for \( \gamma \) and \( \text{EXC} \) is the set of nodes adjacent to \( \text{INC} \). In other words \( p(Y; G) \) is specified as a MRF based on a probit model. We used a scaled Beta as prior for \( \alpha \), i.e.

\[ p(\alpha | \nu, \eta, \theta) \sim B(\nu, \eta) \]

and a similar distribution as prior for \( \nu \) and \( \eta \),

\[ p(\nu, \eta | \nu_0, \eta_0) \sim B(\nu, \eta) \]

MCMC Sampling

The proposals we used to update each parameter \((Y_{INC},G_{INC},G_{EXC},\nu,\eta)\) are the following:

- For \( \nu \), we use a mixture kernel where one component is a random walk reflecting \( \nu \) and \( \overline{\nu} \), the second component is the prior (Equation \( 1 \)) and the third is a uniform distribution over \((0, \overline{\nu})\).
- For \( \eta \), we use a mixture kernel where one component is a random walk reflecting \( \theta \) and \( 0 \), the second component is the prior (Equation \( 1 \)) and the third is a uniform distribution over \((-\infty, 0)\).
- For the extra-sample edges \( G_{EXC} \) (connect a sampled node to a non-sampled one) we take each augmented node and count its number of neighbors \( k \) according to the current version of \( G \), then we delete the corresponding edges and then obtain a random sample of size \( n_k \) over the nodes for which \( k = 1 \). New edges are added which connect those \( n_k \) nodes to it.
- For the intra-sample nodes in \( G_{INC} \) (connect two sampled nodes) we first delete all the current intra-sample edges and then add a new set of edges at random while respecting the restriction that no edge will exist between nodes included in the same wave and preserving the current density of the graph.

Results

Simulation Study

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

- 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 \( k \). 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 model the low dependence \((\phi = 0.82, \zeta = 0.01)\), and high dependence \((\phi = -1.1, \zeta = 0.15)\).
- 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 5 coupons in all cases. The specification of the MRF that we used in combination with the random graph model imply \( Q = 2 \) in all scenarios.

- For each scenario we simulated 500 datasets. The average bias for each regime is shown in the table.

Real Data

- We applied our methodology to the data discussed in de Mello, Pinho, et al.\([2]\). The table summarizes the results they obtained:

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 Camarinas in the state of Sao Paulo, Brazil. Since RDS was used, non-ignorability is an issue. The aim of the study was to infer the prevalence of HIV among gay men in Camarinas, Brazil. The study covered 4595 gay men who have sex with men. The inclusion criteria used for this study, were:

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

RDS was implemented using \( m = 10 \) seeds and 5 coupons per subject (\( m = 3 \)).

Discussion

- Our methodology is the first one that includes a probability model that accounts for ignorance.
- We are able to deal with non-ignorable designs.
- The approach proposed is highly modular: We allow for different distributional specifications for each of the model’s components.

References & Acknowledgments

This research was supported by the National Science Foundation (DMS-1212702) and by the National Institutes of Health (R01MH107315). Simon Lunáñez is supported by the National Science Foundation (DMS-1816544). We would like to thank David Chew and Minyi Hong for their discussions on the issue of non-ignorable sampling for RDS. We are also grateful to the Assistant Editors and the referees for their valuable comments that led to a significant improvement of this paper.