Causal Inference for Dyadic Outcomes in Social Network Analysis

A. N. D'Amour E. M. Airoldi

December 30, 2016

Contents

1 Introduction

In this paper, we consider the analysis of experiments that treat social network structure as an outcome. Letting V be a set of actors and Y be the set of outcomes associated with each pair of actors in $\binom{V}{2}$, the experimental units in these problems are pairs of actors in $\{i, j\} \in \binom{V}{2}$, which we index with ij. In a dataset of this type, the measured outcomes for each pair, Y_{ij}, summarize social activity between the actors in the pair; for example, Y_{ij} could represent the number of emails sent between individuals i and j. Each summary Y_{ij} may live in an arbitrary probability space, for example, we may consider binary interaction networks that represent the presence or absence of an interactions, count-valued interaction networks the record the number of observed interactions, or point-process valued interaction networks that record the timestamps of repeated interactions. In this paper, we call this data structure a random graph, although it is technically a generalization of the standard notion of a random graph, and we call a particular instantiation of a random graph Y defined with respect to a known actor-set V a network sample.

Social network data have unique properties that make causal inference difficult. A critical property of social processes is that they are sparse; put simply, in random graphs Y generated by social processes, the proportion of non-zero outcomes $\sum_{ij} \mathbf{1}_{Y_{ij}>0} \Big/ \binom{|V|}{2}$ among a set of actors V tends to zero as the size of the actor-set V becomes large (Orbanz & Roy, 2013; D'Amour & Airoldi, 2016). D'Amour & Airoldi (2016) showed that sparsity makes it difficult to define estimands that generalize between network samples of different size. This can be particularly problematic in contexts where social scientists hope that a causal effect estimated in a particular experiment can be used to draw conclusions about the more general social process "in the wild".

D'Amour & Airoldi (2016) proposed a model of social network processes called the Conditionally Independent Relationship, or CIR, model, which can be used to define estimands that are useful for generalization in the predictive or associative context, provided that the true generating process admits a particular factorization. This model represents social network generation as a two-stage process, both of which are defined on the dyads $ij \in \binom{V}{2}$: first, a process that generates an unobserved binary relationship graph R that defines the set of actor-dyads ij in the sample that have the potential to generate social activity; and second, a conditional process that generates observed dyad-wise outcomes Y_{ij} independently of each other given the relationship graph R. For social processes that can be represented in this way, so that the conditional distribution $\mathbb{P}_0(Y \mid R)$ factorizes by dyad ij, this conditional distribution has estimable summaries that can be used to make generalizations across networks of different size. D'Amour & Airoldi (2016) also propose a zero-truncated estimation procedure for estimating the conditional distribution $\mathbb{P}_0(Y \mid R)$ that does not depend on specifying a model for the marginal relationship process, $\mathbb{P}_0(R)$.

The purpose of this paper is to extend the estimands and estimation procedures of the CIR model to causal analyses. This paper makes two contributions. First, we specify a class of local causal

estimands that are defined conditionally with respect to the relationship graph R. Second, we extend the zero-truncated estimation procedure proposed by D'Amour & Airoldi (2016) to the task of performing a Bayesian analysis of a randomized experiment. Conceptually, these contributions are similar to the framing and methodology of the principal stratification literature (Frangakis & Rubin, 2002), particularly in censoring-by-death settings (Rubin, 2000; Zhang & Rubin, 2003; Hayden et al., 2005); the main difference is that in this context the zero-truncated estimation procedure makes the treatment assignment mechanisms non-ignorable, even when it is fully randomized. We use a thorough simulation study to highlight several properties of the class of Bayesian treatment effect estimators we propose, which are in line with several known properties of model-based principal stratification estimation procedures. The methods we propose here are not limited to network data and can be applied to many principal stratification problems where the data are zero-truncated.

2 Experiments with Network Outcomes

2.1 Potential Outcomes Setup

Let V be a set of actors, and $\binom{V}{2}$ be the set of pairs of these actors, which we call dyads. In this experimental setup, we consider the dyads, indexed by $ij \in \binom{V}{2}$ to be experimental units. We consider binary treatment that can be applied to individual dyads; for example, in the case of an online social networking platform, the intervention may be a notifiction sent to the actors i and j alerting each of the other's recent activity. Let Z_{ij} be an indicator that represents the application of the treatment, with $Z_{ij} = 1$ indicating that the treatment was applied to dyad ij, and $Z_{ij} = 0$ indicating the treatment was not applied, or that ij was left in the control state.

We frame this experiment using potential outcomes. Let $Y_{ij}(0)$ and $Y_{ij}(1)$ represent the potential outcomes for dyad ij under control and treatment, respectively, and $Y(0)$ and $Y(1)$ represent the full sets of potential outcomes associated with control and treatment. For a random sample of actors V drawn from an actor population $\mathbb{V}$, we observe tuples (Y_{ij}^{obs}, Z_{ij}), for $ij \in \binom{V}{2}$ where $Y_{ij}^{obs} = Y_{ij}(Z_{ij})$. This construction is well-defined under the Single Unit Treatment Value Assumption, or SUTVA.

2.2 Superpopulation Estimands for Network Data

In general, causal effects are defined as measures of discrepancy between the set of potential outcomes $Y(0)$ and $Y(1)$. For example, the most common causal estimand is the average treatment effect, or ATE defined as the sample average of individual treatment effects, conditioning on the

set of units in the sample. In our setting, this has the form

$$\tau_{ATE} = \binom{|V|}{2}^{-1} \sum_{ij \in \binom{|V|}{2}} Y_{ij}(1) - Y_{ij}(0). \tag{1}$$

In cases where the entire population of interest is included in the experiment, for example, in retrospective policy evaluation studies, estimating τ_{ATE} is the ultimate goal. However, often causal effects are estimated for the purpose of understanding the superpopulation from which the experimental outcomes $Y(0), Y(1)$ were sampled. To state this superpopulation estimation goal in terms of observable quantities, the investigator wishes to estimate a causal effect whose numerical value would be the same under different experimental circumstances.

For example, it is common to specify the causal estimand as a population average treatment effect defined in terms of an expectation over the sets of experimental units that could have been included in the analysis. In our setting, one could define the population average treatment effect, or PATE as

$$\tau_{PATE} = E_{\mathbb{P}(V)} \left[\binom{|V|}{2}^{-1} \sum_{ij \in \binom{|V|}{2}} Y_{ij}(1) - Y_{ij}(0) \right], \tag{2}$$

where $E_{\mathbb{P}(V)}$ is an expectation taken over the sampling procedure that yielded the actor set V. By construction, τ_{PATE} is the same for any actor set V drawn according to the sampling distribution $\mathbb{P}(V)$.

A major caveat in the construction of τ_{PATE} is that it depends on the experimental design, specifically the actor sampling policy $\mathbb{P}(V)$. In most applications, it is desirable to define a causal effect that does not depend on the experimental design, which is ancillary to the underlying process of interest. In classical treatments of population effect estimation, this problem is largely resolved by assuming that the experimental units are sampled completely at random from a superpopulation, so that the sampling distribution of each sampled unit ij is identical. In these cases, τ_{PATE} has a particularly simple representation

$$\tau_{PATE} = E_{\mathbb{P}_0}[Y_0(1) - Y_0(0)], \tag{3}$$

where $Y_0(1)$ and $Y_0(0)$ are random variables distributed according to the common marginal distributions of potential outcomes under treatment and control, $\mathbb{P}_0(Y_0(1))$ and $\mathbb{P}_0(Y_0(0))$, for a single sampled unit. This representation makes clear that the estimand is invariant to most details of the design $\mathbb{P}(V)$, for example, the size of the samples that have high probability of being drawn under $\mathbb{P}(V)$ and, in fact, τ_{PATE} is equal to the expected individual treatment across all units in the superpopulation.

2.3 Network Sparsity

Unfortunately, in the case of network data, the dependence of τ_{PATE} on the sampling design $\mathbb{P}(V)$ cannot be resolved so easily. For several reasons, even if the superpopulation of actors is assumed to be identical, the experimental units in this case $ij \in \binom{V}{2}$ cannot be treated as though they were sampled completely at random from a superpopulation. Most obviously, because sampling is performed by choosing an actor set V, the dyads ij included in the sample are drawn in clusters; however, even when this cluster sampling is accounted for, for example, by conditioning on actor-specific covariates, deeper issues remain. We consider one such issue here, which is the sparsity of social network data, discussed in detail in D'Amour & Airoldi (2016). That paper defines sparsity as a property of social network processes where, as the actor set V at which the process is observed becomes large, regardless of how V is selected, the expected fraction of nonzero dyads in the sample, $E\left[\sum_{ij} \mathbf{1}_{Y_{ij}>0} \Big/ \binom{|V|}{2}\right]$ converges to zero in the limit. This corresponds to the sparsity phenomenon ubiquitously observed in real social networks.

