Analytic Methods for Network Data

by
Miles Q. Ott
BA, Smith College, 2001
MPH, University of Minnesota, 2007
ScM, Harvard University, 2009

A Dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Biostatistics at Brown University

Providence, Rhode Island
May 2014

This dissertation by Miles Q. Ott is accepted in its present form
by the Department of Biostatistics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Joseph Hogan, Advisor

Recommended to the Graduate Council

Date _____ _____

Matthew Harrison, Reader

Date _____ _____

Nancy Barnett, Reader

Date _____ _____

Krista Gile, Reader

Approved by the Graduate Council

Date _____ _____

Peter Webber, Dean of the Graduate School

Vitae

Miles Ott
Date of Birth: January 24, 1979
Place of Birth: Ithaca, New York
121 South Main Street, Suite 743A
Providence, Rhode Island 02903
miles_ott@brown.edu 415-203-3656

Education

PhD in Biostatistics, Expected August 2013
Brown University, Providence RI
Thesis: Analytic Methods for Social Network Data
Advisor: Dr. Joseph Hogan

MS in Biostatistics, 2009
Harvard University, Boston MA
Concentration in Women, Gender, and Health
Advisor: Dr. David Wypij

MPH in Epidemiology, 2007
University of Minnesota, Minneapolis MN
Advisor: Dr. J. Michael Oakes

BA in Mathematics, 2001
Smith College, Northampton MA
Minor in Computer Science
Advisor: Dr. Ruth Haas

Teaching Experience

Brown University

Statistical Inference (Teaching Assistant), Fall 2012
Held office hours, graded homework assignments and exams for a class of 9 graduate students.

Exploring Probability and Statistics (Instructor), Summer 2012
Designed and taught an introductory statistics class for 15 incoming undergraduate students.

Statistics Review for Incoming Graduate Students (Instructor), Summer 2010, 2011
Designed and taught modules for incoming Biostatistics graduate students.

Introduction to Biostatistics (Teaching Assistant), Fall 2009
Taught lab sections, held office hours, graded assignments for a class of 40 undergraduate and graduate students.

Harvard University

Linear and Longitudinal Regression (Teaching Assistant), Summer 2008, 2009
Held office hours, graded homework assignments and exams for a class of 50 graduate students.

Analysis of Rates and Proportions (Teaching Assistant), Spring 2009
Taught lab sections, held office hours, graded assignments for a section of 41 graduate students.

Principles of Biostatistics (Teaching Assistant), Fall 2007, 2008
Taught lab sections, held office hours, graded assignments for a section of 43 graduate students.

University of Minnesota

Cancer Epidemiology (Teaching Assistant), Spring 2007
Held office hours, graded homework assignments and exams for a class of 28 graduate students.

Smith College

Calculus I, Calculus II, Discrete Math, Linear Algebra (Teaching Assistant), 1998-2001
Held office hours for a variety of undergraduate classes.

Professional Development

Teaching

Brown University, Harriet W. Sheridan Center for Teaching and Learning, *Sheridan Teaching Seminar*, 2009

Brown University, Harriet W. Sheridan Center for Teaching and Learning, *Professional Development Seminar*, 2012

Brown University, Watson Center for Information Technology, *Canvas Course Development*, 2012

Research

University of Michigan, ICPSR, *A Second Course in Network Analysis*, Summer 2010

Brown University, Division of Biology and Medicine, *Responsible Conduct in Research Training*, Fall 2009

Research Experience

Brown University Department of Community Health, Providence RI
Research Assistant (2011 to Present)
Described the social networks of women with advanced cancer, and analyzed the association between network attributes and advanced care planning decisions.

Brown University Department of Behavioral and Social Sciences, Providence RI
Research Assistant (2010 to Present)
Analyzed a social network drawn from an undergraduate residence hall. Employed network auto-correlation models and community identification methods to identify associations between network characteristics and substance use.

Africa Centre for Health and Population Studies, KwaZulu-Natal South Africa
Research Assistant (Summer 2010)
Investigated age-mixing in sexual relationships with mixing matrices, assortativity indices, and regressions of age-gaps on relationship and individual attributes. Used spatial statistics to identify clusters of highly disparate age-mixing patterns.

Channing Laboratory of Brigham and Womens Hospital, Boston MA
Research Assistant (2008 to 2010)
Collaborated with a multi-disciplinary team of social epidemiologists, psychologists, and statisticians to examine the patterns of sexual orientation mobility in the Growing Up Today Study cohort. Characterized the association between sexual orientation mobility and health risk behaviors.

Publications

Nancy Barnett, **Miles Ott**, Michelle Rogers, Michelle Loxley, Crystal Linkletter, Melissa Clark. Peer associations for substance use and exercise in a college student social network. *Health Psychology.* (In Press).

Miles Ott, David Wypij, Heather Corliss, Margaret Rosario, Sari Reisner, Allegra Gordon, S. Bryn Austin. Repeated changes in reported sexual orientation identity linked to substance use behaviors in youth. *Journal of Adolescent Health.* 2013, 52(4): 465-472.

Miles Ott, Frank Tanser, Mark Lurie, Till Barnighausen, Marie-Louise Newell. Age-gaps in sexual partnerships in rural KwaZulu-Natal: seeing beyond sugar daddies. *AIDS.* 2011, 25(6):861-863.

Miles Ott, Heather L. Corliss, David Wypij, Margaret Rosario, S. Bryn Austin. Stability and Change in Self-Reported Sexual Orientation Identity in Young People: Application of Mobility Metrics. *Archives of Sexual Behavior.* 2011: 1-14.

Miles Ott, Shelly F. Shaw, Richard Danilla, Ruth Lynfield. Lessons Learned from the 1918-1919 Influenza Pandemic in Minneapolis and St. Paul, Minnesota. *Public Health Reports.* 2007, 122: 803-809.

Presentations

Miles Ott, Krista Gile, Anneli Uuskula, Lisa G. Johnston. Spatial Dependencies in Respondent-Driven Sampling Data, Sunbelt XXXII Conference of the International Network for Social Network Analysis, Redondo Beach California, 2012.

Nancy Barnett, **Miles Ott**, Crystal Linkletter, Michelle Rogers, Michelle Loxley, M, Melissa Clark. Alcohol use in one university dormitory: A social network analysis. Annual meeting of the Society for the Study of Emerging Adulthood, Providence Rhode Island, 2011.

Miles Ott, Crystal Linkletter, Nancy Barnett. Bayesian Calibration of Peer Perceptions and Self-Reports on Networks. Graduate Student Research Seminar Series, Brown University, Providence Rhode Island, 2011.

Miles Ott, Shelly F. Shaw, Richard Danilla, Ruth Lynfield. Lessons Learned from the 1918-1919 Influenza Pandemic in Minneapolis and St. Paul, Minnesota. Minnesota Department of Health Immunization Conference, Minneapolis Minnesota, 2007.

Posters

Miles Ott, Joseph Hogan, Krista Gile, Crystal Linkletter, Nancy Barnett. Bayesian Comparative Calibration of Self and Peer Reports of Alcohol Use on a Social Network. ENAR, Orlando Florida, 2013.

Miles Ott, Joseph Hogan, Krista Gile, Crystal Linkletter, Nancy Barnett. Bayesian Comparative Calibration of Self and Peer Reports of Alcohol Use on a Social Network. Workshop on Computational and Online Social Science, Columbia University, New York New York, 2012.

Miles Ott, Crystal Linkletter, Nancy Barnett. Bayesian Calibration of Peer Perceptions and Self-Reports on Networks. Public Health Research Day, Brown University, Providence Rhode Island, 2012.

Miles Ott, Till Barnighaussen, Frank Tanser, Mark Lurie, Mary-Louise Newell. Age-mixing in sexual relationships in rural KwaZulu-Natal: seeing beyond 'sugar daddies'. Conference on Retroviruses and Opportunistic Infections, Boston Massachusetts, 2011.

Miles Ott, Heather L. Corliss, David Wypij, Margaret Rosario, S. Bryn Austin. Changes in Sexual Orientation Linked to Poorer Health Outcomes in Youth, Society of Adolescent Medicine, Toronto Canada, 2010.

Invited Lectures

Brown University, Applied Research Methods, Methods for Sampling Hard to Reach Populations, April 1, 2013.

Carleton College, Mathematics Colloquium, Social Networks and Statistics: It's Who You Know, March 8, 2013.

University of Massachusetts Amherst, Statistics and Probability Seminar, Bayesian Social Network Calibration with Application To Alcohol Use, November 19, 2012.

Brown University, Introduction to Public Health Guest Lecture, Social Networks and Public Health, November 16, 2012.

Interdisciplinary Working Groups

Hard-to-Reach Populations Methods Research Group, Cross-Institutional

Service

Service to Profession

Eastern North American Region of the International Biometric Society Graduate Student and Recent Graduate Council, Council Member, 2012

Brown University

Biostatistics Department Curriculum Committee, Student Member, 2011

Graduate Program Steering Committee, Student Member, 2011

Public Health Journal Club, Coordinator, 2010

Harvard University

Women Gender and Health Steering Committee, 2007 to 2009

Transgender Health Speaker Series, Co-organizer, 2008

Smith College

Mathematics Department Student Liaison, 2000 to 2001

Other Positions

Minnesota Department of Health, St. Paul MN
Senior Student Worker (2005 to 2007)
Assisted with influenza surveillance for the eight county metropolitan area surrounding Minneapolis and St. Paul Minnesota. Conducted chart reviews to gather data on pediatric influenza. Investigated responses to the 1918 influenza pandemic in Minneapolis and St. Paul.

Houghton Mifflin Company: College Division, Boston MA
Mathematics Software Consultant (Summer 2000)
Identified and solved problems in algebra tutorial software using a proprietary language. Proofread software for content and accuracy.

Awards and Honors

2012 Brown University Art of Science: Best of Show Award
Workshop on Computational and Online Social Sciences Travel Scholarship
Framework in Global Health Scholarship, Brown University
Lester Breslow Scholarship, University of Minnesota

Acknowledgements

I am forever indebted to my wonderful teachers, my wise mentors, and my support network of family and friends. These include: Joseph Hogan, Nancy Barnett, Matt Harrison, Krista Gile, Melissa Clark, Crystal Linkletter, David Wypij, Bryn Austin, Owen Marciano, Henry Schneiderman, Edward Ott, Mary Ott and Bill Ott. I would especially like to thank Ben Capistrant for his unfailing support, sage advice, and humor.

Contents

List of Tables

List of Figures

Abstract

Analytic Methods for Network Data

Miles Q. Ott

Chair: Professor Joseph Hogan

This dissertation presents methods for statistical analysis of social network data. First, we develop a Bayesian hierarchical model that calibrates self and peer-reports in order to synthesize information collected from multiple sources in the social network context. Secondly we demonstrate how current methods for analyzing fixed choice design social network data can induce biases. We propose a new survey design which collects information on the total number of relationships. Lastly, we investigate how assumptions in prevalence estimation in network sampling designs can produce biased estimates and introduce a new prevalence estimator for network sampling studies that offers considerable improvement over existing estimators.

CHAPTER 1

INTRODUCTION

Social network analysis is used to investigate potential dependencies between social relationships and behaviors and attitudes (O'Malley and Marsden, 2008). The number of publications on social networks has grown exponentially over the last 30 years (Knoke and Yang, 2008). The dramatic increase in the use of social network analysis methods is largely due to two factors: the adoption of a systemic rather than an individual approach, and the increased ability to collect, store, and manage ever larger datasets (Kolaczyk, 2009). The rapidly growing field of social network analysis holds great promise to integrate multidimensional data, describe complex social structures, and allow for approximations of probability samples. However, these complex social structures are difficult to analyze, and new methodologies must be developed to capitalize on the opportunities that social network analysis offers and to address the limitations that exist in current methods.

In this dissertation, we will develop new methods for analyzing social network data. First, we will introduce a new method for synthesizing information collected from multiple sources in the social network context. Secondly we will improve upon existing methods for analyzing design-induced missingness in social network studies and introduce a novel social network survey design. Lastly, we will introduce a new prevalence estimator for network sampling studies that offers considerable improvement over existing estimators.

1.1 Multiple information sources

One way to sample substance use data on a network is to include survey questions that not only ask the subject to self-report information, but also prompt the subjects to report

on their peers (or alters) in an attempt to gain insight into the structure of relationships and peer-perceptions. Such network surveys may gather information on a complete network (wherein each person in a bounded population reports on all of their associates or alters), or a more local subset of the network centered on each subject (ego-networks). For instance, in the context of studying minors with psychological issues, parents, teachers, and other alters of the child may be asked to report on the child's health and behaviors, forming an ego-network centered on each child in the study called a "multiple informants" data (Horton et al., 2008; Lash et al., 2002). However, surveying individuals about their associates or peers, it is necessary to consider peer-reporting discrepancies, since respondents may have perceptions that over or under-report values of their peers. In the second chapter of the dissertation we introduce a Bayesian hierarchical model that calibrates self-reports and peer-reports.

1.2 Missing data on a closed population

When summarizing information on a network, there is often the possibility that nominations are missing, that individuals in the population are missing, or potentially both are missing (Kossinets, 2006). Earlier work has focused on missingness that is induced by non-response (Robins et al., 2004; Costenbader and Valente, 2003; Huisman and Steglich, 2008; Gile and Handcock, 2006). Due to the complex dependence structure inherent in networks, missing data can pose very difficult problems (Holland and Leinhardt, 1973; Marsden, 1990). For example, it is unclear how to handle this missingness when analyzing the network for peer effects on behavior, or when analyzing structural measures of the network, such as network density, the clustering coefficient, average shortest path on the network, reciprocity of relationships, and transitivity of relationships.

Some missingness is a result of study design. For example, common practice in network studies is to limit the number of possible nominations that each person in the network can make, thus inducing potentially missing nominations (Kossinets, 2006; Holland and Leinhardt, 1973; Yan and Gregory, 2011). Several recent studies of networks have this limited choice design (Witvliet et al., 2010; Mercken et al., 2010; Conlan et al., 2010; Ali

et al., 2011).

Holland and Leinhardt (1973) first introduced the problem of missing ties due to limited choice design (also called fixed choice design and right-censoring of degree). Kossinets (2006) went on to show how the limited choice design can lead to biases in estimates of structural measures of the network.

In the third chapter of the dissertation we demonstrate how current methods for analyzing fixed choice design data can induce biases. We propose a new survey design which collects information on the total number of relationships, and demonstrate through simulation studies the improvement for inference with this new design.

1.3 Respondent-Driven Sampling

In public health, it is often the case that researchers wish to learn about populations that are at an elevated risk for adverse health events. For example, increased attention has been given to injection drug users, commercial sex workers, and men who have sex with men in HIV surveillance (Malekinejad et al., 2008; Magnani et al., 2005). In order to learn about the prevalence of HIV in these populations, it is necessary to construct a sample with a sampling frame. However these populations, commonly termed "hidden populations" or "hard to reach populations", are difficult to construct sampling frames for.

There have been several proposed methods of sampling hidden populations (Semaan, 2010), including snowball sampling, time-space sampling and respondent-driven sampling (RDS). RDS was initially proposed by Heckathorn (Heckathorn, 1997, 2002) as a means of sampling populations without an identifiable sampling frame. Since the initial RDS paper, this method has been widely adopted to sample hidden populations and has been used in hundreds of studies both domestically and internationally (Malekinejad et al., 2008; Johnston et al., 2008).

The RDS method is a variant of a link-tracing procedure (Handcock and Gile, 2011a). In link-tracing, a number of individuals from the target population are enrolled into the study as 'seeds', who are then asked to recruit other people they know to be members of the target population, who are then asked to enroll those they know to be members of the

target population, and so on. Using the sample mean from a link-tracing sample has some inherent flaws, including the fact that the sample mean does not account for non-equal inclusion probabilities (individuals who have more friends are more likely to be included in the sample).

Figure 1.1 shows the sampling structure of an RDS toy example. In this example we see that the sample starts with two seeds, and each participant recruits another two participants. Participants are sampled without replacement. RDS proceeds as a link-tracing sample, with weighting to account for different inclusion probabilities. There are different methods of estimating the inclusion probabilities, and most are a function of the reported number of associates each person has (their degree), thus accounting for the tendency of people who have more friends to have a higher probability of being included in the sample (Gile, 2011; Volz and Heckathorn, 2008; Gile and Handcock, 2011).

In order for RDS to produce consistent estimates, many assumptions must be met. We will primarily concern ourselves with the assumption that sampling is with replacement, demonstrate how violations of this assumptions can incur biased prevalence estimates, and propose an improved prevalence estimator. We will demonstrate the method through simulation studies, and apply the method to a real data set.

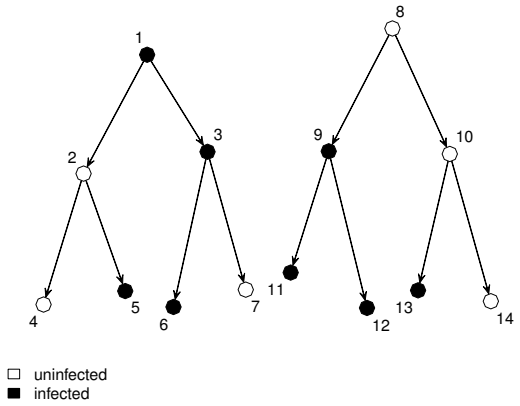

Figure 1.1: *RDS Simple Example*

CHAPTER 2

BAYESIAN PEER CALIBRATION BASED ON NETWORK POSITION WITH APPLICATION TO ALCOHOL USE

2.1 Introduction

Gold-standard measurements of individual-level alcohol consumption are rarely collected (Stahre et al., 2006; Snell et al., 2011; Gmel et al., 2006). In the absence of a gold standard, self-reports or peer-reports are utilized as proxies (Gmel et al., 2006). The utility of peer-reports, termed collateral reports in the alcohol literature, has been extensively studied and reviewed in both non-college settings (Connors and Maisto, 2003) and in the college setting (Borsari and Muellerleile, 2009). In the college setting, while self-reports and peer-reports are often in agreement (Connors and Maisto, 2003; Borsari and Muellerleile, 2009), peer-reports of drinking tend to be higher than self-reports (Hagman et al., 2007). There is general agreement in the alcohol literature that college students' peer-reports of alcohol consumption are biased towards over-reporting (Baer and Carney, 1993; Borsari and Carey, 2003). Hence both self-reports and peer-reports contain information about the true level of alcohol consumption, particularly if the discrepancies (i.e., the amount of over or under-reporting) contained within the peer-reports can be accounted for.

Measurement error due to self-reports of health behaviors, such as physical activity (Nusser et al., 2012), food consumption (Dodd et al., 2006), and alcohol consumption (Kraus and Augustin, 2001), is an important issue because ignoring measurement error may lead to incorrect inference (Buonaccorsi and Lin, 2002; Richardson and Gilks, 1993). Existing measurement error models typically assume self-reports are unbiased (Dodd et al., 2006) or contain systematic biases (Beyler, 2010; Nusser et al., 2012; Spiegelman et al., 1997; Ferrari

et al., 2007). In the presence of self-report error, another type of data on the health-related behavior is collected on each individual. Using both types of data, a method of estimation (i.e. method of moments) is employed to estimate all parameters in the measurement error models as well as the quantity of interest.

Another way of dealing with measurement error is absolute calibration (Brown, 1982; Osborne, 1991; Barnett, 1969), in which both a gold standard and error-prone measures are collected on a limited number of data points in order to infer the relationship between the gold standard and the error-prone measures. This relationship is then exploited to allow for measurement with the error-prone measure. Comparative calibration, which uses information from two or more error-prone measurement instruments, is applied when a gold standard is not available but there is prior information about the structure of the measurement error of each of the instruments (Theobald and Mallinson, 1978; Osborne, 1991). Comparative calibration has been implemented in both the frequentist (Theobald and Mallinson, 1978; Gimenez and Patat, 2005; Kimura, 1992; Lu et al., 1997) and the Bayesian setting (Blangiardo and Richardson, 2008). Osborne (1991) has an extensive review of both absolute and comparative calibration approaches.

Both measurement error and calibration models deal with the situation where we are interested in variable Y, but instead we observe a proxy variable X that is related to Y (Stefanski, 2000), and both types of models allow that X may be biased for Y, prone to error, or both. One contrast between the two methods is that calibration models require more than one proxy variable for Y, say X and W, whereas measurement error models do not require this (Nusser et al., 1996). Further, measurement error models seek to identify and decompose the different sources of bias and variation that are included in the proxy measurements (Fuller, 1995), whereas calibration models aim to learn about the measurement errors, but do not decompose these different sources of bias and variation included in the proxy measurements. Further, measurement error models may investigate measurement error by incorporating covariates into the estimation of the underlying values of interest in a linear mixed effects model (Tooze et al., 2010), akin to a calibration model with a zero-bias assumption. Alternately, when multiple sources of information are available, measurement error models may allow for the measurement error to be modeled with covariates(McIntyre

and Stefanski, 2011). However, these measurement error models do not jointly estimate the relations between covariates and the underlying values of interest and the unknown bias, as we do in our bayesian peer calibration with covariates model which we present below. Further, among measurement error models that make use of multiple sources, none are situated in a social network context which necessarily implies that individuals only report on their associates rather than all members of the network.

Social networks provide a setting in which there are multiple sources of information which are not gold-standards. Consider a social network in which individuals report on both themselves and their peers in the network. Neither self nor peer-reports are ideal measurements, but both types of reports contain information about the quantity of interest. In this paper we present a novel Bayesian comparative calibration model framework: Bayesian Peer Calibration (BPC) which utilizes both self-reports and peer-reports of alcohol consumption. We assume self-reports are unbiased but measured with error. Peer reports may be biased and are also assumed to be measured with error.

2.1.1 Motivating Study

The data motivating this analysis are drawn from a study of alcohol and drug use in a social network formed by college students residing in a freshman dormitory(Barnett et al., 2013). The primary objective of our analysis is to characterize the alcohol consumption of the participants through the use of self and peer-reports of the number of alcoholic drinks consumed on drinking days. In particular, this study seeks to identify characteristics that are associated with alcohol consumption. Though previous studies have collected both self and peer-reports of alcohol consumption, few have used both sources of information to estimate alcohol consumption. This paper is concerned with how to use both information sources which may provide contradictory information on alcohol consumption. In order to address this question we use a Bayesian Peer Calibration (BPC) model on the social network from the motiviating study(Figure 2.1) to generate the posterior distributions of individual-level alcohol consumption while using both sources of information. In Section 2.2 we develop the BPC models (both with and without covariates) for binary, count, and continuous data, in Section 2.3 we apply these models to the data, in Section 2.4 we perform

a simulation to investigate the model's sensitivity to assumptions, and in Section 2.5 we present a brief discussion.

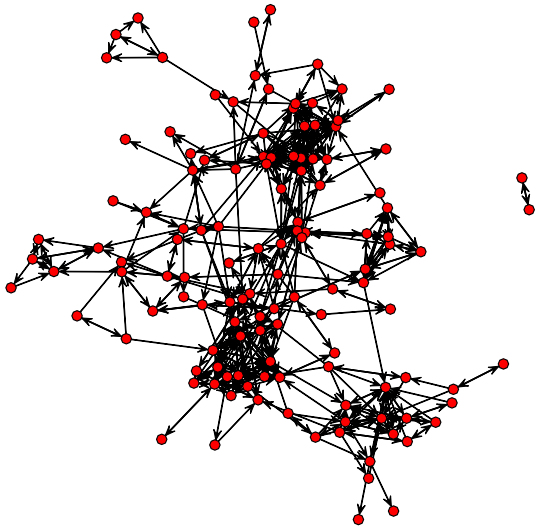

Figure 2.1: Network of College Students

2.2 Bayesian Peer Calibration Models

2.2.1 Overview

We consider data drawn from a network of n nodes (representing members of the social network) with e directed edges (representing the presence or absence of peer nominations), represented by an $n \times n$ adjacency matrix A, where $A_{ij} \in \{0, 1\}$ and equals 1 if node i reports on node j (denoted as $i \rightarrow j$). For each of the n nodes in the network we have a self-report on a quantity of interest, denoted by z_i. Furthermore, for each node, there are reports on the same quantity of interest of their nominated alters, denoted by y_{ij} which is the peer-report of node j on node i. The observed data at each node i is $(z_i, y_{ij} : j \rightarrow i)$. We assume self-reports are made with mean-zero error and peer-reports are potentially

subject to some systematic bias which we call a discrepancy. We assume that the true underlying quantity of interest is θ_i. The self-report z_i is an unbiased but possibly error-prone proxy for θ_i, such that $\mathrm{E}(z_i) = \theta_i$. The peer-reports y_{ij} are subject to a systematic discrepancy that can be additive, multiplicative, or exponential depending on the type of outcome. Specifically, if we denote the discrepancy by γ_{ij}, the peer-reports can follow $\mathrm{E}(y_{ij}) = \theta_i + \gamma_{ij}$, $\mathrm{E}(y_{ij}) = \theta_i\gamma_{ij}$, or $\mathrm{E}(y_{ij}) = \theta_i^{\gamma_{ij}}$.

We now introduce a three-level model which may include covariates, which we call the Bayesian Peer Calibration with Covariates(BPCC) model, or not include covariates, which we call the Bayesian Peer Calibration model (BPC). The first level of the model specifies the joint distribution of the observed data at each node, conditional on parameters θ_i and γ_j that respectively capture the node-specific mean and the alter-specific discrepancy. The second level characterizes variation in θ_i and γ_j, potentially allowing both to be influenced by covariates. At the third level, priors are introduced as needed.

We provide specific model formulations for continuous, binary, and count versions of z and y. In our application to the available data, we have interval-censored count data; hence we focus on posterior inference for this case. The appendices contain information about posterior inference for other cases.

2.2.2 Bayesian Peer Calibration Model

The general form of the first level of the BPC model is

$$\begin{aligned} z_i &\sim F(\theta_i) \\ y_{ij} &\sim G(\theta_i, \gamma_{ij}), \end{aligned}$$

where the distributions F, G and the support of θ_i and γ_{ij} depend on the choice of the model. We treat three parametric formulations here. For Normal y and z, we assume

$$\begin{aligned} z_i &\sim N(\theta_i, {\sigma_z}^2) \\ y_{ij} &\sim N(\theta_i + \gamma_{ij}, {\sigma_y}^2), \end{aligned}$$

where $\sigma_y^2 > 0$ and $\sigma_z^2 > 0$ are the variance parameters, and θ_i and γ_{ij} are real numbers. If y and z are Bernoulli distributed,

$$\begin{aligned} z_i &\sim \text{Ber}(\theta_i) \\ y_{ij} &\sim \text{Ber}(\theta_i^{\gamma_{ij}}), \end{aligned}$$

where θ_i is between 0 and 1, and γ_{ij} is a positive real number. Lastly, if y and z are Poisson distributed,

$$\begin{aligned} z_i &\sim \text{Poi}(\theta_i) \\ y_{ij} &\sim \text{Poi}(\theta_i \gamma_{ij}), \end{aligned}$$

and θ_i and γ_{ij} are positive real numbers. In each of these model formulations, the self-reports are assumed to be unbiased.

Level 1 can be parameterized in terms of generalized linear models for z and y, where

$$\begin{aligned} g\{\text{E}(z_i)\} &= \alpha_i \\ g\{\text{E}(y_{ij})\} &= \alpha_i + \phi_{ij}, \end{aligned}$$

where $g(\cdot)$ is an appropriate link function. The GLM parameters map directly to the parameters in the cases above. For the normal case and identity link $g(x) = x$, $\alpha_i = \theta_i$ and $\phi_{ij} = \gamma_{ij}$. For binary data and log-log link $g(x) = \log\{-\log(x)\}$, we have $\alpha_i = \log\{-\log(\theta_i)\}$ and $\phi_{ij} = \log(\gamma_{ij})$. Finally, in the case of count data, using the link $g(x) = \log(x)$ yields $\alpha_i = \log(\theta_i)$, and $\phi_{ij} = \log(\gamma_{ij})$.

The GLM formulation easily accommodates covariate information, and allows the analyst to put separate priors on the mean and variance parameters. However, except for the normal case, the scaling of the GLM parameters may not always be natural for the application at hand. The parameterizations given above allows for more natural interpretation of model parameters and use conjugate priors for easier computation. Additionally, the posteriors for the model parameters have intuitive forms that transparently show from

where the information is derived. Incorporation of covariates needs to be done carefully, but we provide full details for each formulation of the model.

In addition to assuming that self-reports are unbiased, we also assume that for all i, j $\gamma_{1j} = \gamma_{2j} = \cdots = \gamma_{nj} = \gamma_j$ for $j = 1, \ldots, n$ which implies that each individual j has a constant discrepancy γ_j when reporting on peers. Finally we assume $\theta_i \perp\!\!\!\perp \gamma_j$, $\theta_i \perp\!\!\!\perp \theta_j$ and that $\gamma_i \perp\!\!\!\perp \gamma_j$. These constraints impose structure that is most easily understood in the continuous-data normal distribution case. Specifically, we have $E(z_i \,|\, \theta_i) = \theta_i$, $\mathrm{Var}(z_i \,|\, \theta_i) = \sigma_z^2$, $\mathrm{Cov}(y_{ij}, y_{ik} \,|\, \theta_i) = \mathrm{Var}(\theta_i)$, $\mathrm{Cov}(y_{ij}, z_i \,|\, \theta_i) = \mathrm{Var}(\theta_i)$, $\mathrm{Cov}(y_{ij}, z_i \,|\, \theta_i, \gamma_j) = 0$, and $\mathrm{Cov}(y_{ij}, y_{ik} \,|\, \theta_i, \gamma_j, \gamma_k) = 0$.

We proceed by specifying level two of the BPC model. We focus on the formulation for count data due to our application. We assume that both θ_i and γ_j are each independent and Gamma distributed with shape parameters τ_θ, τ_γ and rate parameters $\kappa_\theta, \kappa_\gamma$ such that $\mathrm{E}(\theta_i) = \tau_\theta/\kappa_\theta, \mathrm{E}(\gamma_j) = \tau_\gamma/\kappa_\gamma$; this is denoted by

$$\begin{aligned} \theta_i &\sim \mathrm{Gam}(\tau_\theta, \kappa_\theta) \\ \gamma_j &\sim \mathrm{Gam}(\tau_\gamma, \kappa_\gamma). \end{aligned}$$

The third level of the model specifies the priors. In our application, we use diffuse Gamma priors with hyperparameters a, b on $\tau_\theta, \kappa_\theta, \tau_\gamma, \kappa_\gamma$ where a is the shape parameter and b is the rate parameter. Details regarding the full conditionals are in the appendices.

The BPCC model is elaborated to include covariate effects in the components characterizing both the quantity of interest θ and the peer-reporting discrepancy γ. Carrying forward the Poisson formulation, we again assume that both θ and γ are Gamma distributed, the difference being that the expected value of both θ and γ are now functions of the covariates:

$$\begin{aligned} \theta_i | X_i, \beta, \omega_\theta &\sim \mathrm{Gam}(\omega_\theta e^{X_i\beta}, \omega_\theta) \\ \gamma_j | X_j, \alpha, \omega_\alpha &\sim \mathrm{Gam}(\omega_\gamma e^{X_i\alpha}, \omega_\gamma). \end{aligned}$$

Hence, $\log \mathrm{E}(\theta_i) = X_i\beta$ and $\log \mathrm{E}(\gamma_j) = X_j\alpha$. The ω_θ and ω_γ parameters help determine the variance of θ and γ respectively such that the $\mathrm{Var}(\theta_i) = e^{X_i\beta}/\omega_\theta$. For instance, if ω_θ is large relative to $e^{X_i\beta}$ then $\mathrm{Var}(\theta_i)$ is small. This BPCC model specification also allows us to gain insight into the relationship between the covariates and θ_i and γ_j

In the third level of the BPCC model, we specify diffuse zero-mean Normal priors for β and α with a large variance σ^2, and diffuse Gamma priors for rate parameters ω_θ and ω_α with hyperparameters f and g. The full conditionals are detailed in the appendix.

