Title
Modeling Preferential Recruitment for Respondent-Driven Sampling

Author
McLaughlin, Katherine Rumjahn

Publication Date
2016

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Modeling Preferential Recruitment for Respondent-Driven Sampling

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Katherine Rumjahn McLaughlin

2016

Abstract of the Dissertation

Modeling Preferential Recruitment for Respondent-Driven Sampling

by

Katherine Rumjahn McLaughlin

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2016

Professor Mark Stephen Handcock, Chair

Respondent-driven sampling (RDS) is a network sampling methodology used worldwide to sample key populations at high risk for HIV/AIDS who often practice stigmatized/illegal behaviors and are not typically reachable by conventional sampling techniques. In RDS, study participants recruit their peers to enroll, resulting in a sampling mechanism that is unknown to researchers. Current estimators for RDS data require many assumptions about the sampling process, including that recruiters choose people from their network uniformly at random to participate in the study. However, this is likely not true in practice. We believe that people recruit based on observable covariates, such as age, frequency of interaction, geography, socioeconomic status, or social capital.

To model preferential recruitment, I develop a sequential two-sided rational-choice framework, referred to as the RCPR model. At each wave of recruitment,

each recruiter has a utility for selecting each peer, and symmetrically each peer has a utility for being recruited by each recruiter. Each person also has utilities for selecting themself (not recruiting or not participating). People in the network behave in a way that maximizes their utility given the constraints of the network and the restrictions on recruitment. Although a person's utility is not observed, it can be modeled as a linear combination of observable nodal or dyadic covariates plus unobserved pair-specific heterogeneities.

This framework allows generative probabilistic network models to be created for the RDS recruitment process. The models can incorporate observable characteristics of the population and have interpretable parameters. It greatly increases the sophistication of the modeling of the RDS sampling mechanism. Inference can be made about the preference coefficients by maximizing the likelihood of the observed recruitment chain given the observed covariates. As the likelihood is computationally intractable, I develop a Bayesian framework where inference is made feasible by approximating the posterior distribution of the preference coefficients via a Markov chain Monte Carlo algorithm. Each update step samples new values of the preference coefficients and utilities via Metropolis-Hastings, subject to constraints. New prevalence estimates can be calculated be generating many recruitment chains from the population using the RCPR coefficients, then directly obtaining the first-order and second-order inclusion probabilities. This framework allows the incorporation of covariates we think effect recruitment into the sample weights.

The dissertation of Katherine Rumjahn McLaughlin is approved.

Frederic R. Paik Schoenberg

Hongquan Xu

Pamina M. Gorbach

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles

2016

To my friends and family.

Table of Contents

List of Figures

List of Tables

Acknowledgments

It has been a long journey to and through graduate school, and I would like to take time to say thank you to all the people who have inspired, encouraged, challenged, and supported me along the way. First and foremost, I would like to say thank you to the entity that is the Department of Statistics at UCLA: graduate school was not just a thing I did, it was a place I belonged and a family I belonged to. Specifically, the incomparable Glenda Jones and Jason Mesa created a warm atmosphere I wanted to be a part of and were ready to help with any problem I encountered.

I am forever indebted to Mark Handcock for the constant support he provided in navigating the research world and academia, and for his patience as I grew as a statistician. Your boundless intelligence, work ethic, and humility are an inspiration, and something I hope to be able to pass on to my own students. I am extremely thankful to Lisa Johnston for giving me the opportunity to travel and learn more about the data I was working with and the importance of collaboration. And your intelligence, compassion, and independence make you a wonderful role model.

I would also like to thank my main mentors during my undergraduate career at UC Berkeley: Philip Stark, for teaching me to work on interesting problems and for an introduction to high-quality statistics research; Deborah Nolan, for inspiring an interest in sampling; and Piera Cirillo and Barbara Cohn, for giving

me an opportunity when I had no experience.

Beyond these professional mentors, so many people supported me personally as I made my journey through graduate school. Most importantly, I would like to thank my parents, Theresa Rumjahn and Bob McLaughlin, for believing in me no matter what, for always supporting me, and always being in my corner. You gave me all the tools I needed for success and asked for nothing in return. I am lucky to have you as my parents.

I could not have done this without the patience and support of my roommate, Angela Hwang, the most flawless person I know. From giving me time to room dwell to always listening to my problems, and from helping me herd our other friends to watching dog videos for hours on end, you make my life brighter and funnier. Finally, I would like to thank my friends, both statistical and non-statistical, without whom I would surely not have survived graduate school: Aromalyn Magrita, Terri Johnson, Josh Embree, James Molyneux, Patrick McCarthy, Peter Skipper, Jon Arfa, Julia Cross, Gregg Murray, Chun Nickles, Bryon Aragam, Damon Alexander, Alex Whitworth, and Philip Wilcox. Finding people with whom I can share both my work life and my personal life is a rare combination, and I never expected to meet so many such people in graduate school. I treasure my memories with all of you.

To all of the aforementioned people, I give my deepest gratitude.

Vita

2011	B.A. Hons. (Statistics) and B.A. *cum laude* (Classical Civilizations), University of California, Berkeley.
2011-2012	Chancellor's Prize Fellowship, UCLA.
2012-2015	National Science Foundation (NSF) Graduate Student Research Fellowship (GRFP).
2013	Most Promising Applied Statistician Award, Department of Statistics, UCLA.
2013-2014	Secretary, UCLA Math and Physical Sciences Council.
2014-2015	Internal President, UCLA Math and Physical Sciences Council.
2015	Finalist, Jeffry L. Hanson Distinguished Service Award, UCLA.
2016	Joint Statistical Meetings (JSM) Joint SRMS, SSS, and GSS Sections Student Paper Award.
2011-present	Graduate Student Researcher, Statistics Department, UCLA.
2015-present	Dissertation Year Fellowship, Statistics Department, UCLA.

Publications and Presentations

McLaughlin, K.R. Modeling Preferential Recruitment for Respondent-Driven Sampling, Presenter at the Joint Statistics Meetings (JSM) Joint SRMS, SSS, and GSS Student Paper Award Session, Chicago, Illinois, USA, 2016.

McLaughlin, K.R. Modeling Preferential Recruitment for Respondent-Driven Sampling, Presnter at the International Network of Social Network Analysis (INSNA) Sunbelt Conference, Session on Networks in Global Health, Newport Beach, California, USA, 2016.

Johnston, L.G., **K.R. McLaughlin**, H. El Rhilani, A. Latifi, A. Toufik, A. Bennani, K. Alami, B. Elomari, and M.S. Handcock, 2015. A novel method for estimating the size of hidden populations using respondent-driven sampling data: Case examples from Morocco. *Epidemiology*, 26(6):846- 852.

McLaughlin, K.R., M.S. Handcock, and L.G. Johnston, 2015. Inference for the visibility distribution for respondent-driven sampling. In *JSM Proceedings*. Alexandria, VA: American Statistical Association. 2259-2267.

McLaughlin, K.R. Sequential Sampling-Population Size Estimation(SS-PSE), Presenter at 3rd Global HIV Surveillance Consultation, Bangkok, Thailand, 2015.

McLaughlin, K.R. Inference for the Visibility Distribution of RDS, Presenter at the Joint Statistics Meetings (JSM), Session on Society and Networks, Seattle, Washington, USA, 2015.

McLaughlin, K.R. and L.G. Johnston. Data Analysis and Population Size Estimations Using RDS-Analyst, Co-Instructor at 5 day workshop at WHO Collaborating Center for HIV Surveillance, Zagreb, Croatia, 2015.

McLaughlin, K.R. and L.G. Johnston. Interpretation and Analysis of Data from Respondent-Driven Sampling Using RDS-Analyst, Co-Instructor at 4 day workshop, Hanoi, Vietnam, 2014.

Sholtz, R.I., **K.R. McLaughlin**, P.M. Cirillo, M. Petreas, J.S. Park, M.S. Wolff, P. Factor-Litvak, B. Eskenazi, N. Krigbaum, and B.A. Cohn, 2011. Assaying organochlorines in archived serum for a large, long-term cohort: Implications of combining assay results from multiple laboratories over time. *Environment International*, 37(4):709-714.

McLaughlin, K.R. and P.B. Stark, 2011. Workload estimates for risk-limiting audits of large contests. UC Berkeley Undergraduate Honors Thesis, Statistics.

CHAPTER 1

Introduction

Respondent-driven sampling (RDS) has been shown to be a cost-effective, culturally sensitive method to sample from hard-to-reach populations throughout the world (Heckathorn, 1997). These are populations which cannot be reached through traditional probability samples and for which the sampling frames are unknown, so innovative methods are needed (Gile and Handcock, 2010). In particular, RDS is typically used for key populations (KPs) that are at high-risk for HIV/AIDS and related diseases. KPs identified by the World Health Organization (WHO) include people who inject drugs (PWID), female sex workers (FSW), men who have sex with men (MSM), and migrants (Johnston, 2008). These populations share much of the burden of the global HIV/AIDS epidemic. Countries report HIV/AIDS prevalence rates and population size estimates among these KPs to the WHO and UNAIDS from samples conducted using RDS (Gile et al., 2015). These estimates are used to inform policy decisions, budgetary considerations, and the allocation of resources for HIV/AIDS treatment and prevention efforts.

A variety of methods and estimators exist for RDS data. In this chapter, I introduce RDS, discuss the current estimators and history of the methodology,

and provide context for my contribution to the field.

1.1 A Brief Overview of Sampling

Sampling is a method to gather information about a population of size N when querying every member of the population is impossible for practical, financial, or other reasons. In a sample of size $n < N$, a subset of units i is selected by some mechanism such that $i \in \boldsymbol{s}$ means unit i is in the sample and $|\boldsymbol{s}| = n$. The goal of sampling is to make inference about population-level characteristics. In an ideal scenario, the sample is representative of the population so that sample summary measures, such as the mean of a characteristic, equal their population equivalent: $\mu = \bar{x} = \frac{1}{n}\sum_{i\in\boldsymbol{s}} x_i$. Because every sample is different, $\mu = \bar{x}$ is often false even in the ideal scenario, so we instead prefer estimators that are *unbiased*, i.e., $\mathbb{E}(\bar{x}) = \mu$ (Thompson, 2012).

In traditional sampling methods, such as simple random sampling, stratified sampling, or cluster sampling, both the *sampling frame* and the *sampling mechanism* are known to researchers. The sampling frame is a list of members of the population from which the sample is drawn. Note that this may not be precisely the population of interest. The sampling mechanism is the method by which each member of the sample is selected. So, for example, in a simple random sample, members of the sample are chosen with equal probability from the population.

In RDS, both the sampling frame and the sampling mechanism are unknown.

The sampling frame is unknown because of the nature of the populations RDS is typically used for: members are often hidden, and the population size N may be unknown. The sampling mechanism is unknown because study participants, rather than researchers, select people to be in the study. The sampling mechanism for RDS will be discussed in more detail in Section 1.1.1.

Traditional samples assume that members of the sample (*units*) are independent. Network sampling is more complex because members of the population are connected by a network of social relations (*edges*). When the structure of the network informs the sample, the units are no longer independent. One commonly used type of network sampling is *link-tracing*, alternatively referred to as *chain-referral* or *snowball sampling* (Goodman, 1961). In this methodology, an initial sample of *seed* members of the population is selected by some mechanism (often a convenience sample). Researchers ask these seeds about members of their social network in the population, and then follow up ('trace') these links. Each of these people is then asked about members of their social network, and again researchers follow up the links. This process continues over several *waves*. An example is shown in Figure 1.1. Note that the part of the population we observe is centered around the initial seed, so if people who are tied together have attributes more similar to each other than to the whole population, we do not observe a representative sample.

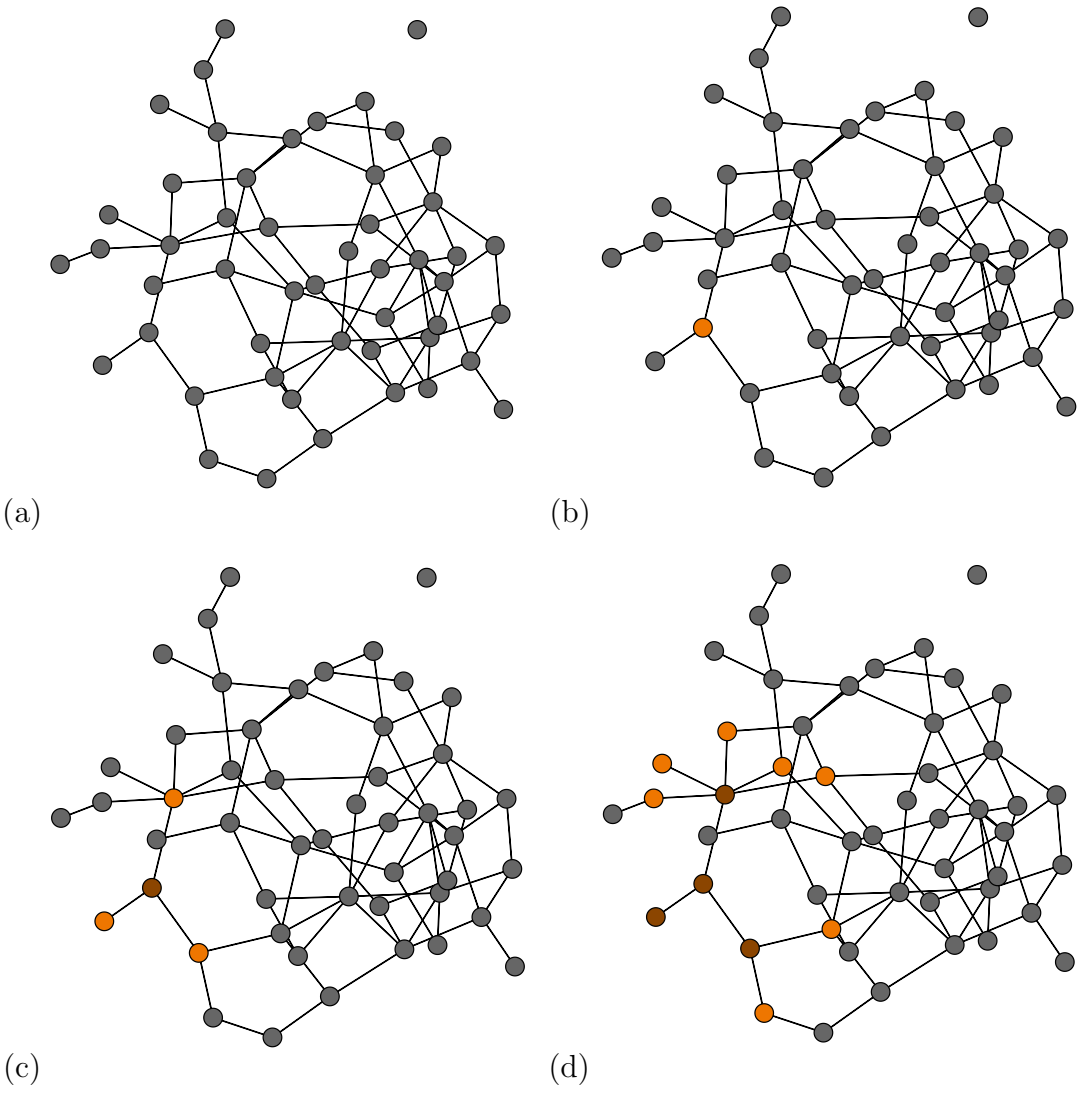


Figure 1.1: Example of a link-tracing network sample showing (a) the population, (b) a seed, (c) one wave of recruitment, and (d) two waves of recruitment.

1.1.1 Respondent-Driven Sampling

Respondent-driven sampling (RDS) is a type of restricted chain-referral sampling that utilizes the underlying *social network* of the population of interest, and relies on participants in the study to recruit their peers (Heckathorn, 1997). RDS proceeds as described above for link-tracing methods, except that instead of researchers asking study participants about their peers and then following up with the peers themselves, these peers' identities remain unknown to the researchers. Because members of the RDS populations frequently studied often practice stigmatized or illegal behaviors, they are reluctant to disclose information about their peers to researchers. RDS maintains privacy and confidentiality by only gathering information about people who voluntarily participate. Therefore, after completing the survey instrument, each seed is given a small number of *coupons* (usually 3) which contain unique identifying information.

They are then instructed to distribute the coupons to members of their social network who meet the study eligibility requirements. Note that this means researchers may observe only part of each person's ego-network and further have no control over which part this is. Peers of the initial seeds who receive coupons can then bring them to the study center and participate themselves. Recruitment is tracked through the unique coupon IDs. Upon participation in the study, the recipients of these coupons form the first wave, and are in turn given their own coupons to distribute. Recruitment continues in this manner for many *waves* until

the desired sample size is attained or no more people participate. People often receive a small primary incentive for participating themselves, and a small secondary incentive for each peer they are able to successfully enroll in the study. As the recruitment chains lengthen over many waves, the composition of the sample begins to reach a point of "equilibrium" where the composition of certain characteristics (e.g., age group, gender, ethnicity, HIV prevalence) within the sample eventually stabilizes, indicating that the final sample is not biased by the purposeful selection of seeds (Heckathorn, 2002).

Population-level inference for RDS studies is difficult for several reasons. Consider the toy example shown in Figure 1.2, with a network of size 16, two seeds, and two waves of recruitment. RDS leads to a *partially-observed network*. The nodes and ties in gray in Figure 1.2(d) are not observed. Therefore, we do not observe all nodes, all edges, or all out-ties from nodes we do observe. Further, some traditional network structures, such as triangles, are impossible to observe in the RDS framework. For example, the tie between nodes 7 and 9 cannot be observed. Ties between nodes in different recruitment chains (e.g., 9 and 11) are also impossible to observe. The only observable network structures are in the recruitment chain, shown for this example in Figure 1.3. Because of the dependence among people in the sample, stemming from the lack of independence among people tied in the underlying social network, the sample is likely not representative of the population. The people we do not observe may be different than the people we do observe. This makes inference for RDS difficult. Many approaches to analyzing

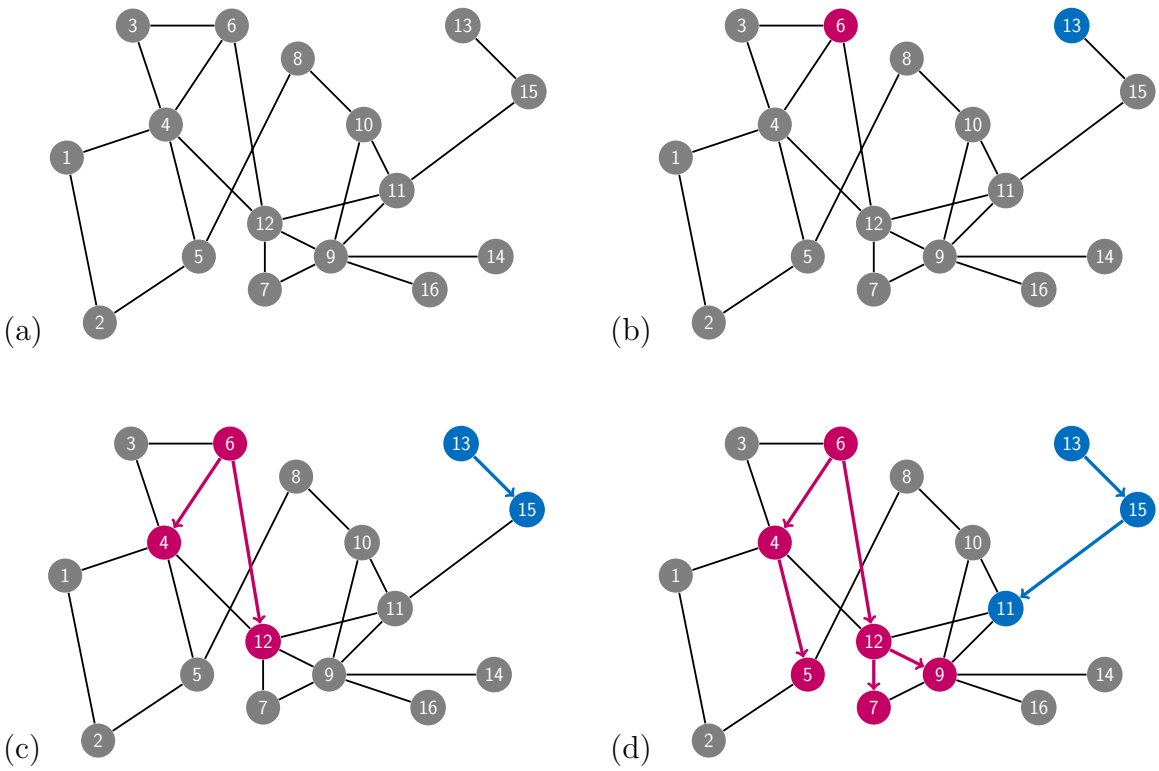


Figure 1.2: Toy RDS example showing (a) the network, (b) selection of two seeds, (c) first wave of recruitment, and (d) second wave of recruitment.

RDS data have been developed, which will be discussed in Sections 1.2 and 1.3.

Further, because study participants rather than researchers control recruitment into the study, RDS almost always has an *unknown sampling mechanism* that is not *ignorable* (Gile and Handcock, 2010). This means that researchers do not control who gets recruited, and do not know the mechanism behind recruitment. In a design-based approach to inference, the lack of knowledge about recruitment decisions, and the impact of this on sample weights, results in a sampling mechanism that is not ignorable. In a model-based approach to inference, the initial convenience sample of seeds results in a sampling mechanism that is not ignorable (Sugden and Smith, 1984). Therefore the sample is a *non-probability*

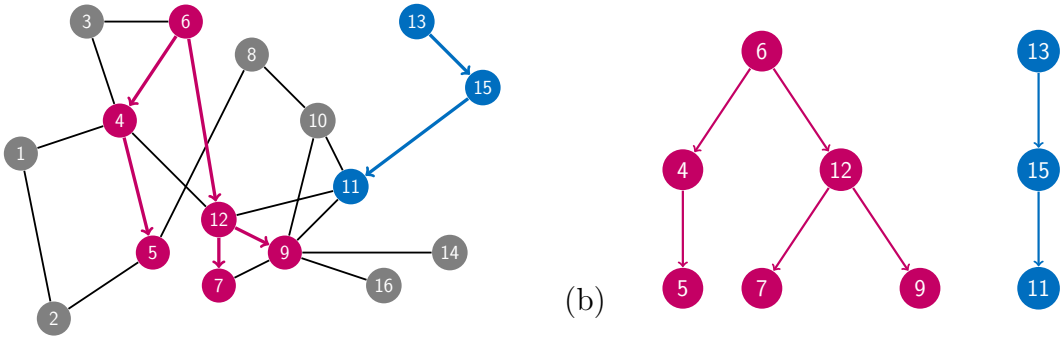


Figure 1.3: Toy RDS example showing (a) recruitment over the underlying network and (b) information observed by researchers during a study, demonstrating that RDS results in a partially-observed sample.

sample and key outcome measures cannot be computed by traditional methods. A variety of estimators exist that attempt to adjust the sampling weights for members of the sample, which will be discussed in Section 1.3. These estimators rely on the assumption that people recruit at random from their personal network. In practice, we do not believe this is true for a variety of reasons. For example, people might prefer to recruit someone they have a stronger tie to (e.g., a close friend rather than an acquaintance), or someone whom they see more frequently for ease of coupon passing. We therefore believe that *preferential recruitment* happens in practice, and more sophisticated models are needed to attempt to understand our data. Evidence for the existence of preferential recruitment is presented in Section 1.5.

1.2 RDS History

Respondent-driven sampling was first used in 1994 to study HIV-related risk behaviors among injecting drug users (IDUs) in the United States (Heckathorn,

1997). RDS was first used outside of the United States for HIV surveillance in 2003 (Wattana et al., 2007). Since then, RDS has been used in hundreds of HIV-related studies in dozens of countries (Malekinejad et al. (2008); Montealegre et al. (2013)). RDS has been adopted by leading public health organizations including the US Centers for Disease Control and Prevention (Lansky et al. (2007); Barbosa Júnior et al. (2011); Wejnert et al. (2012)) and the World Health Organization (Johnston et al., 2013b).

The term "respondent-driven sampling" has proliferated, so attempts have been made to precisely define exactly what constitutes RDS. Importantly, an RDS study must encompass *both* the data gathering protocol and analysis of the data in a way that accounts for the data gathering protocol. Without the second part, a study merely uses peer-referral as a data gathering technique but population-level inference may be poorly posed. Literature reviews have identified several core areas in which studies purporting to be RDS may be methodologically or analytically inconsistent.

First, a careful set of protocols are required for RDS implementation. Studies must (Malekinejad et al., 2008):

1. Initiate recruitment chains with members of the target population, known as seeds;
2. Use a recruitment quota (limit the number of coupons to distribute);
3. Collect data on the size of the social network for all participants using a consistent set of parameters; and
4. Systematically record who recruited whom.

Studies using peer-referral are not necessarily RDS. According to these conditions, reasons of studies not being RDS include: failure of at least one chain to reach three waves of recruitment; not using sample weights; and combining RDS data with data collected using other methods.

(Malekinejad et al., 2008) identify 123 RDS studies conducted from 2003-2007 outside of the United States. For each study, they note the following parameters: principal investigator or contact person or organization; year of study; where the survey was conducted; eligibility criteria; types of biological specimen(s) gathered and laboratory tests performed; whether formative research was conducted prior to the survey; interview method; number of recruitment sites; type of recruitment site; whether mobile recruitment sites were used; whether seeds were diversified (were they selected differently from each other based on key demographic or risk behavior characteristics); total number of seeds used for the study; number of seeds that failed to recruit anyone; whether additional seeds were added after the study began; maximum number of allowable referrals; whether an expiration period was used; primary incentive amount; secondary incentive amount; other services offered during the survey; design effect used to calculate a sample size; desired sample size; actual sample size; maximum number of waves; sampling duration in weeks; whether equilibrium was reported as being reached; whether survey data were adjusted; and description of any operational limitations. In total, during this time period, RDS studies were conducted in 28 different countries and five continents.

A second review, conducted by Montealegre et al. (2013), examined RDS studies in Latin America and the Caribbean between January 1, 2005 and December 31, 2011. The authors found 87 HIV Biological and Behavioral Surveillance Surveys (BBSS) studies that had been conducted using RDS in 15 different countries, using the same eligibility requirements for a study to be considered RDS as Malekinejad et al. (2008).

RDS is widely used in all parts of the world and the plethora of studies have been detailed, analyzed, and discussed well (Malekinejad et al. (2008); Montealegre et al. (2013); and Gile and Handcock (2010)). A similar methodology called peer-driven intervention (PDI) has been less well-studied. I give a more in-depth review of PDI in Chapter 2. I now discuss two particular aspects of RDS pertinent to the discussion that follows of preferential recruitment: what incentivizes people to participate, and social versus sexual networks.

1.2.1 Incentives to Participate

In RDS studies, individuals may choose to participate for a variety of reasons. Fundamentally, RDS relies on the assumption that individuals will be more willing to participate if recruited by their peers rather than a stranger (Heckathorn, 1997). Therefore, the social bonds that they have within the underlying network also form a type of incentive to participate. Other possible incentives are the HIV test that often accompanies the survey questionnaire and knowledge about safe behavior gained as part of participation. Individuals may feel that a member of their social

network could benefit from the HIV test or knowledge, and use this as motivation to recruit them.

In addition to these immaterial incentives, participants in RDS studies often receive economic incentives as well. As noted previously, they often receive a small primary incentive for visiting the study location and completing the survey instrument. Individuals who have participated in the study themselves are then encouraged to recruit their peers using secondary incentives. These incentives should be of at least slight value to participants, but not substantial enough to promote risky behaviors or lead to problems with study ethics.

For example, in PWID populations, there is concern that monetary incentives may be spent directly on drugs. Additionally, incentives that are too high may result in coercion or lying among the target population. This may occur when individuals hear about the study and its high incentive and pressure a friend, acquaintance, or even stranger into giving them a coupon. Individuals may also lie and participate twice in the study to double their incentive reward, or lie and pretend to be a member of the study population when they are not in actuality. To gain the full secondary incentive, recruiters may coerce their peers to participate. Individuals may also attempt to sell their coupons. In a controversial study that has garnered several rebuttals, Scott (2008) notes that the RDS incentive structure could be abused by participants seeking to game the system for their own gain. In particular, he notes that RDS can "foment a stratified market of research participation that reinforces pre-existing economic and social inequalities."

Further, in his study of injection drug users (IDUs) in Chicago, "participants co-opted RDS to develop various "underground" revenue-generating modalities that produced differential risks and benefits among participants. Deleterious outcomes include false advertising regarding the study's risks and benefits, exploitation of relative economic deprivation, generation of sero-discordant social networks, and interpersonal and organised conflict, coercion, and violence" (Scott, 2008).

Although these issues are not common, they have been reported in some studies and raise substantive ethical concerns. The formative research performed before RDS recruitment begins often attempts to address potential issues with incentive structure. A collaboration between researchers, local stakeholders, and members of the target population in focus groups can determine appropriate incentives for the study.

For these reasons, alternatives to monetary incentives have been explored in many populations, such as gifts (e.g., a headscarf, keychain, soap, or school books), vouchers for food or groceries, or telephone cards (Johnston, 2008). Rather than "incentives," these items are often referred to as "tokens of appreciation" by the staff (McCreesh et al., 2013). They are therefore remuneration that shows respect for participants' time and effort in participation and recruitment, and for the transportation costs associated with reaching the study location.

The amount of the incentives will vary based on a number of factors, including the study budget, standard of living in the study country, governmental policies, and population of interest. Incentives should not be so low such that some

segments of the population will not participate, nor so high that they result in sold coupons or coercive behavior. Considerable literature exists on the use of payments as compensation for research in public health studies in general and in HIV-related studies in particular (Semaan et al., 2009). Authors generally conclude that incentives are allowable so long as they are not so high as to coerce low-income participants into consenting to the study.

Nevertheless, differential participation based on socioeconomic status caused by the incentives is a potential problem in RDS studies of many different types of populations. Therefore, when considering an individual's likelihood to be in the sample, we may want to take into account their socioeconomic status. Current RDS weighting schemes do not allow for this option. With this in mind, the RCPR model I introduce in Chapter 3 allows for other nodal or dyadic covariates, in additional to personal network size, to be used when considering recruitment tendencies. Many studies attempt to measure socioeconomic status, as it is a key demographic variable, so it is easy to incorporate. Even if a measure of socioeconomic status is not be available, it is possible to model it as a latent variable. By having the option to incorporate socioeconomic status into an individual's likelihood to be in the sample, the RCPR framework more accurately reflects recruitment realities.

1.2.2 Social vs. Sexual Networks

RDS relies on the assumption that members of the target population are connected by an underlying network of social ties. This social network may be composed of different types of ties, including acquaintances, close friends, and sexual partners, among others. Different populations may have different proportions of these types of ties. For example, women of FSW populations are likely not tied sexually to each other but may instead know each other through their work location. Men in MSM populations would be less likely to work together, but more likely to be tied sexually. This distinction matters because of the nature of HIV transmission. Although certainly particularly high-risk FSW could be clustered together within the network via their workplace, an HIV-positive FSW is less likely to directly transmit the disease to a peer than an HIV-positive MSM. This could result in clustering of HIV status in the recruitment chains observed in the sample, leading to either overestimation or underestimation of HIV prevalence. In addition, if people are more likely to recruit a sexual tie than a social tie, we may observe an over-representation of highly sexually active individuals. If this group also has an elevated HIV prevalence, our final estimate may be too high.

The type of network considered in a study is defined by how the network size question(s) in the survey instrument are posed. For example, even in networks with a high number of sexual ties, passing a coupon to a non-sexual tie is not precluded. The only requirements are that the coupon must be passed to someone

who themself meets the study eligibility requirements. Nevertheless, the likelihood that these peers will be sexually tied to the recruiter will vary from study to study. The worry for MSM populations is that members will tend to pass coupons to longer-term partners rather than shorter-term ones, so the portion of a person's sexual network we observe is not representative. This is an example of preferential recruitment on the type of dyad formed by the recruiter and peer. To gather information about dyad type, it is straightforward to ask participants as part of the survey questionnaire what their relationship is to the person who recruited them. Current RDS weighting schemes do not take into consideration the relationship between the recruiter and peer. With this in mind, the RCPR model I introduce in Chapter 3 allows for the incorporation of relationship type as a dyadic covariate. This will help account for the differences in recruitment behavior in social and sexual networks, and how this may impact HIV prevalence estimates.

1.3 Current RDS Estimators

A number of contributions to the RDS literature have involved the development of new estimators for population-level prevalence of the outcome measure (usually HIV). The foundational paper (Heckathorn, 1997) relied on the naïve estimator (i.e., the sample mean), but it quickly became clear that samples collected using RDS were often not representative of the larger population. Because the sampling mechanism is unknown for RDS, exact sample weights cannot be computed.

The new estimators calculate sample weights in different ways. Several common estimators are detailed and assessed below.

The outcome of interest is the population prevalence μ of a characteristic A, such as HIV. Let $A^C = B$. μ is unknown, so we use information available in the sample to find the estimate $\hat{\mu}$. Let s be the observed sample, where $i \in s$ indicates that individual i was in the observed sample. Let y_i be the response of individual i such that $y_i = 1$ if $i \in A$ and $y_i = 0$ if $i \in B$. Let the self-reported degree of individual i be $\tilde{d}_i$. Let the true inclusion probability of person i in the sample be π_i.

1.3.1 The Salganik-Heckathorn Estimator

The Salganik-Heckathorn estimator (Salganik and Heckathorn, 2004) uses the estimated number of cross-group memberships to adjust for the RDS sampling process. The estimator is

$$\hat{\mu}_{\text{SH}} = \frac{\widehat{C_{BA}}}{\widehat{C_{BA}} + \widehat{C_{AB}}\left(\frac{\widehat{D_B}}{\widehat{D_A}}\right)} \tag{1.1}$$

where $\widehat{C_{AB}}$ is the proportion of all individuals recruited by members of group A who are members of group B. $\widehat{D_A}$ is an estimate of the mean degree of individuals who are part of group A, given by

$$\widehat{D_A} = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} \frac{1}{\tilde{d}_i}}. \tag{1.2}$$

If the mean degree of members of group A is the same as members of group B, then $\hat{\mu}$ is simply the number of people in group A recruited by those in group B, divided by the total number of recruits.

This estimator relies on self-reported values of degree, which are known to be problematic (McLaughlin et al., 2015). Further, if degree mis-reporting is associated with the outcome variable, this estimator will be inaccurate. This can be the case when, for example, people who are HIV positive attempt to minimize their membership in the population of interest by underreporting their number of peers. Therefore, letting A be the group of people who are HIV positive, since people in A in general will have smaller network sizes, the fraction $\frac{\widehat{\overline{D_B}}}{\widehat{\overline{D_A}}}$ will be larger than the 'true' ratio, resulting in the overall estimate $\hat{\mu}_{SH}$ being too small.

1.3.2 The Volz-Heckathorn Estimator

The Volz-Heckathorn estimator (Volz and Heckathorn, 2008) uses the self-reported degree of an individual to weight their response. The general assumption behind this estimator is that people who have higher degree are more likely to be selected to participate in the study, and thus their responses need to be downweighted. The estimator is

$$\hat{\mu}_{\text{VH}} = \frac{\sum_{i \in s} \frac{y_i}{\tilde{d}_i}}{\sum_{i \in s} \frac{1}{\tilde{d}_i}}. \tag{1.3}$$

This is a generalized Hansen-Hurwitz estimator. It is asymptotically unbiased for μ if $\pi_i \propto \tilde{d}_i$ for all i under the assumption of infinite population size.

	Network structure assumptions	**Sampling assumptions**
Random walk model	Network size large ($N \gg n$)	With-replacement sampling Single non-branching chain
Remove seed dependence	Homophily sufficiently weak Bottlenecks limited Connected graph	Enough sample waves
Respondent behavior	All ties reciprocated	Degree accurately measured Random referral

Table 1.1: RDS assumptions for the Volz-Heckathorn estimator. Adapted from Gile (2011).

This estimator again relies on accurately self-reported degree values. Additionally, in some cases, vastly different magnitudes of degree (e.g., 2 and 700) result in individuals being afforded very different sample weights when their reality of inclusion is likely not so drastically different.

Assumptions for the commonly used Volz-Heckathorn estimator are shown in Table 1.3.2. The two-sided rational-choice preferential recruitment model I propose in Chapter 3 relaxes the following assumptions: homophily sufficiently weak; with-replacement sampling; single non-branching chain; degree accurately measured; random referral.

The sample weights in the Volz-Heckathorn estimator are obtained via the assumption that sampling is a random walk over the underlying social network.

1.3.2.1 Random Walk Example

Let $\mathbf{A}$ be the $N \times N$ sociomatrix of ties in the underlying social network, such that $A_{ij} = 1$ if individuals i and j are tied, and $A_{ij} = 0$ if they are not. Assume

that $A_{ij} = A_{ji}$ and $A_{ii} = 0$.

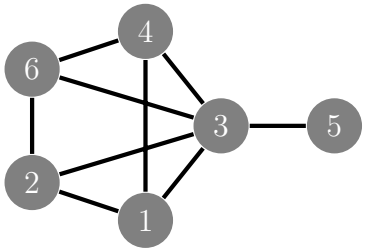
Consider a random walk process on the graph given by $\mathbf{A}$. This process is defined as: Let the vector G represent the successive indices of nodes sampled by the random walk process such that G_k is the index of the node sampled at the kth step. Then the random walk is a Markov process on the space of nodal indices characterized by the transition matrix $\mathbf{T}$. The ijth element of $\mathbf{T}$ is:

$$P(G_{k+1} = j \,|\, G_k = i) = T_{ij} = \begin{cases} \frac{1}{d_i} & y_{ij} = 1 \\ 0 & y_{ij} = 0 \end{cases} \tag{1.4}$$

where $d_i = \sum_{j=1}^{N} A_{ij}$ is the degree of individual i. This constitutes random referral, than is, person i selects person j to give their coupon to *uniformly at random* from their social network. Then if the graph $\mathbf{A}$ is a *connected graph*, this process constitutes an irreducible Markov chain on the space of the nodal indices, characterized by the transition matrix $\mathbf{T}$. The draw-wise selection probabilities p_i are proportional to degree, $p_i = \alpha d_i$ for some constant α. This corresponds to a recruitment procedure with only one seed, where each person can distribute only one coupon with replacement (i.e., a single, non-branching chain).

It is assumed that the stationary distribution of the Markov chain corresponds to the probability distribution of the location of a coupon at a given time after a large number of steps[1].

[1]In practice, it is assumed that the Markov chain has reached equilibrium during the RDS study. This assumption is made so that the initial bias based on the convenience sample used to select the seeds is eradicated. As most RDS studies contain relatively few waves (often < 10),



$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \end{pmatrix}$$

Figure 1.4: Random walk example: graph and adjacency matrix.

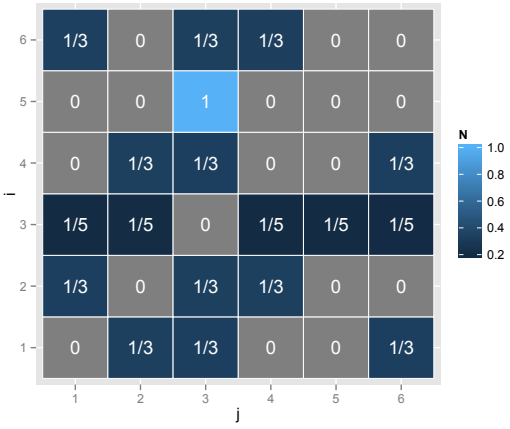


Figure 1.5: Random walk example: transition matrix.

An example graph and adjacency matrix $\mathbf{A}$ are shown in Figure 1.4. Based on the random walk model, the transition matrix $\mathbf{T}$ is shown in Figure 1.5. The sampling probabilities of the nodes after (k) draws from the Markov chain can then be calculated as $\mathbf{T}^{(k)}x^{(0)}$, where $x^{(0)}$ represents the starting point (which node was the seed).

In an RDS sample, the underlying social network is not known. Instead, it is assumed that the nodal inclusion probability π_i of node i is proportional to that individual's self-reported degree $\tilde{d}_i$. These sampling weights are derived from a single non-branching, with replacement, random walk process at equilibrium.

this is a questionable assumption.

However, this is a known simplification that does not mirror the reality of the sampling process. In actuality, the RDS sampling process contains multiple chains, is branching, without replacement, and does not begin at, or even necessarily converge to a fixed equilibrium.

1.3.3 The Successive Sampling Estimator

The Successive Sampling estimator (Gile, 2011) relaxes the infinite population assumption required by the Volz-Heckathorn estimator. In practice, sampling proceeds without replacement (i.e., a person cannot participate twice) from a finite population.

The estimator is

$$\hat{\mu}_{\text{SS}} = \frac{\sum_{i \in s} \frac{y_i}{\tilde{\pi}_i}}{\sum_{i \in s} \frac{1}{\tilde{\pi}_i}} \tag{1.5}$$

where $\tilde{\pi}_i$ is the estimated sampling probability of individual i. This is determined using the successive sampling procedure (Gile, 2011). This estimator also requires knowing the population size, N. The pitfalls based on mis-reported degree that are problematic for the Volz-Heckathorn estimator are also troublesome for the Successive Sampling estimator.

