Title

Design and Analysis of a Novel Respondent-Driven Sampling Methodology for Estimation of Labor Violation Prevalence in Low-Wage Industries

Author
Scott-Curtis, William

Publication Date
2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Design and Analysis of a Novel Respondent-Driven Sampling Methodology for Estimation of Labor Violation Prevalence in Low-Wage Industries

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

William Scott-Curtis

2024

ABSTRACT OF THE THESIS

Design and Analysis of a Novel Respondent-Driven Sampling Methodology for Estimation of Labor Violation Prevalence in Low-Wage Industries

by

William Scott-Curtis
Master of Science in Statistics
University of California, Los Angeles, 2024
Professor Mark S Handcock, Chair

Respondent-driven sampling (RDS) is a network-based sampling strategy useful for studying hard-to-reach populations, such as low-wage workers. Respondent-driven sampling designs prompt respondents to recruit other members of the population of interest in their social network to the survey. RDS methods can collect large samples from hard-to-reach populations by leveraging social ties within these communities to facilitate recruitment. However, these designs are prone to being affected by many sources of bias, including seed bias–the effect of starting the recruitment chains with a biased convenience sample.

Previous work utilizing RDS to sample low-wage workers has suffered from issues of seed bias, making inference difficult. To address this problem, we propose a new design that collects seeds in a probability sample, and study this design's resilience to network homophily, or the tendency for similar people to cluster within social networks. The structure of this design is novel in its focus on estimation within multiple sub-populations of interest (for example, low-wage industries), and in its formulation of complex constraints imposed on recruitment to limit bias. We study and model the population networks and recruitment sampling, propose a modified estimator, and, via simulation, analyze the validity of inference.

Results indicate that inference in this design is feasible, and that modifications to a popular RDS estimator to account for the sampling constraints improve the accuracy of estimation. While the accuracy of the estimator is promising, further improvements to this estimator and the network generation algorithm are likely necessary to properly assess the validity of inference. These improvements include incorporating the sub-population structure of the sampling more fully into the estimator and implementing non-uniform homophily effects estimation and correction within the estimator.

The thesis of William Scott-Curtis is approved.

Jennie Elizabeth Brand

Hongquan Xu

Mark S Handcock, Committee Chair

University of California, Los Angeles

2024

For Daniel

And for Conrad and Letitia

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The prevalence of labor law violations among workers covered by the Fair Labors Standards Act is the subject of many surveys, especially within low-wage industries. The traditional approach to investigate labor violation rates is a nationally-representative address-based survey, but these surveys have low sample sizes of workers within low-wage industries, and likely undersample low-wage workers. This has made it difficult to estimate the prevalence of these violations in these populations, leading to a gap in knowledge. Because of these difficulties, an improved method is needed to increase the sample size of workers within low-wage industries and thus improve the estimates of violation rates within those industries. The 2008 Unregulated Work Survey, which studied the prevalence of labor violations among low-wage workers, utilized respondent-driven sampling (RDS), a common method of sampling for hard-to-reach populations, to generate a large sample of low-wage workers in individual cities [Bernhardt et al., 2009]. However, this work suffered from a variety of issues common to RDS surveys, most notably seed bias. Here, we investigate a new method, representatively-seeded RDS with constrained recruitment, to resolve these issues.

1.1 Respondent-driven sampling

Respondent-driven sampling, first developed by Douglas Heckathorn [Heckathorn, 1997], is a variant of "snowball" sampling that uses social capital within a networked population to sample individuals from rare or stigmatized groups. RDS has been successfully employed to sample many hard-to-reach, or hidden, networked populations including refugees, drug users, and MSM (men who have sex with men) [Gile et al., 2018]. The sampling leverages both the social ties and trust within these communities. Rather than researchers recruiting the social contacts of participants, as in other snowball sampling methods, respondents are recruited by members of that community, making potential recruits more likely to trust and engage with the survey.

A typical RDS design begins with a convenience sample of "seeds" who belong to the population of interest. These seeds are asked about their social contacts who also belong to the population of interest, and are asked to refer those contacts to the survey so they can be part of the sample. These new recruits are subsequently asked to refer their contacts in the population of interest, generating a referral tree that grows from the initial seed over multiple waves of referrals. This respondent-driven approach offers more anonymity for the social contacts of participants than providing researchers with a complete list of those social contacts. Typically, the number of referrals per respondent is capped at 2-5.

The sub-populations of workers in each low-wage industry can be classified as hard-to-reach populations, not only because they represent a small proportion of the total population of covered workers, but also because these workers are less likely to respond to traditional surveys. Therefore, the proposed sampling design, similarly to the Unregulated Work Survey, employs RDS.

Employing RDS to recruit members of the large, widely-distributed, difficult to reach (but not necessarily socially stigmatized) "community" of low-wage workers (and the sub-populations of specific low-wage industries) is different from the traditional settings that RDS was initially implemented in. While beneficial for studying hidden populations, RDS designs have significant issues that make statistical inference difficult. Therefore, taking full advantage of RDS requires the use of a design that addresses and mitigates these issues.

1.2 Modeling issues in respondent-driven sampling

While respondent-driven sampling is typically effective in collecting samples of hard-to-reach populations, the sample is often unrepresentative of the hidden population due to both the dependency in the sampling design and issues with respondent behavior in the recruiting phase. These are common problems in RDS, and their effects on the validity of inference are well-studied [Gile et al., 2018]. Our design seeks to mitigate the effects of seed bias and network homophily.

1.2.1 Design issues

The primary issue our design seeks to mitigate is “seed bias”, which is the tendency for the process of selecting seeds to be unrepresentative of the population of interest. This is a common problem in RDS designs, and was a significant issue in the Unregulated Work Survey, where seeds were active and influential workers in their communities known to the researchers. These seeds were therefore not representative of the population of low-wage workers in general, and their social networks were likely atypical for the population of low wage workers. This non-probability sampling of seeds causes cascading bias into the first few waves of sampling, which can be exacerbated by non-random recruitment of social contacts.

The RDS used in the Unregulated Work Survey was modeled as a variant of a Markov process on the state space of the network, where random recruitment selection implies that the transition probabilities of the Markov process are inversely proportional to personal network size. The validity of inference relies on an asymptotic argument of convergence of the sampling to the stationary distribution of the Markov process. Therefore, RDS designs like the UWS rely on many waves of sampling to eliminate the effects of seed bias, where the initially biased seed distribution asymptotically approaches the unbiased stationary distribution. However, the number of waves required to approach equilibrium in the social network, and therefore independence from the seeds, is often much higher than is practical for most surveys, making this justification unrealistic [Gile et al., 2018].

The issue of seed bias, we argue, can be largely alleviated when, as is the case in the design we propose, the seeds are a representative probability sample from a larger super-population. This lack of seed bias also mitigates (but does not necessarily entirely eliminate) issues arising from using very short recruitment trees that would impact traditional RDS designs. Because we do not need to rely on asymptotic convergence to the stationary distribution, beginning with an unbiased seed distribution at or near equilibrium removes the primary benefit of long recruitment trees.

Another common issue in RDS that also likely impacted the Unregulated Work Survey is is effect of network homophily on the sampling. Homophily is the tendency of people to have social ties to people with whom they share some trait, including race, gender, location, or, as is key here, whether they have experienced labor violations at work. The kinds of strong social ties between low-wage workers most likely to result in recruitment likely exhibit higher homophily on labor violations, especially for recruitment within workplaces. This causes outcome dependence in the recruitment trees, and can both bias estimators and reduce the effective sample size of the survey. This issue is of great concern here, and has therefore significantly impacted the sampling design. Some estimators attempt to correct for this dependence using a bootstrapped estimate of the homophily present in the network, but this approach has limits, as will be shown. In order to assess the validity of inference of this design using these types of estimators under variable levels of homophily, we implement a simulation of the population network and the sampling.

1.2.2 Respondent behavior issues

An important assumption of inference from RDS designs is that recruits are selected randomly by recruiters from a known total number of personal contacts to the population of interest. This requires that referrals are made with uniform probability, and that the total number of contacts is accurately reported. The survey instrument therefore should induce random selection of recruits, and should especially avoid recruitment bias dependent on the outcome variable (violation status). We assume this is the case.

Additionally, if recruitment rates are too low, the sampling can fail to reach a sample size large enough for accurate inference. This is of particular concern in this study because a successful sample requires a large sample from each of the low-wage industries. The sampling design here seeks to maximize the likelihood of collecting a complete sample even when recruitment rates are low. Assessing the effect of the recruitment rate on collecting a complete sample is a secondary goal of the simulations.

CHAPTER 2

Sampling and Estimator Design

To address the issue of seed bias, we propose a hybrid sampling design that generates both a large, nationally-representative sample and oversamples of sub-populations of interest. The process involves two phases: an initial, traditional sampling that will be nationally representative, with a target sample size of about 4000 covered, non-exempt workers, and a subsequent respondent-driven sampling using the initial sample as seeds. Because these seeds will come from a nationally-representative sample, the seeds will not be biased, and the initial waves of recruits in the RDS phase will be more representative than those referred by seeds collected in a convenience sample.

However, although the seeds are representative of the population of covered, non-exempt workers, the recruits they refer from low-wage industries may not be representative of those industries. This is not technically an issue of seed bias because the seeds are representative. Instead, it may be the case that the population of covered workers itself, and by extension the workers referred by these seeds, is not representative of the populations of the low-wage industries. For example, we believe it is likely that the population of covered workers will have significantly lower rates of labor violations than workers in low-wage industries. This would exacerbate network homophily effects in recruitment in a way that is difficult to correct for using traditional estimators.

Over the course of this study, multiple estimator designs were considered. Because we are primarily concerned with mitigating the effects of homophily in the network, we selected and implemented an adaptation of Gile's Sequential Sampling (SS) Estimator [Gile, 2011], which includes a bootstrap of the sampling that corrects for homophily effects.

2.1 Population definition

In order to accurately model the population network, we must explicitly define the populations of workers we plan to study with this design.

2.1.1 Covered, non-exempt workers

The broad population of interest are workers covered by the Fair Labor Standards Act (FLSA), which comprises about 165 million American workers. These workers are employees, not independent contractors, and they are not exempt from the FLSA, which excludes most managerial-level workers. These exemptions to the population would need to be enforced by self-identification in a screener. Workers excluded by this screening could still be asked to refer social contacts from low-wage industries to increase the likelihood of collecting a complete sample, but the effects of these extra seeds are not studied here. It is possible that they could pose problems for inference if their recruits are even more unrepresentative of the low-wage industries with homophily effects significantly different from the covered seeds. In fact, even measuring these homophily effects would be near-impossible because the violation status of the non-covered seeds is not able to be collected, which poses problems for homophily estimation and correction. If used, these seeds would necessitate further alterations in estimator design.

2.1.2 Industry selection

As a hypothetical example, we study a design that seeks to estimate violation rates in 12 low-wage industries. The only relevant statistics of these industries to this hypothetical study are the population sizes of the industries, their relative visibilities to workers outside that industry, and the rate of labor violations in each industry. Of particular interest are industries most at-risk of high rates of labor violations, which includes many small, insular industries. We constrain the size of the industries to by at least 0.2% of the total population of covered, non-exempt workers, meaning that very small industries that would be particularly difficult to collect a large sample from were excluded.

We assume that the social identifiability of workers in these industries is high enough for recruitment to be successful. Industries should be easy to explain to subjects in order to facilitate accurate referrals, and workers in the low-wage industries should have high enough visibility in the social network to be identifiable by subjects. This issue is of unique importance for this study because, unlike most RDS designs, this recruitment structure will not rely exclusively on the social ties within communities of interest.

2.2 Seed sampling

Instead of drawing seeds from a convenience sample, we assume that seeds are sampled from the national population of covered, non-exempt workers, and that this sample is representative of that population. A potential method to accomplish this is an address-based sample, which randomly selects households from the United States Postal Service list of residential addresses, which has very high coverage of people who live in households. This is a common and very high quality method of sampling from US households.

We assume the seed sampling produces 4000 survey completes from covered workers. If respondents who do not meet the eligibility criteria are still asked to refer their social contacts from low-wage industries, the total number of RDS seeds will likely more than double. However, we believe it is reasonable to assume that covered, non-exempt workers are more likely to have social contacts in low-wage industries, and will therefore be more likely to successfully recruit contacts to the survey.

2.3 Referral structure

The purpose of the RDS phase is to collect an oversample of the selected low-wage industries in order to improve estimation of violation prevalence within those industries. A secondary sampling phase is necessary because the population of covered workers in those industries are likely hard-to-sample and will comprise a very small proportion of the seed phase. This design has a somewhat different objective than traditional RDS designs that try to collect a sample from a hidden population using social ties within that population. Because each subject will be asked to make referrals to workers from multiple industries rather than just their own industry in order to reduce homophily in referral chains, the success of the sampling relies heavily on the existence of social ties between sub-populations.

Homophily is further reduced by constraints on recruitment eligibility, most importantly that workers can recruit at most one worker in each industry, and that recruits must not share a workplace with the recruiter, which removes many of the social ties with the highest correlation of violations. We also study a design that allows recruitment of at most two workers in each industry, which increases the likelihood of collecting a complete sample from each industry even under low rates of successful recruitment.

We assume that social ties between the covered workers that will comprise the initial phase and workers in the low-wage industries are common, making an RDS design feasible. However, this is a novel use of RDS, not only because there are many distinct sub-populations of interest, but because, unlike traditional RDS designs, the social ties used to make referrals will often not be within communities of interest. This presents both benefits and risks.

2.4 Potential issues in the RDS

The clearest benefit over traditional RDS designs here is the use of a large number of seeds that themselves are nationally-representative. Because the seed sampling will produce 4000 covered, non-exempt seeds (and more that do not belong to that population), collecting the target 400 samples in each of the dozen low-wage industries will not require many waves of recruitment. This means that there will be many, short trees rather than a few, large, branching trees. While this will reduce the time and therefore costs of sampling, there are potential issues with short trees.