Definition 1 (Sparse Graph Process) *Let $\mathbb{V}$ be a population of actors, and $\binom{\mathbb{V}}{n}$ be the set of actor-sets dran from $\mathbb{V}$ of size n. Let $D(Y) = \sum_{ij} \mathbf{1}_{Y_{ij}>0} \Big/ \binom{|V|}{2}$. Let $D_n := \max_{V \in \binom{\mathbb{V}}{n}} E(D(Y_V))$, where Y_V is the random graph associated with actor-set V. We say the random graph process $Y_{\mathbb{V}}$ is* sparse *if and only if $\lim D_n = 0$.*

Sparsity describes an inhomogeneity in sample size that implies, for example, that additive estimands, such as τ_{ATE} as defined in Equation 1 have magnitudes that depend strongly on sample size. This, in turn, implies that population effects, such as τ_{PATE} as defined in Equation 2 have strong dependence on the sampling design $\mathbb{P}(V)$. For example, if $\mathbb{P}(V)$ puts high probability on selecting actor sets V that are very large, τ_{PATE} would have a smaller upper bound than if $\mathbb{P}(V)$ only put positive probability on small actor sets. This problem is compounded by the fact that the choice of actors V is usually biased toward actors that are *a priori* assumed to be densely connected. Thus, investigators usually design $\mathbb{P}(V)$ to put the most mass on actor-sets whose network density is likely to be close to the upper bound for networks of size $|V|$.

D'Amour & Airoldi (2016) suggest that, under the assumption that network data are distributed according to a generating process that factorizes in a special way, we can obtain sparsity-invariant summaries of a social network process. For processes of this type, which D'Amour & Airoldi (2016) call Conditionally Independent Relationship, or CIR processes, there is a subset of the units $ij \in \binom{V}{2}$ whose outcomes can be treated as though they were drawn completely at random from a larger population, regardless of the sparsity of the social process; however, the identity of this subset of dyads is not completely observable. In the next section, we define a set of estimands that are invariant to sampling policy in terms of this special set of dyads, but first, we give a brief overview of the conditionally independent relationship model.

2.4 The Conditionally Independent Relationship (CIR) model

In this section, provide an overview of several ideas and results from D'Amour & Airoldi (2016) that we will extend in the next section. In the predictive context, D'Amour & Airoldi (2016) propose a Conditionally Independent Relationship model for social networks. The CIR model is framed in terms of a latent binary relationship graph R that underlies the observed network sample Y. For a set of actors V, in addition to the set of observable pairwise outcomes Y, let R be a set of relationship indicators, one for each $ij \in \binom{V}{2}$, which represent prerequisites for social activity. For each $ij \in \binom{V}{2}$, let $R_{ij} = 1$ if there is a relationship between actors i and j, and $R_{ij} = 0$ otherwise. Relationships R_{ij} are related to the observed outcomes Y_{ij} in the following way: actors i and j must have a relationship, so that $R_{ij} = 1$, to generate nonzero social activity, so that $Y_{ij} > 0$; furthermore, conditional on the full set of relationships R, the observable outcomes Y are conditionally independent. Figure 1 provides an illustration. These properties are summarized in the following definition.

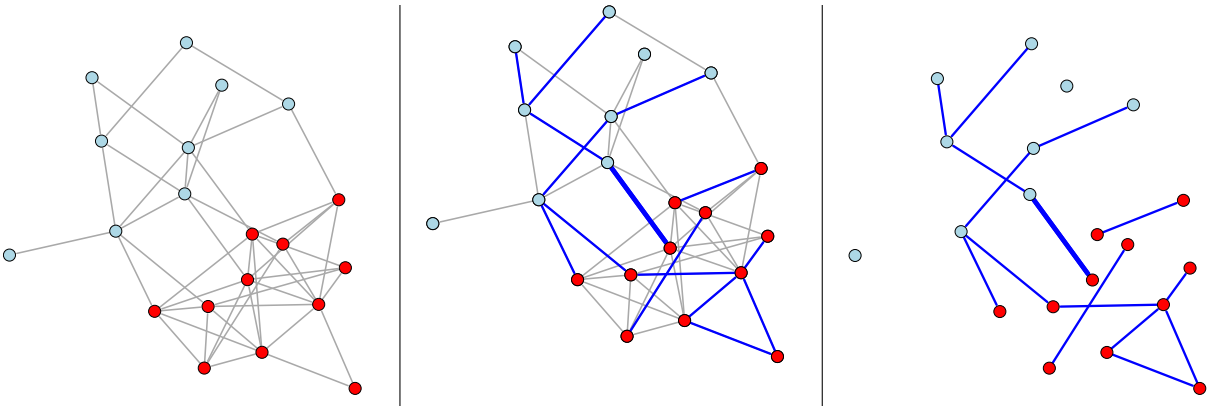
Definition 2 *Let Y be a random graph on an actor set V. We say Y is generated by a Conditionally Indepenent Relationship, or CIR, process if and only if the distribution of Y can be written*

$$\mathbb{P}_0(Y) = \sum_{R \in \mathcal{G}} \left[\mathbb{P}_0(R) \prod_{ij} \mathbf{1}_{\{Y_{ij}=0\}}^{1-R_{ij}} \mathbb{P}_0(Y_{ij} \mid R_{ij} = 1)^{R_{ij}} \right]. \tag{4}$$

where $\mathcal{G}$ is the set of all undirected binary graphs on V.

Remark 1 *The summation in Equation 4 encodes the fact the full set of relationships R is not completely observed. In particular, if the pairwise outcomes Y_{ij} have a discrete component such that the conditional distribution $\mathbb{P}(Y_{ij} \mid R_{ij} = 1)$ assigns some positive probability to $Y_{ij} = 0$, the relationships R are only partially revealed by the observed outcome sample Y: when $Y_{ij} \neq 0$, we know that $R_{ij} = 1$, but when $Y_{ij} = 0$, it is ambiguous whether the actor-pair ij has a relationship such that $R_{ij} = 1$ but failed to interact during the observation period, or whether ij have no relationship at all such that $R_{ij} = 0$. In fact, by definition, when a large network sample is generated by a sparse process, most dyads ij fall into this ambiguous category.*

Based on this representation, D'Amour & Airoldi (2016) have two major results in the context of predictive inference. Together, these results allow for the definition and estimation of sparsity-invariant estimands from sparse network data. First, D'Amour & Airoldi (2016) showed that the sparsity of CIR processes is completely described by the latent relationship process R, implying that parameters summarizing the conditional distribution $\mathbb{P}_0(Y \mid R)$ are invariant to the sparsity of the generating process. Thus, the CIR representation motivates estimands that condition on the relationship graph R.



(a) Unobserved Relationshps

(b) Relationships and Outcomes

(c) Observable Outcomes

Figure 1: Social activity generating process in the Conditionally Independnet Relationship framework. Each panel shows the same set of actors V, represented as dots. Figure 1a shows the unobserved relationship graph, where actors share a tie if they have a relationship, which is a prerequisite for generating observable social activity. Figure 1b shows observable social activity, represented by blue ties, superimposed on the relationship graph; note here that observable activity can only occur between actors that share a grey relationship tie, but that not all relationship ties generate observable interactions. Figure 1c shows the social activity graph that the investigator is able to observe, with all unobservable grey relationship ties removed.

Secondly, D'Amour & Airoldi (2016) describe an estimation procedure for estimands derived from the distribution $\mathbb{P}_0(Y \mid R)$ that does not depend on the nuisance distribution $\mathbb{P}_0(R)$. For a parametric model of the conditional distribution $\mathbb{P}_0(Y \mid R)$, $\{\mathbb{P}_\mu(Y \mid R) : \mu \in M\}$, defined so that for some $\mu_0 \in M$, $\mathbb{P}_0(Y \mid R) = \mathbb{P}_\mu(Y \mid R)$ for all Y and R, D'Amour & Airoldi (2016) proposed a truncated likelihood for estimating μ_0 when Y follows a CIR model. Letting $\mathcal{A} = \{ij : Y_{ij} > 0\}$ and $Y^{\mathcal{A}} = \{Y_{ij} : ij \in \mathcal{A}\}$,

$$\mathbb{P}_\mu(Y^{\mathcal{A}}) = \prod_{\{ij : Y_{ij} > 0\}} \mathbb{P}_\mu(Y_{ij} \mid Y_{ij} > 0) \tag{5}$$

The truncated likelihood is the likelihood for μ under an alternative observation mechanism where the observed dyadic outcomes Y_{ij} are zero-truncated, so that only $Y^{\mathcal{A}}$ is observed. The key properties of this likelihood are that no factors include dependence on the unobservable relationship process R and that the only nontrivial factors correspond to dyadic outcomes Y_{ij} that are nonzero. Estimators derived from Equation 5 are thus robust to the distribution of R. As a bonus, these estimators are computationally efficient because they only require iteration over the active set of dyads $\mathcal{A}$.