2.2.3 Posterior inference for count data

Under the BPC model framework (which does not include covariates), for each iteration of the MCMC we draw from the posterior distribution using full conditionals which have the following Gamma distributions:

$$\begin{aligned} P(\theta_i|\gamma, \theta_{-i}, \tau_\theta, \kappa_\theta, \tau_\gamma, \kappa_\gamma, a, b, z, y) &\sim \mathrm{Gam}(z_i + \sum_{j:j\to i} y_{ij} + \tau_\theta, 1 + \sum_{j:j\to i} \gamma_j + \kappa_\theta) \\ P(\gamma_j|\theta, \gamma_{-j}, \tau_\theta, \kappa_\theta, \tau_\gamma, \kappa_\gamma, a, b, z, y) &\sim \mathrm{Gam}(\sum_{i:j\to i} y_{ij} + \tau_\gamma, \sum_{i:j\to i} \theta_i + \kappa_\gamma). \end{aligned}$$

Full conditionals of the other parameters in this model are detailed in the appendices. The expected value of the full conditionals of θ_i can be instructive:

$$\frac{z_i + \sum_{j:j\to i} y_{ij} + \tau_\theta}{1 + \sum_{j:j\to i} \gamma_j + \kappa_\theta} =$$

$$\frac{z_i}{1 + \sum_{j:j\to i} \gamma_j + \kappa_\theta} + \frac{\sum_{j:j\to i} y_{ij}}{\sum_{j:j\to i} \gamma_j} \frac{\sum_{j:j\to i} \gamma_j}{1 + \sum_{j:j\to i} \gamma_j + \kappa_\theta} + \frac{\tau_\theta}{\kappa_\theta} \frac{\kappa_\theta}{1 + \sum_{j:j\to i} \gamma_j + \kappa_\theta},$$

which is a weighted average of the self-report of person i (z_i), the discrepancy corrected peer-reports on person i, $\frac{\sum_{j:j\to i} y_{ij}}{\sum_{j:j\to i} \gamma_j}$, and the prior information about the mean value of the

θ parameters, $\tau_\theta/\kappa_\theta$. We also obtain full conditionals of γ_j which have the expected value:

$$\frac{\sum_{i:j\to i} y_{ij} + \tau_\gamma}{\sum_{i:j\to i} \theta_i + \kappa_\gamma} = \frac{\sum_{i:j\to i} y_{ij}}{\sum_{i:j\to i} \theta_i} \frac{\sum_{i:j\to i} \theta_i}{\sum_{i:j\to i} \theta_i + \kappa_\gamma} + \frac{\tau_\gamma}{\kappa_\gamma} \frac{\kappa_\gamma}{\sum_{i:j\to i} \theta_i + \kappa_\gamma},$$

which is a weighted average of $\frac{\sum_{i:j\to i} y_{ij}}{\sum_{i:j\to i} \theta_i}$ and $\tau_\gamma/\kappa_\gamma$, the prior information about the mean value of the γ parameters.

In the BPCC model, which includes covariates, for each MCMC step we draw from the posterior distribution using full conditionals having the following Gamma distributions:

$$\begin{aligned} P(\theta_i|\theta_{-\mathbf{i}}, \gamma, \alpha, \beta, \omega_\theta, \omega_\gamma, \mathbf{x}, \mathbf{y}, \mathbf{z}) &\sim \text{Gam}(z_i + \omega_\theta e^{X_i\beta} + \sum_{j:j\to i} y_{ij}, 1 + \omega_\theta + \sum_{j:j\to i} \gamma_j) \\ P(\gamma_j|\theta, \gamma_{-\mathbf{j}}, \alpha, \beta, \omega_\theta, \omega_\gamma, \mathbf{x}, \mathbf{y}, \mathbf{z}) &\sim \text{Gam}(\omega_\gamma e^{X_j\alpha} + \sum_{i:j\to i} y_{ij}, \omega_\gamma + \sum_{i:j\to i} \theta_i). \end{aligned}$$

In the BPCC model, we obtain posterior draws of θ_i which are weighted averages of the self-report of person i (z_i), the discrepancy corrected peer-reports on person i $\frac{\sum_{j:j\to i} y_{ij}}{\sum_{j:j\to i} \gamma_j}$, and the information about the mean value of the θ parameter for those with covariate profile X_i, $e^{X_i\beta}$. We also obtain posterior draws of γ_j which are weighted averages of the peer-reports' discrepancy for person j, $\frac{\sum_{i:j\to i} y_{ij}}{\sum_{i:j\to i} \theta_i}$, with $e^{X_j\alpha}$, the posterior the mean value of the peer-reporting discrepancy for those with covariate profile X_j. Similar to the BPC, the the posterior distributions of the θ parameters are less diffuse when there is more precise information about the γ parameters, and vice-versa.

With the use of covariates in the BPCC model, we can gain insight into the relationships between the covariates with the peer-reporting discrepancies γ and the quantity of interest θ. Additionally, by including covariates to model θ and γ we are relaxing the independence assumption in the BPC model by only imposing conditional independence between θ and γ such that $\theta_i \perp \gamma_j | X_i, X_j$.

It is important to note that these models are subject to several assumptions. The assumption that the θ and γ parameters are independent (in the BPC model) or conditionally independent (in the BPCC model), may not hold. Additionally, the assumption that all self-reports have no systematic bias may not be true. We proceed with these assumptions and will address them in the simulation and discussion sections.

2.3 Application

The objective of this analysis is to learn about the unknown average number of alcoholic drinks consumed on drinking days, and the association between certain personal characteristics and alcohol consumption among study participants. To do so, we formulate Bayesian comparative calibration models (BPC, BPCC) that allow for drawing from the posterior distribution of peer-reporting discrepancies, in reports of alcohol consumption, and use the peer-reports corrected for the posterior draws of the discrepancies along with self-reports to learn about the posterior distribution of the number of alcoholic drinks consumed on drinking days for each participant in the study. We wish to learn how covariates are associated with the number of alcoholic drinks consumed on drinking days as well as which covariates are associated with over or under reporting the alcohol consumption of peers.

We specify diffuse priors on $\omega_\theta, \omega_\gamma, \beta$, and α by specifying the hyper-parameters $f = 0.2$ and $g = 0.2$ and $\sigma^2 = 1000$. The f and g values specify a prior mean of $0.2/0.2 = 1$ and a prior variance of $0.2/0.2^2 = 5$ for both $\omega_\theta, \omega_\gamma$.

We carry out 35,000 iterations of Gibbs Sampling, and of these, discard the first 10,000 as burn-in and use every fifth of the final 25,000 simulations, leaving us with 5,000 simulations to calculate our posterior values of parameters.

2.3.1 Data

In the social network study, data are collected from residents of a freshman dormitory. Each participant is 18 years old or older when the survey is administered and is asked to report the number of alcoholic drinks they consume on drinking days (self-reports) as either 0, 1-2, 3-4, 5-6, or 7 or more drinks. Further, participants are asked to provide demographic information, including their age, sex, race, sexual orientation, and whether or not they intend to join a fraternity or sorority. Additionally, each participant is asked to nominate which of the other participants are important to them. If participant A nominates participant B as important to them, then B is an alter of A. Each participant can nominate alters and serve as an alter for other participants. Lastly, each participant is asked to report the number of alcoholic drinks each of their alters consume on drinking days (0, 1-2, 3-4, 5-6,

or 7 or more drinks), which we refer to as peer-reports. By using nominations to connect participants, we construct a network of social relationships (Figure 2.1).

Of the 129 participants included in the sample, 4 did not nominate alters and were not nominated, and are excluded from the present analysis. This leaves 125 participants who are included in the analysis, with 507 nominations. Of these 507 nominations, there are 366 peer-reports on alcohol consumption on 108 different dorm residents. In the sample of 125 participants, 47.2% are male, 24% report considering joining a fraternity or sorority, 90.4% identify as heterosexual, and 59.2% identify as white. Among the 108 participants whose alters report on their alcohol consumption, 48.1% are male, 23.1% report considering joining a fraternity or sorority, 88.9% identify as heterosexual, and 62.0% identify as as white. Self-reports and peer-reports of alcohol consumption are presented in Figure 2.2, and peer-reports stratified by self-reports are presented in Figure 2.3. The most common self-report of alcohol consumption was 3-4 drinks on drinking days, with 47 out of 129 providing this self-report (Figure 2.2). Further information about the UrWeb Study can be found in Barnett et al. (2013).

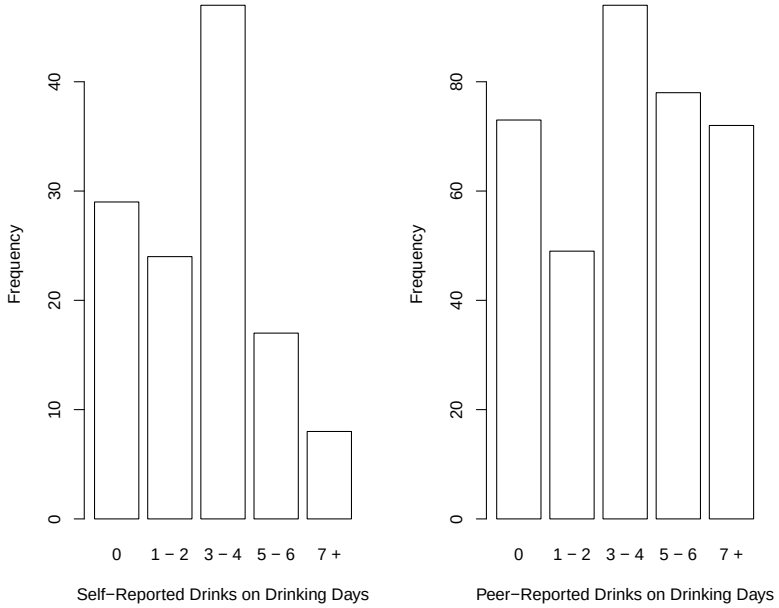

Figure 2.2: Self-reports and peer-reports of number of drinks on drinking days

Because the number of drinks are reported in intervals (0, 1-2, 3-4, 5-6, 7+) we perform a data augmentation step for each of the 35,000 simulations. For simulation t, the self-report

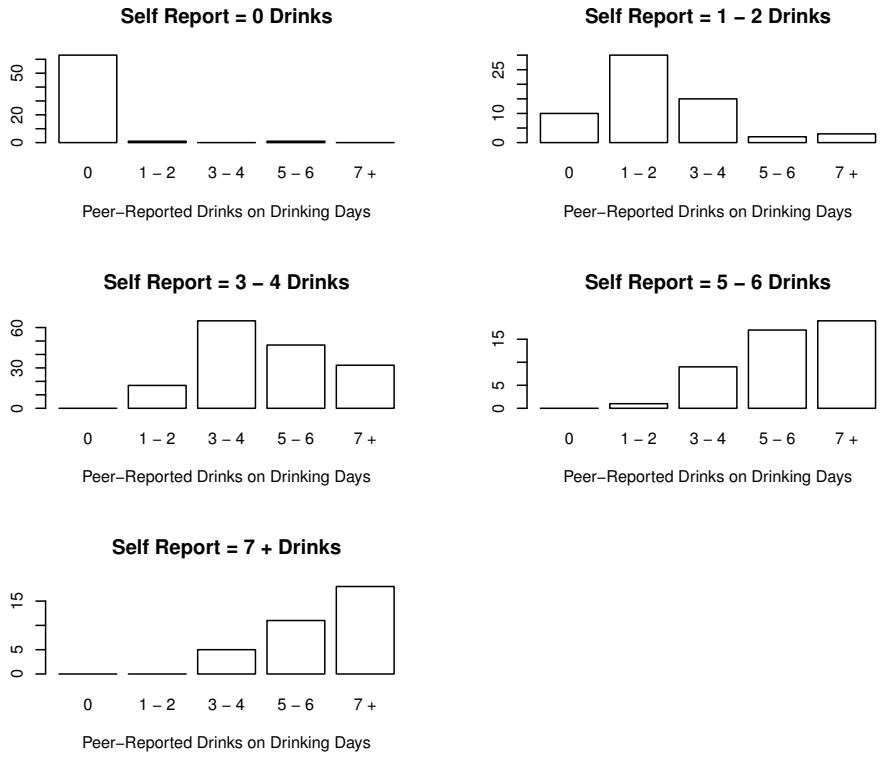

Figure 2.3: Peer-reports on number of drinks on drinking days by self-reported number of drinks on drinking days

for i is generated from a truncated Poisson distribution $z_{it} \sim \text{Poi}(\theta_{it-1})$ such that each z_{it} value is within the specified interval of the self-report. Similarly, for simulation t, the alter report made by j on i is generated $y_{ijt} \sim \text{Poi}(\theta_{it-1}\gamma_{jt-1})$.

2.3.2 Model Specification Without Covariates

In the BPC model θ and γ are the parameters of primary interest. The distribution of posterior means for these parameters is shown in Figure 2.4. The posterior draws of θ stratified by self-reported drinks on drinking days are presented in the first column of Figure 2.5; we see that while the posterior distributions of the θ parameters are related to the self-reported number of drinks on drinking days, there is variation within strata which indicates that the information from the peer-reports and the self-reports has been combined to give us a different result than if we had ignored the peer-report information. We assumed diffuse priors by specifying the hyper-parameters $a = 0.02$ and $b = 0.02$.

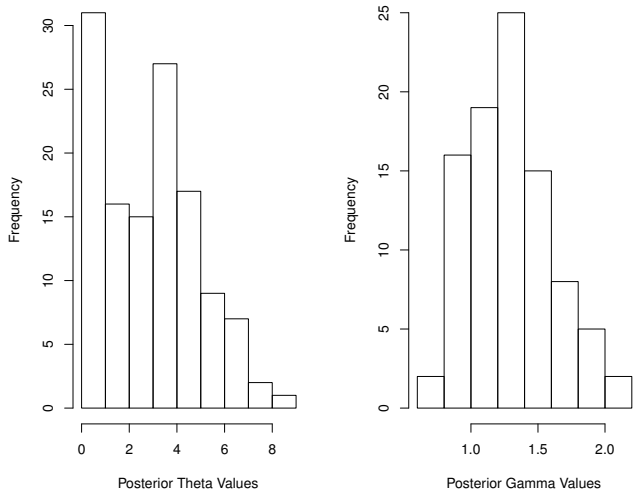

Figure 2.4: Posterior θs and posterior γs from BPC model

2.3.3 Bayesian Peer Calibration with Covariates

The parameters of most interest in the BPCC model are β and α which indicate the extent to which being male, having an interest in joining a fraternity or sorority, having a heterosexual sexual orientation, being white, and self-reporting zero drinks(used for α only) are associated with alcohol consumption and peer-report discrepancies respectively. We present the posterior values of β and α in Table 2.1. For the posterior means of the β parameters, we can see that being male, interest in joining a fraternity or sorority, heterosexual sexual orientation and white race are associated with higher number of drinks on drinking days, though only interest in joining a fraternity or sorority and white race have 95% credible intervals that does not contain zero. For example, average consumption for a white male who is heterosexual and is interested in joining a fraternity is $\exp(0.44 + 0.26 + 0.36 + 0.00 + 0.55) = 5.00$ drinks on drinking days, whereas we expect a *non-white* male who is heterosexual and is interested in joining a fraternity or sorority to drink $\exp(0.44 + 0.26 + 0.36 + 0.00) = 2.89$ drinks on drinking days. For the posterior means of the α parameters, we can see that interest in joining a fraternity or sorority and self-reporting zero drinks on drinking days are both

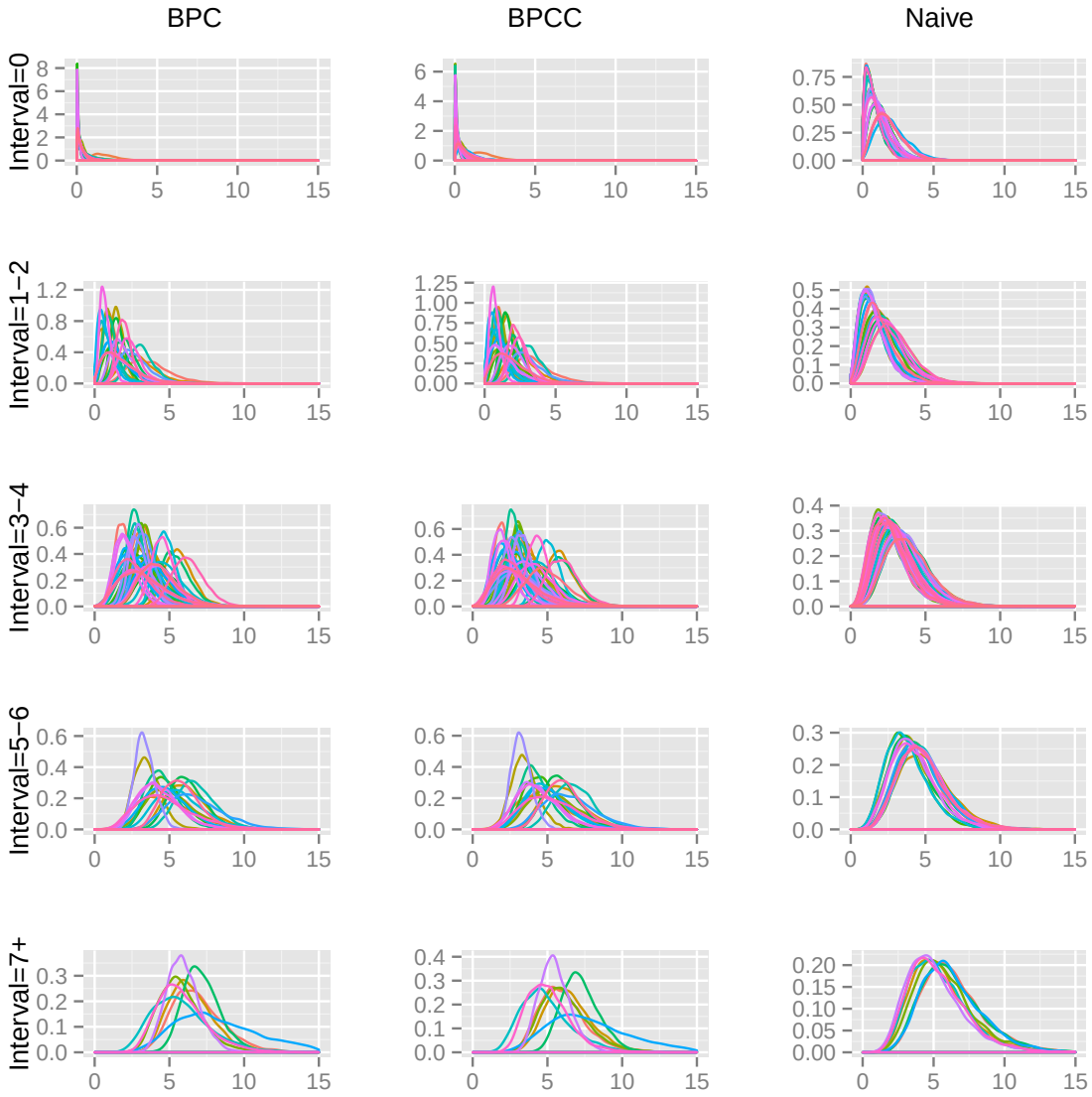

Figure 2.5: Posterior densities of θs stratified by observed self-reports

associated with having discrepancies that could represent under-reporting, assuming that self reports are unbiased. For example, we expect that a white male who is heterosexual, is interested in joining a fraternity, and reported drinking zero drinks would have multiplicative peer-reporting discrepancies of $\exp(0.52 + 0.24 - 0.30 - 0.25 - 0.02 - 0.81) = 0.54$, or under-reporting peer-alcohol consumption by about half.

Next we turn our attention to the distribution of the posterior values of the θ and γ parameters which represent the number of drinks consumed on drinking days, and discrepancies in reporting on an alter's drinks on drinking days respectively. These are displayed

Table 2.1: Posterior Mean Values and 95% CI of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$

Effect	Discrepancy BPCC α	Drinks Per Day BPCC β	Drinks Per Day Naive β
Intercept	0.52 (0.09, 0.97)	0.44 (-0.15, 1.01)	0.53 (-0.10, 1.10)
Male	0.24 (-0.01, 0.50)	0.26 (-0.06, 0.57)	0.28 (-0.03, 0.58)
Frat/Sorority	-0.30 (-0.59, -0.02)	0.36 (0.02, 0.70)	0.12 (-0.21, 0.45)
Heterosexual	-0.25 (-0.64, 0.10)	0.00 (-0.47, 0.52)	-0.06 (-0.55, 0.48)
White Race	-0.02 (-0.29, 0.25)	0.55 (0.22, 0.90)	0.57 (0.25, 0.93)
Self-Reported 0	-0.81 (-1.46, -0.28)		

in Figure 2.6. The posterior distributions of θ stratified by self-reported drinks on drinking days are presented in the second column of Figure 2.5. Similar to the BPC model, we can see that the comparative calibration has utilized information from both the self and peer-reports since dorm-residents with the same self-reports have varying posterior predictive distributions of the θ which reflect the influence of the peer-reports.

2.3.4 Role of Peer-Reports

To understand the role of peer-reports we look at a model with only self-reports (naive model). We present results from the BPCC where we ignore the peer-report information, so that we can learn about the effect of peer reports on both covariates as well as on the model fit. Since only self-reports are considered, this Naive model is not a calibration model. In this model we are primarily interested in the β parameters, and present the posterior means and credible intervals for these parameters in the third column of Table 2.1. For the posterior means of the β parameters, we can see that male sex, interest in joining a fraternity or sorority, heterosexual sexual orientation and white race are associated with higher number of drinks on drinking days, though only white race has a 95% credible interval that does not contain zero. Comparing the posterior distributions for the β parameters from the model without peer-reports to the β parameters from the BPCC model (Table 2.1), we see that there are some differences. Most notably, the 95% credible interval for the β corresponding to interest in joining a fraternity or sorority does not contain zero in the posterior distribution from the BPCC model, but does contain zero in the posterior distribution from the model without peer-reports. In other words, if we only used self-reports we expect that those interested in joining a fraternity or sorority drink $e^{0.12} = 1.13$

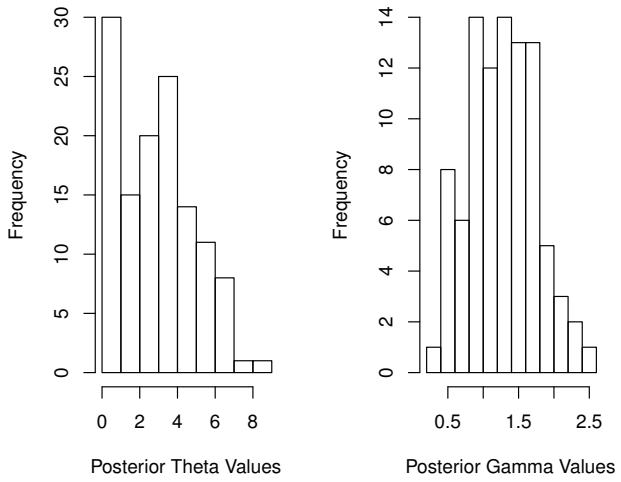

Figure 2.6: Posterior θs and Posterior γs from BPCC Model

times the number of drinks than someone who is not considering joining a fraternity or sorority with a 95% credible interval that includes 1: (0.81, 1.57), whereas if we consider both self-reports and peer reports, we expect that those joining a fraternity or sorority have a $e^{0.36} = 1.43$ times the number of drinks as someone not interested in joining a fraternity or sorority with a 95% credible interval that includes does not include 1: (1.02, 2.01). This demonstrates that the inclusion of the peer-reports may result in different inferences than had we not included peer-reports in our models.

Comparing Model Fit

In order to assess the fit of these models, we generated 5,000 posterior predictive draws of the self-reports z. In order to compare the fit of the BPC and BPCC models, we found the proportion of posterior predictive z values that were within the interval observed in

Table 2.2: Posterior Predictive $\boldsymbol{z}$, generated by Posterior θ

	% of posterior $\boldsymbol{z}$ in interval		
Reported Interval	Without Covariates	With Covariates	Without Peer-Reports
0	75	71	40
1 - 2	46	46	46
3 - 4	33	32	32
5 - 6	25	26	24
7+	45	41	36

the data. These are presented in the first two columns of Table 2.2. We also found the proportion of posterior predictive z and values that were within the interval observed in the data for the model without peer-reports, and present these in the third column of Table 2.2.

By comparing the posterior predictive z's to the observed z's, we can determine whether the fit of the model seems reasonable, and compare the model fits for the models. Here we can see that the BPC and BPCC models have a comparable fit to the data. For example, about 46% of the posterior predictive z values for participants reporting 1-2 drinks were 1 or 2 for both the BPC and BPCC models. While the use of covariates does not seem to markedly change the model fit, the use of peer-reports seems to improve the model fit, particularly for participants who reported zero drinks.

2.4 Simulations Allowing for Self-Report Bias

We carried out simulations to investigate how the BPCC model performs when the assumption of unbiased self-reports is violated. We base these simulations on the UrWeb data, and we only use the Fraternity/Sorority covariate. We then generated simulated data under three conditions: 1.) self-reports are unbiased, 2.) self-reports are positively biased, with bias unrelated to the covariates in the BPCC model, and 3.) self-report bias is a function of the covariates included in the BPCC model. In each simulation we set $\alpha_1 = 0.52, \alpha_2 = -0.30, \beta_1 = 0.44, \beta_2 = 0.35$, which are the posterior means for the intercepts and the Fraternity/Sorority parameters from the analysis in Section 2.33. We set both ω_θ and ω_γ equal to 2.

Table 2.3: Effect of self-report bias on posterior means from BPCC with 100 simulated datasets

Simulation	Empirical mean of posterior means	Absolute Bias	Empirical SD of posterior means
Unbiased self-reports			
α_1	0.521	0.001	0.101
α_2	-0.296	0.004	0.166
β_1	0.433	-0.007	0.097
β_2	0.350	-0.009	0.141
Self-report bias $\perp X$			
α_1	0.728	0.208	0.085
α_2	-0.306	-0.006	0.161
β_1	0.652	0.212	0.085
β_2	0.371	0.011	0.112
Self-report bias function of X			
α_1	0.520	<0.001	0.166
α_2	-0.079	0.221	0.097
β_1	0.432	-0.008	0.093
β_2	0.570	0.210	0.117

In each simulation we first generate θ and γ values from $\theta \sim$ Gamma$(\omega_\theta e^{X\beta}, \omega_\theta)$ and $\gamma \sim$ Gamma$(\omega_\gamma e^{X\beta}, \omega_\gamma)$. Next we generate the self reports (z) and the peer-reports (y) from $z_i \sim \text{Pois}(\delta_i \theta_i)$ and $y_{ij} \sim \text{Pois}(\theta_i \gamma_j)$. In the case of the simulations where self-reports are unbiased, we set $\delta_i = 1$. When self-reports are positively biased but that bias is unrelated to the covariates in the model, we set $\delta_i \sim \text{Unif}(1, 1.5)$. Finally, where self-report bias is a function of the covariate $\delta_i, wehave \sim (1 - X_i) + X_i \text{Unif}(1, 1.5)$. One hundred simulated datasets were generated under each of the 3 different settings of self-report bias.

For each simulated dataset we carried out 5,200 iterations of Gibbs sampling, and discarded the first 200 of these, leaving 5,000 simulations to calculate posterior mean values of the α and β parameters. We present the empirical mean of the posterior means, the bias, and the empirical standard deviation of the posterior means in Table 2.3.

We can see that in the simulations in which self-reports are generated without bias, the BPCC performs well with relatively small biases, and standard deviations. In the second set of simulations, where self-reports are generated with a bias that is unrelated

to the covariates, we find that the posterior means of the intercept terms are biased but the posterior means of α_2 and β_2 seem to have very small biases, indicating that covariate effects are accurately captured. Finally, in the set of simulations where self-report bias is a function of the covariate in the model, we see that the intercept terms seem to have very small biases, but the posterior means of α_2 and β_2 seem to be affected by the self-report bias. This simulation suggests that BPCC properly captures the relationship between covariates and the quantity of interest when there is no self-reporting bias, and when self-reporting bias is not a function of the covariates of interest, but that regression coefficients may be biased when self-report bias is related to the covariates.

2.5 Discussion

We show how to use self and peer-reports to learn about both individual level alcohol consumption and peer-reporting discrepancies. Further, we show that the use of peer-reports improves model fit. Among college students, peer-reports of alcohol consumption are often thought to be overestimates (Hagman et al., 2007). Under the assumptions of the BPC model framework, we have found that a vast majority of the peer-reports indeed had positive discrepancies. Males tended to have larger discrepancies than females, those intending to join a fraternity or sorority, and those who self-reported zero drinks on drinking days having more negative discrepancies than those who do not intend to joining a fraternity or sorority or reported at least one drink on drinking days, respectively. We also found that people who identified as white had higher levels of alcohol consumption than those who did not, and that those intending to join a fraternity or sorority have a higher level of alcohol consumption than those not intending to join.

This method may be applied to any situation where members of a network report measurements on the same quantities of interest for peers as well as themselves. Any network study that asks for self and peer-reports, and has a similar error structure, could make use of this method. Consider, for example, an online rating system wherein users provide ratings of different products or services. By applying a BPC model, the network of users and ratings could be leveraged to learn about both raters' reporting biases as well as an

underlying "true" rating. By applying this method to an online-rating system, products or services that were rated by those with a negative bias could have their over-all rating corrected, thereby reflecting its true quality rather than the biases of the raters.

With the increase of available network data (Knoke and Yang, 2008), there is a growing need to develop methods that allow for the integration of information from different sources within the network. Bayesian comparative calibration offers a coherent way to utilize data on the same quantity which come from many sources. Prior comparative calibration formulations assumed that there are $p > 1$ imperfect measurement instruments, and that each of these instruments are used to measure the same quantity from n object, resulting in $n \times p$ measurements (Kimura, 1992; Blangiardo and Richardson, 2008). In our models, we generalize prior work in Bayesian comparative calibration in three ways 1.) we do not require that each measurement instrument(either a self-report or a peer-report) is applied to measure the quantity of interest from each object (the members of the social network) 2.) we have formulated models for continuous, count, and binary data 3.) we jointly model the covariate effects between both discrepancies and the quantity of interest.

We plan to extend these models in order to relax some of the assumptions that are made. In this paper, we assume that self-reports are unbiased on average. In future work we will allow for self-reports to contain a systematic bias by placing a prior on a self-report discrepancy. Our simulations of data that contain both peer-report and self-report bias show that the BPCC model performs well when self-reporting bias is independent of the covariates included in the models. In the BPCC model, we make the assumption that alcohol consumption and peer-reporting discrepancies are independent given covariate values. Since we know the network structure of this data, we could allow for there to be correlation of these quantities between individuals who are connected in the network.

2.6 Appendix

2.6.1 Model Specification and Full Conditions for Normal Data Bayesian Peer Calibration

We assume that θ and γ are independent and Normally distributed,$\theta \sim N(\mu_\theta, \sigma_\theta^2), \gamma \sim N(\mu_\gamma, \sigma_\gamma^2)$ that the self-reports are on average error-free, and that the peer-reports are subject to an additive systematic discrepancy γ_j.