In traditional RDS, the first few waves of recruits are often excluded from the final sample and are not used for inference because they are likely not representative of hidden communities of interest. This is because many RDS designs use a small, unrepresentative, convenience sample of the population of interest as seeds, and the bias of those seeds infects the first waves of recruits made by those seeds. A common model for RDS is a Markov process over the population of interest, with a starting point at the seed and steps being taken as social ties are used to facilitate recruitment at each wave of sampling [Gile et al., 2018]. While it is not a perfect model (RDS does not allow the same unit to be recruited more than once, and often there are multiple recruits from a single recruiter), the argument for asymptotic convergence to the stationary distribution in sampling is valid, even when the seeds are very biased.

Here, rather than there being bias in the seeds that is eliminated through many waves of sampling, we make the assumption that the seeds are themselves drawn from the stationary distribution, or at least a distribution close enough to the stationary distribution that the sampling is stationary after very few waves.

However, the population the seeds are drawn from could (and we believe likely does) have a significantly different (likely lower) rate of violations. This makes the distribution of violations in the seeds unrepresentative of the sub-populations of workers in the low-wage industries, which, when the social network has significant homophily, would bias the initial waves of sampling. While we start with effectively a simple random sample of the population of covered workers as seeds, we do not necessarily immediately get an unbiased probability sample of workers in the low-wage industries.

Another issue that could cause problems in early waves is differential seed activity among workers in low-wage industries. Activity is a measure of personal network size to a population of interest. Here, we are concerned with the number of covered workers a worker from a low-wage industry knows outside of that worker's own industry. This activity to potential seeds is directly correlated to that worker's sampling probability, especially because the sampling trees will be short, and is therefore important when weighting the sample. If these activity levels differ between workers by violation status, it can cause high variance in the early waves of sampling and bias estimation if personal network size to the population of seeds is not measured accurately.

Currently, differential seed activity is not accounted for in estimation because the sampling does not collect the size of personal networks to the population of potential seeds (all covered, non-exempt workers). This data would be infeasible to collect because the population is so large and broadly defined, and because the kinds of undirected ties that would both facilitate recruitment and be reported by recruits as visibility proxies would be relatively hard to accurately estimate compared to ties to a smaller, specific industry. A potential fix is using the number of ties a worker has to other low-wage industries as a proxy for the activity level to the larger population of covered workers, but this is likely not a perfect representation of the true activity level.

Differential activity is correlated with homophily, but can also exist in excess of homophily, causing further issues. For example, consider low-wage industry workers experiencing violations. They may be less likely to have ties to the super-population of covered workers because workers with violations are less likely to have ties to workers without violations, and the super-population of covered workers has a lower incidence rate of violations than workers in low-wage industries. This is due to homophily, and will cause differential activity to seeds. However, these workers with violations may also have fewer ties to seeds than their colleagues without violations in general, not just because of homophily. Suppose that low-wage workers with violations are more hidden. This could be because low-wage workers with violations have fewer social ties in general, have fewer social ties to the population of covered workers in particular, and/or have social networks that are more insular and concentrated on people who work in their industry but are not necessarily more likely to share their violation status.

A potential causal pathway here is higher vulnerability and therefore decreased visibility of these workers not just to survey researchers but to the broader social network as a whole. This is the exact kind of problem RDS is well suited to resolve. Under traditional RDS circumstances, with many waves of sampling, resulting issues of bias in inference are almost entirely mitigated by the design of the most commonly used RDS estimators, and the increase in variance is accounted for. However, relying on short sampling trees could cause this problem to recur, especially when social proximity to the seeds is highly correlated with violation status.

Consider an extreme example: suppose that there is a very tightly-knit cluster of low-wage industry workers who are disproportionately likely to have violations. Also suppose that no members of the cluster have ties to the larger population of covered workers outside their shared industry, but some do have ties to workers in that industry who themselves have ties to the wider population of seeds. For any seed outside that industry, it is impossible to sample workers from the cluster in the first wave even though the seeds are representative of the population of covered workers, making this cluster hidden from the those seeds. This network would have homophily, but even an accurate measure of the homophily in the network would not indicate the existence of this insular, hidden population, and therefore would not be able to fully explain and correct for the bias in the sampling. A sampling design that utilizes many waves would eventually find these workers, but they would be largely missing from a sample largely comprised of first-wave recruits. Even in the rare case where there is a seed sampled within the cluster, there will still be significant bias in the opposite direction, making estimation in networks with these properties high-variance.

2.4.1 Sampling design mitigation of the issues

We believe that it is reasonable to assume that the sampling network will exhibit homophily on violation status. Additionally, it is at least somewhat plausible that the rate of violations among low-wage industry workers is not uniform across workers with varying numbers of ties to seeds, meaning that the network would also exhibit differential activity. However, we also believe it is reasonable to make some assumptions about the network on account of the design choices we have made that address and mitigate these issues.

While estimating visibility to seeds is likely difficult[1], we believe the design is well-suited to minimize differential seed activity; because the seeds are representative of the population of covered workers, differential seed activity in the network must be due differences in social proximity to the entire population of covered workers, not to a small, highly-unrepresentative group of potential seeds, which is the case in traditional RDS. This is the primary benefit of using an unbiased (although potentially still unrepresentative) seed pool. It is far more plausible that workers in low-wage industries do not have large differences in activity to a very large subset of the American population than a small convenience sample of highly visible workers that would normally comprise the seed pool.

2.5 Estimator design

Gile's SS Estimator builds upon previous RDS estimators, in particular the Volz-Heckathorn estimator [Volz and Heckathorn, 2008], that use personal network size as a proxy for sampling visibility when weighting the sample. The improvement the SS estimator implements that is most relevant here is in adjusting for first-order homophily effects. It is also powerful in estimating and accounting for complex finite population effects resulting from the

[1] [Feehan et al., 2022] provides potentially useful strategies

without-replacement network sampling design when the sampling fraction is large[2].

The estimation is accomplished with a version of a bootstrap algorithm that simulates generating random networks and sampling from them. These networks are generated from a null model called a "configuration model" that replicates the distribution of degrees implied by the degrees observed in the RDS sample. This sampling requires only the distribution of degrees observed in the sample and an estimate of the total size of the population of interest[3]. However, the homophily correction assumes uniform homophily effects across the network, which is not the case here due to varying violation rates between industries, the network constraints, and sampling constraints, so adjustments to the estimator are necessary.

2.5.1 Estimator design mitigation of the issues

First-order effects of personal network size on inclusion probability are accounted for in the Horowitz-Thompson style estimators in use for RDS inference, including the SS estimator. That is to say, in order for there to be a confounding effect from differential activity, workers with labor violations must be less likely to be referred beyond the first order personal network size effects, which are ignorable by the design of the most popular RDS estimators. This means that any sampling bias due to differential activity is caused only by the short chains, assuming that the estimator accurately maps personal network size to general network visibility. Learning this mapping is a strength of the SS estimator, but it could be complicated by the effects of sampling constraints on inclusion probability.

In order to improve this mapping, we implement an adjustment to the estimator to rescale

[2] This is most useful in networks with highly differential activity in the parameter of interest, which could be the case here, but is not modeled directly in the simulation because we believe this is largely mitigated by the unbiased seeds

[3] We already assume that the sizes of each low-wage industry are known, given the reliability of labor statistics. For our simulations, we use the known size of the simulated networks. In a real-world sampling, we would use the known size of the population of covered workers, and would be able to use the sizes of each industry in a version of the estimator that replicated the concentrations of industries in the network sampling.

the network size of each recruit in proportion to the population size of their recruiter's industry. This rescaling also assumes a very large population size for the seeds, and therefore maximal rescaling to each node in the first wave. This relies on an intuition that larger industries themselves are more visible and likely to generate recruiters, so having ties (especially ties that actually facilitate recruitment) to large industries increases a node's visibility by a higher factor than a tie to a smaller industry. While somewhat crude, this adjustment significantly increases the accuracy of the estimator.

Further, assuming that any differential activity that hides workers from initial sampling waves is limited, the bias introduced by the short chains is also limited and can likely be accounted for with a sufficiently sophisticated estimator. The above adjustment made to the estimator is a step in this direction, and further, more refined potential adjustments are discussed below and in Chapter 4.

The adapted SS estimator also bootstraps an estimate of, and corrects for homophily in the recruitment, meaning that in order for these effects to confound inference, the difference in inclusion probability due to homophily must be in excess of the effects of the homophily implied by the homophily observed in the recruitment network as a whole, which provides robust protections against moderate network homophily[4]. For example, it is reasonable that the rate of violations among the covered, non-exempt seeds will be lower than the rate in the low-wage industries. Additionally, it is reasonable to assume that due to positive homophily of labor violations, these seeds will be less likely to have ties to, and therefore are less likely to refer, a low-wage industry worker with a violation than a simple random sample of that industry would generate. However, the first-order effects of this homophily are also ignorable by design of the estimator, meaning that there must be significantly higher homophily between seeds and referred workers than the overall homophily in the referral network that also includes referrals between low-wage industry workers. Any homophily will

[4] This protection relies on the SS estimator's bootstrapped estimate of network homophily to be fairly accurate, which is reliant on seeds that are not too unrepresentative of the population and a good replication of the sampling in the bootstrap. Both of these are issues discussed further in Chapter 4.

increase the standard error of the estimates, but it will take significant, differential homophily across types of recruitment (within or between different industries) to bias the estimator in a manner that cannot be accounted for by the bootstrap estimate. This is why the referral design takes care to exclude the most homophilous social ties.

We believe that it is reasonable to assume that homophily between seeds and their first wave recruits from low-wage industries will be low, and is likely lower than homophily between low-wage workers, especially low-wage workers in the same industry, which are limited to those outside their own workplace. This is beneficial because homophily is low, but could pose problems if that homophily varies across different types of recruitment. While the estimator corrects for homophily in the network, the design of the bootstrap assumes uniform homophily across the network, and the effects are corrected for uniformly across industries. This is exacerbated by varying violation rates between industries because the magnitude of the effects of homophily are dependent on the distribution of the homophilous variable. These effects have a particularly large impact on industries that have violation rates that vary significantly from the mean violation rate in the overall sample.

Significant modifications to the estimator that introduce variable homophily by recruitment type that is corrected for individually within each industry are possible, but require work beyond the scope of this study, including the restructuring of the bootstrap to reflect the complexities of this recruitment constraints. These modifications would also significantly increase the degrees of freedom of the homophily estimation, potentially increasing variance of the estimator. Instead, we examine the effectiveness of an estimator that maintains the assumption of uniform homophily effects with the understanding that there is likely bias in the implementation of this estimator. This is likely a significant source of the bias in the results because it fits the observed pattern of increasing positive bias in industries that have significantly higher violation rates than the sample as a whole.

CHAPTER 3

Population and sampling simulation

In order to test the validity of the design, a simulation of the population network of covered workers and the novel RDS methods was built to study the resilience of estimators and design choices under a variety of potentially problematic circumstances. Of particular concern were scenarios where the sampling network had high homophily, and scenarios where the rate of successfully recruiting new subjects was very low. Because we believe that the effects of differential activity are thoroughly mitigated by the design, the network model does not include differential activity. Because the recruitment structure is novel, extensive investigation of homophily and constraint effects is necessary to determine the validity of potential estimators. A simulation study was chosen to do this investigation due to the complexity of both the network and the sampling design. The simulation could also be used to inform decisions about recruitment methods and potential modifications to existing RDS estimators to account for the recruitment structure and homophily.

Two models were built to simulate the design: a social network model, which generates a network containing the seeds and each of the low-wage industries, and an RDS model that samples along the ties in the simulated network according to the design rules. The simulations were designed to include parameters likely to affect the sampling process in order to assess the reliability and resilience of estimators. These parameters determine the structure of the network and the behavior of subjects. The network model parameters fix the distributions of ties between industries and seeds which model visibility, the density of ties within sub-industries, and violation homphily of the overall network. The sampling model parameters set the rate of successful recruitment and the restrictiveness of the recruitment process.

3.1 Simulation of the population network

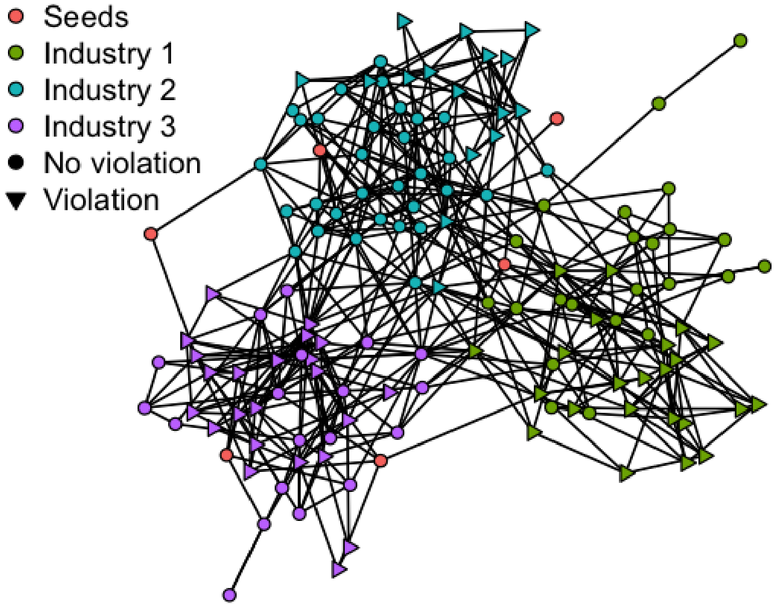

Figure 3.1: A small sample network to illustrate industry structure

The true network we model here is the social network of all covered, non-exempt workers that will be used to recruit the sample. A tie exists between two workers if those workers would identify each other as meeting the criteria established in the RDS design for "knowing", which is that they know each other, have spoken recently, and that the subject has the potential recruit's contact information. Ties are also restricted to workers who do not work in the subject's own workplace, which is a design choice to reduce the homophily of the network and avoid recruitment where violation status is highly correlated. A more definite definition of "knowing" should be established in the specific design of the survey that balances the needs for a sufficiently dense network and a network with strong ties that maximize the likelihood of successful recruitment of workers in the targeted populations.