1.4 Homophily

RDS estimators commonly assume that no preferential recruitment occurs; that is, that people recruit members of their social network uniformly at random. To

assess the validity of this claim, I introduce the notion of *recruitment homophily*. Homophily is a concept commonly used in sociology and network science (Mcpherson et al., 2001) to describe the propensity for people in a network to be tied to others like them. Homophily has different definitions in different scenarios. Here, I consider recruitment homophily $\mathcal{H}_R$ on a categorical covariate. For a covariate with n_l levels, the recruitment homophily is

$$\mathcal{H}_R := \frac{\sum_{i=1}^{n_l} o_{ii}}{\frac{1}{n_d}\sum_{i=1}^{n_l} o_{i\cdot}o_{\cdot i}} \tag{1.6}$$

where o_{ij} is the observed cell count for dyads between recruiters of category i ($i = 1, \ldots, n_l$) and peers of category j ($j = 1, \ldots, n_l$). $o_{i\cdot} = \sum_{j=1}^{n_l} o_{ij}$ and $o_{\cdot i} = \sum_{j=1}^{n_l} o_{ji}$ are the marginal totals for, respectively, dyads from a recruiter of type i to any peer and dyads from any type of recruiter to a peer of type i. $\mathcal{H}_R$ is a ratio of the observed count of matching dyads (both the recruiter and the peer have the same covariate value) to the expected count under the assumption that no preferential recruitment occurs. Therefore, $\mathcal{H}_R > 1$ indicates homophily on the covariate, $\mathcal{H}_R = 0$ indicates no preferential recruitment, and $\mathcal{H}_R < 1$ indicates *heterophily*. Heterophily is the propensity for people to form ties with those different from themselves.

1.5 Evidence of Preferential Recruitment

The commonly used Volz-Heckathorn and Successive Sampling estimators described in Sections 1.3.2 and 1.3.3 rely on the assumption that people distribute their coupons randomly to members of their social network. However, there is evidence that this is not true in practice and that *preferential recruitment* does occur.

Several studies have also attempted to discern the recruitment decisions of participants by asking additional questions after the completion of the study. Gile et al. (2015) consider 12 parallel RDS studies conducted in the Dominican Republic in 2008 where respondents were encouraged to set an appointment to return 2 weeks later for a follow-up visit. With the caveat that 43% of participants made a follow-up visit, and that those people tended to be those who recruited more often and participated earlier in the study, they examine recruiter behavior. First, when asked "How many people did you try to give a coupon but they had already participated in the study?" they found wide variance based on study site. However in 6 of the 12 sites, more than 25% of respondents said they attempted to distribute a coupon to someone who had already participated in the study. In these cases, recruiting decisions were affected by earlier parts of the sample. Regarding specific variables on which preferential recruitment occurred, they found that recruiters preferred to select peers who were employed.

Recruitment decisions may be more complex than generally assumed in stan-

dard RDS models (Bengtsson and Thorson (2010); Kerr et al. (2011); McCreesh et al. (2012)). For example, a study of MSM in Brazil found that some people tended to recruit their riskiest friends because they were thought to need safe sex counseling (de Mello et al., 2008). A study of PWID in Hartford, Connecticut found that people had a tendency to recruit men over women, people who used crack over those who did not, and people who were not homeless over those who were (Crawford et al., 2015). Wejnert and Heckathorn (2008) found that men and Asian Americans demonstrated preference for recruiting peers of their own gender and racial group in a web-based RDS study of college students.

Iguchi et al. (2009) considered a dual high-risk group sampling approach that relied on peer recruitment for a combined, overlapping sample of MSM and DU (drug users). They found that in St. Petersburg, Russia, homophily was very high among both MSM and DU recruits, with MSM demonstrating an almost complete preference for recruiting other MSM. When out-group recruitment occurred, it almost always was the recruitment of MSM/DU and rarely included heterosexual DU. Further, they used DU respondents' reports on the gender composition of their entire networks to calculate what recruitments would have been had participants recruited in a manner consistent with their self-reported network composition. Comparison of these expected recruitments with actual recruitments revealed that both men and women over-recruited peers of their own gender leading to upward-biased Salganik-Heckathorn estimates of group proportions.

McCreesh et al. (2012) examined a RDS study of a non-hidden population of

Ugandan villagers and found evidence of preferential recruitment by age, tribe, socioeconomic status, village, and sexual activity. Yamanis et al. (2013) found that for FSWs in Shanghai, the RDS referral process steered recruitment away from high-tier sex workers who are less vulnerable than middle-/low-tier sex workers to risky behaviors. This biased the sample and proportion estimates away from low-risk segments in favor of high-risk segments of the population.

Each different study context will have a different set of variables (if any) upon which preferential recruitment occurs. However it is clear that more sophisticated estimators that allow for adjustments based on covariates people may recruit preferentially on is needed. Tomas and Gile (2011) note that "there is no evidence that any of the estimators are adjusting for differential recruitment."

I introduce a two-sided rational-choice preferential recruitment (RCPR) framework to relax many of the assumptions about the sampling process that current RDS estimators make. First, I provide a brief introduction to and a literature review of peer-driven interventions, which build on the RDS framework, and to which the RCPR framework can be applied in Chapter 2. Next, I introduce the RCPR model in Chapter 3, and assess its performance using a variety of simulation studies in Chapter 4. I apply the model to RDS studies of hidden populations at high risk for HIV/AIDS in Chapter 5.

CHAPTER 2

Peer-Driven Interventions

Peer-driven interventions build on the RDS framework. As originally described by Robert S. Broadhead and Douglas D. Heckathorn, a Peer-Driven Intervention (PDI) is a "chain-referral" outreach methodology. It was developed in the 1990s as an attempt to prevent the spread of HIV and related diseases, and was first applied to populations of injection drug users (IDUs). The RDS framework is usually intended to estimate population averages of outcome measures, while PDIs allow researchers to attempt to affect change on the population of interest through an intervention. In addition to recruiting peers into a study through the use of coupons as in RDS, recruiters in a PDI are also asked to pass along an "intervention" message, usually targeted toward increasing knowledge of safe behavior practices. To test the effectiveness of the information transmission, study participants take a short knowledge test as part of the survey instrument. A thorough description of the PDI methodology will be given in Section 2.1.

Many advances have been made in RDS analysis focused around computing sample weights with the goal of calculating population-level prevalences (e.g., HIV prevalence). In a PDI, the goal is often to compare two groups of people, those

who received an "intervention" and those who did not, and determine if the intervention had an effect on some outcome measure. Unlike a traditional randomized control trial (RCT), in a PDI we cannot assume that people were randomly assigned to either the treatment or the control. The underlying network structure, recruitment structure, homophily on characteristics affecting the outcome measure, and potential for spillover between groups all complicate analysis.

PDIs were developed as an alternative to traditional "provider-client" outreach models,[1] which rely on a staff of salaried outreach workers. In contrast, the PDI model relies on the premise that people who are part of the experiment are much more capable of reaching and communicating with one another on matters of mutual concern than are salaried "peer-educators," outreach workers, or professionals.

Based on sociological underpinnings, Broadhead and Heckathorn designed PDIs to tap into six critical elements of behavior change: (1) knowledge, (2) skill building, (3) motivation, (4) peer influence, (5) social norms, and (6) repetition (Heckathorn and Broadhead, 1996). Individuals participate in facilitated intervention activities targeting critical mediators of behavior change, and then independently educate peers on selected core messages, for which compensation is provided. It is hypothesized that through peer education an individual's own commitment to engage in targeted outcome behavior is strengthened because the

[1]This class of model is also referred to as "traditional outreach," "traditional outreach intervention" (TOI), the "NIDA Standard model," the "Indigenous Leader" model, the "peer-educator model," and the "Community Health Outreach Worker" (CHOW) model.

act of educating peers is a public affirmation of the outcome behavior. Ideally, through successive waves of recruitment and peer education, network social norms are altered.

Since the original studies involving peer-driven interventions, the methodology has been used to study many different populations. Most studies focus on determining the feasibility or efficacy of PDIs and therefore compare the desired outcome under the PDI model to the same outcome when a different type of intervention, usually the traditional "provider-client" model, is used. Although these studies are essential in establishing PDI as an effective, cost-efficient, and practical methodology, they do not provide opportunities to develop more complex analyses of PDI data that can be obtained when randomization into treatment/intervention and control/placebo groups is internal to the study, or when a statistical model is employed to understand the underlying processes of recruitment and behavior change. Notably, the traditional "provider-client" studies that are frequently used as control groups often occur at different times (often, several years prior to the PDI) or in different, not necessarily comparable cities, with seeds drawn from a different population. Even when careful design-based decisions are made to create identical scenarios[2], the demographic and endogenous network characteristics of these populations are subject to natural variation on account of both time and place. For example, sometimes the choice drug of IDUs is different, or one area is more urban, or one group has on average higher incomes (Broadhead et al., 2006).

[2]For example, placing the health clinic in the same type of building in a similar neighborhood and choosing health educators with similar backgrounds.

These types of comparisons, which often offer careful design-based choices followed by a simple statistical t-test, establish PDI as a valid methodology in comparison to the traditional "provider-client" outreach models for increased knowledge or behavior change. However, we wish to make many other types of comparisons. For example, in the previously discussed comparison, there is no true "placebo" and it is hard to assess the true impact of the intervention because accurate population sizes are usually hard to gather for the groups of interest. The different profiles of the populations used for comparison make it difficult to attribute change solely to the intervention.

Using a t-test to consider changes in an outcome measure may miss part of the bigger picture. For example, looking at changes in drug injection incidents helps us determine the effect of the intervention on this measure, but does not address the underlying mechanisms that determine how effective the intervention actually is. We may know on average that it worked, but we still do not understand how the intervention was transmitted through the population. As an illustration, consider an intervention aimed at smoking cessation. It is possible that overall the intervention significantly reduced the number of cigarettes smoked per person per day, but that behavior change actually only occurred if the intervention message was given by a family member instead of a friend or acquaintance. The t-test would miss this last fact, but a model examining the underlying process of the PDI would capture it. As researchers, we would gain valuable information from this, and tailor future PDIs either to target family members or to explore

different intervention strategies to influence friends and acquaintances. Therefore, a statistical modeling approach to PDI analysis can provide more nuanced and policy-relevant information. The RCPR model I present in Chapter 3 allows more careful examination of covariates that may be associated with the efficacy of the intervention.

Section 2.1 provides a more precise definition of all PDI components. Using these definitions, a review of five different variants of PDIs that have been implemented throughout the world is given in Section 2.2.

2.1 A Formal Definition of Peer-Driven Intervention

Many studies promote peer-support as an effective method to affect behavior change. However a peer-driven intervention is a very specific study design easily incorporated into the framework of respondent-driven sampling (RDS) and penetrating hard-to-reach populations. In creating this definition, precise language is important.

Linguistically, note that herein a *seed* refers to an individual recruited by a health official or other non-population individual, who initiates a recruiting chain. A *peer* is any person in the population of interest who is eligible to be recruited. A *recruiter* is someone who has a coupon to give out to an eligible peer. A *participant* is any person who took part in the study. This group comprises the initial seeds and all recruits who chose to participate.

PDIs have several essential components, which separate them from other forms of intervention that rely on peer helpers:

1. Initial seeds are reached by any means possible, often a sample of convenience.
2. Recruitment is limited to a small subset of a person's network by the use of unique coupons (usually 3 to 5 per subject).
3. The initial seeds recruit a set of their peers and educate them on a predetermined set of knowledge (usually 8 to 10 pieces of information), with the aid of a "crib card" they are given. These recruits in turn recruit their own peers.
4. Recruitment continues for either a set number of *waves* w, until a desired sample size n is reached, or until no more participants recruit.
5. Once a person is recruited, if they choose to participate, they visit a community outreach center where they are screened for eligibility and take a knowledge test on the information their recruiter passed on.
6. Participants complete the survey instrument and other activities, such as a blood draw.
7. Participants receive training as part of an "intervention session" that delivers the core intervention message and provides information on how to recruit and educate peers.
8. Primary incentives are given for one's own participation in screening, questionnaires, and intervention sessions. Secondary incentives are given for successfully recruiting peers and for peers' performance on the knowledge test.

These components are shown schematically in Figure 2.1, with the numbers representing distinct time points. Examples provided in the later sections will demonstrate these essential PDI components, but also the variability and flexibility of the framework when implemented in different manners.

A respondent-driven sample would proceed in the same manner, except without the education, knowledge test, or intervention session activities.

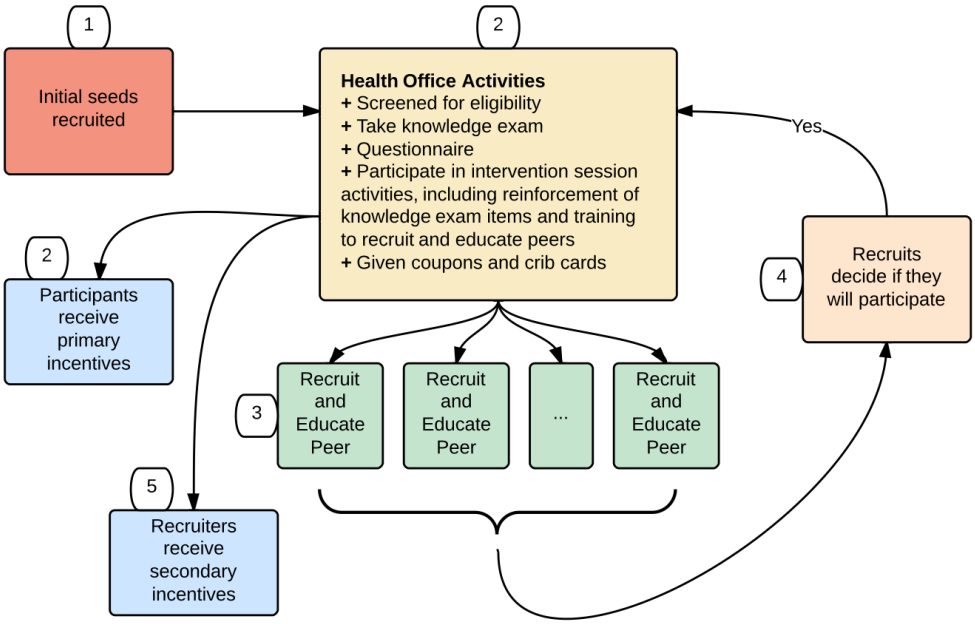


Figure 2.1: Schematic diagram of a peer-driven intervention. The numbers (1)-(5) refer to distinct time points. Primary and secondary incentives are shown in blue boxes.

There are many fine points of PDI implementation which will not be discussed at length here. For a more in-depth discussion, refer to Broadhead (2009); Broadhead et al. (1998); and Broadhead et al. (2009). Two points are worth nothing, however. First, the "crib cards" that the recruiters use when educating their peers have cue-words that are lost on anyone who has not undergone the intervention training. For example, in a PDI of IDUs in Vietnam and China (Broadhead et al., 2009), one cue-word was "Outward Circling Motion." The meaning of this phrase is not immediately apparent, but those who had undergone the intervention training would be reminded that it was referring to the proper way to use an alcohol wipe in preparing an injection site. These cue-words on the crib cards force the recruiter to educate their peer if they wish for them to do well on the knowledge test; they cannot simply give them reading material and instruct them to study because the meaning of the phrases will be lost on a recruit without instruction by their recruiter.

Second, although by definition the incentives are created to mold behaviors in the recruiters and participants, they are not substantial enough to alter or create risk-behavior for even one day. Some studies prefer to use non-cash monetary incentives, which may take the form of, for example, gift cards or pre-paid phone cards. In particular, in studies targeting IDUs, researchers need to be careful ethically that any money provided to participants will not negatively impact their drug-use habits. Another concern if incentives are too high is that recruiters will bully or coerce their recruits into participating and scoring highly on the

knowledge test. Broadhead et al. (2009) note that they kept rewards nominal enough not to induce these behaviors. The incentives are sometimes seen, then, as acknowledgements of participation and effort that are memorable to participants, even if they are not large sums of money.

2.2 Literature Review and Current Statistical Analysis

Peer-driven interventions have been carried out in many populations throughout the world, including Russia, Latvia, Lithuania, Ukraine, China, Vietnam, and the United States. Characteristics of a demonstrative selection of these PDIs are discussed in more detail in the sections that follow.

2.2.1 Peer-Driven Interventions without a Direct Comparison

Some PDIs are implemented without a comparison group, making it difficult to access the efficacy of the intervention. Because it is difficult to understand the impact of the intervention, most of the information gleaned from these comparisons is similar to what would result from a respondent-driven sample, and there is still merit to these studies.

A minimal PDI was carried out in El Paso, Texas and Chaparral, New Mexico to increase Latinas' access to and utilization of HIV prevention and testing services (Ramos et al., 2009). This study only considered Latinas at moderate to high risk of HIV infection, and only had one wave of recruitment. The initial seeds

were given 5 coupons and received $5 for each successful referral. The 56 seeds were able to recruit 61 peers, for a total sample size of 117. The analysis of this study was mainly anecdotal, but the authors did note that initially only 31/56 seeds had information about HIV prevention, but by the end all seeds and all recruits had been given this information. However, without a direct comparison to another population or within the same population, it is difficult to quantify the efficacy of a PDI. Further, without the knowledge test, it is difficult to determine to what extent the participants were internalizing the information they were presented with, rather than just being exposed to it.

2.2.2 Comparisons of Peer-Driven Interventions to Traditional Outreach Methods

Many early PDIs, and PDIs used as pilot studies to establish the efficacy of the methodology in a new population, focused on comparisons with traditional outreach methods. These studies often attempt to establish two things: first, that the PDI is effective as a recruiting strategy; and second, that the intervention itself is effective. The first comparison is the same as one that would be made in RDS pilot studies. The second is unique to PDIs but is sometimes harder to quantify because the outcome measure may not have been measured in a similar population under a TOI.

A typical example of a PDI to TOI comparison is that by Broadhead et al. (1998). This study compared injection drug users (IDUs) in two medium-sized

towns in eastern and central Connecticut. A PDI was conducted in New London, and then moved to Middletown after a year for logistic reasons, while a TOI was conducted in Windham. For primary incentives, participants earned $20 for their initial interview, and $30 for a follow-up interview. Additionally, as secondary incentives, participants were given $10 per recruit, up to a maximum of three. In addition to this, they were given up to $10 per recruit for the recruit's performance on the knowledge test. Therefore each participant could earn up to $50 in primary incentives and up to $60 in secondary incentives, for a total maximum of $110. The storefronts where participants were interviewed and, in the case of the PDI, trained to educate and recruit peers, were selected and designed to be as identical as possible. The health officials working at these sites were also chosen by the same criteria.

The study found that the PDI was more successful in recruiting than the TOI, with the PDI getting 317 recruits compared to 233 for the TOI in the same length of time. Additionally, a t-test comparing the mean scores on the 8 question knowledge test showed that those in the PDI scored higher, with a p-value < 0.001. Additional t-tests and chi-square tests were conducted between the two populations for syringe sharing, cooker and filter sharing, rinse water sharing, injection frequency, and incidences of unsafe sex. The average estimated cost per recruit was also calculated, with the PDI being substantially cheaper at $16 per recruit, compared to $470 per recruit for the TOI.

Studies such as this establish many important facts about PDI efficacy and

implementation feasibility. Not only was the PDI more effective at getting recruits, it was able to show a greater change in risk behaviors from baseline, and it was substantially cheaper. It should be noted, however, that the PDI and TOI took place in different populations. Thus, despite the best efforts of the researchers, it is still possible that population differences accounted for some of these significant results.

Another type of comparison to a TOI that can be made is in the same population, but at different times. An example of this type of PDI was carried out in Ukraine by Smyrnov et al. (2012). In this study, t-tests for differences in means and z-tests for differences in proportions of factors were conducted amongst PDI and pre-PDI recruits. The pre-PDI recruits were collected in the 6 months prior to the beginning of the PDI. Again, this type of comparison would face problems if either the endogenous or exogenous the characteristics of the population changed over time.

2.2.3 Comparisons of Varied Incentive Structures within a Peer-Driven Intervention

An interesting PDI extension by Broadhead et al. (2006) examines the effect of the incentive structure. This PDI was carried out on IDUs in two cities in western Russia's Yaroslavl province. IDUs in Bragino were given a "Standard" PDI, while those in Rybinsk received a "Simplified" PDI. This study also had a second stage of recruiting, which will be discussed in the next section. The following description

of the incentive structure refers only to the first stage of recruiting.

Participants were given 100 rubles for completing their baseline questionnaire, and up to 30 rubles for each recruit's score on the knowledge test[3]. They received an additional 20 rubles for each woman they recruited or for the return of a postcard verifying re-locator information. The preceding incentives were provided for both the Standard and Simplified PDIs. In addition to these rewards, recruiters in the Standard PDI also received 30 rubles per successful recruit.

Over the first year of the study, the Standard PDI recruited 493 people, while the Simplified PDI recruited only 364. Again remember that despite the researcher's best efforts to choose comparable populations, the size or accessibility of the populations may be different, resulting in different recruiting counts. Broadhead et al. note that although the Standard PDI out-recruited the Simplified PDI, participants in the Simplified PDI out-scored their counterparts in the Standard PDI. A t-test for the average score on the knowledge test was significant with p-value < 0.001. Additionally, the average cost per recruit was higher for the Standard PDI, at 56.24 rubles per recruit compared to 25.42 rubles per recruit for the Simplified PDI.

The outcome of this study is logical: if recruiters only got paid for their efforts based on the recruit's score on the knowledge test, they might (1) make more of an effort to educate them, resulting in higher test scores on average; or (2) decide

[3]The test had 8 questions. Based on their recruit's score, recruiters received 3 rubles per correct question, and then a bonus of 3 rubles for each of questions 7 and 8 being correct given that the first six were all correct.

the effort of recruiting was not worth their time, resulting in overall fewer recruits. The incentive structure is not necessarily the causal force behind these outcomes, but it is consistent with our understanding of the situation.

2.2.4 Second-Time Recruiting

The study of IDUs in Yaroslavl province, Russia (Broadhead et al., 2006) also incorporated second-time recruiting. In this instance, after six months had passed, seeds and participants were asked to go back into the field and continue recruiting. They could either recruit a new participant, in which case they would educate them in the first body of information, or an old participant, whom they would be asked to educate on a second body of information.

For both the Standard and Simplified PDIs, people's follow-up test scores were significantly higher than their first test score (p-value < 0.001). However, it should be noted that as this test was on a different body of information, the questions could just have been scaled to be easier. The same patterns regarding recruitment counts and test scores for Standard versus Simplified PDI participants manifested themselves in this second stage as well.

The fact that PDI participants can successfully re-recruit and educate prior participants is promising for new or modified methodologies as well. Perhaps estimators could be adjusted using ideas from the capture-recapture sampling literature.

Another study that employed second-time recruiting was carried out in Vietnam and China targeting IDUs (Broadhead et al., 2009). These PDIs were conducted as feasibility studies for the two areas, and the authors note that the two locations had very different characteristics. Ha Giang town, in Ha Giang province, northern Vietnam is a small town of population 32,000, while Guigang township in Guangxi, China is larger at 300,000 people. In Vietnam, recruiters were given 30K VND for each male and 50K VND for each female recruit, up to 3 total, and a maximum of 30K VND per recruit based on their scores on the knowledge test. In China, recruiters were given 20 CNY per recruit regardless of gender and up to 16 CNY per recruit based on their scores on the knowledge test. In Vietnam, 282 people were recruited, compared to 610 in China.

Although outcome measures such as behavior with used syringes were measured, problems were found with the administration of the knowledge test. Participants in Guigang received almost perfect scores, and scores that were higher than those in Ha Giang, even though members of the Guigang population were significantly less educated. Again, this highlights the problem with making comparisons between populations with different characteristics.

2.2.5 Seed Randomization and Treatment/Control Groups

To address the problem of potentially different populations, Gwadz et al. (2011) developed a method whereby seeds are selected and then randomized into either treatment/intervention or control/placebo arms of the study. This enables them

to draw the seeds from the same population at the same time.

This PDI was targeted at people living with HIV/AIDS of color (PLHA of color) in New York City, and was aimed at increasing rates of screening for AIDS clinical trials among African Americans and Hispanics. Recruiters were given 3 coupons each, one of which was allotted for a female, and recruitment was conducted for 5 waves. The 49 seeds were assigned in a 2:1 ratio to treatment or control, and all recruits stemming from the initial seed were assigned to the same arm as the initial seed. All participants received $15 for the screening interview, $25 for each of two follow-up questionnaires and three or four intervention sessions[4], $15 per successful recruit, and up to $10 for each recruit's score on the knowledge test ($1 per question on a 10 question test).

The recruitment structure from each seed is shown in Figure 2.2[5]. Chains in the intervention arm are colored blue, while those in the control arm are colored orange. The seeds who did not recruit anyone are shown in the last panel. These recruitment chains demonstrate the great variability in recruitment patterns, which any model attempting to represent recruitment must capture. Note, for example, the chain from Seed 3 does not branch at all, with each person recruiting only one peer successfully. In contrast, the chain initiated by Seed 6

[4]These were small group activities of 6-9 persons, lasting approximately 1.5 hours each. Intervention arm participants could be in up to four sessions, while control arm participants were limited to at most three.

[5]These recruitment chains cover the full number of participants in the study, generated from data Gwadz et al. generously provided. The numbers of recruits cited textually are from (Gwadz et al., 2011), which puts an end-date on participation. After this date, intervention activities continued, resulting in the longer chains seen in the intervention arm, for example from Seeds 6 and 14.

encompasses many more people, with instances of one, two, and three successful recruits.

Overall, 342 people were recruited, 127 into the control arm and 215 into the intervention arm. Analysis was conducted via a logistic generalized linear mixed-model with a random intercept for the seed, adding many layers of sophistication to the analysis used in previous studies. Recall, the outcome of interest was screening for an AIDS clinical trial, which was gathered by self-report from the participants and then verified with clinical records. The terms used in the final model were age, number of peers recruited (0-3), and indicators for full intervention dose (four sessions for intervention arm participants, three for control), being African American, being Hispanic, having been screened for an AIDS clinical trial in the past, and being assigned to the intervention arm. The full intervention dose and assignment to the intervention arm were both found to be significant at the 0.05 level.

This approach is intriguing for several reasons. First, since we randomize seeds into intervention and control from the same population, the comparisons on the outcome variable are more relevant than when PDIs are compared to TOIs of a different population. Second, since the control sessions were time- and attention-matched to the intervention sessions, the effect is more directly attributable to the intervention itself, rather than a nebulous set of factors, such as taking time to come in to the health center or seeing any promotional material displayed there. Third, by assigning recruits to the same group as the root seed, the dependence

between the observations is acknowledged. Conducting t-tests as in some of the previous studies discussed overlooks this dependence.

One issue that arises with this design is how many seeds to select. Ideally, for a fixed sample size, we would prefer to choose fewer seeds and get longer chains. This would hopefully reduce effects resulting from the initial convenience sample of seeds. Conversely, we need enough seeds to guarantee that recruitment continues until the desired sample size is reached, and too few seeds risks the early termination of all chains. This creates a tension, and picking the best number of seeds is difficult. Since randomization occurs at the level of the seeds, we need more seeds to increase the power of our statistical test, which would result in shorter chains for the same fixed sample size. There is no easy solution to this problem, and more studies that use seed randomization would need to be carried out to get a sense of viable numbers of seeds.

In addition to providing better assessment methods for PDIs, this study demonstrates the great benefits that come with fitting a statistical model to the PDI. Examining coefficients on each of the parameters fit provides insight into the direction and magnitude of effects of factors that were frequently overlooked in comparisons to a TOI. A more complicated model which takes into account the discrete choices made by each person in the network is presented in Chapter 3.

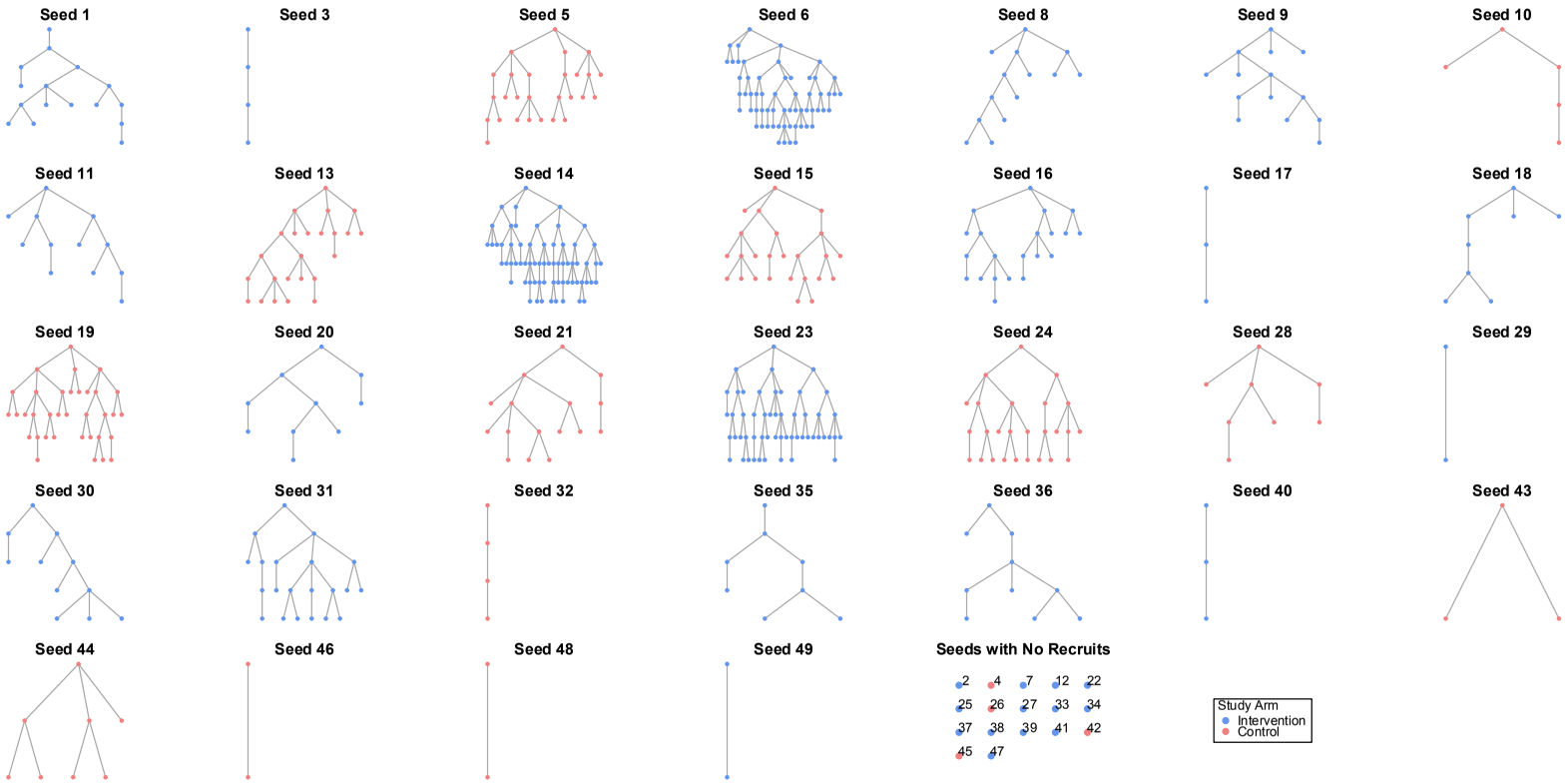


Figure 2.2: Recruitment chains for PDI targeted at people living with HIV/AIDS of color in New York City. Chains in the intervention arm are colored blue, while those in the control arm are colored orange. The seeds who did not recruit anyone into the study are shown in the last panel. There were 49 total seeds.

CHAPTER 3

Two-Sided Rational-Choice Preferential Recruitment Model

In this chapter I present the two-sided rational-choice preferential recruitment (RCPR) model. I begin by providing some motivating context and describing the background of two-sided matching models in Section 3.1. I then introduce some notation in Section 3.2, detail the matching mechanism used to generate recruitment chains from preferences in Section 3.3, and describe the complete inference procedure in Section 3.4. Finally, the complete Gibbs sampler over multiple waves of recruitment is provided in Section 3.5. Model selection is discussed in Section 3.6. The methodology to use the RCPR model to generate inclusion probabilities for members of the sample that can be used to estimate population prevalences is provided given in Section 3.7.

3.1 Introduction to Two-Sided Matching Models

A model for preferential recruitment should reflect the way we think coupon distribution occurs in reality. Consider a recruiter $i \in \mathcal{R}$, where $\mathcal{R}$ is the set of

Outcome	**Notation**
i does not recruit anyone	$c(i) = \{\emptyset\}$
i recruits one person	$c(i) = \{c_1\}$
$\dots$	$\dots$
i recruits n_c people	$c(i) = \{c_1, \dots, c_{n_c}\}$

Table 3.1: Possible recruitment outcomes.

recruiters, with the maximum n_c coupons in hand. After a recruitment action occurs, there are several different observable matching outcomes, shown in Table 3.1. $c(i)$ indicates the set of i's coupons that were returned when the peers they were given to participate in the study.

Note that the reality may be more nuanced that the observable outcome. For example, $c(i) = \{c_1\}$ and $n_c > 1$ implies that recruiter i had coupons to distribute that were not used. But we do not know if this is because recruiter i never distributed them, or if recruiter i did give them away but the peer(s) chose not to participate. Further, for $c(i) = \{c_1\}$, we do not know if the peer who accepted coupon c_1 was recruiter i's first choice, or if i tried to distribute the coupon, was immediately rebuffed, and then gave the coupon away to a peer who did participate. These uncertainties reflect the complicated nature of data collection.

Current RDS estimators assume that recruiters distribute their coupons uniformly at random. That is, each person in their social network has equal probability of receiving a coupon. This is likely not true in practice. I develop a recruitment model for RDS that allows for preferential recruitment on observed nodal and dyadic covariates in a manner that reflects how recruiters make coupon distribution decisions. This framework also relaxes many of the assumptions noted

in Table 1.3.2, notably: homophily sufficiently weak; with-replacement sampling; single non-branching chain; degree accurately measured; and random referral.

To model preferential recruitment, I use a *rational-choice framework* where actors in the system make decisions based on their utility for that action given the constraints that exist. The foundation for this work was developed in the context of two-sided matching models in the economics literature (Gale and Shapley (1962); Roth (1984); Roth and Sotomayor (1990)).

Gale and Shapley (1962) originally developed the *deferred acceptance algorithm* in the context of a marriage market. They envisioned a set of men and a set of women, and wanted to figure out an algorithm to create marriages between men and women based on individuals' preferences. This method considered one-to-one heterosexual matchings (i.e., no polygamous marriages, no homosexual marriages). Each man had ranked preferences for each women, and each woman had ranked preferences for each man. Additionally, each person also had a 'self' preference, meaning the preference for not getting married.

The original ideas of Gale and Shapley have been extended in numerous ways, notably to include many-to-one and many-to-many matchings (Roth and Sotomayor, 1990). The many-to-one scenario, which is most similar to RDS coupon distribution, was originally developed in the context of a college admissions process. It has also been applied to firms employing many workers and hospitals employing many interns (Roth, 1984). I discuss the methodology in the context of a college admissions process, where universities can grant admission to many

students (often with a quota), while students can accept at most one university. I will then adapt this process to the RDS framework.

Let $\mathbf{C} = \{C_1, \ldots, C_n\}$ and $\mathbf{S} = \{s_1, \ldots, s_n\}$ be finite and disjoint sets of colleges and students, respectively. Each student has preferences for each college, and each college has preferences for each student. Additionally, each student and college has a preference for themself, denoted 0. These preferences are complete and transitive, so they can be represented as ordered lists. Let $P(C_i) = s_1, s_2, 0, s_3, \ldots$ denote the order of college C_i's preferences, so that they prefer student s_1 over student s_2, but would prefer either of them to leaving the position unfilled. Student s_3 and all others ranked lower are unacceptable to college C_i. Likewise, student s_j's preferences can be written as $P(s_j) = C_3, C_1, C_2, 0, \ldots$, indicating that college C_3 is s_j's first choice, but they would be willing to accept an offer from any of the first three colleges.

An outcome of the college admissions model is a matching of students to colleges such that each student is matched to at most one college, and each college is matched to at most its quota of students. A student who is not matched to any college is 'self-matched'. A college with quota q_C who is matched to $n_s < q_C$ students is said to have self-matched $q_C - n_s$ times. Formally,

Definition 3.1.1. A matching m is a function from the set $\mathbf{C} \cup \mathbf{S}$ into the set of unordered families of elements of $\mathbf{C} \cup \mathbf{S}$ such that:

1. $|m(s)| = 1$ for every student s and $m(s) = 0$ if $m(s) \notin \mathbf{C}$;
2. $|m(C)| = q_C$ for every college C, and if the number of students in $m(C)$, say

n_s, is less than q_C, then $m(C)$ contains $q_C - n_s$ copies of 0;

3. $m(s) = C$ if and only if s is in $m(C)$.

When considering many-to-one matches, it is also necessary to define how entities (in this case, colleges) prefer different groups. In one-to-one matchings this is solved by directly comparing the relative preferences of two options. In the many-to-one scenario, consider two matchings m and m' where $m(C_i)$ assigns C_i its first and second choices. If $m'(C_i)$ assigns C_i its second and third choices, then college C_i prefers matching m over m'. This is known as *responsiveness*. Formally,

Definition 3.1.2. The preference relation $P^{\#}(C)$ over sets of students is *responsive* if, whenever $m'(C) = m(C) \cup \{s_k\} \setminus \{\sigma\}$ for σ in $m(C)$ and s_k not in $m(C)$, then C prefers $m'(C)$ to $m(C)$ if and only if C prefers s_k to σ.

Many different matchings are possible between a set of colleges and a set of students. Because of the complex series of rankings, picking an 'optimal' matching is difficult, so we focus on matchings that are *stable*. This is an important concept which will be used in deriving the full conditional distributions for the inference procedure discussed in Section 3.4.

The original deferred acceptance algorithm (Gale and Shapley, 1962) was designed to create stable matchings. In a stable match, each student does not need to be matched to the college for which they have the highest preference (and this is likely impossible) and each college does not need to be matched to the students for whom they have the highest preference. Instead, the matching needs to not

be *blocked* by any individual agent or college-student pair. A student blocks a matching if $m(s) = C$ but C is unacceptable to s. Similarly, a college blocks a matching if $m(C) = s$ but s is unacceptable to C. A college-student pair (C, s) blocks a matching if $m(s) \neq C$ but s prefers C over the college they are matched to and C prefers s over one of the students they are matched to. College C and student s are then part of a *coalition* A that blocks the matching. Formally,

Definition 3.1.3. A matching m is *stable* if $\nexists$ another matching m' such that for all students s and all colleges C in a coalition A

1. $m'(s) \in A$ (every student in the coalition is matched to a college in the coalition);
2. $m'(s) > m(s)$ (every student in the coalition prefers the college in match m' over the college in m);
3. $\sigma \in m'(C)$ implies $\sigma \in A \cup m(C)$ (every college in the coalition is matched to at least one student in the coalition, and the students not in the coalition that the college is matched to must also be matches of C in m);
4. $m'(C) > m(C)$ (every college in the coalition prefers its set of students in m' to those in m)

That is, m is blocked by a coalition A of colleges and students if, by matching among themselves, the students and colleges in A could all get an assignment preferable to m. Put another way, the following scenarios would all result in an unstable match: s prefers themself over C but $m(s) = C$; C prefers themself over s but $m(C) = s$; $m(s_1) = C_1$ and $m(s_2) = C_2$ but s_1 prefers C_2 over C_1 and C_2 prefers s_1 over s_2. In the last scenario, there is a college-student pair who would both prefer to dissolve an existing tie and form a tie with each other instead. Figure 3.1 gives an example of matchings that are stable and unstable.

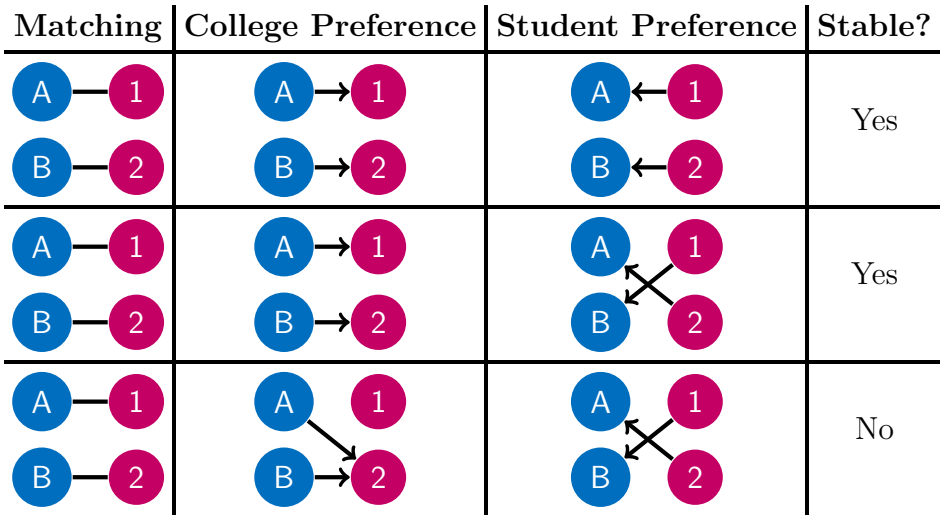


Figure 3.1: Stability example. Colleges A and B are shown in blue; students 1 and 2 are shown in red.