We assume conjugate prior distributions for $\mu_\theta, \mu_\gamma, \sigma_\theta^2, \sigma_\gamma^2, \sigma_y^2$ and σ_z^2 specifying diffuse prior distributions. Normal zero-centered priors are placed on μ_θ and μ_γ, with variance term ϕ. The variance terms $\sigma_\theta^2, \sigma_\gamma^2, \sigma_y^2$ and σ_z^2 are given Inverse-Gamma prior with shape a and rate s. These prior specifications result in the following full conditionals, where $t_i = \frac{\sum_j y_{ij} - \gamma_j}{p_i}$, $w_j = \frac{\sum_i y_{ij} - \theta_i}{r_j}$, p_i = number of peer-reports on person i, r_j = number of reports that person j gave, $\bar{\theta} = \frac{\sum_i \theta_i}{n}$, $\bar{\gamma} = \frac{\sum_j \gamma_j}{g}$, and g = the number of people who made reports. This yields full conditionals:

$$
\begin{aligned}
P(\theta_i|\boldsymbol{\gamma}, \sigma_y^2, \sigma_z^2, \sigma_\theta^2, \mu_\theta, y_{i.}, z_i) \sim \\
\mathrm{N}\left[\frac{t_i\sigma_z^2\sigma_\theta^2 + z_i\sigma_\theta^2\sigma_y^2/p_i + \mu_\theta\sigma_y^2\sigma_z^2/p_i}{\sigma_z^2\sigma_\theta^2 + \sigma_\theta^2\sigma_y^2/p_i + \sigma_z^2\sigma_y^2/p_i}, \frac{\sigma_z^2\sigma_\theta^2\sigma_y^2/p_i}{\sigma_z^2\sigma_\theta^2 + \sigma_\theta^2\sigma_y^2/p_i + \sigma_z^2\sigma_y^2/p_i}\right] \\
P(\gamma_j|\boldsymbol{\theta}, \sigma_z^2, \sigma_\gamma^2, \mu_\gamma, y_{.j}) \sim \mathrm{N}\left[\frac{w_j\sigma_\gamma^2 + \mu_\gamma\sigma_y^2/r_j}{\sigma_\gamma^2 + \sigma_y^2/r_j}, \frac{\sigma_\gamma^2\sigma_y^2/r_j}{\sigma_\gamma^2 + \sigma_y^2/r_j}\right] \\
P(\mu_\theta|\phi, \bar{\theta}, \sigma_\theta^2) \sim \mathrm{N}\left[\frac{\bar{\theta}\phi}{\phi + \sigma_\theta^2/n}, \frac{\phi\sigma_\theta^2/n}{\phi + \sigma_\theta^2/n}\right] \\
P(\mu_\gamma|\phi, \bar{\gamma}, \sigma_\gamma^2,) \sim \mathrm{N}\left[\frac{\bar{\gamma}\phi}{\phi + \sigma_\gamma^2/g}, \frac{\phi\sigma_\gamma^2/g}{\phi + \sigma_\gamma^2/g}\right] \\
P(\sigma_z^2|\boldsymbol{\theta}, \mathbf{y}, \boldsymbol{\gamma}) \sim \text{Inv-Gam}\left[a + e/2, \left(\frac{1}{s} + \frac{1}{2}\sum_i (y_{i,j} - \theta_i - \gamma_j)^2\right)^{-1}\right] \\
P(\sigma_\theta^2|\boldsymbol{\theta}, \mathbf{z}) \sim \text{Inv-Gam}\left[a + n/2, \left(\frac{1}{s} + \frac{1}{2}\sum_i (\theta_i - \mu_\theta)^2\right)^{-1}\right] \\
P(\sigma_\gamma^2|\boldsymbol{\gamma}, \mathbf{z}) \sim \text{Inv-Gam}\left[a + g/2, \left(\frac{1}{s} + \frac{1}{2}\sum_j (\gamma_j - \mu_\gamma)^2\right)^{-1}\right].
\end{aligned}
$$

2.6.2 Model Specification and Full Conditionals for Normal Data Bayesian Peer Calibration with Covariates

Here, we specify θ_i and γ_j as being Normally distributed with means and variances $X_i\beta_\theta, \sigma^2_\beta$ and $X_j\alpha_\gamma, \sigma^2_\alpha$ respectively. Assuming diffuse prior distributions such that: $\mathrm{P}(\beta_\theta, \sigma^2_\beta) \propto \sigma^{-2}_\beta, \mathrm{P}(\alpha_\gamma, \sigma^2_\alpha) \propto \sigma^{-2}_\alpha, \mathrm{P}(\sigma^2_y) \sim \text{Inv-Gam}(a, s),$ and $\mathrm{P}(\sigma^2_z) \sim \text{Inv-Gam}(a, s)$ we get the following full conditionals:

$$P(\gamma_j|\boldsymbol{\theta}, \sigma^2_z, \sigma^2_\alpha, y_{.j}, \alpha_\gamma, X_j) \sim \mathrm{N}\left[\frac{w_j\sigma^2_\alpha + X_j\alpha_\gamma\sigma^2_y/r_j}{\sigma^2_\alpha + \sigma^2_y/r_j}, \frac{\sigma^2_\alpha\sigma^2_y/r_j}{\sigma^2_\alpha + \sigma^2_y/r_j}\right]$$

$$P(\theta_i|\boldsymbol{\gamma}, \sigma^2_z, \sigma^2_y, \sigma^2_\beta, \beta_\theta, y_{i.}, z_i, X_i) \sim$$

$$\mathrm{N}\left[\frac{t_i\sigma^2_z\sigma^2_\beta + z_i\sigma^2_\beta\sigma^2_y/p_i + X_i\beta_\theta\sigma^2_z\sigma^2_y/p_i}{\sigma^2_z\sigma^2_\beta + \sigma^2_\beta\sigma^2_y/p_i + \sigma^2_z\tau^2_y/p_i}, \frac{\sigma^2_z\sigma^2_\beta\sigma^2_y/p_i}{\sigma^2_z\sigma^2_\beta + \sigma^2_\beta\sigma^2_y/p_i + \sigma^2_z\sigma^2_y/p_i}\right]$$

$$P(\beta_\theta|\boldsymbol{\theta}, \sigma^2_\beta, \mathbf{X}) \sim \mathrm{MVN}\left[\hat{\beta}_\theta, (X'X)^{-1}\sigma^2_\beta\right]$$

$$P(\alpha_\gamma|\boldsymbol{\gamma}, \sigma^2_\alpha, \mathbf{X}) \sim \mathrm{MVN}\left[\hat{\alpha_\gamma}, (X'X)^{-1}\sigma^2_\gamma\right]$$

$$P(\sigma^2_z|\boldsymbol{\theta}, \mathbf{z}) \sim \text{Inv-Gam}\left[a + n/2, \left(\frac{1}{s} + \frac{1}{2}\sum_i (z_i - \theta_i)^2\right)^{-1}\right]$$

$$P(\sigma^2_y|\boldsymbol{\theta}, \mathbf{y}, \boldsymbol{\gamma}) \sim \text{Inv-Gam}\left[a + e/2, \left(\frac{1}{s} + \frac{1}{2}\sum_i (y_{i,j} - \theta_i - \gamma_j)^2\right)^{-1}\right]$$

$$P(\sigma^2_\beta|\boldsymbol{\theta}, \mathbf{X}) \sim \text{Inv-}\chi^2\left[n - k, s^2_\beta\right]$$

$$P(\sigma^2_\alpha|\boldsymbol{\gamma}, \mathbf{X}) \sim \text{Inv} - \chi^2\left[g - k, s^2_\alpha\right],$$

where $\hat{\beta}_\theta = (X'X)^{-1}X'\boldsymbol{\theta}$, $\hat{\alpha_\gamma} = (X'X)^{-1}X'\boldsymbol{\gamma}$, k is the number of parameters used in the model, $s^2_\beta = \frac{1}{n-k}(\boldsymbol{\theta} - \mathbf{X}\hat{\beta}_\theta)'(\boldsymbol{\theta} - \mathbf{X}\hat{\beta}_\theta)$, and $s^2_\alpha = \frac{1}{g-k}(\boldsymbol{\gamma} - \mathbf{X}\hat{\alpha_\gamma})'(\boldsymbol{\gamma} - \mathbf{X}\hat{\alpha_\gamma})$.

2.6.3 Model Specification and Full Conditionals for Bayesian Peer Calibration with Bernoulli data

We assume that θ and γ are independent and Bernoulli distributed, that the self-reports are on average error-free, and that the peer-reports are subject to an exponential systematic discrepancy given by γ_j such that: $z_i \sim \text{Bern}(\theta_i)$ and $y_{ij} \sim \text{Bern}(\theta_i^{\gamma_j})$. Further, $\theta_i \sim$

Beta(a, b) and $\gamma_j \sim \text{Gam}(c, d)$. We specify diffuse Gamma priors for a, b, c, d with shape and rate parameters ω and ψ. This results in the following full conditionals:

$$\begin{aligned}
P(\theta_i|\boldsymbol{\gamma}, a, b, c, d, y_{i\cdot}, z_i) &\propto \theta_i^{a+z_i+\sum_{j:j\to i}\gamma_j y_{ij}-1}(1-\theta_i)^{b-z_i}\prod_{j:j\to i}(1-\theta_i^{\gamma_j})^{1-y_{ij}} \\
P(\gamma_j|\boldsymbol{\theta}, a, b, c, d, y, z) &\propto \gamma_j^{c-1}e^{-d\gamma_j}\prod_{i:j\to i}\theta_i^{\gamma_j y_{ij}}(1-\theta_i^{\gamma_j})^{1-y_{ij}} \\
P(a|\boldsymbol{\gamma}, \boldsymbol{\theta}, z, y, b, c, d, \omega, \psi) &\propto a^{\omega-1}e^{-\psi a}\prod_i \frac{\theta_i^a}{\text{Beta}(a,b)} \\
P(b|\boldsymbol{\gamma}, \boldsymbol{\theta}, z, y, a, c, d, \omega, \psi) &\propto b^{\omega-1}e^{-\psi b}\prod_i \frac{(1-\theta_i)^b}{\text{Beta}(a,b)} \\
P(c|\boldsymbol{\gamma}, \boldsymbol{\theta}, z, y, a, b, d, \omega, \psi) &\propto c^{\omega-1}e^{-\psi c}\prod_j \frac{(d\gamma_j)^c}{\Gamma(c)} \\
P(d|\boldsymbol{\gamma}, \boldsymbol{\theta}, z, y, a, b, c, \omega, \psi) &\sim \text{Gam}(\omega + \sum_j c, \psi + \sum_j \gamma_j).
\end{aligned}$$

2.6.4 Model Specification and Full Conditionals for Bayesian Peer Calibration with Covariates with Bernoulli data

We assume that θ and γ are independent and Bernoulli distributed, that the self-reports are on average error-free, and that the peer-reports are subject to an exponential systematic discrepancy given by γ_j, as with the BPC model. We specify distributions on θ_i and γ_j that make use of covariate information: $\theta_i \sim \text{Beta}(a, \frac{a}{e^{X\beta}}), \gamma_j \sim \text{Gam}(\omega_\gamma e^{X\alpha}, \omega_\gamma)$, such that the expected value of θ_i equals $\frac{e^{X_i\beta}}{1+e^{X_i\beta}}$. We specify diffuse Gamma priors for a, ω_γ with shape and rate parameters, and diffuse Normal priors on α, β which result in the following full conditionals:

$$\begin{aligned}
P(\theta_i|\boldsymbol{\gamma}, a, \beta, x_i, \omega_\gamma, y_{i\cdot}, z_i) &\propto \theta_i^{a+z_i+\sum_{j:j\to i}\gamma_j y_{ij}-1}(1-\theta_i)^{\frac{a}{e^{X_i\beta}}-z_i}\prod_{j:j\to i}(1-\theta_i^{\gamma_j})^{1-y_{ij}} \\
P(\gamma_j|\boldsymbol{\theta}, a, \beta, \omega_\gamma, \alpha, y, z, x) &\propto \gamma_j^{\omega_\gamma e^{X-j\alpha}-1}e^{-\omega_\gamma\gamma_j}\prod_{i:j\to i}\theta_i^{\gamma_j y_{ij}}(1-\theta_i^{\gamma_j})^{1-y_{ij}}.
\end{aligned}$$

2.6.5 Full Conditionals of Bayesian Peer Calibration for Count Data

$$
\begin{aligned}
P(\theta_i|\gamma,\theta_{-i},\tau_\theta,\kappa_\theta,\tau_\gamma,\kappa_\gamma,a,b) &\sim \text{Gam}(z_i+\sum_{j:j\to i} y_{ij}+\tau_\theta, 1+\sum_{j:j\to i}\gamma_j+\kappa_\theta)\\
P(\gamma_j|\theta,\gamma_{-j},\tau_\theta,\kappa_\theta,\tau_\gamma,\kappa_\gamma,a,b) &\sim \text{Gam}(\sum_{i:j\to i} y_{ij}+\tau_\gamma, \sum_{i:j\to i}\theta_i\kappa_\gamma)\\
P(\kappa_\theta|\theta,\gamma,\tau_\theta,\tau_\gamma,\kappa_\gamma,a,b) &\sim \text{Gam}(\tau_\theta n+b, \sum_i \theta_i+a)\\
P(\kappa_\gamma|\theta,\gamma,\tau_\theta,\kappa_\theta,\tau_\gamma,a,b) &\sim \text{Gam}(\tau_\gamma n_e+b, \sum_j \gamma_j+a)\\
P(\tau_\theta|\theta,\gamma,\kappa_\theta,\tau_\gamma,\kappa_\gamma,a,b) &\propto \tau_\theta^{b-1}e^{-a\tau_\theta}\prod_i \frac{(\kappa_\theta\theta_i)_\theta^\tau}{\Gamma(\tau_\theta)}\\
P(\tau_\gamma|\theta,\gamma,\tau_\theta,\kappa_\theta,\kappa_\gamma,a,b) &\propto \tau_\gamma^{b-1}e^{-a\tau_\gamma}\prod_j \frac{(\kappa_\gamma\gamma_j)_\gamma^\tau}{\Gamma(\tau_\gamma)}
\end{aligned}
$$

2.6.6 Full Conditionals of Bayesian Peer Calibration with Covariates for Count Data

$$
\begin{aligned}
P(\theta_i|\theta_{-\mathbf{i}},\gamma,\alpha,\beta,\omega_\theta,\omega_\gamma,\mathbf{X},\mathbf{y},\mathbf{z},f,g) &\sim \text{Gam}(z_i+\omega_\theta e^{X_i\beta}+\sum_{j:j\to i} y_{ij}, 1+\omega_\theta+\sum_{j:j\to i}\gamma_j)\\
P(\gamma_j|\theta,\gamma_{-\mathbf{j}},\alpha,\beta,\omega_\theta,\omega_\gamma,\mathbf{X},\mathbf{y},\mathbf{z}f,g) &\sim \text{Gam}(\omega_\gamma e^{X_j\alpha}+\sum_{i:j\to i} y_{ij}, \omega_\gamma+\sum_{i:j\to i}\theta_i)\\
P(\beta_k|\theta,\gamma,\alpha,\beta_{-\mathbf{k}},\omega_\theta,\omega_\gamma,\mathbf{X},\mathbf{y},\mathbf{z},f,g) &\propto e^{-\frac{\beta^2_k}{2\sigma^2}}\prod_i \frac{{\omega_\theta}^{\omega_\theta e^{X_i\beta}}\theta_i^{\omega_\theta e^{X_i\beta}}}{\Gamma\omega_\theta e^{X_i\beta}}\\
P(\alpha_l|\theta,\gamma,\alpha_{-\mathbf{l}},\beta,\omega_\theta,\omega_\gamma,\mathbf{X},\mathbf{y},\mathbf{z},f,g) &\propto e^{-\frac{\alpha^2_l}{2\sigma^2}}\prod_j \frac{{\omega_\gamma}^{\omega_\gamma e^{X_i\alpha}}\gamma_j^{\omega_\gamma e^{X_j\alpha}}}{\Gamma\omega_\gamma e^{X_j\alpha}}\\
P(\omega_\theta|\theta,\gamma,\alpha,\beta,\omega_\gamma,\mathbf{X},\mathbf{y},\mathbf{z},f,g) &\propto {\omega_\theta}^{f-1}e^{-g\omega_\theta}\prod_i \frac{{\omega_\theta}^{\omega_\theta e^{X_i\beta}}}{\Gamma\omega_\theta e^{X_i\beta}}\theta_i^{\omega_\theta e^{X_i\beta}}e^{-\omega_\theta\theta_i}\\
P(\omega_\gamma|\theta,\gamma,\alpha,\beta,\omega_\theta,\mathbf{X},\mathbf{y},\mathbf{z},f,g) &\propto {\omega_\gamma}^{f-1}e^{-g\omega_\gamma}\prod_j \frac{{\omega_\gamma}^{\omega_\gamma e^{X_j\alpha}}}{\Gamma\omega_\gamma e^{X_j\alpha}}\gamma_j^{\omega_\gamma e^{X_j\alpha}}e^{-\omega_\gamma\gamma_j}
\end{aligned}
$$

CHAPTER 3

FIXED CHOICE DESIGN AND AUGMENTED FIXED CHOICE DESIGN FOR NETWORK DATA

3.1 Introduction

The statistical analysis of social networks is increasingly used to understand social processes and patterns (Knoke and Yang, 2008). Of particular interest to sociologists, psychologists, and public health researchers is the association between social relationships and individual behaviors. When analyzing a network, there is often the possibility that nominations are missing, that individuals in the population are missing, or potentially both are missing (Kossinets, 2006). Earlier work has focused on missingness that is induced by nonresponse (Robins et al., 2004; Costenbader and Valente, 2003; Huisman and Steglich, 2008; Gile and Handcock, 2006). Due to the complex dependence structure inherent in networks, missing data can pose very difficult problems (Holland and Leinhardt, 1973; Marsden, 1990; Kossinets, 2006). For example, it is unclear how to handle missingness when analyzing the network for peer effects on behavior.

Some missingness is a result of study design. Several recent studies on networks in school settings make use of the fixed choice design (FCD), which may induce missing edges. In FCD, the number of possible nominations that each person in the network can make is capped at a maximum, m, inducing missing nominations (Kossinets, 2006; Holland and Leinhardt, 1973; Yan and Gregory, 2011). For example, studies have variously restricted the number of friends in a classroom to 4 when studying depression (Witvliet et al., 2010), the number of best friends to 5 when studying smoking (Mercken et al., 2010), another study on infectious disease transmission on social networks allowed participants to name up to 6

within class contacts and up to 4 outside of class contacts (Conlan et al., 2010), whereas the National Longitudinal Study of Adolescent Health (Add Health) allowed participants to nominate and rank up to 5 boys and 5 girls as friends (Resnick et al., 1997; Goodreau, 2007).

Holland and Leinhardt (1973) first introduced the problem of missing ties due to the FCD (also called limited choice design and right-censoring of degree). Kossinets (2006) went on to show how the FCD can lead to biases in estimates of structural measures of the network.

Hoff et al. (2012) has developed a likelihood based approach for fixed rank network data (where there is a maximum number of nominations that could be made, and those nominations are ranked) as well as for FCD, and then used these likelihoods in Bayesian estimation of latent variables which are assumed to govern the nominations and their ranks.

Here we are concerned with FCD and suggest a novel approach that can improve inference: that the number of nominations that would have been made had there not been a fixed choice design is collected in addition to the regular FCD data. We develop a method that accounts for the missingness resulting from FCD, given the number of nominations that would have been made, and show how different cut-offs affect parameter estimation using a dyad-independent exponential random graph model. Biases resulting from missingness in the FCD can be alleviated by incorporating information on the number of nominations that would have been made had there been no maximum number of nominations imposed by the study survey. In Section 3.2 we introduce novel methods for accounting for the fixed choice censoring and introduce a new survey design. In Section 3.3 we present simulation studies wherein we demonstrate our methods for handling fixed choice data and compare the effects of different cut-offs on parameter estimation. In Section 3.4 we demonstrate our method in an analysis of a network of undergraduates living in a residence hall. A discussion is presented in Section 3.5.

3.2 Parameter Estimation

In this section we develop a method that accounts for missingness resulting from imposing a maximum number of nominations that a network survey respondent can endorse. First we assume that the number of nominations that would have been made is known and call this an augmented fixed choice design (AFCD). We next consider the FCD setting where the number of nominations that would have been made had there been censored and unknown.

In each of these cases we consider a network of n individuals, represented by an $n \times n$ sociomatrix Y, where $Y_{ij} = 1$ if i nominates j, and $Y_{ii} = 0$ for all i. Note that $Y_{ij} = 1$ does not imply that $Y_{ji} = 1$, meaning that these relationships may be unreciprocated. Hence, each individual's nominations are represented by a $1 \times n$ vector:

$$Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in}),$$

with $Y_{ii} = 0$ with probability 1. For individual i, let

$$V_i = (V_{i1}, V_{i2}, \ldots, V_{iq})$$

denote a $1 \times q$ vector of covariates. For all i, j pairs where $i \neq j$, we also have the $1 \times q$ vector

$$D_{ij} = (|V_{i1} - V_{j1}|, |V_{i2} - V_{j2}|, \ldots, |V_{iq} - V_{jq}|),$$

which represents the differences of a covariate of interest V. For all pairs i, j, we have the $1 \times (2q + 1)$ vector $X_{ij} = (1, V_i, D_{ij})$, and define the $(n - 1) \times (2q + 1)$ matrix

$$\begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iq} \end{pmatrix} = \begin{pmatrix} 1 & V_i & D_{i1} \\ 1 & V_i & D_{i2} \\ \vdots & \vdots & \vdots \\ 1 & V_i & D_{iq} \end{pmatrix}.$$

. All probabilities are conditional on covariates $\mathbf{x}$ and parameters β. These are suppressed

in the notation. $Y = (Y_{ij})$ are independent Bernoulli random variables with $\mathbb{P}(Y_{ij} = 1) = \pi_{ij} = \pi_{ij}(\mathbf{x}, \beta)$. Later we will specialize to a logistic regression formulation for π_{ij}.

When $\mathbf{Y}$ is fully observed we can infer associations between the presence or absence of directed relationships and the covariates of interest $\mathbf{x}$. However, in the FCD setting we can only observe a fixed number of nominations m for each individual. In the AFCD setting we also observe $N_i = \sum_{j:j\neq i} Y_{ij}$, the number of nominations that would have been made by person i if $m = n$. $W = (W_{ij})$ is a randomly censored version of Y, meaning that W is created from Y by switching some of the ones in Y to zeros. Hence $W \leq Y$ (meaning each $W_{ij} \leq Y_{ij}$). Then censoring is done so that

- rows of Y are censored independently
- if $N_i(Y) \leq m$, there is no censoring and $W_i = Y_i$
- if $N_i(Y) > m$, then a subset of the ones in Y_i is observed in W_i.

We assume that for person i given that $N_i > m$ each of the N_i relationships have an equal probability m/N_i of being observed. Using this structure we first derive the observed data likelihood for the AFCD, and next the observed data likelihood for the FCD.

3.2.1 AFCD Observed Data Likelihood

When deriving the likelihood of the observed data, it may be helpful to consider the data for one row at a time of the sociomatrix $\mathbf{Y}$. While we do not fully observe $\mathbf{Y}$, we do observe $\mathbf{w}$ and $N_i, i \in 1, 2, \ldots, n$. As such, we are interested in maximizing

$$\mathcal{L}(\beta; \mathbf{w}, \mathbf{x}, \mathbf{N}) = \prod_i p_i(\mathbf{w}; \mathbf{x}, \mathbf{N}, \beta) = \prod_i \sum_{\mathbf{y}} p_i(\mathbf{w}; \mathbf{y}) p_i(\mathbf{y}; \mathbf{x}, \beta)$$

where $p_i(\mathbf{w}; \mathbf{x}, \mathbf{N}, \beta)$ is the probability that the i^{th} row of the random matrix W is equal to w_i given covariates $\mathbf{x}$, and N_i, $p_i(\mathbf{w}; \mathbf{y})$ is the probability that $W_i = w_i$ given $Y_i = y_i$, and $p_i(\mathbf{y}; \mathbf{x}, \beta)$ is the probability that $Y_i = y_i$ given covariates x and parameters β.
Under the AFCD censoring mechanism, we have:

$$p_i(\mathbf{w};\mathbf{y},\mathbf{N}) = \begin{cases} 1 & \mathbf{w}_i = \mathbf{y}_i, N_i \leq m \\ \binom{N_i}{m}^{-1} & w_{ij} \leq y_{ij}, N_i > m, \sum_j w_{ij} = m \\ 0 & otherwise. \end{cases}$$

When $N_i \leq m$, the AFCD sampling scheme implies $\mathbf{w}_i = \mathbf{y}_i$ with probability 1. In other words, if an individual i nominates less than or equal to the maximum number of possible nominations m, then we fully observe $\mathbf{y}_i$. If $N_i > m$, the AFCD sampling scheme and the assumption that each nomination made by person i has an equal probability of being observed gives rise to the probability of observing $\mathbf{w}_i$ as $\binom{N_i}{m}^{-1}$ since there are $\binom{N_i}{m}$ ways to for the AFCD censoring mechanism to produce the observed data $\mathbf{w}_i$ such that it is consistent with $\mathbf{y}_i$. In other words, there are $\binom{N_i}{m}$ ways to choose which m nominations of the N_i that will be observed in $\mathbf{w}_i$. Therefore we have

$$\mathcal{L}(\beta;\mathbf{w},\mathbf{x},\mathbf{N}) \propto \prod_i p_i(\mathbf{w};\mathbf{x},\beta,\mathbf{N})$$

where

$$p_i(\mathbf{w};\mathbf{x},\beta,\mathbf{N}) = \begin{cases} \prod_j \pi_{ij}^{w_{ij}}(1-\pi_{ij})^{(1-w_{ij})} & ;N_i \leq m \\ \prod_j \pi_{ij}^{w_{ij}} \binom{N_i}{m}^{-1} P(\sum_{j\neq i} Y_{ij}(1-w_{ij}) = N_i - m; x_i, \beta) & ;N_i > m, \sum_j w_{ij} = m \end{cases}.$$

Here $p_i(\mathbf{w};\mathbf{x},\beta,\mathbf{N}) = \pi_{ij}^{w_{ij}}(1-\pi_{ij})^{(1-w_{ij})}$ when $N_i \leq m$ because when $N_i \leq m$, we fully observe the row i. When $N_i > m$ there are m edges that are observed, $N_i - m$ edges that are censored, and $n - N_i - 1$ non-edges that are censored.

Because there are $\binom{n-m-1}{N_i-m}$ ways to obtain the equality $\sum_{j\neq i} \mathbf{Y}_{ij}(1-w_{ij}) = N_i - m$, calculating $P(\sum_{j\neq i} \mathbf{Y}_{ij}(1-w_{ij}) = N_i - m)$ may seem computationally infeasible. However, this probability can be calculated rather quickly as a convolution of independent non-identical Bernoulli variables. First, relabel the covariate values for unobserved x_i as $z_{i1}, z_{i2}, \ldots, z_{i(n-m-1)}$, and relabel the unobserved values for Y_i as $v_{i1}, v_{i2}, \ldots, v_{i,(n-m-1)}$. Let $S(k,l,i) = P(\sum_{h=0}^{k} v_{ih} = l)$, where $S(0,l,i) = 0$ for all $l > 0$ and $S(0,1,i) = 1$ otherwise. Therefore $P(\sum_{j\neq i} \mathbf{Y}_{ij}(1-r_{ij}) = N_i - m; \beta, x) = S(n-m-1, N_i-m, i)$. Having relabeled the

missing values, we proceed by assigning $S(k,l,i) = S(k-1,l-1,i)\pi_{ik} + S(k-1,l,i)(\pi_{ik}-1)$ for $k \in (1,..n-m-1)$ and for $l \in 1,\ldots N_i - m$, first starting with $k = 1$, and cycling through $l \in 0,\ldots N_i - m$ for each k. Here $\pi_{ik} = \frac{e^{z_{ik}\beta}}{1+e^{z_{ik}\beta}}$.

To estimate the β parameters we maximize the above likelihood using Nelder and Mead's simplex method, which is available in the R function optim (Nelder and Mead, 1965; Team, 2012).

3.2.2 FCD Likelihood

We next consider the FCD where N_i is unknown, though the maximum number of possible nominations m is known. As with the AFCD, it is helpful to consider the likelihood contribution for one individual at a time. Here the likelihood is

$$\mathcal{L}(\beta; \mathbf{w}, \mathbf{x}) = \prod_i p_i(\mathbf{w}; \mathbf{x}, \beta) = \prod_i \sum_{\mathbf{y}} p_i(\mathbf{w}; \mathbf{y}) p_i(\mathbf{y}; \mathbf{x}, \beta).$$

Under the FCD censoring mechanism, we have:

$$p_i(\mathbf{wy}) = \begin{cases} 1 & \mathbf{w}_i = \mathbf{y}_i, \sum_j y_{ij} < m \\ \binom{\sum_j y_{ij}}{m}^{-1} & ; \sum_j y_{ij} \geq \sum_j w_{ij} = m, \\ 0 & otherwise. \end{cases}$$

Therefore we have:

$$\mathcal{L}(\beta; \mathbf{w}, \mathbf{x}) \propto \prod_i p_i(\mathbf{w}; \mathbf{x}, \beta)$$

where

$$p_i(\mathbf{wx}, \beta) = \begin{cases} \prod_j \pi_{ij}^{w_{ij}} (1-\pi_{ij})^{(1-w_{ij})} & ; m > \sum_{j \neq i} w_{ij} \\ \prod_j \pi_{ij}^{w_{ij}} \sum_{k=0}^{n-m-1} \left[\binom{m+k}{m}^{-1} \mathrm{P}_i(\sum_{j \neq i} Y_{ij}(1-w_{ij}) = k) \right] & m = \sum_{j \neq i} w_{ij} \\ 0 & otherwise \end{cases} .$$

The i^{th} row of Y is fully observed when $m > \sum_{j \neq i} w_{ij}$, and we treat them as $n-1$ independent Bernoulli random variables. When $m = \sum_{j \neq i} w_{ij}$, we only observe the m edges, and since we do not have any information on how many of the missing values are edges or non-edges, we sum over the probability that the number of edges is $m+k$, where $k \in 0, \ldots, n-m-1$. Here the key difference in the likelihood for FCD as opposed to AFCD is that with the FCD we do not know N_i, and therefore need to sum over all possible values of N_i, where $N_i \in \{m, m+1, \ldots, n-1\}$. With AFCD it is unnecessary to sum over all possible values of N_i.

Proceeding in a similar fashion to the AFCD, we perform a convolution of independent non-identical Bernoulli variables in order to calculate the likelihood function, and find the maximum likelihood estimates using the simplex optimization method.

3.2.3 Variance Estimation

The variance estimation for $\hat{\beta}$ is non-trivial. Simply using the observed information will not incorporate the uncertainty of the censored values. In order to quantify the variance of the maximum likelihood estimates of the β parameters, we describe a parametric bootstrap by following these steps. Let B deonte the number of bootstrap samples.

1. Maximize the appropriate likelihood as described above to produce $\hat{\beta}$.
2. Using $\hat{\beta}$ and the same x covariates as observed in the data and, generate sociomatrix $Y^b, b \in 1, 2, \ldots B$.
3. Randomly delete the observations in rows of Y^b as described above to generate $\mathbf{w}_b$.
4. Maximize the appropriate likelihood using $\mathbf{w}_b$, $N_i \forall i$ (in the case of the AFCD), and x as the observed data, to get $\widehat{\beta}_b$
5. Repeat steps 2-4 B times to generate a distribution of $\widehat{\beta}_b$ which can be used to get bootstrapped standard errors and confidence intervals, using a Normal approximation.

