Estimating Average Causal Effects Under Interference Between Units

Peter M. Aronow and Cyrus Samii*

August 14, 2015

*Peter M. Aronow is Assistant Professor, Departments of Political Science and Biostatistics, Yale University, 77 Prospect St., New Haven, CT 06520 (Email: peter.aronow@yale.edu). Cyrus Samii is Assistant Professor, Department of Politics, New York University, 19 West 4th St., New York, NY 10012 (Email: cds2083@nyu.edu).

Estimating Average Causal Effects Under Interference Between Units

Abstract

This paper presents a randomization-based framework for estimating causal effects under interference between units. The framework integrates three components: (i) an experimental design that defines the probability distribution of treatment assignments, (ii) a mapping that relates experimental treatment assignments to exposures received by units in the experiment, and (iii) estimands that make use of the experiment to answer questions of substantive interest. Using this framework, we develop the case of estimating average unit-level causal effects from a randomized experiment with interference of arbitrary but known form. The resulting estimators are based on inverse probability weighting. We provide randomization-based variance estimators that account for the complex clustering that can occur when interference is present. We also establish consistency and asymptotic normality under local dependence assumptions. We discuss refinements including covariate-adjusted effect estimators and ratio estimation. We illustrate and assess empirical performance with a naturalistic simulation using network data from American high schools.

Keywords: causal inference; SUTVA; randomization inference; potential outcomes.

1 Introduction

Experimental and observational studies often involve treatments with effects that "interfere" (Cox, 1958) across units through spillover or other forms of dependency. Such interference is sometimes considered a nuisance, and researchers may strive to design studies that isolate units as much as possible from interference. However, such designs are not always possible. Furthermore, researchers may be interested in estimation of the spillover effects themselves, as these effects may be of substantive importance. Treatments may be applied to individuals in a social network, and we may wish to study how effects transmit to peers in the network. An urban renewal program applied to one town may divert capital from other towns, in which case the overall effect of the program may be ambiguous. Treatment effects may carry over from one time period to another, and units have some chance of receiving treatment at any one of a set of points in time. In these cases, we need a method to estimate effects of both direct and *indirect* exposure to a treatment.

This paper presents a general, randomization-based framework for estimating causal effects under these and other forms of interference. Interference represents a departure from the traditional causal inference scenario wherein units are assigned directly to treatment or control, and the potential outcomes that would be observed for a unit in either the treatment or control condition are fixed (Cole and Frangakis, 2009) and do not depend on the overall set of treatment assignments. The latter condition is what Rubin (1990) refers to as the "stable unit treatment value assumption" (SUTVA). In the examples above, the traditional scenario is clearly an inadequate characterization, as SUTVA would be violated. A more sophisticated characterization of treatment exposure and associated potential outcomes must be specified.

Our estimation framework consists of three components: (i) the experimental (or quasi-experimental) "design," which characterizes precisely the probability distribution of treat-

ments *assigned*; (ii) an "exposure mapping," which relates treatments assigned to exposures *received*; and (iii) a set of causal estimands selected to make use of the experiment to answer questions of substantive interest. For the case of a randomized experiment under arbitrary but known forms of interference, we provide unbiased estimators of average unit-level causal effects induced by treatment exposure. We also provide estimators for the randomization variance of the estimated average causal effects. These variance estimators are assured of being conservative (that is, nonnegatively biased). We establish conditions for consistency and large-N confidence intervals based on a normal approximation. We propose ratio-estimator-based and covariate-adjusted refinements for increased efficiency. Finally, we sketch how one could apply this framework to more irregular forms of interference, alternative estimands, observational data, and situations where there is uncertainty about the form of interference.

2 Related Literature

Our framework extends from the foundational work of Hudgens and Halloran (2008), who study two-stage, hierarchical randomized trials in which some groups are randomly assigned to host treatments; treatments are then assigned at random to units within the selected groups, and interference is presumed to operate only within groups. Hudgens and Halloran provide randomization-based estimators for group-average causal effects, conditional on assignment strategies that determine the density of treatment within groups. Tchetgen-Tchetgen and VanderWeele (2010) extend Hudgens and Halloran's results, providing conservative variance estimators, a framework for finite sample inference with binary outcomes, and extensions to observational studies. Related to these contributions is work by Rosenbaum (2007), which provides methods for inference with exact tests under partial interference. Under hierarchical treatment assignment and partial interference, esti-

mation and inference can proceed assuming independence across groups. In some settings, however, the hierarchical structuring may not be valid, as with experiments carried out over networks of actors that share links as a result of a complex, endogenous process.