In the two-sided matching model literature, it is assumed that the preferences of the agents are known, and the goal is to create a matching that is stable. In RDS recruitment, however, the preferences are unknown but we observe the structure of the recruitment chain (matching). We would like to use the observed matching and be able to make inferences about the preferences of the actors. Inference for covariates affecting matching preferences, based on an observed matching, was first developed by (Logan et al., 2008). They again consider the one-to-one heterosexual marriage matching scenario, and develop the *revealed preferences model* to make inference about the variables affecting the matching. The revealed preferences model will be explained as I adapt it for use in the RDS scenario.

Briefly, many-to-one matches are allowed so that recruiters can select multiple peers, up to the number of coupons they have to distribute. To accommodate this adaptation, I developed a new matching mechanism, derived new constraints and

updating steps, and created a separate model that integrates more control over recruitment robustness via the handling of self-matches. Additionally, I allow the sequential application of the model across multiple waves of recruitment. These steps will be clarified as the model is explained below.

Using this background from the two-sided matching model literature, I now develop the two-sided rational-choice preferential recruitment model, referred to as the RCPR model. I first lay out some preliminary notation in Section 3.2. I then discuss the matching mechanism which can be used to generate stable many-to-one matchings from known preferences in Section 3.3. The complete inference procedure is described in Section 3.4 for one wave of recruitment. Sequential application over many waves of recruitment is discussed in Section 3.5.

3.2 Notation

To present the model, we first lay out some basic notation. For a specific wave, let $\mathcal{R}$ be the set of recruiters and $\mathcal{P}$ be the set of peers such that $\mathcal{R} \cup \mathcal{P} = \emptyset$. Let $\mathcal{S}_R$ be the set of possible recruiter self-matches such that $|\mathcal{S}_R| = n_c$, the number of coupons each has to distribute. The implicit difference here from the many-to-one matchings of Gale and Shapley (1962) and Roth and Sotomayor (1990) is that recruiters have different preferences for self-matching with their different coupons, rather than just one self-match value. Let $\mathcal{S}_P$ be the set of possible peer self-matches such that $|\mathcal{S}_P| = 1$.

Let $i \in \mathcal{R} \cup \mathcal{S}_P$ be a recruiter, $j \in \mathcal{P} \cup \mathcal{S}_R$ be a peer, and m be a matching. Then $m(i) = j$ indicates that i recruited j in matching m. If $j \in \mathcal{P}$ then a successful recruitment occurred; if $j \in \mathcal{S}_R$, then recruiter i self-matched. Likewise, if $m(j) = i$ and $i \in \mathcal{R}$ then j was recruited and participated; if $i \in \mathcal{S}_P$, then j self-matched. A recruiter i can select up to n_c peers or self-matches. A peer j can be recruited by only one recruiter, or not be recruited by self-matching.

At wave w, let $\mathcal{R}_w$ be the set of recruiters, where $|\mathcal{R}_w| = n_{r,w}$, and let $\mathcal{P}_w$ be the set of peers, where $|\mathcal{P}_w| = n_{p,w}$. The recruiters in wave w are those who were recruited during wave $w-1$, and the peers are all those who were recruited during waves $w, \ldots, n_w$, where n_w is the total number of waves[1]. Let n_s be the number of seeds and n_c be the maximum number of coupons each recruiter could distribute. (This is typically 2-3 in most RDS studies.) Then the maximum sample size is

$$n^* = n_s \left(\sum_{w=0}^{n_w} n_c^w \right). \tag{3.1}$$

We form a bipartite network $\mathcal{A}_w$ at wave w between recruiters in $\mathcal{R}_w \cup \mathcal{S}_P$ and peers in $\mathcal{P}_w \cup \mathcal{S}_R$. Assume that each recruiter $i \in \mathcal{R}_w$ has a utility $u_{ij;w}$ for each peer $j \in \mathcal{P}_w \cup \mathcal{S}_R$[2]. Symmetrically, assume that each peer $j \in \mathcal{P}_w$ has a utility $v_{ji;w}$

[1] We only observe covariate information about people who participated at some point in the study, and not other members of the population who did not enroll in the study. However, we still glean information about preferential recruitment based on people who were recruited in later waves, but not earlier ones. If full population information was known, then $\mathcal{P}_w$ would be the set of all members of the population who had not yet participated in the RDS study by wave w.

[2] We can assume that if i and j are not tied in the underlying social network, then $u_{ij} = -\infty$. Therefore, since we assume that people have a finite utility for recruiting themself, they will never recruit people they are not tied to.

for each recruiter $i \in \mathcal{R}_w \cup \mathcal{S}_P$. These utilities are continuous, but can be ranked to get the preferences needed to generate a matching. Let $\mathbf{U}_w$ and $\mathbf{V}_w$ be the utility matrices where $\mathbf{U}_w$ has dimension $n_{r;w} \times (n_{p;w} + n_c)$ and $\mathbf{V}_w$ has dimension $n_{p;w} \times (n_{r;w} + 1)$. The row-ranked matrices with preferences are denoted $\mathbf{U}_w^*$ and $\mathbf{V}_w^*$.

In this chapter the subscript w is often dropped for conciseness, since the same method is applied in turn to each wave of the study. Each wave is represented by a different bipartite network. Thus, $\mathcal{R}$ is the set of recruiters of size n_r, $\mathcal{P}$ is the set of peers of size n_p, u_{ij} is the utility that recruiter $i \in \mathcal{R}$ has for peer $j \in \mathcal{P} \cup \mathcal{S}_R$, and v_{ji} is the utility that peer $j \in \mathcal{P}$ has for recruiter $i \in \mathcal{R} \cup \mathcal{S}_P$.

These utilities $\mathbf{U}$ and $\mathbf{V}$ are unobserved (latent), but we assume they can be represented by linear combinations of nodal or dyadic covariates that we can observe. More precisely,

$$u_{ij} = \alpha \mathbf{X}_{ij} + \xi \mathbf{S}_{ij}^R + \epsilon_{ij} \tag{3.2}$$

$$v_{ji} = \beta \mathbf{Y}_{ji} + \zeta \mathbf{S}_{ji}^P + \gamma_{ji} \tag{3.3}$$

where α and β are preference coefficients governing the covariates, ξ and ζ are preference coefficients governing self-matching, $\mathbf{X}$ and $\mathbf{Y}$ are matrices of observed covariate matches between dyads (where $\mathbf{X}_{ij} = 0$ for $j \in \mathcal{S}_r$ and $\mathbf{Y}_{ji} = 0$ for $i \in \mathcal{S}_p$ by definition), $\mathbf{S}^R$ and $\mathbf{S}^P$ are indicator arrays of self-match opportunity, and ϵ and γ are unobserved pair-specific heterogeneities. Therefore we want to make

inference about the preference coefficients α, β, ξ, and ζ given $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{S}^R$, $\mathbf{S}^P$, and the observed recruitment structure. Note that α and β will have length equal to n_v, the number of variables; ξ will have length n_c, the number of coupons; and ζ will have length one.

First the matching mechanism to generate a matching from utilities is discussed, then the inference procedure is detailed.

3.3 Matching Mechanism

We assume that the observed recruitment structure represents a stable match in the bipartite network between recruiters and peers at a given wave. We further assume that this stable match can be generated from a complete set of individuals' preferences via a matching algorithm. The elements in each row of $\mathbf{U}$ and $\mathbf{V}$ can be ranked to get each individual's preferences.

Therefore we construct Algorithm 1, which returns a stable matching given utilities. To begin, each recruiter $i \in \mathcal{R}$ selects the n_c peers $j \in \mathcal{P} \cup \mathcal{S}_R$ for whom they have the highest utility u_{ij} and places them in their current choice set, $\mathcal{C}(i)$. Self-selections are possible (i.e., $j \in \mathcal{S}_R$) and always accepted. Each peer $j \in \mathcal{P}$ considers all recruiters who selected them (denoted $\mathcal{J}(j)$) and chooses the one for whom they have the highest utility v_{ji}, denoted $\tilde{\mathcal{J}}(j)$. They also have the option of selecting themself ($\mathcal{S}_P \subseteq \mathcal{J}(j)$). If j did not select i, j is removed from i's current choice set $\mathcal{C}(i)$ and is added to i's discard set $\mathcal{D}(i)$. Recruiter i then adds

peers with the next highest utilities u_{ij} who they have not already tried to recruit (i.e., $j \notin \mathcal{D}(i)$) to $\mathcal{C}(i)$ until $|\mathcal{C}(i)| = n_c$. The peers then respond to their offers again. The procedure continues until every recruiter i has n_c tentative accepts or self-matches.

Algorithm 1 Matching mechanism: Recruitment from utilities, maximum coupons n_c

initialize for each recruiter $i \in \mathcal{R}$
  current choices $\mathcal{C}(i) = \emptyset$, and let $c_i = |\mathcal{C}(i)|$
  discarded peers $\mathcal{D}(i) = \emptyset$
for each recruiter $i \in \mathcal{R}$ **do**
  while $c_i < n_c$ for at least one i **do**
    Add j with largest u_{ij} to $\mathcal{C}(i)$, where $j \in (\mathcal{P} \cup \mathcal{S}_R) \setminus (\mathcal{C}(i) \cup \mathcal{D}(i))$
  end while
end for
for each peer $j \in \mathcal{P}$ **do**
  $\mathcal{J}(j) = \mathcal{S}_P \cup \{i\}$ such that $j \in \mathcal{C}(i)$
  $\tilde{\mathcal{J}}(j)$ is i that maximizes v_{ji}, where $i \in \mathcal{J}(j)$
  for each $i \in \mathcal{J}(j)$, $i \neq \tilde{\mathcal{J}}(j)$ **do**
    Add j to $\mathcal{D}(i)$
    Remove j from $\mathcal{C}(i)$
  end for
end for
while $c_i < n_c$ for at least one i **do**
  Repeat (4)-(16)
end while

This algorithm is applied sequentially, beginning with the first wave where the recruiters are the initial seeds. The people who were recruited in the previous wave become the recruiters in the current wave. This procedure can be carried out for a set number of waves.

3.3.1 Matching Algorithm Example 1

As an example, consider a small example with $n_r = 3$, $n_p = 8$, and $n_c = 3$. Figure 3.2 provides a visual example (here, the utilities are not explicitly stated; an example where they are can be found in the following section). In the first recruitment attempt, recruiter A gives coupons to peers 1, 3, and 4; B gives coupons to 4, 5, and 8; and C gives coupons to 7 and themself twice. The peers then respond: 1 and 3 select A; 4 selects B; 7 selects C; and 2, 5, 6, and 8 all select themself. Note that, for example, 5 selected themself over a recruiter, while 2 selected themself because that was their only option (they never received a coupon). The recruiters that had rejected coupons now distribute them again[3]. In the final matching, A successfully recruits 1, 3, and 5; B successfully recruits 4 and 6; and C successfully recruits 7. This means that A has self-matched zero times, B once, and C twice.

A more detailed example, with specific utilities, is provided in Example 2.

3.3.2 Matching Algorithm Example 2

Consider a scenario with three recruiters ($\mathcal{R} = \{r_1, r_2, r_3\}$) and five peers ($\mathcal{P} = \{p_1, p_2, p_3, p_4, p_5\}$), where each recruiter has $n_c = 2$ coupons. Therefore $\mathcal{S}_\mathcal{R} = \{\dot{s}_1, \dot{s}_2\}$ and $\mathcal{S}_P = \{\breve{s}_1\}$. Assume the utility matrices $\mathbf{U}$ and $\mathbf{V}$ are known and can be ranked across each row to produce the row-ranked matrixes $\mathbf{U}^*$ and $\mathbf{V}^*$.

[3]In RDS implementation, we view this as a recruitment attempt with immediate refusal – so the recruiter retains the coupon. The scenario where peers take a coupon but never use it to participate gets conflated into the first scenario.

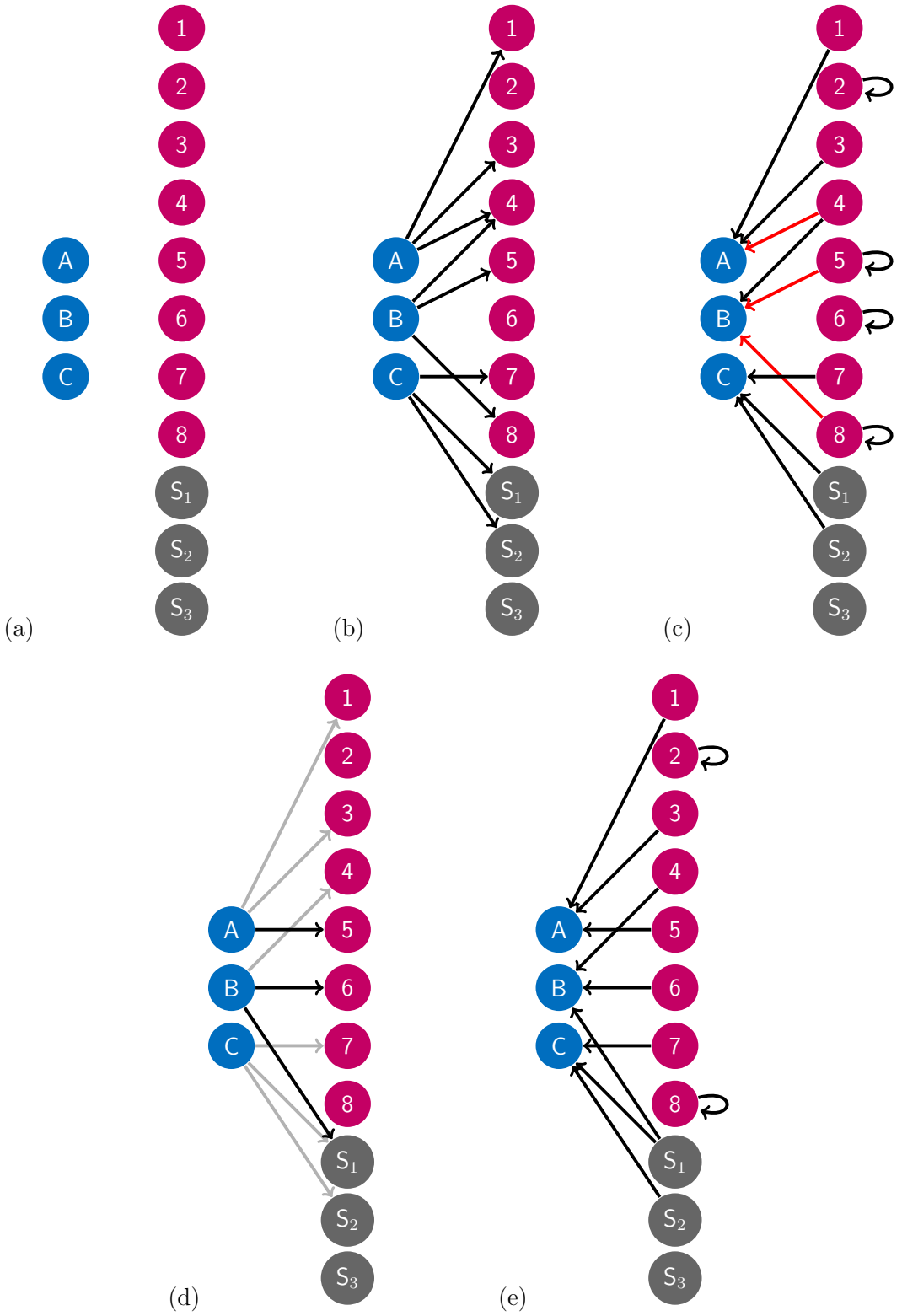


Figure 3.2: Schematic of matching mechanism with $n_r = 3$ recruiters, $n_p = 8$ peers, and $n_c = 3$ coupons. (b) and (d) show recruiter coupon distribution attempts, and (c) and (e) show peer responses.

Let

$$\mathbf{U}^* = \begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \end{array} \begin{array}{c} \begin{array}{ccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & \dot{s}_1 & \dot{s}_2 \end{array} \\ \begin{pmatrix} 1 & 3 & 6 & 5 & 7 & 2 & 4 \\ 6 & 3 & 5 & 7 & 2 & 1 & 4 \\ 6 & 7 & 2 & 4 & 1 & 3 & 5 \end{pmatrix} \end{array} \tag{3.4}$$

$$\mathbf{V}^* = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{array} \begin{array}{c} \begin{array}{cccc} r_1 & r_2 & r_3 & \breve{s}_1 \end{array} \\ \begin{pmatrix} 1 & 2 & 4 & 3 \\ 3 & 4 & 2 & 1 \\ 4 & 1 & 3 & 2 \\ 2 & 4 & 3 & 1 \\ 3 & 4 & 1 & 2 \end{pmatrix} \end{array} \tag{3.5}$$

where '1' indicates the highest preference. Table 3.2 shows how the matching algorithm applies in this example. At each iteration, new choices are shown as filled in circles, while previously tentatively accepted choices are shown as open circles.

The final outcome is that r_1 selects p_1, r_2 does not recruit anyone, and r_3 selects p_5.

	Recruiter Choices	Peer Options	Peer Choices	Recruiter Discard
1	$\mathcal{C}(r_1) = p_1\ \dot{s}_1$ $\mathcal{C}(r_2) = \dot{s}_1\ p_5$ $\mathcal{C}(r_3) = p_5\ p_3$	$\mathcal{J}(p_1) = r_1\ \breve{s}_1$ $\mathcal{J}(p_2) = \breve{s}_1$ $\mathcal{J}(p_3) = r_3\ \breve{s}_1$ $\mathcal{J}(p_4) = \breve{s}_1$ $\mathcal{J}(p_5) = r_2\ r_3\ \breve{s}_1$	$\widetilde{\mathcal{J}}(p_1) = r_1$ $\widetilde{\mathcal{J}}(p_2) = \breve{s}_1$ $\widetilde{\mathcal{J}}(p_3) = \breve{s}_1$ $\widetilde{\mathcal{J}}(p_4) = \breve{s}_1$ $\widetilde{\mathcal{J}}(p_5) = r_3$	$\mathcal{D}(r_1) =$ $\mathcal{D}(r_2) = p_5$ $\mathcal{D}(r_3) = p_3$
2	$\mathcal{C}(r_1) = p_1\ \dot{s}_1$ $\mathcal{C}(r_2) = \dot{s}_1\ p_2$ $\mathcal{C}(r_3) = p_5\ \dot{s}_1$	$\mathcal{J}(p_1) = r_1$ $\mathcal{J}(p_2) = r_2\ \breve{s}_1$ $\mathcal{J}(p_3) = \breve{s}_1$ $\mathcal{J}(p_4) = \breve{s}_1$ $\mathcal{J}(p_5) = r_3$	$\widetilde{\mathcal{J}}(p_1) = r_1$ $\widetilde{\mathcal{J}}(p_2) = \breve{s}_1$ $\widetilde{\mathcal{J}}(p_3) = \breve{s}_1$ $\widetilde{\mathcal{J}}(p_4) = \breve{s}_1$ $\widetilde{\mathcal{J}}(p_5) = r_3$	$\mathcal{D}(r_1) =$ $\mathcal{D}(r_2) = p_2\ p_5$ $\mathcal{D}(r_3) = p_3$
3	$\mathcal{C}(r_1) = p_1\ \dot{s}_1$ $\mathcal{C}(r_2) = \dot{s}_1\ \dot{s}_2$ $\mathcal{C}(r_3) = p_5\ \dot{s}_1$	$\mathcal{J}(p_1) = r_1$ $\mathcal{J}(p_2) = \breve{s}_1$ $\mathcal{J}(p_3) = \breve{s}_1$ $\mathcal{J}(p_4) = \breve{s}_1$ $\mathcal{J}(p_5) = r_3$	$\widetilde{\mathcal{J}}(p_1) = r_1$ $\widetilde{\mathcal{J}}(p_2) = \breve{s}_1$ $\widetilde{\mathcal{J}}(p_3) = \breve{s}_1$ $\widetilde{\mathcal{J}}(p_4) = \breve{s}_1$ $\widetilde{\mathcal{J}}(p_5) = r_3$	$\mathcal{D}(r_1) =$ $\mathcal{D}(r_2) = p_2\ p_5$ $\mathcal{D}(r_3) = p_3$

Table 3.2: Matching algorithm example where the row-ranked preference matrices $\mathbf{U}^*$ and $\mathbf{V}^*$ are given in Equations 3.4 and 3.5, respectively.

3.4 Inference

One primary goal of modeling preferential recruitment is to make inference about the preference coefficients α, β, ξ, and ζ given the observed covariates $\mathbf{Z}$ and the observed recruitment structure, including self-match information $\mathbf{S}$. In practice, we do not observe α, β, ξ, and ζ, but we do observe $\mathbf{Z}$, $\mathbf{S}$, and the recruitment chain.

Recall that $\mathbf{X}$ and $\mathbf{Y}$ are indicator arrays of dyadwise matches on covariates $\mathbf{Z} = (Z_1, \ldots, Z_{n_v})$, and are related to the utilities $\mathbf{U}$ and $\mathbf{V}$ via the linear forms given in Equations (3.2) and (3.3). $\mathbf{X}$ has dimension $(n_r) \times (n_p + n_c) \times (n_v)$ and $\mathbf{Y}$ has dimension $(n_p) \times (n_r + 1) \times (n_v)$. For covariate $c \in \{Z_1, \ldots, Z_{n_v}\}$, let

$$X_{ij;c} = Y_{ji;c} = \begin{cases} 1 & z_{i;c} = z_{j;c} \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

where $z_{i;c}$ is recruiter i's value of covariate c and $z_{j;c}$ is peer j's value of covariate c. For recruiter self matches ($j \in \mathcal{S}_R$), by definition $X_{ij;c} = 0$. Likewise, for peer self matches ($i \in \mathcal{S}_P$), by definition $Y_{ji;c} = 0$.

Recall that $\mathbf{S}^R$ and $\mathbf{S}^P$ are indicator arrays of self match opportunity. $\mathbf{S}^R$ has

dimension $(n_r) \times (n_p + n_c)$ and $\mathbf{S}^P$ has dimension $(n_p) \times (n_r + 1)$. Thus

$$\mathbf{S}^R_{ij} = \begin{cases} 1 & \text{if } j \in \mathcal{S}_R \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \mathbf{S}^P_{ji} = \begin{cases} 1 & \text{if } i \in \mathcal{S}_P \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

Let m be a matching, encompassing all the information we know about the recruitment chain: the precise recruitment structure, number of waves, and number of coupons to distribute. We are interested in finding

$$\arg\max_{\alpha,\beta,\xi,\zeta} p(m \in \mathcal{M} \,|\, \alpha, \beta, \xi, \zeta) \tag{3.8}$$

where $\mathcal{M}$ is the set of all possible stable matchings. We treat the utilities $\mathbf{U}$ and $\mathbf{V}$ as auxiliary variables and examine the marginal probability of stability

$$\begin{aligned} &\arg\max_{\alpha,\beta,\xi,\zeta} p(m \in \mathcal{M} \,|\, \alpha, \beta, \xi, \zeta) = \\ &\qquad \arg\max_{\alpha,\beta,\xi,\zeta} \iint p(m \in \mathcal{M} \,|\, \mathbf{U}, \mathbf{V})\, p(\mathbf{U}, \mathbf{V} \,|\, \alpha, \beta, \xi, \zeta)\, d\mathbf{U}\, d\mathbf{V}. \end{aligned} \tag{3.9}$$

Note that the probability of a specific matching m given the utilities $\mathbf{U}$ and $\mathbf{V}$ is either 0 or 1, so the first term in Equation (3.9) is an indicator of whether the utilities produce m via Algorithm 1. We therefore focus on the second term in Equation (3.9). However, calculating $\arg\max_{\alpha,\beta,\xi,\zeta} \iint p(\mathbf{U}, \mathbf{V} | \alpha, \beta, \xi, \zeta)\, d\mathbf{U}\, d\mathbf{V}$ involves an integral over a high-dimensional space, and is computationally intractable.

We therefore take a Bayesian approach, and draw inference based on the pos-

terior distribution

$$\pi(\alpha, \beta, \xi, \zeta \,|\, m \in \mathcal{M}) \propto \pi(\alpha, \beta, \xi, \zeta)\, p(m \in \mathcal{M} \,|\, \alpha, \beta, \xi, \zeta) \tag{3.10}$$

where $\pi(\alpha, \beta, \xi, \zeta)$ is a prior distribution for α, β, ξ, and ζ representing our uncertainty about the parameters. Although the second term in Equation 3.10 involves the same intractable integral as Equation 3.9, inference is made feasible by approximating the posterior via samples of α, β, ξ, and ζ from a Markov chain.

Unlike many likelihood models, in this model the contributions from different individuals to the likelihood $p(m \in \mathcal{M} \,|\, \alpha, \beta, \xi, \zeta)$ are not multiplicative. This is due to the complicated way in which the assumption of stability constrains the utilities. Since we cannot calculate or sample directly from $\pi(\alpha, \beta, \xi, \zeta \,|\, m \in \mathcal{M})$, we instead sample values of α, β, ξ, ζ, $\mathbf{U}$, and $\mathbf{V}$ from a distribution approximating $\pi(\alpha, \beta, \xi, \zeta, \mathbf{U}, \mathbf{V} \,|\, m \in \mathcal{M})$. We then examine the sampled values α, β, ξ, and ζ marginally to estimate $\pi(\alpha, \beta, \xi, \zeta \,|\, m \in \mathcal{M})$. This is an example of using data augmentation to simplify posterior sampling (Tanner and Wong, 1987).

We can then make draws from the joint posterior of $\pi(\alpha, \beta, \xi, \zeta, \mathbf{U}, \mathbf{V} \,|\, m \in \mathcal{M})$ using Gibbs sampling because we can specify the full conditional distributions. The Gibbs sampler is described in full in Section 3.5. The full conditional distri-

butions are:

$$p(\alpha \,|\, \beta, \xi, \zeta, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) \tag{3.11}$$

$$p(\beta \,|\, \alpha, \xi, \zeta, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) \tag{3.12}$$

$$p(\xi \,|\, \alpha, \beta, \zeta, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) \tag{3.13}$$

$$p(\zeta \,|\, \alpha, \xi, \zeta, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) \tag{3.14}$$

$$p(\mathbf{U} \,|\, \alpha, \beta, \xi, \zeta, \mathbf{V}, m \in \mathcal{M}) \tag{3.15}$$

$$p(\mathbf{V} \,|\, \alpha, \beta, \xi, \zeta, \mathbf{U}, m \in \mathcal{M}) \tag{3.16}$$

These distributions will be specified in Sections 3.4.1 through 3.4.6.

Let $\phi = (\alpha, \beta, \xi, \zeta, \mathbf{U}, \mathbf{V})$ represent a state of the unknowns in the model, which includes both the unobserved preference parameters and the unknown utilities. We use the Metropolis-Hastings algorithm to iteratively construct sequences of ϕ from which we can approximate $\pi(\alpha, \beta, \xi, \zeta \,|\, m \in \mathcal{M})$. For example, given the most recent state $\phi^{(k)}$, a new value $\phi^{(k+1)}$ is constructed by separately updating each unobserved utility $\mathbf{U}$ and $\mathbf{V}$, and each component of the preference parameters α, β, ξ, and ζ. Sampling from a full conditional distribution is Gibbs sampling, and the acceptance probability of such a sample is always unity. The procedures for updating the utilities and preference parameters are described in the following sections.

As $k \to \infty$, the sampling distribution of $\phi^{(k)}$ approaches the desired posterior $\pi(\phi \,|\, m \in \mathcal{M})$. Therefore, the empirical distribution of the samples of ϕ from the

Markov chain can be used to approximate $\pi(\phi \,|\, m \in \mathcal{M})$.

To specify the full conditional distributions for α, β, ξ, and ζ, recall that we assume $u_{ij} = \alpha \mathbf{X}_{ij} + \xi \mathbf{S}^R_{ij} + \epsilon_{ij}$ and $v_{ji} = \beta \mathbf{Y}_{ji} + \zeta \mathbf{S}^P_{ji} + \gamma_{ji}$. If we further assume $\epsilon_{ij} \sim N(\mu_\epsilon, \sigma^2_\epsilon)$, $\mu_\epsilon = 0$, $\gamma_{ji} \sim N(\mu_\gamma, \sigma^2_\gamma)$, and $\mu_\gamma = 0$, (i.e., normal mean zero unobserved heterogeneity), then we can construct a two-sided probit model where the conjugate priors for α, β, ξ, and ζ are multivariate normal.

3.4.1 Full Conditional Distribution for α

Based on the independence of the preference parameters given specific utilities $\mathbf{U}$ and $\mathbf{V}$, and specifically conditioning on the observed dyadwise covariate matches, which have been left implicit to this point,

$$\begin{aligned} p(\alpha \,|\, \beta, \xi, \zeta, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) &= p(\alpha \,|\, \mathbf{U}, \mathbf{X}) \\ &\propto \pi(\alpha)\, p(\mathbf{U} \,|\, \alpha, \mathbf{X}) \end{aligned} \tag{3.17}$$

This is the format for a probit model using the multivariate normal as the conjugate prior: $\pi(\alpha) = \mathcal{MVN}(\mu_\alpha, \mathbf{\Sigma}_\alpha)$.

Thus by multiplying together the terms

$$
\begin{aligned}
p(\alpha \,|\, \mathbf{U}, \mathbf{X}) &\propto \pi(\alpha)\, p(\mathbf{U} \,|\, \alpha, \mathbf{X}) \qquad (3.18)\\
&= \pi(\alpha) \prod_{i=1}^{n_r} \prod_{j=1}^{n_p} p(u_{ij} \,|\, \alpha, \mathbf{X}_{ij}) \\
&\propto \mathcal{MVN}(\mu_\alpha, \mathbf{\Sigma}_\alpha) \prod_{i=1}^{n_r} \prod_{j=1}^{n_p} N(\alpha \mathbf{X}_{ij}, \sigma_\epsilon^2) \\
&\propto \left(\exp\{(\alpha - \mu_\alpha)' \mathbf{\Sigma}_\alpha^{-1} (\alpha - \mu_\alpha)\}\right) \left(\exp\left\{\sum_{i=1}^{n_r} \sum_{j=1}^{n_p} (u_{ij} - \alpha \mathbf{X}_{ij})^2\right\}\right) \\
&\propto \exp\left\{\alpha' \mathbf{\Sigma}_\alpha^{-1} \alpha - 2\alpha' \mathbf{\Sigma}_\alpha^{-1} \mu_\alpha + \sum_{i=1}^{n_r} \sum_{j=1}^{n_p} \left(\alpha' \mathbf{X}_{ij}' \mathbf{X}_{ij} \alpha - 2 u_{ij} \alpha \mathbf{X}_{ij}\right)\right\} \\
&\propto \exp\left\{\alpha' \left(\sum_{i=1}^{n_r} \sum_{j=1}^{n_p} \mathbf{X}_{ij} \mathbf{X}_{ij}' + \mathbf{\Sigma}_\alpha^{-1}\right) \alpha - 2\alpha' \left(\sum_{i=1}^{n_r} \sum_{j=1}^{n_p} \mathbf{X}_{ij} u_{ij} + \mathbf{\Sigma}_\alpha^{-1} \mu_\alpha\right)\right\}
\end{aligned}
$$

Let $\Upsilon^{-1} = \sum_{i=1}^{n_r} \sum_{j=1}^{n_p} \mathbf{X}_{ij} \mathbf{X}_{ij}' + \mathbf{\Sigma}_\alpha^{-1}$, the first term in parentheses, and $\Lambda = \sum_{i=1}^{n_r} \sum_{j=1}^{n_p} \mathbf{X}_{ij} u_{ij} + \mathbf{\Sigma}_\alpha^{-1} \mu_\alpha$, the second term in parentheses. Then by completing the square

$$
\begin{aligned}
\pi(\alpha)\, p(\mathbf{U} \,|\, \alpha, \mathbf{X}) &\propto \exp\left\{\alpha'(\Upsilon^{-1})\alpha - 2\alpha'(\Lambda)\right\} \\
&\propto \exp\left\{(\alpha - \Upsilon\Lambda)' \Upsilon^{-1} (\alpha - \Upsilon\Lambda)\right\} \\
&\propto \mathcal{MVN}(\Upsilon\Lambda, \Upsilon)
\end{aligned}
$$

So if $\pi(\alpha) = \mathcal{MVN}(\mu_\alpha, \mathbf{\Sigma}_\alpha)$, then

$$
\alpha^{(k+1)} \,|\, \mathbf{U}^{(k)}, \mathbf{X} \sim \mathcal{MVN}(\tilde{\mu}_\alpha, \tilde{\mathbf{\Sigma}}_\alpha) \qquad (3.19)
$$

where

$$\tilde{\boldsymbol{\Sigma}}_{\alpha}^{-1} = \sum_{i=1}^{n_r} \sum_{j=1}^{n_p} \mathbf{X}_{ij} \mathbf{X}_{ij}' + \boldsymbol{\Sigma}_{\alpha}^{-1} \tag{3.20}$$

$$\tilde{\mu}_{\alpha} = \tilde{\boldsymbol{\Sigma}}_{\alpha} \left(\sum_{i=1}^{n_r} \sum_{j=1}^{n_p} \mathbf{X}_{ij} u_{ij}^{(k)} + \boldsymbol{\Sigma}_{\alpha}^{-1} \mu_{\alpha} \right). \tag{3.21}$$

Updated values of α at time $(k+1)$ depend on the utilities at the previous time (k), as well as the observed covariates and the original prior specified. Note that if we know $\mathbf{U}^{(k)}$ then $\alpha^{(k+1)}$ does not depend on $\mathbf{V}^{(k)}$ or m, but that $\mathbf{U}^{(k)}$ does depend on $\mathbf{V}^{(k)}$ and m, as will be shown in Section 3.4.5.

3.4.2 Full Conditional Distribution for β

Based on the independence of the preference parameters given specific utilities $\mathbf{U}$ and $\mathbf{V}$, and specifically conditioning on the observed dyadwise covariate matches, which have been left implicit to this point,

$$\begin{aligned} p(\beta \,|\, \alpha, \xi, \zeta, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) &= p(\beta \,|\, \mathbf{V}, \mathbf{Y}) \\ &\propto \pi(\beta)\, p(\mathbf{V} \,|\, \beta, \mathbf{Y}) \end{aligned} \tag{3.22}$$

This is the format for a probit model using the multivariate normal as the conjugate prior: $\pi(\beta) = \mathcal{MVN}(\mu_{\beta}, \boldsymbol{\Sigma}_{\beta})$. The derivation for the full conditional distribution is identical to the one for α shown in Section 3.4.1 and is not repeated here.

Thus

$$\beta^{(k+1)} \,|\, \mathbf{V}^{(k)}, \mathbf{Y} \sim \mathcal{MVN}(\tilde{\mu}_\beta, \tilde{\mathbf{\Sigma}}_\beta) \tag{3.23}$$

where

$$\tilde{\mathbf{\Sigma}}_\beta^{-1} = \sum_{j=1}^{n_p} \sum_{i=1}^{n_r} \mathbf{Y}_{ji} \mathbf{Y}'_{ji} + \mathbf{\Sigma}_\beta^{-1} \tag{3.24}$$

$$\tilde{\mu}_\beta = \tilde{\mathbf{\Sigma}}_\beta \left(\sum_{j=1}^{n_p} \sum_{i=1}^{n_r} \mathbf{Y}_{ji} v_{ji}^{(k)} + \mathbf{\Sigma}_\beta^{-1} \mu_\beta \right). \tag{3.25}$$

3.4.3 Full Conditional Distribution for ξ

Based on the independence of the preference parameters given specific utilities $\mathbf{U}$ and $\mathbf{V}$, and specifically conditioning on the observed dyadwise covariate matches, which have been left implicit to this point,

$$\begin{aligned} p(\xi \,|\, \alpha, \beta, \zeta, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) &= p(\xi \,|\, \mathbf{U}, \mathbf{S}^R) \\ &\propto \pi(\xi)\, p(\mathbf{U} \,|\, \xi, \mathbf{S}^R) \end{aligned} \tag{3.26}$$

This is the format for a probit model using the multivariate normal as the conjugate prior: $\pi(\xi) = \mathcal{MVN}(\mu_\xi, \mathbf{\Sigma}_\xi)$. The derivation for the full conditional distribution is identical to the one for α shown in Section 3.4.1 and is not repeated here.

Thus

$$\xi^{(k+1)} \,|\, \mathbf{U}^{(k)}, \mathbf{S}^R \sim \mathcal{MVN}(\tilde{\mu}_\xi, \tilde{\mathbf{\Sigma}}_\xi) \tag{3.27}$$

where

$$\tilde{\boldsymbol{\Sigma}}_{\xi}^{-1} = \sum_{i=1}^{n_r} \sum_{j=n_p+1}^{n_p+n_c} \mathbf{S}_{ij}^{R} \mathbf{S}_{ij}^{R\prime} + \boldsymbol{\Sigma}_{\xi}^{-1} \tag{3.28}$$

$$\tilde{\mu}_{\xi} = \tilde{\boldsymbol{\Sigma}}_{\xi} \left(\sum_{i=1}^{n_r} \sum_{j=n_p+1}^{n_p+n_c} \mathbf{S}_{ij}^{R} u_{ij}^{(k)} + \boldsymbol{\Sigma}_{\xi}^{-1} \mu_{\xi} \right). \tag{3.29}$$

Note the change of the indices on the summation for self matches.

3.4.4 Full Conditional Distribution for ζ

Based on the independence of the preference parameters given specific utilities $\mathbf{U}$ and $\mathbf{V}$, and specifically conditioning on the observed dyadwise covariate matches, which have been left implicit to this point,

$$\begin{aligned} p(\zeta \,|\, \alpha, \beta, \xi, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) &= p(\zeta \,|\, \mathbf{V}, \mathbf{S}^P) \\ &\propto \pi(\zeta)\, p(\mathbf{V} \,|\, \zeta, \mathbf{S}^P) \end{aligned} \tag{3.30}$$

This is the format for a probit model using the multivariate normal as the conjugate prior: $\pi(\zeta) = \mathcal{MVN}(\mu_{\zeta}, \boldsymbol{\Sigma}_{\zeta})$. The derivation for the full conditional distribution is identical to the one for α shown in Section 3.4.1 and is not repeated here.

Thus

$$\zeta^{(k+1)} \,|\, \mathbf{V}^{(k)}, \mathbf{S}^P \sim \mathcal{MVN}(\tilde{\mu}_{\zeta}, \tilde{\boldsymbol{\Sigma}}_{\zeta}) \tag{3.31}$$

where

$$\tilde{\boldsymbol{\Sigma}}_{\zeta}^{-1} = \sum_{j=1}^{n_p} \sum_{i=n_r+1}^{n_r+1} \mathbf{S}_{ji}^{P} \mathbf{S}_{ji}^{P\prime} + \boldsymbol{\Sigma}_{\zeta}^{-1} \tag{3.32}$$

$$\tilde{\mu}_{\zeta} = \tilde{\boldsymbol{\Sigma}}_{\zeta} \left(\sum_{j=1}^{n_p} \sum_{i=n_r+1}^{n_r+1} \mathbf{S}_{ji}^{P} v_{ji}^{(k)} + \boldsymbol{\Sigma}_{\zeta}^{-1} \mu_{\zeta} \right). \tag{3.33}$$

Note the change of the indices on the summation for self matches.

3.4.5 Full Conditional Distribution for U

Unlike the preference parameters α, β, ξ, and ζ, the conditional distributions of $\mathbf{U}$ and $\mathbf{V}$ depend on each other. In particular, for $\mathbf{U}$, again explicitly stating the observed covariates $\mathbf{X}$ and self-matching structure $\mathbf{S}^R$:

$$p(\mathbf{U} \,|\, \alpha, \beta, \xi, \zeta, \mathbf{V}, m \in \mathcal{M}) = p(\mathbf{U} \,|\, \alpha, \xi, \mathbf{V}, \mathbf{X}, \mathbf{S}^R, m \in \mathcal{M}). \tag{3.34}$$

Recall that $u_{ij} = \alpha \mathbf{X}_{ij} + \xi \mathbf{S}_{ij}^R + \epsilon_{ij}$. Let $\mathbf{U}_{-\{ij\}}$ indicate all entries of $\mathbf{U}$ excluding the (i,j)th entry. Therefore assuming $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_{\epsilon}^2)$ (i.e., normal mean zero pairwise heterogeneity),

$$p(u_{ij} \,|\, \mathbf{U}_{-\{ij\}}, \alpha, \xi, \mathbf{V}, \mathbf{X}_{ij}, \mathbf{S}_{ij}^R, m \in \mathcal{M}) = \mathrm{N}(\alpha \mathbf{X}_{ij} + \xi \mathbf{S}_{ij}^R, \sigma_{\epsilon}^2) \tag{3.35}$$

constrained such that the updated $\mathbf{U}$ and $\mathbf{V}$ reproduce the observed matching via the matching mechanism in Algorithm 1.