3.3 Simulation Studies

3.3.1 Simulation Design

In order to contrast the AFCD, FCD, and a naive analysis (where we assume that all unobserved values of $\mathbf{w}$ are non-edges), we performed a simulation study.

For a simulated population of n, we first generated a continuous covariate V for all n members of the simulated population, which stays fixed for all simulations. In keeping with the previously defined notation, we next generated directed edges between members of the network such that the probability that individual i has a directed relationship with individual j, denoted by $Y_{ij} = 1$, is independent given $\mathbf{x}$:

$$\text{Logit}(P(Y_{ij} = 1); \beta, \mathbf{x}) = \mathbf{x}_{ij}\beta$$

where

$$\mathbf{X}_i = \begin{pmatrix} 1 & V_i & |V_i - V_1| \\ 1 & V_i & |V_i - V_2| \\ \vdots & \vdots & \vdots \\ 1 & V_i & |V_i - V_n| \end{pmatrix}$$

and β is a 3×1 vector. We next simulate the censoring processes of the AFCD and the FCD for maximum number of nominations $m \in \{0, 2, 4, 6, 8\}$. Under the AFCD, in row i, given N_i and m, when $N_i > m$, $N_i - m$ edges are censored, where each of the N_i edges have an equal probability of being censored. Moreover, all non-edges are censored. This gives us the observed AFCD data $\mathbf{w}_{AFCD}$. For the FCD, for a maximum m in each row where $N_i \geq m$, $N_i - m$ edges are censored, and all non-edges are censored, to give us the observed FCD data $\mathbf{w}_{FCD}$.

We then obtain maximum likelihood estimators under the AFCD and FCD with varying m values. We also carry out the naive analysis by carrying out logistic regression on the

full data $\mathbf{y}$, and on the $\mathbf{w}_{FCD}$ data where we treat all censored values as zero, which is the current practice for analyzing FCD.

3.3.2 Simulation Results

The distribution of the covariate value V is displayed in Figure 3.1. We generated the network using $\beta_0 = 2, \beta_1 = -0.1, \beta_2 = -0.05$. These V and β values were chosen to create a distribution of N_i that are similar to the network which we analyze in Section 3.4. We choose a population size of $n = 50$ to allow for computational speed. We simulated 200 different networks using these covariate and β values. The distribution of N_i for all 200 simulations is plotted in Figure 3.2, where green bars indicate m values used to censor the simulated data according to FCD and AFCD censoring schemes. Boxplots of the estimated β's are displayed in Figures 3.3, 3.4, and 3.5, where the red horizontal line represents the value that was used to generate the simulated networks. We present the MSE, empirical bias, and SEs of the estimated β's are in Tables 3.1a, 3.1b and 3.1c.

For a given maximum m, the AFCD maximum likelihood estimates tended to outperform the FCD and naive analyses in terms of both bias and variance. In fact, having an AFCD with $m = 0$ could be a better choice than a FCD where $m > 0$. For the β_0 parameter, having an AFCD with $m = 0$ had a lower MSE than FCD with $m = 6$, and a naive analysis with $m = 8$. For the β_1 parameter, having an AFCD with $m = 0$ outperformed a FCD with $m = 2$, and a naive analysis with $m = 6$ in terms of MSE. However, for higher values of m, the relative improvement in MSE for AFCD over FCD tended to diminish, and when the full data is observed the AFCD, FCD and naive analyses necessarily produce the same estimates.

3.4 Analysis of UrWeb Study

We next apply the method to the UrWeb dataset, in which data were collected from residents of a freshman dormitory (Barnett et al., 2013). Each participant is 18 years old or older

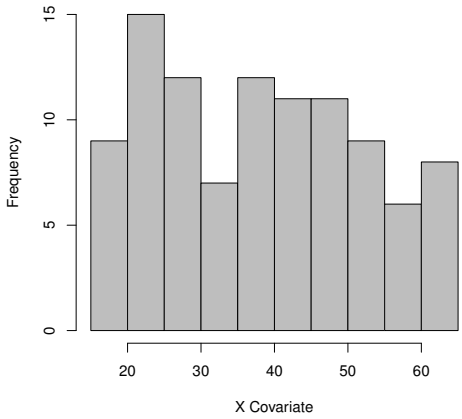

Figure 3.1: Distribution of V Covariate Used to Generate Simulated Networks

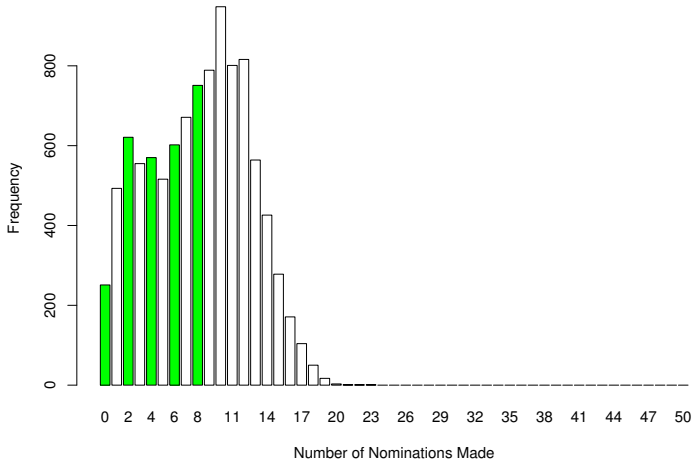

Figure 3.2: Distribution of the Actual Number of Nominations Made from 200 Simulations. The green bars denote values of m used to impose AFCD and FCD censoring in 200 different simulations.

Table 3.1: Mean Squared Error, Empirical Bias, and Standard Deviation of Estimated β_0 from 200 Simulations

Maximum	MSE $\times 10^2$			$\lvert$Bias$\rvert \times 10$			SD$\times 10$		
	AFCD	FCD	Naive	AFCD	FCD	Naive	AFCD	FCD	Naive
0	12.60	—	—	1.41	—	—	3.26	—	—
2	8.49	90.12	903.99	0.16	1.02	29.98	2.92	9.46	2.30
4	6.56	34.79	404.52	0.03	1.61	20.02	2.57	5.69	1.89
6	5.53	12.95	172.43	0.06	0.83	13.01	2.36	3.51	1.79
8	4.85	8.03	62.28	0.11	0.47	7.70	2.20	2.80	1.74
Full	3.68	3.68	3.68	0.03	0.03	0.03	1.92	1.92	1.92

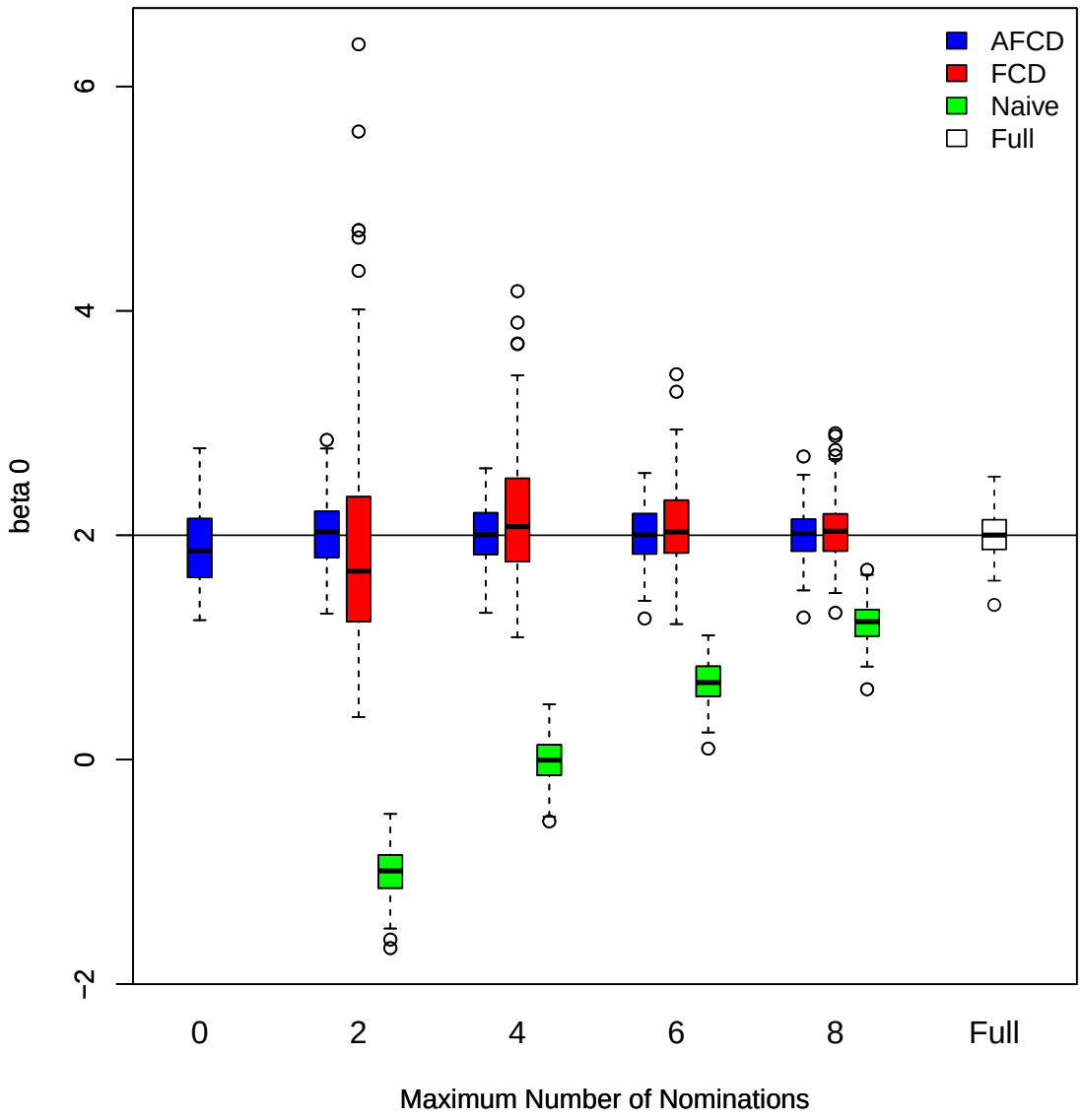

Figure 3.3: Boxplots of $\hat{\beta}_0$, varying maximum number of nominations m and AFCD, FCD, or naive analysis. The black horizontal line is the true β_0 value used to generate the simulated data.

Table 3.2: Mean Squared Error, Empirical Bias, and Standard Deviation of Estimated β_1 from 200 Simulations

Maximum	MSE $\times 10^5$			$\lvert\text{Bias}\rvert \times 10^4$			SD$\times 10^3$		
	AFCD	FCD	Naive	AFCD	FCD	Naive	AFCD	FCD	Naive
0	27.97	—	—	46.66	—	—	16.10	—	—
2	9.55	52.41	153.46	5.91	51.22	384.61	9.78	22.37	7.46
4	7.52	17.98	94.14	2.21	30.16	299.70	8.69	13.10	6.59
6	6.13	9.02	49.22	1.32	15.85	213.00	7.85	9.39	6.22
8	5.25	6.96	22.11	0.39	8.09	135.78	7.26	8.32	6.07
Full	3.71	3.71	3.71	0.48	0.48	0.48	6.11	6.11	6.11

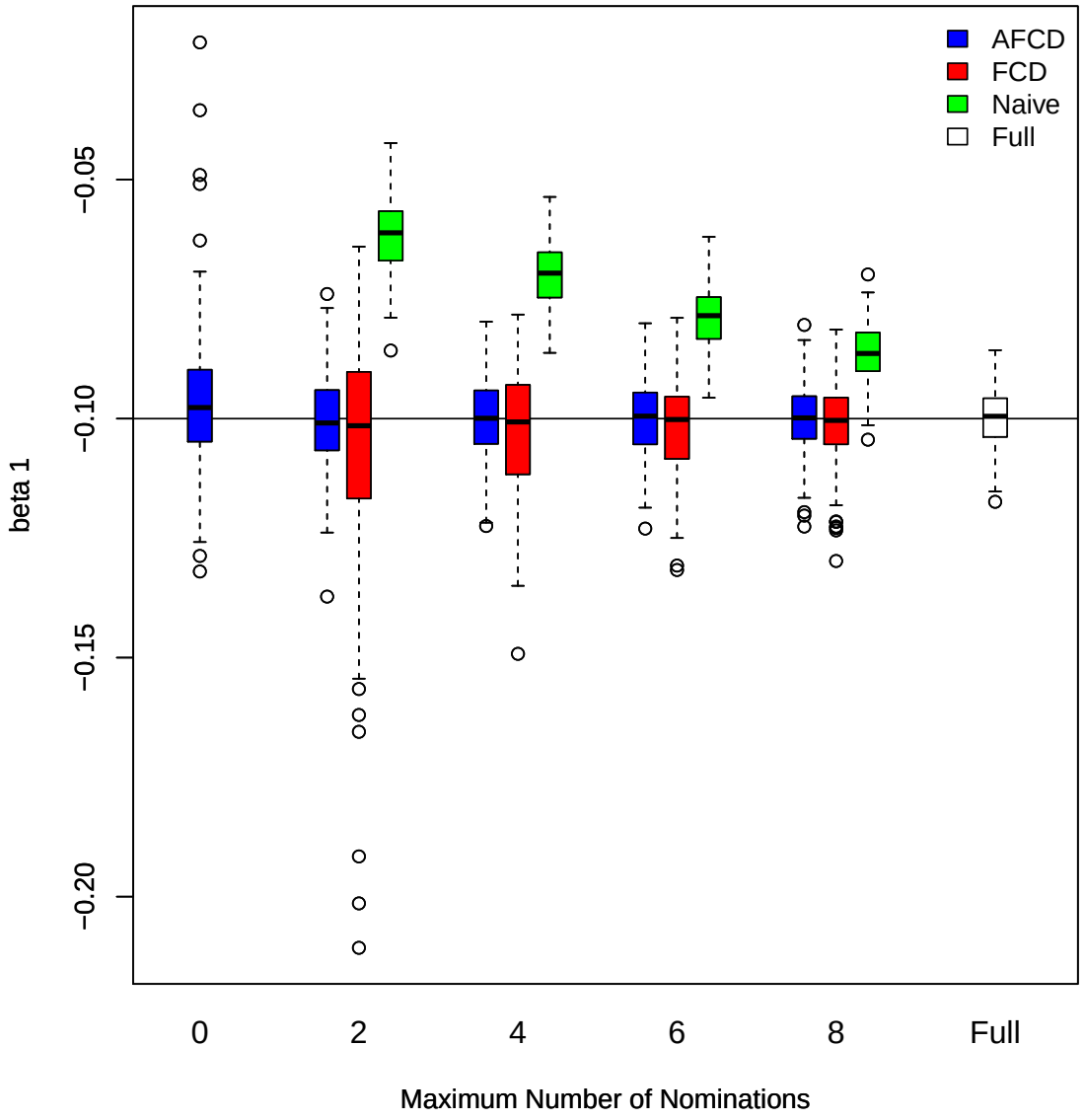

Figure 3.4: Boxplots of $\hat{\beta}_1$, varying maximum number of nominations m and AFCD, FCD, or naive analysis. The black horizontal line is the true β_1 value used to generate the simulated data.

Table 3.3: Mean Squared Error, Empirical Bias, and Standard Deviation of Estimated β_2 from 200 Simulations

Maximum	MSE $\times 10^4$			$\lvert\text{Bias}\rvert \times 10^3$			SD$\times 10^3$		
	AFCD	FCD	Naive	AFCD	FCD	Naive	AFCD	FCD	Naive
0	73.19	—	—	4.28	—	—	8.57	—	—
2	4.54	8.77	11.11	0.21	2.28	27.22	2.14	2.96	1.93
4	3.24	4.89	6.09	0.35	1.18	18.68	1.80	2.21	1.62
6	2.24	2.87	3.06	0.33	0.98	10.58	1.50	1.70	1.40
8	1.89	2.04	1.88	1.36	1.45	4.30	1.37	1.43	1.30
Full	1.48	1.48	1.48	0.41	0.41	0.41	1.22	1.22	1.22

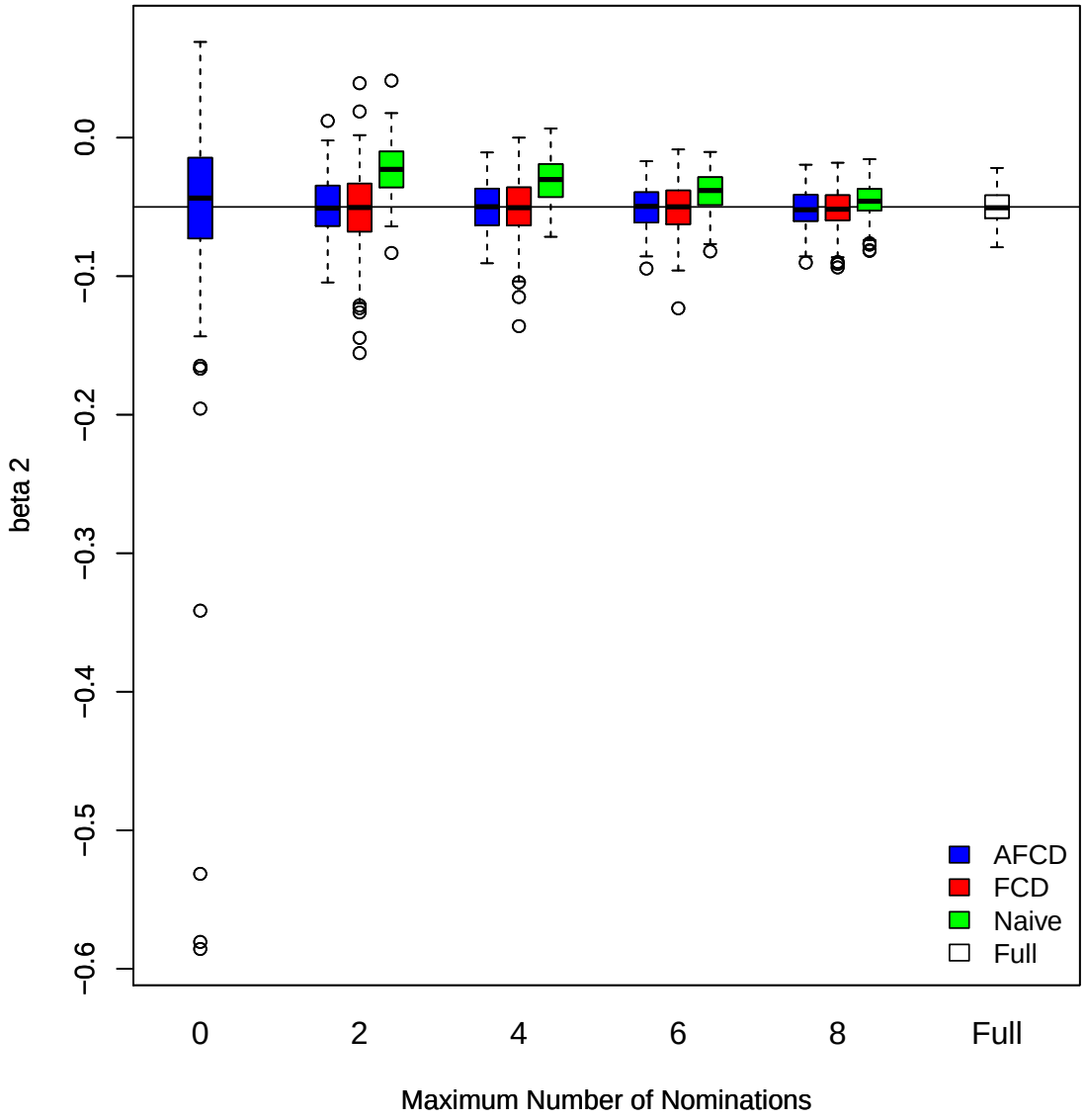

Figure 3.5: Boxplots of $\hat{\beta}_2$, varying maximum number of nominations m and AFCD, FCD, or naive analysis. The black horizontal line is the true β_2 value used to generate the simulated data.

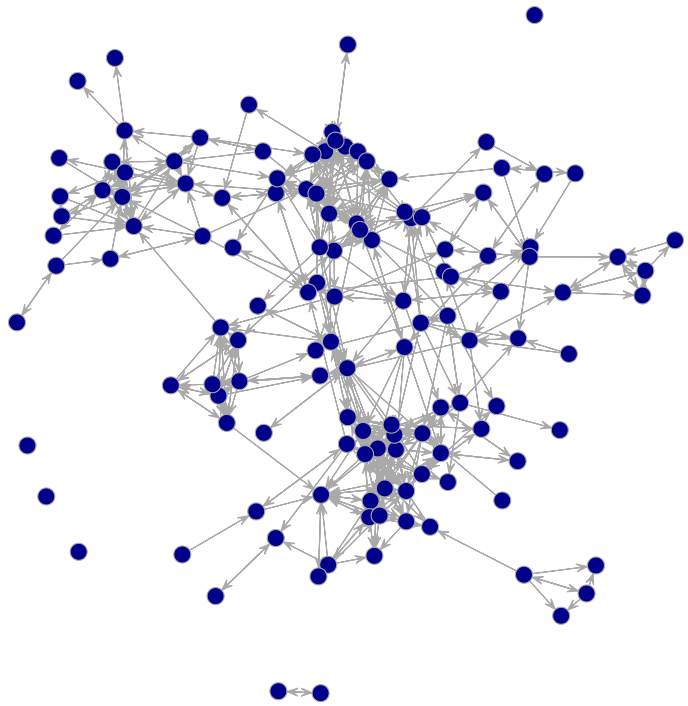

Figure 3.6: UrWeb Social Network

when the survey is administered and is asked to report the number of days in a month that they consume alcohol. Further, participants are asked to provide demographic information, including their age, sex, race, sexual orientation, and whether or not they intend to join a fraternity or sorority. Additionally, each participant is asked to nominate which of the other participants are important to them. This network is pictured in Figure 3.6. Among the 129 participants included in the sample, 507 nominations were made; 4 participants did not nominate alters and were not nominated.

The UrWeb data were collected under a FCD, with $m = 10$. In this dataset, only one person endorsed the maximum number of nominations, which suggests that there is little design-induced missingness of nominations. We will proceed to show the utility of the methods introduced here by artificially inducing a $m < 10$ in the UrWeb dataset, estimating parameters, and comparing the estimated parameters when $m \in \{0, 2, 4, 6, 8\}$ to the parameters estimated with logistic regression using the full dataset. We will artificially induce $m \leq 8$ by deleting edges of individuals with more than m in two different ways, first by randomly deleting edges (as above in the simulation study), and secondly by deleting edges with regard to the order that each individual made their nominations. Figure 3.7

displays realizations of the UrWeb network variously imposing m equal to 2,4,6,8. Figure 3.8 displays the distribution of the number of nominations that were made by each participant in the UrWeb study. We will proceed assuming that the UrWeb data is fully observed, and will demonstrate how this method will work in practice, and compare the information loss for different m values, and contrast AFCD, FCD, and naive analyses.

Here we use the number of days in a month that the subjects consume alcohol as the V covariate in this model

$$\text{Logit}(P(Y_{ij} = 1)|\beta, \mathbf{x}_i) = \beta_0 + \beta_1 v_i + \beta_2 |v_i - v_j|$$

where $Y_{ij} = 1$ indicates that participant i nominated j.

Having artificially induced $m \in \{0, 2, 4, 6, 8\}$ 200 times, we estimated $\beta_0, \beta_1, \beta_2$ using the AFCD, FCD, and naive methods. We present boxplots of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ in Figures 3.9, 3.10, and 3.11. In each of these figures we denote the estimate from the fully observed UrWeb data with a solid horizontal line, and the estimate from the full data $\pm$ the estimated standard error with dotted horizontal lines.

When $m = 0$ we are only able to obtain maximum likelihood estimates for the AFCD. Because there is only one way to impose that there is no nomination data observed, Figures 3.9, 3.10, and 3.11 present a point estimate rather than a distribution of estimates when $m = 0$.

The analyses in this section differ from those in section 3.3 in a few important ways. First, rather than simulating several networks, we are analyzing a single real network $\mathbf{y}_{UrWeb}$, and repeatedly removing edges at random to form many different realizations of $\mathbf{w}_{UrWeb}$. In these analyses, we do not know the true β values, and so we cannot speak to the bias of the estimates. However, we can use these analyses to investigate information loss due to the design induced censoring by comparing AFCD, FCD, or naive estimates when $m < 10$ to estimates when we observe $m = 10$. For example in Figure 3.8, excluding when $m = 0$ the estimates of β_0 from the AFCD are all within one standard error of the estimate when the full UrWeb data is observed ($m = 10$). This is in sharp contrast to the FCD and

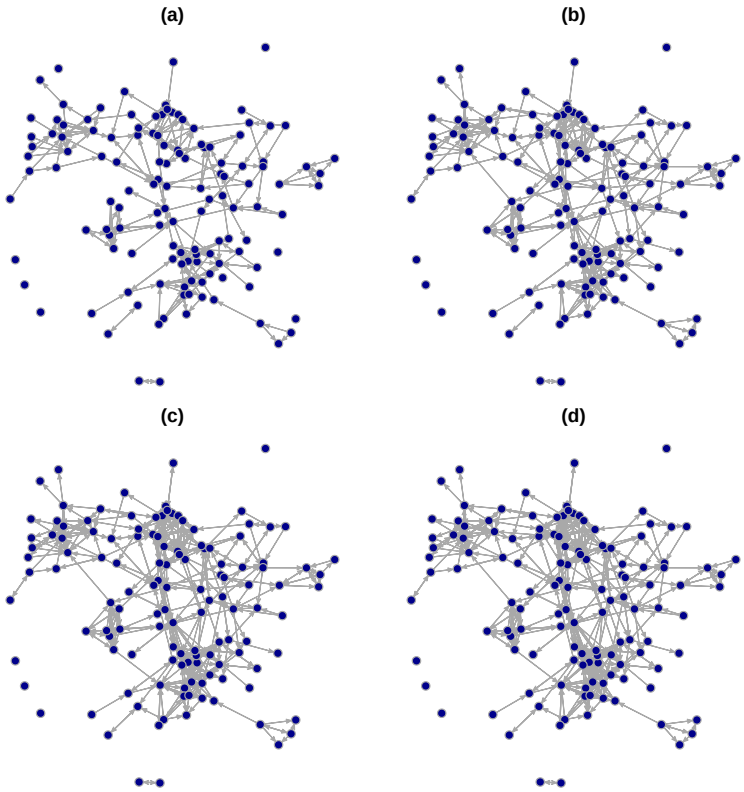

Figure 3.7: UrWeb Social Network with Simulated Censoring of Edges, $m = \{2, 4, 6, 8\}$ for a,b,c, and d respectively.

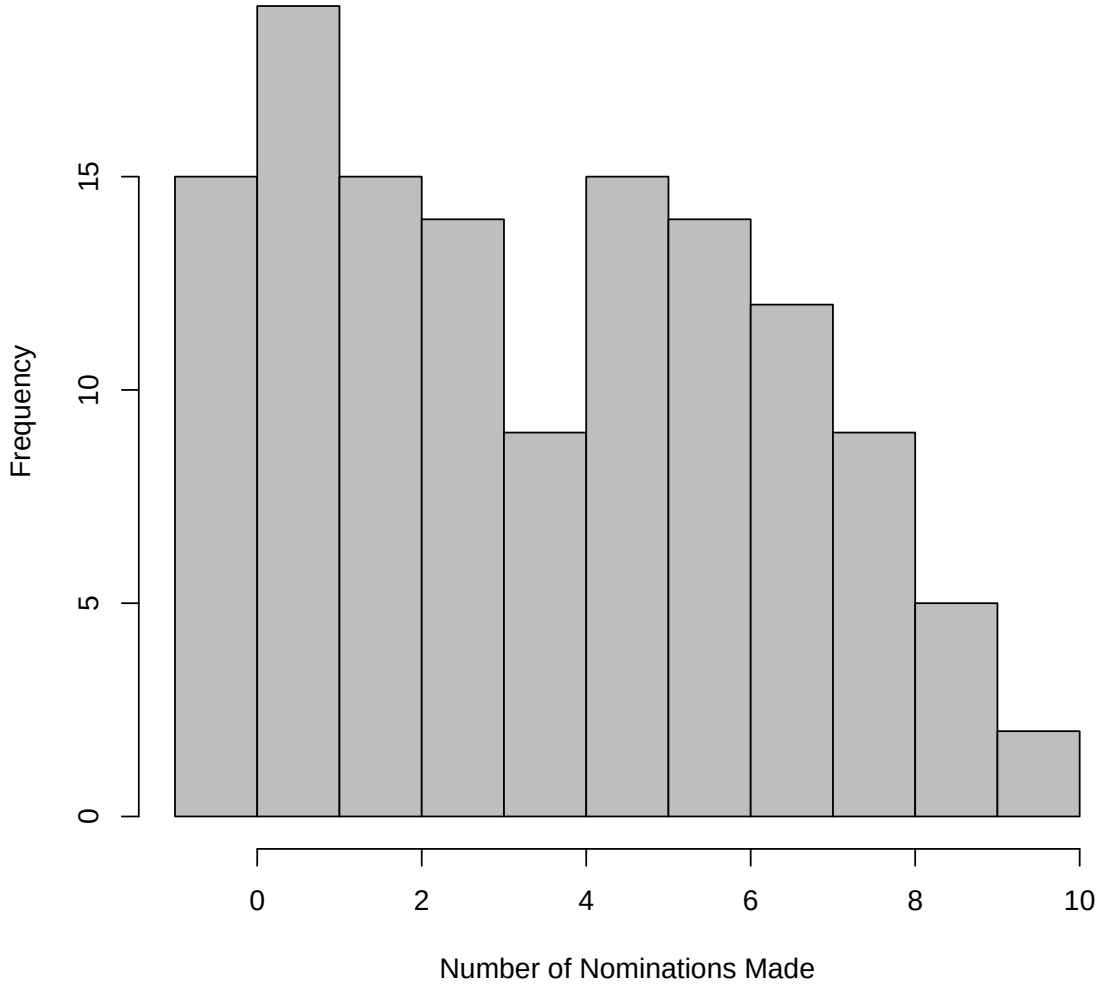

Figure 3.8: Distribution of nominations made by UrWeb study participants

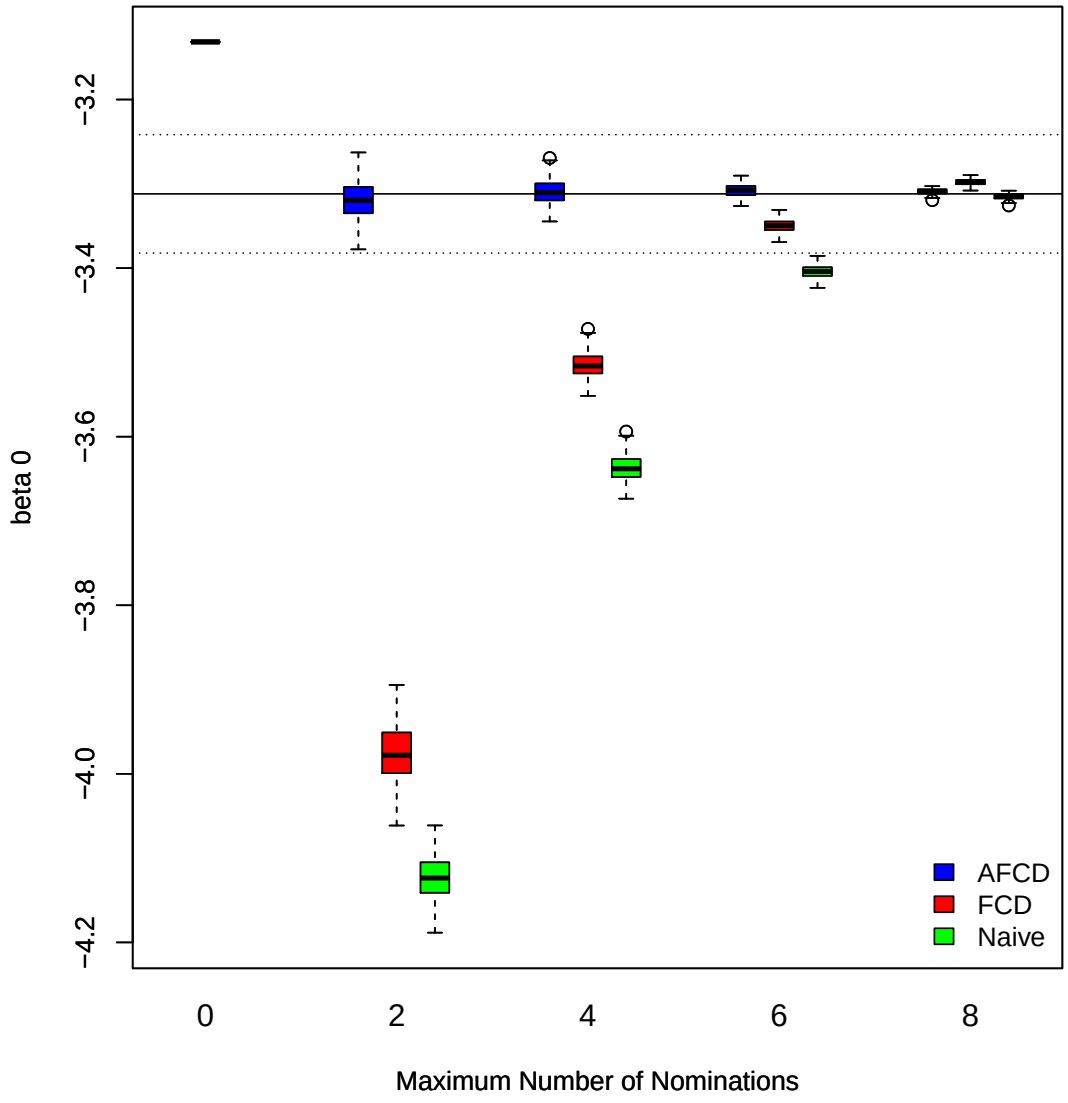

Figure 3.9: Boxplots of $\hat{\beta}_0$, varying maximum number of nominations m, and AFCD, FCD, or naive analysis in the UrWeb data set. The black horizontal line is the $\hat{\beta}_0$ value when using the full UrWeb data. The dotted lines are the $\hat{\beta}_0 \pm \widehat{SE}(\beta_0)$ computed from the full UrWeb data.