A key contribution of this paper is to go beyond the setting of hierarchical experiments with partial interference, and to generalize estimation and inference theory to settings that exhibit arbitrary forms of interference and treatment assignment dependencies. In addition, our framework allows the analyst to work with different estimands, including both types of group-average causal effects defined by the authors above as well as average unit-level causal effects. Average unit-level causal effects are often the estimand of primary interest, as is the case, for example, when exploring unit-level characteristics that moderate the magnitude of treatment effects.

3 Treatment Assignment and Exposure Mappings

In this section, we define the first two components of our analytical framework: the experiment design and exposure mapping. We focus on the case of a randomized experiment with an arbitrary but known exposure mapping. The first step is to distinguish between (i) treatment assignments over the set of experimental units and (ii) each unit's treatment exposure under a given assignment. Treatment assignments can be manipulated arbitrarily with the experimental design. However, treatment exposures may be constrained on the basis of the varying potential for interference of different experimental units. For example, interference or spillover effects may spread over a spatial gradient. If so, different treatment assignments may result in different patterns of interference depending on where treatments are applied on the spatial plane.

Formally, suppose we have a finite population U of units indexed by $i = 1, ..., N$ on which a randomized experiment is performed. Define a treatment assignment vector, $\mathbf{z} =$

$(z_1, ..., z_N)'$, where $z_i \in \{1, .., M\}$ specifies which of M possible treatment values that unit i receives. An *experimental design* contains a plan for randomly selecting a particular value of $\mathbf{z}$ from the M^N different possibilities with predetermined probability $p_{\mathbf{z}}$. Restricting our attention only to treatment assignments that can be generated by a given experimental design, define $\Omega = \{\mathbf{z} : p_{\mathbf{z}} > 0\}$, so that $\mathbf{Z} = (Z_1, ..., Z_N)'$ is a random vector with support Ω and $\Pr(\mathbf{Z} = \mathbf{z}) = p_{\mathbf{z}}$.

We define an *exposure mapping* as a function that maps an assignment vector and unit-specific traits to an exposure value: $f : \Omega \times \Theta \to \Delta$, where $\theta_i \in \Theta$ quantifies relevant traits of unit i. The exposure mapping construction is functionally equivalent to the "effective treatments" function used by Manski (2013), though we find it helpful to denote separately the unit-specific attributes, θ_i, that feed into the exposure mapping, $f(\cdot)$. The codomain Δ contains all of the possible treatment exposures that might be induced in the experiment. The contents of Δ depend on the nature of interference or treatment heterogeneity. These exposures may be represented as vectors, discrete classes, or real numbers. As we will show formally below, each of the distinct exposures in Δ may give rise to distinct potential outcomes for each unit in U. The estimation of causal effects under interference or treatment heterogeneity amounts to using information about treatment *assignments*, which come from the experiment's design, to estimate effects defined in terms of *treatment exposures*, which result from the interaction of the design (captured by $\mathbf{Z}$) and other underlying features of the population (captured by f and the θ_is).

To make things more concrete, consider some examples of exposure mappings. The Neyman-Rubin causal model under SUTVA corresponds to assuming an exposure mapping in which we set $\Delta = \{1, ..., M\}$ and $f(\mathbf{z}, \theta_i) = f(\mathbf{z}) = z_i$ for all i. This model has been a workhorse for much of the causal inference literature (Neyman, 1923; Rubin, 1978; Holland, 1986; Imbens and Rubin, 2011). An exposure mapping that allowed for completely arbitrary interference or treatment heterogeneity would be one for which $|\Delta| = |\Omega| \times N$, in

which case each unit has a unique type of exposure under each treatment assignment, and $f(\mathbf{z}, \theta_i)$ would be unique for each $\mathbf{z}$. If such an exposure mapping were valid, then it is clear that there would be no meaningful way to use the results of the experiment. Instead, the analyst must use substantive judgment about the extent of interference to fix a mapping somewhere between the traditional randomized experiment and completely arbitrary exposure mappings in order to carry out analyses under interference or treatment heterogeneity. For example, Hudgens and Halloran (2008) consider a setting that allows unit i's exposure to vary with each possible treatment assignment within i's group, but, where conditional on the assignment for i's group, i's exposure does not vary in the treatment assignments of other groups. Then, θ_i would be unit i's group index, and $|\Delta|$ would equal the largest number of assignment possibilities for any group. In the simulation study below and illustrative applications, we provide more examples of exposure mappings.