Therefore, given $\alpha^{(k)}$ and $\xi^{(k)}$, new draws of $\mathbf{U}^{(k+1)}$ are obtained from truncated normal distributions where the constraints are based on $\mathbf{V}^{(k)}$, the rules of the sampling process, and the observed recruitment chain. Let $\psi(\mu, \sigma, a, b)$ be the truncated normal distribution with mean μ, variance σ^2, lower bound a, and upper bound b[4]. Then

$$u_{ij}^{(k+1)} \sim \psi(\alpha^{(k)}\mathbf{X}_{ij} + \xi^{(k)}\mathbf{S}_{ij}^{R}, \sigma_{\epsilon}^{2}, a_{\mathbf{U};ij}, b_{\mathbf{U};ij}) \tag{3.36}$$

where $a_{\mathbf{U};ij}$ and $b_{\mathbf{U};ij}$ enforce stability and that $\mathbf{U}$ and $\mathbf{V}$ return the observed recruitment chain via the matching mechanism in Algorithm 1. These constraints are shown in Table 3.3 and can be split into four types for recruiters i depending on whether $j \in \mathcal{P}$ (j is a peer) or $j \in \mathcal{S}_R$ (j represents a recruiter self match) and on whether or not i and j are tied in the observed recruitment chain.

Consider a recruiter $i \in \mathcal{R}$ and peer $j \in \mathcal{P}$ who are tied. We know that i must have a higher utility for j than for any self-match they do not select: $u_{ij} \geq \max(\mathcal{U}_{i,s^-})$, where $\mathcal{U}_{i,s}$ is the set of utilities i has for their self match opportunities (of size n_c), $\mathcal{U}_{i,s^+}$ is the set of utilities i has for self matches that were selected, and $\mathcal{U}_{i,s^-}$ is the set of utilities i has for self matches that were not selected. $\mathcal{U}_{i,s} =$

[4]Suppose $X \sim N(\mu, \sigma^2)$ has a normal distribution and lies within the interval $X \in (a, b), -\infty \leq a < b \leq \infty$. Then X conditional on $a < X < b$ has a truncated normal distribution. The probability density function for $a \leq x \leq b$ is given by

$$\psi(\mu, \sigma^2, a, b) = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the probability density function and the cumulative distribution function, of the standard normal distribution.

Dyad Type	**Match**	$a_{\mathbf{U};ij}$	$b_{\mathbf{U};ij}$
Recruiter-Peer	Yes	$\max(\mathcal{U}_{i,s^-} \cup \mathcal{U}_{i,h\in\mathcal{O}(i)})$	∞
Recruiter-Peer	No	$-\infty$	$\begin{cases}\min(\{u_{i,m(i)}\}) & \text{if } v_{ji} > v_{j,m(j)} \\ \infty & \text{otherwise}\end{cases}$
Recruiter-Self	Yes	$\max(\mathcal{U}_{i,s^-} \cup \mathcal{U}_{i,h\in\mathcal{O}(i)})$	∞
Recruiter-Self	No	$-\infty$	$\min(\{u_{i,m(i)}\})$

Table 3.3: Upper and lower bounds for truncated normal draws for u_{ij} where $\mathcal{U}_{i,h\in\mathcal{O}(i)}$ is the set of utilities i has for peers $h \neq j$ such that h is in i's opportunity set $\mathcal{O}(i)$. $\mathcal{O}(i)$ is the set of peers $h \in \mathcal{P}$, $h \notin m(i)$ such that $v_{hi} > v_{h,r(h)}$. $\mathcal{U}_{i,s^-}$ is the set of i's self-match utilities that i does not select. $m(i)$ is the set of peers that i is matched to. $m(j)$ is the recruiter of j.

$\mathcal{U}_{i,s^-} \cup \mathcal{U}_{i,s^+}$ and $\mathcal{U}_{i,s^-} \cap \mathcal{U}_{i,s^+} = \emptyset$. Precisely,

$$\mathcal{U}_{i,s} = \{u_{ih} : h \in \mathcal{S}_R\} \tag{3.37}$$

$$\mathcal{U}_{i,s^+} = \{u_{ih} : h \in \mathcal{S}_R, h \in m(i)\} \tag{3.38}$$

$$\mathcal{U}_{i,s^-} = \{u_{ih} : h \in \mathcal{S}_R, h \notin m(i)\} \tag{3.39}$$

Additionally, i must have a higher utility for j than they do for any peers who prefer i over their own match. These are the peers $h \neq j$, $h \notin m(i)$ that i has in their opportunity set $\mathcal{O}(i)$:

$$\mathcal{O}(i) = \{h : v_{hi} \geq v_{h,m(h)}; h \in \mathcal{P}, h \notin m(i)\}. \tag{3.40}$$

Therefore, $u_{ij} \geq \max(\mathcal{U}_{i,h\in\mathcal{O}(i)})$ where $\mathcal{U}_{i,h\in\mathcal{O}(i)} = \{u_{ih} : h \in \mathcal{O}(i)\}$. These two bounds, required to satisfy stability constraints, result in $a_{\mathbf{U};ij} = \max(\mathcal{U}_{i,s^-} \cup$

$\mathcal{U}_{i,h\in\mathcal{O}(i)}$). Since i and j are matched, u_{ij} has no upper bound so $b_{\mathbf{U};ij} = \infty$. This is the scenario in the first row of Table 3.3.

The constraints are the same if $j \in \mathcal{S}_R$ and there is match (j is a self-match that is selected). This is the scenario in the third row of Table 3.3.

When i and j are not matched and $j \in \mathcal{P}$ (j is a peer), the utility that i has for j must be less than the utility i has for any of their matches if j prefers i over their own match. This is the scenario in the second row of Table 3.3. In particular, $u_{ij} \leq \min(\{u_{i,m(i)}\})$ if $v_{ji} > v_{j,m(j)}$. Therefore, $a_{\mathbf{U};ij} = -\infty$ and $b_{\mathbf{U};ij}$ is either $\min(\{u_{i,m(i)}\})$ or ∞. If peer j prefers their own match over recruiter i, the utility that i has for j is irrelevant, so the draw is unconstrained.

Finally, when i and j are not matched and $j \in \mathcal{S}_R$ (j is a self-match), the utility that i has for j must be less than the utility i has for any of their matches. This is the scenario in the fourth row of Table 3.3. In particular, $u_{ij} \leq \min(\{u_{i,m(i)}\})$. Thus, $a_{\mathbf{U};ij} = -\infty$ and $b_{\mathbf{U};ij} = \min(\{u_{i,m(i)}\})$.

Based on these constraints, $u_{ij}^{(k+1)} \sim \psi(\alpha^{(k)}\mathbf{X}_{ij} + \xi^{(k)}\mathbf{S}_{ij}^R, \sigma_\epsilon^2, a_{\mathbf{U};ij}, b_{\mathbf{U};ij})$. Figure 3.3 depicts how the constraints of the stable matching affect the normal distributions that updated elements $u_{ij}^{(k+1)}$ are drawn from. For all draws of $u_{ij}^{(k+1)}$, the normal distribution shown is $\psi(\alpha^{(k)}\mathbf{X}_{ij} + \xi^{(k)}\mathbf{S}_{ij}^R, \epsilon_{ij}, a_{\mathbf{U};ij}, b_{\mathbf{U};ij})$.

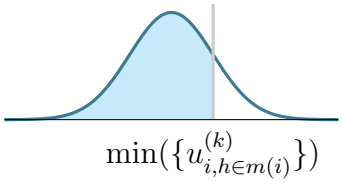
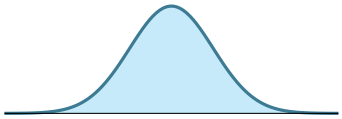
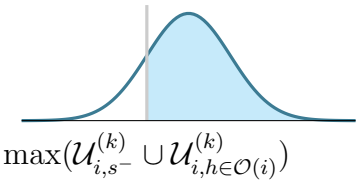
Distribution	Use for $u_{ij}^{(k+1)}$ when
	i is not matched to j and $j \in \mathcal{S}_R$; or i is not matched to j and j has a higher utility for i than for their own match
	i is not matched to j and j has a higher utility for their own match
	i and j are matched

Figure 3.3: Illustration of distributions from which new values of $u_{ij}^{(k+1)}$ are drawn, based on recruitment chain constraints.

3.4.6 Full Conditional Distribution for V

For $\mathbf{V}$, again explicitly stating the observed covariates $\mathbf{Y}$ and self-matching structure $\mathbf{S}^P$:

$$p(\mathbf{V} \mid \alpha, \beta, \xi, \zeta, \mathbf{U}, m \in \mathcal{M}) = p(\mathbf{V} \mid \beta, \zeta, \mathbf{U}, \mathbf{Y}, \mathbf{S}^P, m \in \mathcal{M}) \tag{3.41}$$

Recall that $v_{ij} = \beta \mathbf{Y}_{ji} + \zeta \mathbf{S}_{ji}^P + \gamma_{ji}$. Let $\mathbf{V}_{-\{ji\}}$ indicate all entries of $\mathbf{V}$ excluding the (j,i)th entry. Therefore assuming $\gamma_{ji} \sim \mathrm{N}(0, \sigma_\gamma^2)$ (i.e., normal mean zero pairwise heterogeneity),

$$p(v_{ji} \mid \mathbf{V}_{-\{ji\}}, \beta, \zeta, \mathbf{U}, \mathbf{Y}_{ji}, \mathbf{S}_{ji}^P, m \in \mathcal{M}) = \mathrm{N}(\beta \mathbf{Y}_{ji} + \zeta \mathbf{S}_{ji}^P, \sigma_\gamma^2) \tag{3.42}$$

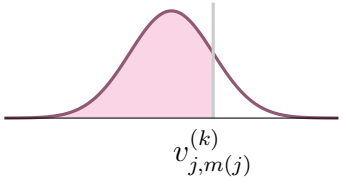
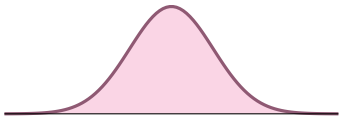
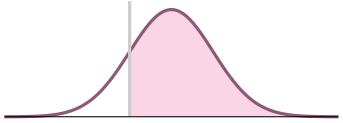
Distribution	**Use for $v_{ji}^{(k+1)}$ when**
	j is not matched to i and $i \in \mathcal{S}_P$; or j is not matched to i and i has a higher utility for j than for their own match
	j is not matched to i and i has a higher utility for their own match
	j and i are matched and $i \in \mathcal{R}$
	j and i are matched and $i \in \mathcal{S}_P$

Figure 3.4: Illustration of distributions from which new values of $v_{ji}^{(k+1)}$ are drawn, based on recruitment chain constraints.

constrained such that $\mathbf{U}$ and the updated $\mathbf{V}$ reproduce the observed matching via the matching mechanism in Algorithm 1.

Therefore, given $\beta^{(k)}$ and $\zeta^{(k)}$, new draws of $\mathbf{V}^{(k+1)}$ are obtained from truncated normal distributions where the constraints are based on $\mathbf{U}^{(k)}$, the rules of the sampling process, and the observed recruitment chain. Let $\psi(\mu, \sigma, a, b)$ be the truncated normal distribution with mean μ, variance σ^2, lower bound a, and upper bound b. Then

$$v_{ji}^{(k+1)} \sim \psi(\beta^{(k)}\mathbf{Y}_{ji} + \zeta^{(k)}\mathbf{S}_{ji}^{P}, \sigma_{\gamma}^{2}, a_{\mathbf{V};ji}, b_{\mathbf{V};ji}) \tag{3.43}$$

where $a_{\mathbf{V};ji}$ and $b_{\mathbf{V};ji}$ enforce stability and that $\mathbf{U}$ and $\mathbf{V}$ return the observed recruitment chain via the matching mechanism in Algorithm 1. These constraints are shown in Table 3.4 and can be split into four types for peers j depending on whether $i \in \mathcal{R}$ (i is a recruiter) or $i \in \mathcal{S}_P$ (i represents a peer self match) and on whether or not j and i are tied in the observed recruitment chain. Note that these constraints are similar to those for draws of $\mathbf{U}^{(k+1)}$, but differ because of the many-to-one nature of matching.

Consider a peer $j \in \mathcal{P}$ and recruiter $i \in \mathcal{R}$ who are tied. We know that j must have a higher utility for i than their self match: $v_{ji} \geq v_{js}$. Additionally, j must have a higher utility for i than they do for any recruiters who prefer j over their own match. These are the peers $\ell \neq i$, $\ell \neq m(j)$ that j has in their opportunity

Dyad Type	**Match**	$a_{\mathbf{V};ji}$	$b_{\mathbf{V};ji}$
Peer-Recruiter	Yes	$\max(v_{j,s} \cup \mathcal{V}_{j,\ell\in\mathcal{O}(j)})$	∞
Peer-Recruiter	No	$-\infty$	$\begin{cases} v_{j,m(j)} & \text{if } u_{ij} > \min(\{u_{i,m(i)}\}) \\ \infty & \text{otherwise} \end{cases}$
Peer-Self	Yes	$\max(\mathcal{V}_{j,\ell\in\mathcal{O}(j)})$	∞
Peer-Self	No	$-\infty$	$v_{j,m(j)}$

Table 3.4: Upper and lower bounds for truncated draws of v_{ji} where $\mathcal{V}_{j,\ell\in\mathcal{O}(j)}$ is the set of utilities j has for recruiters $\ell \neq i$ such that ℓ is in j's opportunity set $\mathcal{O}(j)$. $\mathcal{O}(j)$ is the set of recruiters $\ell \in \mathcal{R}$, $\ell \neq m(j)$ such that $u_{\ell j} > \min(\{u_{\ell,m(\ell)}\})$. v_{js} is j's self-match utility. $m(i)$ is the set of peers that i is matched to. $m(j)$ is the recruiter of j.

set $\mathcal{O}(j)$:

$$\mathcal{O}(j) = \{\ell : u_{\ell j} \geq \min(\{u_{\ell,m(\ell)}\}); \ell \in \mathcal{R}, \ell \neq m(j)\}. \tag{3.44}$$

Therefore, $v_{ji} \geq \max(\mathcal{V}_{j,\ell\in\mathcal{O}(j)})$ where $\mathcal{V}_{j,\ell\in\mathcal{O}(j)} = \{v_{j\ell} : \ell \in \mathcal{O}(j)\}$. These two bounds, required to satisfy stability constraints, result in $a_{\mathbf{V};ji} = \max(v_{js} \cup \mathcal{V}_{j,\ell\in\mathcal{O}(j)})$. Since i and j are matched, v_{ji} has no upper bound so $b_{\mathbf{V};ji} = \infty$. This is the scenario in the first row of Table 3.4.

If $i \in \mathcal{S}_P$ and there is match (j self matches), then $v_{ji} \geq \max(\mathcal{V}_{j,\ell\in\mathcal{O}(j)})$. This results in $a_{\mathbf{V};ji} = \max(\mathcal{V}_{j,\ell\in\mathcal{O}(j)})$. Since i and j are matched, v_{ji} has no upper bound so $b_{\mathbf{V};ji} = \infty$. This is the scenario in the third row of Table 3.4.

When i and j are not matched and $i \in \mathcal{R}$ (i is a recruiter), the utility that j has for i must be less than the utility j has for their match if i prefers j over their own match. This is the scenario in the second row of Table 3.4. In particular, $v_{ji} \leq v_{j,m(j)}$ if $u_{ij} > \min(\{u_{i,m(i)}\})$. Therefore, $a_{\mathbf{V};ji} = -\infty$ and $b_{\mathbf{V};ji}$ is either

$v_{j,m(j)}$ or ∞. If recruiter i prefers their own match over peer j, the utility that j has for i is irrelevant, so the draw is unconstrained.

Finally, when i and j are not matched and $i \in \mathcal{S}_R$ (i is a self match), the utility that j has for i must be less than the utility j has for their match. This is the scenario in the fourth row of Table 3.4. In particular, $v_{ji} \leq v_{j,m(j)}$. Thus, $a_{\mathbf{V};ji} = -\infty$ and $b_{\mathbf{V};ji} = v_{j,m(j)}$.

Based on these constraints, $v_{ji}^{(k+1)} \sim \psi(\beta^{(k)}\mathbf{Y}_{ji} + \zeta^{(k)}\mathbf{S}_{ji}^{P}, \sigma_{\gamma}^{2}, a_{\mathbf{V};ji}, b_{\mathbf{V};ji})$. Figure 3.4 depicts how the constraints of the stable matching affect the normal distributions that updated elements $v_{ji}^{(k+1)}$ are drawn from. For all draws of $v_{ji}^{(k+1)}$, the normal distribution shown is $\psi(\beta^{(k)}\mathbf{Y}_{ji} + \zeta^{(k)}\mathbf{S}_{ji}^{P}, \gamma_{ji}, a_{\mathbf{V};ji}, b_{\mathbf{V};ji})$.

■

Practically for the implementation of the Gibbs sampler, for each iteration k to update $\mathbf{U}^{(k+1)}$ and $\mathbf{V}^{(k+1)}$, a random (i, j) pair is drawn. Then either u or v is chosen with equal probability to be updated first. If u is chosen, $u_{ij}^{(k+1)}$ and then $v_{ji}^{(k+1)}$ are updated from truncated normal distributions according to constraints detailed above. If v is chosen, $v_{ji}^{(k+1)}$ and then $u_{ij}^{(k+1)}$ are updated from the truncated normal distributions. This process is done for all (i, j) pairs before the sampler returns to updating the preference coefficients α, β, ξ, and ζ. After all pairs u_{ij} and v_{ji} have been updated at step k, a check is carried out to guarantee that the $\mathbf{U}^{(k+1)}$ and $\mathbf{V}^{(k+1)}$ return the observed recruitment chain via Algorithm 1.

3.5 Gibbs Sampler Overview

For a desired number of iterations K the Markov chain is updated as described in Algorithm 2. Note that the procedure described in Section 3.4 is only for one wave (that is, for one bipartite network between one set of recruiters and one set of peers). In practice, we observe multiple waves of recruitment, and thus we sequentially apply the method where the α, β, ξ, and ζ from wave $w-1$ during iteration (k) are used as the starting values of α, β, ξ, and ζ during wave w of iteration (k). Values of α, β, ξ, and ζ for wave 1 of iteration $(k+1)$ are informed by wave n_w of iteration (k). Algorithm 2 includes this sequential application.

Inputs to the inference function are: the observed recruitment chain m; the observed covariate array $\mathbf{Z}$; priors μ_α, Σ_α, μ_β, Σ_β, μ_ξ, Σ_ξ, μ_ζ, and Σ_ζ for the preference coefficients; and the desired number of iterations K.

3.6 Model Selection

Values of α and β can be examined to determine whether the corresponding covariates seem to impact recruitment. In particular, the posterior probability that the coefficient is above zero for homophilous recruitment will be close to 1. Although there are no set cut-off points for when a covariate is 'important' for recruitment, potential comparison values are 0.9 and 0.95. Thus, for example, if the posterior probability that α is greater than zero is 0.93, we would say that this covariate seems to impact recruitment. For heterophilous recruitment, we

Algorithm 2 Gibbs sampler for $\pi(\alpha, \beta, \xi, \zeta, \mathbf{U}, \mathbf{V} \,|\, m \in \mathcal{M})$

initialize $\alpha_0^{(0)}$, $\beta_0^{(0)}$, $\xi_0^{(0)}$, and $\zeta_0^{(0)}$
  $\alpha_0^{(0)} \sim \mathcal{MVN}(\mu_\alpha, \Sigma_\alpha)$
  $\beta_0^{(0)} \sim \mathcal{MVN}(\mu_\beta, \Sigma_\beta)$
  $\xi^{(0)} \sim \mathcal{MVN}(\mu_\xi, \Sigma_\xi)$
  $\zeta_0^{(0)} \sim \mathcal{MVN}(\mu_\zeta, \Sigma_\zeta)$
for $w = 1 \ldots, n_w$ **do**
  initialize ϵ_w and γ_w
    $\epsilon_w[i, j] \sim N(\mu_\epsilon, \sigma_\epsilon^2)$
    $\gamma_w[j, i] \sim N(\mu_\gamma, \sigma_\gamma^2)$
  calculate initial values $\mathbf{U}_w^{(0)}$ and $\mathbf{V}_w^{(0)}$
    $u_w^{(0)}[i, j] \,|\, \alpha^{(0)}, \xi^{(0)}, \mathbf{U}_{w;-\{ij\}}, \mathbf{V}_w, m \in \mathcal{M} = \alpha_{w-1}^{(0)} \mathbf{X}_{w;ij} + \xi_{w-1}^{(0)} \mathbf{S}_{w;ij}^R + \epsilon_w[i, j]$
    $v_w^{(0)}[i, j] \,|\, \beta^{(0)}, \zeta^{(0)}, \mathbf{U}_w, \mathbf{V}_{w;-\{ji\}}, m \in \mathcal{M} = \beta_{w-1}^{(0)} \mathbf{Y}_{w;ji} + \zeta_{w-1}^{(0)} \mathbf{S}_{w;ji}^P + \gamma_w[j, i]$
  draw initial values $\alpha_w^{(1)}$, $\beta_w^{(1)}$, $\xi^{(1)}$, and $\zeta^{(1)}$
    $\alpha_w^{(1)} \,|\, \mathbf{U}_w^{(0)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\alpha^{(1)}, \tilde{\Sigma}_{\alpha;w})$
    $\beta_w^{(1)} \,|\, \mathbf{V}_w^{(0)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\beta^{(1)}, \tilde{\Sigma}_{\beta;w})$
    $\xi^{(1)} \,|\, \mathbf{U}_w^{(0)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\xi^{(1)}, \tilde{\Sigma}_{\xi;w})$
    $\zeta^{(1)} \,|\, \mathbf{V}_w^{(0)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\zeta^{(1)}, \tilde{\Sigma}_{\zeta;w})$
  $k = 1$
end for
while $k < K$ **do**
  for $w = 1, \ldots, n_w$ **do**
    check = 0
    while check = 0 **do**
      update $\mathbf{U}_w^{(k+1)}$ and $\mathbf{V}_w^{(k+1)}$
        $u_w^{(k+1)}[i, j] \,|\, \alpha^{(k)}, \xi^{(k)}, \mathbf{U}_{w;-\{ij\}}, \mathbf{V}_w, m \in \mathcal{M}$
          $\sim \psi(\alpha_{w-1}^{(k)} \mathbf{X}_{w;ij} + \xi_{w-1}^{(k)} \mathbf{S}_{w;ij}^R, \sigma_\epsilon^2, a_{\mathbf{U};w;ij}^{(k+1)}, b_{\mathbf{U};w;ij}^{(k+1)})$
        $v_w^{(k+1)}[j, i] \,|\, \beta^{(k)}, \zeta^{(k)}, \mathbf{U}_w, \mathbf{V}_{w;-\{ji\}}, m \in \mathcal{M}$
          $\sim \psi(\beta_{w-1}^{(k)} \mathbf{Y}_{w;ji} + \zeta_{w-1}^{(k)} \mathbf{S}_{w;ji}^P, \sigma_\gamma^2, a_{\mathbf{V};w;ji}^{(k+1)}, b_{\mathbf{V};w;ji}^{(k+1)})$
      if $\mathbf{U}_w^{(k+1)}$ and $\mathbf{V}_w^{(k+1)}$ reproduce the observed matching **then**
        check = 1
      end if
    end while
    update $\alpha_w^{(k+1)}$ and $\beta_w^{(k+1)}$
      $\alpha_w^{(k+1)} \,|\, \mathbf{U}_w^{(k)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\alpha^{(k+1)}, \tilde{\Sigma}_{\alpha;w})$
      $\beta_w^{(k+1)} \,|\, \mathbf{V}_w^{(k)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\beta^{(k+1)}, \tilde{\Sigma}_{\beta;w})$
      $\xi^{(k+1)} \,|\, \mathbf{U}_w^{(k)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\xi^{(k+1)}, \tilde{\Sigma}_{\xi;w})$
      $\zeta^{(k+1)} \,|\, \mathbf{V}_w^{(k)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\zeta^{(k+1)}, \tilde{\Sigma}_{\zeta;w})$
  end for
  $k = k + 1$
end while

would consider the posterior probability that α and β are less than zero. In order to select a final model (i.e., a final set of covariates upon which participants recruit preferentially), a larger set of covariates can be pruned by considering their posterior probability above (or below) zero.

To compare whether sets of covariates impact preferential recruitment, we could compare them to the null model where $\alpha = 0$ and $\beta = 0$ for all covariates. Then *Bayes factors* can be used to compare the models (Gelman et al., 2004). The Bayes factor is:

$$B_{10} = \frac{p(m \in \mathcal{M} \,|\, \alpha_1, \beta_1, \xi, \zeta)}{p(m \in \mathcal{M} \,|\, \alpha_0, \beta_0, \xi, \zeta)} \tag{3.45}$$

where α_1 and β_1 refer to the covariate coefficients from the preferential recruitment model, and α_0 and β_0 refer to the covariate coefficients from the null model. Values of the Bayes factor greater than 1 indicate that the preferential recruitment model is more strongly supported by the observed recruitment chain than the null model. In particular, Kass and Raftery (1995) discuss guidelines for the interpretation of B_{10}, noting that twice the natural logarithm of the Bayes factor (i.e., $2\log_e(B_{10})$) is on the same scale as the familiar deviance and likelihood ratio test statistics. By this metric, values of $2\log_e(B_{10})$ from 6 to 10 provide strong evidence against the null model, and values greater than 10 provide very strong evidence.

This method can also be used to compare different sets of covariates as a method of model selection. In this case, instead of the null model, consider an

alternative model with covariate coefficients α_2 and β_2. Then:

$$B_{12} = \frac{p(m \in \mathcal{M} \,|\, \alpha_1, \beta_1, \xi, \zeta)}{p(m \in \mathcal{M} \,|\, \alpha_2, \beta_2, \xi, \zeta)} \tag{3.46}$$

where values of the Bayes factor greater than 1 indicate that the first model is more strongly supported by the observed recruitment chain. Values less than 1 indicate that the second model is more strongly supported by the observed recruitment chain. The same guidelines from Kass and Raftery (1995) can be used to evaluate B_{12}. For model selection, consider either adding or pruning a covariate from an existing set. The new model can then be compared to the old model, and the one more strongly supported by the data retained.

3.7 Generating Model-Based Inclusion Probabilties

Section 3.4 detailed inference for preference parameters α, β, ξ, and ζ using the RCPR model. Section 3.5 described how to apply this procedure sequentially over multiple waves of recruitment. Although the values of the preference parameters are of interest, and in particular noting which values are far from zero to determine which covariates people recruit preferentially on, the key outcome of RDS studies is usually an estimate of population prevalence for HIV or related diseases (Heckathorn, 1997).

Different estimators exist, as discussed in Section 1.3. Prevalence estimates are often obtained using design-based inference relying on a person's self-reported

network size as a proxy for their inclusion probability. For example, the commonly used Volz-Heckathorn (VH) estimator for population prevalence is:

$$\hat{\mu}_{\text{VH}} = \frac{\sum_{i \in s} \frac{y_i}{\tilde{d}_i}}{\sum_{i \in s} \frac{1}{\tilde{d}_i}} \tag{3.47}$$

where $i \in s$ indicates that person i is in the sample s, y_i is person i's outcome measure (e.g., HIV status), and $\tilde{d}_i$ is person i's self-reported degree. This is a generalized Horvitz-Thompson estimator in which $\tilde{d}_i \propto \pi_i$, where π_i is the inclusion probability of person i.

The RCPR model can be used to directly generate inclusion probabilities using values for the preference parameters obtained via inference. These new inclusion probabilities depend on covariates affecting preferential recruitment (potentially including network size), rather than just self-reported network size. Additionally, using the generative model framework allows for the exact calculation of second-order inclusion probabilities π_{ij}, $i \neq j$, $i, j \in s$. Common RDS estimation techniques rely on bootstrap methods to calculate π_{ij}, which are required for the variance of the estimator. The procedure to generate these inclusion probabilities from the RCPR is described below.

Let s_0 be the observed sample, where $i \in s_0$ indicates that person i was in the observed sample. Let $n = |s_0|$ be the observed sample size. Let N be the size of the population, which may be approximated using a method such as successive sampling-population size estimation (SS-PSE) if it is unknown (Handcock et al.

(2014), Handcock et al. (2015))[5]. Then a standard RDS method such as the successive sampling (SS) estimator can be used to calculate weights to scale up the observed sample to be of size N. This is done by replicating individuals proportional to their sample weight. Note that this is a very approximate procedure, as the sample weights are based just on the network size variable and the goal is to get more sophisticated sample weights. However this method is acceptable as a rough means of constructing a population when that population is unknown.

1. Generate a population of size N by replicating covariate values of individuals $i \in s_0$. n of the N people must match those in the sample s_0.
2. Draw α, β, ξ, and ζ from their posterior distributions.
3. Simulate a recruitment chain s from the generated population using α, β, ξ, ζ and the same seeds as in s_0.
4. For each $i \in s_0$, record whether $i \in s$.
5. For each $(i, j) \in s_0$, record whether both $i \in s$ and $j \in s$.

Let D be the number of new recruitment chains drawn, and let $s_d, d = 1 \ldots, D$ denote one of these recruitment chains. Let c_i be the number of recruitment chains $i \in s_0$ appeared in: $c_i = \sum_{d=1}^{D} \mathbb{1}(i \in s_d)$. Let c_{ij} be the number of recruitment chains $i, j \in s_0$ both appeared in: $c_{ij} = \sum_{d=1}^{D} \mathbb{1}(i \in s_d)\mathbb{1}(j \in s_d)$. Then:

$$\tilde{\pi}_i = \frac{c_i}{D} \tag{3.48}$$

$$\tilde{\pi}_{ij} = \frac{c_{ij}}{D} \tag{3.49}$$

[5]SS-PSE requires only information routinely collected as part of an RDS study. A population size estimation may also be available using a unique object or service multiplier, or a network scale-up method. These methods are commonly performed in conjunction with RDS studies.

where $\tilde{\pi}_i$ is the estimated first-order inclusion probability of individual i and $\tilde{\pi}_{ij}$ is the estimated second-order inclusion probability of (i, j).

Using the generalized Horvitz-Thompson estimator, the RCPR-based prevalence estimate is then:

$$\hat{\mu}_{\text{RC}} = \frac{\sum_{i\{\in s_0 \setminus f\}} \frac{y_i}{\tilde{\pi}_i}}{\sum_{i \in \{s_0 \setminus f\}} \frac{1}{\tilde{\pi}_i}} \tag{3.50}$$

where f are the indices of the seeds. Because recruitment chains are always generated from the same seeds, $\tilde{\pi}_{i \in f} = 1$ so these values are excluded from the calculation.

An estimate for the variance of this estimator is:

$$\widehat{\text{var}}(\hat{\mu}_{\text{RC}}) = \frac{1}{N^2} \left[\sum_{i \in \{s_0 \setminus f\}} \left(\frac{1 - \tilde{\pi}_i}{\tilde{\pi}_i^2} \right) (y_i - \hat{\mu}_{\text{RC}})^2 \right. \tag{3.51}$$
$$\left. + \sum_{i \in \{s_0 \setminus f\}} \sum_{\substack{j \neq i, \\ j \in \{s_0 \setminus f\}}} \left(\frac{\tilde{\pi}_{ij} - \tilde{\pi}_i \tilde{\pi}_j}{\tilde{\pi}_i \tilde{\pi}_j} \right) \frac{(y_i - \hat{\mu}_{\text{RC}})(y_j - \hat{\mu}_{\text{RC}})}{\tilde{\pi}_{ij}} \right]$$

■

This chapter described the two-sided rational-choice preferential recruitment (RCPR) model and fully specified it for inference of the preference parameters governing the matching. This work builds on the foundation of two-sided matching models and the one-to-one revealed preferences model for the marriage market. Extensions for application to the RDS scenario include parameters to govern recruitment robustness, the many-to-one matching mechanism with multiple self-

match utilities, inference for a many-to-one scenario, and the sequential application of the model over multiple waves of recruitment. In addition, the procedure to use the RCPR model to generate inclusion probabilities used to calculate population prevalences for outcome measures like HIV is detailed. The performance of the RCPR model is analyzed in Chapter 4 by demonstrating that the inference procedure can recover the original values of the preference parameters used to set up simulation. The RCPR model is applied to several RDS studies of populations at high risk for HIV/AIDS in Chapter 5.

CHAPTER 4

Simulation Studies

To demonstrate the performance of the RCPR model, I perform a variety of simulation studies. In each simulation, a recruitment chain is generated from a known population using the preferential recruitment model with user-specified values for the preference coefficients (α and β) and the self-matching coefficients (ξ and ζ). This recruitment chain is assumed to be the 'observed' sample. The RCPR model is then used on this recruitment chain to make inferences about α, β, ξ, and ζ, with the goal of returning the original values used as inputs to the simulation. For α and β, being able to precisely return the original values is important for determining which covariates recruiters and peers use to preferentially make decisions. For ξ and ζ, the precise values do not matter; all that matters is being able to generate recruitment chains with the same distribution of number of recruits as the original recruitment chain. This is due to potential non-identifiability (discussed further in Section 4.2) among the self-matching parameters.

Input parameters for the simulations are shown in Table 4.1. These parameters can be varied to obtain a wide range of recruitment chains, encompassing many different types we observe in real data. Examples are shown in Section 4.1. I detail

Variable	Description
N	Population size
n_v	Number of covariates
$\mathbf{Z}$	$N \times n_v$ matrix of covariate values for each member of the population
n_s	Number of seeds
n_w	Number of waves
n_c	Number of coupons
α	Preference coefficients for recruiters to peers
β	Preference coefficients for peers to recruiters
ξ	Preference coefficients for recruiter self-matching
ζ	Preference coefficients for peer self-matching

Table 4.1: Simulation input parameters.

the inference procedure in a variety of simulation studies, starting with one-to-one matchings in Section 4.2 and extending to many-to-one matchings in Section 4.3. All simulations were carried out using the `prefrecruit` package in the `R` programming language, which I developed as part of my dissertation. The functions are incorporated with those developed as part of the `RDS` package (Handcock et al., 2012).

For a population of size N and for each of $Z_1, \ldots, Z_{n_v}$ covariates, assume that all values $z_{g;c}$, $g \in \{1 \ldots, N\}$ and $c \in \{1, \ldots, n_v\}$ are known. Unless otherwise noted, assume that $z_{g;c} \sim$ Bernoulli(0.5). These covariate values form the matrix $\mathbf{Z}_{N \times n_v}$.

From the list of populations members, n_s seeds are randomly selected, and all other members of the population are their eligible peers. In this first wave, the number of recruiters is $n_{r,w=1} = n_s$ and $n_{p,w=1} = N - n_s$. The matrices $\mathbf{X}$ and $\mathbf{Y}$ can then be constructed.

Recall that $\mathbf{X}$ and $\mathbf{Y}$ are indicator arrays of dyadwise matches on observed

covariates and $\mathbf{S}^R$ and $\mathbf{S}^P$ are indicator arrays of self match opportunity. Using the input preference coefficients α and β, and the input self-matching coefficients ξ and ζ, individual entries of the utility matrices $\mathbf{U}$ and $\mathbf{V}$ are calculated directly as

$$u_{ij} = \alpha \mathbf{X}_{ij} + \xi \mathbf{S}^R_{ij} + \epsilon_{ij} \tag{4.1}$$

$$v_{ji} = \beta \mathbf{Y}_{ji} + \zeta \mathbf{S}^P_{ji} + \gamma_{ji} \tag{4.2}$$

where $\epsilon_{ij} \sim \mathrm{N}(0,1)$ and $\gamma_{ji} \sim \mathrm{N}(0,1)$ (standard normal unobserved pairwise heterogeneity).

The values of ξ and ζ can then be thought of as additive shifts in the utilities recruiters and peers have for themselves. The utility that a recruiter i has for a peer j will be drawn from $\mathrm{N}(\alpha \mathbf{X}_{ij}, 1)$, where $\alpha \mathbf{X}_{ij}$ is the sum of values of α for which the recruiter and peer share the corresponding covariate level. The self-match values are drawn from $\mathrm{N}(\xi, 1)$, with the element of ξ used corresponding to the self-match number. Therefore, the value of ξ in relation to $\alpha \mathbf{X}$, rather than its absolute value, is what controls recruitment robustness. Values of ξ that are large in relation to $\alpha \mathbf{X}$ will result in many self matches, while values of ξ that are small in relation to $\alpha \mathbf{X}$ will result in few self matches. The same relationships are true for ζ and $\beta \mathbf{Y}$.

4.1 Simulation Examples: Recruitment Chain Variety

The input parameters for the simulations shown in Table 4.1 can be varied to obtain a wide range of recruitment chains, encompassing many different types we observe in real data. Some examples are given in this section. For each simulation, the input parameters are given, and the resulting recruitment chain, sample size, distribution of the number of recruits, and recruitment homophily are provided.

Robust but not preferential recruitment Figure 4.1 shows an example of robust but not preferential recruitment. Here, for the population $N = 2000$, $n_v = 1$, and $z_g \sim$ Bernoulli(0.5). For the simulation, $n_s = 2$, $n_w = 4$, and $n_c = 3$. To indicate no preferential recruitment, set $\alpha = 0$ and $\beta = 0$. To indicate robust recruitment (i.e., a tendency for each recruiter to distribute their full $n_c = 3$ coupons and peers to accept some coupon), set $\xi = (-2, -2, -2)$ and $\zeta = -2$.

The sample size is $n = 242$, with each of the 80 recruiters distributing their full three coupons successfully. The recruitment homophily is $\mathcal{H}_R = 0.942$. This value is close to 1, so we believe there is not strong preferential recruitment.

Robust and homophilous preferential recruitment Figure 4.2 shows an example of robust preferential recruitment with strong homophily on the covariate. Here, for the population $N = 2000$, $n_v = 1$, and $z_g \sim$ Bernoulli(0.5) as in the previous example. For the simulation, $n_s = 2$, $n_w = 4$, and $n_c = 3$, again the same as in the previous example. To indicate homophilous preferential recruitment, set

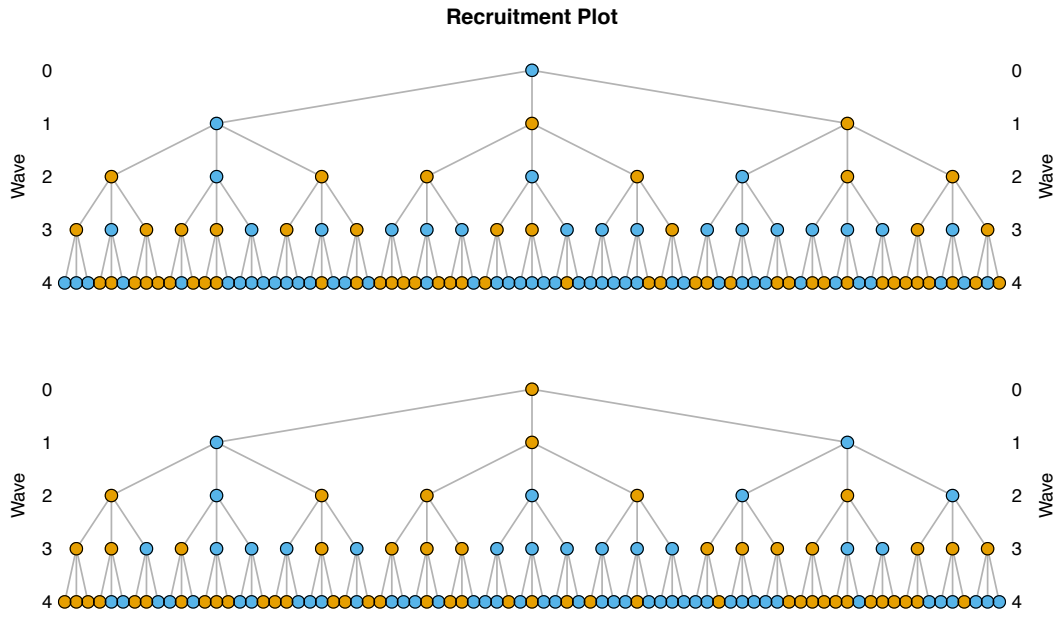


Figure 4.1: Simulation with robust but not preferential recruitment with $n_s = 2$, $n_w = 4$, $\alpha = 0$, $\beta = 0$, $\xi = (-2, -2, -2)$, and $\zeta = -2$. The color of the node represents the covariate value.

$\alpha = 2$ and $\beta = 2$. To indicate robust recruitment, set $\xi = (-2, -2, -2)$ and $\zeta = -2$, as in the previous example.