This network is likely very complex, with many significant variables and parameters determining the structure. However, much of the complexity in this network, including geographical clustering, local transitivity, and hierarchical structures will likely not have a significant impact on the sampling, both because restrictions in the sampling design reduce the degree of transitivity in the true network relative to the larger, less strict national social network of workers, and because their will be many, short sampling chains, reducing the effects of transitivity on inference. We focus on the aspects of the network we believe will have the largest impact on the validity of inference: visibility of industries in the network, defined as the distribution of ties between industries and from seeds to industries, which is enforced by the constraint set ω, and the homophily in the network, which is injected into the network by the model parameters θ. [1] In the example network shown in figure 3.1, note the number of connections between industries, which are fixed by the constraints ω, and the clustering of violations within each industry, due to homophily caused by θ. [2].

[1] The inclusion of hierarchical structures and transitivity in the network model were considered, but we determined that the simplicity and efficiency of sampling a network were more important for investigating the validity of estimators that retain the form of standard RDS estimators. In the future, the testing of more sophisticated, specialized estimators that more comprehensively account for the sampling design could potentially benefit from more complex network models.

[2] This network is much smaller than the networks used to simulate the sampling, and therefore has significantly different generating parameters

3.1.1 ERGM models

Networks are simulated by sampling from a constrained Exponential-family Random Graph Model (ERGM) [Fellows and Handcock, 2012]. ERGMs model the ties of a network with a fixed number of nodes. Each ERGM can be specified by the terms it is comprised of. Like an Exponential-family random variable, the mass function of an ERGM $P(Y)$ for a random network Y can be represented in the form

$$P(Y = y) = \frac{\exp\left(\theta^T s(y)\right)}{c(y)}$$

where θ are the model parameters, $s(y)$ are the associated network statistics, and $c(y)$ is a normalizing constant of the form

$$c(y) = \sum_{y \in \mathcal{Y}} \exp\left(\theta^T s(y)\right)$$

Where $\mathcal{Y}$ is the set of all networks that fit the constraint set.

Common network statistics used in ERGMs are counts of the total number of edges, counts of more complex structures like "triangles" (sets of three nodes each connected to the others) and "k-stars" (a node that is connected to k other nodes), and parameters that count ties between nodes with different levels of some categorical variable. A non-exhaustive list of statistics and their definitions is available in the ERGM package for R [Hunter et al., 2008].

We model the social network between workers Y, an undirected ERGM with n nodes, parameters θ, network attributes X, and constraint set ω sampled from a superset Ω such that $\omega \in \Omega$. The model density is then

$$P(Y = y|\omega, \theta, X) = \mathbb{1}\left(y \in \mathcal{Y}(\omega)\right) \frac{\exp\left(\theta^T s(y)\right)}{c(y)}$$

The network model contains two categorical node variables, industry X_c and violation status X_v, which comprise the network attributes X. The industry of a worker X_{ci} corresponds to the low-wage industry that worker belongs to, or an additional "industry" that categorizes that worker as a seed. The structure of the network between industries is set by the constraint ω. The violation status X_{vi} models whether that worker has experienced a labor violation. The variety and severity of that violation is not modeled in order to simplify and focus the simulation on the primary goal of the survey, estimating the rate of labor violations within each industry.

First, the industry of each node X_{ci} is randomly sampled in proportion to the size of each industry, except for the seeds. Because the number of seeds n_{seeds} is determined by the initial sample and not the RDS, we fix n_{seeds} at 4000, and therefore set the industry of 4000 nodes in the network to the "seed" category. The weights for generating industries were set to model industries of a wide variety of sizes. The weights ϕ are proportional to the population of covered workers from each industry, so the distribution of industries in the network will reflect the true distribution of workers. The model for (non-seed) X_{ci} is then

$$X_{ci} \sim \text{Multinomial}(1, \phi)$$

and the number of nodes in each non-seed industry is also multinomial-distributed with $n - n_{\text{seeds}}$ trials and probabilities of success scaled to the population sizes (because only workers in industries of interest are modeled).

The violation statuses X_v are each a random Bernoulli trial, with unique probabilities of success in each industry, so

$$X_{vi} \sim \text{Bernoulli}(\rho(x_{ci}))$$

For simplicity, we designate v_i the sampled violation status of a node i. These probabilities ρ are essentially[3] the target of our estimation (Add a footnote pointing to the later discussion of finite population effects, etc).

[3] see section 3.1.9

The rates themselves are sampled uniformly for each non-seed industry I_i,

$$\rho_i \sim \text{Unif}(0.1, 0.5)$$

This reflects the wide range of violation rates that may exist in these industries according to limited prior research, and therefore is a relatively high-variance prior. This also will allow our assessment of the viability of estimators to include changes in viability that may occur at different violation rates. For example, lower rates may magnify the confounding effects of homophily, or higher rates in smaller industries could suffer from increased bias due to the uniform network homophily assumption.

3.1.2 Modeling visibility in the network

We assume a symmetric definition of "knowing", so the network model is undirected. This significantly impacts the way in which visibility into each industry, which we define here as the distribution of ties from nodes in one industry to another[4], is modeled. The most significant consequence of using undirected networks is that modeling visibility from one industry into another also fixes the visibility in the reverse direction, which is not just reasonable but necessary for symmetric "knowing".

The assumption of undirectedness is also supported by the sampling constraints. Because referrals will be between covered workers who do not share a workplace, any social hierarchies that may limit the willingness of two workers to both identify each other for purposes of the survey are minimized. For the purposes of this network model, in concordance with the assumption of no differential activity, visibility is independent of violation status. Ties will still exhibit homophily, but the distribution of the number of ties into an industry from a node with violations is equal to the distribution from a node without violations.

[4] Note that this is related is not exactly the same as the concept of visibility in the RDS, which is the model for sample inclusion probability of nodes [McLaughlin et al., 2015]

Analysis of a specific sampling design should rely on exploratory research of the visibility of each industry in the network not only to ensure that the selected industries are visible enough to successfully recruit workers from, but also to provide data to inform the visibility prior for each industry for network simulation. For now, we fix visibility parameters that are roughly but not exactly proportional to industry size, and we assume that data regarding the visibility of larger industries is more reliable.

Visibility is operationalized in the model via the construction of the constraints set ω with each network. ω reflects the visibility of each industry to workers outside that industry by randomly generating the number of ties each node has to each of the other industries. The constraints ω enforce the industry structure of the network, fixing the number of ties between industries but not the precise distribution of those ties in the network, which allows the ERGM sampling process to draw them with the amount of homophily specified by the parameters.

Here, the visibility of an industry A is the distribution of the number of ties a worker (not necessarily in A) has to workers in A. We make the simplifying assumption that the visibility of an industry is uniform for workers from other industries, including the seeds. Additionally, because the network is undirected, the total number of ties from any industry A to any other industry B must be equal to the number of ties from B to A. Further, if the maximum number of ties from a single worker in industry A is x, then there must be at least x workers in industry B with at least one tie to industry A. In fact, this is just the most obvious requirement for a constraint set of degrees between sub-populations to be consistent, meaning that the sample set of networks that fulfill the constraints is non-empty; this principle, that there must be enough nodes in B with ties into A to draw all the ties from A into B causes cascading dependencies that complexly restrict the set of consistent constraints. For this reason, we simplify the generation of constraints to increase the likelihood that the constraints are consistent using the principle that the visibility of sub-population B from sub-population A in an undirected network is proportional to the visibility of A to B, and that the proportionality constant is the ratio of the sizes of the sub-populations.

We define visibility here as the distribution of a random variable $X_{a\to B}$, the number of ties from a node a in industry A to nodes in industry B, or the degree of node a to industry B, and express the undirected constraint as

$$\sum_{a\in A} x_{a\to B} = \sum_{b\in B} x_{b\to A}$$

Which, using independence of nodes, given by the assumption of no transitivity, no differential activity, and uniform visibility into B, yields

$$N_A\mathbb{E}\left[X_{a\to B}\right] = N_B\mathbb{E}\left[X_{b\to A}\right]$$

This way, even if A is a much larger industry than B, as is the case with many combinations of industries of interest here, the constraints imposed on these degrees to model visibility are respected.

This also implies that for any two industries, a difference in mean visibility exists if and only if the sizes of the industries are different, and the difference in mean visibility is directly proportional to the ratio of the population sizes.

Additionally, the simplifying assumption of uniform visibility can be expressed

$$\forall C,\ \forall A, B \neq C,\ p(X_{a\to C}) = p(X_{b\to C}),\ \text{where } a \in A,\ b \in B$$

where A, B, C are industries.

This uniformity assumption does not mean that visibility is modeled merely as a constant multiple of population size. Rather, a prior on visibility is formed through exploratory research of the visibility of the specific industries of interest, and is set through the choice of parameters to model the visibility of each industry. Here, we model visibility as a zero-inflated negative binomial distribution, with the zero-inflation parameter $\phi_z(\cdot)$, such that

$$X_{a \to B} \sim \mathrm{NB}_{\mathrm{zi}}(\phi_z(B), r, p)$$

The zero-inflation, which artificially increases the proportion of the probability mass of the distribution at zero, models the assumption that many workers do not have ties to an industry. The rest of the negative binomial models the degree distribution for the workers who have at least one tie (The negative binomial still places some mass on 0, but this is accounted for when setting the zero-inflation parameter).

The zero-inflation parameter $\phi_z(\cdot)$ is unique for each industry, and roughly but not exactly tracks with the inverse of the industry size. The negative binomial parameters were set to $r = 5, p = 0.7$ to reflect assumptions of the distribution of ties to industries. The negative binomial parameters were set constant across industries because of a lack of good data about differences in the shapes of the visibility distributions between industries. We do, however, expect the distributions to be further right-shifted for larger industries, but this intuition would have been needlessly complex to include in the model. The negative binomial distribution was chosen because it is the result of a Poisson variable with a gamma-distributed rate fits the model of first sampling the “affinity” a worker has with an industry, and then sampling how many contacts that worker has to that industry given that “affinity”. These assumptions should be further informed by the previously mentioned exploratory research.

This model of visibility alone does not ensure that the constraint set is consistent, however, so further steps must be taken. First, because the visibility from A to B is so closely related to the visibility from B to A, we will only sample one set of degrees for the pair. Using the assumption that the reliability of the theoretical visibility data and the accuracy of this visibility model for large industries is higher than for smaller ones, the visibility of the larger industry within the smaller industry is modeled, and the set of degrees from the smaller industry to the larger is sampled. In this way, the constraints on ties from smaller industries to larger industries is set first. This means that the models for the visibility of the larger industries will have a much greater impact on the network structure. As an exception, this process begins with setting constraints on ties from seeds to each industry, because of the lack of data about the relative visibility of the wider population of covered workers.

Then, in order to meet the requirement that the total number of ties from the large industry A to a small one B is equivalent to the total from B to A, we must fix the sum of ties from the A to B to be the sum of the ties from B to A. Finally, using the assumption that the negative-binomial parameters (but not the zero-inflation parameters) are roughly equal across industries, the degree constraints for each node a in A to B are sampled without replacement from the set of non-zero degree constraints from B to A, ensuring equal sums. This means that, because A is larger than B, there will be far more nodes in B with zero ties to A than vice-versa, which aligns with the intuition that visibility is roughly proportional to industry size.

We believe that the assumption of equal negative-binomial parameters does little to affect the sampling. Because referrals are limited to one per industry, whether or not you have any ties to an industry is far more important in the sampling design than the total number of ties given you have at least one. In fact, the total number of ties past the first to a particular industry will only determine the visibility of that worker, and therefore its weight in the estimator, not how that worker is able to recruit. Any nuances in the particular shapes of these distributions would add very little power to the model.

The independent sampling of degree constraints poses a potentially larger issue, however. Because the sampling is independent even for the same node to different industries, the model lacks a network property that likely exists: concentration of ties to an industry among nodes that already have ties to other industries. It is well known that knowing some people makes it more likely that you know others, but the likelihood of knowing a class of people given you know an unrelated class of people is likely a smaller effect. For now, we assume these effects are also minimal.

Undirected network models are the norm in RDS modeling because social ties, especially strong social ties, are usually bi-directional, and because RDS inference requires the assumption that the number of contacts a subject reports (which would be out-ties in a directed network) can be used to estimate the inclusion probability of that subject because the number of ties can be used to estimate the visibility of the subject in the network (which is only true of in-ties). This modeling choice necessitates some approximations and simplifications in the constraint generation.

3.1.3 Sampling industry-specific degree constraints, ω

Let Ω be the set of consistent constraints on degrees between industries. An element $\omega \in \Omega$ is a matrix such that $\omega \in \mathbb{R}^{n \times m}$ where n is the network size and m is the number of industries, including seeds. The element ω_{ij} is a non-negative integer equaling the number of edges from node i to industry j. Let $\mathcal{Y}(\omega)$ be the set of all networks y that fulfill the constraint ω, where y is a symmetric $n \times n$ indicator matrix representing the existence of ties between each pair of nodes in the network. ω is consistent is there exists some network y such that the degree sequences of the nodes to each industry exactly match the degree constraints enumerated in ω, which is equivalent to $\mathcal{Y}(\omega)$ being non-empty.

Let the distribution of the generating process of specific constraints on degrees ω using zero-inflated negative binomials and resampling non-zero degrees without replacement be $f(\omega)$.

In order to ensure that sampling ω produces a plausibly consistent constraint, ω_{iJ}, the degree constraint from node i, a member of industry I, to industry J is sampled in one of two steps performed in order. For the purpose of generating constraints for seeds, seeds are treated as nodes from the smallest "industry".

First, if I is smaller than J,

$$\omega_{iJ} \sim \mathrm{NB}_{\mathrm{zi}}(\phi_z(J), r, p)$$

Second, if I is larger than J, ω_{iJ} is sampled for all i from industry I without replacement from the non-zero entries in the constraint vector ω_{jI} for all j from industry J. This ensures that

$$\sum_{i \in I} \omega_{iJ} = \sum_{j \in J} \omega_{jI}$$

which is necessary for constraints to be consistent for an undirected network. Any constraints ω_{iI} for within industry degrees are left free, and are modeled separately. In this way, we draw a constraint set $\omega_0 \sim f(\omega)$.

In order to initialize the Metropolis-Hastings algorithm used to model from the constrained ERGM, we need a seed network y_0 that fulfills the constraints ω_0. Therefore, the inter-industry edges of the network y are drawn to fulfill the constraint ω_0, which would ideally sample y_0 from the set $\mathcal{Y}(\omega_0)$. However, because the generation of ω_0 does not guarantee $\omega_0 \in \Omega$, $\mathcal{Y}(\omega_0)$ could be empty. Additionally, defining an edge-drawing process $f'(Y|\omega_0)$ that is guaranteed to sample y_0 from any nonempty $\mathcal{Y}(\omega_0)$ is complex, and the process itself would be computationally intensive.