3 The Causal CIR model

To extend the CIR model to the causal context, we assume that the sampling distribution of the potential outcomes $Y(0)$ and $Y(1)$ can be represented as a CIR processes, where the randomness is induced by the sampling mechanism $\mathbb{P}(V)$. We also define intermediate outcomes $R(0)$ and $R(1)$, representing the unobservable relationship indicators for each dyad ij under control and treatment, respectively, and write the relationship indicator for unit ij under treatment z as $R_{ij}(z)$.

In this representation, each dyad ij can be assigned to a straum based on its relationship status under treatment and control, $(R_{ij}(0), R_{ij}(1))$, which we call its *relationship profile*. We use superscripts to indicate the relationship profile with which an object is associated. We will use both set and indicator notation to indicate the membership of units in relationship strata. In set notation, let $\mathcal{U}^{r_0 r_1}$ denote the set of indices such that $R_{ij} = (r_0, r_1)$ for all $ij \in \mathcal{U}^{r_0 r_1}$. For example, under this notation $\mathcal{U}^{11}$ is the set of indices referencing all units ij such that $R_{ij}(0) = R_{ij}(1) = 1$. In indicator notation, let $U_{ij}^{r_0 r_1}$ be an indicator associated with each ij, which takes the value 1 if and only if $ij \in \mathcal{U}^{r_0 r_1}$. Additionally, define $N^{r_0 r_1}$ as the total number of units indexed by $\mathcal{U}^{r_0 r_1}$ such that $N^{r_0 r_1} = |\mathcal{U}^{r_0 r_1}| = \sum_{ij} U^{r_0 r_1}$. Finally, let $n^{r_0 r_1}(z)$ denote the number of units in $\mathcal{U}^{r_0 r_1}$ assigned to treatment z, or $\sum_{\{ij: Z_{ij} = z\}} U^{r_0 r_1}$. For an illustration of this notation, see Figure 2.

3.1 Local Estimands

For the remainder of the paper we focus on estimating the causal effect of a treatment within the stratum where $(R_{ij}(0), R_{ij}(1)) = (1, 1)$, or the set of actor-pair units that would have a relationship under both treatment and control. Therefore, we define the average treatment effect τ as a function of average discrepancy between $Y_{ij}(0)$ and $Y_{ij}(1)$ among all units indexed by $\mathcal{U}^{11}$. Formally, for an arbitrary discrepancy measure D,

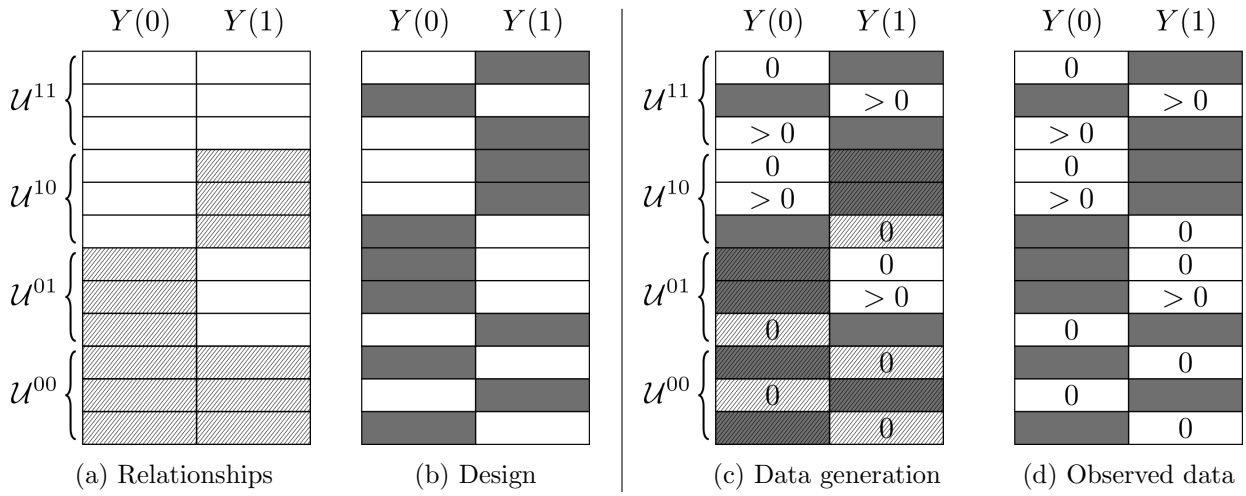
$$\tau_D^{11} = E_{\mathbb{P}(V)} \left[\frac{1}{N^{11}} \sum_{ij \in \mathcal{U}^{11}} D(Y_{ij}(0), Y_{ij}(1)) \right]. \tag{6}$$

This estimand is analogous to the survivor average causal effect defined in censoring-by-death applications of principal stratification (Rubin, 2000; Robins & Greenland, 2000; Zhang & Rubin, 2003).

Under the assumption that $Y(0)$ and $Y(1)$ are CIR processes, this estimand has a simple representation

$$\tau_D^{11} = E_{\mathbb{P}_0} \left[D(Y_0(1), Y_0(0)) \mid U_0^{11} = 1 \right], \tag{7}$$

where $Y_0(1)$, $Y_0(0)$ are random variables distributed according to the sampling distribution of



(a) Relationships (b) Design (c) Data generation (d) Observed data

Figure 2: Potential outcomes table with relationship profiles. Figures 2a and 2b show the underlying relationship structure and experimental design, respectively, while figures 2c and 2d show the complete and observed data, respectively. In Figure 2a, those potential outcomes that correspond to a "no relationship" state are line-shaded. In Figure 2b, the potential outcomes missing by design due to treatment assignment are colored gray. These two shadings are overlaid on the complete potential outcomes table on figure 2c. Note that line-shaded cells (with no relationship) only produce potential outcomes equal to 0, while cells with underlying relationships can produce both zero and nonzero data. Figure 2d highlights the two inferential challenges in this problem: imputing the data in the gray-shaded cells and distinguishing between zeros generated by line-shaded and non-line-shaded cells.

potential outcomes $\mathbb{P}_0(Y_0(1) \mid U_0^{11} = 1)$ and $\mathbb{P}_0(Y_0(0) \mid U_0^{11} = 1)$, respectively, for a single unit ij sampled completely at random from the subpopulation $\{ij : U_{ij}^{11} = 1\} \subset \binom{\mathbb{V}}{2}$.

For clarity of exposition, we focus on the average treatment effect for the remainder of the paper, and define the estimand to be

$$\tau_{PATE}^{11} = E_{\mathbb{P}_0}\left[Y_0(1) - Y_0(0) \mid U_0^{11} = 1\right]. \tag{8}$$

From this representation, the identification strategy is clear: from the observed data, we wish to estimate the conditional distributions of the potential outcomes $\mathbb{P}_0(Y_0(1) \mid U_0^{11} = 1)$ and $\mathbb{P}_0(Y_0(0) \mid U_0^{11} = 1)$, noting that the joint conditional distribution of $Y_0(1)$ and $Y_0(0)$ is not identifiable because for each unit only one potential outcome can be observed. We will proceed with the standard working assumption that $Y_0(1)$ and $Y_0(0)$ are independent; in the case of the average treatment effect, this assumption produces confidence intervals that are conservative (Neyman, 1923).

4 Estimation by Bayesian Inference

Estimating the potential outcome distributions $\mathbb{P}_0(Y_0(1) \mid U_0^{11} = 1)$ and $\mathbb{P}_0(Y_0(0) \mid U_0^{11} = 1)$ requires that we identify units in the relationship stratum $\mathcal{U}^{11}$. Unfortunately, in the best case, the relationship stratum for a particular unit ij is at best half-observed. For example, an active observed unit assigned to the control treatment, such that $Z_{ij} = 0$ and $Y_{ij}(0) > 0$, would be known to have a relationship under control with $R_{ij}(0) = 1$, but could belong to either the stratum $\mathcal{U}^{11}$ or $\mathcal{U}^{10}$, depending on the value of $R_{ij}(1)$, about which we have no direct information. This introduces a missing data problem, which we approach from a Bayesian perspective.