Units' probabilities of falling into one or another exposure condition are crucial for the estimation strategy that we develop below. Define $D_i = f(\mathbf{Z}, \theta_i)$, a random variable with support $\Delta_i \subseteq \Delta$ and for which $\Pr(D_i = d) = \pi_i(d)$. Note that because $|\Delta| \leq |\Omega| \times N$, Δ is a finite set of $K \leq |\Omega| \times N$ values, such that $\Delta = \{d_1, ..., d_K\}$. Then for each unit, i, we have a vector of probabilities, $(\pi_i(d_1), ..., \pi_i(d_K))' = \pi_i$. Invoking Imbens (2000)'s *generalized propensity score*, we call π_i the *generalized probability of exposure* for i. A unit i's generalized probability of exposure tells us the probability of i being subject to each of the possible exposures in $\{d_1, ..., d_K\}$. We have

$$\pi_i(d_k) = \sum_{\mathbf{z} \in \Omega} \mathbf{I}(f(\mathbf{z}, \theta_i) = d_k) \Pr(\mathbf{Z} = \mathbf{z}) = \sum_{\mathbf{z} \in \Omega} p_{\mathbf{z}} \mathbf{I}(f(\mathbf{z}, \theta_i) = d_k).$$

Given an experiment in which the design is known exactly (that is, $\Pr(\mathbf{Z} = \mathbf{z})$ for all $z \in \Omega$ is known exactly), the generalized probability of exposure for unit i is also known exactly. Each component probability, $\pi_i(d_k)$, is equal to the expected proportion of treatment as-

signments that induce exposure d_k for unit i.

Below, we will refer to joint exposure probabilities when discussing variance estimators. That is, we define $\pi_{ij}(d_k)$ as the probability of the joint event that both units i and j are subject to exposure d_k, and we define $\pi_{ij}(d_k, d_l)$ as the probability of the joint event that units i and j are subject to exposures d_k and d_l, respectively. To compute both individual and joint exposure probabilities from the experiment's design, first define the $N \times |\Omega|$ matrix

$$\mathbf{I}_k = [\mathbf{I}(f(\mathbf{z}, \theta_i) = d_k)]_{\substack{\mathbf{z}\in\Omega \\ i=1,\ldots,N}} =$$

$$\begin{bmatrix} \mathbf{I}(f(\mathbf{z}_1, \theta_1) = d_k) & \mathbf{I}(f(\mathbf{z}_2, \theta_1) = d_k) & \ldots & \mathbf{I}(f(\mathbf{z}_{|\Omega|}, \theta_1) = d_k) \\ \mathbf{I}(f(\mathbf{z}_1, \theta_2) = d_k) & \mathbf{I}(f(\mathbf{z}_2, \theta_2) = d_k) & \ldots & \mathbf{I}(f(\mathbf{z}_{|\Omega|}, \theta_2) = d_k) \\ \vdots & \vdots & \ddots & \\ \mathbf{I}(f(\mathbf{z}_1, \theta_N) = d_k) & \mathbf{I}(f(\mathbf{z}_2, \theta_N) = d_k) & & \mathbf{I}(f(\mathbf{z}_{|\Omega|}, \theta_N) = d_k) \end{bmatrix},$$

which is a matrix of indicators for whether units are in exposure condition k over possible assignment vectors. Define the $|\Omega| \times |\Omega|$ diagonal matrix $\mathbf{P} = \text{diag}(p_{\mathbf{z}_1}, p_{\mathbf{z}_2}, ..., p_{\mathbf{z}_{|\Omega|}})$. Then

$$\mathbf{I}_k \mathbf{P} \mathbf{I}_k' = \begin{bmatrix} \pi_1(d_k) & \pi_{12}(d_k) & \ldots & \pi_{1N}(d_k) \\ \pi_{21}(d_k) & \pi_2(d_k) & \ldots & \pi_{2N}(d_k) \\ \vdots & \vdots & \ddots & \\ \pi_{N1}(d_k) & \pi_{N2}(d_k) & & \pi_N(d_k) \end{bmatrix},$$