Therefore, we loosen and redefine our constraints by first sampling y_0 by drawing edges using a simple, deterministic, but imperfect process $f(Y|\omega_0)$, and then making a small, very sparse adjustment to the constraints, $\varepsilon \in \mathbb{R}^{n \times m}$ so that $y \in \mathcal{Y}(\omega)$ where $\omega = \omega_0 + \varepsilon$. For each pair of industries I, J, the magnitude of the adjustment across all nodes in I, $\sum_{i \in I} \varepsilon_{iJ}$ is usually between 1 and 6, and almost always less than 10. These deviations ε from the generated constraints ω_0 are small enough that, given the lack of precision in the choice of the generating parameters $\phi_{zi}(\cdot), r$, and p, there should be little impact on the sampling. Therefore, we slightly perturb the constraints to ω after the edges are drawn to contain the newly-generated network y_0.

3.1.4 Metropolis-Hastings Sampling from the ERGM

The Metropolis-Hastings algorithm proposes perturbations of the network y to y' sampled from $g(Y'|Y = y, \omega)$ such that y' remains within the constraint ω, and accepts those changes in accordance with the acceptance probability $\alpha(y, y')$, where

$$\alpha(y, y') = \frac{P(Y' = y'|\omega)}{P(Y = y|\omega)} \frac{g(Y = y|Y' = y', \omega)}{g(Y' = y'|Y = y, \omega)} = \exp\left[\theta^T(s(y') - s(y)) \frac{g(Y = y|Y' = y', \omega)}{g(Y' = y'|Y = y, \omega)}\right]$$

The Metropolis-Hastings algorithm is well-suited to sample from ERGMs because the acceptance ratio is easily calculable from the model density via the change in log-odds of the network under the proposed perturbation [Fellows and Handcock, 2012]. This simplification of the acceptance ratio makes ERGMs particularly useful, especially when the statistics of the perturbation $s(y') - s(y)$ are simple to calculate. Therefore, in order to maximize the speed of the algorithm, $g(Y'|Y = y, \omega)$ should produce only simple, sparse perturbations $\epsilon = y' - y$.

The simplest perturbation of a network is a single edge flip, where ϵ_{ij} is a matrix of zeros except for the entries at (i, j) and (j, i), which are 1 if $y_{ij} = 0$ and -1 if $y_{ij} = 1$. This perturbation flips e_{ij} from “off” to “on” or “on” to “off”. This is also the default method of sampling from an ERGM without edge independence in the ERGM package [Hunter et al., 2008] because of its simplicity and speed. However, when the network constraints are this tight, this method does not work, so we must develop a new method to sample perturbations.

One approach to sampling from $g(Y'|Y = y, \omega)$ first samples an edge to flip $e_{ij} \in E(\omega)$, the set of edges that could exist in a network according to the constraints ω. Constructing $E(\omega)$ here, however, is difficult. Consider a case where the edge e_{ij} connects nodes i, j from respective industries I, J and $I \neq J$. It is not necessarily true that if the constraints $\omega_{iJ} > 0$ and $\omega_{jI} > 0$, then $e_{ij} \in E(\omega)$. This is a necessary but not sufficient condition for an inter-industry edge $e_{ij} \in E(\omega)$, and determining which edges meet this condition but are not in $E(\omega)$ is difficult to do efficiently.

Therefore, we adapt $g(Y'|Y=y,\omega)$ to sample e_{ij} from a superset of $E(\omega)$ defined as the union of all edges $e_{ij}|I=J$ and $e_{ij}|\omega_{iJ}>0,\ \omega_{jI}>0$. This superset, while large, is simple to construct, and, because it is only depends on ω, will remain constant, meaning it only needs to be calculated once. If the sampled $e_{ij}\notin E(\omega)$, then $y'\notin\mathcal{Y}(\omega)$, and the perturbation is rejected. This occurs much less frequently than the naive edge sampling method.

If the nodes i,j share an industry, $e_{ij}\in E(\omega)$ (so long as they are not seeds). Therefore, $g(Y'|Y=y,\omega)$ is uniform over this set of edges within non-seed industries, and when such an edge is sampled initially, the perturbation is just the flip of that edge, $y'=y+\epsilon_{ij}$. Because seeds come from a national sample, we make the assumption that there are no ties between the seeds, and consequently $g(Y'|Y=y,\omega)$ never samples an edge between two seeds.

However, if the nodes i,j do not share an industry, a single edge flip, even if $e_{ij}\in E(\omega)$, would force $y'\notin\mathcal{Y}(\omega)$. It is important to construct $g(Y'|Y=y,\omega)$ so that $y'\in\mathcal{Y}(\omega)$ because if the proposed perturbation forces the network outside the constraints, it is rejected. Because the constraints fix the degree of each node to each other industry exactly, proposing a flip of a single edge e_{ij} where i,j do not share an industry will always be rejected for falling outside the constraints, leading to no mixing of ties between industries, and no introduction of homophily. In fact, ensuring that y' remains within ω is more complex and requires flipping at least a quartet of edges together.

The set of quartets of edge flips between industries that stay within the constraints is much smaller than the set of all possible quartets of flips, meaning that edges between industries will be very stable even after extensive computation time if we do not precisely define the selection of a quartet. Therefore, $g(Y'|Y=y,\omega)$ only samples quartets of flips that keep $y'\in\mathcal{Y}(\omega)$. This restriction increases the computational time to sample from $g(Y'|Y=y,\omega)$ because it requires, in addition to sampling the primary edge e_{ij}, generating the set of all other edges e_{kl} such that flipping $e_{ij},e_{il},e_{kj},e_{kl}$ keeps $y'\in\mathcal{Y}(\omega)$. This increase in computational time per iteration is worthwhile because it improves the convergence rate of the network sampling[5].

Therefore, when an edge e_{ij} between industries I, J is selected to be perturbed, the algorithm selects two other ties e_{il}, e_{kj} between the same two industries to flip in the opposite direction to maintain the degree constraints on i, j. These two edges are selected by sampling new nodes $k \in I, l \in J$ such that $\omega_{kJ} > 0, \omega_{lI} > 0$ and $Y_{ij} \neq Y_{il}, Y_{ij} \neq Y_{kj}$. Additionally, in order to maintain the constraints on k, l, the edge between them must also be flipped, specifically in the same direction as e_{ij} is being flipped. This means that in order for a quartet of flips to maintain consistency with the constraints, $Y_{kl} = Y_{ij}$. This constraint is also imposed on the sampling of k, l, and makes the sampling dependent. For this reason, instead of sampling k, l individually and potentially selecting edges that when flipped would force the network out of the constraints, we sample from the set of pairs (k, l) that meet the conditions[6].

Figure 3.2 shows the structure of these perturbations, where the dashed lines are flipped in one direction and the solid lines are flipped in the other direction.

So long as e_{kl} has the same initial status as the e_{ij}, the total number of ties between the industries is preserved, and the perturbed network remains within the constraint set. Because the set of ties between nodes of the two industries with degree constraints greater than 0 is sparse (most do not exist in the network)[7], it is much more likely that this final

[5] We considered loosening the constraints to have a floor and ceiling on the degrees between industries, but almost all proposed changes would still force the network out of the constraints, and would interfere with the model of inter-industry ties.

[6] It is not clear what the optimal method of this sampling is to thoroughly mix the Markov Chain across the set of networks consistent with the constraint set. Further study could examine methods of drawing the initial network to maximize the perturbation potential of edges between industries given any known stagnancy of the Markov Chain with certain edge sampling methods that may not significantly retard in the convergence of the network to the ERGM parameters, but do not thoroughly mix across the set of constraint consistent networks. This “maximization of perturbation potential” could be called biasing the initial network drawing toward “hot” networks with high “potential energy” that can be perturbed in many ways before settling into one of many cooler structures that tend to maintain the general structure of inter-industry ties. The contours of the model space are largely unknown, so the existence of this dynamic is speculative

[7] This is different than the set of edges that *could* exist in a network that fulfills the constraint set. But,

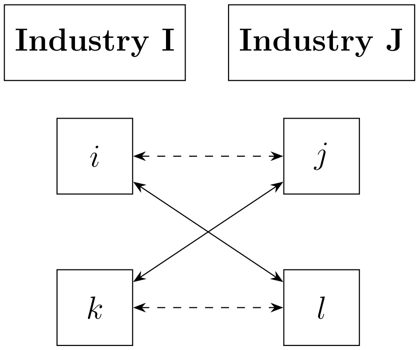

Figure 3.2: Perturbation ϵ_{ijkl} of a quartet of edges

condition of the perturbation is met if the initially selected edge is "off" (because it is likely that the edge between these other two nodes is also "off"). However, this means that, in order for the proposed quartet of flips to be consistent with the constraints, if an initially selected edge is already on, the two other nodes to be sampled from the set of nodes whose ties to the initial nodes are off must be connected by an edge that is on. This would cause most proposed perturbations of a primary edge that exists to fail before the Metropolis-Hastings step, potentially biasing the Markov Chain mixing toward sampling ties to turn on, and only choosing to turn off ties that are connected to the dual pair of nodes, which are likely not a representative sample of existing ties.

Instead, we modify $g(Y'|Y = y, \omega)$ to condense the two samplings of nodes to form the dual pair into a single sampling of a dual pair (or quartet) of nodes that are connected to each other and are not connected by inter-industry ties to the primary pair of nodes. This will ensure that the rare cases of edge drawing failing because the sampled dual pair are already connected are avoided. Sampling from the distribution of proposed quartet perturbations at each Metropolis-Hastings step is accomplished by marginalizing the quartet over the primary pair.

given industries with large enough populations of nodes with non-zero constraints (meaning a large set of edge permutations that fulfill the constraints), this set is also likely sparse

3.1.5 Modifying Metropolis-Hastings perturbation sampling, $g(Y'|Y = y, \omega)$

Because the networks are large and sparse, it is more efficient to store them as "edgelists" rather than $n \times n$ matrices of edges like y. We define $E(y)$ as the set of edges that exist in the network. The edgelist is a matrix of n_e rows and 2 columns, whose rows enumerate the elements of $E(y)$ by the pair of nodes they connect. This structure makes sampling edges that exist in the network simple, and reduces the complexity of enumerating and sampling from the space of all possible quartets of flips. Therefore, an adaptation to $g(Y'|Y = y, \omega)$ is implemented.

First, we use the independence of inter-industry and intra-industry edges to separate the drawing of these two types of edges. Because y is initially drawn from $f(Y|\omega)$, which only draws edges between industries, $E(y)$ initially only contains edges between industries. Therefore, we can run the Metropolis-Hastings sampling on just the inter-industry edges. This sampling takes two steps that can be expressed as

$$g(Y'|Y = y, \omega) = g_1(e_{kl}|e_{ij}, Y = y, \omega) g_0(e_{ij}|Y = y, \omega)$$

where e_{ij}, e_{kl} uniquely define a quartet of flips ϵ_{ijkl} such that $Y' = y + \epsilon_{ijkl}$, which flips e_{ij}, e_{kl} from "on" to "off", and flips e_{il}, e_{jk} from "off" to "on".

$g_0(e_{ij}|Y = y, \omega)$ uniformly samples an inter-industry edge e_{ij} from the network edgelist. $g_1(e_{kl}|e_{ij}, Y = y, \omega)$ samples another edge between the same two industries e_{kl} from the edgelist. These two edges form the quartet, and drawing e_{il} and e_{kj} while erasing e_{ij} and e_{kl} will maintain the constraints so long as e_{il} and e_{kj} are not already in the network. $g_1(e_{kj}|e_{ij}, Y, \omega)$ utilizes rejection sampling to uniformly sample from $E(e_{ij})$, the set of all edges that could form a quartet with e_{ij}, so

$$e_{kl} \in E(e_{ij}) \text{ if } e_{kl} \in E(Y), k \in I, l \in J, \; k \neq i, l \neq j, \text{ and } e_{il}, e_{jk} \notin E(Y)$$

This enforces the requirement that e_{il} and e_{jk} begin “off”. All edges in $E(Y)$ from the are sampled, and only edges e_{kl} such that $k \in I, l \in J, k \neq i, l \neq j$ are accepted. e_{kl} is also rejected if e_{il} or e_{kj} already exist in the network, because these edges must be flipped on to maintain the constraints.

The rejection sampling is abandoned after 100000 iterations, meaning that proposed perturbations between the smallest industries have a slightly inflated chance of failing even when there exist quartets that can flip. Currently, the smallest industries comprise about 3% of the network. Given that the density of edges between industries is roughly equal across the network (the probability of knowing a worker in a particular industry is roughly proportional to the size of that industry), about 0.08% of inter-industry edges in the network are edges between the two smallest industries. The probability that $g_1(e_{kl}|e_{ij}, Y, \omega)$ will fail to find a quartet after 100000 sampled edges is binomial, specifically

$$X \sim \text{Binom}(p = 0.0008, n = 100000), p(X = 0) < 1.75 \times 10^{-35}$$

This is rare enough that any impact on the RDS sampling can be safely ignored.

Finally, because sampling from the edgelist will only flip quartets in one direction (e_{ij}, e_{kl} off, e_{il}, e_{kj} on), we must address concerns about convergence of the sampling to the model. In order for the Metropolis-Hastings sampling to converge, any edge $e \in E(\omega)$ must be able to be drawn by $g(Y'|Y, \omega)$. This is possible when only flipping quartets one direction because edges are still being drawn, but is not immediately obvious because we are only sampling edges to erase, and edges selected to draw could systematically miss some edges $e \in E(\omega)$.