To simplify the probability modeling necessary for estimation, we employ the truncated data model from D'Amour & Airoldi (2016). The truncated data model restricts the data to units ij for which one of $Y_{ij}(0)$ or $Y_{ij}(1)$ was observed to be nonzero. This corresponds to a modification of the experiment with a truncated observation mechanism, wherein only units with nonzero outcomes in response to their assigned treatment appear in the sample. Formally, let A_{ij} be an indicator defined

$$A_{ij} = \begin{cases} 1 & \text{if } Y_{ij}^{obs} > 0 \\ 0 & \text{otherwise,} \end{cases}$$

and let $\mathcal{A} = \{ij : A_{ij} = 1\}$ be the set of active observed units under the truncated data model. Let $Y^{obs,\mathcal{A}}$ be the outcomes observed under the truncated data model, defined as $Y^{obs,\mathcal{A}} = \{Y_{ij}^{obs} : ij \in \mathcal{A}\}$, and likewise, let $Z^{obs,\mathcal{A}}$ be the treatment assignments of observed active units. Under this data model, only units in relationship strata are $11, 10, 01$ have positive probability of being included in $Y^{obs,\mathcal{A}}$, so the 00 stratum can be ignored.

Formally, let $\mathcal{P}_\mu$ be a model family indexed by $\mu \equiv \{\mu^{r_0 r_1}(z) : z \in \{0,1\}, r_0 r_1 \in \{11, 10, 01\}\}$

and composed of conditional potential outcomes distributions $\{\mathbb{P}_{\mu^{r_0r_1}(z)}(Y_0(z) \mid \mathcal{U}_0^{r_0r_1} = 1) : z \in \{0,1\}, r_0r_1 \in \{11,10,01\}\}$. We are interested in estimating $\mu_0^{11}(z)$ for $z \in \{0,1\}$ such that $\mathbb{P}_0(Y_0(1) \mid U_0^{11} = 1) = \mathbb{P}_{\mu^{11}(1)}(Y_0(1) \mid U_0^{11} = 1)$ and $\mathbb{P}_0(Y_0(0) \mid U_0^{11} = 1) = \mathbb{P}_{\mu^{11}(0)}(Y_0(0) \mid U_0^{11} = 1)$. For the remainder of the paper, we suppose that the model is well specified so that $\mu_0^{11}(z)$ exists for $z \in \{0,1\}$. We will derive a posterior distribution $\pi(\mu \mid Y^{obs,\mathcal{A}}, Z^{obs,\mathcal{A}})$.

Obtaining this posterior requires integrating over ambiguous stratum memberships, with corresponding nuisance parameters. Let $N^+ \equiv N^{11} + N^{10} + N^{01}$ be the number of units that have positive probability of being included under the truncated data model. For each of these strata, define $\zeta^{r_0r_1} \equiv \frac{N^{r_0r_1}}{N^+}$. In addition, let $\alpha(z)$ be the probability that any unit is assigned to treatment z; for binary treatment $\alpha(1) = 1 - \alpha(0)$. In the case of a randomized experiment, α is a known part of the experimental design.

Under the causal CIR model, the likelihood of the truncated data model has factors that correspond to finite mixtures with mixture weights $\{w(r_0r_1, z) : z \in \{0,1\}, r_0r_1 \in \{11,10,01\}\}$, which are functions of μ, ζ, and α. These mixtures encode the ambiguity of stratum membership for each observed active unit.

$$\mathbb{P}(Y^{obs,\mathcal{A}}, Z^{obs,\mathcal{A}} \mid \mu, \zeta; \alpha) = \prod_{ij \in \mathcal{A}} \left[\sum_{r_0r_1 \in \{11,10\}} w(r_0r_1, 0) \mathbb{P}_{\mu^{r_0r_1}(0)}(Y_0(0) = Y_{ij}(0) \mid Y_0(0) > 0) \right]^{(1-Z_{ij})} \left[\sum_{r_0r_1 \in \{11,01\}} w(r_0r_1, 1) \mathbb{P}_{\mu^{r_0r_1}(1)}(Y_0(1) = Y_{ij}(1) \mid Y_0(1) > 0) \right]^{Z_{ij}} \tag{9}$$

Let $p(\mu^{r_0r_1}(z)) = \mathbb{P}_{\mu^{r_0r_1}(z)}(Y_0(z) > 0)$ be the probability that the potential outcome corresponding to treatment z in stratum r_0r_1 is nonzero. For each stratum r_0r_1 and each treatment assignment z, the mixture weights satisfy

$$w(r_0r_1, z) = \frac{\alpha(z)\zeta^{r_0r_1} p(\mu^{r_0r_1}(z))}{\sum_{\substack{z \in \{0,1\}, \\ r_0r_1 \in \{11,10,01\}}} w(r_0r_1, z)} \tag{10}$$

Intuitively, these encode the probability that a randomly chosen unit ij from the active observed samples $\mathcal{A}$ would belong to stratum r_0r_1 and be observed under treatment assignment z.

Remark 2 *A key feature of this likelihood under the truncated data model is that the treatment assignment mechanism is not ignorable. This is because the probability of a unit ij being included in the active observed set $\mathcal{A}$ depends on the probability that the potential outcome under the assigned treatment is nonzero.*

5 Mixture Identification

The above modeling statements are generic, and will yield valid draws from a posterior predictive distribution of the potential outcomes table without further assumptions. However, this posterior distribution may have symmetries that lead to substantively different causal conclusions, for example, the sign of τ^{11}_{PATE}, that cannot be resolved with any amount of data. This is the non-identifiability problem treated extensively in the principal stratification and instrumental variables literatures.

Specifically, identification without additional restrictions requires identifying the dependence between the potential outcomes $Y(0)$ and $Y(1)$, which is unobservable. This results in an inability to match mixture components inferred under the treatment and control conditions to each other. Here, we consider two kinds of identifying assumptions about the composition of relationship profiles in the finite population that allow us to do away with this label-switching problem. We call these invariance and monotonicity assumptions.

5.1 Case 1: Invariance

Invariance posits that the relationship status for any dyad is the same under both treatment assignments, so that $R_{ij}(0) = R_{ij}(1)$ for all ij. This implies that all dyads belong to the strata $\mathcal{U}^{00}$ or $\mathcal{U}^{11}$, and implies $\zeta^{01} = \zeta^{10} = 0$. Thus, under the truncated data model, all observed active dyads must belong to stratum $\mathcal{U}^{11}$, and there is no ambiguity of stratum membership. Without the problem of labeling mixture components, μ^{11} is well-identified. The simplified likelihood for μ^{11} is:

$$\mathbb{P}(Y^{obs,\mathcal{A}}, Z^{obs,\mathcal{A}} \mid \mu, \zeta; \alpha) = \prod_{ij \in \mathcal{A}} \begin{array}{l} \mathbb{P}_{\mu^{11}(0)}(Y_0(0) = Y_{ij}(0) \mid Y_0(0) > 0)^{(1-Z_{ij})} \\ \mathbb{P}_{\mu^{11}(1)}(Y_0(1) = Y_{ij}(1) \mid Y_0(1) > 0)^{Z_{ij}} \end{array} \tag{11}$$

Estimation and inference under the monotonicity assumption is largely trivial, so we do not treat it any further in this paper.

5.2 Case 2: Monotonicity

The invariance assumption may be too strong in many circumstances. A weaker identifying assumption, often called monotonicity (Zhang & Rubin, 2003; Rubin, 2006), assumes that the relationship strata interact with the treatment in only one direction. Without loss of generality, consider the case where $R_{ij}(0) \geq R_{ij}(1)$, so that the treatment can only eliminate relationships. Under this assumption, there are three valid relationship strata: $\mathcal{U}^{00}$, $\mathcal{U}^{10}$, and $\mathcal{U}^{11}$. Equivalently, $\zeta^{01} = 0$.

Under this assumption, all active observed dyads under treatment, so that $ij \in \mathcal{A}$ and $Z_{ij} = 1$, are known to be in $\mathcal{U}^{11}$ as in the invariant case, but active observed dyads under control, so that $ij \in \mathcal{A}$ and $Z_{ij} = 0$ can belong to either $\mathcal{U}^{10}$ or $\mathcal{U}^{11}$, retaining the mixture component in Equation 9. The resulting likelihood is

$$\mathbb{P}(Y^{obs,\mathcal{A}}, Z^{obs,\mathcal{A}} \mid \mu, \zeta; \alpha) = \prod_{ij \in \mathcal{A}} \left[\sum_{r_0 r_1 \in \{11,10\}} w(r_0 r_1, 0) \mathbb{P}_{\mu^{r_0 r_1}(0)}(Y_0(0) = Y_{ij}(0) \mid Y_0(0) > 0) \right]^{(1-Z_{ij})} \left[w(11,1) \mathbb{P}_{\mu^{r_0 r_1}(1)}(Y_0(1) = Y_{ij}(1) \mid Y_0(1) > 0) \right]^{Z_{ij}}. \tag{12}$$

Under this likelihood, $\mu^{11}(1)$ is easily identified from the active observed outcomes under control. With this information, the mixture weights defined in Equation 10 identify $\mu^{11}(0)$ and $\mu^{10}(0)$.