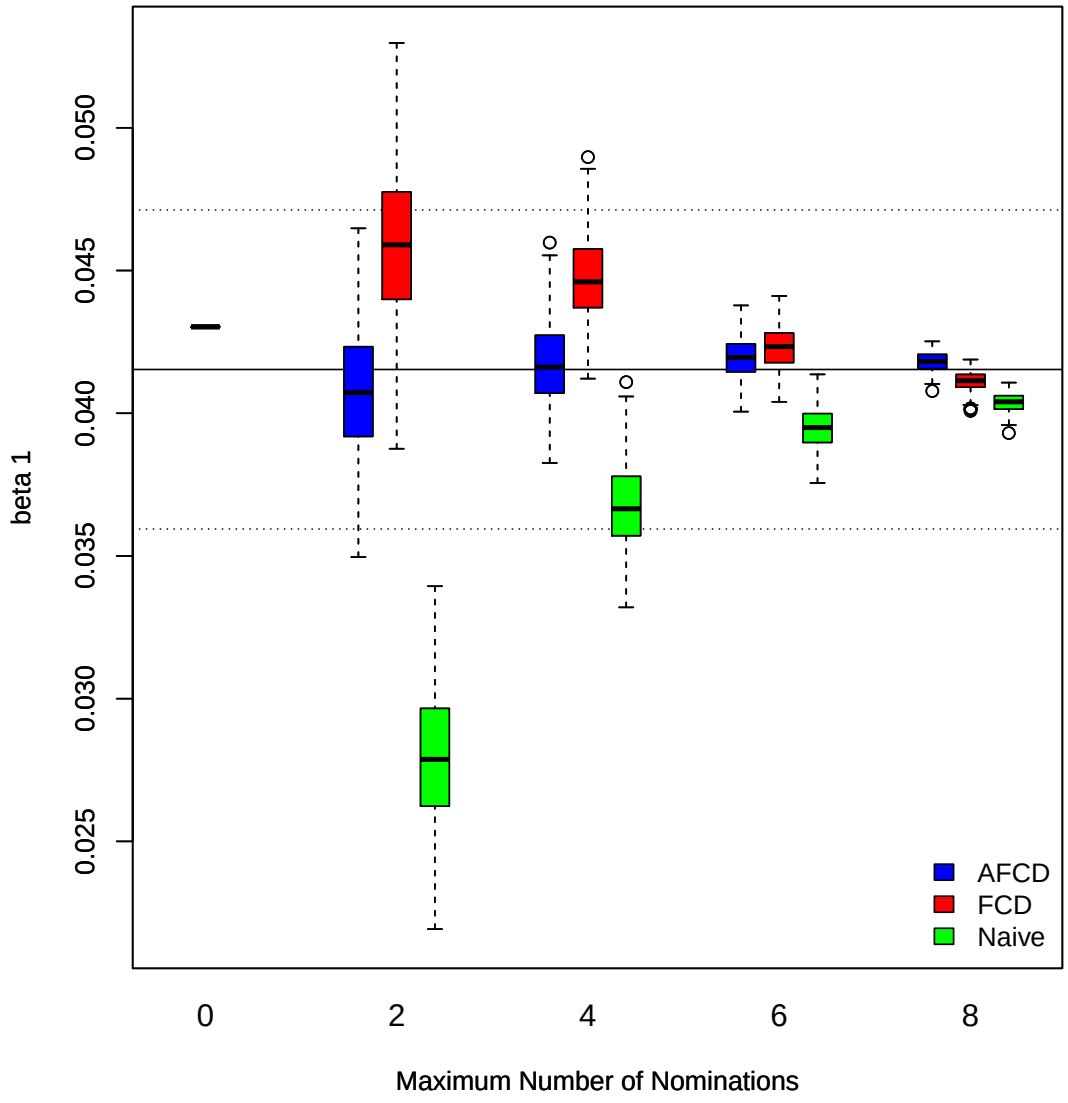

Figure 3.10: Boxplots of $\hat{\beta}_1$, varying maximum number of nominations m, and AFCD, FCD, or naive analysis in the UrWeb data set. The black horizontal line is the $\hat{\beta}_1$ value when using the full UrWeb data. The dotted lines are the $\hat{\beta}_1 \pm \widehat{SE}(\beta_1)$ computed from the full UrWeb data.

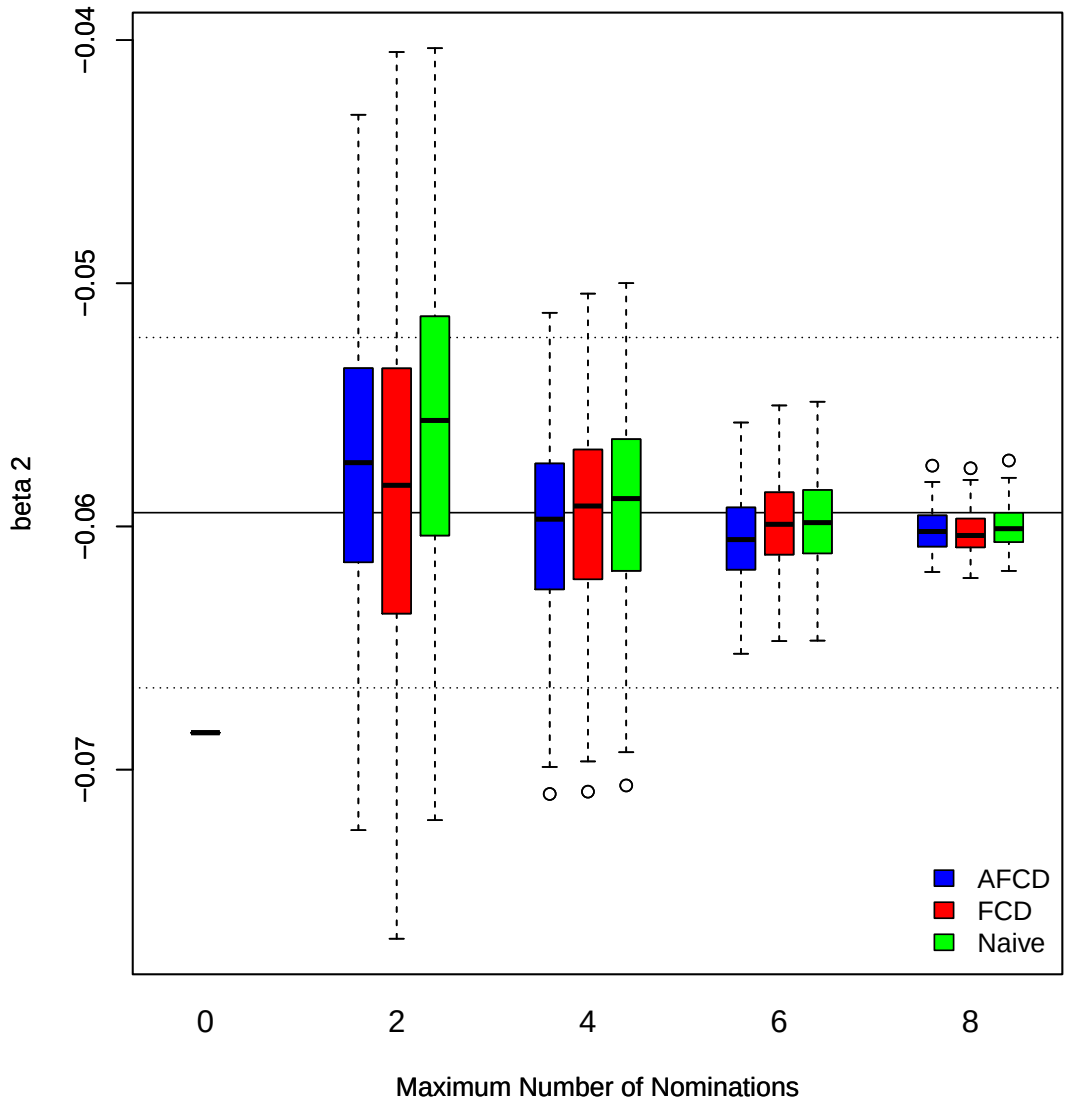

Figure 3.11: Boxplots of $\hat{\beta}_2$, varying maximum number of nominations m, and AFCD, FCD, or naive analysis in the UrWeb data set. The black horizontal line is the $\hat{\beta}_2$ value when using the full UrWeb data. The dotted lines are the $\hat{\beta}_2 \pm \widehat{SE}(\beta_2)$ computed from the full UrWeb data.

naive estimates of β_0 when $m = 2, 4$. In general, the AFCD seems to lose less information than the FCD, which in turn looses less information than the naive analysis.

In this analysis of the UrWeb network, the AFCD produces estimates of β that are roughly centered on the full data estimate for β. The FCD method produces estimates that diverge from the estimate when the data is fully observed, especially when m is small. This result in in contrast to the simulation study which showed that the FCD on average produces approximately unbiased estimates. We postulate that there is considerable between-network variation that could produce this phenomenon.

Next, we deleted nominations in the reverse order in which they were made by the participants in the UrWeb study so that $m \in (0, 2, 4, 6, 8)$. We estimated β using AFCD, FCD, and naive methods. For the AFCD and FCD we calculated standard error estimates from 500 bootstrap samples. For the naive analyses, we simply used the standard error from a regular logistic regression model. These are presented in Figures 3.12, 3.13, and 3.14. In each of Figures 3.12,3.13, and 3.14 we also include the middle 90% of the distributions for the corresponding plots in Figures 3.9, 3.10, and 3.11 which allow us to compare the results. The UrWeb study was not explicitly designed to accommodate the AFCD or FCD design in that participants were not prompted to name a random sample of the people who were important to them. It is possible that study participants chose their nominations non-uniformly, for example nominating peers in the order of their importance. By deleting nominations in reverse order, we seek to investigate whether this could impact inference. For instance, we compare Figure 3.11 where the edges were deleted in reverse order to Figure 3.9 where edges were deleted without regard to the order in which they were made. The maximum likelihood estimates of β_0 in Figure 3.11 are are not in the middle of the distributions of the maximum likelihood estimates of Figure 3.9. This suggests that the order in which nominations were made in this dataset may impact the inferences made in these analyses, and that in order to use AFCD and FCD, nominations should be solicited with uniform probability.

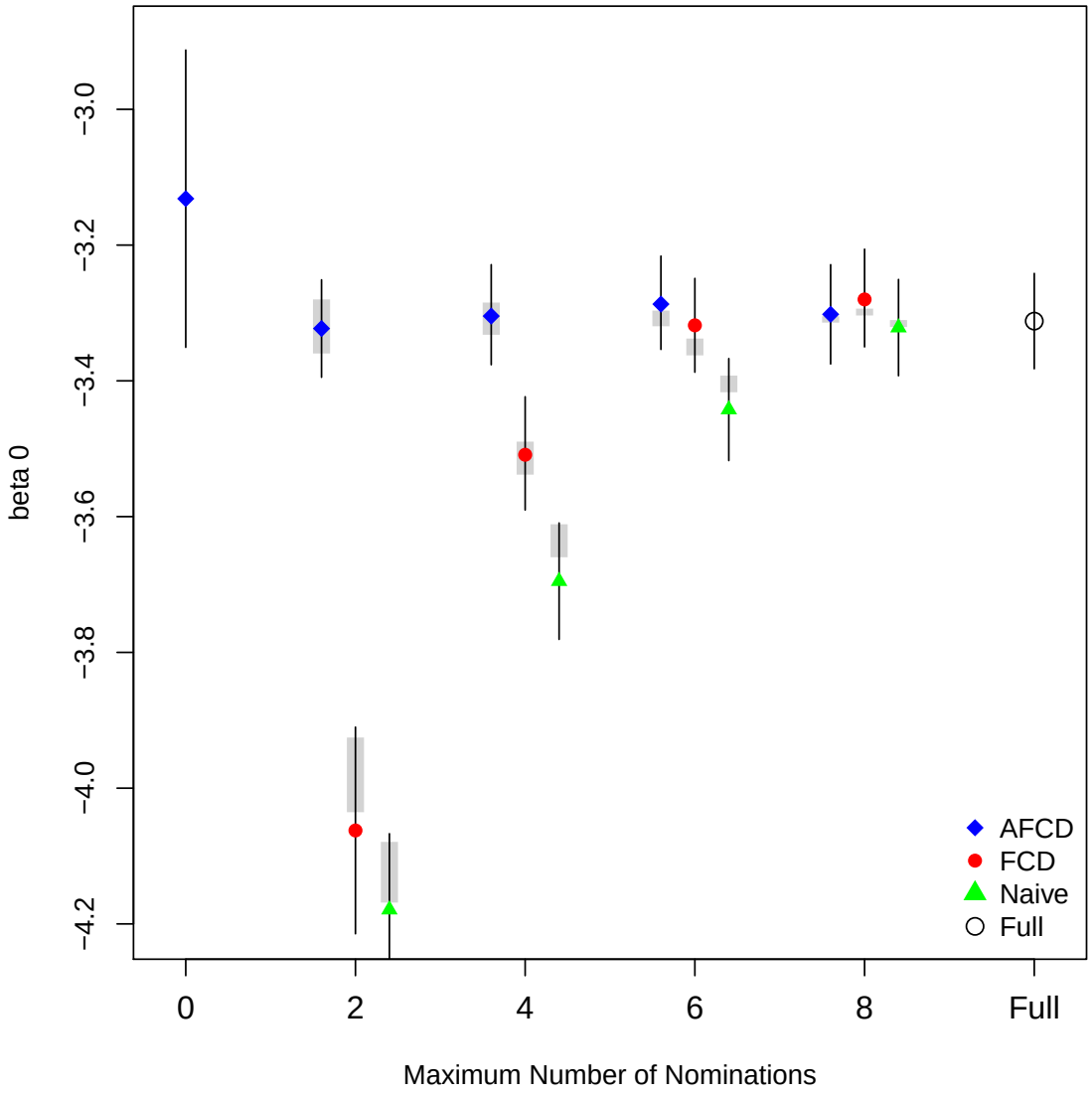

Figure 3.12: Plots of $\hat{\beta}_0 \pm$ 1 bootstrap standard error for the AFCD and FCD and the logistic regression standard error for the naive analysis, deleting nominations in the reverse order in which they were made.

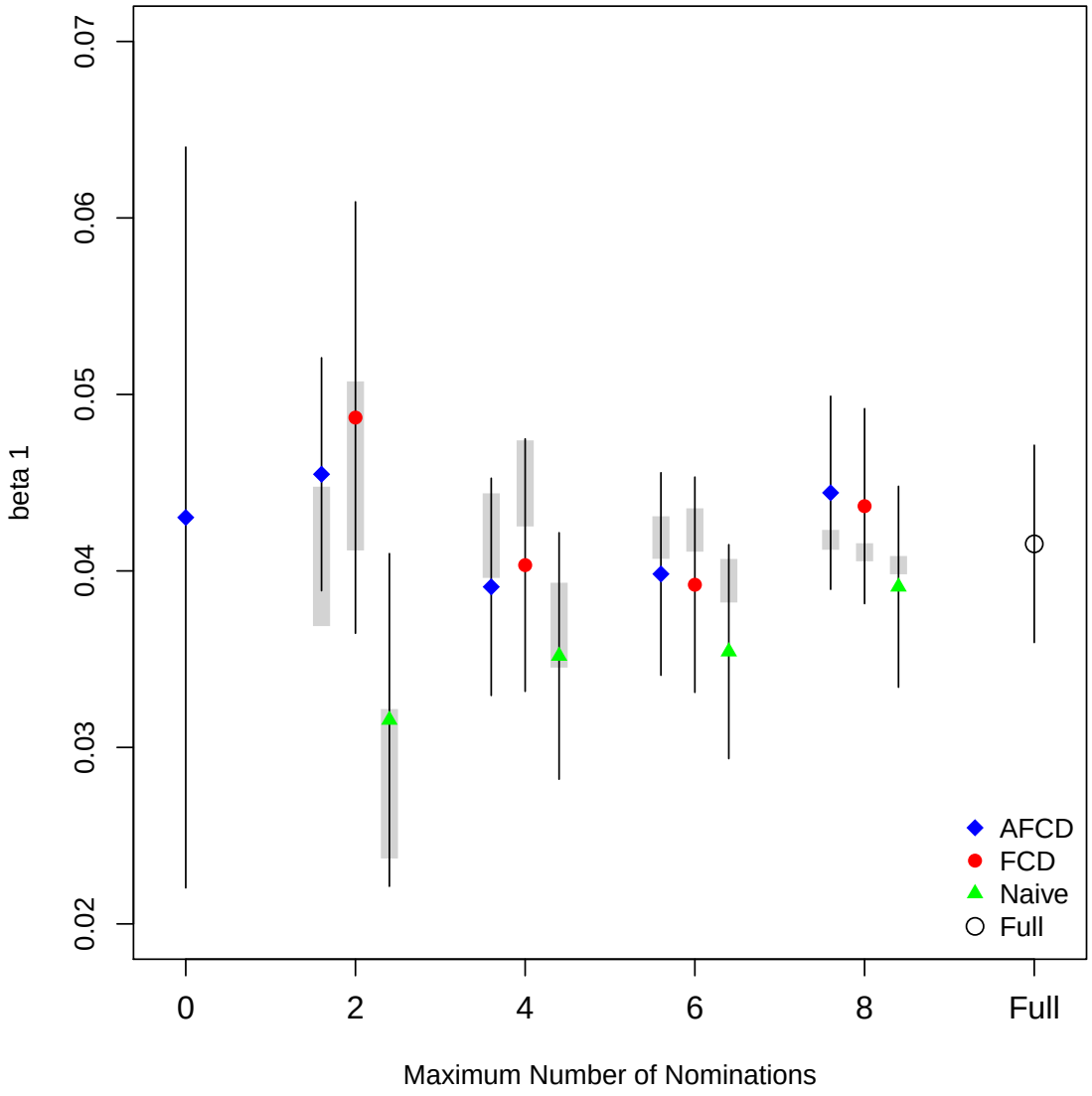

Figure 3.13: Plots of $\hat{\beta}_1 \pm$ 1 bootstrap standard error for the AFCD and FCD and the logistic regression standard error for the naive analysis, deleting nominations in the reverse order in which they were made.

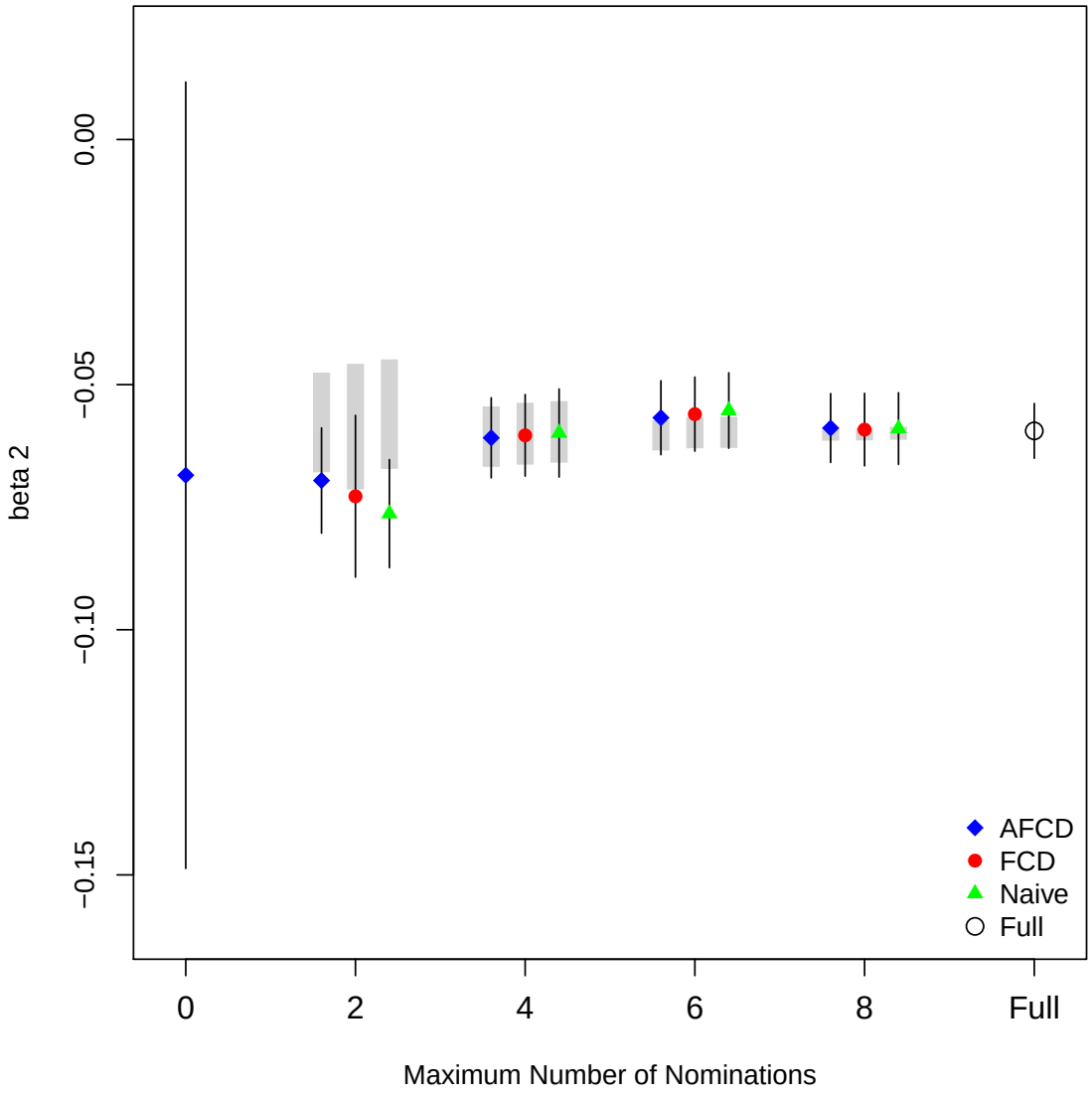

Figure 3.14: Plots of $\hat{\beta}_2 \pm$ 1 bootstrap standard error for the AFCD and FCD and the logistic regression standard error for the naive analysis, deleting nominations in the reverse order in which they were made.

3.5 Discussion

Collecting complete network information in a closed network may be difficult as the network survey may impose an unreasonable amount of respondent burden. The FCD seeks to ameliorate respondent burden by asking respondents to nominate up to m individuals in the population that they have a particular relationship with.

In this work we demonstrate that estimating associations between behaviors and social relationships from social network data arising from a fixed choice survey design as though it was fully observed (as is the current standard practice) can result in severely biased estimates. We introduce observed data likelihoods for FCD data. We demonstrated that maximizing the observed data likelihood for the FCD improves the MSE in comparison to estimates where the data is (falsely) assumed to be fully observed.

We also introduce the AFCD, a new network survey sampling design and method of analysis which collects information on the total number of relationships each individual in the network has, in addition to the data collected with the standard FCD. This innovative study design can add considerable information to analyses without unduly burdening the survey respondent, resulting in improvements over the FCD and naive analyses. We demonstrate that the AFCD is superior to both the FCD and naive analyses in terms of MSE. The improvement of the AFCD's MSE relative to the FCD and naive analyses is particularly pronounced when the m is small relative the number of nominations that would be endorsed if no maximum is imposed. Unsurprisingly, for larger m, there is less variation and bias all three estimators exhibit in simulation. However, the AFCD estimator when no specific edge information is collected ($m = 0$) is an improvement over the FCD and naive analyses when m is small. In other words, if one were to choose between an AFCD where $m = 0$ and an FCD where $m > 0$, an AFCD where $m = 0$ may be preferred. This finding suggests that there may be instances when surveys should collect information on the number of relationships rather than specific nominations. While collecting all possible nominations from each survey respondent would be optimal in terms of minimizing variance and bias, the AFCD can provide a way to improve estimation while keeping respondent burden low.

Since the AFCD utilizes information on the number of nominations each person would

make without censoring, the estimates of the intercepts are much better with the AFCD than the FCD or the naive analyses. This suggests that the AFCD should be implemented when edge prediction is a goal of the analyses.

Limitations are acknowledged. In this work we assume that nominations are randomly censored. Violation of this assumption may lead to incorrect inference.Indeed, in this work we find that the order in which respondents nominated their peers is related to parameter estimation. In order to address this assumption, further research into survey methodology for AFCD and FCD data is necessary.

The analyses and simulations we present use a dyad independent exponential random graph model. This model does not incorporate important network characteristics including reciprocity, transitivity, and clustering. Modeling network structural characteristics, and allowing for complex dependencies is particularly important when the goal of the model is to impute missing edges, or provide a realistic network model. Alternative models that incorporate the more network characteristics and dependencies include the social relations model (Warnter et al., 1979), the exponential random graph family of models (Frank and Strauss, 1986; Goodreau, 2007; Robins et al., 2004), and the latent space and factor models (Hoff et al., 2002; Hoff, 2009).

Hoff et al. (2012) presented likelihoods for fixed rank and fixed choice design data. Hoff et al. assume that there is an underlying parametric model for the network that generates the ranked or binary social relations data. Using a social relations model, Hoff et al. perform estimation in the Bayesian framework. A benefit of that approach is the ability to accommodate both ranked and binary nominations. However, that method relies upon an underlying parametric model, requiring more stringent assumptions.

CHAPTER 4

UNEQUAL EDGE INCLUSION PROBABILITIES IN RESPONDENT-DRIVEN SAMPLING: IMPLICATIONS FOR INFERENCE

4.1 Introduction

4.1.1 Respondent Driven Sampling Background

Respondent Driven Sampling (RDS) is a method that has been used to estimate the population proportion of some binary trait. RDS has been widely adopted to estimate disease prevalence within populations that are difficult to sample from. The RDS method is a variant of a link-tracing procedure (Goodman, 1961; Handcock and Gile, 2011a). In link-tracing, a number of individuals from the target population are enrolled into the study as 'seeds', who are then asked to recruit other people they know to be members of the target population, who are then asked to enroll those they know to be members of the target population, and so on. Using the sample mean from a link-tracing sample has some inherent flaws, including the fact that the sample mean does not account for non-equal inclusion probabilities. Instead of simply using the mean of the outcome of interest from the sample, RDS prevalence estimates incorporate information about how the sample was collected to estimate the sampling probability π, which is often a function of the reported number of associates each person has in the population, i.e. their degree. In practice, RDS diverges from its theoretical basis in that nodes (and therefore edges) are sampled without replacement and RDS is a branching process (each person may refer more than one other person into the sample). Despite this, much of the current theory underlying RDS methodology assumes that the RDS sampling procedure is with-replacement and non-branching.

4.1.2 Network Terminology

Networks are used to represent systems of inter-related entities. In social networks, people (or groups of people) are represented by nodes, and their inter-relationships are represented by edges. Frank (1977) presented an overview of network concepts, which we summarize here. Edges may be either directed, which implies that the relationship has the potential to not be reciprocated, or undirected in which every relationship is reciprocated. An edge is said to be incident to the two nodes that share a relationship. The degree for each node is equal to the number of edges incident to the node. A walk on an undirected network consists of an ordered series of edges where consecutive edges are incident with a common node. When a walk originating at node A and ending at node B exists, then A is connected to B even if there are no edges incident to both A and B. In this work we will focus exclusively on undirected networks in which each node is connected to every other node in the network, forming a single connected component.

4.1.3 Edge Inclusion Probabilities

Social networks tend to have complex structure, and are often difficult to observe in their entirety. Frank (1981) reviewed statistical approaches to networks, including network sampling. As in other settings, network sampling may either be adaptive (the sampling procedure uses information from those previously included in the sample to choose subsequent observations to sample), or non-adaptive (Thompson, 2006). Examples of adaptive network sampling include snowball sampling (Goodman, 1961; Handcock and Gile, 2011b), Bayesian adaptive link tracing (St Clair and O'Connell, 2012; Chow and Thompson, 2003), and Respondent-Driven Sampling (Heckathorn, 1997, 2002; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008; Gile, 2010, 2011; Gile and Handcock, 2011). In each of these adaptive network sampling strategies, initial nodes are chosen in some fashion, and then some subset of the nodes that are a walk of length 1 away from the sampled nodes are then sampled, and so on until some stopping point is reached. In this way, both nodes and edges are sampled.

Often these adaptive network sampling approaches build on the theory of random-walks

(Lovasz, 1993), where the random walk forms a first-order Markov chain (Goel and Salganik, 2009). A random walk on a network begins when a node is randomly sampled. Next, another node which shares an incident edge with the initial node is chosen with uniform probability as the next step in the walk, and this procedure continues to produce a sequence of nodes and edges which compose the random walk. Under the assumptions that 1.) the network is undirected, 2.) nodes are selected uniformly and with replacement, 3.) the initial node is chosen with probability proportional to degree, then the draw-wise probability of sampling an edge is uniform.