The sample size is $n = 242$, with each of the 80 recruiters distributing their full three coupons successfully. The recruitment homophily is $\mathcal{H}_R = 2.00$. This value is much bigger than 1, so we believe there is strong homophilous preferential recruitment. This is backed up by the recruitment graph, where each seed initiates a chain where everyone has the same value as themself.

Robust and heterophilous preferential recruitment Figure 4.3 shows an example of robust preferential recruitment with strong heterophily on the covariate. Here, for the population $N = 2000$, $n_v = 1$, and $z_g \sim \text{Bernoulli}(0.5)$ as in the

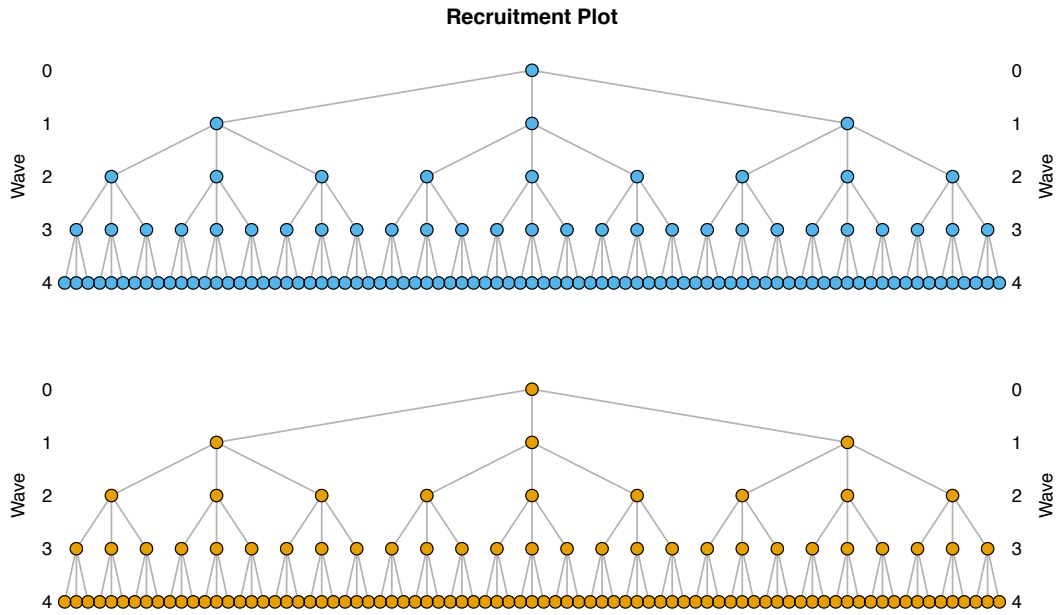


Figure 4.2: Simulation with robust and homophilous preferential recruitment with $n_s = 2$, $n_w = 4$, $\alpha = 2$, $\beta = 2$, $\xi = (-2, -2, -2)$, and $\zeta = -2$. The color of the node represents the covariate value.

previous examples. For the simulation, $n_s = 2$, $n_w = 4$, and $n_c = 3$, again the same as in the previous examples. To indicate heterophilous preferential recruitment, set $\alpha = -2$ and $\beta = -2$. To indicate robust recruitment, set $\xi = (-2, -2, -2)$ and $\zeta = -2$, as in the previous example.

The sample size is $n = 242$, with each of the 80 recruiters distributing their full three coupons successfully. The recruitment homophily is $\mathcal{H}_R = 0.00$. This value is much smaller than 1, so we believe there is strong heterophilous preferential recruitment. This is backed up by the recruitment graph, where each recruiter selects someone with the opposite covariate value as themself so the waves alternate in color.

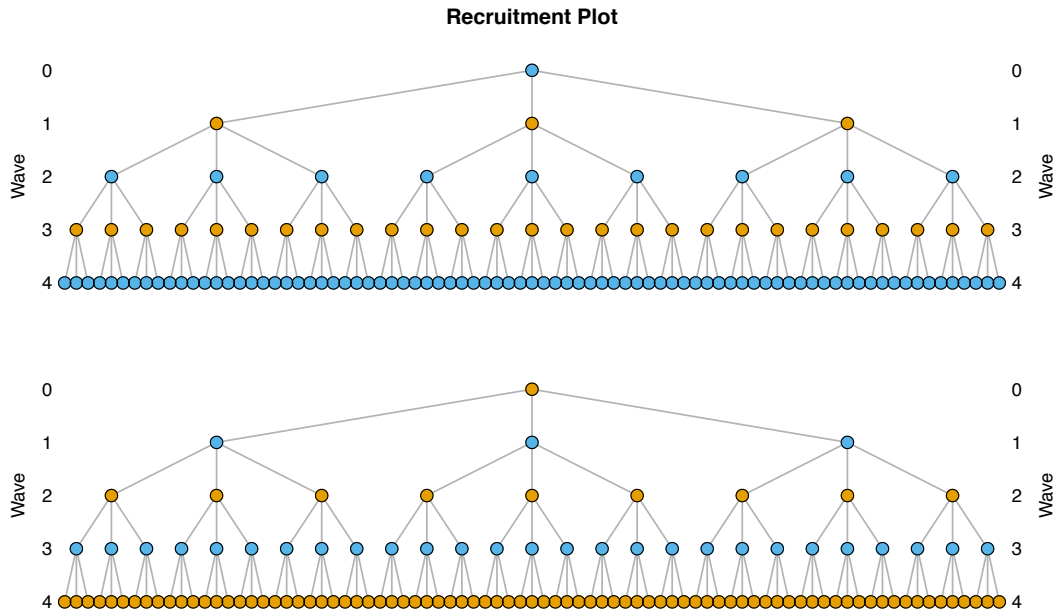


Figure 4.3: Simulation with robust and heterophilous preferential recruitment with $n_s = 2$, $n_w = 4$, $\alpha = -2$, $\beta = -2$, $\xi = (-2, -2, -2)$, and $\zeta = -2$. The color of the node represents the covariate value.

The examples showing perfect homophilous and heterophilous recruitment are extremes. In practice, even for instances of preferential recruitment on covariates, the recruitment tendencies will likely fall between these two extremes.

Limited and not preferential recruitment Figure 4.4 shows an example of limited and not preferential recruitment. Here, for the population $N = 2000$, $n_v = 1$, and $z_g \sim \text{Bernoulli}(0.5)$ as in the previous examples. For the simulation, $n_s = 10$, $n_w = 5$, and $n_c = 3$. These changes are made to increase the sample size, as making recruitment less robust will decrease n. To indicate no preferential recruitment, set $\alpha = 0$ and $\beta = 0$. To indicate limited recruitment, set $\xi = (6, 6, -2)$ and $\zeta = -2$. The value of $\zeta = -2$ indicates that most peers will still

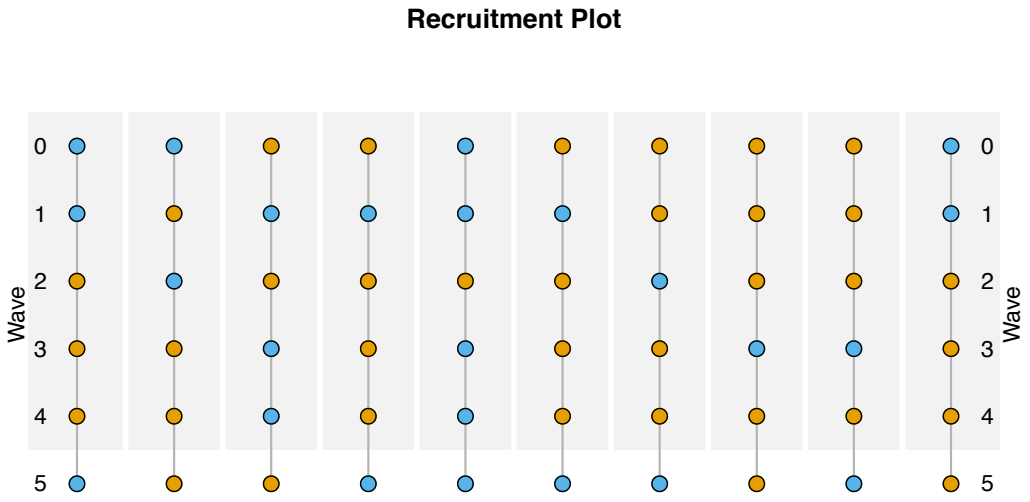


Figure 4.4: Simulation with limited and not preferential recruitment with $n_s = 10$, $n_w = 5$, $\alpha = 0$, $\beta = 0$, $\xi = (6, 6, -2)$, and $\zeta = -2$. The color of the node represents the covariate value.

accept a coupon they are offered (i.e., have a low utility for selecting themself). Setting $\xi = (6, 6, -2)$ indicates that recruiters will have high utilities for selecting themself twice, but a low utility for selecting themself a third time. This indicates that they will likely distribute only one of their three coupons. Since peers want to accept coupons they are given, these settings of ξ and ζ will result in a recruitment chain where each recruiter selects exactly one peer. This can be thought of as a simplification of the many-to-one matching scenario with limited matching into robust one-to-one matching.

The sample size is $n = 60$, with each of the 50 recruiters distributing exactly one of their three coupons successfully. The recruitment homophily is $\mathcal{H}_R = 0.910$. This value is close to 1, so we believe there is not strong preferential recruitment.

Recruitment chains where some people do not recruit, with some homophilous recruitment Figure 4.5 shows an example of recruitment chains where some people do not recruit (and thus possibly, as here, not all seeds initiate chains that reach the full n_w waves). This simulation also has slight homophilous recruitment. Here, for the population $N = 3000$, $n_v = 1$, and $z_g \sim \text{Bernoulli}(0.5)$. For the simulation, $n_s = 4$, $n_w = 6$, and $n_c = 3$. To indicate some homophilous recruitment, set $\alpha = 0.5$ and $\beta = 0.5$. Note that this results in most dyads matching on the covariate value, although it is not the complete set of dyads as in Figure 4.2. To indicate that some people may self match all $n_c = 3$ times, set $\xi = (2.25, 2.25, 2.25)$ and $\zeta = 3.5$. Here the values of ξ and ζ must be big enough that some people will choose to self-match. And in particular for ξ the values must all be high otherwise recruiters will all select at least one recruit. But the values of ξ and ζ cannot be so high such that no recruiters ever distribute coupons or no peers ever accept coupons.

The sample size is $n = 69$, where 4 of the 51 recruiters distributed zero coupons successfully, 30 distributed one successfully, 16 distributed two successfully, and 1 distributed all three successfully. The recruitment homophily is $\mathcal{H}_R = 1.684$. This value is higher than 1, indicating there is homophilous recruitment.

Longer recruitment chains with no preferential recruitment Figure 4.6 shows an example of longer recruitment chains with no preferential recruitment. Here, for the population $N = 3000$, $n_v = 1$, and $z_g \sim \text{Bernoulli}(0.5)$. For the

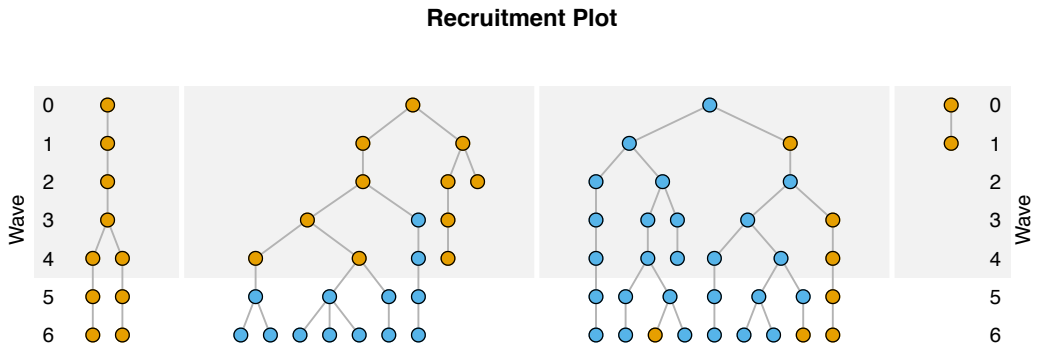


Figure 4.5: Simulation where some people do not recruit anyone, and some recruit homophilously. For this simulation, $n_s = 4$, $n_w = 6$, $\alpha = 0.5$, $\beta = 0.5$, $\xi = (2.25, 2.25, 2.25)$, and $\zeta = 3.5$. The color of the node represents the covariate value.

simulation, $n_s = 3$, $n_w = 10$, and $n_c = 3$. To indicate no preferential recruitment, set $\alpha = 0$ and $\beta = 0$. Setting $\xi = (3, 2.5, 2)$ and $\zeta = 2.5$ indicates that recruiters are slightly less likely to distribute each additional coupon, and overall have some tendency to select themself. This set up results in the recruitment chains from the three original seeds appearing very different: one seed recruited no one, one seed initiates a fairly dense chain, and the final a fairly sparse chain. Different simulations with the same parameters have the possibility to generate graphs with different combinations of these types of recruitment chains. Therefore, these input parameters allow for a lot of variation in the structure of the chain, while those presented earlier had extreme values resulting in nearly identical structure.

The sample size is $n = 189$, where 18 of the 171 recruiters distributed zero coupons successfully, 81 distributed one successfully, 48 distributed two successfully, and 3 distributed all three successfully The recruitment homophily is $\mathcal{H}_R =$

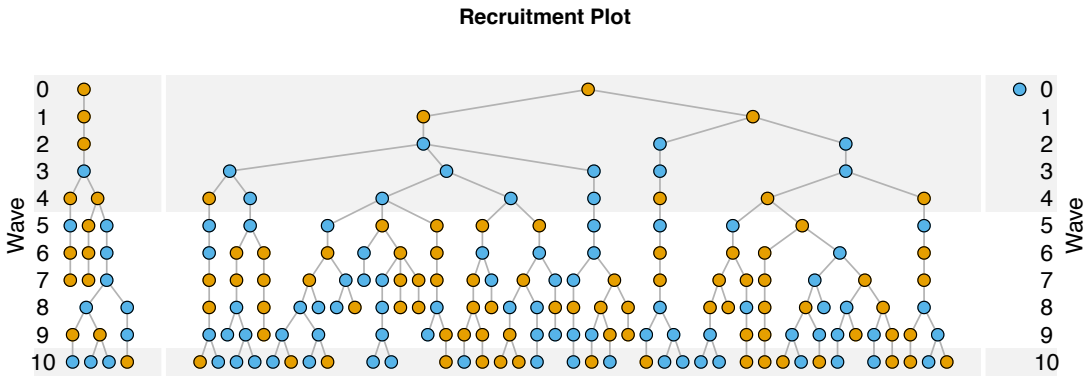


Figure 4.6: Simulation with no preferential recruitment and more waves. For this simulation, $n_s = 3$, $n_w = 10$, $\alpha = 0$, $\beta = 0$, $\xi = (3, 2.5, 2)$, and $\zeta = 2.5$. The color of the node represents the covariate value.

1.118. This value is still relatively close to 1, indicating that there is not strong evidence for preferential recruitment.

Multiple covariates and more covariate categories Figure 4.7 shows an example of a recruitment chain determined by preferences on three different covariates ($n_v = 3$), along with self-matching preferences. Unlike the previous examples, $n_v = 3$ and additionally each variable has a different number of categories such that $z_{g,1} \in \{0, 1, 2, 3, 4\}$, $z_{g,2} \in \{0, 1\}$, and $z_{g,3} \in \{0, 1, 2\}$ for $g \in \{1, \ldots, N\}$. For each variable, an individual is equally likely to be in each of the categories (although this certainly is not a requirement). For the simulation, $N = 500$, $n_s = 3$, $n_w = 5$, and $n_c = 3$. Let recruitment be homophilous with respect to the first covariate, non-preferential with respect to the second covariate, and heterophilous with respect to the third covariate: $\alpha = (1, 0, -1)$ and $\beta = (1, 0, -1)$. Setting $\xi = (3, 3, 3)$ and $\zeta = -1$ indicates that recruiters have a tendency for

selecting themself, but that peers generally want to participate. The three panels of Figure 4.7 color nodes of the resulting recruitment chain by the three different covariates.

The sample size is $n = 98$, where 5 of the 61 recruiters distributed zero coupons successfully, 22 distributed one successfully, 29 distributed two successfully, and 5 distributed all three successfully. The recruitment homophily is $\mathcal{H}_{R,1} = 2.753$ on the first covariate, $\mathcal{H}_{R,2} = 1.051$ on the second covariate, and $\mathcal{H}_{R,3} = 0.125$ on the third covariate. These values reflect the goal of homophilous recruitment on the first covariate, non-preferential recruitment on the second, and heterophilous recruitment on the third.

■

The simulations presented in this section demonstrate the range of recruitment chains that can be simulated using the RCPR model. The model allows the preferences for covariates to be varied from strong heterophily to no preferential recruitment to strong homophily. Further, the model allows the structure of the recruitment chains to be varied from very sparse to very dense, representing robust recruitment.

In Sections 4.2 and 4.3 I demonstrate that the inference procedure can recover the input values of α, β, ξ, and ζ using the simulated recruitment chain as 'observed' data.

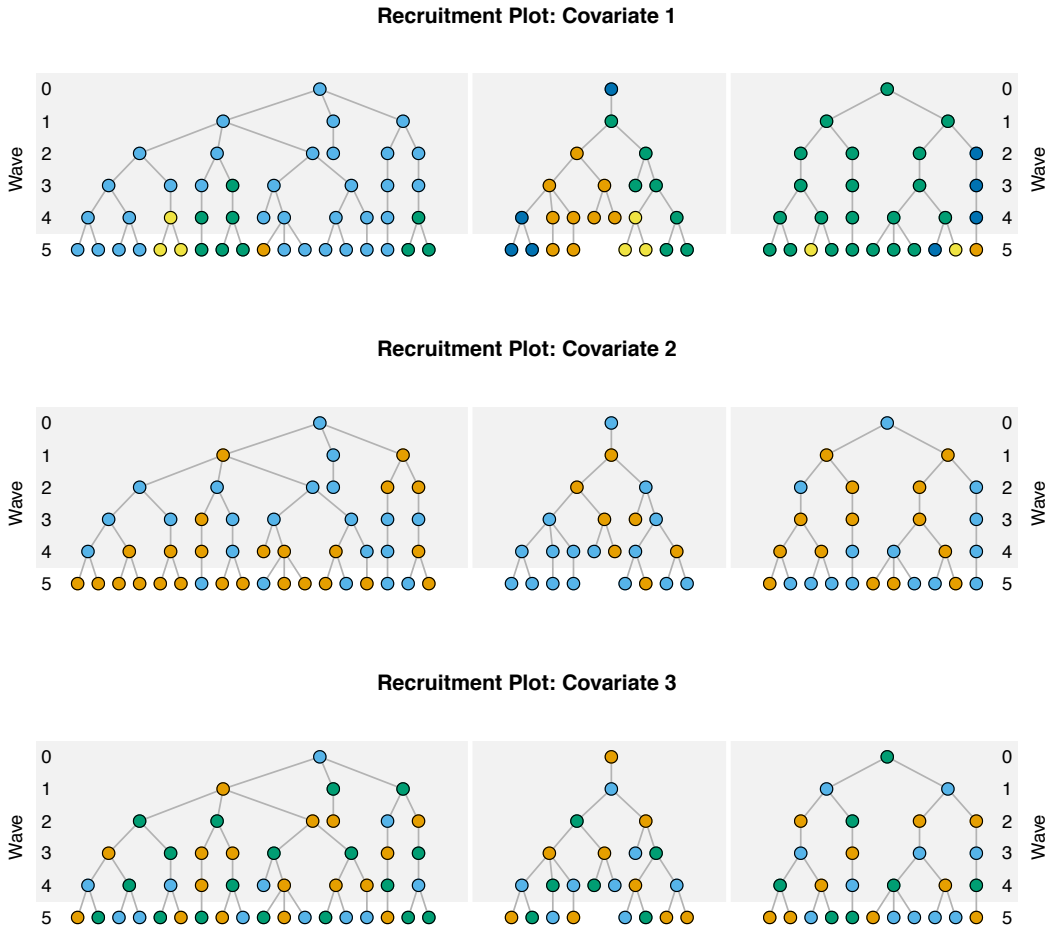


Figure 4.7: Simulation with multiple covariates with more categories. For this simulation, $n_s = 3$, $n_w = 5$, and $n_v = 3$, where the first covariate has five categories, the second has two categories, and the third has three categories. $\alpha = (1, 0, -1)$, $\beta = (1, 0, -1)$, $\xi = (3, 3, 3)$, and $\zeta = -1$. The color of the node represents the covariate value for each of the three different covariates.

4.2 Simulations of One-To-One Matchings

I first present the case of one-to-one matchings: each recruiter only has one coupon. Each recruiter can select either a peer or themself, and each peer can only be recruited once or choose not to participate. The methodology builds from the original Logan et al. (2008) reveled preferences model, with the additional complexity of the self-matching parameters to govern recruitment efficacy. In this case, the matching mechanism employed for simulation and as the rejection sampler for the check during each iteration of the Markov chain is the one-to-one *deferred acceptance algorithm* of Gale and Shapley (1962).

To demonstrate the strength of the methodology, I use many seeds and only one wave of recruitment, a framework as close as possible to the original marriage model. Each of the following simulations uses $n_s = 50$ and $N = 100$, with $n_w = 1$ and $n_c = 1$. As in Section 4.1, $n_v = 1$ and $z_g \sim$ Bernoulli(0.5). The original adjacency matrix between recruiters and peers is retained, so in this case covariate information is known about the full population.

For each of the following examples, the recruitment chains generated from the simulation input parameters can be found in Table 4.3.

No preferential recruitment and no increased propensity to self-match

In this simulation, $\alpha = 0$, $\beta = 0$, $\xi = 0$, and $\zeta = 0$. This means that all utilities will be draws from $\mathrm{N}(0, 1)$. Figure 4.8 shows the MCMC draws and the density of

the posterior distribution for α, β, ξ, and ζ. A burn-in period of 1000 iterations was used, along with a thinning interval of 5. The true values are indicated by red lines, and are recovered well by the inference procedure.

Homophilous recruitment and no increased propensity to self-match In this simulation, $\alpha = 1$, $\beta = 1$, $\xi = 0$, and $\zeta = 0$. This means that people will have a slight preference for matching with someone they share a covariate value with (homophily) over either someone they do not share a covariate value with or themself. Figure 4.9 shows the MCMC draws and the density of the posterior distribution for α, β, ξ, and ζ. A burn-in period of 1000 iterations was used, along with a thinning interval of 5. The true values are indicated by red lines, and are recovered well by the inference procedure.

Heterophilous recruitment and no increased propensity to self-match In this simulation, $\alpha = -1$, $\beta = -1$, $\xi = 0$, and $\zeta = 0$. This means that people will have a slight preference for matching with either someone they do not share a covariate value with (heterophily) or themself over a peer they do share a covariate value with. Figure 4.10 shows the MCMC draws and the density of the posterior distribution for α, β, ξ, and ζ. A burn-in period of 1000 iterations was used, along with a thinning interval of 5. The true values are indicated by red lines, and are recovered well by the inference procedure. Note, however, that in this case there may be local maxima which could lead to potential non-identifiability of the parameters.

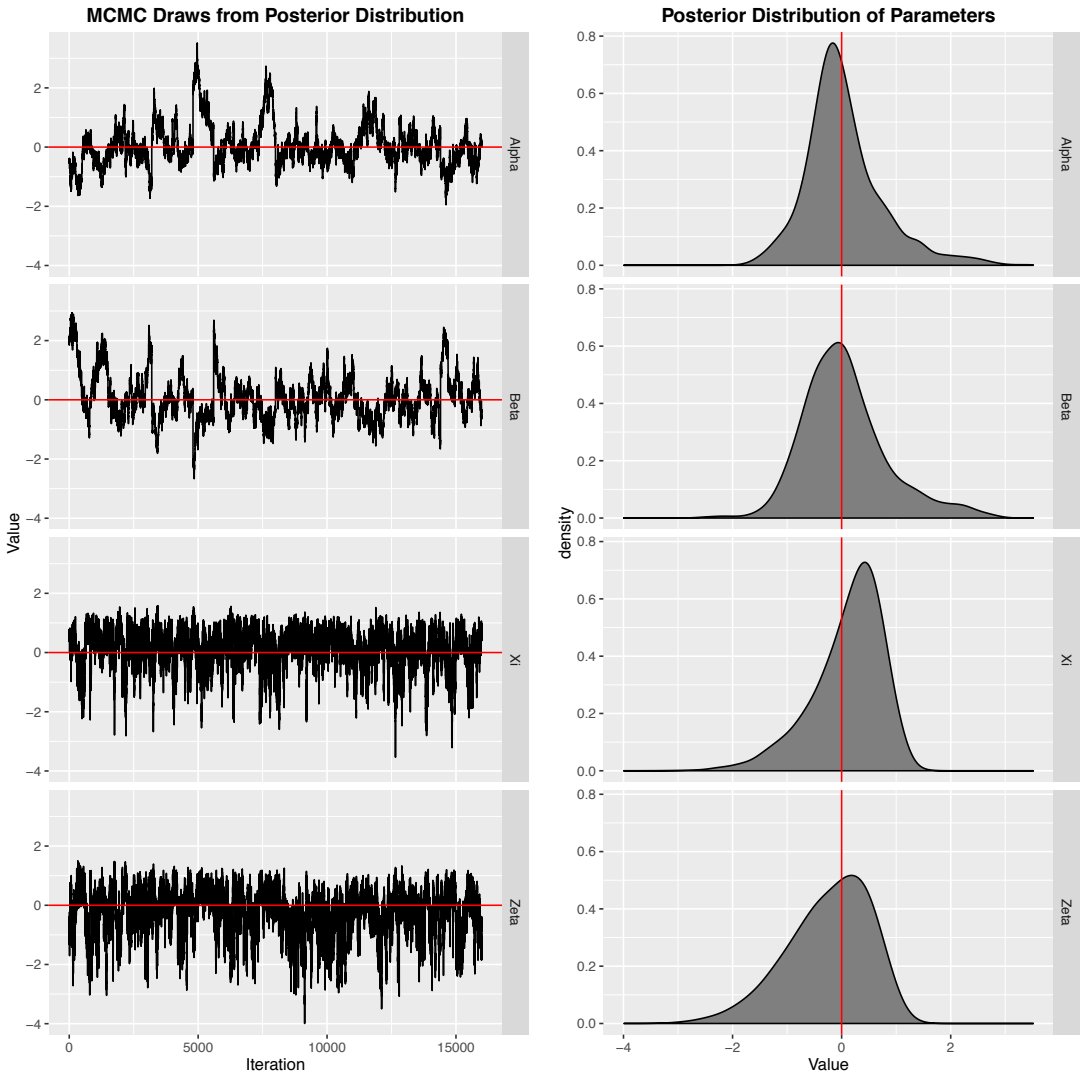


Figure 4.8: MCMC and posterior distributions for one-to-one inference with no preferential recruitment and no increased propensity to self-match. True values (used to set up simulation) are shown in red.

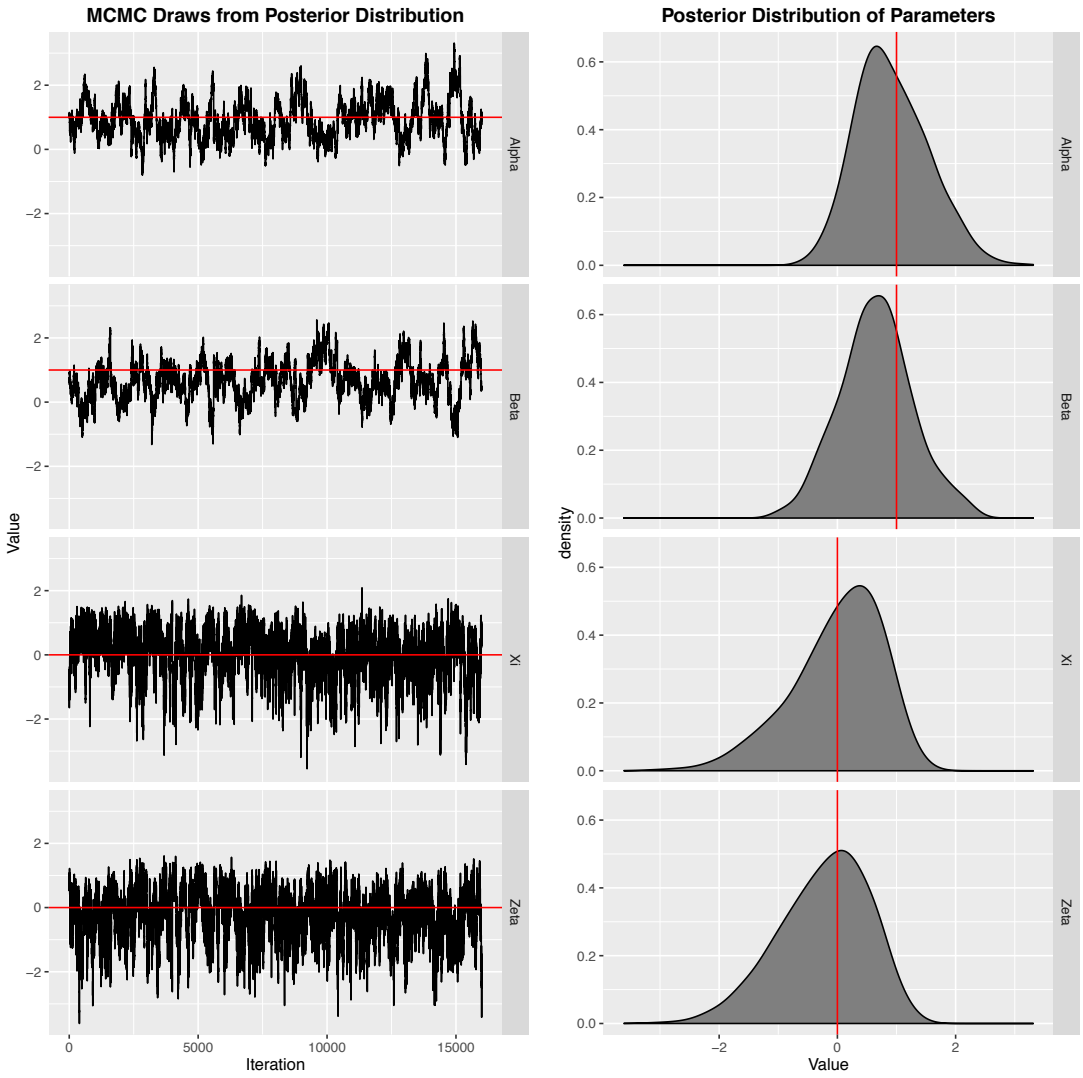


Figure 4.9: MCMC and posterior distributions for one-to-one inference with homophilous recruitment and no increased propensity to self-match. True values (used to set up simulation) are shown in red.

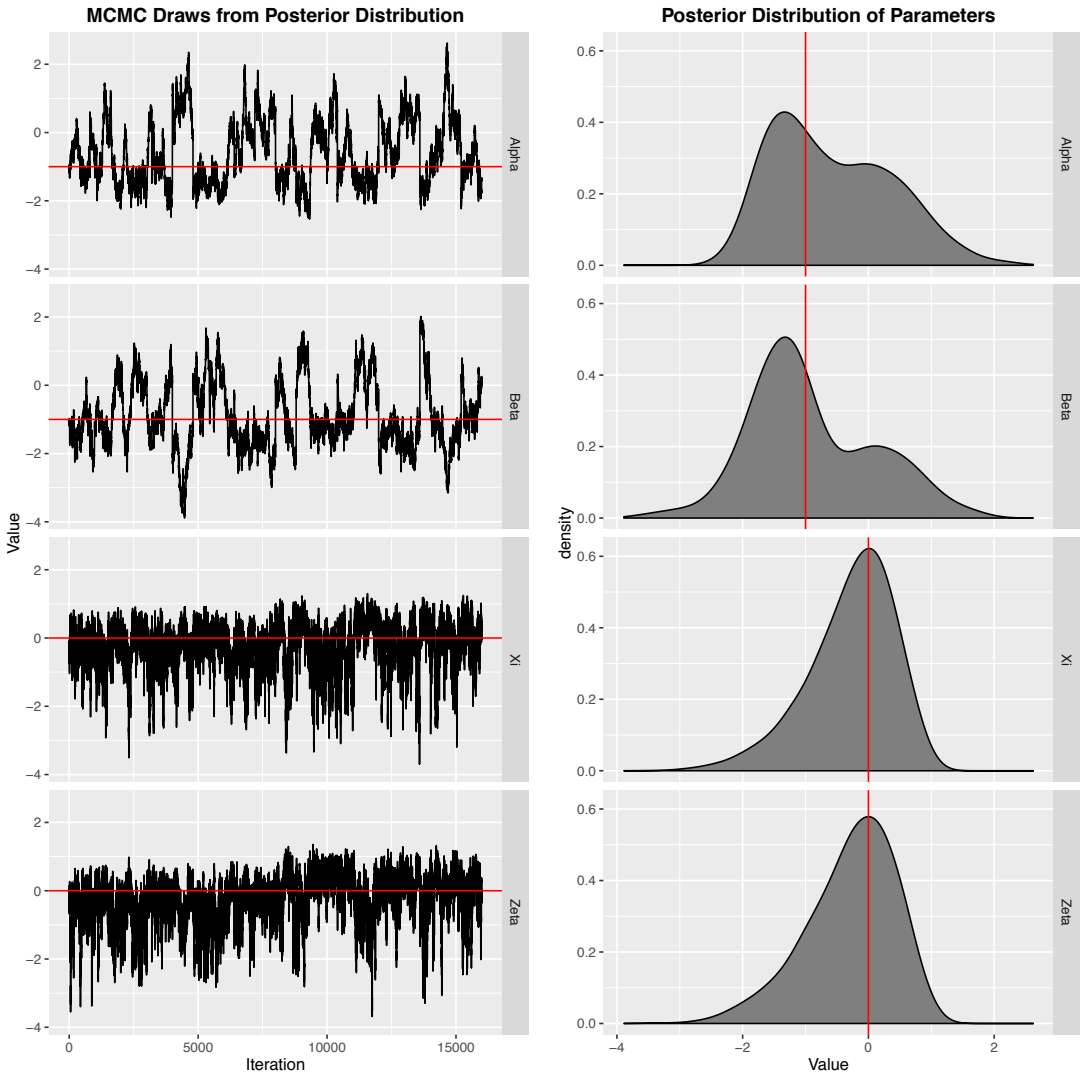


Figure 4.10: MCMC and posterior distributions for one-to-one inference with heterophilous recruitment and no increased propensity to self-match. True values (used to set up simulation) are shown in red.

Different recruiter and peer preferences and no increased propensity to self-match In this simulation, $\alpha = 0.5$, $\beta = -0.5$, $\xi = 0$, and $\zeta = 0$. This means that recruiters will have a slight preference for selecting someone with whom they share a covariate value, but peers prefer to be recruited by someone with whom they do not share a covariate value. Figure 4.11 shows the MCMC draws and the density of the posterior distribution for α, β, ξ, and ζ. A burn-in period of 1000 iterations was used, along with a thinning interval of 5. The true values are indicated by red lines. In this case, the model does not seem to be picking up the differences in α and β. Rather, both distributions appear to be centered around zero, the average of the original α and β used to set up simulation.

Because of the complex dependencies in the recruitment process, there seem to be identifiability problems when the preference parameters take on different values. For example, in this scenario recruiters will prefer peers they share a covariate category with. This means that peers may not get coupons from people of opposite covariate category, so their only options may be self-matching or accepting a coupon from someone in the same covariate category. They have a slightly higher tendency to self-match ($\zeta = 0$) than choosing someone from the same covariate category ($\beta = -0.5$), so they may more often than not choose to self-match. This may result in the self-match values being slightly higher than the original simulation value. Finally, because these peers would have a slight tendency to prefer self-matches, the recruiters will continue attempting to give a coupon away, resulting in attempts to less-preferred peers (who they do not share

a covariate category with) or self-matches. Here, the size of the population affects the outcome: how many people does a recruiter share a covariate category with? Recruiters having to make many coupon distribution attempts may result in both recruiter self-match values that are slightly higher than the original simulation values or homophilous recruitment values lower than in the original simulation. These tendencies are reflected in the posterior distributions in Figure 4.11.

Homophilous recruitment, recruiters have an increased propenstity to self-match In this simulation, $\alpha = 0.5$ and $\beta = 0.5$, indicating slight homophilous recruitment, and $\xi = 2$, and $\zeta = 0$, indicating that recruiters have an increased tendency to self-match. Figure 4.12 shows the MCMC draws and the density of the posterior distribution for α, β, ξ, and ζ. A burn-in period of 1000 iterations was used, along with a thinning interval of 5. The true values are indicated by red lines.

As in the previous example, there appears to be an identifiability problem with the self-match values ξ and ζ, where the values recovered through inference for both appear to be dragged toward each other from their simulation input values.

Homophilous recruitment, peers have an increased propenstity to self-match In this simulation, $\alpha = 0.5$ and $\beta = 0.5$, indicating slight homophilous recruitment, and $\xi = 0$, and $\zeta = 2$, indicating that peers have an increased tendency to self-match. Figure 4.13 shows the MCMC draws and the density of

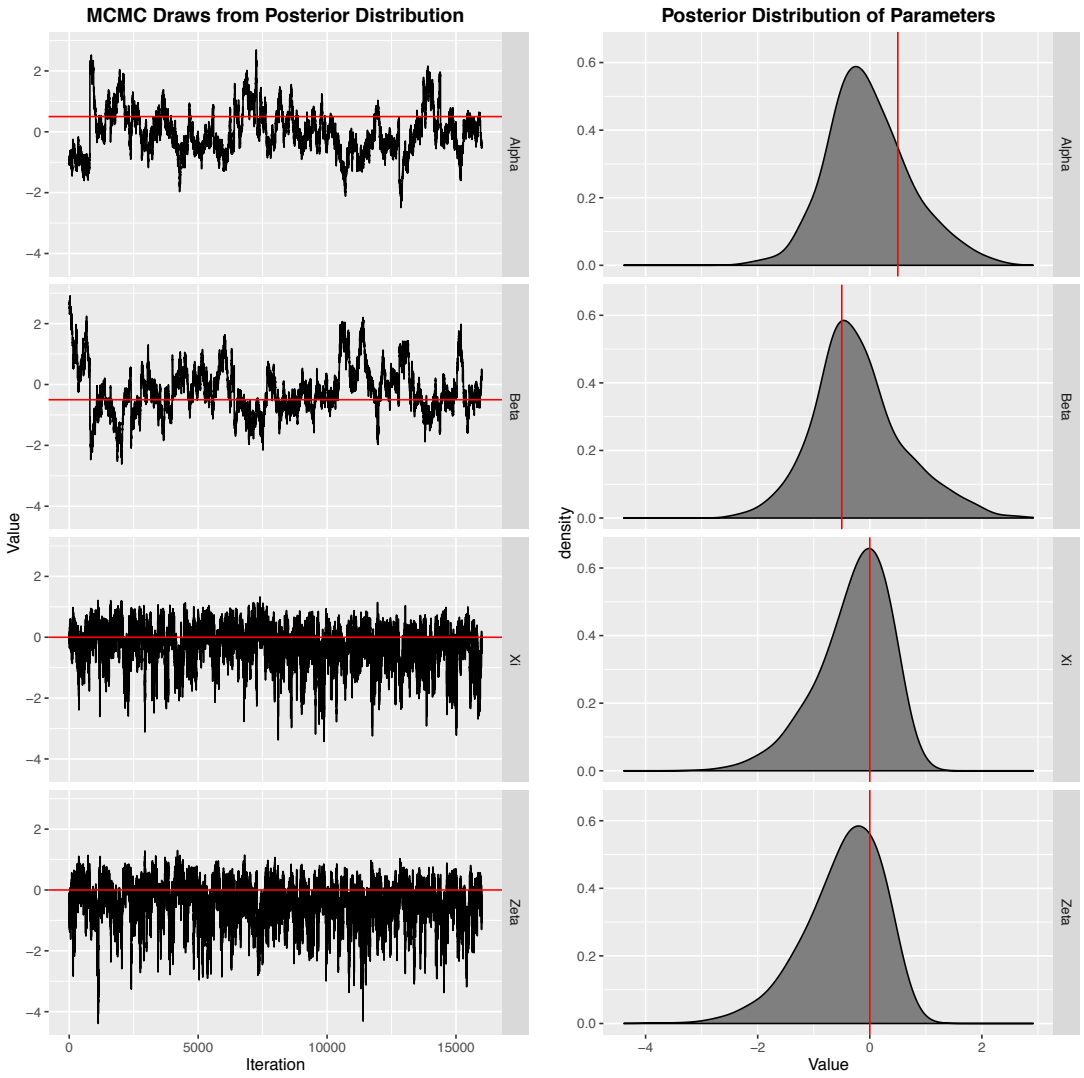


Figure 4.11: MCMC and posterior distributions for one-to-one inference with different recruiter and peer preferences and no increased propensity to self-match. True values (used to set up simulation) are shown in red.