is an $N \times N$ symmetric matrix with individual exposure probabilities, the $\pi_i(d_k)$s, on the diagonal and joint exposure probabilities, the $\pi_{ij}(d_k)$s, on the off-diagonals. The non-

symmetric $N \times N$ matrix

$$\mathbf{I}_k \mathbf{P} \mathbf{I}_l' = \begin{bmatrix} 0 & \pi_{12}(d_k, d_l) & \dots & \pi_{1N}(d_k, d_l) \\ \pi_{21}(d_k, d_l) & 0 & \dots & \pi_{2N}(d_k, d_l) \\ \vdots & \vdots & \ddots & \\ \pi_{N1}(d_k, d_l) & \pi_{N2}(d_k, d_l) & & 0 \end{bmatrix},$$

yields all joint probabilities across exposure conditions k and l. The zeroes on the diagonal are due to the fact that a unit cannot be subject to multiple exposure conditions at once.

In practice, $|\Omega|$ may be so large that it is impractical to construct Ω to compute the π_is and the joint probability matrices exactly. One may nonetheless approximate the π_is and joint probabilities with arbitrary precision through simulation; that is, produce R random replicate $\mathbf{z}$s based on the randomization plan. From these R replicates, we can construct an $N \times R$ indicator matrix, $\widehat{\mathbf{I}}_k$, for each of the $k = 1, ..., K$ exposure conditions. Then an estimator for $\mathbf{I}_k \mathbf{P} \mathbf{I}_k'$ is $\widehat{\mathbf{I}}_k \widehat{\mathbf{I}}_k' / R$, and similarly for $\mathbf{I}_k \mathbf{P} \mathbf{I}_l'$, an estimator is $\widehat{\mathbf{I}}_k \widehat{\mathbf{I}}_l' / R$.

Proposition 3.1.

$$\mathrm{E}_R[\widehat{\mathbf{I}}_k \widehat{\mathbf{I}}_k' / R] = \mathbf{I}_k \mathbf{P} \mathbf{I}_k', \quad \mathrm{E}_R[\widehat{\mathbf{I}}_k \widehat{\mathbf{I}}_l' / R] = \mathbf{I}_k \mathbf{P} \mathbf{I}_l',$$

where $\mathrm{E}_R[.]$ *denotes the expectation across R random replicates. And, as* $R \to \infty$*,*

$$\widehat{\mathbf{I}}_k \widehat{\mathbf{I}}_k' / R \xrightarrow{a.s.} \mathbf{I}_k \mathbf{P} \mathbf{I}_k', \quad \widehat{\mathbf{I}}_k \widehat{\mathbf{I}}_l' / R \xrightarrow{a.s.} \mathbf{I}_k \mathbf{P} \mathbf{I}_l'.$$

Proof. The replication procedure is equivalent to drawing a random sample without replacement from Ω, with probabilities of selection equal to those which are defined in the randomization plan. The first result follows from unbiasedness of the sample mean, the second result follows from the Strong Law of Large Numbers. □

Rates of convergence of $\widehat{\mathbf{I}}_k \widehat{\mathbf{I}}_k' / R$ are discussed in Fattorini (2006) and Aronow (2013). Chen

et al. (2010) apply a similar approach.

4 Average Potential Outcomes and Causal Effects

We develop the case of estimating average unit-level causal effects of exposures. An average unit-level causal effect is defined in terms of a difference between the average of units' potential outcomes under one exposure versus the average under another exposure. The starting point is the estimation of average potential outcomes under each of the exposure conditions. With that, the analyst is in principle free to compute a variety of causal quantities of interest, not just average unit-level causal effects. For example, one could consider effects that are defined as differences between the average of potential outcomes under one set of exposures versus the average under another set of exposures. The direct, indirect and overall effects of Hudgens and Halloran (2008) are defined in this way using the construction of the "individual average potential outcome." The hierarchical designs that they consider are specifically tailored to ensure that estimators for such effects are non-parametrically identified. While our focus is on estimation of unit-level causal effects that are defined for arbitrary designs, such design-specific estimators can certainly be derived and analyzed using the framework developed here. Our focus on the average unit-level causal effect is due to its being the natural extension of the "average treatment effect" that is the focus of much current causal inference and program evaluation literature (e.g., Imbens and Wooldridge, 2009).