This posterior is well-identified, although the likelihood remains multimodal in finite samples (Redner et al., 1984; Jasra et al., 2005; Frühwirth-Schnatter, 2006). However, as opposed to non-identifiability concerns, these pathologies diminish with sample size and can be overcome by well-designed posterior summaries (Stephens, 2000).

As demonstrated in Feller et al. (2016), the method of moments is a useful heuristic for understanding high-density regions in this multimodal posterior distribution of finite mixtures in principal stratification problems. For example, in the case where $\mu^{r_0 r_1}(z)$ is one-dimensional for each $r_0 r_1$ and z, we can identify the three-dimensional parameter of interest μ and the nuisance parameter ζ with the following four identities for observable moments from the truncated data model.

$$E\left[Z_0 \mid Y_0(Z_0) > 0\right] = \frac{w(11,1)}{w(11,1) + w(11,0) + w(10,0)} \tag{13}$$

$$E[Y_0(1) \mid Y_0(1) > 0] = \frac{E_{\mu^{11}(1)}[Y_0(1) \mid U_0^{11} = 1]}{p(\mu^{11}(1))} \tag{14}$$

$$E[Y_0(0) \mid Y_0(0) > 0] = \frac{w(11,0)}{w(11,0) + w(10,0)} E_{\mu^{11}(0)}[Y_0(0) \mid U_0^{11} = 1] + \frac{w(10,0)}{w(11,0) + w(10,0)} E_{\mu^{10}(0)}[Y_0(0) \mid U_0^{10} = 1] \tag{15}$$

$$\begin{aligned} \mathrm{Var}[Y_0(0) \mid Y_0(0) > 0] = {} & \frac{w(11,0)}{w(11,0) + w(10,0)} \mathrm{Var}_{\mu^{11}(0)}[Y_0(0) \mid U_0^{11} = 1] + \\ & \frac{w(10,0)}{w(11,0) + w(10,0)} \mathrm{Var}_{\mu^{10}(0)}[Y_0(0) \mid U_0^{10} = 1] + \\ & \frac{w(11,0) w(10,0)}{(w(10,0) + w(11,0))^2} \left(\frac{E_{\mu^{11}(0)}[Y_0(0) \mid U_0^{11} = 1]}{p(\mu^{11}(0))} - \frac{E_{\mu^{10}(0)}[Y_0(0) \mid U_0^{10} = 1]}{p(\mu^{10}(0))} \right)^2 \end{aligned} \tag{16}$$

We demonstrate the usefulness of these moment equations for posterior summarization in the results

section.

6 Inferential Procedure Under Monotonicity

6.1 Working Example: Negative Binomial Outcomes

We demonstrate our inferential procedure and the complications it is designed to handle in the context of a simulated example that obeys the monotonicity assumption. Suppose that we have a network sample in which we are able to apply treatment to particular dyads, and in which we may assume that standard SUTVA conditions hold when we define each dyad in the network as an experimental unit. Suppose that social activity in this network is recorded using count-valued random variables. For example, these could count the number of messages sent between actors i and j. Here, we assume that each unit with an underlying relationship, so that $R_{ij} = 1$, generates an independent Poisson-distributed outcome with rates varying across the dyads according to an exponential distribution; marginally these outcomes are independent and identically distributed negative binomial variates.

In this setup, the parameter vector is effectively four-dimensional: $\theta \equiv (\mu, \zeta)$, where $\mu \equiv (\mu^{10}(0), \mu^{11}(0), \mu^{11}(1))$ are parameters of conditional potential outcome distributions, and $\zeta \equiv (\zeta^{10}, \zeta^{11})$ is constrained to sum to one, and defines the proportions of units in the observable strata $\mathcal{U}^{10}$ and $\mathcal{U}^{11}$. From the likelihood Equation 12, it is clear that $\mu^{11}(1)$ is straightforwardly identified by the distribution of outcomes from active observed units assigned to treatment. On the other hand, the control outcome parameters $\mu^{11}(0)$, $\mu^{11}(1)$, and ζ are only identified through a finite mixture, which makes the posterior distribution more complex.

For both experiments, we set $N^{10} = 2000$, $N^{11} = 4000$, so that $\zeta^{10} = 1/3$, and set $\alpha(1) = 0.5$. For the first experiment, $\mu^{10}(0) = 6$, $\mu^{11}(0) = 12$, $\mu^{11}(1) = 8$. For the second experiment, we change $\mu^{11}(0) = 7$ so that the two potential outcome distributions under the control condition, which are confounded under the observation mechanism, are more difficult to tell apart.

6.2 Procedure

Here, we outline the steps we recommend for drawing approximately calibrated inferences about the estimand τ^{11}_{PATE} using Bayesian machinery. The procedure is not as straightforward as the Bayesian approach in more familiar settings because of the multimodality of the posterior distribution. We illustrate each step of our procedure on two example datasets drawn from the working example generating process described above.

1. **Obtain method of moments solution.** The moment solutions give a heuristic understanding of the posterior distribution. Figure 3 illustrates moment solutions derived from the example samples. We see immediately that data generated by both parameter settings admit three solutions to the moment equations, one of which, in each case, is close to the truth.

2. **Initialize sampling chains at each of the method of moments solutions.** Because the posterior distribution is multimodal, approximating the full posterior distribution can require combining chains initialized in different places of the parameter space. We apply a heuristic of initializing multiple chains at each of the moment solutions, giving each chain an opportunity to explore each of the three modes.

3. **Sample from $\pi(\theta \mid Y^{obs,\mathcal{A}}, Z^{obs,\mathcal{A}})$.** We obtain posterior samples for the parameters $\theta \equiv (\mu, \zeta)$ using Markov Chain Monte Carlo. Our sampler is implemented in Stan, the code for which is included in the appendix. Figure 4a and Figure 4c illustrate the marginal posterior distributions of the mixture-identified parameters $\mu^{11}(0)$ and $\mu^{10}(0)$, which have modes close to the moment solutions, as expected.

4. **Partition samples into disjoint posterior modes using k-means.** Stephens (2000) proposed summarizing such multimodal posteriors using an optimization-based clustering algorithm like k-means to compute centroids in the parameter space when the number of regions of posterior mass is known, and assigning each sample to one of these centroids. This is a generalization of the posterior mean, which is the soluton to this optimization using only one centroid. We set k to number of moment solutions obtained in the first step, and initialize the k-means optimization algorithm at the moment solutions.

   Each posterior mode corresponds to a plausible explanation of the observed data. As such, it is useful to summarize each mode separately. In particular, the number of posterior samples assigned to each mode approximates the posterior mass in the region of the mode can be used to determine the relative support that the data express for each mode. Figure 4a and Figure 4c have centroid locations and attributed samples overplotted for each mode.

5. **(Optional) Weight posterior samples to obtain accurate mode masses.** Constructing a sampler that samples from all posterior modes proportionally is non-trivial. The initialization scheme described above ensures that samples will explore the key regions of the parameter space with significant posterior mass, but if these regions are too disconnected, general-purpose MCMC methods can give incorrect estimates of the relative posterior mass in each of these regions, even if run for a large but finite number of iterations, and even if they appear to mix well within the region of high posterior mass near the initialization point. Here, we propose a simply implemented procedure that does not require a custom-designed sampling algorithm.

   We partition the parameter space $\Pi(\Theta) = \{\mathcal{S}_k : k \in K\}$. For this section only, we adopt some shorthand. Let $\pi(\theta)$ be the true posterior, $\tilde{\pi}(\theta)$ be the unnormalized posterior, and $\hat{\pi}(\theta)$

be the posterior approximation obtained from the sampler. We require a weighting function $w(\theta)$ such that for each $\mathcal{S}_k \in \Pi(\Theta)$

$$\int \mathbf{1}\{\theta \in \mathcal{S}_k\} w(\theta) d\hat{\pi}(\theta) \approx \int \mathbf{1}\{\theta \in \mathcal{S}_k\} d\pi(\theta).$$

The simplest weighting scheme assigns a constant weight w_k to samples within each partition $\mathcal{S}_k$, and satisfies

$$w_k \propto \frac{\mathbb{P}_{\pi(\theta)}(\theta \in \mathcal{S}_k)}{\mathbb{P}_{\hat{\pi}(\theta)}(\theta \in \mathcal{S}_k)}.$$

The denominator in this weight is the number of samples in the partition, $\#\{\theta^{(i)} : \theta^{(i)} \in \mathcal{S}_k\}$. The numerator can be approximated up to a constant by importance sampling

$$\mathbb{P}_{\pi(\theta)}(\theta \in \mathcal{S}_k) \tilde{\propto} \int \mathbf{1}\{\theta \in \mathcal{S}_k\} \frac{\tilde{\pi}(\theta)}{\hat{\pi}(\theta)} d\hat{\pi}(\theta).$$

In terms of posterior samples, we write the weights as

$$w_k \propto \frac{\frac{1}{M} \sum_{i=1}^{M} \mathbf{1}\{\theta^{(i)} \in \mathcal{S}_k\} \tilde{\pi}(\theta^{(i)})}{\#\{\theta^{(i)} : \theta^{(i)} \in \mathcal{S}_k\}}.$$

As $\hat{\pi}(\theta) \to \pi(\theta)$, the weights are approximately uniform. On the other hand, if $\hat{\pi}(\theta \mid \theta \in \mathcal{S}_k) \to \pi(\theta \mid \theta \in \mathcal{S}_k)$ for each $\mathcal{S}_k$, the reweighted samples define a measure that converges to $\pi(\theta)$.