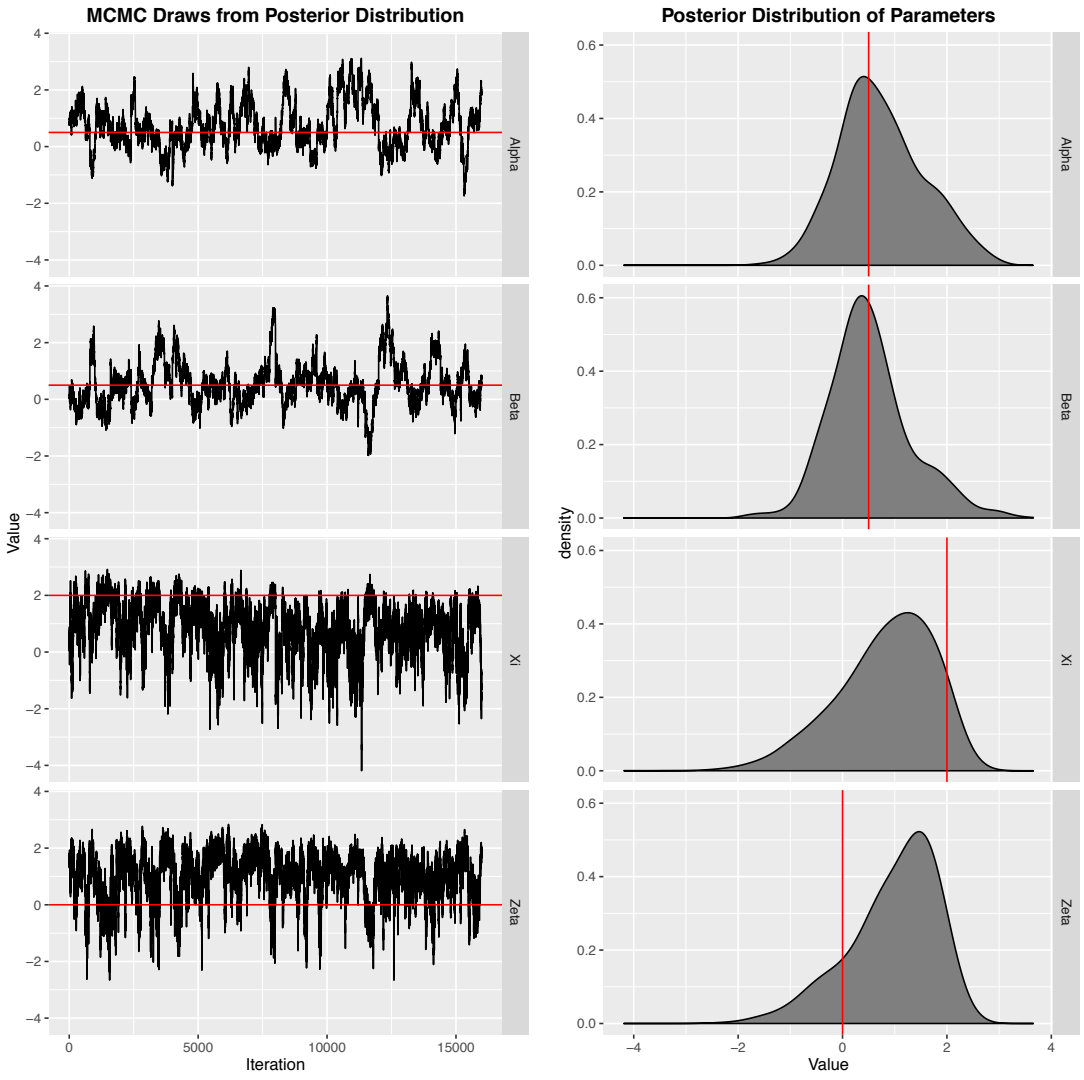


Figure 4.12: MCMC and posterior distributions for one-to-one inference with homophilous recruitment and recruiters having an increased propensity to self–match. True values (used to set up simulation) are shown in red.

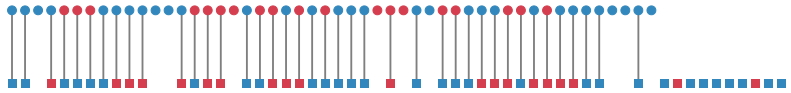
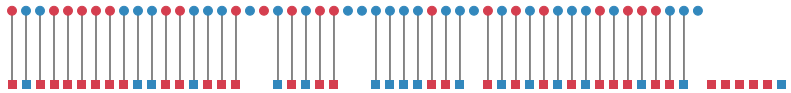
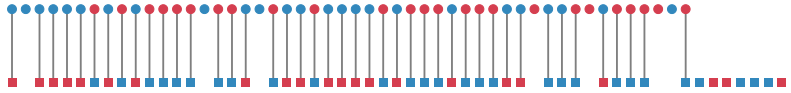
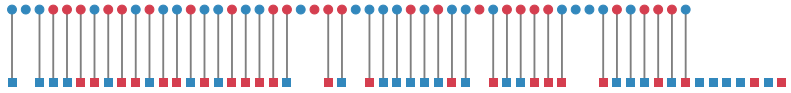
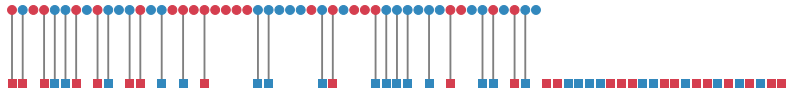
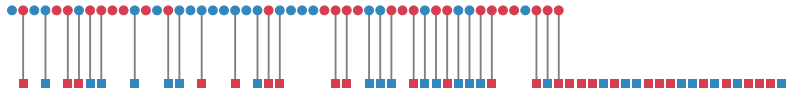
Parameters	Recruitment Chain
$\alpha = 0$, $\beta = 0$, $\xi = 0$, $\zeta = 0$	
$\alpha = 1$, $\beta = 1$, $\xi = 0$, $\zeta = 0$	
$\alpha = -1$, $\beta = -1$, $\xi = 0$, $\zeta = 0$	
$\alpha = 0.5$, $\beta = -0.5$, $\xi = 0$, $\zeta = 0$	
$\alpha = 0.5$, $\beta = 0.5$, $\xi = 2$, $\zeta = 0$	
$\alpha = 0.5$, $\beta = 0.5$, $\xi = 0$, $\zeta = 2$	

Table 4.3: One-to-one simulated recruitment chains. Recruiters are indicated by circles, peers by squares. The color of the node indicates covariate category.

the posterior distribution for α, β, ξ, and ζ. A burn-in period of 1000 iterations was used, along with a thinning interval of 5. The true values are indicated by red lines.

As in the previous two examples, there appears to be an identifiability problem with the self-match values ξ and ζ, where the values recovered through inference for both appear to be dragged toward each other from their simulation input values.

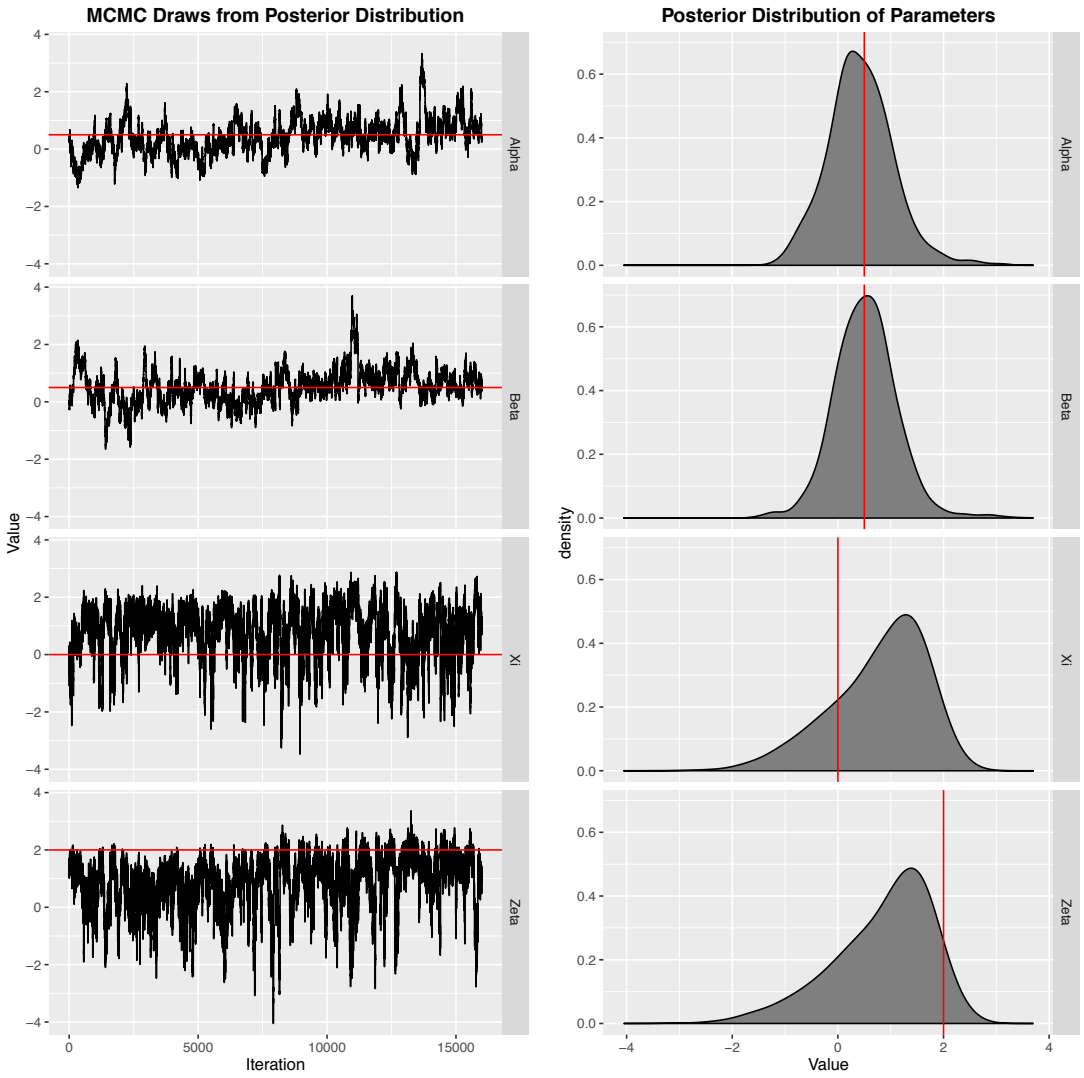


Figure 4.13: MCMC and posterior distributions for one-to-one inference with homophilous recruitment and peers having an increased propensity to self-match. True values (used to set up simulation) are shown in red.

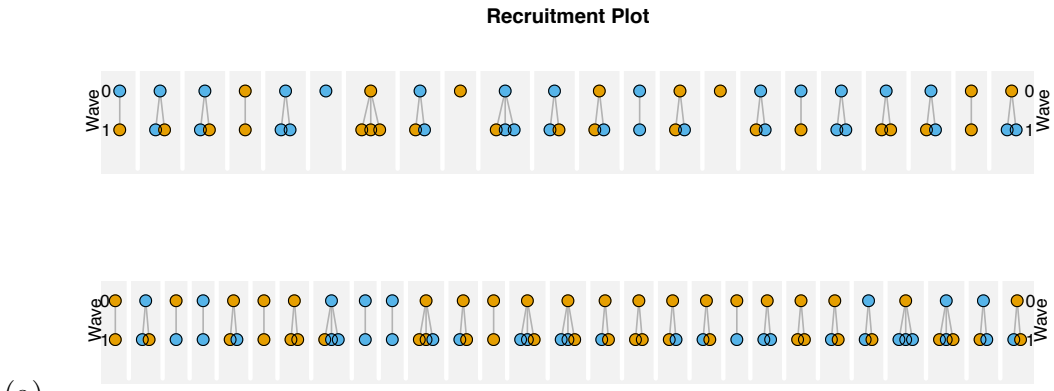
4.3 Simulations of Many-To-One Matchings

I now demonstrate model performance for the case of many-to-one matchings. In particular, I focus on cases where $n_c = 3$, the typical value used in RDS studies. In these simulations I focus on different values of the covariate preference coefficients α and β and the ability of the self match coefficients ξ and ζ to generate new recruitment chains that look similar to the original one. The specific values of ξ and ζ are not important, however, and were chosen such that the recruitment chains appeared realistic.

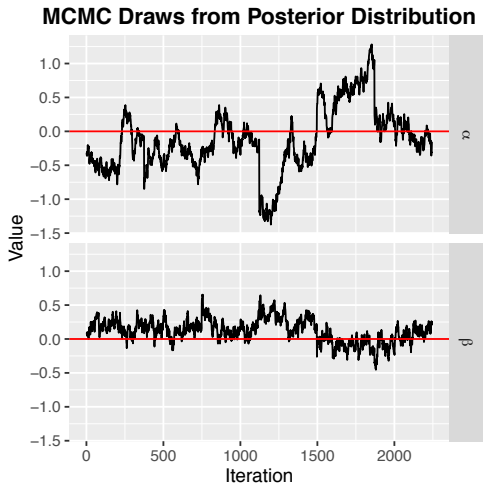
No preferential recruitment, $n_w = 1$. First consider the case where there is no preferential recruitment: $\alpha = 0$ and $\beta = 0$. This simulation uses $n_w = 1$, $n_s = 50$, $N = 200$, $n_v = 1$, $z_g \sim$ Bernoulli(0.5), $\xi = (1, 1, 1)$, and $\zeta = 1$. The recruitment graph for successful recruitments is shown in Figure 4.14(a).

For inference, information about the complete population was retained. Figure 4.14(b)-(c) show the MCMC and posterior distributions for α and β, with the original simulation values indicated by a red line. A burn-in period of 500 iterations and thinning interval of 4 were used. The model does an adequate job of recovering these values, especially when considering the net effect of the recruiter and peer preference values.

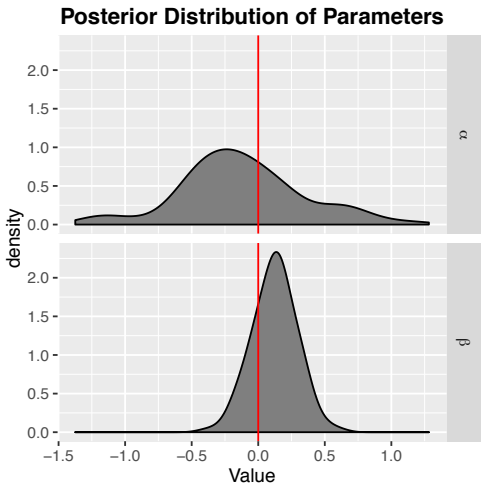
Because of the identifiability issue with ξ and ζ, exacerbated by the many-to-one nature of matching, I instead consider whether or not the median values of ξ



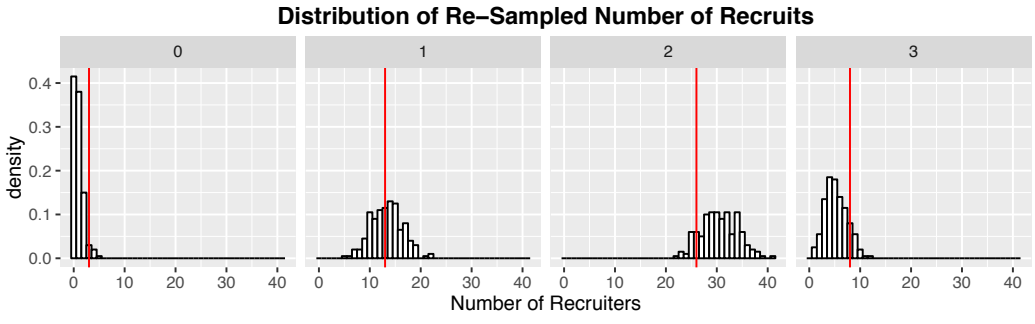
(a)



(b)



(c)



(d)

Figure 4.14: (a) Recruitment chain, (b) MCMC, (c) posterior distribution, and (d) re-simulated recruitment chains for $n_w = 1$, $\alpha = 0$, and $\beta = 0$.

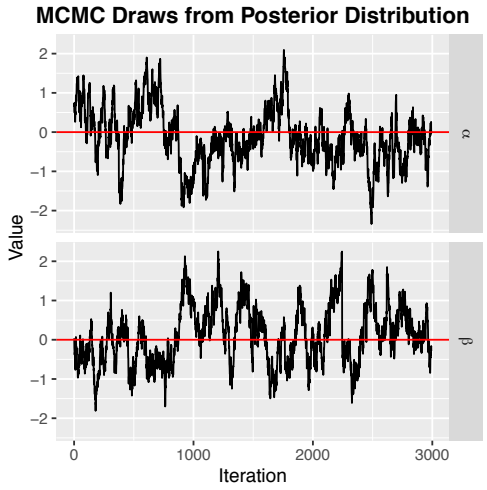
and ζ recovered via inference simulate recruitment chains similar to the original simulation. Figure 4.14(d) shows the distribution of the number of recruiters who successfully recruited 0, 1, 2, or 3 peers in 200 new simulations. The red line indicates the count for the original simulated recruitment chain. The values of ξ and ζ appear plausible.

No preferential recruitment, $n_w = 5$. Now consider the same case of no preferential recruitment, but a more realistic scenario with multiple waves of recruitment. Here, the simulation uses $\alpha = 0$, $\beta = 0$, $n_w = 5$, $n_s = 3$, $N = 300$, $n_v = 1$, $z_g \sim$ Bernoulli(0.5), $\xi = (3, 2, 1)$, and $\zeta = 2$. The recruitment graph for successful recruitments is shown in Figure 4.15(a). Notice that this combination of self-matching preference parameters resulted in a very sparse recruitment chain.

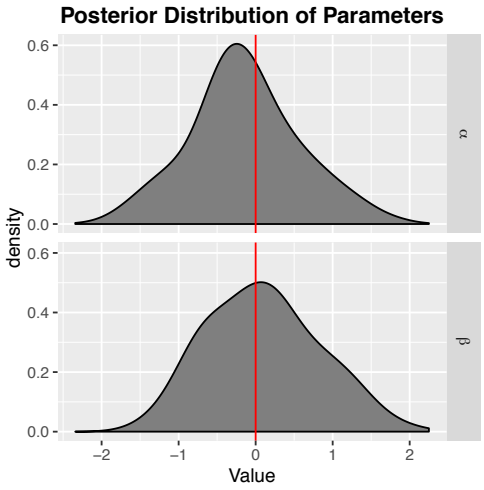
For inference, information about the complete population was retained. Figure 4.15(b)-(c) show the MCMC and posterior distributions for α and β, with the original simulation values indicated by a red line. A burn-in period of 500 iterations and thinning interval of 4 were used. The model does an adequate job of recovering these values.

Because of the identifiability issue with ξ and ζ, exacerbated by the many-to-one nature of matching, I instead consider whether or not the median values of ξ and ζ recovered via inference simulate recruitment chains similar to the original simulation. Figure 4.15(d) shows the distribution of the number of recruiters who successfully recruited 0, 1, 2, or 3 peers in 200 new simulations. The red line

(a)



(b)



(c)

(d)

Figure 4.15: (a) Recruitment chain, (b) MCMC, (c) posterior distribution, and (d) re-simulated recruitment chains for $n_w = 5$, $\alpha = 0$, and $\beta = 0$.

indicates the count for the original simulated recruitment chain. In this case, the values of ξ and ζ do not appear the most plausible. However, this could be due to the small number of people who were successfully recruited.

No preferential recruitment, network size term. Now consider the same case of no preferential recruitment, with the added complexity of a second term in the model to capture the network size (degree) of participants. Instead of using $n_v = 1$ where $z_g \sim \text{Bernoulli}(0.5)$, use $n_v = 2$ where $z_{g;1} \sim \text{Bernoulli}(0.5)$ as before representing the covariate, and $z_{g;2} \sim \text{NB}(5, 0.4)$ conditional on $z_{g;2} > 0$. $\text{NB}(r, p)$ is the negative binomial distribution with parameters $r > 0$ (the number of failures until the experiment is stopped) and $p \in (0, 1)$ (the probability of success in each experiment).

In this case, for covariate D representing degree, let $\mathbf{X}_{ij;D} = \tilde{d}_j / \max(\tilde{d}_{h \in \mathcal{P}})$ and $\mathbf{Y}_{ji;D} = \tilde{d}_i / \max(\tilde{d}_{\ell \in \mathcal{R}})$. Because $\tilde{d} > 0$ (everyone in the study must have degree larger than zero since they were recruited by someone or agreed to serve as a seed), $0 < \mathbf{X}_{ij;D}, \mathbf{Y}_{ji;D} \leq 1$. Therefore the network size covariate is on the same scale as the categorical covariates discussed previously. Large positive values of α_D and β_D indicate that recruiters want to give coupons to peers with higher degree, and peers want to accept coupons from recruiters with higher degree. Values of α_D and β_D close to zero indicate that degree does not have a large effect on recruitment decisions. Large negative values of α_D and β_D indicate that recruiters want to give coupons to peers with lower degree, and peers want to

accept coupons from recruiters with lower degree.

This simulation uses $\alpha = (0, 0)$, $\beta = (0, 0)$, $n_w = 5$, $n_s = 3$, $N = 300$, $n_v = 2$, $z_{g;1} \sim \text{Bernoulli}(0.5)$, $z_{g;2} \sim \text{NB}(5, 0.4)$, $\xi = (3, 2, 1)$, and $\zeta = 2$. Because the preference coefficients for the network terms are zero, the network size should not have an effect on recruitment and we should observe a similar scenario to Section 4.3. The recruitment graph for successful recruitments is shown in Figure 4.16(a).

For inference, information about the complete population was retained. Figure 4.16(b)-(c) show the MCMC and posterior distributions for α and β, with the original simulation values indicated by a red line. A burn-in period of 500 iterations and thinning interval of 4 were used. The model does an adequate job of recovering these values for the covariate. The values for α_D and β_D are not as good, although this could be due to a sample size issue.

Because of the identifiability issue with ξ and ζ, exacerbated by the many-to-one nature of matching, I instead consider whether or not the median values of ξ and ζ recovered via inference simulate recruitment chains similar to the original simulation. Figure 4.16(d) shows the distribution of the number of recruiters who successfully recruited 0, 1, 2, or 3 peers in 200 new simulations. The red line indicates the count for the original simulated recruitment chain. In this case, the values of ξ and ζ do an excellent job of returning recruitment chains that look like the original.

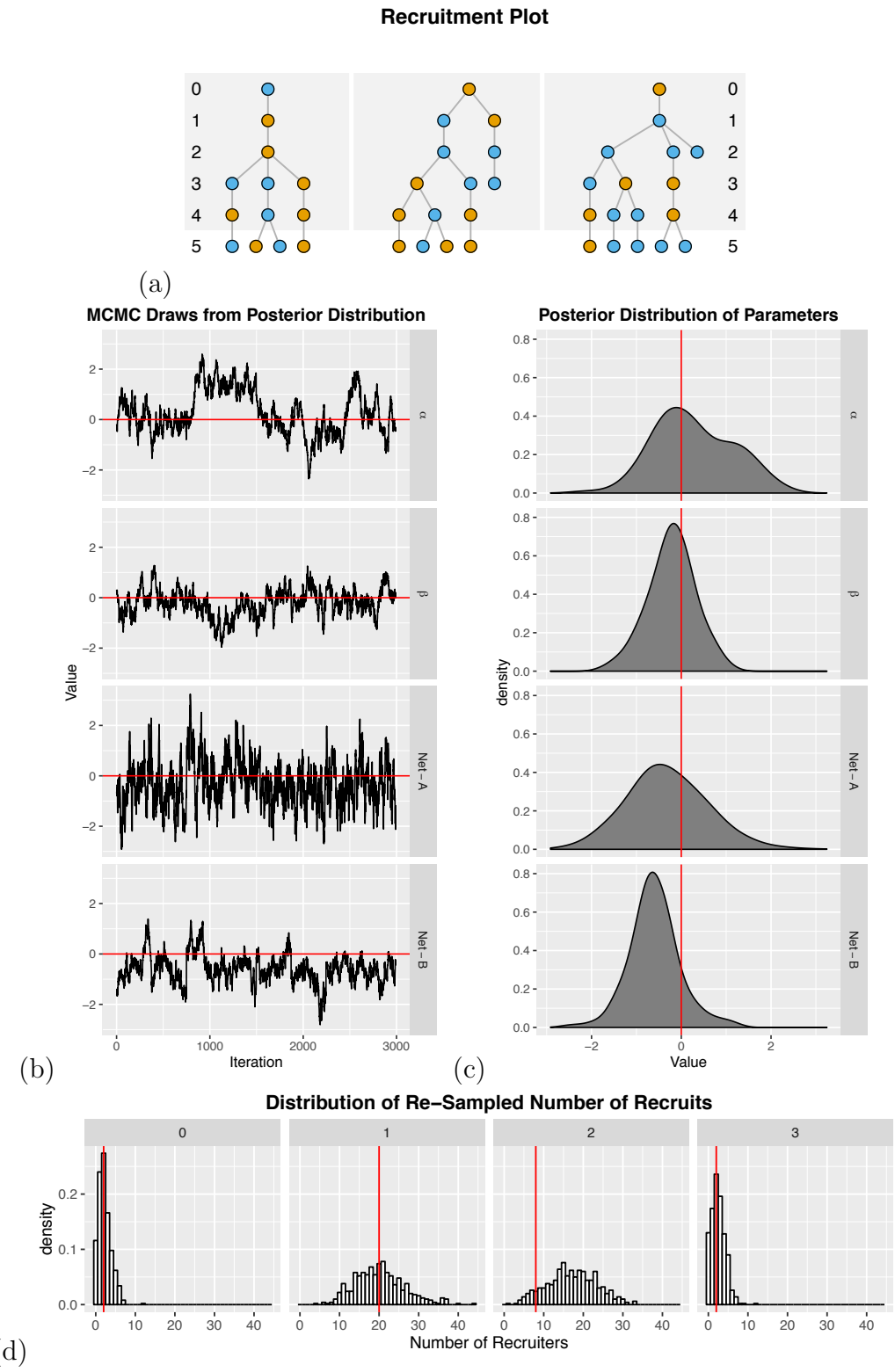


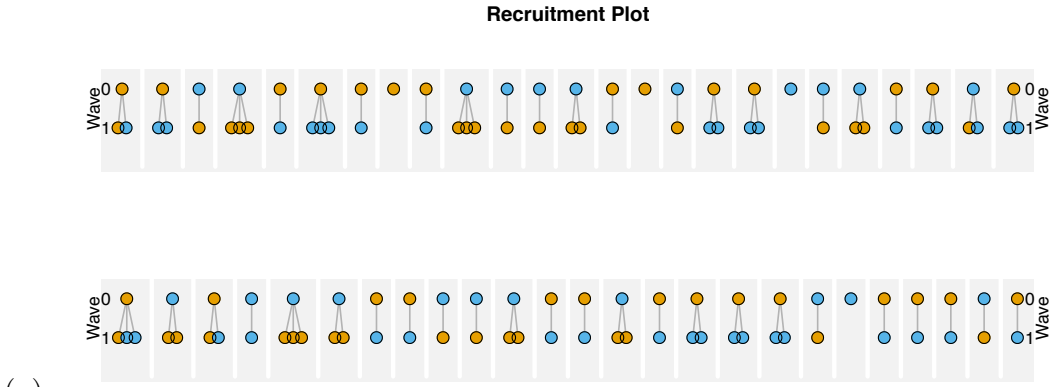
Figure 4.16: (a) Recruitment chain, (b) MCMC, (c) posterior distribution, and (d) re-simulated recruitment chains for $n_w = 5$, $\alpha = (0, 0)$, $\beta = (0, 0)$, using a netwok size term.

Heterophilous recruitment, $n_w = 1$. I now consider the case where there is heterophilous recruitment: $\alpha = -1$ and $\beta = -1$. This simulation uses $n_w = 1$, $n_s = 50$, $N = 200$, $n_v = 1$, $z_g \sim$ Bernoulli(0.5), $\xi = (1, 1, 1)$, and $\zeta = 1$. The recruitment graph for successful recruitments is shown in Figure 4.17(a).

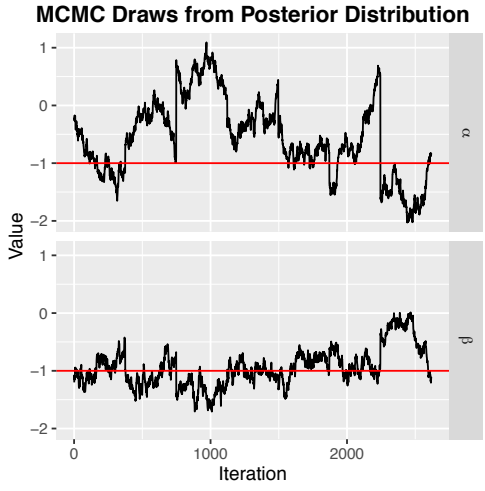
For inference, information about the complete population was retained. Figure 4.17(b)-(c) show the MCMC and posterior distributions for α and β, with the original simulation values indicated by a red line. A burn-in period of 500 iterations and thinning interval of 4 were used. The model does a decent job of recovering these values.

Because of the identifiability issue with ξ and ζ, exacerbated by the many-to-one nature of matching, I instead consider whether or not the median values of ξ and ζ recovered via inference simulate recruitment chains similar to the original simulation. Figure 4.17(d) shows the distribution of the number of recruiters who successfully recruited 0, 1, 2, or 3 peers in 200 new simulations. The red line indicates the count for the original simulated recruitment chain. The values of ξ and ζ appear plausible.

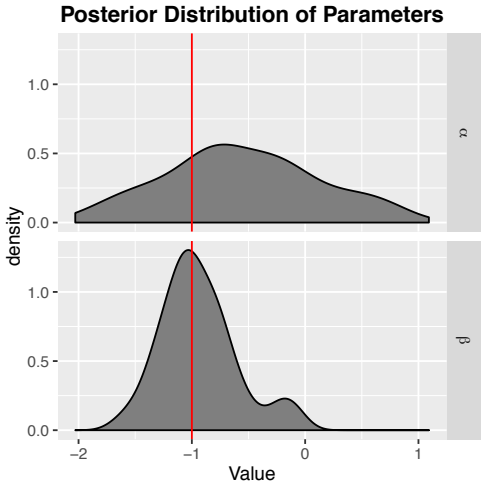
Heterophilous recruitment, $n_w = 4$. I now consider the same heterophilous recruitment scenario with $\alpha = -1$ and $\beta = -1$, and with more waves of recruitment. This simulation uses $n_w = 4$, $n_s = 3$, $N = 200$, $n_v = 1$, $z_g \sim$ Bernoulli(0.5), $\xi = (1, 1, 1)$, and $\zeta = 1$. The recruitment graph for successful recruitments is shown in Figure 4.18(a).



(a)



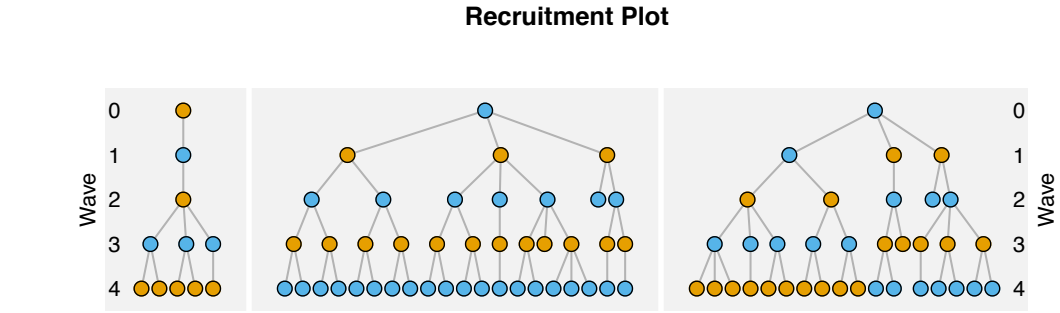
(b)



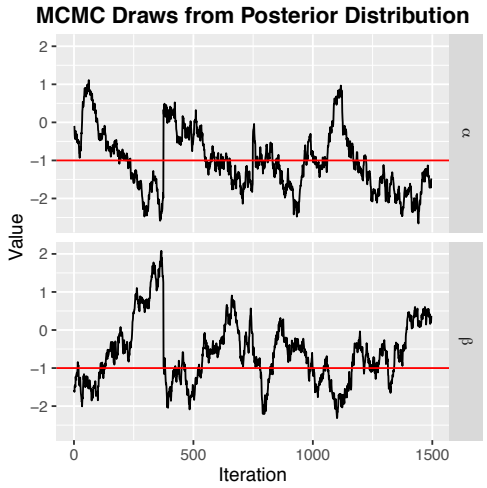
(c)

(d)

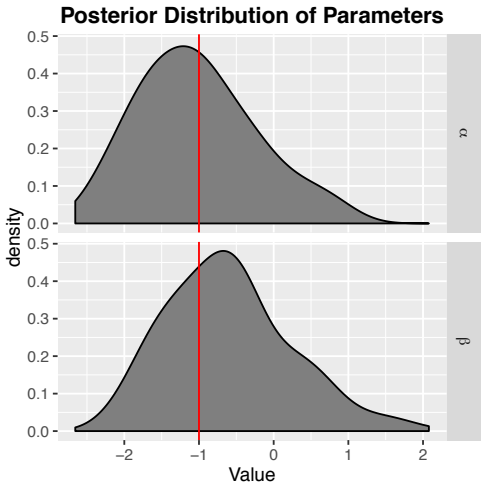
Figure 4.17: (a) Recruitment chain, (b) MCMC, (c) posterior distribution, and (d) re-simulated recruitment chains for $n_w = 1$, $\alpha = -1$, and $\beta = -1$.



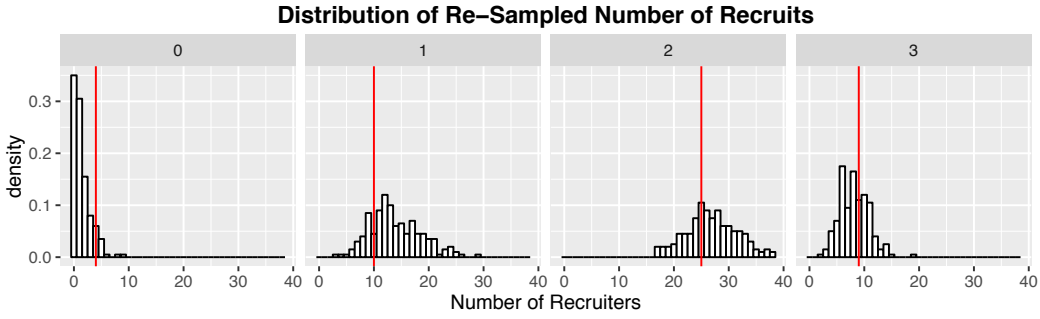
(a)



(b)



(c)



(d)

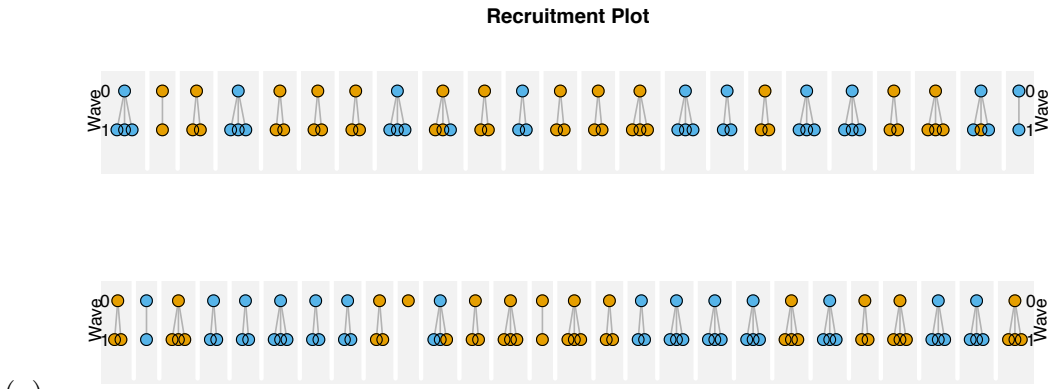
Figure 4.18: (a) Recruitment chain, (b) MCMC, (c) posterior distribution, and (d) re-simulated recruitment chains for $n_w = 4$, $\alpha = -1$, and $\beta = -1$.

For inference, information about the complete population was retained. Figure 4.18(b)-(c) show the MCMC and posterior distributions for α and β, with the original simulation values indicated by a red line. A burn-in period of 500 iterations and thinning interval of 4 were used. The model does a decent job of recovering these values, and definitely picks up the negative trend.

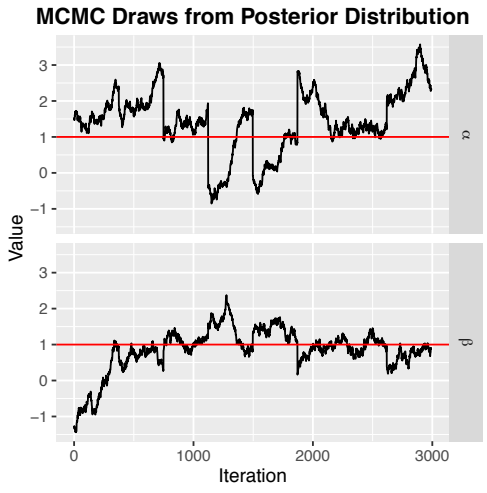
Because of the identifiability issue with ξ and ζ, exacerbated by the many-to-one nature of matching, I instead consider whether or not the median values of ξ and ζ recovered via inference simulate recruitment chains similar to the original simulation. Figure 4.18(d) shows the distribution of the number of recruiters who successfully recruited 0, 1, 2, or 3 peers in 200 new simulations. The red line indicates the count for the original simulated recruitment chain. The values of ξ and ζ appear plausible.

Homophilous recruitment, $n_w = 1$. I now consider the case where there is homophilous recruitment: $\alpha = 1$ and $\beta = 1$. This simulation uses $n_w = 1$, $n_s = 50$, $N = 200$, $n_v = 1$, $z_g \sim$ Bernoulli(0.5), $\xi = (1, 1, 1)$, and $\zeta = 1$. The recruitment graph for successful recruitments is shown in Figure 4.19(a).

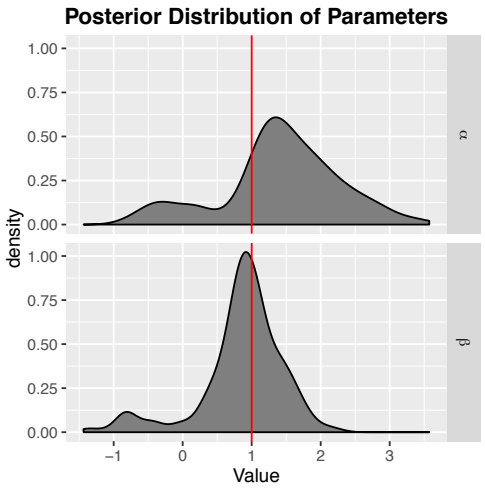
For inference, information about the complete population was retained. Figure 4.19(b)-(c) show the MCMC and posterior distributions for α and β, with the original simulation values indicated by a red line. A burn-in period of 500 iterations and thinning interval of 4 were used. The model does a decent job of recovering these values.



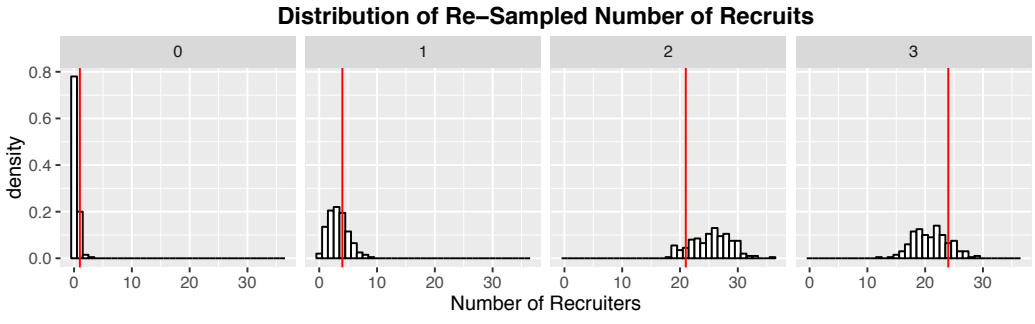
(a)



(b)



(c)



(d)

Figure 4.19: (a) Recruitment chain, (b) MCMC, (c) posterior distribution, and (d) re-simulated recruitment chains for $n_w = 1$, $\alpha = 1$, and $\beta = 1$.

Because of the identifiability issue with ξ and ζ, exacerbated by the many-to-one nature of matching, I instead consider whether or not the median values of ξ and ζ recovered via inference simulate recruitment chains similar to the original simulation. Figure 4.19(d) shows the distribution of the number of recruiters who successfully recruited 0, 1, 2, or 3 peers in 200 new simulations. The red line indicates the count for the original simulated recruitment chain. The values of ξ and ζ appear plausible.