First, consider some edge $e_{ij} \in E(\omega)$ between industries I, J but $Y_{ij} = 0$, meaning that e_{ij} is not in the current network. We must show that a sequence of perturbations $g(Y'|Y = y, \omega)$ can draw e_{ij}. Because $e_{ij} \in E(\omega)$, $\omega_{iJ} > 0, \omega_{jI} > 0$. Additionally, because $y \in \mathcal{Y}(\omega)$, there exists at least one edge e_{il} from i to some node $l \in J$. Similarly, there exists at least one e_{kj} from j to some node $k \in I$. Because these two edges both exist in y, they can be sampled together by the process enumerated above. If these two edges are sampled, then e_{ij} must be one of the two edges that is drawn (along with e_{kl}). This shows not only that any $e_{ij} \in E(\omega)$ can be drawn by a sequence of $g(Y'|Y = y, \omega)$, but that e_{ij} can be drawn at any single step in the sequence regardless of the structure of y.

Further, we can say that $g(Y'|Y = y, \omega)$ is effectively uniform over the space of all possible quartets $\{e_{ij}, e_{kl}, e_{il}, e_{kj}\}$.

Each quartet requires two edges that already exist in the network e_{ij}, e_{kl}. These two edges uniquely define a quartet. e_{ij} is uniformly sampled from the set of edges that can exist in a quartet, and e_{kl} is sampled effectively uniformly from the set of edges that can define a quartet with e_{ij}. This sampling is not truly uniform because the sampling may reach the iteration cap before finding e_{kl}, and this is more likely to occur for some edges e_{ij} than others, but the effect is very small (see above). Therefore, the sampling is effectively uniform over all possible quartets.

While the single edge e_{il} may be more likely to be drawn if ω_{iJ} or ω_{jI} is large, that increase in likelihood corresponds to the number of unique quartets that involve drawing e_{il}. Because there are more unique quartets that draw edges between nodes with large degrees, the uniform sampling over the set of quartets pushes $g(Y', |Y = y, \omega)$ toward networks where nodes with high degrees are connected to each other. This corresponds to the trend in the space $\mathcal{Y}(\omega)$, where there are more unique ways to connect nodes with high degrees to each other than to nodes with lower degrees, and therefore more networks with that property. This property also exists in real social networks, and is therefore desirable. $f(Y|\omega)$, the process of drawing the initial network to seed the Metropolis-Hastings algorithm, is not a uniform random sample from $\mathcal{Y}(\omega)$ because it deterministically chooses nodes to draw edges to, and this is biased away from drawing edges between nodes with high degrees.

Finally, we can also say that $g(Y'|Y = y, \omega)$ is symmetric, which simplifies the calculation of the acceptance ratio because $g(Y'|Y, \omega) = g(Y|Y', \omega)$. To show this, consider a quartet of perturbations ϵ_{ijkl} such that $Y' = Y + \epsilon_{ijkl}$, and ϵ_{ijkl} flips e_{ij}, e_{kl} from "on" to "off", and

e_{il}, e_{jk} from "off" to "on". Therefore, the reverse perturbation would flip e_{ij}, e_{kl} from "off" to "on", and e_{il}, e_{jk} from "on" to "off". This is the quartet of flips ϵ_{iljk}. Without loss of generality by the symmetry of the quartets, let e_{ij} and e_{il} be the edges sampled by g_0. The total number of edges in the network is constant and the initial edge is sampled uniformly, so $p(e_{ij}|Y = y, \omega) = p(e_{il}|Y' = y', \omega)$.

Next, $g_1(e_{kl}|e_{ij}, Y = y, \omega)$ is (effectively, notwithstanding the iteration cap) uniform over the set of all edges that form a quartet. Because they both form quartets, $p(e_{kl}|e_{ij}, Y = y, \omega) > 0$ and $p(e_{kj}|e_{il}, Y' = y', \omega) > 0$. All edges that complete a quartet with e_{ij} in y also complete a quartet with e_{il} in y' except for e_{kl}, which exists in y but not y', and e_{kj}, which exists in y' but not y. Therefore, the cardinalities of the sets of edges that complete a quartet are equal, and so the uniform sampling probability is equal, $p(e_{kl}|e_{ij}, Y = y, \omega) = p(e_{kj}|e_{il}, Y' = y', \omega)$. So $g(Y'|Y = y, \omega) = g(Y|Y' = y', \omega)$.

Therefore, we can update the Metropolis-Hastings acceptance ratio to be

$$\alpha(y, y') = \exp\left[\theta^T(s(y') - s(y))\right]$$

$s(y') - s(y)$ is very simple to calculate because each quartet changes only 4 ties. In fact, the number of violation-concordant edges either increases by two (where i and l share a status and j and k share the opposite status), decreases by two (where i and j share a status and k and l share the opposite status), or does not change (every other case). The same is true of violation discordant edges.

Intra-industry edges are unconstrained and independent of all other edges, so are drawn as independent Bernoulli trials with probability of success determined by the ERGM parameters.

3.1.6 ERGM parameters

We use three node mixing parameters, $\theta_1, \theta_2, \theta_3$, to inject homophily on the violation variable into the model. The associated network statistics, denoted $s_1(y), s_2(y), s_3(y)$ respectively, are the number of edges between nodes without violations, between a node without violations and one with violations, and between nodes with violations. Their explicit definitions are

$$s_1(y) = \sum_{e_{ij} \in E(y)} \mathbb{1}\left[v(y_i) = 0 \wedge v(y_j) = 0\right]$$

$$s_2(y) = \sum_{e_{ij} \in E(y)} \mathbb{1}\left[(v(y_i) = 1 \wedge v(y_j) = 0) \vee (v(y_i) = 0 \wedge v(y_j) = 1)\right]$$

$$s_3(y) = \sum_{e_{ij} \in E(y)} \mathbb{1}\left[v(y_i) = 1 \wedge v(y_j) = 1\right]$$

where $v(y_i)$ is the violation status of the node y_i

These parameters fix the probability that an edge within an industry exists. These parameters also determine the expectation of the total number of edges within an industry. Currently, we primarily use two sets of parameters, one for high homophily, $\theta_{\cdot h}$, and one for low homophily $\theta_{\cdot l}$. We also implement a very high homophily model that uses parameters $\theta_{\cdot vh}$, and a combination of the low and high homophily settings where there is low homophily using $\theta_{\cdot l}$ only for ties to seeds, and high homophily using $\theta_{\cdot h}$ for the rest of the network.

These ERGM parameters define a model that, when unconstrained, is edge-independent, meaning that the existence of any individual edge is independent of the existence of any other edge. This is the case for all (non-seed) intra-industry edges. However, using the same ERGM parameters for all industries will cause nodes in larger industries to have significantly higher degrees than within smaller industries; the probability of any edge being in the network is only dependent on the violation statuses of the nodes, meaning that the larger industries that have more potential edges will also have more edges.

This is obviously a discrepancy between the network model and the true social network of workers; it is not likely that, for example, schoolteachers know more schoolteachers than nuclear physicists know other nuclear physicists merely because there are more schoolteachers than nuclear physicists in the country, let alone that the scale of this discrepancy scales one-to-one (if violation rates are equal) with the difference in the sizes of those populations. But the discrepancy is largely due to the model not accounting for hierarchical structures of the network, like clustering. Because at most 2 workers are being referred within industries, the total number of ties to workers within an industry has little impact on the sampling so long as most workers have at least two connections within their industry, as most do given the negative binomial visibility parameters, and that the distribution of those ties accounts for the network homophily, which is the case. The total number of ties of a node is used as a proxy for the visibility of that node, which itself is used as a proxy for the inclusion probability of that node, but this scaling reflects the actual mechanics of sampling in the simulated network, and should function similarly in the true network even with significantly more ties within the largest industries.

The use of the same ERGM parameters for each industry also causes homophily effects within each industry (and between each pair of industries) to vary due to varying rates of violation prevalence. This effect is more problematic because it biases estimators like Gile's SS estimator that attempt to estimate and correct for uniform homophily effects across the network. Adding more ERGM parameters that enable varying edge existence probabilities by industry and adapting the SS estimator to correct for non-uniform homophily could resolve these issues, but this is not attempted here.

Another likely advantageous adaptation to the SS estimator is using more granular personal network information. In particular, if the inverse probability weights were determined not just by the total number of social ties, but by the specific industries those ties were to, estimation could likely be improved.

3.1.7 Modeling homophily

We define population homophily on violations as the ratio of the expected number of violation-discordant contacts “absent homophily” to the expected number of violation discordant contacts with homophily. Population homophily is therefore not a property of an individual network, but of the network model itself. This can be roughly thought of as a measure of how many times less likely a violation-discordant edge, and can be expressed as

$$h_v(\omega,\theta) = \frac{\mathbb{E}_{y\sim Q(Y)}\left[s_2(y)\right]}{\mathbb{E}_{y\sim P(Y|\omega,\theta)}[s_2(y)]}$$

where $Q(Y)$ is a network model “absent homophily”. The exact choice of $Q(Y)$ is relevant and non-trivial because the ERGM parameters θ are not the only potential source of homophily. The ERGM sets homophily uniformly across all industries by fixing the probability of a homophilous and a discordant edge existing, but the expected number of homophilous and discordant edges within an industry is dependent not just on these probabilities, but also the size of the set of discordant edges within that industry. Here we face a fundamental problem in network statistics: the edge density of a network modeled by independently sampling edges at fixed rates is dependent on the size of the network, and the dependency is $O(n^2)$. Further, in this case, the expected number of discordant edges within an industry i also depends on the violation rate within that industry $\rho(x_{ci})$ itself, with the expectation growing as $\rho(x_{ci})$ approaches 0.5. This way, industries with different violation rates will have different amounts of homophily using the same ERGM paramterers for each industry, and those different amounts of homophily will bias any estimator for the homophily of the whole network that fails to account for the industry structure. These problems all exist for the choice of Q used by default in the RDS package [Handcock et al., 2024] that ignores the industry structure of the network, namely

$$\mathbb{E}_{Q(Y|\theta)}[s_2(y)] = \frac{\mathbb{E}\left[n(E(Y))|\theta,\rho\right]}{n(n-1)}\left(\sum_{i=1}^{n}\mathbb{1}(v_i=0)\right)\left(\sum_{i=1}^{n}\mathbb{1}(v_i=1)\right)$$

where $n(E(Y))$ is the size of the edgelist and $\mathbb{E}\left[n(E(y))|\theta,\rho\right]/(n(n-1))$ is the edge density of the network, and is fixed given θ, the ERGM parameters, and the rate of violations across the whole network ρ. This uniform model of homophily is also the basis for the corrections for homophily effects bootstrapped by Gile's SS estimator. Let $\hat{h}_{\text{vunif}}(y)$ be this standard estimator of homophily for a network y, so

$$\hat{h}_{\text{vunif}}(y) = \frac{n(E(y))}{s_2(y)n(n-1)} \left(\sum_{i=1}^{n} \mathbb{1}(v_i = 0)\right) \left(\sum_{i=1}^{n} \mathbb{1}(v_i = 1)\right)$$

This choice of Q assumes that a network absent homophily is a network where $p(e_{ij} \in y|v_i \neq v_j) = p(e_{ij} \in y|v_i = v_j)$. This is derived from the more fundamental idea that $p(e_{ij} \in y) \perp (v_i, v_j)$. Ideally, in a network without homophily, the probability of any edge existing in the network should be independent of the violation status of the nodes in the edge. However, in this network, even if $p(e_{ij} \in y|v_i \neq v_j) = p(e_{ij} \in y|v_i = v_j)$, there is still dependency because of the constraints. There are many edges excluded by the constraints, and the probability that an edge is excluded by the constraints is dependent on the violation status because violation status is not uniform across industries. This causal pathway between $p(e_{ij} \in y)$ and (v_i, v_j) can only be closed by also conditioning on the sizes of the industries I, J, the constraints ω_{iJ}, ω_{jI}, and the violation rates within each industry $\rho(X_c)$.

As previously noted, we could build a more intelligent homophily structure into the model that accounts for these effects, and similarly adapt the estimator so that the bootstrapped homophily estimates account for this, but this is beyond the scope of this work.

Instead, to partially address these shortfalls, we derive a different model for estimating homophily. This model of homophily assumes that the network absent homophily still maintains the general structure of the homophilous network. It can therefore be used to inform and verify the choice of ERGM parameters θ.

Let Q, the model for a non-homophilous network, be the permutation distribution of violations within industry. This distribution does not have a simple closed form mass function, but it is estimable by sampling new permuted networks $\pi(y)$, and crucially, it respects the structure and constraints of the network. The sampling process generates new violation statuses $\pi(v_I)$ for each node in an industry I that is a simple permutation of the violations in y by industry. This way, the constraints are respected and the edge structure of the network remains unchanged, as do the violation rates within industries, but we can say that $p(e_{ij} \in y') \perp (v_i, v_j)|\rho(X_{cI})$. The form of the permutation homophily estimator $\hat{h}_v(y)$ is

$$\hat{h}_{v\mathrm{perm}}(y) = \frac{\mathbb{E}\left[s_2(y)|\left[p(e_{ij} \in y) \perp (v_i, v_j)|\omega, \theta, \rho\right]\right]}{s_2(y)} = \frac{s_2(\pi(y))}{s_2(y)}$$

In order to reduce the sampling noise of the estimator, we sample multiple permutations and use the mean. The variation of this sampling distribution should be small given the size of each of the industries. This intuition is supported by very small observed sample variance.

Using the overall rate of violations in the network ρ, we can say that about $(1-\rho)^2$ potential edges in the network are between nodes without violations, ρ^2 edges are between nodes with violations, and $2\rho(1-\rho)$ edges are between one node with violations and one without. This is a rough estimate that neglects the constraint structure of the network, and therefore likely underestimates the number of edges between nodes with violations and overestimates the number of edges without violations (and slightly underestimates the number of violation-discordant edges). These discrepancies are acceptable for homophily estimation because we only need parameters that generally fix high and low homophily, and generally fix the edge density. The edge density is fixed so that the nodes in the smallest industry will have on average 5 ties to other nodes in the same industry. This edge density is constant within all industries, and so large industries will have on average more than 5 ties to other nodes within the industry. This means that, for a network with 60000 nodes and the smallest industry in that network comprising about 3% of the network (as is the case here), the mean edge should exist with probability about 0.0027, meaning that

$$(1-\rho)^2\exp(\theta_{1\cdot}) + 2\rho(1-\rho)\exp(\theta_{2\cdot}) + \rho^2\exp(\theta_{3\cdot}) = 0.0027$$

Which is a weighted average of the probabilities that an edge of each type exists, $\exp(\theta_{\cdot\cdot})$, where the weights are the proportion of potential ties of each type in a network with an overall violation rate ρ.