Our purpose in this paper is to show how the deviations from the with-replacement non-branching random walk setting result in un-equal edge sampling probabilities in RDS samples, and introduce a new estimation approach to address the biases in prevalence estimation that these violations incur. We will demonstrate that edge sampling probabilities are non-uniform when sampling of individuals is done without replacement in Section 4.2. In Section 4.3, we will present how RDS prevalence estimators are sensitive to the violation of this assumption. In Section 4.4, we propose a method for estimating edge inclusion probabilities, and we will address limitations in current methodology by presenting a prevelence estimator which utilizes estimated edge inclusion probabilities in Section 4.5. In Section 4.6 we compare this new RDS prevalence estimator to existing estimators through simulation studies. In Section 4.7 we will apply this novel method to the New York Jazz dataset. In Section 4.8 we present a brief discussion.

4.2 Unequal Inclusion Probabilities

4.2.1 Rationale for Uniform Edge Inclusion Assumption

Previous work in RDS has treated the RDS sampling procedure as a Markov Chain (Volz and Heckathorn, 2008; Salganik and Heckathorn, 2004), where the probability that a node i is sampled at stage x is proportional to the degree of that node, d_i. Next, one (or more) of the d_i edges incident to node i is randomly chosen, leading to another node j to be sampled at stage $x + 1$. In this way the sampling process alternately samples between nodes and edges. Assuming that this sampling process is made with-replacement and is at stationarity,

the probability that node i is sampled at any given time is in proportion to d_i, and the probability that any edge e_{ij} is sampled at any time is equal to the probability that node i is sampled times the probability that e_{ij} is chosen of the d_i possible edges connected to node i. Therefore, the draw-wise probability of choosing any edge $e_{kl} = d_k / \sum_j d_j \times d_k^{-1} = (\sum_j d_j)^{-1}$ for any nodes k, l that are connected by an edge, where $\sum_j d_j$ is equal to 2 times the number of undirected edges in the network. Therefore, under these assumptions, the draw-wise edge sampling probabilities are all equal.

Salganik and Heckathorn (2004) claim that while RDS is a without-replacement sampling process, it is valid to assume equal edge inclusion probabilities since generally the sampling proportion of RDS studies is relatively low. We will show that this claim does not hold in many situations.

4.2.2 Simple Example of Unequal Edge Sampling Probabilities

Consider the network composed of five nodes and eight undirected edges in Figure 4.1. Using this network as an example, we will show that random non-repeating walks may result in non-equal edge sampling probabilities. In this example we consider walks that visit four nodes and three edges, or walks of length three. One such walk could start at node A and proceed to nodes B, C, and D in that order. This walk is displayed in Figure 4.2, where the labels refer to the order in which each node is sampled. Since node A was sampled first, node A is labeled with a zero, node B is labeled with a 1, and so on. Assuming that the seeds are chosen with probability proportional to degree, the probability that we would observe the random-walk sequence:$\{A, B, C, D\} = 3/16 \times 1/3 \times 1/2 \times 1/2 = 1/64$.

By enumerating each possible random-walk sequence and finding their corresponding probabilities, we can also find the probability of sampling the edges and nodes. The edge sampling probabilities for this simple network are displayed in Table 4.1. Since we find for example that the edge sampling probability of AC does not equal the edge sampling probability of edge CA we have enumerated all directed edges.

Table 4.1 shows that the edges that are incident to node C, which has the highest degree in the network, have a lower probability of being sampled.

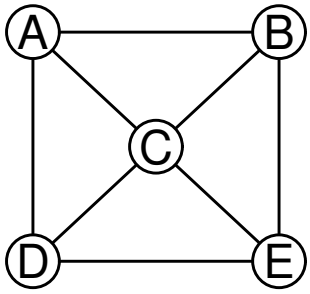

Figure 4.1: Network Example

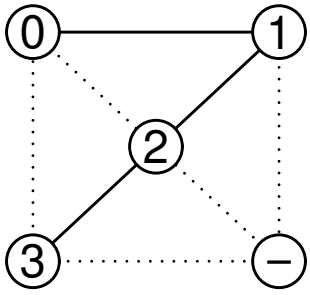

Figure 4.2: Random Walk Example

Table 4.1: Edge Inclusion Probabilities for Random-Walks on Network in Figure 4.1

Edge	P(sampled first)	P(sampled second)	P(sampled third)	Total Sampling Probability
A,B	0.06250	0.06250	0.078125	0.203125
A,C	0.06250	0.06250	0.031250	0.156250
A,D	0.06250	0.06250	0.078125	0.203125
B,A	0.06250	0.06250	0.078125	0.203125
B,C	0.06250	0.06250	0.031250	0.156250
B,E	0.06250	0.06250	0.078125	0.203125
C,A	0.06250	0.06250	0.062500	0.187500
C,B	0.06250	0.06250	0.062500	0.187500
C,D	0.06250	0.06250	0.062500	0.187500
C,E	0.06250	0.06250	0.062500	0.187500
D,A	0.06250	0.06250	0.078125	0.203125
D,C	0.06250	0.06250	0.031250	0.156250
D,E	0.06250	0.06250	0.078125	0.203125
E,B	0.06250	0.06250	0.078125	0.203125
E,C	0.06250	0.06250	0.031250	0.156250
E,D	0.06250	0.06250	0.078125	0.203125

4.2.3 Simulated Edge Sampling Probabilities

In the example with the simple network (Figure 4.1), we showed that edges incident to low-degree nodes have a higher probability of being sampled when a random-walk is made without replacement on that particular network. To demonstrate that this phenomenon is not specific to that one simple network, we performed simulations of non-repeating random walks on undirected networks with 50 nodes. The simulations proceeded as follows:

1. Generate an undirected Bernoulli network where each node pair i, j has a 40% chance of sharing an incident edge if $i \neq j$
2. Keep track of the number of edges in the network and their incident degrees
3. Choose a node in proportion to degree to start the random walk on the network generated in the previous step
4. With uniform probability, choose an edge from that node that does not go to any previous sampled nodes
5. Repeat steps 2 and 3 until 20 nodes (and 19 edges) have been sampled
6. Keep track of edges that were sampled and their incident degrees

Using the above simulation scheme, we simulated 150,000 networks and took one without-replacement random walk of length 19 on each simulated network. In Figure 4.3 we present the distribution of edge sampling probabilities for edges by incident degree, for incident degree pairs that were observed at least 10,000 times. In the left panel of Figure 4.3, we can see that edges with low-degree incident nodes have a higher probability of being sampled, while edges with higher-degree incident nodes have a lower probability of being sampled. Indeed, it seems that some edges are twice as likely to be sampled than other edges, and that this variation in edge-sampling probability is associated with the degree of the nodes incident to the edges.

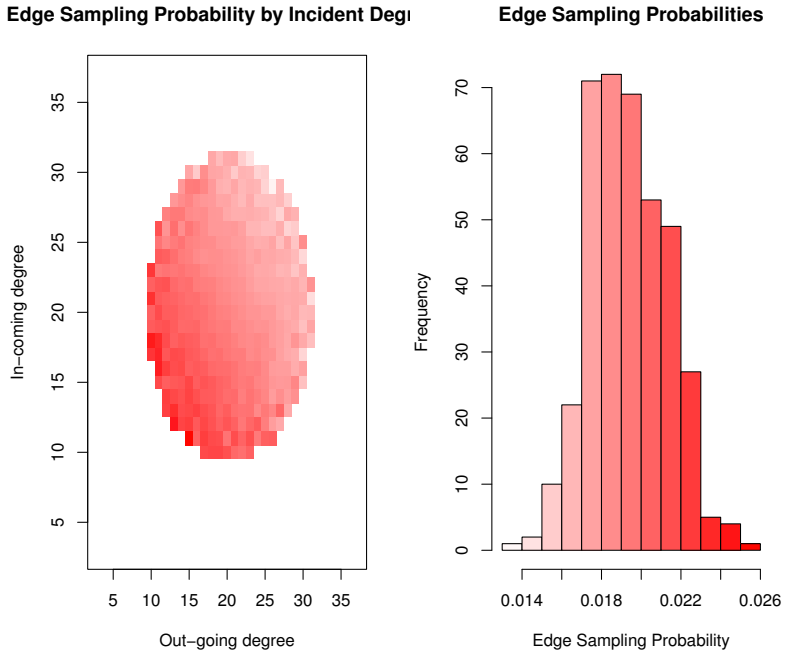

Figure 4.3: Edge Sampling Probabilities for Non-Repeating Random Walk of Length 20 on a Bernoulli Networks with 50 nodes

4.2.4 Developing Intuition for Unequal Sampling Probabilities

The distribution of edge sampling probabilities are a function of the overall structure of the network (including the degree distribution of the nodes), and the self-avoiding quality of a without-replacement random-walk. As we are unable to observe the overall structure of the network, we will focus exclusively on the effect of the degree distributions on sampling probabilities, and will leave the related subject of the overall structure of the network as an area for future research.

Edges that are incident to high-degree nodes tend to have a lower sampling probability than nodes incident to low-degree nodes. Though this might be counterintuitive at first, since nodes with higher degree have a higher probability of being sampled, intuition may be gained by examining two extreme cases. First consider a non-repeating walk on a network with n nodes of length $n-1$. Here $n-1$ edges have been traversed in the walk, and each node has the same probability of being sampled. In this extreme situation, edges that are incident to nodes with degree 2 have a probability of being sampled equal to one, edges

incident to nodes with degree 3 will have an average probability of 2/3 of being sampled, and edges incident to nodes with degree 50 have an average probability 2/50 of being sampled and so on, given that those nodes are not the first or last node in the walk. Now consider a random walk of length 1 where the seed is chosen in proportion to degree, with higher degree nodes having a higher probability of being sampled as the seed. The probability of sampling any given edge is equal to the probability that the node incident to that edge is sampled, times the probability that that edge is chosen given their incident node is sampled. Considering this, when there is a non-repeating random-walk of length 1, all edges have an equal probability of being sampled.

Most non-repeating random walks will lie somewhere between the two extremes described above. In general, when a low proportion of nodes are sampled, we will be closer to the extreme where each edge has equal inclusion probabilities, and when a large proportion of nodes are sampled, we will be closer to the extreme where the probability of sampling an edge is highly dependent on the incident degrees of that edge.

4.3 Implications of Unequal Edge Sampling Probabilities

The premise of uniform edge sampling probability underlies many different facets of estimation and diagnostic assessments of RDS. In this section we will first introduce the most commonly implemented RDS prevalence estimators, then explore how falsely assuming uniform edge sampling probabilities can bias RDS estimators of prevalence, with a particular focus on the Salganik-Heckathorn (SH) estimator (Salganik and Heckathorn, 2004).

4.3.1 RDS prevalence estimators

Several RDS prevalence estimators have been proposed and implemented (Heckathorn, 1997, 2002; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008; Gile, 2011; Gile and Handcock, 2011), and it has been demonstrated that no one estimator is the superior estimator (Tomas and Gile, 2011). Here we will provide a brief review of the three most commonly implemented estimators: the Volz-Heckathorn estimator (VH), the successive sampling estimator (SS), and the SH. The VH and SS estimators estimate the probability of observing

individuals based on their degree, and use this probability to perform inverse-probability weighting. The SH estimator relies on the assumption that each edge has an equal probability of being included in the sample, and leverages edge-wise information to estimate prevalence.

One such RDS prevalence estimator was put forth by Volz and Heckathorn (Volz and Heckathorn, 2008). For each person that is included in the sample, the indicator of having a particular characteristic (such as being HIV positive), z_i (z_i=1 indicating that person i is HIV positive, z_i=0 indicating otherwise), and the degree, d_i is ascertained. The numerator of this estimator sums over all z_i (where $i \in 1 \dots n$ are included in the sample), weighted by the probability of sample inclusion which is assumed to be proportional to their degree d_i. Ideally, we would divide this sum by the number of individuals in the population, however this quantity is unknown and is estimated by the sum of the inverse of the degrees for all those included in the study, which is drawn from the generalized Horvitz-Thompson estimator (Thompson, 2002) ratio estimator:

$$\widehat{\mu_{VH}} = \frac{\sum_{i=1}^{n} \frac{z_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}}.$$

While VH performs well in many settings it is subject to bias when there is differential recruitment (individuals passing along coupons do so with a preference for certain groups), differential recruitment effectiveness (individuals passing coupons in one group are more likely to disperse all their coupons than individuals in the other group), and differential activity (individuals in one group tend to have a higher degree than individuals in the other group) in the presence of a large sample fraction(Tomas and Gile, 2011). Further the VH estimator is prone to be affected by large sampling proportions.

The SS is in a very similar form to the VH estimator, though the VH estimator assumes that sampling probabilities are proportional to degree, the SS estimator models these probabilities as a without replacement process (Gile, 2011). The SS estimator addresses many of the short-comings of the VH estimator (Gile, 2011), in that it is not subject to bias when there are large sampling proportions, and that it has lower bias induced by differential activity, though it is still subject to bias resulting from differential recruitment effectiveness.

Here we see that the formula for the SS estimator,

$$\widehat{\mu_{SS}} = \frac{\sum_{i=1}^{n} \frac{z_i}{\tilde{\pi}(d_i)}}{\sum_{i=1}^{n} \frac{1}{\tilde{\pi}(d_i)}},$$

differs from the formula for the VH estimator as the estimated sampling probability is a function of the degree, rather than directly proportional to degree.

Salganik and Heckathorn (2004) developed a RDS design-based estimator of prevalence. In contrast to the VH and SS estimators, the SH estimator uses information on the number of between group ties in an RDS sample, rather than the number of nodes from the group of interest. While we assume that the underlying network is undirected, in the RDS sample we observe directed edges. For instance, consider the case where we are interested estimating the proportion of a networked-populations that is HIV positive $P_{(+)}$, and we know $T_{(-+)}$ is the number of ties that go from an HIV negative person to someone who is HIV positive, and $T_{(+-)}$ is the number of ties that go from an HIV positive person to an HIV negative person. Since we assume the network is undirected the total number of ties from someone who is HIV negative to someone HIV positive must equal the total number of ties from someone HIV positive to someone HIV negative. Therefore, if $T_{(+-)} = a \times T_{(-+)}$ then there must be a times as many people who are HIV negative than HIV positive. Using this equality, we can use the number of cross-ties from each group to find the prevalence:

$$P_{(+)} = \frac{T_{(-+)}}{T_{(-+)} + T_{(+-)}}.$$

By symmetry we also have that:

$$P_{(-)} = \frac{T_{(+-)}}{T_{(-+)} + T_{(+-)}}.$$

Next we substitute $D_{(-)} \times C_{(-+)}$ for $T_{(-+)}$ and $D_{(+)} \times C_{(+-)}$ for $T_{(+-)}$ where $D_{(-)}$ is the average degree for those who are HIV negative and $C_{(-+)}$ is the number of cross-group ties divided by the total number of ties incident to someone who is HIV negative:

$$P_{(+)} = \frac{D_{(-)} \times C_{(-+)}}{D_{(-)} \times C_{(-+)} + D_{(+)} \times C_{(+-)}}.$$

In Figure 4.4 we display a simple network where 3 nodes represent HIV positive people and 2 nodes represent HIV negative people. In the figure the cross-group edges are represented by dashed lines, and the within-group edges are represented by solid lines. In this example, $D_{(+)} = 10/3, D_{(-)} = 3, C_{(+-)} = 4/10$, and $C_{(-+)} = 4/6$. Using these values in the above formula, we find that

$$P_{(+)} = \frac{3\frac{4}{6}}{3\frac{4}{6} + \frac{10}{3}\frac{4}{10}} = 0.60$$

$$P_{(-)} = \frac{\frac{10}{3}\frac{4}{10}}{\frac{10}{3}\frac{4}{10} + 3\frac{4}{6}} = 0.40.$$

However, if we knew the entire network structure and the HIV status of each member of the population, we wouldn't need to perform RDS. Salganik and Heckathorn (2004)'s method involves performing RDS, while keeping track of the between and within group referrals and the degree for each participant sampled. They estimate:

$$\widehat{C}_{(+-)} = \frac{r_{(+-)}}{r_{(+-)} + r_{(++)}}$$

$$\widehat{C}_{(-+)} = \frac{r_{(-+)}}{r_{(-+)} + r_{(--)}},$$

where $r_{(+-)}$ is the number of referrals from someone who is HIV positive to someone who is HIV negative, $r_{(++)}$ is the number of referrals from someone who is HIV positive to another person who is HIV positive, and so on. Further, $D_{(+)}$ and $D_{(-)}$ are also unknown and need to be estimated in order to compute prevalence estimates. Like the VH estimator, the SH estimator makes use of the generalized Horvitz-Thompson estimator (Thompson, 2002), but rather than estimating the prevalence, the SH estimator uses the generalized Horvitz-Thompson estimator to estimate the average degree for the positive and negative groups:

$$\widehat{D}_{(+)} = \frac{\sum_{i=1}^{n} z_i}{\sum_{i=1}^{n} \frac{1}{d_i z_i}}$$

$$\widehat{D}_{(-)} = \frac{\sum_{i=1}^{n}(1-z_i)}{\sum_{i=1}^{n}\frac{1}{d_i(1-z_i)}}$$

$$\widehat{P}_{(+)} = \frac{\widehat{D}_{(-)} \times \widehat{C}_{(-+)}}{\widehat{D}_{(-)} \times \widehat{C}_{(-+)} + \widehat{D}_{(+)} \times \widehat{C}_{(+-)}}$$

$$\widehat{P}_{(-)} = \frac{\widehat{D}_{(+)} \times \widehat{C}_{(+-)}}{\widehat{D}_{(+)} \times \widehat{C}_{(+-)} + \widehat{D}_{(-)} \times \widehat{C}_{(-+)}}.$$

The SH estimator out performs other estimators in certain situations, particularly when there is differential recruitment effectiveness (Tomas and Gile, 2011). However, the SH estimator has been noted to perform poorly in the presence of differential activity (when infected individuals have different average degree than non-infected individuals), and when there is a large sample fraction.

4.3.2 Implication for Inference on Nodal Characteristics: Salganik-Heckathorn Estimator

Though the SH estimator is approximately unbiased under the condition of a non-branching random-walk with replacement, when each edge has an equal probability of being sampled, $\widehat{C}_{(+-)}$ is not unbiased for $C_{(+-)}$ and $\widehat{C}_{(-+)}$ is not unbiased for $C_{(-+)}$ when we are sampling without replacement, as we are in RDS. The bias of these estimates will be most pronounced when there are large differences in the average degree for those who are HIV positive and those who are HIV negative. For instance, $\widehat{C}_{(+-)}$ will tend to be an overestimate when those who are HIV negative have a lower degree than those who are HIV positive. Taking this to the logical conclusion, when $\widehat{C}_{(+-)}$ is an overestimate of $C_{(+-)}$, and when $\widehat{C}_{(-+)}$ is an underestimate of $C_{(-+)}$, $\widehat{P}_{(+)}$ will tend to be an underestimate. Referring to the probabilities of the ordered edges in Table 4.1 and the network depicted in Figure 4.4, we can illustrate this phenomenon by calculating the SH estimate of HIV prevalence for this network, using the full network data. First considering $C_{(+-)}$:

$$\widehat{C}_{(+-)} = \frac{n_{(+-)} \times F_{(+-)}}{n_{(+-)} \times F_{(+-)} + n_{(++)} \times F_{(++)}}$$

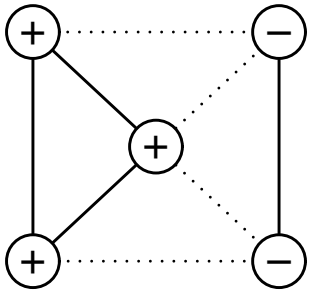

Figure 4.4: Example of Salganik Heckathorn Prevalence Estimation on a Simple Network

Where $n_{(++)}$ is the number of directed edges that begin and end with an HIV positive person in the whole population, $n_{(+-)}$ is the number of directed edges that begin with someone who is HIV positive and end with someone who is HIV negative, $F_{(+-)}$ equals the average probability of observing an edge that originates with someone who is HIV positive and ends with someone who is HIV negative, and $F_{(++)}$ equals the average probability of observing an edge that originates and ends with someone who is HIV positive. The set of edges that originate with an HIV positive person and end with an HIV negative person is $\{(A,B),(C,B),(C,E),(D,E)\}$ and have an average sampling probability of 0.1953125. The set of edges that originate and end with someone who is HIV positive is $\{(A,C),(C,A),(A,D),(D,A),(C,D),(D,C)\}$ and has an average sampling probability of 0.1822917. In order to find the value of $\widehat{C}_{(+-)}$ we plug in the above values:

$$\widehat{C}_{(+-)} = \frac{4 \times .1953125}{4 \times .1953125 + 6 \times .1822917} = 0.416667,$$

whereas $C_{(+-)} = 0.40$. In a similar fashion, we would find $\widehat{C}_{(-+)} \approx 0.6388889$ when $C_{(-+)} = \frac{2}{3} \approx 0.666667$. Assuming that we know $D_{(+)}, D_{(-)}$ we find that:

$$\widehat{P}_{(+)} = \frac{3 \times 0.6388889}{3 \times 0.6388889 + \frac{10}{3} \times 0.416667} = 0.369863 \neq P_{(+)} = 0.40.$$

This illustrates that the SH estimator may produce biased results even in the simplest network setting.

Table 4.2: Simulated RDS With and Without Replacement

	MSE $\times 10^3$		$\lvert\text{Bias}\rvert$ $\times 10^3$		SE$\times 10^5$	
Estimator	w/ R	w/o R	w/ R	w/o R	w/R	w/o R
CAB	7.22	6.93	7.86	17.34	15.49	14.86
CBA	0.41	0.53	0.44	11.94	3.71	3.60
DA	1217.68	945.08	44.10	16.75	201.31	177.46
DB	383.20	283.51	3.68	150.08	113.31	93.27
SH	2.52	3.01	2.24	29.25	9.16	8.48

Table 4.3: Simulated RDS With and Without Replacement

4.3.3 Simulation Study on Effect of Without-Replacement Sampling on SH Estimator

In order to further demonstrate the bias that is incurred in the SH prevalence estimator, we simulated a network on 1000 nodes, where 300 nodes belong to group A, and 700 nodes belong to group B. We generated an edge between nodes both in group A with probability with a probability of 0.02, edges between nodes in different group were generated with a probability of 0.03, and edges between nodes both in group B were generated with probability 0.12. Using these edge probabilities induces differential activity.

Next we simulated RDS on this network, starting with 2 seeds, and allowing each node to refer to two other nodes, until the sample size reached 250. We repeated this procedure 300,000 times. For each of the 300,000 simulations, we kept track of which nodes were sampled, which edges were sampled, and calculated the SH prevalence estimates. We repeated these simulations with the key difference that we allowed for visiting the same node more than once, in other words, we allowed RDS to proceed with-replacement.

In Figure 4.5 and Table 4.2 we compare $\widehat{P}_{(A)}, \widehat{C}_{(AB)}, \widehat{C}_{(BA)}, \widehat{D}_{(A)}, \widehat{D}_{(B)}$, from the RDS simulations with and without replacement. In the first row of Figure 4.5, we see that that $\widehat{C}_{(AB)}$ and $\widehat{C}_{(BA)}$ are less biased for $C_{(AB)}$ and $C_{(BA)}$ when sampling is with replacement as compared to when sampling is with replacement, which is confirmed in Table 4.2. Looking at Table 4.2, we see that the standard errors for $\widehat{C}_{(AB)}$ and $\widehat{C}_{(BA)}$ are larger when sampling is with replacement as opposed to without replacement, which we would expect since the without replacement process involves sampling more of the network.

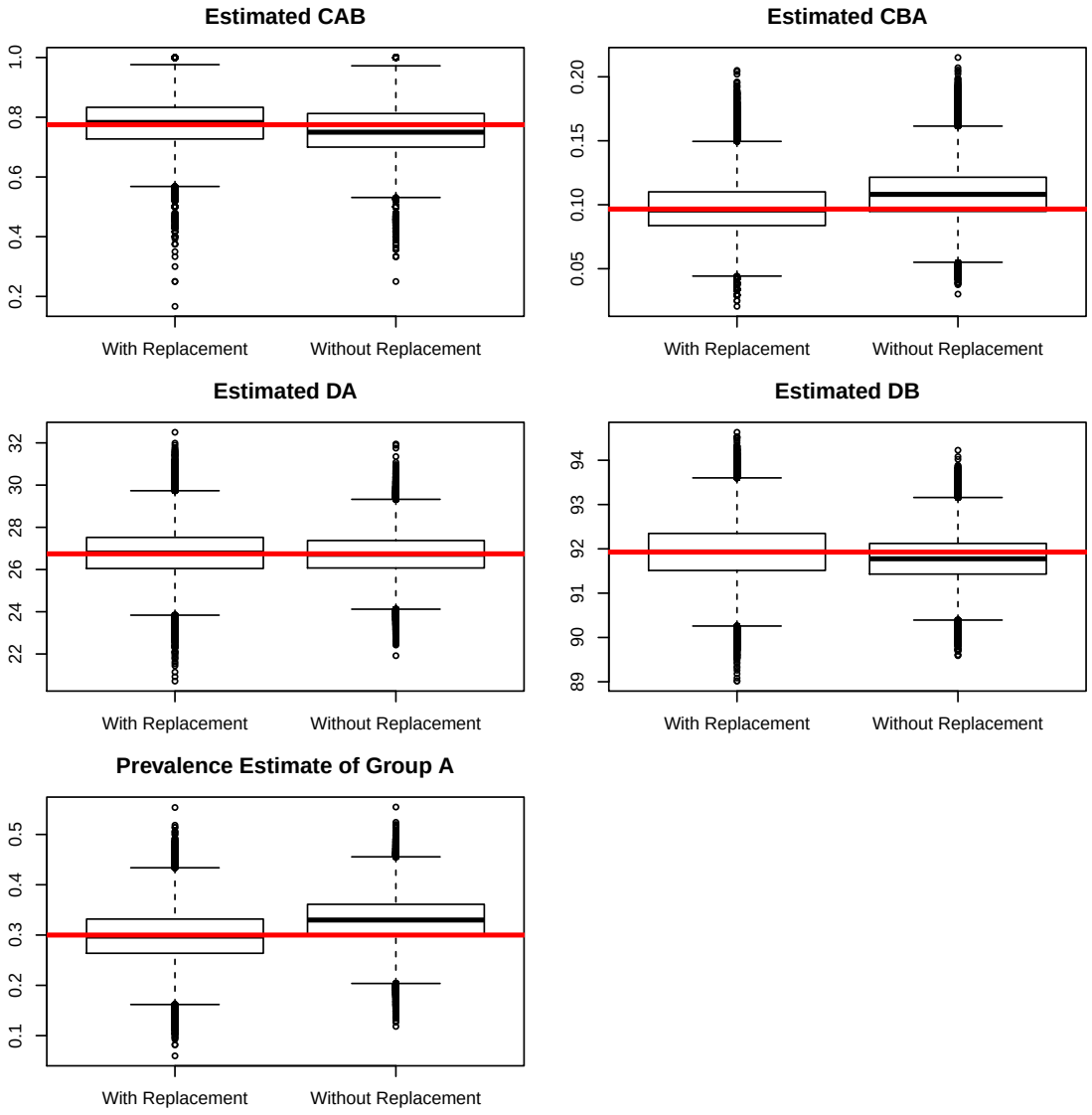

Figure 4.5: Boxplots of $\widehat{C}_{(AB)}, \widehat{C}_{(BA)}, \widehat{D}_{(A)}, \widehat{D}_{(B)}, \widehat{P}_{(A)}$ when sampling is with replacement and without replacement.

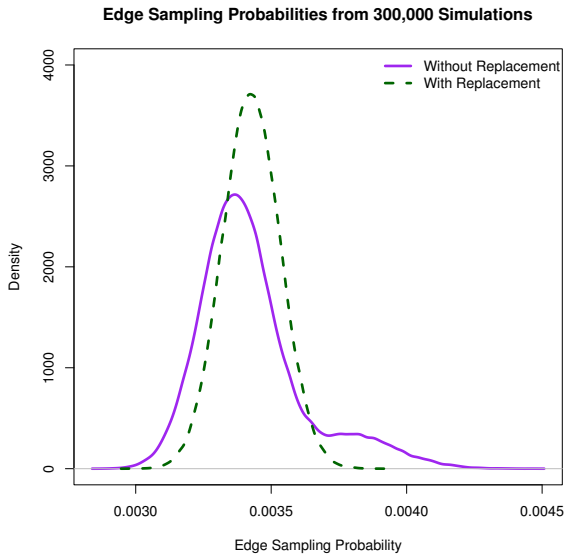

Figure 4.6: Edge Sampling Densities on the Same Network when sampling is with and without replacement.

Figure 4.6 displays the densities of the edge sampling probabilities from the 300,000 simulations for with and without replacement on the same network. The distribution of edge sampling probabilities is more narrow and uni-modal when sampling with replacement, and wider and bimodal when sampling without replacement. Figures 4.7 and 4.8 present heatmaps of the edge sampling probabilities for the without replacement and with replacement simulations by recruiter degree on the x-axis and recruitee degree on the y-axis. In these heatmaps the legend is the same, and higher probability edges are denoted by darker colors, while lower probability edges are denoted by lighter colors. In Figure 4.7 it is apparent that edge sampling probabilities do not vary by the recruiter degree, but do vary by recruitee degree, where edges incident to recruitees with lower degrees have a higher edge sampling probability. In figure 4.8 there is no discernible relationship between edge sampling probabilities and the degrees of the incident nodes.