Homophilous recruitment, $n_w = 4$. I now consider the same homophilous recruitment scenario with $\alpha = 1$ and $\beta = 1$, and with more waves of recruitment. This simulation uses $n_w = 4$, $n_s = 3$, $N = 200$, $n_v = 1$, $z_g \sim$ Bernoulli(0.5), $\xi = (2, 2, 2)$, and $\zeta = 1$. The recruitment graph for successful recruitments is shown in Figure 4.20(a).

For inference, information about the complete population was retained. Figure 4.20(b)-(c) show the MCMC and posterior distributions for α and β, with the original simulation values indicated by a red line. A burn-in period of 500 iterations and thinning interval of 4 were used. The model does a decent job of recovering these values, and definitely picks up the positive trend.

Because of the identifiability issue with ξ and ζ, exacerbated by the many-to-one nature of matching, I instead consider whether or not the median values of ξ and ζ recovered via inference simulate recruitment chains similar to the original simulation. Figure 4.20(d) shows the distribution of the number of recruiters who

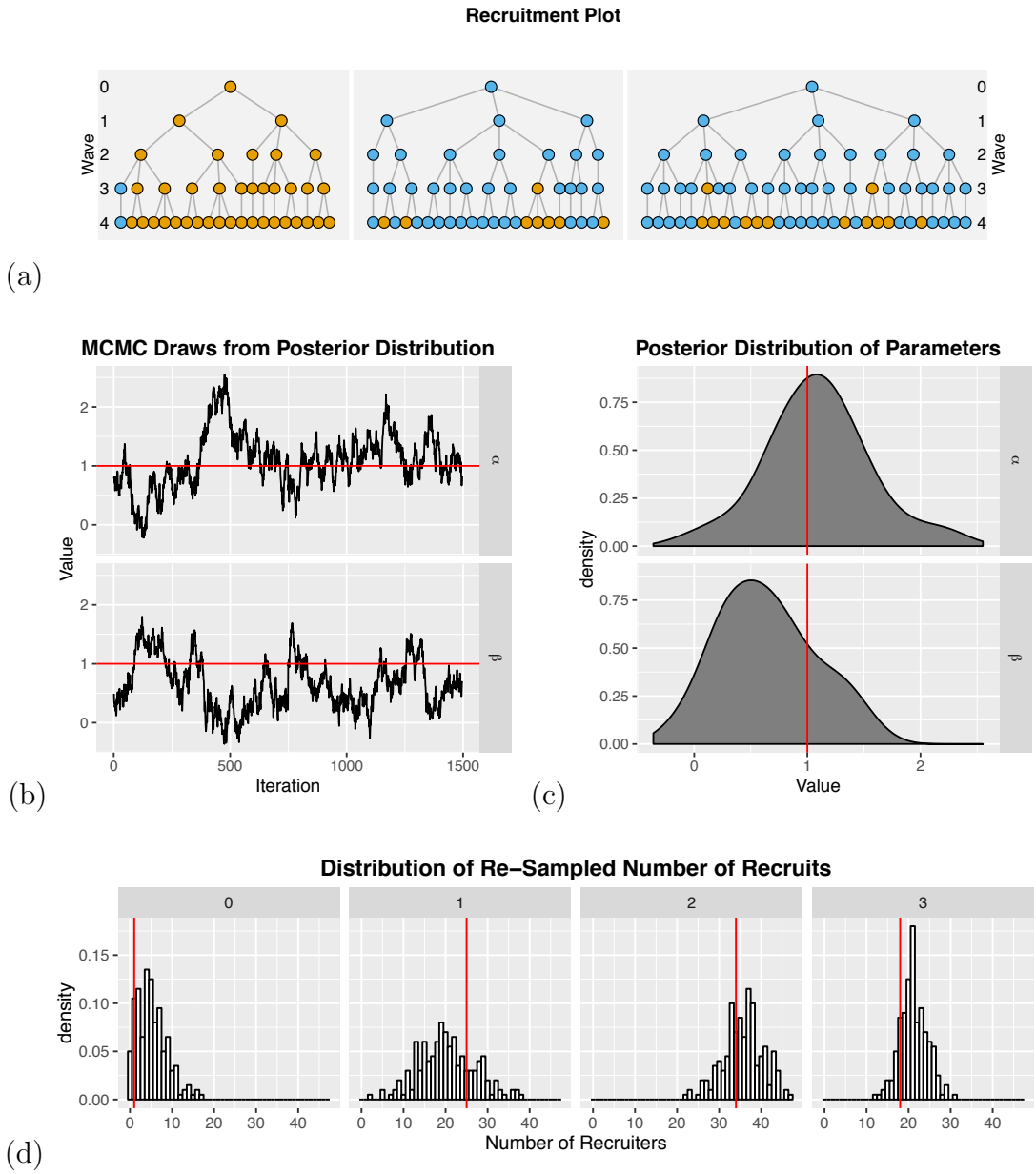


Figure 4.20: (a) Recruitment chain, (b) MCMC, (c) posterior distribution, and (d) re-simulated recruitment chains for $n_w = 4$, $\alpha = 1$, and $\beta = 1$.

successfully recruited 0, 1, 2, or 3 peers in 200 new simulations. The red line indicates the count for the original simulated recruitment chain. The values of ξ and ζ appear plausible.

■

In this chapter I presented a variety of simulations demonstrating the range of the RCPR model and its ability to generate recruitment chains that look like those we would see in real RDS studies. Further, for both one-to-one and many-to-one matchings, I demonstrate that simulation input parameters for the covariate preference coefficients α and β can be recovered using the inference procedure. There are issues with identifiability of the self match preference coefficients ξ and ζ, particularly in the many-to-one case of recruiters having multiple coupons. Nevertheless, the function of ξ and ζ is to control recruitment robustness. Therefore, it is not as important to recover the precise values of ξ and ζ as it is to use those values to generate recruitment chains similar to the original one. Indeed, this is possible for the examples presented in Section 4.3. The RCPR model helps make insights into preferential recruitment for RDS. Examples using data from several RDS studies are provided in Chapter 5.

CHAPTER 5

Application to RDS Studies of Populations at High Risk for HIV/AIDS

In addition to the simulation studies presented in Chapter 4, the preferential recruitment model performs well on data from RDS studies conducted in several different contexts. Examples are presented for: people who inject drugs (PWID) in Prizren, Kosovo; Francophone migrants in Rabat, Morocco; and female sex workers (FSW) in Agadir, Morocco. Summary measures of these populations, gathered from the literature, are presented in Table 5.1, and each study is discussed in more detail in its corresponding section.

Each example considers one covariate (with either high homophily or negligible homophily), along with either the inclusion or exclusion of a network size term. For recruiter $i \in \mathcal{R}$ and peer $j \in \mathcal{P}$ on covariate c, recall that $\mathbf{X}_{ij;c} = \mathbf{Y}_{ji;c} = 1$ if $z_{i;c} =$

Population	**Study Period**	N	n	n_s	n_w
PWID, Prizren, Kosovo	July-Sept., 2014	1728	199	4	12
Francophone migrants, Rabat, Morocco	March-April, 2013	4300	410	6	6
FSW, Agadir, Morocco	Dec., 2011-Jan., 2012	4900	372	10	8

Table 5.1: Summary measures for populations, gathered from the literature.

$z_{j;c}$ and 0 otherwise. The traditional intuition for RDS is that people with higher degree (personal network size) are more likely to be included in the sample than those with lower degree (Volz and Heckathorn (2008); Gile (2011); Handcock et al. (2014); Handcock et al. (2015)). To capture this, a network size term can be added to the model as an element of α, β, $\mathbf{X}$, and $\mathbf{Y}$. In this case, for covariate D representing degree, let $\mathbf{X}_{ij;D} = \tilde{d}_j / \max(\tilde{d}_{h \in \mathcal{P}})$ and $\mathbf{Y}_{ji;D} = \tilde{d}_i / \max(\tilde{d}_{\ell \in \mathcal{R}})$. Because $\tilde{d} > 0$ (everyone in the study must have degree larger than zero since they were recruited by someone or agreed to serve as a seed), $0 < \mathbf{X}_{ij;D}, \mathbf{Y}_{ji;D} \leq 1$. Therefore the network size covariate is on the same scale as the categorical covariates discussed previously. Large positive values of α_D and β_D indicate that recruiters want to give coupons to peers with higher degree, and peers want to accept coupons from recruiters with higher degree. Values of α_D and β_D close to zero indicate that degree does not have a large effect on recruitment decisions. Large negative values of α_D and β_D indicate that recruiters want to give coupons to peers with lower degree, and peers want to accept coupons from recruiters with lower degree.

5.1 People Who Inject Drugs in Prizren, Kosovo

An RDS study of people who inject drugs (PWID) in Prizren, Kosovo was conducted as part of the HIV Integrated Behavioral and Biological Surveillance (IBBS) Survey in July through September, 2014 (Kosovo HIV Integrated Behav-

ioral and Biological Surveillance Survey Reference Group, 2014). To be eligible, participants were required to be 18 years or older, live and/or work in Prizren, and have injected drugs in the past month. This study, including formative research, was funded by the Global Fund to fight HIV/AIDS, Tuberculosis and Malaria (GFATM). The sample size attained was $n = 199$. After providing informed consent, respondents completed an interview and provided blood specimens to be tested for HIV, hepatitis B (HBV), hepatitis C (HCV), and syphilis.

One of the covariates collected with the highest recruitment homophily is education category, with $\mathcal{H}_R = 1.274$. It is plausible that education level could affect recruitment, as people tend to be tied to people with similar educational attainments in their social network. The recruitment plot for the Prizren, Kosovo PWID RDS is shown in Figure 5.1 with nodes colored by a person's highest education level completed (4 levels: no formal education, primary, secondary, and college/university). This plot also shows graphically the $n_s = 4$ seeds and $n_w = 12$ waves of the study. Note that one of the seeds was much more prolific at recruiting than the others. Although the population size N is unknown, Kosovo HIV Integrated Behavioral and Biological Surveillance Survey Reference Group (2014) estimate $N = 1113$ using a combination of unique object, service multiplier, and SS-PSE methods.

For inference, I consider two models: one using only education category, and one using education category along with a network size term.

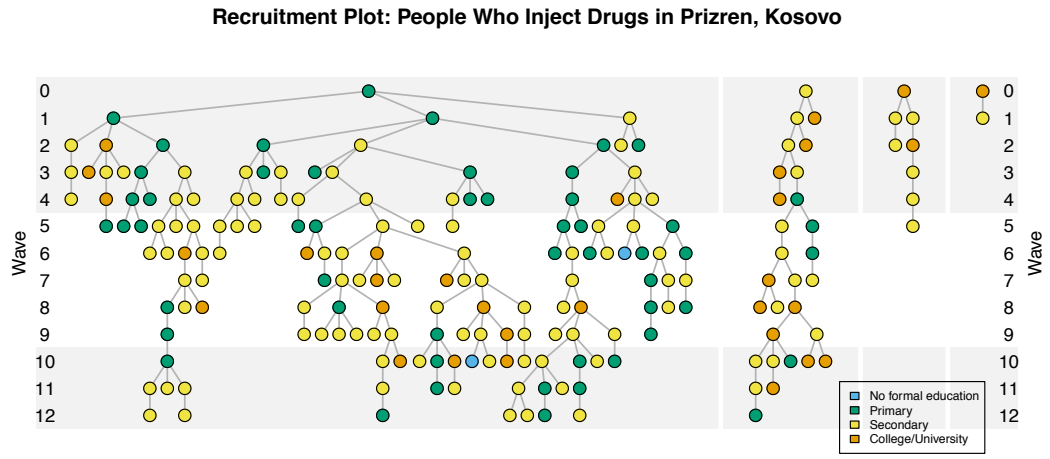


Figure 5.1: Recruitment plot for PWID in Prizren, Kosovo. Points are colored by highest education level completed.

5.1.1 Education Only Model

Figure 5.2 shows the MCMC and density of the posterior distributions. A burn-in of 500 and thinning interval of 5 were used. The medians of the posterior distributions are indicated by red lines as the true values are unknown. The posterior medians for education are -0.047 for recruiters and 0.566 for peers. 95% probability intervals for both of these terms contain zero, so it does not seem that education level has a strong influence on preferential recruitment. The general trend, though, is for people to recruit and be recruited by those in the same education category as themself. This is reflected in the original recruitment homophily on education category value $\mathcal{H}_R = 1.274$.

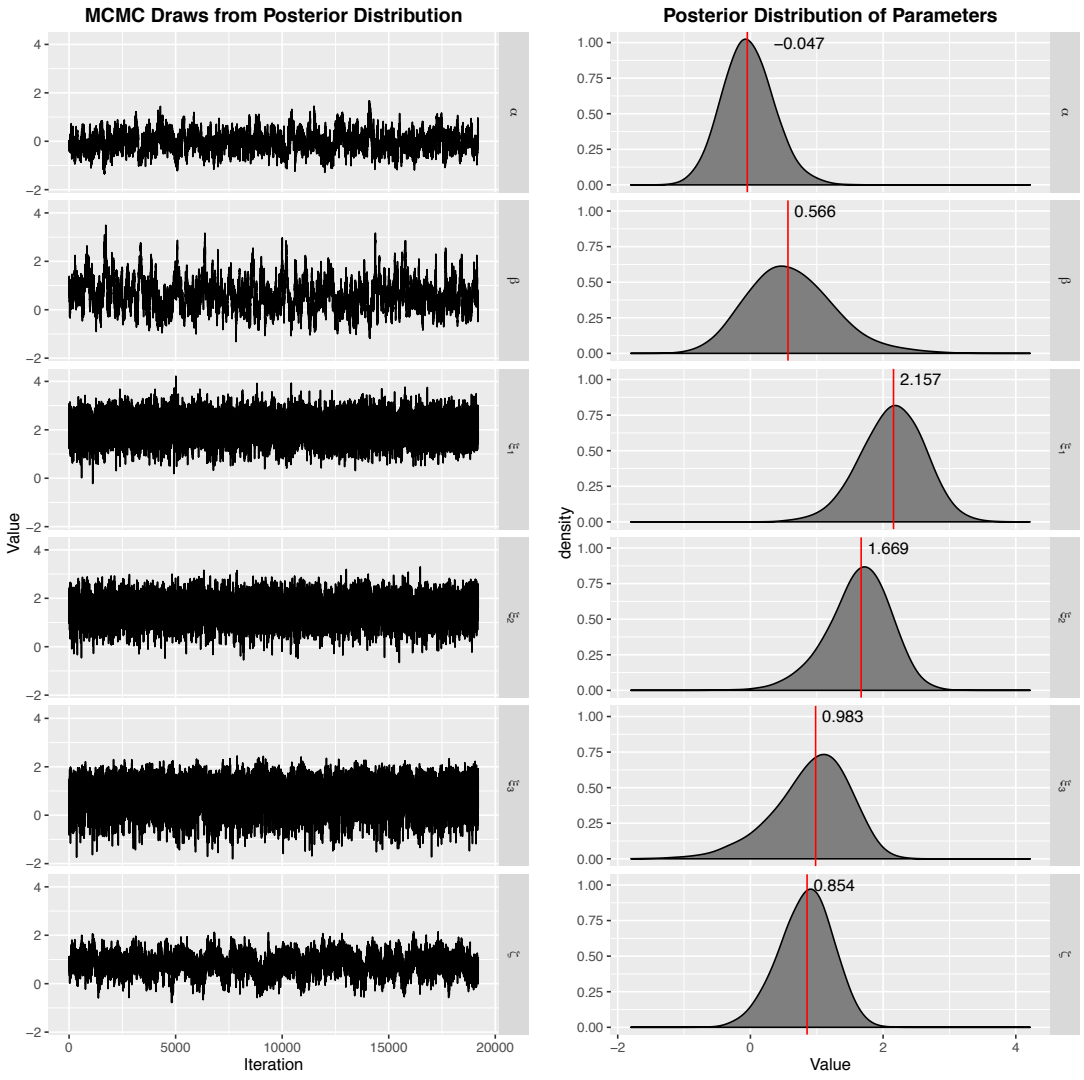


Figure 5.2: MCMC and density of posterior distrubtions for education only model for PWID in Prizren, Kosovo. The red lines show the medians of the posterior distributions.

5.1.2 Education and Network Size Model

Figure 5.3 shows the MCMC and density of the posterior distributions. A burn-in of 500 and thinning interval of 5 were used. The medians of the posterior distributions are indicated by red lines as the true values are unknown. The posterior medians for education are 0.029 for recruiters and 0.402 for peers. 95% probability intervals for both of these terms contain zero, so it does not seem that education level has a strong influence on preferential recruitment. This is consistent with the model without network size. Similarly, the posterior medians for the network size term are 0.099 for recruiters and -0.124 for peers, indicating that network size does not appear to have a strong influence on preferential recruitment.

To investigate the idea that network size does not preferentially affect recruitment, consider the plot of network size by wave in Figure 5.4(a). Network sizes hold relatively constant across waves and do not exhibit the decrease in size due to the depletion of high-degree individuals in the population that commonly occurs in RDS (Handcock et al. (2014); Handcock et al. (2015); Johnston et al. (2015)). For this population consider hepatitis C (HCV) as the outcome of interest. This is because in this particular RDS study, the number of cases of HIV was zero, so estimating a population prevalence does not make sense. The proportion of people in the sample who are HCV positive is 0.201. Figure 5.4(b) shows that HCV positive people tended to have slightly larger network sizes. In the Volz-Heckathorn and Successive Sampling estimators, this will result in a population prevalence

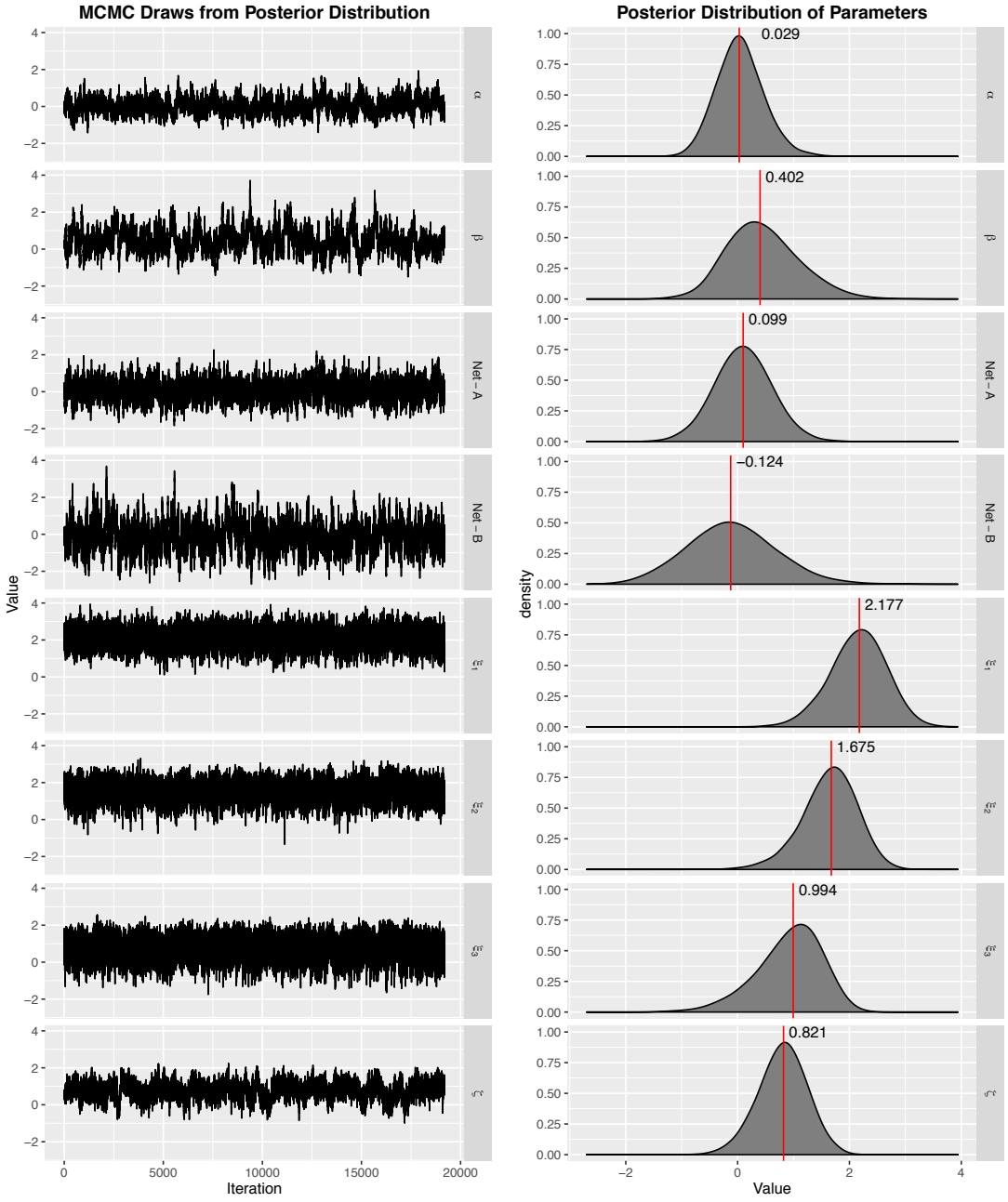


Figure 5.3: MCMC and density of posterior distrubtions for education and network size model for PWID in Prizren, Kosovo. The red lines show the medians of the posterior distributions.

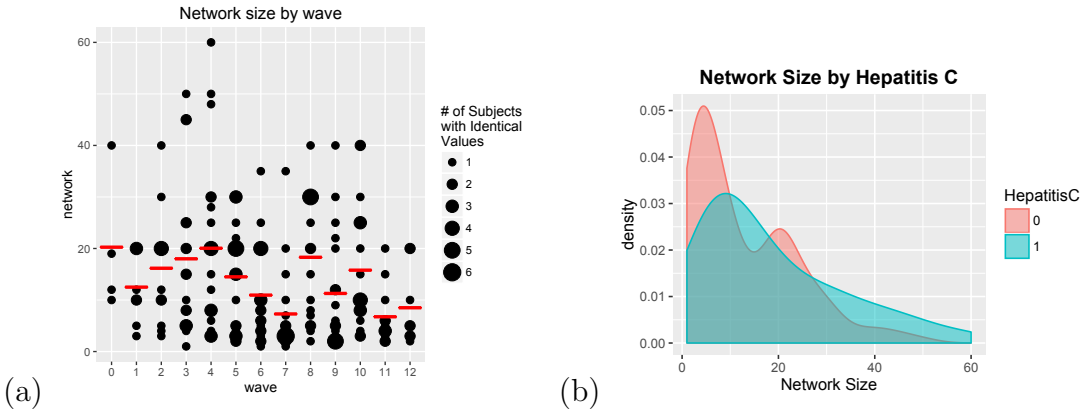


Figure 5.4: Plot of network size by (a) wave and (b) by Hepatitis C for PWID in Prizren, Kosovo.

estimate less than 0.201 since the HCV positive people's responses will be down-weighted more than those who are HCV negative. Because the RCPR model finds that network size is not a large influencer of recruitment, people with larger network sizes will likely have inclusion probabilities that are larger than those using sample weights calculating according to the Volz-Heckathorn or Successive Sampling estimators.

Recall that the specific values of the self-matching parameters are not important, but they will be used to generate new recruitment chains from the population for the calculation of updated inclusion probabilities.

5.1.3 Prevalence Comparison

Table 5.2 provides a comparison of population prevalence estimates for HCV among PWID in Prizren, Kosovo. The RCPR estimates were calculated by simulating 30,000 recruitment chains from the population using the preference pa-

Estimator	**Prevalence Estimate**	95% **Interval**
Sample mean	0.2000	(0.1489, 0.2511)
Salganik-Heckathorn	0.1122	(0.0839, 0.1405)
Volz-Heckathorn	0.1195	(0.0764, 0.1626)
Successive Sampling	0.1280	(0.0844, 0.1717)
RCPR Education Only	0.1898	(0.1137, 0.2659)
RCPR Educ + Net	0.1882	(0.1118, 0.2646)

Table 5.2: Comparison of HCV prevalence estimates for PWID in Prizren, Kosovo.

rameters inferred from the observed recruitment chain. The RCPR estimates are higher, which we would expect as network size does not seem to be important for preferential recruitment. Additionally the education only and education plus network size models give very similar results, again supported by the claim that network size does not seem important for preferential recruitment.

5.2 Francophone Migrants in Rabat, Morocco

An RDS study of Francophone migrants in Rabat, Morocco was conducted during March through April, 2013 after the inclusion of this group in the Morocco National Strategic Plans on HIV/TB for 2012-2016 (Johnston et al., 2016). To be eligible, participants were required to be 18 years or older, have lived in Rabat for at least 3 months, have originated from a sub-Saharan African country (e.g., Senegal, Cameroon, Mali, Cote d'Ivoire, Democratic Republic of the Congo, Guinea), and speak French. The sample size attained was $n = 410$. After providing informed consent, respondents completed an interview and provided blood specimens to be tested for HIV and syphilis.

One of the covariates collected with the highest recruitment homophily is the region in which someone lived in the city (referred to as 'living location' hereafter), with $\mathcal{H}_R = 2.399$. It is very plausible that people will recruit preferentially based on living location, both because they are more likely to be tied to these individuals and because it will likely be easier for them to pass along a coupon. Additionally, in particular for migrant populations, communities in a new country tend to be very insular. The recruitment plot for the Rabat, Morocco Francophone migrant RDS is shown in Figure 5.5 with nodes colored by a person's living location (7 levels; not specified to protect confidentiality). This plot also shows graphically the $n_s = 6$ seeds and $n_w = 6$ waves of the study. Note that no one seed was much more prolific at recruiting than the others. Although the population size N is unknown, Johnston et al. (2015) estimate $N = 4300$ using the SS-PSE method with imputed degree.

For inference, I consider two models: one using only living location, and one using living location along with a network size term.

5.2.1 Living Location Only Model

Figure 5.6 shows the MCMC and density of the posterior distributions. A burn-in of 500 and thinning interval of 5 were used. The medians of the posterior distributions are indicated by red lines as the true values are unknown. The posterior medians for living location are 0.564 for recruiters and 0.516 for peers. The posterior probability that α for living location is above 0 is 0.951, and for

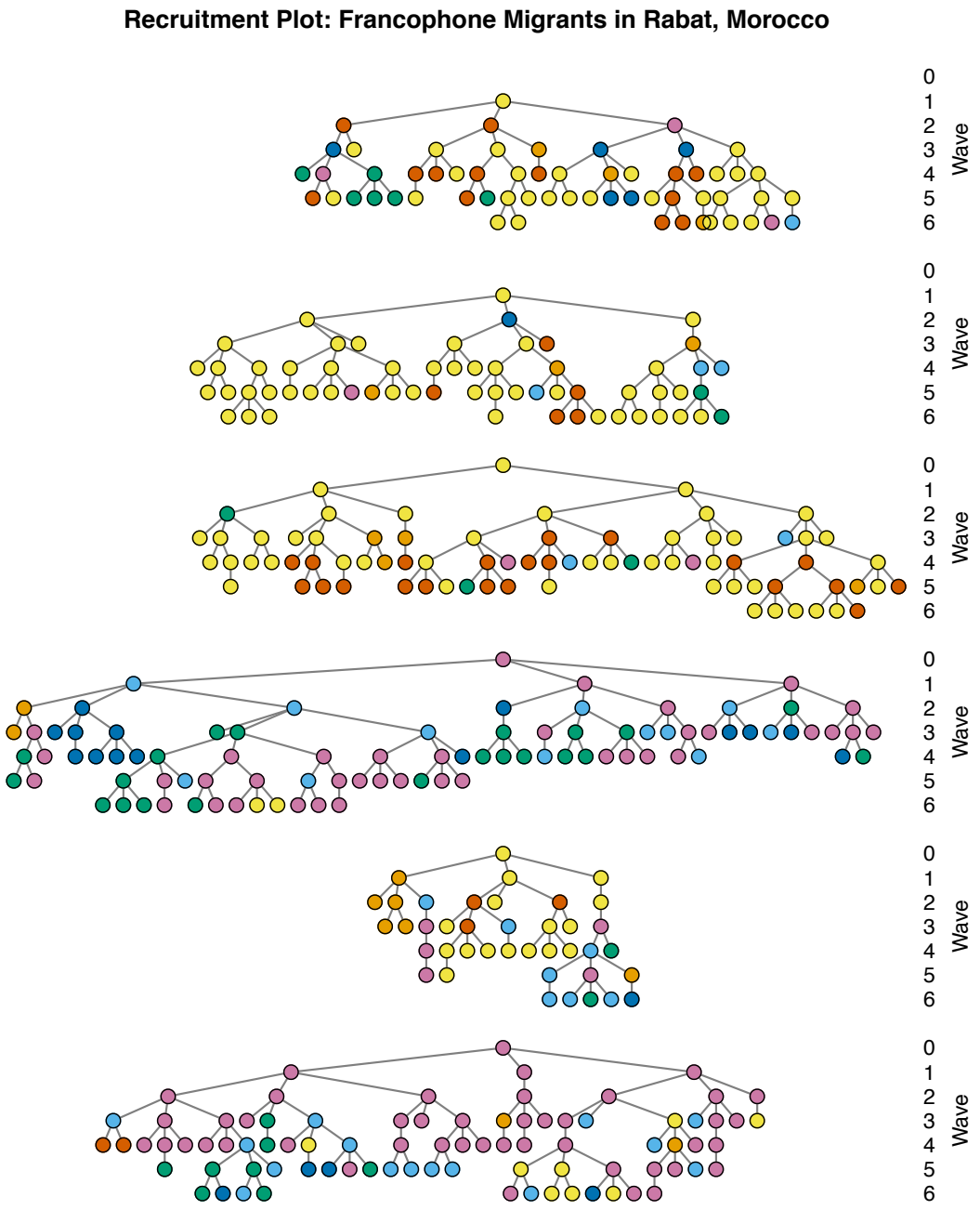


Figure 5.5: Recruitment plot for Francophone migrants in Rabat, Morocco. Points are colored by the region of the city a person lives in.

β it is 0.999. Therefore, living location does seem to have strong influence on recruitment behavior, with people tending to recruit and be recruited by those who live in the same region of the city as themself. This is reflected in the original recruitment homophily on education category value $\mathcal{H}_R = 2.399$.

5.2.2 Living Location and Network Size Model

Figure 5.7 shows the MCMC and density of the posterior distributions. A burn-in of 500 and thinning interval of 5 were used. The medians of the posterior distributions are indicated by red lines as the true values are unknown. The posterior medians for living location are 0.417 for recruiters and 0.552 for peers. The posterior probability that α for living location is above 0 is 0.881, and for β it is 0.998. Therefore, living location does seem to have strong influence on recruitment behavior, with people tending to recruit and be recruited by those who live in the same region of the city as themself. This is reflected in the original recruitment homophily on education category value $\mathcal{H}_R = 2.399$. Similarly, the posterior medians for the network size term are 0.083 for recruiters and -0.339 for peers, indicating that network size does not appear to have a strong influence on preferential recruitment.

To investigate the idea that network size does not preferentially affect recruitment, consider the plot of network size by wave in Figure 5.8(a). Network sizes hold relatively constant across waves (except for the seeds, whose value is greatly inflated by one outlier) and do not exhibit the decrease in size due to the deple-

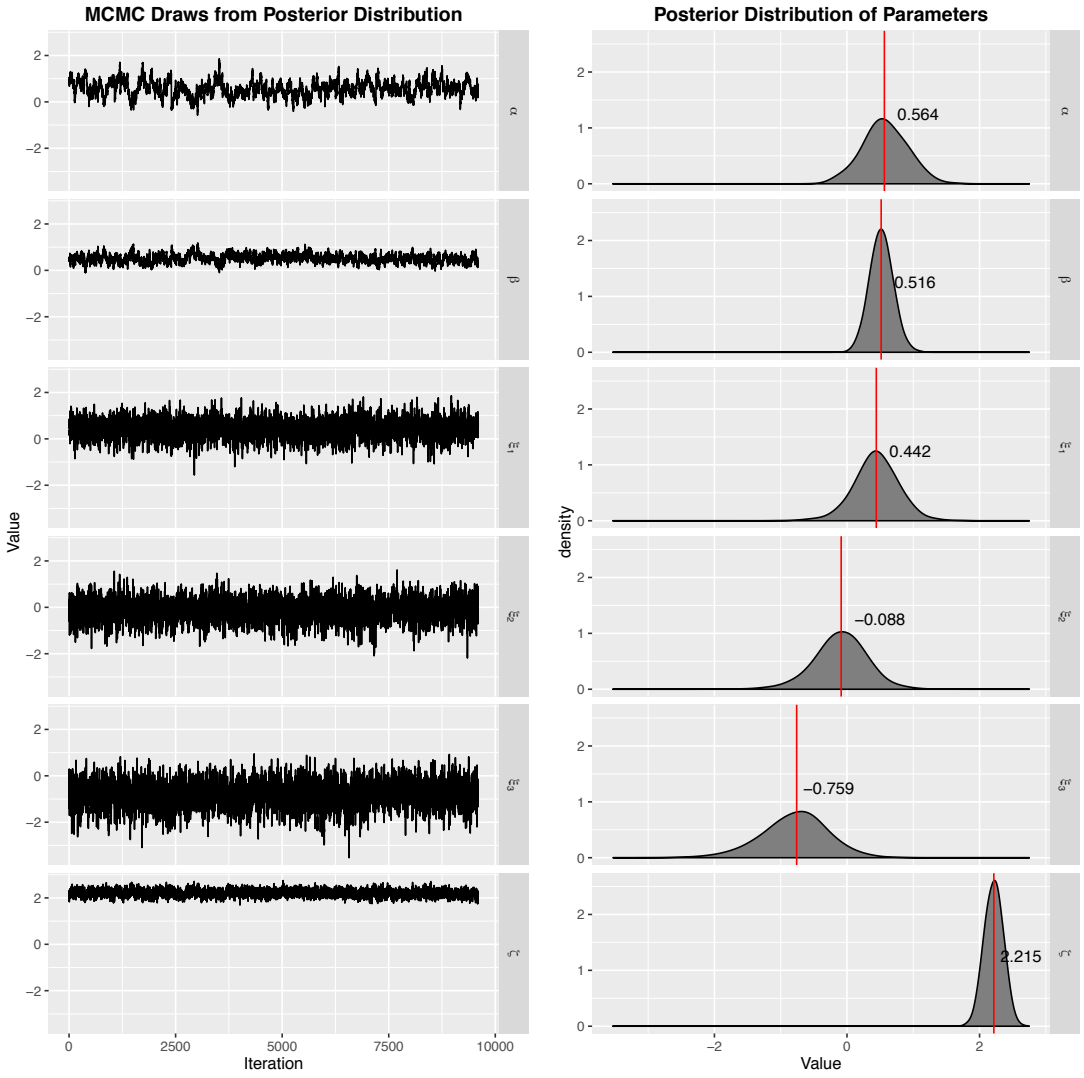


Figure 5.6: MCMC and density of posterior distrubtions for living location only model for Francophone migrants in Rabat, Morocco. The red lines show the medians of the posterior distributions.

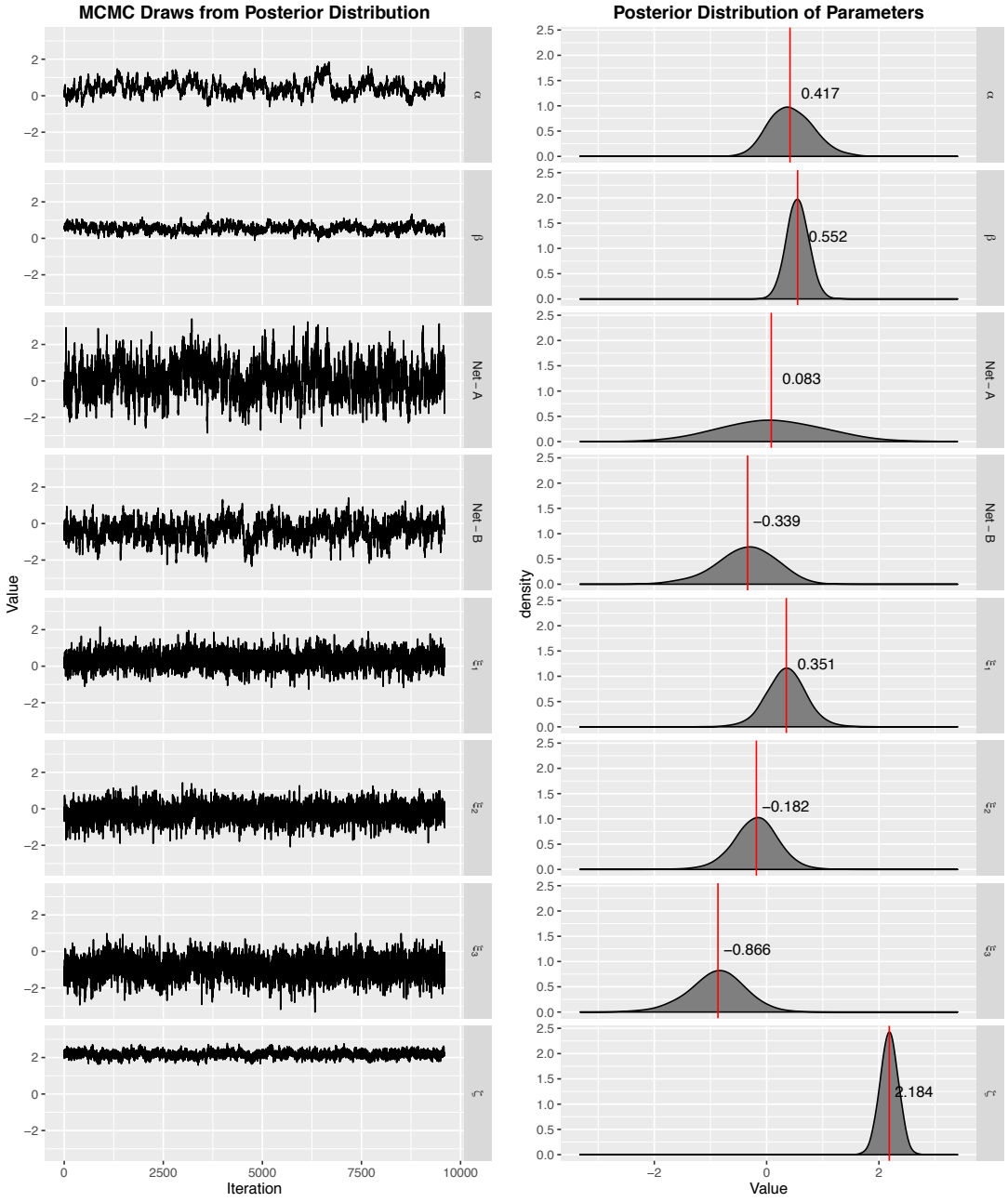


Figure 5.7: MCMC and density of posterior distrubtions for living location and network size model for Francophone migrants in Rabat, Morocco. The red lines show the medians of the posterior distributions.

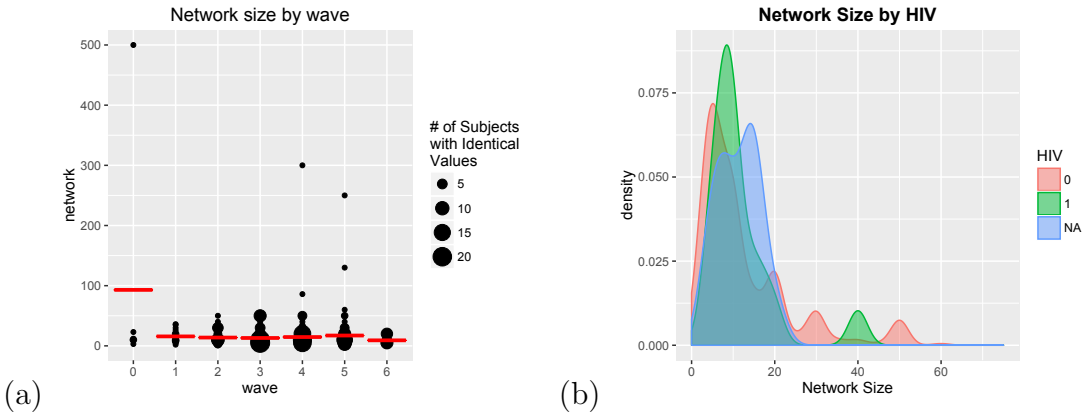


Figure 5.8: Plot of network size by (a) wave and (b) by HIV for Francophone migrants in Rabat, Morocco.

tion of high-degree individuals in the population that commonly occurs in RDS (Handcock et al. (2014); Handcock et al. (2015); Johnston et al. (2015)). For this population consider HIV as the outcome of interest. The proportion of people in the sample who are HIV positive is 0.042, excluding 12 people with missing values. Figure 5.8(b) shows people who were HIV positive have slightly larger network sizes than those who were HIV negative. Note that this plot is zoomed in to focus on the shape of the distribution, so five people with very large network sizes are excluded. These people were all HIV negative. In the Volz-Heckathorn and Successive Sampling estimators, this will result in a population prevalence estimate less than 0.042 since the HIV positive people's responses will be downweighted more than those who are HIV negative. Because the RCPR model finds that network size is not a large influencer of recruitment, people with larger network sizes will likely have inclusion probabilities that are larger than those using sample weights calculating according to the Volz-Heckathorn and Successive Sampling estimators.