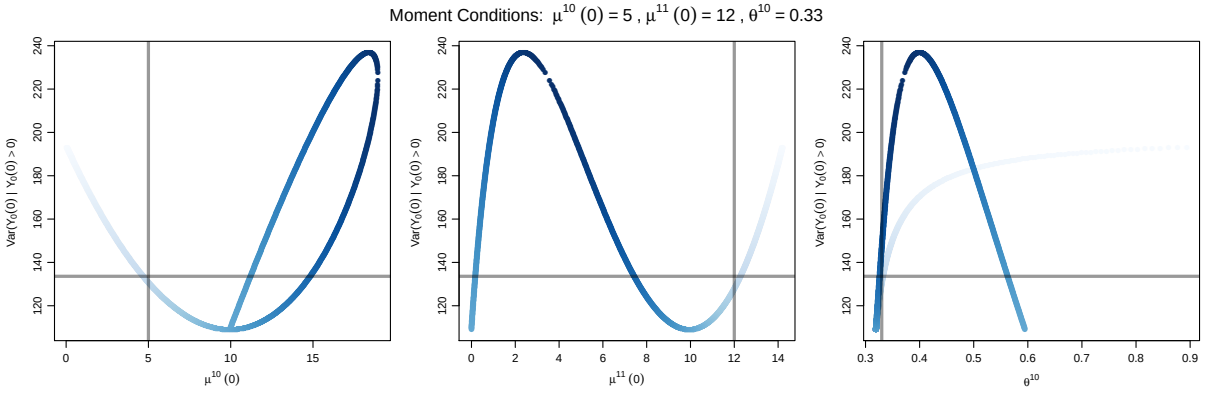
Figure 4a and Figure 4c have the weighted attributed samples overplotted for each mode.

6. **Summarize the (weighted) posterior distribution of τ^{11}_{PATE}.** For a point estimate, we estimate the median value of τ^{11}_{PATE} from samples assigned to the the k-means centroid with the highest weighted posterior mass. We call this the picked mode median. For approximately calibrated credible intervals, we report highest posterior density (HPD) interval of τ^{11}_{PATE}, constructed from the weighted sample, which is often disjoint. Figure 4b and Figure 4d show example posterior distributions of τ^{11}_{PATE}, with the true value of τ^{11}_{PATE}, the picked mode median estimate, and HPD interval overplotted.

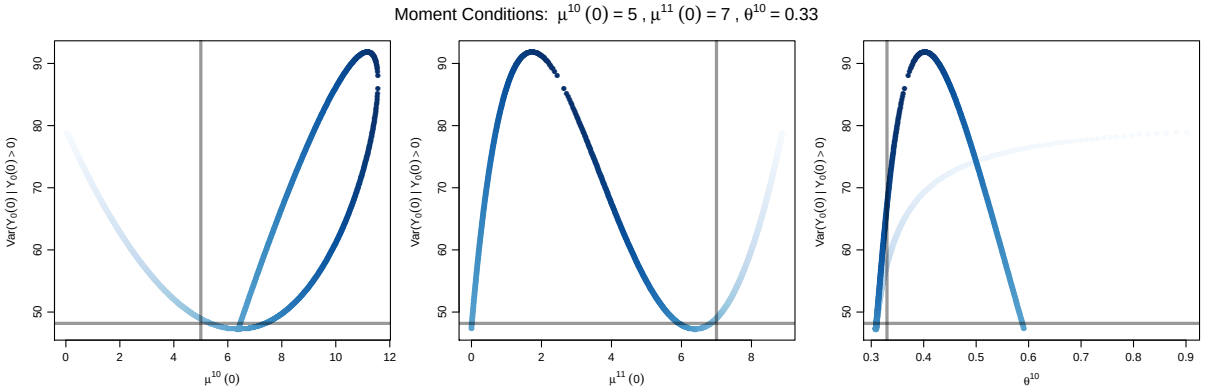
7 Simulation Study

7.1 Purpose and Design

To demonstrate the effectiveness of the above inferential framework in addition to the practical challenges that can arise in its practical implementation, we present a simulation study. The study

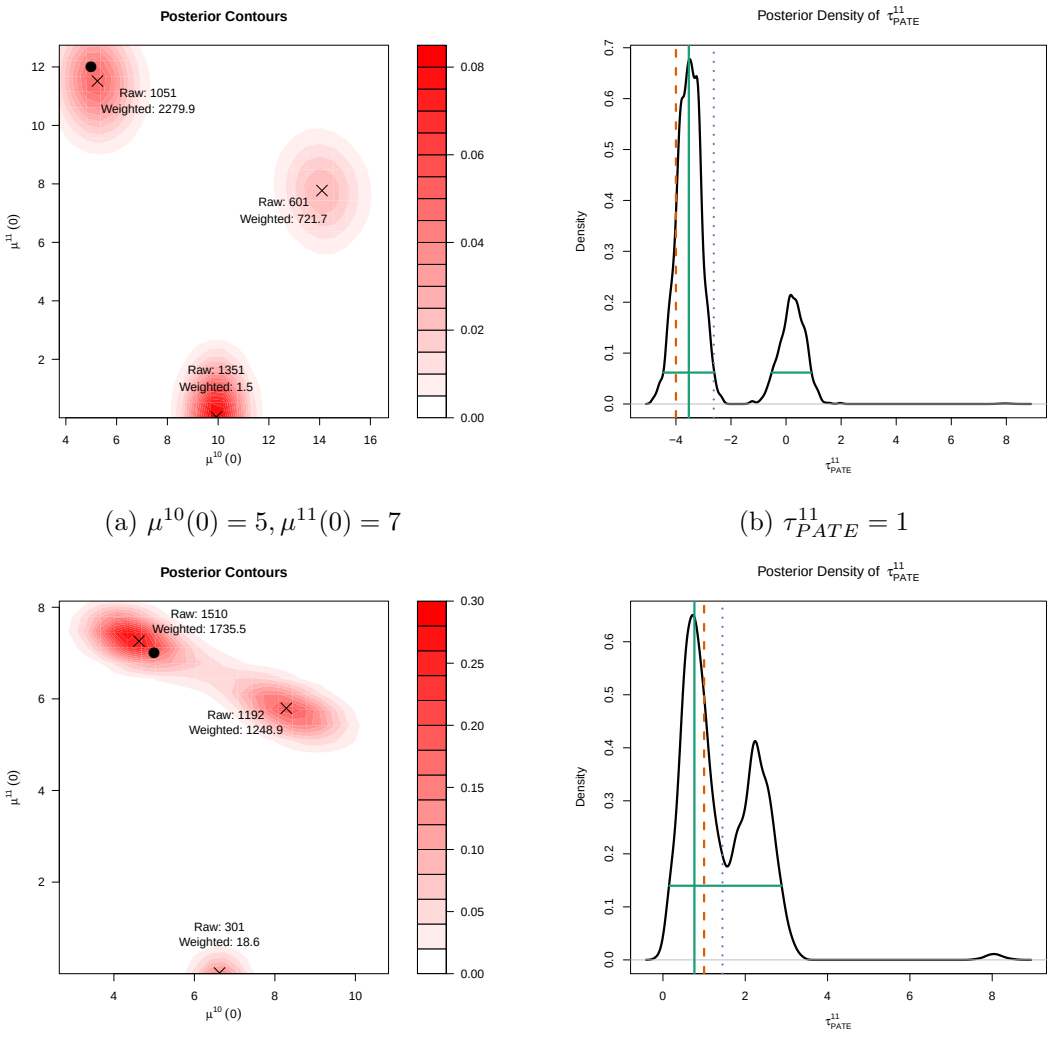


(a) $\mu^{10}(0) = 6$, $\mu^{11}(0) = 12$.