Additionally, we define the high-homophily scenario, denoted $\theta_{\cdot h}$ to have about half as many discordant edges as you would expect without homophily, meaning the probability

$$\exp(\theta_{2h}) = \frac{1}{2}0.0027$$

Further, we define the low-homophily scenario, denoted $\theta_{\cdot l}$, to have about 0.9 as many discordant edges as you would expect without homophily, meaning roughly

$$\exp(\theta_{2l}) = \frac{9}{10}0.0027$$

We also test a "very high" homophily scenario, denoted $\theta_{\cdot vh}$, where discordant edges appear only 1/10 as often as you would expect without homophily, and we expect the survey to fail to measure violation rates, meaning roughly

$$\exp(\theta_{2vh}) = \frac{1}{10}0.0027$$

And finally, we want about even edge density (but not equal probability of edges existing) among nodes sharing violation status, whether that status is "no violation" or "violation", meaning

$$(1-\rho)^2\exp(\theta_{1\cdot}) = \rho^2\exp(\theta_{3\cdot})$$

Solving the system, we find

$$\theta_{1h} = \exp\left(\frac{0.0027(\rho^2 - \rho + 1)}{2(1-\rho)^2}\right),\ \theta_{2h} = \exp\left(0.0027\frac{1}{2}\right),\ \theta_{3h} = \exp\left(\frac{0.0027(\rho^2 - \rho + 1)}{2\rho^2}\right)$$

$$\theta_{1l} = \exp\left(\frac{0.0027(9\rho^2 - 9\rho + 5)}{10(1-\rho)^2}\right),\ \theta_{2h} = \exp\left(0.0027\frac{9}{10}\right),\ \theta_{3h} = \exp\left(\frac{0.0027(9\rho^2 - 9\rho + 5)}{10\rho^2}\right)$$

Using $\rho = 1/3$, which is the average violation rate for the networks we simulate, the coefficients are

$$\theta_{\cdot h} = [-6.1, -6.6, -4.7]^T$$

$$\theta_{\cdot l} = [-6.3, -6.0, -4.9]^T$$

$$\theta_{\cdot vh} = [-5.8, -8.2, -4.5]^T$$

This also fixes the acceptance probabilities α of the Metropolis-Hastings sampler for edges between industries, as enumerated in Table 3.1. The structure of the quartets limits the net changes in total number of discordant edges. It is impossible to draw a quartet with a net change of plus or minus 1 discordant edges. The only quartets that change the total number of discordant edges are those that add or subtract a net 2 discordant edges.

It may be useful to construct an adjustment to the estimator to account for variable homophily within and between each industry, but this is beyond the scope of this study. Because the purpose of this study is to investigate the resilience of the sampling method in homophilous networks, and because we have little data to build an intelligent prior about how homophilous the social network of interest is, we make a compromise in modeling homophily as uniform across the network. While this simplifies the choice of parameters θ, the effects on edge density within industries of various sizes may cause problems. Specifically, the number

Δ discordant edges	α	$\alpha,\ \rho = 1/3, \theta_{\cdot\cdot} = \theta_{\cdot h}$	$\alpha,\ \rho = 1/3, \theta_{\cdot\cdot} = \theta_{\cdot l}$
2	$\exp(2\theta_{2\cdot} - \theta_{1\cdot} - \theta_{3\cdot}))$	0.09	0.45
0	1	1	1
-2	> 1	> 1	> 1

Table 3.1: Acceptance probabilities for perturbations that add and subtract discordant edges

of intra-industry edges for a node in the largest industry will be about 10 times higher than for a node in the smallest industry (because the largest industry is about 10 times larger than the smallest and intra-industry edge density is held constant by industry).

We believe these issues will not significantly impact the sampling because the vast majority of referrals will utilize inter-industry ties, which do have calibrated degrees, and because the ratio of violation-concordant and violation-discordant edges for a particular node remains the same across industries of different sizes due to the fixed probabilities imposed by θ, given equal violation rates in the two industries. The potentially more concerning impact of non-uniform edge density is the effect on estimation. RDS estimators use nodal degree to weight the sample, meaning that workers from large industries with many ties could have artificially lower weights than workers in smaller industries. However, this effect should significantly lessened for estimation of rates within industries, which is the primary focus of this study.

Any more complicated effects on ratios of concordant and discordant edges caused by different rates of violations is outside the scope of this study, but a potential model for these effects could involve generating a large set of model parameters $\theta(\rho_{\cdot})$, which intelligently set the probabilities that an edge exists within or between any industry(ies), perhaps in some symmetric tensor $\Theta(\rho_{\cdot})$ such that $\theta_{ij\cdot}$ is the set of three parameters that fix probabilities of concordant and discordant edges from industry i to industry j to maintain uniform edge density and homophily (or to inflate or deflate them). This model would add significantly more degrees of freedom, making not just parameter estimation for setting homophily noisier, but also further complicating inference from the sample using an RDS estimator that estimates and corrects for variable homophily in referrals. We do investigate a very simple version of this idea for network simulation with lower homophily for ties to seeds than the network contains as a whole. (put in conclusions?)

We make the assumption that, for the purposes of exploring the rough contours of the parameter space, the simplified estimator of homophily we use to calibrate the ERGM parameters (and used by the SS bootstrap to estimate violation rates corrected for homophily from the sample, which is the bigger problem here) are adequate to assess the validity of the RDS methods under generally low and high homophily. (Also in conclusions)

3.1.8 Network size

The process of sampling from the network model is computationally intensive, and the seeding and running of the Metropolis-Hastings algorithm is at least $O(n^2)$, so we use networks that have 60000 nodes. 60000-node networks are small enough to be computationally reasonable while large enough to be useful for validating a survey of this scale. Larger networks would be more desirable, especially for validating estimation among the smallest industries where the target sample size can be as high as 1/3 of the population of that industry in the network, but these effects can be mitigated by adjusting the size of the confidence intervals of the estimators for the finite population effects.[8]

3.1.9 Finite population effects

This finite population issue relates to the earlier question of how to evaluate the accuracy of the estimators; the probability of having a violation in in each industry ρ_c is sampled from a uniform distribution, but the empirical violation rate varies from ρ_c because the total

[8] Because the seeds are unique, the potential of generating separate networks for each individual seed was considered, but the structure of existing RDS modeling and sampling functions in the RDS package [Handcock et al., 2024] made a single, large network with all seeds more efficient.

number of violations n_{vc} in the industry is binomial with $p = \rho_c$, specifically

$$n_{vc} = \sum_{i=1}^{n_c} X_{vi} \sim \text{Binom}(\rho_c, n_c)$$

The question is then whether to use the generating parameter ρ_c or the empirical frequency n_{vc}/n_c as the target of the estimator when assessing its validity. The RDS estimators will converge to n_{vc}/n_c, but n_{vc}/n_c itself converges to ρ_c with variance $\frac{\rho_c(1-\rho_c)}{N_c}$.

The concern is that when constructing confidence intervals from the simulated samples, the estimate of the variance will be smaller than from the true sample which will be from an effectively infinite population. Consider the variance of the RDS estimator, $\text{Var}(\hat{\rho}_c)$. By the law of total variance,

$$\text{Var}(\hat{\rho}_c) = \mathbb{E}[\text{Var}(\hat{\rho}_c|n_{vc}/n_c)] + \text{Var}(\mathbb{E}[\hat{\rho}_c|n_{vc}/n_c])$$

The first term is just the variance of the estimator given the finite population rate, which can be adjusted using the finite population correction, so letting N_c be the (fixed) size of the RDS sample from industry c,

$$\text{Var}(\hat{\rho}_c) = \frac{n_c - N_c}{n_c - 1}\mathbb{E}[\text{Var}(\hat{\rho}_c|\rho_c)] + \text{Var}(\mathbb{E}[\hat{\rho}_c|n_{vc}/n_c])$$

Now, consider the variance conditioning on the infinite population rate ρ_c.

$$\text{Var}(\hat{\rho}_c) = \mathbb{E}[\text{Var}(\hat{\rho}_c|\rho_c)] + \text{Var}(\mathbb{E}[\hat{\rho}_c|\rho_c])$$

This is the same estimator, so

$$\mathbb{E}[\text{Var}(\hat{\rho}_c|n_{vc}/n_c)] - \mathbb{E}[\text{Var}(\hat{\rho}_c|\rho_c] = \text{Var}(\mathbb{E}[\hat{\rho}_c|\rho_c]) - \text{Var}(\mathbb{E}[\hat{\rho}_c|n_{vc}/n_c])$$

The right-hand side represents the improvement in the estimation we receive in the simulation because sampling from a finite population explains more variance in the sampling than from an infinite population.

The unexplained variance $\mathbb{E}[\mathrm{Var}(\hat{\rho}_c|\rho_c)]$ is the sampling variance from an infinite population using the RDS methodology. This is the sampling variance we estimate when constructing confidence intervals in the real survey. The simulation instead estimates $\mathbb{E}[\mathrm{Var}(\hat{\rho}_c|n_{vc}/n_c)]$.

Therefore, we implement the finite population correction for this variance

$$\mathrm{Var}(\hat{\rho}_c|\rho_c) = \frac{n_c - 1}{n_c - N_c}\mathrm{Var}(\hat{\rho}_c|n_{vc}/n_c)$$

so that we can simplify the above equation (LABEL) to

$$\left(1 - \frac{n_c - 1}{n_c - N_c}\right)\mathbb{E}[\mathrm{Var}(\hat{\rho}_c|\rho_c)] = \mathrm{Var}(\mathbb{E}[\hat{\rho}_c|n_{vc}/n_c]) - \mathrm{Var}(\mathbb{E}[\hat{\rho}_c|\rho_c])$$

So using the finite population correction when constructing confidence intervals will exactly correct the variance estimation to account for the finite sampling effects in the simulation. Therefore, we use this correction when constructing intervals, and we test the validity against the infinite population rate ρ_c, which simulates the true estimation of ρ_c from a sample of an effectively-infinite population.

3.1.10 Finite population effects in the SS estimator

The SS estimator is particularly useful when the sampling fraction is large. This is the case in the simulations (for some industries), but is not the case for the true social network of workers. The version of Gile's SS estimator used here, however, does not discriminate by industry, so any finite population effects in small industries are mostly not accounted for. Therefore, when using Gile's SS estimator, we set the population size parameter to the size of the network and still implement the standard finite population correction to the confidence interval for each industry, and rely on any mitigation of these effects by the estimator being very small due to the relatively small sampling fraction for the network as a whole, which is about 1/6, which, compared to the smallest industries that often have sampling fractions in excess of 1/3, will have much lower impact on the confidence interval. This choice will slightly inflate the length of the confidence intervals, however.

If we were to adapt the SS estimator to account for the sampling restrictions, it may also be worthwhile to measure the accuracy of this estimator against the empirical violation rate within each network rather than against the generating Bernoulli parameter; because the design of the estimator accounts for finite population effects in the sampling likelihood, one can think of it as more realistically representing estimation from a much larger network with a very low sampling fraction, meaning that there is no need to artificially replicate the variance by measuring accuracy against the generating parameter.

3.2 Simulation of the respondent-driven sampling

In order to simulate the constraints on the recruitment structure, edits were made to the RDS sampling function in the RDS package [Handcock et al., 2024]. These edits enforced the constraint of a maximum of two referrals per industry, set a target sample size for each industry, and adjusted the order of industries that units were asked to recruit from to efficiently reach the target sample size within each industry.