Estimator	**Prevalence Estimate**	95% **Interval**
Sample mean	0.0428	(0.0238, 0.0618)
Salganik-Heckathorn	0.0338	(0.0183, 0.0493)
Volz-Heckathorn	0.0328	(0.0144, 0.0512)
Successive Sampling	0.0332	(0.0149, 0.0515)
RCPR Living Only	0.0410	(0.0206, 0.0615)
RCPR Living + Net	0.0416	(0.0190, 0.0642)

Table 5.3: Comparison of HIV prevalence estimates for Francophone migrants in Rabat, Morocco.

Recall that the specific values of the self-matching parameters are not important, but they will be used to generate new recruitment chains from the population for the calculation of updated inclusion probabilities.

5.2.3 Prevalence Comparison

Table 5.3 provides a comparison of population prevalence estimates for HIV among Francophone migrants in Rabat, Morocco. The RCPR estimates were calculated by simulating 20,000 recruitment chains from the population using the preference parameters inferred from the observed recruitment chain. The RCPR estimates are higher, which we would expect as network size does not seem to be important for preferential recruitment. Additionally the living location only and living location plus network size models give very similar results, again supported by the claim that network size does not seem important for preferential recruitment.

5.3 Female Sex Workers in Agadir, Morocco

An RDS study of female sex workers (FSW) in Agadir, Morocco was conducted during December, 2011 through January, 2012 (Johnston et al., 2013a). To be eligible, participants were required to be female, 18 years or older, of Moroccan nationality, work in Agadir, and have reported exchanging penetrative (vaginal/anal) sex for money with more than one male client in the past six months. The sample size attained was $n = 372$. After providing informed consent, respondents completed an interview and provided blood specimens to be tested for HIV and syphilis.

A covariate with only slight recruitment homophily is the woman's number of years of sex work (condensed into three categories; there were three missing values), with $\mathcal{H}_R = 1.097$. It is plausible that FSW may recruit preferentially based on number of years of sex work, as women may be friends with and work in similar places to those with similar experience. The recruitment plot for the Agadir, Morocco FSW RDS is shown in Figure 5.9 with nodes colored by the number of years of sex work. This plot also shows graphically the $n_s = 10$ seeds and $n_w = 8$ waves of the study. Although the population size N is unknown, Johnston et al. (2015) estimate $N = 4900$ using the SS-PSE method with imputed degree.

For inference, I consider two models: one using only number of years of sex work, and one using number of years of sex work along with a network size term.

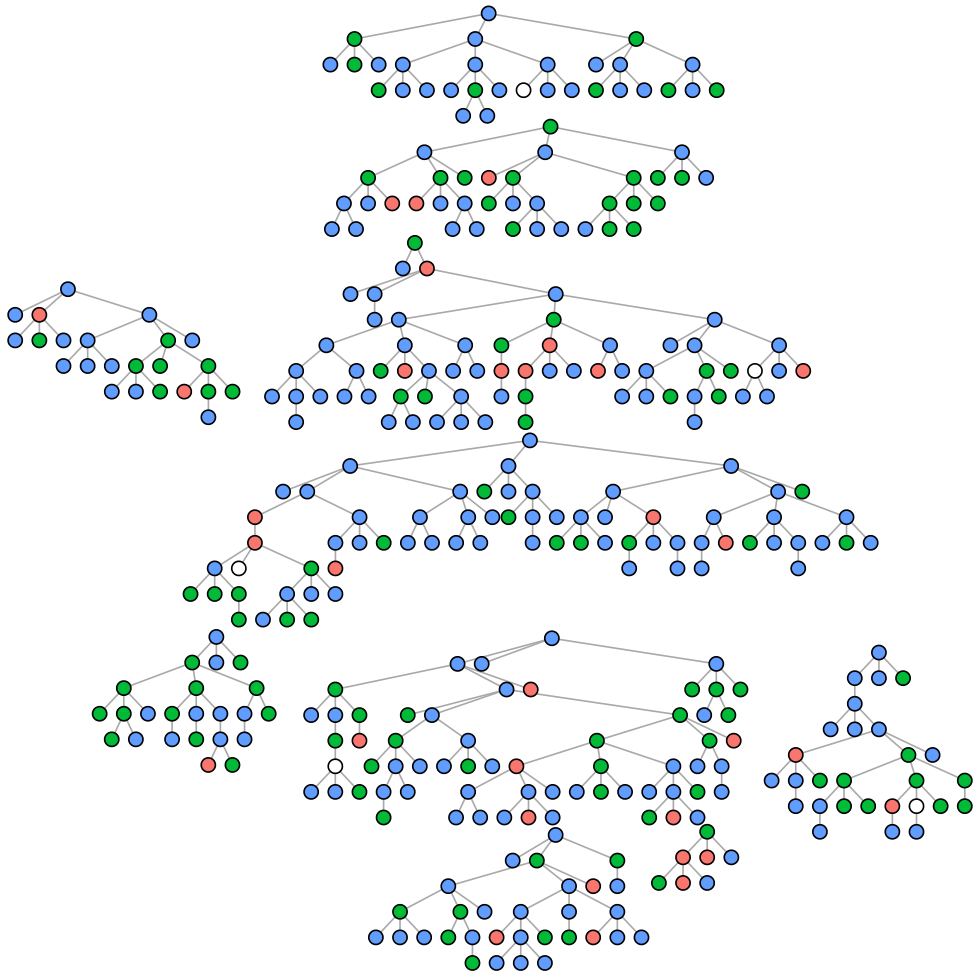


Figure 5.9: Recruitment plot for FSW in Agadir, Morocco. Points are colored by number of years of sex work (condensed into three categories). White nodes indicate missing value.

5.3.1 Number of Years of Sex Work Only model

Figure 5.10 shows the MCMC and density of the posterior distributions. A burn-in of 500 and thinning interval of 5 were used. The medians of the posterior distributions are indicated by red lines as the true values are unknown. The posterior medians for living location are 0.137 for recruiters and 0.139 for peers. 95% probability intervals for both of these terms contain zero, so it does not seem that number of years of sex work has a strong influence on preferential recruitment. This is reflected in the original recruitment homophily on education category value $\mathcal{H}_R = 1.097$, not very different from zero.

5.3.2 Number of Years of Sex Work and Network Size Model

Figure 5.11 shows the MCMC and density of the posterior distributions. A burn-in of 500 and thinning interval of 5 were used. The medians of the posterior distributions are indicated by red lines as the true values are unknown. The posterior medians for living location are 0.232 for recruiters and 0.128 for peers. 95% probability intervals for both of these terms contain zero, so it does not seem that number of years of sex work has a strong influence on preferential recruitment. This is consistent with the model without network size.

The posterior medians for the network size term are 0.898 for recruiters and 0.480 for peers. These values are larger than in the examples in Sections 5.1.2 and 5.2.2. The posterior probability that α for network size is above 0 is 0.887,

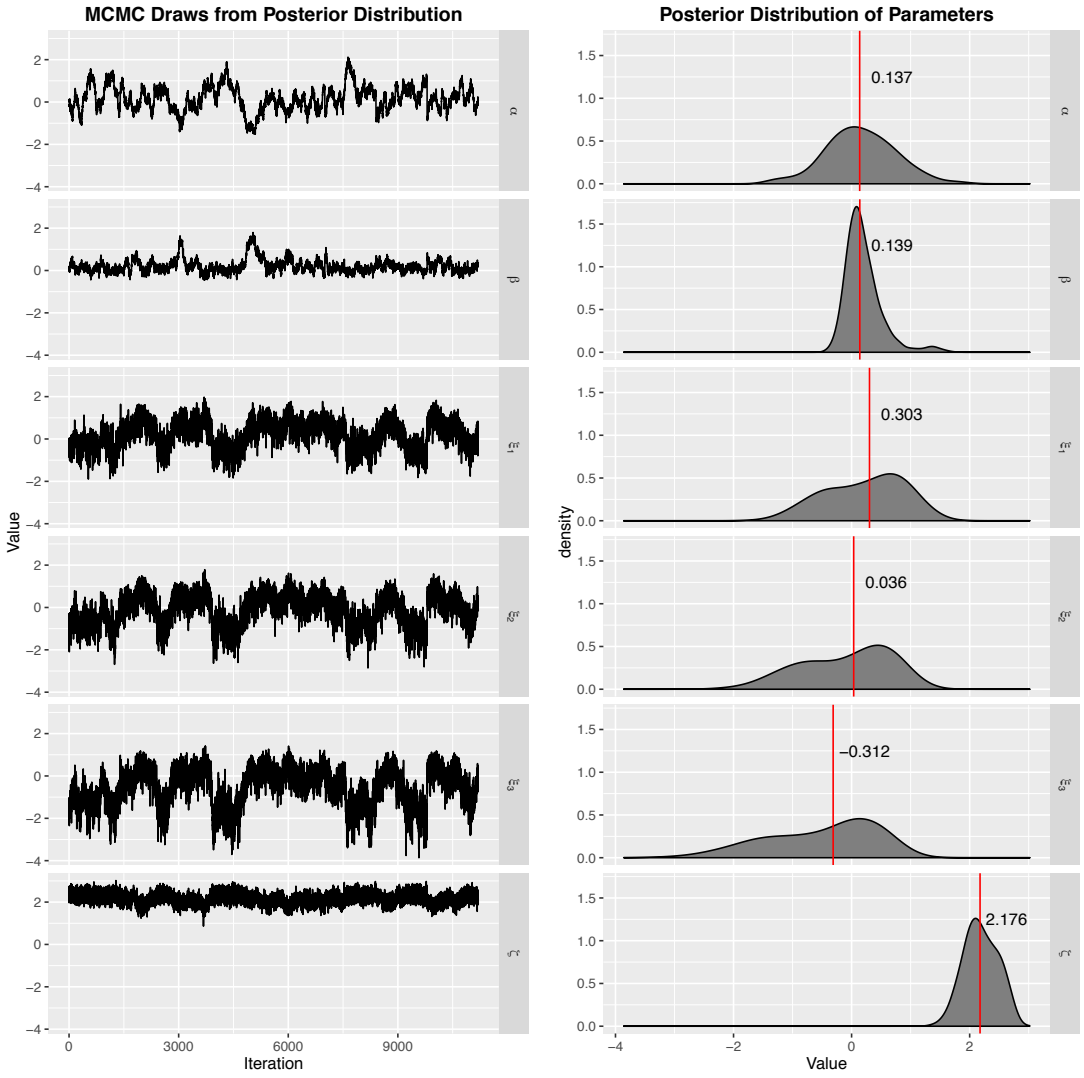


Figure 5.10: MCMC and density of posterior distrubtions for number of years of sex work only model for FSW in Agadir, Morocco. The red lines show the medians of the posterior distributions.

and for β for network size it is 0.906. This suggests that network size does play a role in recruitment for FSW in Agadir Morocco.

To investigate the idea that network size does preferentially affect recruitment in this case, consider the plot of network size by wave in Figure 5.12(a). Network sizes hold seem to decrease slightly across waves, as hypothesized commonly occurs in RDS (Handcock et al. (2014); Handcock et al. (2015); Johnston et al. (2015)). For this population consider HIV as the outcome of interest. The proportion of people in the sample who are HIV positive is 0.048, excluding 8 people with missing values. Figure 5.12(b) shows no major difference in network size among people who are HIV positive and negative. In the Volz-Heckathorn and Successive Sampling estimators, this will result in a population prevalence estimate that are similar to 0.048. Because the RCPR model finds that network size is a large influencer of recruitment, people with larger network sizes will likely have inclusion probabilities that are similar to those using sample weights calculating according to the Volz-Heckathorn and Successive Sampling estimators.

Recall that the specific values of the self-matching parameters are not important, but they will be used to generate new recruitment chains from the population for the calculation of updated inclusion probabilities.

5.3.3 Prevalence Comparison

Table 5.4 provides a comparison of population prevalence estimates for HIV among FSW in Agadir, Morocco. The RCPR estimates were calculated by simulating

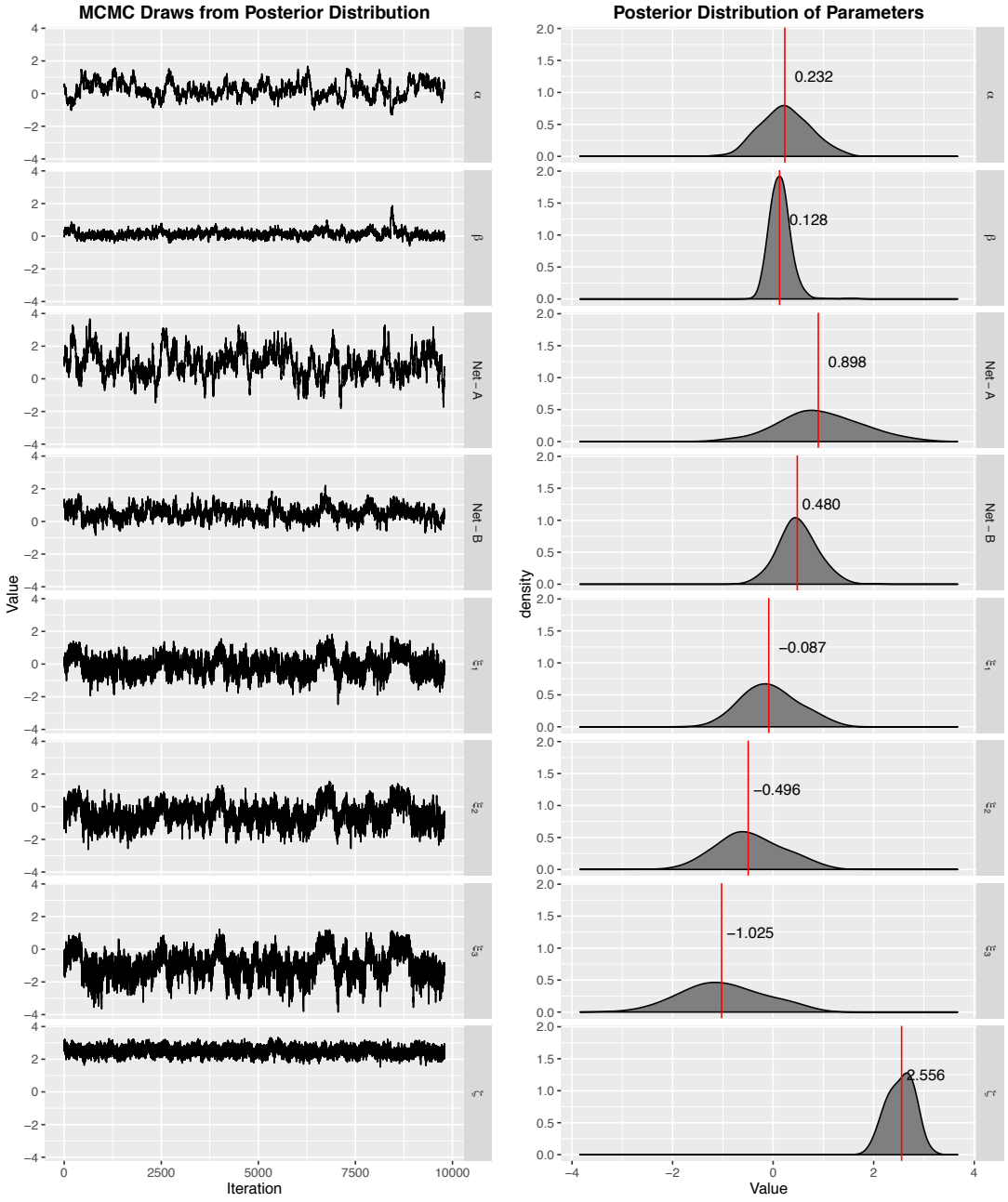
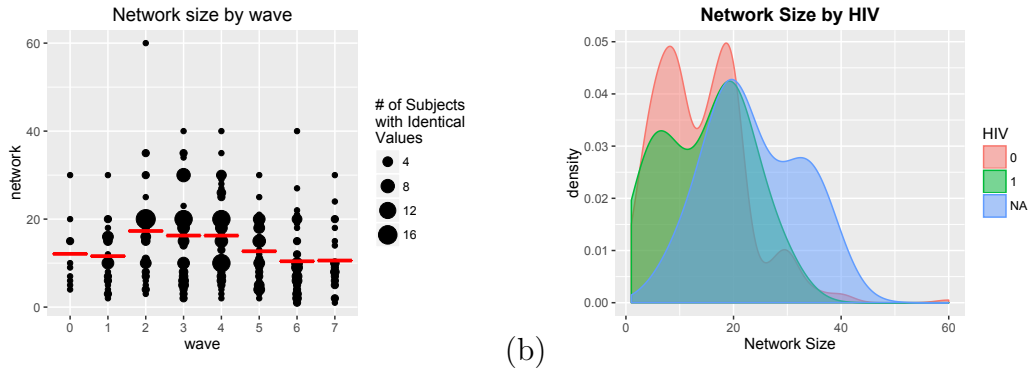


Figure 5.11: MCMC and density of posterior distrubtions for number of years of sex work and network size model for FSW in Agadir, Morocco. The red lines show the medians of the posterior distributions.



(a) (b)

Figure 5.12: Plot of network size by (a) wave and (b) by HIV for FSW in Agadir, Morocco.

Estimator	**Prevalence Estimate**	95% **Interval**
Sample mean	0.0500	(0.0283, 0.0717)
Salganik-Heckathorn	0.0494	(0.0231, 0.0757)
Volz-Heckathorn	0.0488	(-0.0090, 0.1066)
Successive Sampling	0.0494	(-0.0105, 0.1093)
RCPR Yrs Sex Work Only	0.0496	(0.0301, 0.0691)
RCPR Yrs Sex Work + Net	0.0486	(0.0289, 0.0682)

Table 5.4: Comparison of HIV prevalence estimates for FSW in Agadir, Morocco.

20,000 recruitment chains from the population using the preference parameters inferred from the observed recruitment chain. Observe that unlike the examples in Sections 5.1.3 and 5.2.3, here the estimates using the RCPR model are almost identical to those using the other estimators. This could be because for this population degree seemed to influence preferential recruitment. In the other scenarios, degree did not seem as important, so the estimators that adjust sample weights based on degree could have produced an inaccurate estimate.

Figure 5.13 shows a comparison of the inclusion probabilities using the Successive Sampling weights and the inclusion probabilities using the RCPR model for the years of sex work only model and the years of sex work plus network

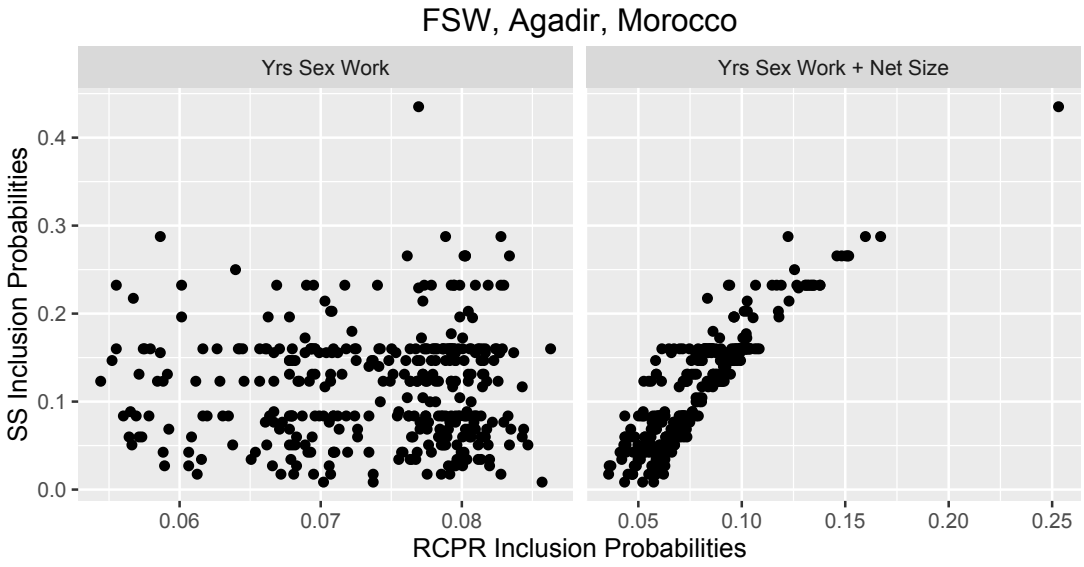


Figure 5.13: Comparison of SS and RCPR inclusion probabilities in the years of sex work only model and the years of sex work plus network size model for FSW in Agadir, Morocco.

size model. The years of sex work only model plot shows no strong relationship between the two sets of inclusion probabilities, but the years of sex work plus network size model demonstrates a clear positive relationship between the two. This indicates that the RCPR model with a network size term does a good job of retaining the influence of network size on the sample weights of individuals in the study.

CHAPTER 6

Conclusion

Respondent-driven sampling (RDS) is used throughout the world to gather information on key populations at high risk for HIV/AIDS. A full RDS study incorporates both the data gathering mechanism and the analysis methodology. Common RDS estimators rely on a series of assumptions about the sampling process that may not hold in practice. One of these is the assumption that people recruit their peers uniformly at random from their social network. This assumption is used to obtain sample weights for individuals in the study based on the assumption that recruitment is a random walk over the underlying social network. In this method, a person's network size (degree) is the only variable that affects their sample weight. In practice, we think that other factors also may influence an individual's inclusion probability and people likely recruit with preference for certain members of their network based on nodal or dyadic characteristics.

I therefore propose a two-sided rational choice framework as a method to model preferential recruitment. This model, referred to as the RCPR model, mimics the way we think people actually recruit their peers into the study. Further, it and allows us to incorporate and identify variables important for preferential

recruitment, potentially including network size, rather than using only network size. This framework is a generative network model for the RDS sampling process, and thus the preferences for coefficients obtained through inference can be used in the simulation of additional recruitment chains with the same properties from a larger population. By measuring the frequency with which individuals and dyads in the original sample appear in these simulated recruitment chains, we can get direct estimates of both first-order and second-order inclusion probabilities. These inclusion probabilities can then be used in a generalized Horvitz-Thompson estimator to calculate prevalence estimates for desired outcome measures such as HIV prevalence.

References

Admiraal, R. *Models for Heterogeneity in Heterosexual Partnership Networks*. Ph.d., University of Washington, 2009.

Albert, J. H. and Chib, S. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Barbosa Júnior, A., Pascom, A. R. P., Szwarcwald, C. L., Kendall, C., and McFarland, W. Transfer of Sampling Methods for Studies on Most-At-Risk Populations (MARPs) in Brazil. *Cadernos de Saúde Pública*, 27:36–44, 2011.

Bengtsson, L. and Thorson, A. Global HIV Surveillance Among MSM: Is Risk Behavior Seriously Underestimated? *AIDS*, 24:2301–2303, 2010.

Bengtsson, L., Lu, X., Nguyen, Q. C., Camitz, M., Hoang, N. L., Nguyen, T. A., Liljeros, F., and Thorson, A. Implementation of Web-Based Respondent-Driven Sampling Among Men Who Have Sex With Men in Vietnam. *PloS ONE*, 7(11): e49417, 2012.

Broadhead, R. S. Peer-Driven Intervention. Technical report, 2009.

Broadhead, R. S., Heckathorn, D. D., Weakliem, D. L., Anthony, D. L., Madray, H., Mills, R. J., and Hughes, J. Harnessing Peer Networks as an Instrument for AIDS Prevention: Results from a Peer-Driven Intervention. *Public Health Reports*, 113(Supplement I):42–57, 1998.

Broadhead, R. S., Heckathorn, D. D., Altice, F. L., van Hulst, Y., Carbone, M., Friedland, G. H., O'Connor, P. G., and Selwyn, P. A. Increasing Drug Users' Adherence to HIV Treatment: Results of a Peer-Driven Intervention Feasibility Study. *Social Science & Medicine*, 55:235–246, 2002.

Broadhead, R. S., Volkanevsky, V. L., Rydanova, T., Ryabkova, M., Borch, C., van Hulst, Y., Fullerton, A., Sergeyev, B., and Heckathorn, D. D. Peer-Driven HIV Interventions for Drug Injectors in Russia: First Year Impact Results of a Field Experiment. *International Journal of Drug Policy*, 17(5):379–392, 2006.

Broadhead, R. S., Hammett, T. M., Kling, R., Ngu, D., Liu, W., Chen, Y., Quyen, H. N., Binh, K. T., Van Tren, H., Borch, C., Gauchat, G., and Des Jarlais, D. C. Peer-Driven Interventions in Vietnam and China to Prevent HIV: A Pilot Study Targeting Injection Drug Users. *Journal of Drug Issues*, 39(4):829–850, 2009.

Casella, G. and George, E. I. Explaining the Gibbs Sampler. *The American*

Statistician, 46(3):167–174, 1992.

Crawford, F. W., Aronow, P. M., Zeng, L., and Li, J. Identification of Homophily and Preferential Recruitment in Respondent-Driven Sampling. 2015.

de Mello, M., de Pinho, A. A., Chinaglia, M., Tun, W., Júnior, A. B., Ilário, M. C. F. J., Reis, P., Salles, R. C. S., Westman, S., and Díaz, J. Assessment of Risk Factors for HIV Infection among Men Who have Sex with Men in the Metropolitan Area of Campinas City, Brazil, using Respondent-Driven Sampling. Technical report, Population Council, New York, 2008.

Deering, K. N., Shannon, K., Sinclair, H., Parsad, D., Gilbert, E., and Tyndall, M. W. Piloting a Peer-Driven Intervention Model to Increase Access and Adherence to Antiretroviral Therapy and HIV Care among Street-Entrenched HIV-Positive Women in Vancouver. *AIDS Patient Care and STDs*, 23(8), 2009.

Dombrowski, K., Khan, B., Wendel, T., Mclean, K., Misshula, E., and Curtis, R. Estimating the Size of the Methamphetamine-Using Population in New York City Using Network Sampling Techniques. *Advances in Applied Sociology*, 2(4): 245–252, 2012.

Engel, S., Gwadz, M. V., and Cleland, C. M. The Experience of Screening for HIV/AIDS Medical Studies among African-American/Black and Latino/Hispanic Persons Living with HIV/ AIDS: A Mixed-Methods Exploration. *Journal of AIDS & Clinical Research*, 4(7), 2013.

Frank, O. *International Encyclopedia of Statistical Science*, chapter Network Sampling, pages 941–942. Springer Berlin Heidelberg, 2011.

Gale, D. and Shapley, L. S. College Admissions and the Stability of Marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, 2nd edition, 2004.

Gile, K. J. Improved Inference for Respondent-Driven Sampling Data With Application to HIV Prevalence Estimation. *Journal of the American Statistical Association*, 106(493):135–146, 2011.

Gile, K. J. and Handcock, M. S. Respondent-Driven Sampling: An Assessment of Current Methodology. *Sociological Methodology*, 40(1):285–327, 2010.

Gile, K. J., Johnston, L. G., and Salganik, M. J. Diagnostics for Respondent-Driven Sampling. *Journal of the Royal Statistical Society: Series A*, 178(1): 241–269, 2015.

Goel, S. and Salganik, M. J. Respondent-Driven Sampling as Markov Chain Monte Carlo. *Statistics in Medicine*, 28:2202–2229, 2009.

Goel, S. and Salganik, M. J. Assessing Respondent-Driven Sampling. *Proceedings of the National Academy of Sciences*, 107(15):6743–6747, 2010.

Goodman, L. A. Snowball Sampling. *The Annals of Mathematical Statistics*, 32 (1):148–170, 1961.

Granovetter, M. Network Sampling: Some First Steps. *The American Journal of Sociology*, 81(6):1287–1303, 1976.

Gwadz, M., Cleland, C. M., Leonard, N. R., Ritchie, A. S., Banfield, A., Riedel, M., Colon, P., and Mildvan, D. Predictors of Screening for AIDS Clinical Trials Among African-Americans and Latino/Hispanics Enrolled in an Efficacious Peer-Driven Intervention: Uncovering Socio-Demographic, Health, and Substance Use-Related Factors that Promote or Impede Screening. *AIDS and Behavior*, 17(2):801–12, 2013.

Gwadz, M. V., Leonard, N. R., Cleland, C. M., Riedel, M., Banfield, A., and Mildvan, D. The Effect of Peer-Driven Intervention on Rates of Screening for AIDS Clinical Trials Among African Americans and Hispanics. *American Journal of Public Health*, 101(6):1096–102, 2011.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *Journal of Statistical Software*, 24(1), 2008.

Handcock, M. S., Fellows, I. E., and Gile, K. J. RDS: Respondent-Driven Sampling, 2012.

Handcock, M. S., Gile, K. J., and Mar, C. M. Estimating Hidden Population Size Using Respondent-Driven Sampling Data. *Electronic Journal of Statistics*, 8 (1):1491–1521, 2014.

Handcock, M. S., Gile, K. J., and Mar, C. M. Estimating the Size of Populations at High Risk for HIV Using Respondent-Driven Sampling Data. *Biometrics*, 71 (1):258–266, 2015.

Heckathorn, D. D. Collective Sanctions and Compliance Norms: A Formal Theory of Group-Mediated Social Control. *American Sociological Review*, 55(3):366–384, 1990.

Heckathorn, D. D. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*, 44(2):174–199, 1997.

Heckathorn, D. D. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Social Problems*, 49(1):11–34, 2002.

Heckathorn, D. D. and Broadhead, R. S. Rational Choice, Public Policy and AIDS. *Rationality and Society*, 8(3):235–260, 1996.

Heckathorn, D. D., Broadhead, R. S., Anthony, D. L., and Weakliem, D. L. AIDS and Social Networks: HIV Prevention Through Network Mobilization. *Sociological Focus*, 32(2):159–179, 1999.

Heimer, R. Critical Issues and Further Questions About Respondent-Driven Sampling: Comment on Ramirez-Valles, et al. (2005). *AIDS and Behavior*, 9(4): 403–408, 2005.

Horvitz, D. G. and D.J., T. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47 (260):663–685, 1952.

Iguchi, M. Y., Ober, A. J., Berry, S. H., Fain, T., Heckathorn, D. D., Gorbach, P. M., Heimer, R., Kozlov, A., Ouellet, L. J., Shoptaw, S., and Zule, W. A. Simultaneous Recruitment of Drug Users and Men Who Have Sex with Men in the United States and Russia using Respondent-Driven Sampling: Sampling Methods and Implications. *Journal of Urban Health*, 86(SUPPL. 1):5–31, 2009.

Johnston, L., Bennani, A., Latifi, A., Oumzil, H., El Omari, B., El Rhoufrani, F., Ouarsas, L., Alami, K., and El Rhilani, H. Using Respondent-Driven Sampling to Estimate HIV and Syphilis Prevalence Among Female Sex Workers in Agadir, Fes, Rabat and Tangier, Morocco. *Sexually Transmitted Infections*, 89(Suppl 1):A180, 2013a.

Johnston, L., Oumzil, H., El Rhilani, H., Latifi, A., Bennani, A., and Alami, K. Sex Differences in HIV Prevalence, Behavioral Risks and Prevention Needs Among Anglophone and Francophone Sub-Saharan African Migrants Living in Rabat, Morocco. *AIDS and Behavior*, 20(4):746–753, 2016.

Johnston, L. G. Behavioural Surveillance: Introduction to Respondent Driven Sampling, 2008.

Johnston, L. G., Chen, Y.-H., Silva-Santisteban, A., and Raymond, H. F. An Empirical Examination of Respondent Driven Sampling Design Effects Among HIV Risk Groups from Studies Conducted Around the World. *AIDS and Behavior*, 17:2202–2210, 2013b.

Johnston, L. G., McLaughlin, K. R., El Rhilani, H., Latifi, A., Toufik, A., Ben-

nani, A., Alami, K., Elomari, B., and Handcock, M. S. A Novel Method for Estimating the Size of Hidden Populations Ssing Respondent-Driven Sampling Data: Case Examples from Morocco. *Epidemiology*, 26(6):846–852, 2015.

Kass, R. E. and Raftery, A. E. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Kerr, L. R. F. S., Kendall, C., Pontes, M. K., Werneck, G. L., McFarland, W., Mello, M. B., Martins, T. A., and Macena, R. H. M. Selective Participation in a RDS Survey Among MSM in Ceara, Brazil: A Qualitative and Quantitative Assessment. *Jornal Brasileiro de Doenças Sexualmente Transmissíveis*, 23(3): 126–133, 2011.

Kosovo HIV Integrated Behavioral and Biological Surveillance Survey Reference Group. HIV Integrated Behavioral and Biological Surveillance Surveys - Kosovo. Technical report, 2014.

Lansky, A., Abdul-Quader, A. S., Cribbin, M., Hall, T., Finlayson, T. J., Garfein, R. S., Lin, L. S., and Sullivan, P. S. Developing an HIV Behavioral Surveillance System for Injecting Drug Users: The National HIV Behavioral Surveillance System. *Public Health Reports*, 122(Supplement 1):48–55, 2007.

Leonard, N. R., Banfield, A., Riedel, M., Ritchie, A. S., Mildvan, D., Arredondo, G., Cleland, C. M., and Gwadz, M. V. Description of an Efficacious Behavioral Peer-Driven Intervention to Reduce Racial/Ethnic Disparities in AIDS Clinical Trials. *Health Education Research*, 28(4):574–90, 2013.

Lin, N. Building a Network Theory of Social Capital. *Connections*, 22(1):28–51, 1999.

Logan, J. A. Estimating Two-Sided Logit Models. *Sociological Methodology*, 28 (1):139–173, 1998.

Logan, J. A., Hoff, P. D., and Newton, M. A. Two-Sided Estimation of Mate Preferences for Similarities in Age, Education, and Religion. *Journal of the American Statistical Association*, 103(482):559–569, 2008.

Long, S. J. *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, Thousand Oaks, CA, 1997.

Lu, X. Linked Ego Networks: Improving Estimate Reliability and Validity with Respondent-Driven Sampling. *Social Networks*, 35(4):669–685, 2013.

Magnani, R., Sabin, K., Saidel, T., and Heckathorn, D. Review of Sampling Hard-to-Reach and Hidden Populations for HIV Surveillance. *AIDS*, 19(Suppl

2):S67–S72, 2005.

Malekinejad, M., Johnston, L. G., Kendall, C., Kerr, L. R. F. S., Rifkin, M. R., and Rutherford, G. W. Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review. *AIDS and Behavior*, 12(SUPPL. 1):105–130, 2008.

Manski, C. F. Identification of Binary Response Models. *Journal of the American Statistical Association*, 83(403):729–738, 1988.

Marsden, P. V. Network Data and Measurement. *Annual Review of Sociology*, 16: 435–463, 1990.

Martin, J. L., Wiley, J., and Osmond, D. Social Networks and Unobserved Heterogeneity in Risk for AIDS. *Population Research and Policy Review*, 22:65–90, 2003.

McCreesh, N., Frost, S., Seeley, J., Katongole, J., Tarsh, M. N., Ndunguse, R., Jichi, F., Lunel, N. L., Maher, D., Johnston, L. G., Sonnenberg, P., Copas, A. J., Hayes, R. J., and White, R. G. Evaluation of Respondent-Driven Sampling. *Epidemiology*, 23(1):138–47, 2012.

McCreesh, N., Tarsh, M. N., Seeley, J., Katongole, J., and White, R. G. Community Understanding of Respondent-Driven Sampling in a Medical Research Setting in Uganda: Importance for the Use of RDS for Public Health Research. *International Journal of Social Research Methodology*, 16(4):269–284, 2013.

McLaughlin, K. R., Handcock, M. S., and Johnston, L. G. Inference for the Visibility Distribution for Respondent-Driven Sampling. In *Proceedings of the Joint Statistical Meetings, Social Statistics Section*, pages 2259–2267, Alexandria, VA, 2015.

Mcpherson, M., Smith-lovin, L., and Cook, J. M. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27:415–444, 2001.

Montealegre, J. R., Johnston, L. G., Murrill, C., and Monterroso, E. Respondent Driven Sampling for HIV Biological and Behavioral Surveillance in Latin America and the Caribbean. *AIDS and Behavior*, 17(7):2313–2340, 2013.

Onyx, J. and Bullen, P. Measuring Social Capital in Five Communities. *The Journal of Applied Behavioral Science*, 36(1):23–42, 2000.

Ramos, R. L., Green, N. L., and Shulman, L. C. Pasa la Voz: Using Peer Driven Interventions to Increase Latinas' Access to and Utilization of HIV Prevention and Testing Services. *Journal of Health Care for the Poor and Underserved*, 20:

29–35, 2009.

Reingold, E. and Tilford, J. Tidier Drawings of Trees. *IEEE Transactions on Software Engineering*, SE-7(2):223–228, 1981.

Roth, A. The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory. *The Journal of Political Economy*, 92(6):991–1016, 1984.

Roth, A. E. and Sotomayor, M. A. O. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge University Press, 1990.

Roth, A. E. and Vande Vate, J. H. Incentives in Two-Sided Matching with Random Stable Mechanisms. *Economic Theory*, 1:31–44, 1991.

Rudolph, A. E., Fuller, C. M., and Latkin, C. The Importance of Measuring and Accounting for Potential Biases in Respondent-Driven Samples. *AIDS and Behavior*, 17(6):2244–52, 2013.

Sadasivam, R. S., Volz, E. M., Kinney, R. L., Rao, S. R., and Houston, T. K. Share2Quit: Web-Based Peer-Driven Referrals for Smoking Cessation. *JMIR Research Protocols*, 2(2), 2013.

Salganik, M. J. and Heckathorn, D. D. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34 (1):193–240, 2004.

Scott, G. "They Got Their Program, and I Got Mine": A Cautionary Tale Concerning the Ethical Implications of Using Respondent-Driven Sampling to Study Injection Drug Users. *International Journal of Drug Policy*, 19(1):42–51, 2008.

Semaan, S., Santibanez, S., Garfein, R. S., Heckathorn, D. D., and Des Jarlais, D. C. Ethical and Regulatory Considerations in HIV Prevention Studies Employing Respondent-Driven Sampling. *International Journal of Drug Policy*, 20 (1):14–27, 2009.

Smyrnov, P., Broadhead, R. S., Datsenko, O., and Matiyash, O. Rejuvenating Harm Reduction Projects for Injection Drug Users: Ukraine's Nationwide Introduction of Peer-Driven Interventions. *The International Journal on Drug Policy*, 23(2):141–7, 2012.

Solomon, S. S., Lucas, G. M., Celentano, D. D., Sifakis, F., and Mehta, S. H. Beyond Surveillance: A Role for Respondent-Driven Sampling in Implementation Science. *American Journal of Epidemiology*, 178(2):260–7, 2013.

Stein, M. L., van Steenbergen, J. E., Chanyasanha, C., Tipayamongkholgul, M., Buskens, V., van der Heijden, P. G. M., Sabaiwan, W., Bengtsson, L., Lu, X., Thorson, A. E., and Kretzschmar, M. E. E. Online Respondent-Driven Sampling for Studying Contact Patterns Relevant for the Spread of Close-Contact Pathogens: A Pilot Study in Thailand. *PloS ONE*, 9(1):e85256, 2014.

Sugden, R. and Smith, T. M. F. Ignorable and Informative Designs in Survey Sampling Inference. *Biometrika*, 71(3):495–506, 1984.

Tanner, M. A. and Wong, W. H. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398): 528–540, 1987.

Thompson, S. K. *Sampling*. Wiley, 3rd edition, 2012.

Tierney, L. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

Tomas, A. and Gile, K. J. The Effect of Differential Recruitment, Non-Response and Non-Recruitment on Estimators for Respondent-Driven Sampling. *Electronic Journal of Statistics*, 5:899–934, 2011.

Volz, E. and Heckathorn, D. D. Probability Based Estimation Theory for Respondent-Driven Sampling. *Journal of Official Statistics*, 24(1):79–97, 2008.

Wattana, W., van Griensven, F., Rhucharoenpornpanich, O., Manopaiboon, C., Thienkrua, W., Bannatham, R., Fox, K., Mock, P. A., Tappero, J. W., and Levine, W. C. Respondent-Driven Sampling to Assess Characteristics and Estimate the Number of Injection Drug Users in Bangkok, Thailand. *Drug and Alcohol Dependence*, 90:228–233, 2007.

Wejnert, C. An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and Out-of-Equilibrium Data. *Sociological Methodology*, 39(1):73–116, 2009.

Wejnert, C. and Heckathorn, D. D. Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling. *Sociological Methods & Research*, 37 (1):105–134, 2008.

Wejnert, C., Pham, H., Krishna, N., Le, B., and DiNenno, E. Estimating Design Effect and Calculating Sample Size for Respondent-Driven Sampling Studies of Injection Drug Users in the United States. *AIDS and Behavior*, 16(4):797–806, 2012.

Yamanis, T. J., Merli, M. G., Neely, W. W., Tian, F. F., Moody, J., Tu, X.,

and Gao, E. An Empirical Analysis of the Impact of Recruitment Patterns on RDS Estimates Among a Socially Ordered Population of Female Sex Workers in China. *Sociological Methods & Research*, 42(3):392–425, 2013.