4.4 Estimating Edge Inclusion Probabilities

We have shown that the without replacement sampling design in RDS results in unequal edge sampling probabilities which have previously been unaccounted for. In order to address

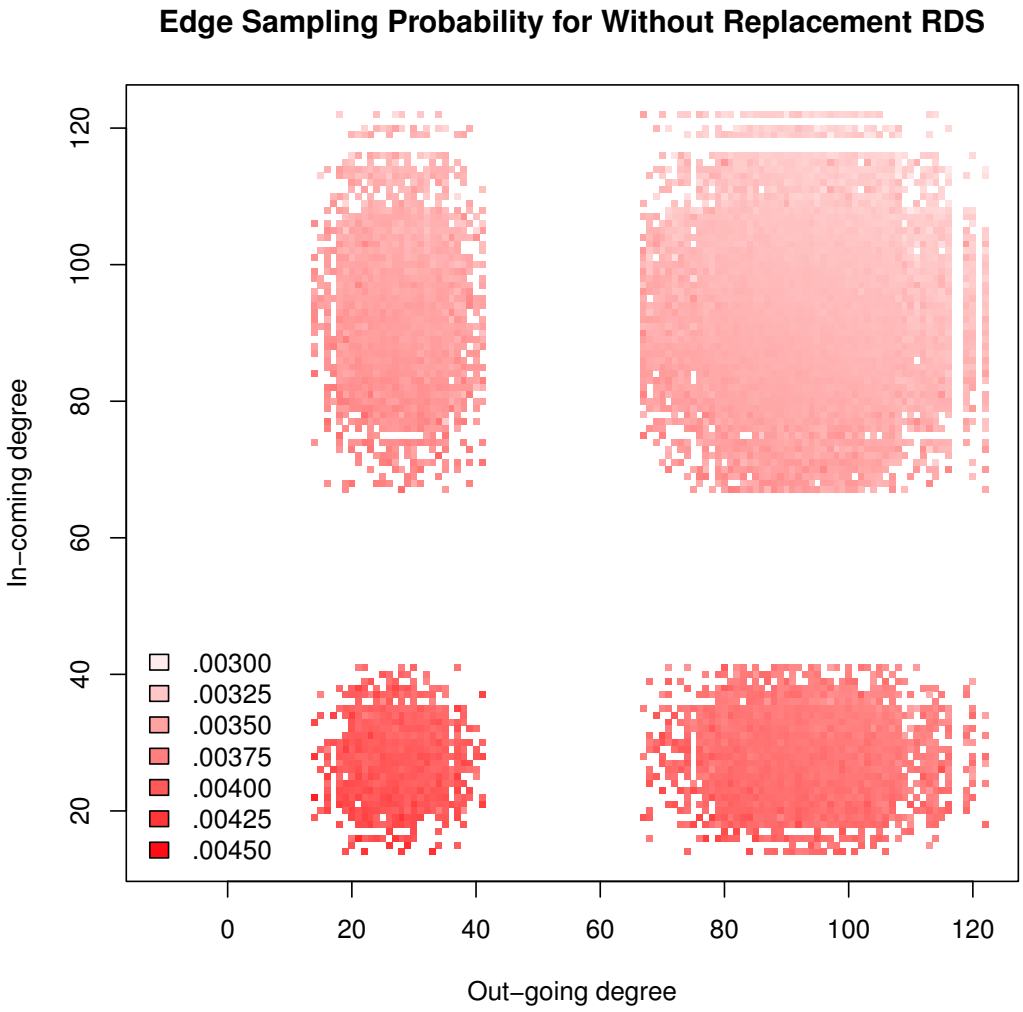

Figure 4.7: Heat Map of Without Replacement Edge Sampling Probabilites

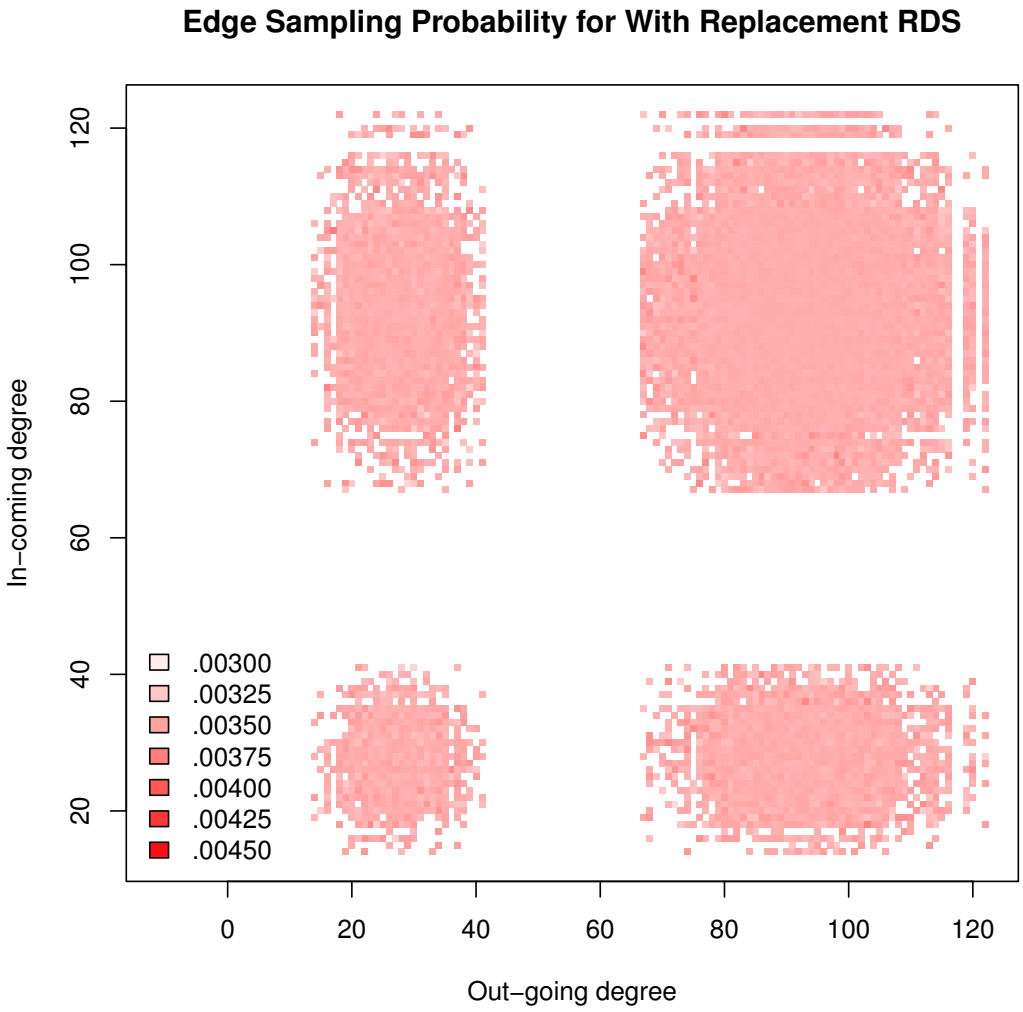

Figure 4.8: Heat Map of With Replacement Edge Sampling Probabilities

the estimation problem resulting from unequal edge inclusion probabilities, we propose to estimate edge inclusion probabilities and use inverse-probability weights. We will estimate the probability that we observe a referral from person i with degree d_i to person j with degree d_j. We specify that $S_{d_i,d_j} = 1$ if a person with degree d_i is sampled as a non-terminal node such that a person with degree d_j is available to be sampled. Given this, the probability person i recruits person j into the sample, given the degree distribution of the underlying population $P(d_i \rightarrow d_j|\mathbf{d})$ is calculated as follows:

$$P(d_i \rightarrow d_j|\mathbf{d}) = P(d_i \rightarrow d_j|S_{d_i,d_j} = 1, \mathbf{d}) \times P(S_{d_i,d_j} = 1).$$

Here we assume that:

$$P(d_i \rightarrow d_j|S_{d_i,d_j} = 1, \mathbf{d}) = \frac{n_c}{g(d_i)},$$

where $g(d_i)$ is the average number of connections incident to person i that are unsampled when i is prepared to pass out coupons. Further, n_c is the maximum number of coupons that each participant may pass on.

The less trivial aspects of estimating $P(d_i \rightarrow d_j|\mathbf{d})$ are estimating $P(S_{d_i,d_j} = 1)$ and $g(d_i)$. Following Gile (2011) we propose a successive sampling method which we describe here as a broad overview and we will follow with a detailed explanation of each step. First, given the observed RDS sample of size n, and the size of the population we are sampling from N(which we assume to be known) we estimate the degree distribution for the underlying population($\mathbb{N} = \mathbb{N}_h, \mathbb{N}_{h+1}, \ldots, \mathbb{N}_K$ where $\mathbb{N}_j$ is the number of nodes with degree j, and h and K are the minimum and maximum degrees), and the degree sampling distribution, f($k, n, \mathbb{N}$, which gives the probability that node i is included in a sample of size n from a population of size N with a degree distribution $\mathbb{N}$ given that $d_i = k$. Next we use our estimates of $\mathbb{N}$ and f($k, n, \mathbb{N}$) to try to recreate the RDS sampling procedure, and simulate the order in which nodes of different degrees are sampled. To simulate the order in which nodes of different degrees are sampled, we repeatedly sample without replacement from $\mathbb{N}$ with probability proportional to degree.

Initially, we estimate f($k, n, \mathbb{N}$) for k in $h \ldots K$ assuming that f($k, n, \mathbb{N}$) is proportional to k:

$$\mathrm{f}(k, n, \mathbb{N}^0) = \frac{k}{N} \sum_{l=h}^{K} \frac{v_l}{l},$$

where v_l is the observed number of individuals sampled with degree l. Next we estimate the distribution of degrees for the underlying population of size N:

$$\mathbb{N}_k^g = N \frac{v_k}{\mathrm{f}(k, n, \mathbb{N}^{g-1})} / \sum_{l=1}^{K} \frac{v_l}{\mathrm{f}(l, n, \mathbb{N}^{g-1})}.$$

Next, we conduct successive sampling and simulate M samples of size n from the degree distribution $\mathbb{N}$ such that nodes are chosen proportional to degree. Using the M samples we next recalculate $\mathrm{f}(k, n, \mathbb{N})$ and $\mathbb{N}$:

$$\mathrm{f}(k, n, \mathbb{N}^g) = \frac{U^k + 1}{M \times \mathbb{N}_k^g + 1}$$

$$\mathbb{N}_k^g = N \frac{v_k}{\mathrm{f}/(k, n, \mathbb{N}^{g-1})} / \sum_{l=1}^{K} \frac{v_l}{\mathrm{f}(l, n, \mathbb{N}^{g-1})}$$

Where U^k is the number of times an observation with degree k was sampled in the M simulated samples. This process is repeated r times so that we may approximate the degree distribution of the underlying population with $\mathbb{N}^r$, and the probability that a node with degree k is included in the sample:

$$\hat{\pi}_k = \mathrm{f}(k, n, \mathbb{N}^r).$$

Having estimated the degree distribution $\mathbb{N}^r$ through successive sampling, we again simulate M resamplings of the population, maintaining the sample size, n, and the number of seeds n_0 that were observed in the actual RDS sample:

1. Draw a sample of size n from $\mathbb{N}$ where nodes are sampled without replacement with probability proportion to degree

2. For all ordered pairs i, j find the number of times a node with degree i is sampled before a node with degree j such that the node with degree i may pass a coupon on

to the node with degree j, forming a $K \times K$ matrix C.

3. Repeat steps 1 and 2 M times

4. Let W be a $K \times K$ matrix, where $W_{ij} = \frac{C_{ij}+1}{\mathbb{N}_i \times \mathbb{N}_j \times M+1}$

5. For the diagonal elements of W: $W_{ii} = \frac{C_{ii}+1}{(\mathbb{N}_i-1) \times \mathbb{N}_i \times M+1}$

Then we estimate $P(S_{d_i,d_j} = 1)$ as $W_{d_i d_j}$.

Again using the M resamplings with degree distribution $\mathbb{N}^r$, we estimate $g(d_i)$ depending on the order in which nodes are sampled, so that if a node i is resampled as a seed, we assign $g(d_i) = d_i$, since all all connections of a node sampled as a seed could be available to be passed on to. If node i is the x^{th} node sampled, and not a seed, in a population of N nodes, we assign $g(d_i) = d_i \times (N - x)/N$, and take the average of these across the M simulations.

4.5 Improving upon the SH estimator

Recall that the SH estimator estimates prevalence by estimating the average number of cross-ties from someone who is HIV negative to HIV positive, and the average number of cross-ties from someone who is HIV positive to someone who is HIV negative. Here we use the new inclusion probabilities to improve upon the SH estimator. Let $\hat{q}_{di,dj}$ be the estimated probability that an edge originating with someone with degree d_i and terminating with someone with degree d_j is included in the sample. Let r_{ij} be the indicator that person i passed a coupon to person j, and recall that $z_i = 1$ if i is HIV positive, and $z_i = 0$ if person i is HIV negative. Then we have:

$$\widehat{C}_{W(+-)} = \frac{\sum_i \sum_{j:i \neq j} r_{ij} z_i (1 - z_j)/\hat{q}_{di,dj}}{\sum_i \sum_{j:i \neq j} r_{ij} z_i (1 - z_j)/\hat{q}_{di,dj} + \sum_i \sum_{j:i \neq j} r_{ij} z_i z_j/\hat{q}_{di,dj}},$$

$$\widehat{C}_{W(-+)} = \frac{\sum_i \sum_{j:i \neq j} r_{ij} (1 - z_i) z_j/\hat{q}_{di,dj}}{\sum_i \sum_{j:i \neq j} r_{ij} (1 - z_i) z_j/\hat{q}_{di,dj} + \sum_i \sum_{j:i \neq j} r_{ij} (1 - z_i)(1 - z_j)/\hat{q}_{di,dj}}.$$

The SH relies on the assumption that individuals are sampled in proportion to their degree in order to estimate the average degree for those who are HIV positive and those who are HIV negative. Here we follow Gile (2011) and make use of the successive sampling estimator to estimate $D_{(+)}$ and $D_{(-)}$ using $\hat{\pi}_{di}$, the estimated probability that someone with degree d_i is included in the sample:

$$\widehat{D}_{W(+)} = \frac{\sum_i z_i d_i \hat{\pi}_{di}^{-1}}{\sum_i z_i \hat{\pi}_{di}^{-1}},$$

$$\widehat{D}_{W(-)} = \frac{\sum_i (1 - z_i) d_i \hat{\pi}_{di}^{-1}}{\sum_i (1 - z_i) \hat{\pi}_{di}^{-1}}.$$

Now we have

$$\widehat{P}_{(+)WSH} = \frac{\widehat{D}_{W(-)} \times \widehat{C}_{W(-+)}}{\widehat{D}_{W(-)} \times \widehat{C}_{W(-+)} + \widehat{D}_{W(+)} \times \widehat{C}_{W(+-)}}$$

which we use to estimate prevalence. Note that the weighted SH prevalence estimator is in the same form as the original SH estimator, but includes adjustments for unequal edge sampling probabilities.

4.5.1 Variance of Weighted SH Estimator

Variance estimation of RDS prevalence estimators is an open area of research. Since RDS may have increased variance due to the inherently correlated observations, the variance for an RDS prevalence estimate will in general be larger than a mean taken from a simple random sample. Currently the standard variance estimator for RDS prevalence estimates is the bootstrap variance estimator put forth by Salganik (2006). However, the Salganik bootstrap variance estimate for RDS point estimators tends to be under-estimate the variance (Goel and Salganik, 2010). Understanding the variance of the weighted SH estimator that we propose here, as well as the variance of other RDS estimators, is important though we do not address it in this work. Here we use the Salganik bootstrap variance estimate, which we describe here:

1. Categorize nodes in the model as either referred by infected, or referred by uninfected.

2. Sample uniformly and with replacement from the RDS sample n_s seeds, where n_s = the number of seeds in the RDS sample. Theese become our bootstrap seeds.

3. For each bootstrap seed that is infected, sample n_c with replacement from the observed rds sample that were referred by someone who is infected.

4. For each bootstrap seed that is uninfected, sample n_c with replacement from the observed rds sample that were referred by someone who is uninfected.

5. Repeat steps 3 and 4 until the sample size in the original RDS sample is reached.

6. Calculate the RDS estimate of disease prevalence $\hat{P}_{bs}$.

7. Repeat steps 2-6 many times to get a distribution of $\hat{P}_{bs}$.

8. Use the distribution of $\hat{P}_{bs}$ to form confidence intervals or calculate standard errors using a Normal approximation.

4.6 Simulation Studies

4.6.1 Simulating and Estimating Edge Inclusion Probabilities

To allow for comparisons with previous RDS work, we used networks from Gile and Handcock (2011). These models are composed of 1000 nodes composed of two groups: $infected$ (200 nodes) and $uninfected$ (800 nodes). The infected group has average degree 11 and the uninfected group has average degree 5.5. The single network that we performed these simulations on has *differential activity* such that the mean degree in the infected group is twice that of the mean degree in the uninfected group. Further, this network has a *homophily* ratio of 1 meaning that the expected number of cross-group ties is equal to the number of cross-group ties we would observe if ties were formed at random given the same proportion of individuals who are infected, and the same degree distributions. These terms are rigorously defined in Table 4.4 for some network Y with N nodes where N^1 is the number of people in the network with the attribute we are trying to find the prevalence of, and $N^0 = N - N^1$:

Table 4.4: Network Parameters for Simulations

Parameter	Definition
Number of Nodes	N
Prevalence	$\mu = \sum_{i=1}^{N} z_i/N$
Mean degree	$\bar{d} = 1/N \sum_{i=1}^{N} d_i/2$
Homophily	$H = \frac{2/(N^1(N^1-1)) \sum_{i,j} z_i z_j \mathbb{E}(y_{ij})}{1/(N^1 N^0) \sum_{i,j} z_i(1-z_j)\mathbb{E}(y_{ij})}$
Differential Activity	$DA = \bar{d}^1/\bar{d}^0$

We performed simulations of respondent driven sampling with a sample size of 200 and 10 seeds and a maximum of 2 coupons per person on a single network of 1000 nodes, 50,000 times and kept track of the number of times someone with degree d_i made a referral to someone with degree d_j given that the person with degree d_i was sampled as a non-terminal node and the person with d_j was available to be sampled, and the number of times someone with degree d_i was sampled as a non-terminal node and the person with d_j was available to be sampled. We then compute the proportion of times these events occur for each degree pair and use these simulated values as a close approximation of $P(d_i \to d_j | S_{d_i,d_j} = 1), P(S_{d_i,d_j} = 1)$.

On the same network of 1000 nodes, we performed simulations of respondent driven sampling (again with sample size of 200, 10 seeds, and 2 coupons per person) 100 times and calculated our estimates of $P(d_i \to d_j | S_{d_i,d_j} = 1)$ and $P(S_{d_i,d_j} = 1)$ for each degree pair for each of the 100 simulated RDS samples. Here we compare our estimates to the simulated values.

Figure 4.9 displays estimates of $P(S_{d_i,d_j} = 1)$ on the y-axis, and simulated values of $P(S_{d_i,d_j} = 1)$ on the x-axis. Figure 4.9 is colored with rainbows(ROY G BIV) with the left panel colored according to the degree of the recruiter and the right panel according to the degree of the recruitee. In both figures the red line denotes the ideal situation where the simulated edge inclusion probability would equal the estimated edge inclusion probability. In the left panel, edges with recruiter degree of 1 are colored dark red, degree 2 is red, degree 3 is light red, degree 4 is dark orange, and so on. Similarly, Figure 4.10 displays the estimates of $P(d_i \to d_j | S_{d_i,d_j} = 1)$ on the x-axis, and the simulated values of $P(d_i \to d_j | S_{d_i,d_j} = 1)$ on the y-axis. In both Figures 4.9 and 4.10, the estimated probabilities are linearly related, though not exactly equal to the simulated probabilities. In particular, when the recruitee

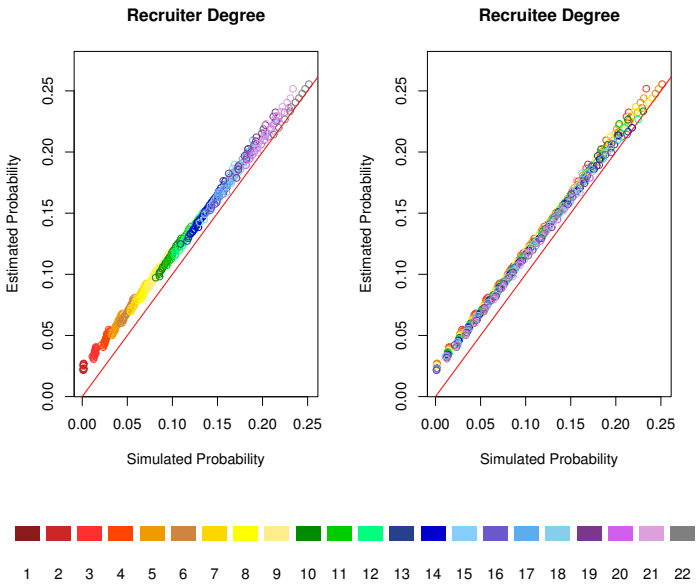

Figure 4.9: Simulated vs. Estimated $P(d_i \rightarrow d_j | S_{d_i,d_j} = 1)$

degree is low, the estimated probabilities are notably higher than the simulated probabilities.

We also simulated the edge sampling probability $P(i \rightarrow j)$ for each edge in the network (without marginalizing over degree pair) based on 100,000 simulations (as above), and compared those simulated edge sampling probabilities to the edge sampling probabilities we calculated after performing simulations of respondent driven sampling 100 times. We plot the simulated edge sampling probabilities on the y-axis and the estimated edge sampling probabilities on the x-axis of Figure 4.11. Figure 4.11 is colored with rainbows(ROY G BIV) with the left panel colored according to the degree of the recruiter and the right panel according to the degree of the recruitee. In the left panel, edges with recruiter degree of 1 are colored dark red, degree 2 is red, degree 3 is light red, degree 4 is dark orange, and so on. By inspecting these plots, we see that we are generally overestimating the $P(d_i \rightarrow d_j)$, where $P(d_i \rightarrow d_j)$ is overestimated more when d_i is small rather than when d_i is large. While our estimates of $P(d_i \rightarrow d_j)$ are certainly not perfect, this still provides a substantial improvement over assuming that all edges have an equal probability of being sampled.

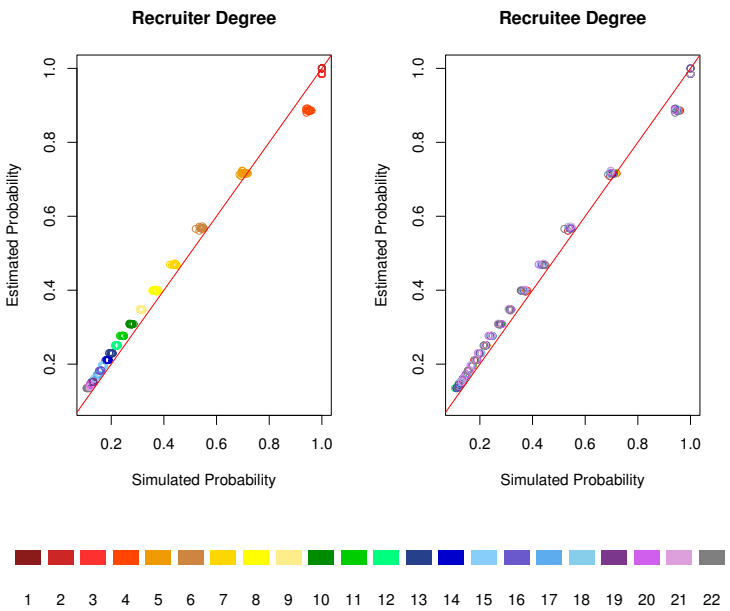

Figure 4.10: Simulated vs. Estimated $P(S_{d_i,d_j} = 1)$

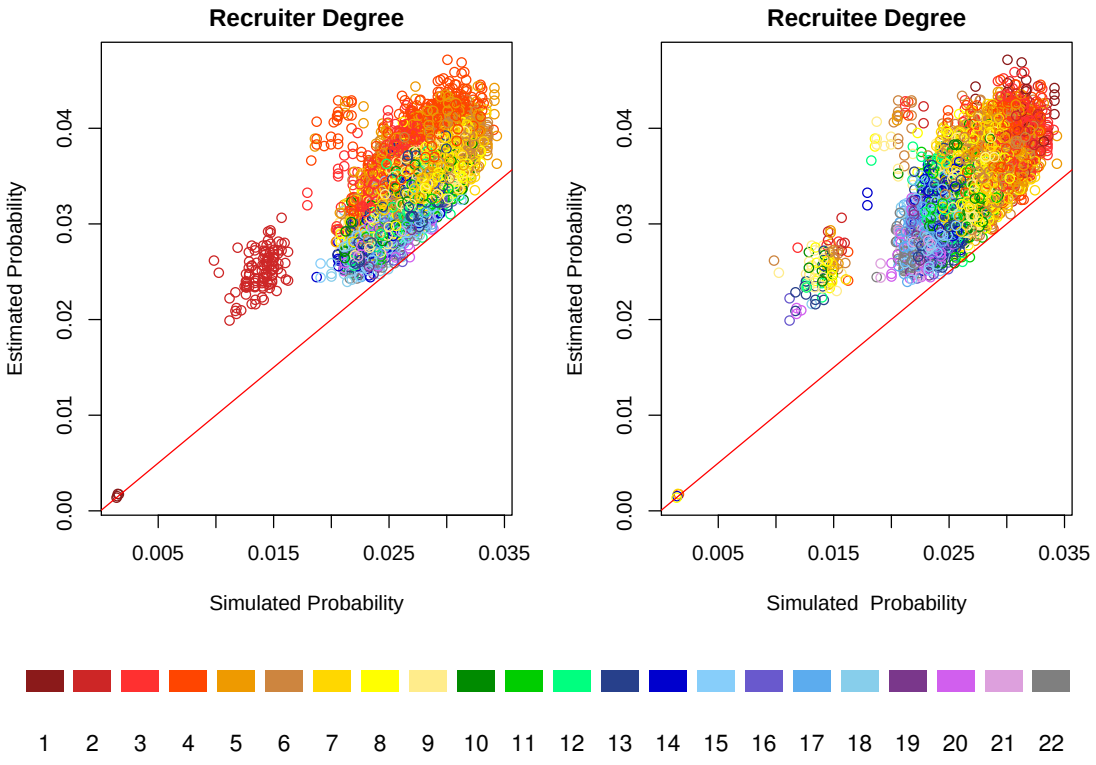

Figure 4.11: Simulated vs. Estimated $P(d_i \rightarrow d_j)$

4.6.2 Simulations on a single network

First we performed 100 simulations of RDS on the same network described in the previous subsection, and using both the simulated and estimated edge inclusion probabilities, re-weighted the $C_{(+-)}, C_{(-+)}$ which we display in the first two panels of Figure 4.12. In this figure, the red horizontal line denotes the true value that we would like to estimate. In each of these first two panels of Figure 4.12 we see that the estimates of $C_{(+-)}, C_{(-+)}$ using the estimated inverse probability weights perform similarly to the those weighted by simulated probabilities, and that the SH estimates of $C_{(+-)}, C_{(-+)}$ perform worse than those inversely weighted by the estimated and simulated edge inclusion probabilities. In the last panel of Figure 4.12 we display the estimated proportion of those who are infected, where the true proportion is 20%. Here we see that by estimating the edge inclusion probabilities and using them to form inverse probability weights we are able to estimate the prevalence of infected individuals just as well as the simulated edge sampling weights, and are able to make a large improvement over the original SH estimator.

4.6.3 Simulations to compare RDS prevalence estimators

Here we simulated RDS on networks and compared the mean of the sample, the SS, the VH, and the original SH to the re-weighted SH. These networks all have 1000 nodes, and a prevalence of 20%. In the simulations we vary the sampling proportion, homophily, and differential recruitment effectiveness, and differential activity.

Differential recruitment effectiveness occurs when individuals in one group are more likely to successfully recruit people into the sample. In these simulations differential recruitment effectiveness was set to either (1,1) or (.9,.6). Differential recruitment effectiveness = (1,1) when individuals in both the infected and uninfected groups would recruit as many people into the sample as they were allowed. Differential recruitment effectiveness = (.9, .6) when individuals in the infected group have a 90% chance of successfully recruiting someone for each coupon they are given, and those in the uninfected group have a 60% chance of successfully recruiting someone for each coupon they are given. In these simulations the number of coupons, $n_c = 2$, so each person in the sample recruited up to two others given

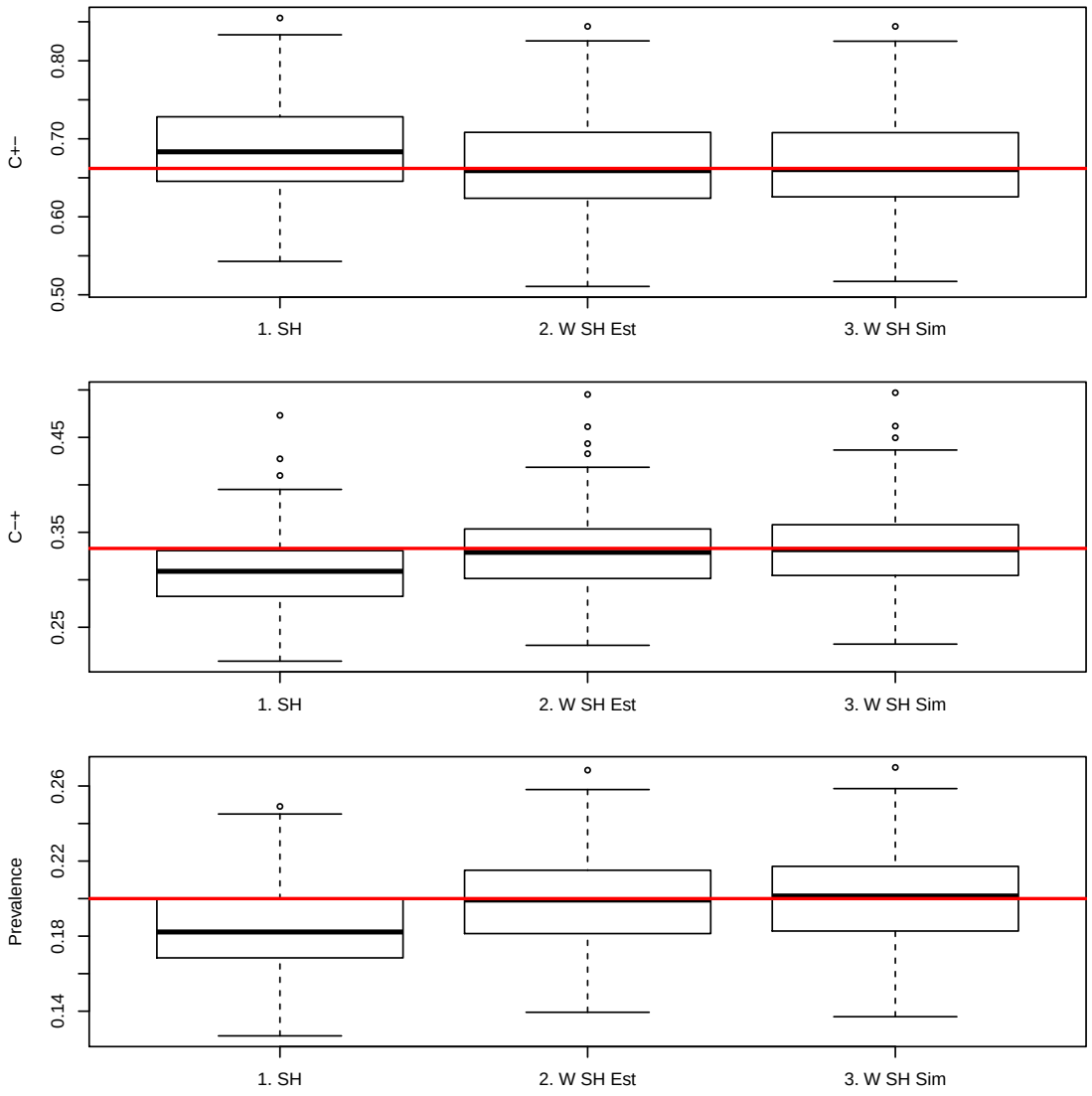

Figure 4.12: Boxplots of stimates of $C_{(+-)}, C_{(-+)}, P_{(+)}$ using the SH estimator, the SH estimator weighted by estimated edge inclusion probabilities, and the SH estimator weighted by simulated edge inclusion probabilities. Here differential activity =2, homophily=1, and differential recruitment effectivenes =(1,1). The true values are denoted by horizontal red lines.