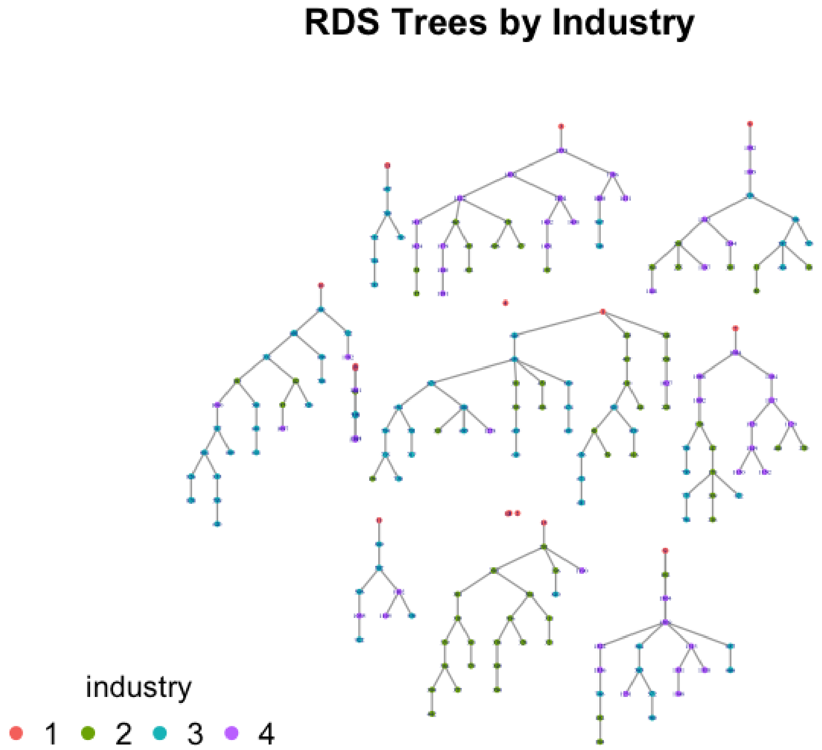

Figure 3.3: Recruitment trees by industry

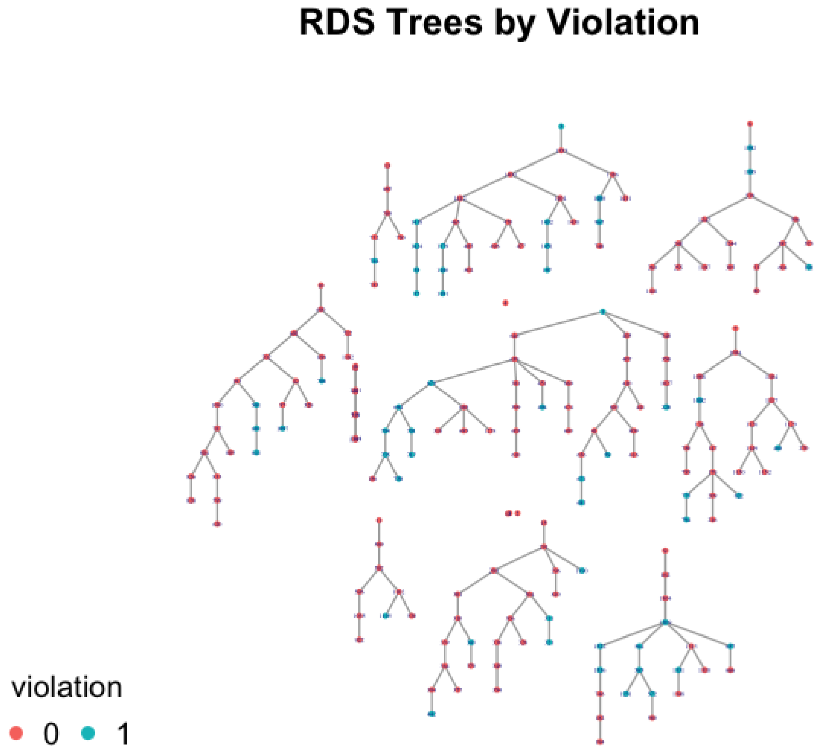

Figure 3.4: Recruitment trees by violation status

The sampling model includes the referral rate parameter ϕ, which is the likelihood that an existing tie in the network results in a successful recruitment. When asked to make a referral to a particular industry, a worker checks their social ties to see if they have any to that industry, randomly selects one (or two, if enabled), and that worker is recruited successfully with probability ϕ. We use $\phi = 0.20$ as a default, which is informed by actual referral rates in RDS surveys. 0.2 is at the lower end of empirical referral rates, and was chosen to investigate sampling in poor conditions. The 0.8 probability of an attempted recruit failing represents both the scenario where the referred worker never responds and the scenario where the referred worker responds but is ineligible due to an error in recruitment, either because they do not work in the industry they were referred for, or they are an exempt worker. We examine the likelihood of successfully recruiting a complete sample and the accuracy of inference as ϕ varies.

The sampling trees shown in figures 3.3 and 3.4 use a referral rate of $\phi = 0.5$ and a lower concentration of seeds than the actual simulations to better visualize the sampling constraints and homophily. Note how no unit refers more than two workers from a single industry, and the concentration of violations within the trees. These patterns also occur in the larger samples, but the trees are on average much shorter and have less branching.

We also examine the effect of raising the limit on recruits per industry from one to two. This can increase the likelihood of collecting a complete sample even when the recruitment rate is low, and has minimal effects on inference assuming the two recruits are from the same workplaces. Even if the two recruits are highly correlated, it is unlikely that both are actually recruited. The benefit is instead the increased probability that at least one is recruited in a smaller, low-visibility industry that is more difficult to collect a complete sample from.

3.3 Parameters of interest

Simulation parameters	Values
Network Size (n)	60000
Seeds	4000
Seeds in low-wage industries	200
Industries	12
NB parameters (r, p)	$(5, 0.7)$
$\theta_{\cdot l}$	$[-6.3, -6.0, -4.9]^T$
$\theta_{\cdot h}$	$[-6.1, -6.6, -4.7]^T$
$\theta_{\cdot vh}$	$[-5.8, -8.2, -4.5]^T$
Metropolis-Hastings iterations	10^7
Recruit rate (ϕ)	0.2
Max recruits per industry	2
Max recruits	5

Table 3.2: Simulation parameters

Industry	Proportion of network	NB Inflation parameter (ϕ_z)	Violation rate (ρ)
Seeds	0.067	NA	0.050
1	0.031	0.80	0.313
2	0.031	0.80	0.335
3	0.031	0.75	0.416
4	0.041	0.75	0.246
5	0.041	0.70	0.384
6	0.041	0.70	0.351
7	0.041	0.70	0.217
8	0.051	0.65	0.428
9	0.062	0.65	0.217
10	0.072	0.60	0.157
11	0.246	0.55	0.164
12	0.246	0.50	0.209

Table 3.3: Industry parameters

CHAPTER 4

Results and Discussion

To investigate the performance of the adapted SS estimator, we simulated networks with varying levels of homophily and sampled from them using a model of the RDS design.

In general, the adapted SS estimator is fairly accurate, but the simulations show that the sample mean is a more accurate and lower variance estimator with better-calibrated confidence intervals. For example, in the low homophily model, we observed an 85.4% coverage rate for the SS estimator and an 89.7% coverage rate for the sample mean. We believe that the most likely cause of this is the previously articulated issues arising from the SS estimator's inaccurate assumption of and adjustment for uniform homophily, which can reduce its effectiveness.

Homophily effects, as shown in table 4.1 are, at first examination, opposite in direction and magnitude than expected. Networks generated from the high-homophily parameters $\theta_{\cdot h}$ produce samples and estimators with higher confidence interval coverage rates, lower bias, and lower variance than networks generated from low-homophily parameters $\theta_{\cdot l}$. These patterns exist for both the SS estimator and the sample mean. Further investigation indicates that this result is due, at least in part, to lower convergence rates of the Metropolis-Hastings network sampler for high-homophily models and lower homophily effects for the ties most likely to facilitate referrals, the inter-industry ties to large industries and seeds. These effects disproportionately lower homophily in sampling relative to the homophily observed in the network overall and the homophily implied by the network model. This means that the notionally high-homophily networks, those sampled from high-homophily models with high observed homophily in both the traditional and permutation measures, produce samples with significantly lower-than-expected homophily.

4.1 Results

The primary measure of validity of the the methodology and inference is the rate at which confidence intervals capture the true parameter from the simulated network, or the coverage rate. We also examine the length of the confidence intervals, which affects the power of the inference, and the systematic bias of the estimators within each industry.

Simulation runs for each set of ERGM parameters $\theta_{..}$ consisted of 10 networks sampled from the constrained ERGM. 100 RDS samples were produced for each network, for a total of 1000 samples. For each sample, both the modified SS estimator and the sample mean were used to construct point estimates and 95% confidence intervals. The rates at which the true parameter was captured by the confidence intervals, average confidence interval lengths, and both the bias and mean absolute errors of the point estimates are reported in the tables below. We also simulated networks with hybrid ERGM parameters, where ties to seeds were sampled with $\theta_{\cdot l}$ and all other ties were sampled with $\theta_{\cdot h}$, to investigate the possibility that the weaker ties to the non-low-wage seed population had less homophily.

Homophily	$\hat{h}_{\text{vunif}}(y)$	$\hat{h}_{\text{vperm}}(y)$	Mean Coverage		Bias		Mean abs. error	
Parameters			$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$	$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$	$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$
Low	1.23	0.98	0.854	0.897	0.0293	0.0196	0.0351	0.0234
Hybrid	1.93	1.65	0.913	0.923	0.0166	0.0119	0.0289	0.0207
High	1.95	1.67	0.923	0.967	0.0139	0.0095	0.0274	0.0194
Very High	5.98	5.83	0.940	0.937	0.0030	0.0024	0.0250	0.0190

Table 4.1: Estimator performance by ERGM parameters

Industry	Size	ρ	Coverage		CI length		Bias		Mean abs. error	
			$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$	$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$	$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$	$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$
1	1829	0.313	0.938	0.939	0.127	0.096	0.0195	0.0124	0.0287	0.0208
2	1864	0.335	0.873	0.876	0.138	0.105	0.0330	0.0240	0.0383	0.0281
3	1806	0.416	0.912	0.954	0.141	0.108	0.0246	0.0118	0.0333	0.0224
4	2450	0.246	0.956	0.962	0.120	0.089	0.0156	0.0104	0.0248	0.0176
5	2466	0.384	0.875	0.881	0.129	0.095	0.0278	0.0175	0.0342	0.0243
6	2463	0.351	0.938	0.968	0.129	0.094	0.0169	0.0056	0.0278	0.0172
7	2440	0.217	0.962	0.965	0.115	0.084	0.0084	0.0033	0.0234	0.0166
8	2974	0.428	0.897	0.901	0.137	0.098	0.0270	0.0199	0.0338	0.0249
9	3680	0.217	0.954	0.977	0.114	0.082	0.0043	0.0019	0.0213	0.0150
10	4341	0.157	0.945	0.939	0.101	0.071	0.0020	0.0120	0.0206	0.0150
11	14863	0.164	0.920	0.943	0.090	0.067	-0.0069	0.0005	0.0201	0.0140
12	14824	0.209	0.907	0.934	0.098	0.074	-0.0055	0.0061	0.0228	0.0163
Totals	56000	0.223	0.923	0.967	0.120	0.089	0.0139	0.0095	0.0274	0.0194

Table 4.2: Estimator performance with $\theta_{\cdot h}$

Industry	Size	ρ	Coverage		CI length		Bias		Mean abs. error	
			$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$	$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$	$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$	$\hat{\rho}_{SS}$	$\hat{\rho}_{mean}$
1	1829	0.313	0.747	0.816	0.126	0.097	0.0438	0.0299	0.0453	0.0314
2	1864	0.335	0.814	0.890	0.136	0.105	0.0408	0.0253	0.0428	0.0279
3	1806	0.416	0.786	0.879	0.138	0.108	0.0437	0.0281	0.0456	0.0306
4	2450	0.246	0.925	0.967	0.120	0.090	0.0268	0.0135	0.0304	0.0184
5	2466	0.384	0.777	0.883	0.127	0.095	0.0407	0.0247	0.0430	0.0271
6	2463	0.351	0.800	0.907	0.128	0.095	0.0371	0.0191	0.0402	0.0236
7	2440	0.217	0.898	0.929	0.117	0.086	0.0266	0.0165	0.0312	0.0209
8	2974	0.428	0.793	0.858	0.135	0.098	0.0422	0.0231	0.0449	0.0270
9	3680	0.217	0.920	0.952	0.116	0.083	0.0222	0.0083	0.0287	0.0168
10	4341	0.157	0.915	0.923	0.106	0.073	0.0211	0.0136	0.0273	0.0181
11	14863	0.164	0.946	0.913	0.091	0.069	0.0008	0.0133	0.0182	0.0170
12	14824	0.209	0.921	0.846	0.099	0.076	0.0060	0.0194	0.0232	0.0223
Totals	56000	0.223	0.854	0.897	0.120	0.090	0.0293	0.0196	0.0351	0.0234

Table 4.3: Estimator performance with $\theta_{\cdot l}$

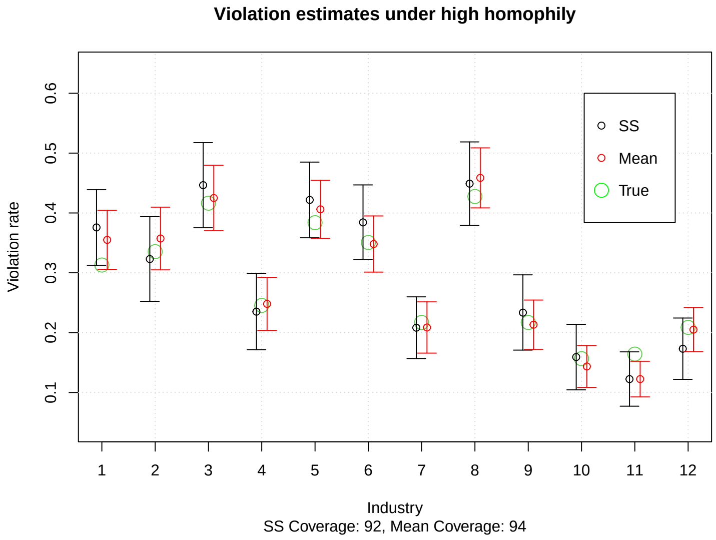

Figure 4.1: High Homophily Estimates

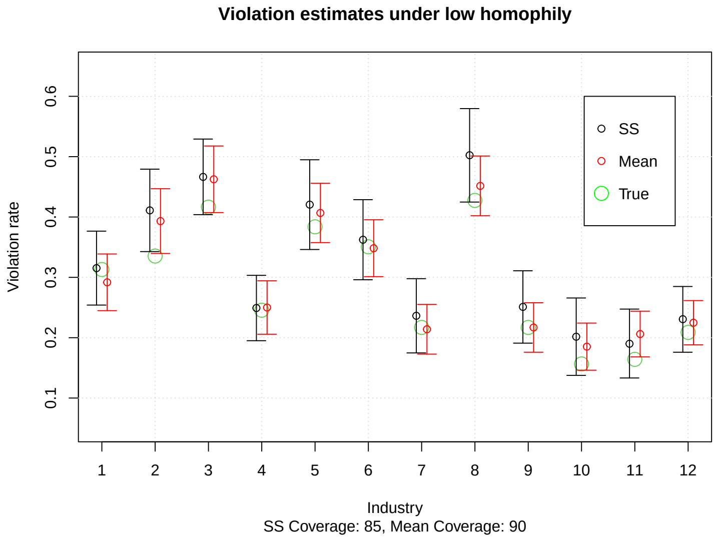

Figure 4.2: Low Homophily Estimates

4.2 Discussion

While these results initially appear counterintuitive, a plausible explanation emerges when we observe that there are two distinct fairly non-intuitive patterns in these results: lower accuracy of the SS estimator compared to the sample mean in every level of homophily tested, and increased accuracy of both estimators in networks sampled from high-homophily models. These distinct patterns can be plausibly explained by two separate phenomena, both affecting homophily: First, the uniform homophily effects from the SS estimator does more harm than not correcting for homophily at all. Second, convergence of Metropolis-Hastings is significantly slower in high-homophily networks[1], and so the high homophily models actually produce networks that have very little to no homophily because the seed network has null homophily.

[1] This effect could be complex and related to a very non-convex likelihood space for the network models, where the MH algorithm does cannot sufficiently explore the space of high-homophily models because it cannot escape the local minimum it is initiated in.