(b) $\mu^{10}(0) = 6$, $\mu^{11}(0) = 7$.

Figure 3: Parameter combinations satisfying moment conditions in method of moments analysis. The curve in each panel illustrates combinations of parameters that satisfy the first three moment conditions, with the y-axis representing the theoretical variance of observed nonzero control outcomes implied by feasible parameter combinations. Parameter combinations that occur together have the same color in all three panels. The horizontal line marks the sample variance among active observed control outcomes; intersections of this line with the feasible parameter curve in each panel identifies parameter values that satisfy the final moment condition. In both cases, there are three solutions to the full set of moment equations, which correspond to the modes of the posterior distribution. The vertical line marks the true value of each parameter, and occurs close to a moment solution.



(a) $\mu^{10}(0) = 5, \mu^{11}(0) = 7$

(b) $\tau^{11}_{PATE} = 1$

(c) $\mu^{10}(0) = 5, \mu^{11}(0) = 12$

(d) $\tau^{11}_{PATE} = -4$

Figure 4: Margins of posterior distribution derived from samples drawn from the working simulated example. (a) and (c) show contours of the joint posterior distribution of $\mu^{10}(0)$ and $\mu^{11}(0)$. The true parameter values are marked with a circle. Centroids computed using k-means are marked with 'x's. The number of posterior samples assigned to each centroid and the equivalent number of reweighted samples are overlaid. (b) and (d) show the marginal density of $\tau^{11}_{PATE} = \mu^{11}(1) - \mu^{11}(0)$. The horizontal lines show the highest posterior density (HPD) interval, which is disconnected in the case of (b). The solid vertical line is the picked mode median point estimate. The true value of τ^{11}_{PATE} (dashed vertical line) and the posterior mean (dotted vertical line) are provided for comparison.

is designed to confirm several two types of claims in this paper. The first claims are statistical, and assert that the posterior distribution derived from the truncated data model extracts useful information for recovering parameters of interest and the estimand. To this end, the simulation study shows that the picked mode median summary has desirable frequentist risk properties, and that the highest posterior density interval has desirable frequentist coverage properties for this data

model. The second claims are computational, and assert that the reweighted posterior sampling approach described in the previous section is an effective computational tool for approximating these posterior summaries. The simulation study shows that computational summaries we obtained agree qualitatively with theoretical predictions about the behavior of the posterior distribution.

The generative model in the simulation study is the same as in the example in the previous section: we assume that the monotonicity condition holds, so that the only observable strata are $\mathcal{U}^{11}$ and $\mathcal{U}^{10}$, and that within strata, potential outcomes $Y_{ij}(z)$conditional on a nonzero relationship $R_{ij}(z) = 1$ are distributed negative binomially. As in the previous example, we set the convolution parameter of the negative binomial in all cases to 1, so that distribution of the potential outcomes in each stratum in each condition is completely characterized by the mean $\mu^{r_0 r_1}(z)$. As in the previous section, the parameters of interest are effectively four-dimensional: $\theta = (\mu, \zeta)$, where the set of conditional means μ is 3-dimensional, and the stratum proportions ζ are effectively one-dimensional. Recall that under this parameterization, the estimand can be represented as $\tau^{11}_{PATE} = \mu^{11}(0) - \mu^{11}(1)$.

The study is designed as a factorial experiment. We simulate data from parameter settings that are defined in terms of two factors f_1 and f_2, which define, respectively, the distance between the mixture components in the observed control outcomes and the sample size. Specifically, we set the mean parameters so that $\mu^{10}(0) = 5$, $\mu^{11}_{f_1}(0) = \mu^{10}(0) + f_1\sqrt{V(\mu^{10}(0))}$, and $\mu^{11}(1) = 8$. For the sample size, we set $N_{f_2} = 3 \cdot 10^{f_2}$. For all values of f_1 and f_2, we set $\zeta^{10} = 0.33$, $\zeta^{11} = 0.67$, and $\alpha(1) = 0.5$. Finally, we let the levels of each factor to be $f_1 \in \{0.5, 1.125, 1.75, 2.375, 3\}$ and $f_2 \in \{2, 2.5, 3, 3.5, 4\}$.

These factors influence how well-separateed the two mixture components are in the observed distribution of control outcomes, so for analysis it is natural to project the properties of point estimates and interval estimates onto the number of standard errors separating the two components, defined as

$$\Delta(f_1, f_2) = \frac{\mu^{11}_{f_1}(0) - \mu^{10}(0)}{\sqrt{\frac{V(\mu^{10}(0)) + V(\mu^{11}_{f_1}(0))}{10^{f_2}}}}.$$

For each value of the factors, we simulated 100 data replicates, for a total of 2500 simulated datasets. For each dataset, we sampled from the posterior distribution using Stan, and summarzied the posterior samples using the reweighted scheme described in the previous section.

7.2 Point Estimate Results

Figure 5a and Figure 5b summarize the experimental results for point estimates for τ^{11}_{PATE} derived from weighted posterior samples. Recall that this point estimate is constructed by picking the mode in the posterior distribution with the largest mass, and taking the median of samples closest

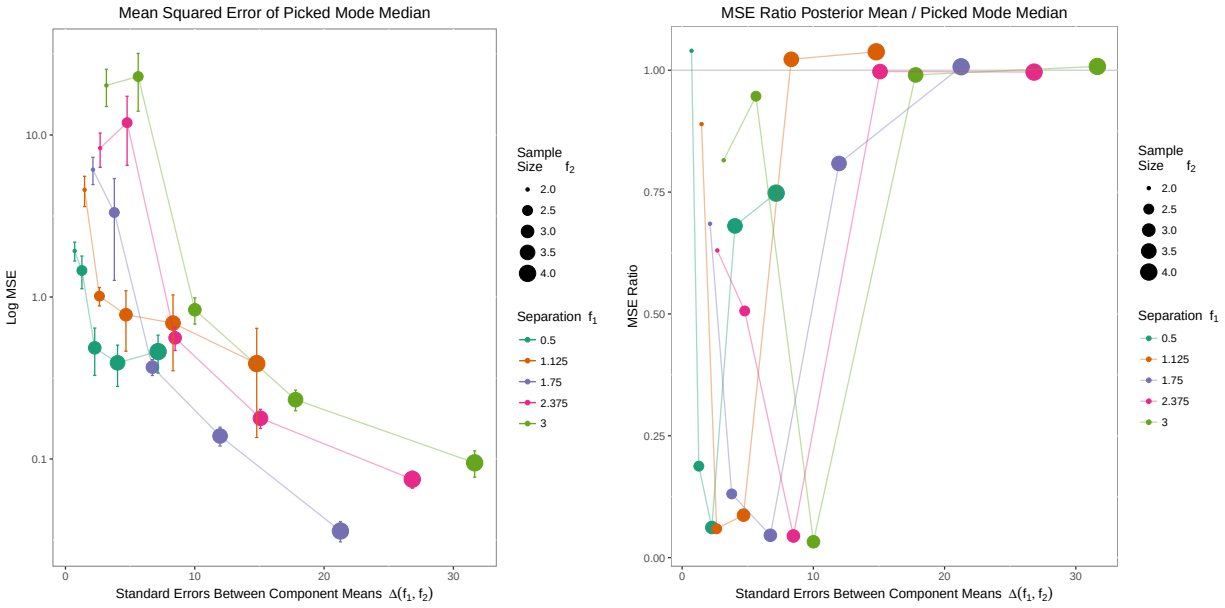
to that mode. Figure 5a shows that the mean squared error of this point estimation procedure is generally higher when the means of the mixture components of the control outcomes are far apart; within levels of the sample size factor f_2, the mean square error for generating processes with larger for larger values of the separation factor f_1. This is expected because, in finite samples, the posterior distribution may assign larger mass to spurious modes because of randomness in the data generation, as described in Feller et al. (2016). However, as expected, within levels of the separation factor f_1, as the sample size factor f_2 increases, the mode corresponding the the true parameters begins to dominate the posteiror distribution, and the mean square error falls.

It is notable, that when the mixture components are moderately separated, even though the mean squared error decreases with sample size, the variability of the squared error of any single point estimate is can increase with sample size. We attribute this to a phase change in the posterior distribution where the modes in the posterior become more distinct from one another, and errors in estimation move from being driven by the locations of modes in the posterior to the masses corresponding to each mode.