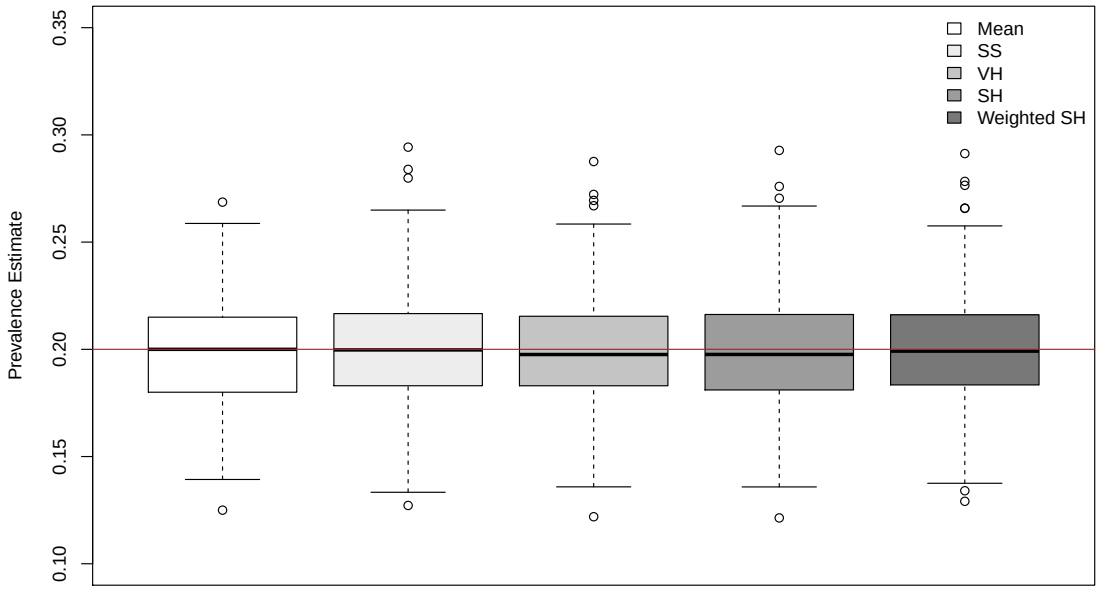

Figure 4.13: Prevalence estimates using the sample mean, SS, VH, SH, and weighted SH estimators without differential recruitment effectiveness, differential activity, or homophily effects. Sampling fraction is 20%. The horizontal red line represents the true prevalence.

Table 4.5: Simulated Prevalence Estimates on Single Network with 1000 Nodes

Estimator	MSE $\times 10^4$	$\lvert$Bias$\rvert$ $\times 10^4$	SE$\times 10^3$
Mean	5.90	11.15	1.72
SS	7.38	9.23	1.92
VH	7.72	9.18	1.97
SH	7.67	3.80	1.96
Weighted SH	7.22	4.84	1.91

the sample size had not yet been attained.

In the first set of simulations, we demonstrate how the weighted SH estimator compared to the other RDS prevalence estimators in the absence of differential activity, differential recruitment effectiveness, and homophily effects. We simulated RDS on 200 networks of size 1000, with prevalence 0.20, and a sampling proportion of 20%. The prevalence estimates from these 200 simulations are presented in Figure 4.13, and the MSE, SE, and empirical bias of these estimates are in Table 4.5. All five estimators, including the naive mean perform comparably when there is no differential activity, differential recruitment effectiveness or homophily present, and the sampling proportion is relatively small.

In the second set of simulations, we used 200 networks where homophily was set to 1 and

differential activity was held constant at 2. In the simulated RDS sampling, the differential recruitment effectiveness was set to (1,1), and we varied the sampling proportion (20%, 50%, 70%). On each of the 200 networks we drew three RDS samples of size 200, 500, and 700, to produce the desired sampling proportions. Boxplots of the estimated prevalence are displayed in Figure 4.14, and Table 4.6 contains the MSE, standard deviations, and bias of the various estimators. In these simulations, regardless of the sampling proportion, the SS estimator performs the best in terms of MSE, and the Weighted SH performing comparably to the SS when the sampling proportion is 20% or 50%. The VH and SH estimators display increasing bias as the sampling proportion increases, and perform similarly in terms of MSE. The mean performs very poorly relative to all the other estimators when the sample size is small, but has a smaller MSE than the VH and SH estimators when the sampling proportion is 70%.

The final set of simulations were performed on 400 networks where homophily was varied to be either 1 or 2 and differential activity was held constant at 2. Among the 200 networks where homophily was set to 1, RDS was carried out with differential recruitment effectiveness set to (1,1) and (.9,.6). Among the 200 networks where homophily was set to 2, RDS was carried out with differential recruitment effectiveness set to (1,1) and (.9,.6). In this set of simulations, the sampling proportion was held constant at 20%. We compare the performance of the estimators from these simulations in Figure 4.15, and in Table 4.7. Regardless of homophily level, when the differential recruitment effectiveness is set to (1,1) all five estimators perform similarly in terms of MSE, except the mean which is positively biased. Similarly,when homophily is set to 1, and differential recruitment effectiveness is set to (.9,.6) the mean has the highest MSE, and the other four estimators perform comparably. However, when differential recruitment effectiveness is set to (.9,.6) and homophily is set to 2, the weighted SH estimator has the lowest MSE, followed by the SH.

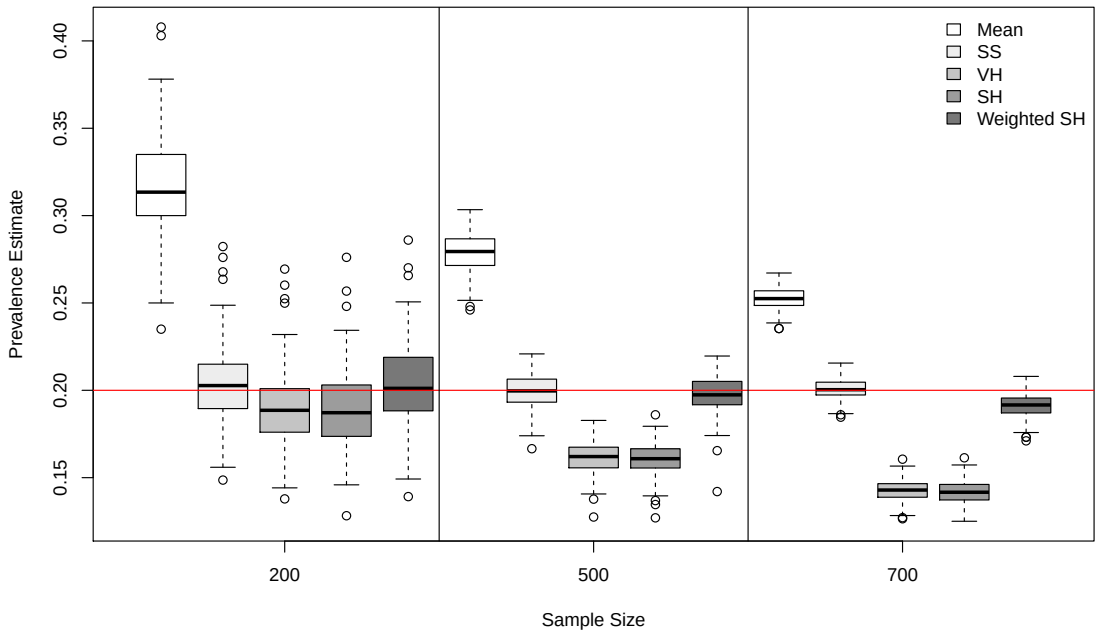

Figure 4.14: Prevalence estimates using the sample mean, SS, VH, SH, and weighted SH estimators for samples of size 200, 500, and 700 from simulated networks of 1000. Here differential activity =2, homophily=1, and differential recruitment effectiveness =(1,1). The horizontal red line represents the true prevalence.

Table 4.6: Simulated Prevalence Estimates on Network with 1000 Nodes, Varying Sample Proportion

	MSE $\times 10^4$			$\lvert$Bias$\rvert$ $\times 10^3$			SE$\times 10^4$		
Estimator	200	500	700	200	500	700	200	500	700
Mean	144.23	63.64	27.89	116.98	78.97	52.48	19.27	8.04	4.22
SS	4.79	0.86	0.28	3.74	0.11	0.73	15.28	6.57	3.73
VH	5.52	15.67	33.41	10.22	38.57	57.49	15.00	6.32	4.20
SH	6.12	16.26	34.51	10.80	39.23	58.40	15.77	6.61	4.54
Weighted SH	5.62	1.16	1.17	3.76	2.26	8.69	16.59	7.45	4.58

Table 4.7: Simulated Prevalence Estimates on Network with 1000 Nodes, Varying Recruitment Effectiveness and Homophily

	MSE $\times 10^3$				$\lvert$Bias$\rvert$ $\times 10^2$				SE$\times 10^3$			
RE:	1,1		.9,.6		1,1		.9,.6		1,1		.9,.6	
H:	1	.2	1	2	1	2	1	2	1	2	1	2
Estimator												
Mean	14.42	17.40	15.23	36.67	11.70	12.11	12.00	18.40	1.93	3.71	2.06	3.75
SS	0.48	1.65	0.72	4.99	0.37	0.75	0.87	5.62	1.53	2.83	1.80	3.04
VH	0.55	1.52	0.55	3.41	1.02	0.66	0.93	4.06	1.50	2.73	1.52	2.97
SH	0.61	1.83	0.64	2.41	1.08	2.14	1.03	3.33	1.58	2.63	1.63	2.56
Weighted SH	0.56	1.63	0.72	1.65	0.38	0.10	0.88	1.07	1.66	2.86	1.80	2.78

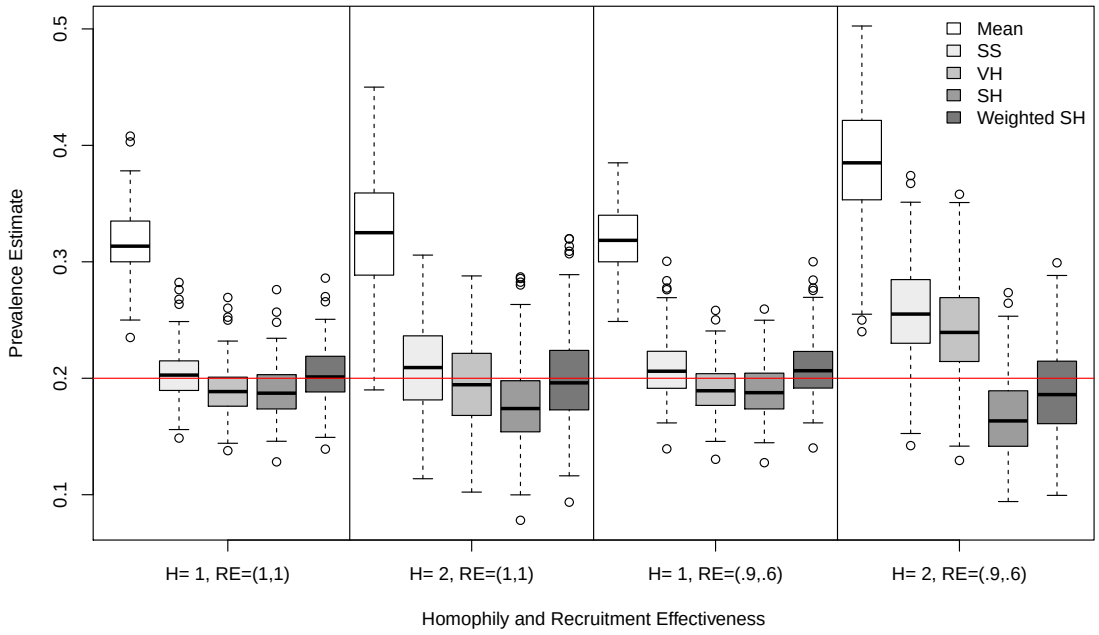

Figure 4.15: Prevalence estimates using the sample mean, SS, VH, SH, and weighted SH estimators for varying homophily $\in (1, 2)$, recruitment effectiveness $\in [(1, 1); (.9, .6)]$, and differential activity =2. Sampling fraction is 20%. The horizontal red line represents the true prevalence.

4.7 Application to NY Jazz dataset

We applied this method to the New York city jazz dataset, which is freely available for download with the RDSAT software (Volz et al., 2012; Heckathorn, 1997). In this dataset, 264 Jazz musicians living in New York City were recruited into the sample with 13 seeds. Subject level variables included sex, race, age, whether their music has been played on the radio, and whether they were members of a musicians union, among other variables. Here we focus on the variable sex, 191 in the sample were coded as male, 68 were coded as female, and 5 had missing data on sex. Here we coded all missing values as males. The average degree in the sample was 206.06, and the average sample degree among males was 201.69, while the average sample degree among females was 218.63. As there seems to be differential activity in this sample, the weighted SH estimator developed in this work may be of use.

In order to apply the weighted SH estimator, we must estimate the total population size of Jazz musicians in New York City. It has been estimated that there are 33,003

Jazz musicians in New York City at the time of this study (Heckathorn and Jeffri, 2003), which we used in our estimation procedure. Using this data set and the SH method we estimate $\widehat{C}_{(MF)} = 0.160, \widehat{C}_{(FM)} = 0.579, \widehat{D}_{(F)} = 10.515, \widehat{D}_{(M)} = 14.755$. For the weighted SH method, we estimate $\widehat{C}_{(MF)} = 0.165, \widehat{C}_{(FM)} = 0.631, \widehat{D}_{(F)} = 10.515, \widehat{D}_{(M)} = 14.755$. Since the estimate of average degree for the weighted SH estimate corrects for finite sample size, and the estimated sampling proportion is less than 1%, it is not surprising that the estimated average degree used to compute the SH estimate and the weighted SH estimate are the same. The estimated proportions of male jazz musicians in New York City are 83.52% (SH), 84.31% (Weighted SH), 80.18% (SS), 80.18% (VH), and 74.24% (naive mean).

Figure 4.16 displays the estimated proportions of male jazz musicians and a bootstrap 95% confidence interval Salganik (2006) (from 10,000 bootstrap samples) for all five estimators that we discuss here. In the figure we can see that the naive estimator has both the lowest estimate of the proportion of New York jazz musicians who are male and the narrowest 95% confidence interval. The weighted SH has the highest estimate of the proportion of New York jazz musicians that are male, though none of these estimates are significantly different than each other. Indeed, the weighted SH and SH are very similar, as is the SS and VH estimates. One remarkable aspect of these results is that the 95% confidence intervals of all five of the estimates are so wide.

We also estimated the SS and weighted SH when specifying the underlying population to be 20,000, 30,000, and 40,000 since these two estimators depend on the underlying population size. In this example, both the SS estimators were the same (80.18% for all three different population sizes), as were the weighted SH estimator (84.31). This indicates that the SS and weighted SH were not sensitive to the specified population size.

4.8 Discussion

In this paper we have shown that the assumption that Respondent Driven Sampling results in equal edge sampling probabilities is false when the network has differential activity and finite sampling effects, and that estimators of prevalence will be biased when unequal edge sampling probabilities are unaccounted for. We provide a solution by way of introducing

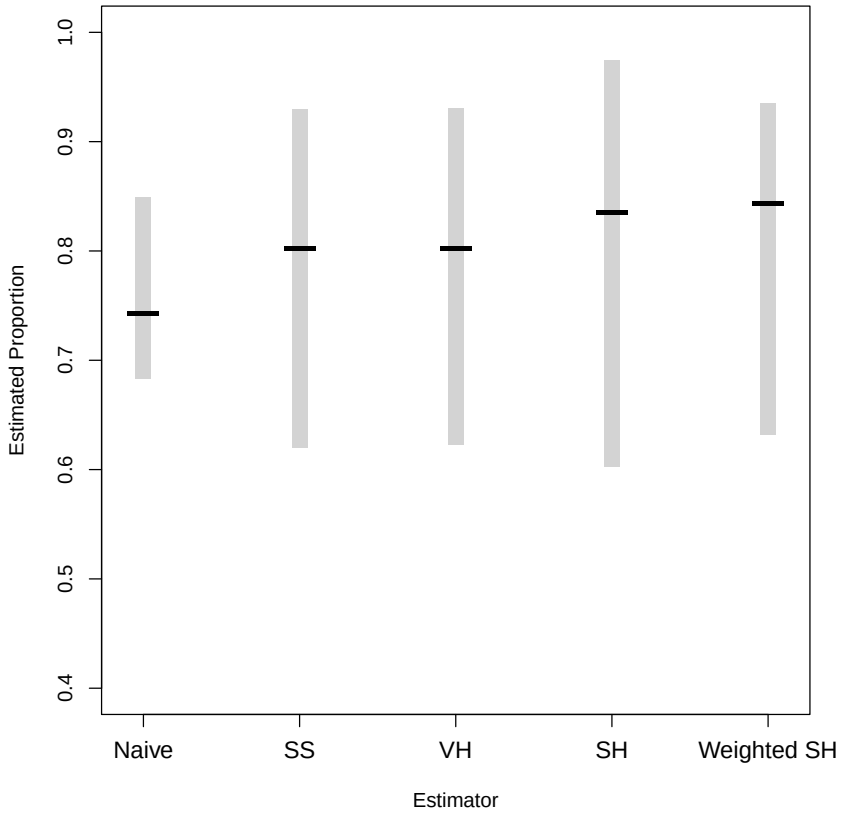

Figure 4.16: Estimated Proportion of New York City Jazz Musician who are Male, with 95% Bootstrap Confidence Intervals

a new prevalence estimator that weights edges inversely to an estimated edge sampling probability. This new prevalence estimator will be particularly useful when the degree distributions vary by the outcome of interest, and also since RDS is often used for prevalence estimation of infectious disease (Lansky et al., 2007; Johnston et al., 2008), and we generally expect those who are infected to have larger degrees than those who are uninfected, we recommend that future analyses of data collected according to an respondent-driven sampling scheme take unequal edge sampling probabilities into account, as not doing so may severely bias prevalence estimates.

While the weighted SH estimator that we propose here improves over the SH in many ways, the SH does not require that the underlying population size is known and takes less time to compute. Like the SH, the weighted SH may be subject to biases introduced by biased seeds. The weighted SH estimator, like the original SH estimator tends to exhibit less bias than estimators that weight the nodal attributes (SS, VH, mean) in the presence of differential recruitment effectiveness. However, as the weighted SH is currently formulated, it assumes that there is no differential recruitment effectiveness when calculating the edge inclusion probabilities. The weighted SH could then be improved by incorporating estimates of differential recruitment effectiveness.

One of the main contributions of this work is to improve upon a widely adopted prevalence estimator for hidden and networked populations. There are many different RDS prevalence estimators that are now in use. It is not apparent which estimator should be used in which setting. Future work should find ways to identify the best estimator to be used in a given setting. We think the new weighted SH estimator is best when there is differential activity, differential recruitment effectiveness, and homophily effects, which is what we would expect to see in a realistic network setting. As such, we recommend the adoption of this new estimator.

More broadly, this work contributes to the understanding of the estimation of edge-wise characteristics when sampling according to the RDS scheme. The assumption that all edges have equal inclusion probabilities also impacts other RDS estimators and diagnostic criteria. For example, homophily is a characteristic of edges that quantifies the tendency for individuals who are in the same group to have relationships with each other as opposed

with individuals not in the same group. Homophily is an edge-wise characteristic, and when homophily on the outcome of interest is strong in the underlying population (people who are HIV positive are much more likely to be associates with others who are HIV positive than HIV negative and people who are HIV negative are much more likely to be associates with others who are HIV negative than those who are HIV positive), RDS prvalence estimators are known to perform poorly (Goel and Salganik, 2009, 2010). RDSAT (Volz et al., 2012) is the most commonly used software for analyzing RDS data sets, and includes a function to estimate population homophily using the RDS sample. However, the RDSAT software assumes that all edges have an equal inclusion probability, and therefore assume that the group mixing observed in the sample can be used as the basis of an unbiased estimator of the population homophily, which may produce biased estimates of population homophily.

We recommend that population homophily is estimated accounting for differing edge inclusion probabilities. Similar to our suggestions for how to address unequal edge inclusion probabilities for the SH estimator, we recommend that estimates of population homophily utilize inverse probability weights. Future work should compare the estimation of homophily which takes into account unequal edge sampling probabilities, as we propose here, and current methods for estimating homophily.

Bibliography

Ali, M. M., Amialchuk, A., and Dwyer, D. S. (2011). The social contagion effect of marijuana use among adolescents. *PLoS One*, 6.

Baer, J. S. and Carney, M. M. (1993). Biases in the perceptions of the consequences of alcohol use among college students. *Journal of Studies on Alcohol*, 54:54–60.

Barnett, N., Ott, M. Q., Rogers, M., Loxley, M., Linkletter, C., and Clark, M. (2013). Peer associations for substance use and exercise in a college student social network. *Health Psychology*, In Press.

Barnett, V. D. (1969). Simultaneous pairwise linear structural relationships. *Biometrics*, 25:129–142.

Beyler, N. K. (2010). Statistical methods for analyzing physical activity data. unpublished doctoral dissertation. *Unpublished Dissertation*.

Blangiardo, M. and Richardson, S. (2008). A bayesian calibration model for combining different pre-processing methods in affymetrix chips. *BMC Bioinformatics*, 9:512.

Borsari, B. and Carey, K. (2003). Descriptive and injunctive norms in college drinking a meta-analytic integration. *Journal of Studies on Alcohol*, 64:331–341.

Borsari, B. and Muellerleile, P. (2009). Collateral reports in the college setting: A meta-analytic integration. *Alcoholism: Clinical and Experimental Research*, 33:826–838.

Brown, P. (1982). Multivariate calibration. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 44:287–321.

Buonaccorsi, J. and Lin, C.-D. (2002). Berkson measurement error in designed repeated measures studies with random coefficients. *Journal of Statistical Planning and Inference*, 104:53–72.

Chow, M. and Thompson, S. K. (2003). A bayesian approach to estimation with link-tracing sampling designs. *Survey Methodology*, 29:197–205.

Conlan, A., Eames, K., Gage, J., von Kirchbach, J., Ross, J., Saenz, R., and Gog, J. (2010). Measuring social networks in british primary schools through scientific engagement. *Proceedings of the Royal Society Series B*, 278:1467–1475.

Connors, G. J. and Maisto, S. A. (2003). Drinking reports from collateral individuals. *Addiction*, 14:Suppl. 2 21–29.

Costenbader, E. and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25:283–307.

Dodd, K. W., Guenther, P. M., Freedman, Laurence, S., Subar, A. F., Kipnis, V., Midthune, D., Tooze, J. A., and Krebs-Smith, S. M. (2006). Statistical methods for estimating usual intake of nutrients and foods: A review of the theory. *Journal of the American Dietetic Association*, 10:1640–1650.

Ferrari, P., Friedenreich, C., and Matthews, C. E. (2007). The role of measurement error in estimating levels of physical activity. *American Journal of Epidemiology*, 166:832–840.

Frank, O. (1977). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4:81–89.

Frank, O. (1981). A survey of statistical methods for graph analysis. *Sociological Methodology*, 12:110–155.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81:832–842.

Fuller, W. A. (1995). Estimation in the presence of measurement error. *International Statistical Review*, 63:121–141.

Gile (2010). Respondent-driven sampling: An assessment of current methodology. *SOCIOLOGICAL METHODOLOGY*, 40:285–327.

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

Gile, K. J. and Handcock, M. S. (2006). Model-based assessment of the impact of missing dat on inference for networks. *CSSS Working Paper 66*.

Gile, K. J. and Handcock, M. S. (2011). Network model-assisted inference from respondent-driven sampling data.

Gimenez, P. and Patat, M. (2005). Estimation in comparative calibration models with replicate measurement,. *Statistics and Probability Letters*, 71:155–164.

Gmel, G., Graham, K., Kuendig, H., and Kuntsche, S. (2006). Measuring alcohol consumption-should the graduated frequency approcah become the norm in survey research? *Addiction*, 101:16–30.

Goel, S. and Salganik, M. J. (2009). Respondent-driven sampling as markov chain monte carlo. *Stat Med*, 28(17):2202–2229.

Goel, S. and Salganik, M. J. (2010). Assessing respondent-driven sampling. *Proc Natl Acad Sci U S A*, 107(15):6743–6747.

Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170.

Goodreau, S. M. (2007). Advances in exponential random graph (p*) models applied to a large social network. *Soc Networks*, 29:231–248.

Hagman, B. T., Patrick, C. R., Noel, N. E., Davis, C. M., and Cramond, A. J. (2007). The utility of collateral informants in substance use research involving college students. *Addictive Behaviors*, 32:2317–2323.

Handcock, M. S. and Gile, K. (2011a). Comment: on the concept of snowball sampling. *Social Methodology*, 41:367–371.

Handcock, M. S. and Gile, K. J. (2011b). On the concept of snowball sampling. *Sociological Methodology*, 41:367–371.

Heckathorn, D. and Jeffri, J. (2003). *Changing the Beat: A Study of the Worklife of Jazz Musicians*, chapter IV Social Networks of Jazz Musicians, pages 48–61. National Endowment for the Arts.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):pp. 174–199.

Heckathorn, D. D. (2002). Respondent-driven sampling ii: Deriving valid population estimates from chain referral samples of hidden populations. *Social Problems*, 49:11–34.

Hoff, P., Fosdick, B., Volfovsky, A., and Stovel, K. (2012). Likelihoods for fixed rank nomination networks. *arXiv:1212.6234 [stat.ME]*.

Hoff, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Origanization Theory*, 15:261–272.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space apapproach to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098.

Holland, P. W. and Leinhardt, S. (1973). The structural implications of meausrement error in sociometry. *Journal of Mathematical Sociology*, 3:85–111.

Horton, N., Roberts, K., Ryan, L., Suglia, S., and Wright, R. (2008). A maximum likelihood latent variable regression model for multiple informants. *Statistics in Medicine*, 27:4992–5004.

Huisman, M. and Steglich, C. (2008). Treatment of non-response in longitudinal network studies. *Social Networks*, 30:297–308.

Johnston, L. G., Malekinejad, M., Kendall, C., Iuppa, I. M., and Rutherford, G. W. (2008). Implementation challenges to using respondent-driven sampling methodology for hiv biological and behavioral surveillance: field experiences in international settings. *AIDS Behav*, 12(4 Suppl):S131–S141.

Kimura, D. (1992). Functional comparative calibration using an em algorithm. *Biometrics*, 48:1263–1271.

Knoke, D. and Yang, S. (2008). *Social Network Analysis*, volume 154 of *Quantitative Applications in the Social Sciences*. Sage Publichations, 2 edition.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data*. Springer.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28:247–268.

Kraus, L. and Augustin, R. (2001). Measuring alcohol consumption and alcohol-related problems: comparison of responses from self-administred questionnaires and telephone interviews. *Addiction*, 96:459–471.

Lansky, A., Abdul-Quader, A. S., Cribbin, M., Hall, T., Finlayson, T. J., Garfein, R. S., Lin, L. S., and Sullivan, P. S. (2007). Developing an hiv behavioral surveillance system for injecting drug users: The national hiv behavioral surveillance system. *Public Health Reports (1974-)*, 122:pp. 48–55.

Lash, T., Thwin, S., Horton, N., Guadagnoli, E., and Silliman, R. (2002). Multiple informants: A new method to assess breast cancer patients' comorbidity. *American Journal of Epidemiology*, 157:249–257.

Lovasz, L. (1993). Random walks on graphs: A survey. *Combinatorics: Paul Erdos is Eighty*, 2:1–46.

Lu, Y., Ye, K., Mathur, A. K., Hui, S., Fuerst, T., and Genant, H. K. (1997). Comparative calibration without a gold standard. *Statistics in Medicine*, 16:1889–905.

Magnani, R., Sabin, K., Saidel, T., and Heckathorn, D. (2005). Review of sampling hard-to-reach and hidden populations for hiv surveillance. *AIDS*, 19 Suppl 2:S67–S72.

Malekinejad, M., Johnston, L. G., Kendall, C., Kerr, L. R. F. S., Rifkin, M. R., and Rutherford, G. W. (2008). Using respondent-driven sampling methodology for hiv biological and

behavioral surveillance in international settings: a systematic review. *AIDS Behav*, 12(4 Suppl):S105–S130.

Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology*, 16:435–463.

McIntyre, J. and Stefanski, L. A. (2011). Regression-assisted deconvolution. *Statistics in Medicine*, 30:1722–1734.

Mercken, L., Snijders, T., Steglich, C., Vartianen, E., and de Vries, H. (2010). Dynamics of adolescent friendship networks and smoking behavior. *Social Networks*, 32:72–81.

Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, 7:308313.

Nusser, S. M., Beyler, N. K., J, W. G., Carriquiry, A. L., Fuller, W. A., and King, B. M. (2012). Modeling errors in physical activity recall data. *Journal of Physical Activity and Health*, 9 (Suppl 1):S56–S67.

Nusser, S. M., Carriquiry, A., Dodd, K. W., and Fuller, W. (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91:1440–1449.

O'Malley, J. A. and Marsden, P. V. (2008). The analysis of social networks. *Health Serv Outcomes Res Methods*, 8:222–269.

Osborne, C. (1991). Statistical calibration: A review. *International Statistical Review*, 59:309–336.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., Ireland, M., Bearinger, L. H., and Udry, J. R. (1997). Protecting adolescents from harm. findings from the national longitudinal study on adolescent health. *JAMA*, 278(10):823–832.

Richardson, S. and Gilks, W. R. (1993). A bayesian aprroach to measurement error problems

in epidemiology using conditional independence models. *American Journal of Epidemiology*, 138:430–442.

Robins, G., Pattison, P., and Woolcock, J. (2004). Missing data in networks: exponential random graph (p*) models for networks with non-respondents. *Social Networks*, 26:257–283.

Salganik, M. J. (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *J Urban Health*, 83(6 Suppl):i98–112.

Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:pp. 193–239.

Semaan, S. (2010). Time-space sampling and respondent-driven sampling with hard-to-reach populations. *Methodological Innovations Online*, 5:60–75.

Snell, L. D., Ramchandani, V. A., Saba, L., Herion, D., Heilig, M., George, D. T., Pridzun, L., Helander, A., Schwandt, M. L., Philips, M. J., Hoffman, P. L., Tabakoff, B., the WHO/ISBRA Study on State, of Alcohol Use, T. M., and Investigators, D. (2011). The biometric measurement of alcohol consumption. *Alcoholism: Clinical and Experimental Research*, Epub ahead of print.

Spiegelman, D., Schneeweiss, S., and McDermott, A. (1997). Measurement error correction for logistic regression models with an "alloyed gold standard". *American Journal of Epidemiology*, 145:184–196.

St Clair, K. and O'Connell, D. (2012). A bayesian model for estimating population means using a link-tracing sampling design. *Biometrics*, 68:165–173.

Stahre, M., Naimi, T., Brewer, R., and Holt, J. (2006). Measuring average alcohol consumption: the impact of including binge drinks in quantity-frequency calculations. *Addiction*, 101:1711–1718.

Stefanski, L. (2000). Measurement error models. *Journal of the American Statistical Association*, 95:1353–1358.

Team, R. C. (2012). R: A language and environment for statistical computing. Technical report, R Foundation for Statistical Computing.

Theobald, C. M. and Mallinson, J. R. (1978). Comparative calibration, linear structural relationships and congeneric measurements. *Biometrics*, 34:39–45.

Thompson, S. K. (2002). *Sampling*. Wiley.

Thompson, S. K. (2006). Adaptive web sampling. *Biometrics*, 62:1224–1234.

Tomas, A. and Gile, K. J. (2011). The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *ELECTRONIC JOURNAL OF STATISTICS*, 5:899–934.

Tooze, J. A., Kipnis, V., Buckman, D. W., Carroll, R. J., Freedman, L. S., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., and Dodd, K. W. (2010). A mixed-effects model approach for estimating the distribution of usual intake of nutrients: The nci method. *Statistics in Medicine*, 29:2857–2868.

Volz, E. and Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24:79–97.

Volz, E., Wejnert, C., Cameron, C., Spiller, M., Barash, V., Degani, I., , and Heckathorn, D. (2012). Respondent-driven sampling analysis tool (rdsat). Version 7.1.

Warnter, R., Kenny, D., and Soto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and*, 37:1742–1757.

Witvliet, M., Brendgen, M., van Lier, P., Koot, H. M., and Vitaro, F. (2010). Early adolescent depressive symptoms: Prediction from clique isolation, loneliness, and perceived social acceptance. *Journal of Abnormal Child Psychology*, 38:1045–1056.

Yan, B. and Gregory, S. (2011). Finding missing edges and communities in incomplete networks. *J. Phys. A: Math. Theor.*, 44:495102–495119.