4.2.1 Homophily effects

The most troubling result is that both estimators appear to perform better when the network being sampled from has higher homophily, which is contrary to our understanding of homophily effects, especially for the sample mean. Examining the homophily of the inter-industry edges, however, we observe significantly lower homophily than expected. The reported homophily estimates used to verify the network sampling appear to be in accordance with the model parameters only because of intra-industry ties. Further investigation indicates that the Metropolis-Hastings convergence rate is much lower for network models with high homophily than those with low homophily, which explains why the inter-industry ties, the only ones sampled using Metropolis-Hastings, have lower homophily; the sampling does not deviate far enough from the seed network that exhibits null homophily. When the Metropolis-Hastings sampler was initially tested, convergence was only verified for the low-homophily case, and further increasing the runtime of the sampling was therefore deemed unnecessary and impractical. This was clearly a mistake, and further simulations should be run either with more time for network sampling or a better network sampling model. A simple test of the homophily of inter-industry ties among networks with significantly higher runtime reveals that homophily parameters are still not consistent with expectations, see table 4.4, meaning that changes to the optimization algorithm are likely necessary.

Homophily	All edges		MH edges		Seed edges	
Parameters	$\hat{h}_{v\text{unif}}(y)$	$\hat{h}_{v\text{perm}}(y)$	$\hat{h}_{v\text{unif}}(y)$	$\hat{h}_{v\text{perm}}(y)$	$\hat{h}_{v\text{unif}}(y)$	$\hat{h}_{v\text{perm}}(y)$
Low	1.24	0.98	1.15	1.16	0.97	1.05
Hybrid	1.94	1.66	1.60	1.61	0.97	1.06
High	1.95	1.66	1.62	1.63	1.09	1.18
Very High	6.03	5.83	2.65	2.65	1.27	1.37

Table 4.4: Measures of homophily for networks with increased sampling time

These effects disproportionately lower homophily in sampling relative to the homophily observed in the network overall and the homophily implied by the network model. This means that while the networks sampled from are technically high-homophily, and certainly come from a model with high homophily parameters, the effective homophily in recruitment chains is not high. This explains why the expected homophily effects were not observed. Preliminary investigation of the homophily in the recruitment trees supports this theory.

However, table 4.4 indicates that this is not the only problem. While the homophily of inter-industry ties in high homophily networks is significantly lower than would be implied by the model, it is not lower than the homophily of the same ties in the low homophily models. Therefore, this phenomena would only explain why higher homophily models do not perform worse than low homophily models, not why they, in fact, perform better[2]. The performance of the sample mean, especially the very low bias for the very high homophily model, would indicate that the as homophily in the model increases, the samples grow closer to simple random samples[3].

For the sake of argument, suppose that this is the case, and the effective homophily in the samples from low homophily networks is in fact higher than the samples from high homophily networks. In addition to explaining the patterns in the coverage rates and bias, this would also explain the particular pattern of biases observed in what would now be the highest homophily samples (those from the low homophily model). The only industries where the SS estimator outperformed the sample mean in coverage and bias were the largest industries, 11 and 12. These industries make up about half the network together, and are also overrepresented in the RDS samples, which explains the lower errors and shorter

[2] One possible explanation is that the low-homophily model causes effectively negative homophily in the samples of some industries, which causes more bias than the higher effectively null homophily of the higher-homophily models. However, this is merely speculation.

[3] Another explanation of the extremely good performance of the sample mean is that the primary source of bias in RDS designs is differential activity causing differential visibility. This is the source of bias RDS estimator designs are primarily concerned with, but our network models do not contain differential activity. It is possible that the incidental differential activity/visibility decreases in high homophily models. This is also just speculation, though

confidence intervals for those industries. They also happen to have violation rates that are close to the overall violation rate of the networks. When the SS estimator calibrates the homophily adjustment, it will be highly informed by the samples from the largest industries and closely match the ideal homophily adjustment for those industries. This implies that when homophily is a problem, the SS estimator's correction for it, when well calibrated, is an improvement on the sample mean. This is speculative though, and reliant on these samples actually having the supposed homophily patterns.

4.2.2 Sequential-sampling estimator performance

In general, the SS estimator slightly underperforms the sample mean in coverage, bias, and absolute error, all while the sample mean produces smaller confidence intervals (because it does not consider network effects when calculating the standard error). This implies that the effective homophily in the samples is not significant enough to greatly bias the sample mean, which does nothing to account for it. It also implies, separately, that the adjustments for homophily and sampling visibility made by the SS estimator are generally doing more harm than good. There are a variety of potential explanations for this, two of which are particularly promising: that the SS estimator falls short in replicating the complex sampling mechanism, and that the uniform homophily correction being applied to every industry fails to anticipate varying homophily effects.

The unadjusted SS estimator assumes no constraints or targeting in the recruitment process, and therefore the weights it generates ignore the sub-population structure of the sampling. This means that the successive sampling process makes no distinction between nodes of different industries, and the differences in degree distribution between industries will cause errors in the weighting of these nodes, including the previously mentioned inflation of degrees within large industries. The estimator imputes equal probability for each edge in a personal network to facilitate recruitment, but this is not the case.

This is partially corrected for by the adjustment to the degree of each node to account for higher likelihoods of recruiting a sample if the recruiter is from a larger industry. This adjustment is rough and reliant on the accuracy of the prior knowledge of industry sizes, and while it significantly reduced the bias of the estimator, a more precise adjustment that implemented a more nuanced model of visibility in the sampling would likely improve the estimator further. This correction may have different effects depending on the specifics of how industries are prioritized as the survey progresses, and whether multiple recruits are used, as these changes will effect the sampling probability of workers in particular industries. For example, even though workers in the rarest industries have fewer social ties, meaning that this adjustment would increase the weight of any workers referred by a worker in a rare industry, the true probability of sampling a particular worker in a rare industry could be much larger than a worker in a larger industry because workers in rare industries have higher priority of being sampled, and referrals to those industries are prioritized when soliciting referrals.

In addition to a more nuanced estimation of referrals occurring from a particular industry, the use of more granular personal network data, specifically the number of connections to each industry, could also improve the estimator. A simple change to the estimator that would likely improve accuracy is a further rescaling of the personal network size parameter to incorporate this data. Instead of using the size of an industry as a proxy for the likelihood of the industry initiating a referral, we could use the actual number of referrals made by workers in each industry to bootstrap a more precise estimate of that likelihood. This simple change would only require restructuring the sampling simulation to record the industries ties are connected to, and we therefore recommend that it should be studied and, if beneficial, implemented for a survey utilizing this design.

Other changes that would respect the successive sampling model instead of merely rescaling personal network size would be more complicated. They would involve not just restructuring the RDS sampling code to record that information, but also restructuring the fairly complex bootstrap of the sampling the SS estimator utilizes to accurately reconstruct networks with the proper degree constraints. Here, the sampling simulation itself could be a powerful tool in bootstrapping this mapping of sets of nodal degrees to inclusion probability, making estimation of both visibility and homophily effects more precise [4].

These adjustments, along with applying different homophily adjustments to each type of recruitment, could significantly improve this estimation. It is plausible that, with these adjustments, performance of the SS estimator would surpass the sample mean, particularly in coverage and bias. However, the feasibility of these rather significant modifications to the estimator are unclear, especially considering that the successive sampling implemented by Gile relies on assumptions of the structure of the configuration model that no longer hold for the constrained RDS.

The theory that the uniform homophily adjustment causes bias is corroborated by a consistent pattern of bias being higher in the SS estimator than in the sample mean for industries with high violation rates. The magnitude of homophily correction applied to the sample depends on the overall prevalence of violations in the sample. Because the sample includes seeds with very low rates of violations and the largest industries happen to have low violation rates, the correction is calibrated for a lower prevalence of violations, increasing the weights of nodes with violations more than the case where similar homophily but overall higher prevalence of violations was observed. This is because homophily effects on the point estimate have higher magnitude the further the overall prevalence of the homophilous variable is from 0.5. This correction is therefore about right for the industries with smaller violation rates, but is too high for industries with higher violation rates, biasing the estimator in the positive direction.

[4] The design of the network simulation is well-suited to taking sequences of degrees to other industries and constructing a network that very closely represents those degrees, exactly like successive sampling. These networks are not randomly sampled from a prior null network model, but the process is remarkably similar to the configuration model used by Gile's SS, and random edge sampling could easily be implemented. It may also be worth considering the construction of some null network model that maintains the industry constraints but has sampling equivalence to a version of successive sampling, which would make the bootstrap considerably faster. This, or some other bootstrap of the simulation to estimate the distribution of degrees and sampling inclusion probabilities via the simulation used here, could yield significant improvements for inference for complex RDS designs

The homophily adjustment is also known to perform poorly and have low coverage rates when the initial seeds are biased [Gile, 2011]. Here, even though the seeds are not technically biased, the effect of having significantly lower rates of violations among the seeds replicate many of the issues seed bias causes. Incorrect homophily adjustments caused by quasi-"seed bias" and short chains could also cause the observed lower coverage rates than the sample mean, which does not attempt to adjust for observed homophily. It is also relevant to note that these homophily adjustments cannot discriminate between homophily in the network and bias in recruit selection. Any bias in recruitment of a unit's contacts correlated with the violation status of those contacts is modeled by neither the sampling model nor the estimator. We have assumed that recruit selection is uniform across the set of relevant ties in the sampling model, but it is possible that this may not hold for real referrals, which could cause the homophily adjustment to perform even more poorly.

4.2.3 Estimator selection

While we ultimately chose to adapt Gile's SS estimator, other estimator designs and strategies were considered. The strategies for estimating weak-tie personal network size presented in [Feehan et al., 2022] were considered useful due to the applicability to recruitment from many sub-populations, especially for inferring visibility to the seed population. This visibility is difficult to measure absent the likely difficult-to-collect data of respondents' mutual ties to the broad population of covered workers from which the seeds are drawn. However, we decided that simplest and most readily available (given the sampling design) implementation of these visibility inference strategies for use in network scale-up estimation [Feehan and Salganik, 2016] relied too heavily on the assumption that the number of ties to low-wage industries was a good approximation for ties to covered workers in general.

We later considered using a homophily configuration graph model [Fellows, 2018] because of its robustness to homophily in sampling, but this proved complex when applying the model to this specific problem, and Gile's SS estimator already included a homophily correction.

In addition to further adaptation of the SS estimator, future research into estimation for short-chain RDS with probability-sampled seeds and multiple sub-populations of interest using either network scale-up methods or the homophily configuration graph could prove fruitful.

4.2.4 Recruitment rate effects

The recruitment rate used in the simulations was 20%, which is at the low end of recruitment rate observed in real RDS surveys. When loosening the recruitment constraints to a maximum of 2 recruits per industry, no simulation failed to recruit a complete sample.

Simulations run with a maximum of 1 recruit per industry at a 20% recruitment rate failed to reach a complete sample almost always, often leaving the smallest industries with samples of less than 250. When the recruitment rate was raised to 30%, all samples were complete 95% of the time, even with only 1 recruit per industry.

When the recruitment rate was further reduced to 15%, even allowing 2 recruits per industry failed to reach a complete sample approximately 10% of the time. This indicates that if the true recruitment rate drops below 15%, there is a significant risk of the sampling failing. However, even in the cases where some industries had a sample size of less than 400, there were more than 350 respondents, meaning that inference could still proceed with only minor increases in the standard error.

4.3 Conclusion

We observe that inference using this design and the adapted SS estimator is fairly accurate. Despite the odd patterns in coverage rates and homophily, the accuracy of the adapted SS estimator is promising. While it is not perfect, a minimal coverage rate of 85% is evidence that this design is worth pursuing, especially with further improvements to the estimator. We recommend additional adjustments to the network size parameter that incorporate the number of ties workers have to each industry to fully analyze the validity of the inference. The simplest forms of these adjustments do not require changes to the sequential sampling bootstrap.

Additionally, the value of the simulation in testing and refining the sampling design is high. The need to loosen recruiting constraints in order to avoid a failure of the sampling was an important insight that would have been difficult to assess without the simulation. With more precise data of the particular visibilities and recruitment behavior of workers in each low-wage industry, the simulation can be better calibrated and further investigation will be possible.

We also believe that the simulation may be valuable for use in estimation. With a well-calibrated model, the simulation can be an effective bootstrap of the sampling, and therefore improve estimation of homophily effects and inclusion probability.

This study indicates that the proposed design is an improvement on methods previously used to investigate labor violations among low-wage workers. We believe not only that this design is feasible for this problem, but also that there is significant potential for use of representatively-seeded RDS to investigate many hard-to-sample populations.

REFERENCES

[Bernhardt et al., 2009] Bernhardt, A., Milkman, R., Theodore, N., Heckathorn, D. D., Auer, M., DeFilippis, J., González, A., Narro, V., and Perelshteyn, J. (2009). Broken laws, unprotected workers: Violations of employment and labor laws in america's cities. *UCLA: Institute for Research on Labor and Employment.*

[Feehan et al., 2022] Feehan, D., Hai Son, V., and Abdul-Quader, A. (2022). Survey methods for estimating the size of weak-tie personal networks. *Sociological Methodology*, 52.

[Feehan and Salganik, 2016] Feehan, D. M. and Salganik, M. J. (2016). Generalizing the network scale-up method: A new estimator for the size of hidden populations. *Sociological Methodology*, 46(1):153–186. PMID: 29375167.

[Fellows and Handcock, 2012] Fellows, I. and Handcock, M. S. (2012). Exponential-family random network models.

[Fellows, 2018] Fellows, I. E. (2018). Respondent-driven sampling and the homophily configuration graph. *Statistics in Medicine*, 38(1):131–150.

[Gile, 2011] Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association*, 106(493):135–146.

[Gile et al., 2018] Gile, K. J., Beaudry, I. S., Handcock, M. S., and Ott, M. Q. (2018). Methods for inference from respondent-driven sampling data. *Annual Review of Statistics and Its Application*, 5(Volume 5, 2018):65–93.

[Handcock et al., 2024] Handcock, M. S., Gile, K., Fellows, I. E., and Neely, W. W. (2024). Rds: Respondent driven sampling (version 0.9-10).

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

[Hunter et al., 2008] Hunter, D., Handcock, M., Butts, C., Goodreau, S., and Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3).

[McLaughlin et al., 2015] McLaughlin, K. R., Handcock, M. S., and Johnston, L. G. (2015). Inference for the visibility distribution for respondent-driven sampling. In *JSM Proceedings 2015, Statistical Computing Section.*

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