Figure 5b compares the mean square error of the picked mode median procedure to the mean square error of the more standard posterior mean procedure, and confirms the intuition that the posterior mean is not an appropriate point summary of the posterior when the posterior is known to be multimodal. The picked mode median procedure has the lowest relative error in moderately-sized samples, when the mode corresponding to the true parameter values begins to dominate, but the spurious posterior modes still have non-trivial mass. For small and large samples, the mean-squared error is comparable: in small samples, the picked mode median will have more variable squared error, but its mean squared error is comparable to the posterior mean; in large samples where the true posterior mode dominates, the posteiror mean and the median of the picked mode converge.

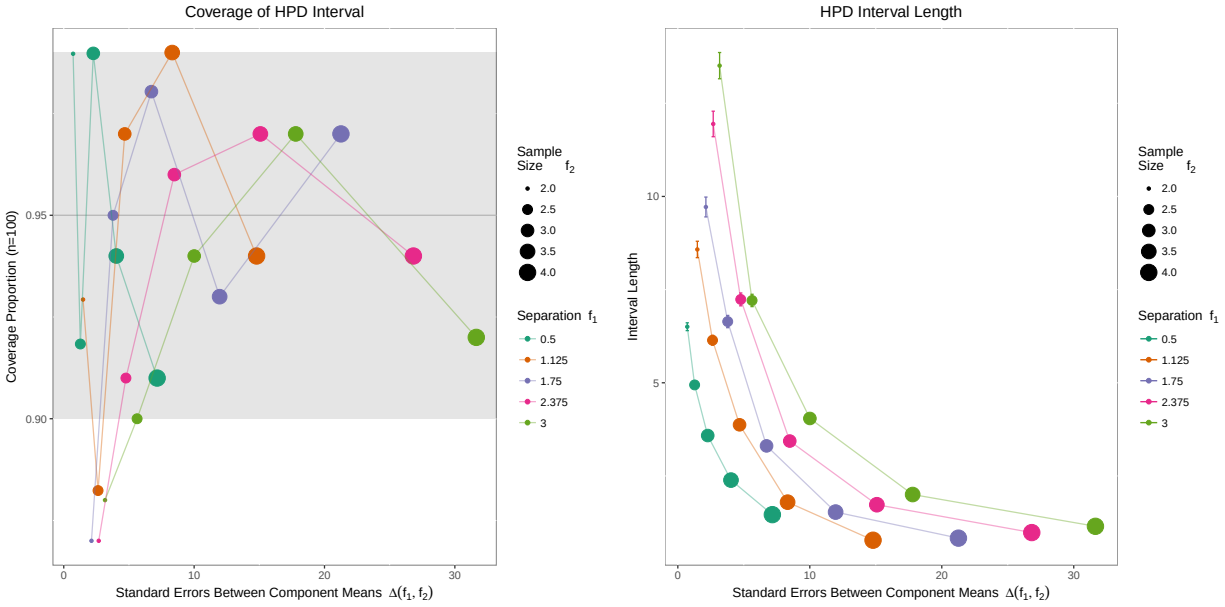
7.3 Credible Interval Results

Figure 6a and Figure 6b summarize the experimental results for highest posterior density credible regions for τ^{11}_{PATE} derived from weighted posterior samples. Figure 6a shows the rate at which the HPD interval covers the true value of τ^{11}_{PATE}, with points inside the grey band being indistinguishable from a procedure with nominal 95% coverage in an experiment with 100 replications. As expected, for large values of the sample size factor f_2, the HPD interval has nominal frequentist coverage. Within smaller levels of the separation factor f_1, there is a noticeable drop in coverage at small to moderate sample sizes. This again corresponds to a phase change as the modes of the posterior become more distinct with larger sample size, but the data do not contain enough information about the relative sizes of these modes. Figure 6b shows the mean total length of the HPD interval at each factor setting. The interval lengths behave as expected: the lengths increase in the separation factor f_1 because the varinace of the control outcomes is increasing in $\mu^{11}(0)$, while the lengths decreas in the sample size factor f_2.



(a) (b)

Figure 5: Point estimate results. (a) shows the mean squared error of the point estimate constructed by identifying the posterior mode with the highest assigned mass, then taking the median of samples in that region. (b) comparse the MSE of this estimator to the more standard posterior mean summary.



(a) (b)

Figure 6: Interval results. (a) shows the coverage rate of the HPD interval constructed from weighted posterior samples. (b) shows the average total length of those intervals, constructed by summing the length of disjoint regions, if applicable.

We note that these interval properties can be sensitive to the choice of bandwidth in the density estimate used to construct the HPD estimate. In these experiments, we used the `bw.SJ` bandwidth selection method in `R`, which implements the method of Sheather & Jones (1991).

8 Discussion

In this paper, we developed several ideas for defining and estimating causal superpopulation estimands for experiments where the outcomes summarize pairwise social interactions. There are several opportunities for extending these ideas further to answer questions from retrospective and observational studies.

For retrospective studies, such as policy evaluations where the treatment effect on a specific finite population is of interest, the methodology developed in this paper needs to be extended to estimate finite population estimands. In this case, the estimand would not be defined in terms of a sampling expectation, but would instead condition on the set of actors V included in the sample:

$$\tau^{11}_{ATE} = \frac{1}{N^{11}} \sum_{ij \in \mathcal{U}^{11}} D(Y_{ij}(0), Y_{ij}(1)). \tag{17}$$

The Bayesian framework in this paper could be extended for this estimand. The primary difference would be that posterior predictive samples of τ^{11}_{PATE} would not be a simple function of $\mu^{r_0 r_1}(z)$, but would also include variation from the posterior predictive distribution of N^{11}, the number of units, both observed and unobserved, in stratum $\mathcal{U}^{11}$. Under the truncated data model presented in this paper, this posterior predictive distribution is easily shown to be negative binomial. This finite sample estimand is closely related to the census undercount problem treated in Meng & Zaslavsky (2002), and the truncated data model presented here can be derived by marginalizing over the Single Observation Unbiased Prior presented in that paper.

For observational studies, the design parameter α would need to be estimated to apply the machinery developed in this paper. Propensity score (Rosenbaum & Rubin, 1983) or instrumental variable methods (Imbens, 2014) could be used for this purpose.

Finally, there is a connection between the approach in this paper and the curse of dimensionality appropriate asymptotics argument presented in Robins & Ritov (1997). In both cases, the investigator is interested in a relatively simple estimand, but a complex nuisance process stands in the way; in both cases, a pure likelihood-based approach would have the investigator model and marginalize this process at the cost of introducing major sensitivity to misspecification; and in both cases, the proposed solution is to use a simple alternative estimation procedure that incorporates ignorable design information, namely, the treatment assignment probabilities. In this paper, we justified our

final step as the likelihood-based approach derived from an alternative truncated data model. It can be shown that the family of Horvitz-Thompson estimators advocated in Robins & Ritov (1997) have a similar interpretation as a likelihood-based approach under an alternative data model. This connection hints at the possibility of a more general principle for estimator construction in causal inference problems with complex or high-dimensional nuisance parameters.

References

D'Amour, A. & Airoldi, E. (2016). Misspecification, sparsity, and superpopulation inference for sparse social networks. Ongoing work for publication and inclusion in dissertation.

Feller, A., Greif, E., Miratrix, L. & Pillai, N. (2016). Principal stratification in the Twilight Zone: Weakly separated components in finite mixture models .

Frangakis, C. E. & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer New York.

Hayden, D., Pauler, D. K. & Schoenfeld, D. (2005). An estimator for treatment comparisons among survivors in randomized trials. *Biometrics* **61**, 305–310.

Imbens, G. W. (2014). Instrumental Variables: An Econometrician's Perspective. *Statistical Science* **29**, 323–358.

Jasra, A., Holmes, C. C. & Stephens, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science* **20**, 50–67.

Meng, X. L. & Zaslavsky, A. M. (2002). Single observation unbiased priors. *Annals of Statistics* **30**, 1345–1375.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: principles (in Polish with German summary). *Roczniki Nauk Rolniczch* **10**, 21–51.

Orbanz, P. & Roy, D. M. (2013). Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *arXiv* **37**, 1–25.

Redner, R. A., Walker, H. F., Mathematics, A. & Review, S. (1984). Mixture Densities, Maximum Likelihood and the Em Algorithm. *SIAM Review* **26**, 195–239.

Robins, J. M. & Greenland, S. (2000). Causal Inference Without Counterfactuals: Comment. *Journal of the American Statistical Association* **95**, 431.

Robins, J. M. & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in medicine* **16**, 285–319.

Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* , 41–55.

Rubin, D. B. (2000). Causal Inference Without Counterfactuals: Comment. *Journal of the American Statistical Association* **95**, 435.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with censoring due to death. *Statistical Science* **21**, 299–309.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, B Methodology* **62**, 795–809.

Zhang, J. L. & Rubin, D. B. (2003). Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death". *Journal of Educational and Behavioral Statistics* **28**, 353–368.