Suppose all units have non-zero probabilities of being subject to each of the K exposures: $0 < \pi_i(d_k) < 1$ for all i and k. (When $\pi_i(d_k) = 0$ for some units, then design-based estimation of average potential outcomes and causal effects must be restricted to the subset of units for which $\pi_i(d_k) > 0$.) Each unit i has K potential outcomes, which we denote by $(y_i(d_1), ..., y_i(d_K))$, that do not depend on the value of $\mathbf{Z}$. We seek estimates for all k of

$\mu(d_k) = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k) = \frac{1}{N} y^T(d_k)$, where $y^T(d_k)$ is the total of the potential outcomes under d_k. The number of units in the population, N, is fixed, but we cannot estimate $y^T(d_k)$ directly, as we only observe $y_i(d_k)$ for those with $D_i = d_k$. However, by design, the collection of units for which we observe $y_i(d_k)$ is an unequal-probability without-replacement sample from $(y_1(d_k), ..., y_N(d_k))$, with the sampling probabilities known exactly. By Horvitz and Thompson (1952), a natural estimator for $y^T(d_k)$ is the inverse probability weighted estimator

$$\widehat{y^T_{HT}}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{y_i(d_k)}{\pi_i(d_k)}. \tag{1}$$

Estimator 1 is unbiased, and its variance is characterized in Lemma 4.1.

Lemma 4.1.

$$\mathrm{E}[\widehat{y^T_{HT}}(d_k)] = \sum_{i=1}^{N} y_i(d_k)$$

$$\begin{aligned}\mathrm{Var}[\widehat{y^T_{HT}}(d_k)] = &\sum_{i=1}^{N} \pi_i(d_k)[1-\pi_i(d_k)] \left[\frac{y_i(d_k)}{\pi_i(d_k)}\right]^2 \\ &+ \sum_{i=1}^{N}\sum_{j\neq i} [\pi_{ij}(d_k) - \pi_i(d_k)\pi_j(d_k)] \frac{y_i(d_k)}{\pi_i(d_k)} \frac{y_j(d_k)}{\pi_j(d_k)}.\end{aligned} \tag{2}$$

The result follows directly from the fact that $\widehat{y^T_{HT}}(d_k)$ is a sum of correlated random variables. Given the estimator of the total of the N potential outcomes under exposure d_k, a natural estimator for the mean is thus $\widehat{\mu_{HT}}(d_k) = (1/N)\widehat{y^T_{HT}}(d_k)$, with variance $\mathrm{Var}(\widehat{\mu_{HT}}(d_k)) = (1/N^2)\mathrm{Var}[\widehat{y^T_{HT}}(d_k)]$. This allows us to construct the difference in estimated means

$$\widehat{\tau_{HT}}(d_k, d_l) = \widehat{\mu_{HT}}(d_k) - \widehat{\mu_{HT}}(d_l) = \frac{1}{N}\left[\widehat{y^T_{HT}}(d_k) - \widehat{y^T_{HT}}(d_l)\right] \tag{3}$$

which is an estimator of $\tau(d_k, d_l) = \frac{1}{N}\sum_{i=1}^{N}[y_i(d_k) - y_i(d_l)]$, the average unit-level causal

effect of exposure k versus exposure l.

Proposition 4.1.

$$\mathrm{E}\left[\widehat{\tau_{HT}}(d_k,d_l)\right] = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k) - \frac{1}{N}\sum_{i=1}^{N} y_i(d_l)$$

$$\begin{aligned}\mathrm{Var}\left(\widehat{\tau_{HT}}(d_k,d_l)\right) = &\frac{1}{N^2}\Big\{\mathrm{Var}\left[\widehat{y^T_{HT}}(d_k)\right] + \mathrm{Var}\left[\widehat{y^T_{HT}}(d_l)\right] \\ &-2\mathrm{Cov}\left[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)\right]\Big\}, \end{aligned} \tag{4}$$

where

$$\begin{aligned}\mathrm{Cov}\left[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)\right] = &\sum_{i=1}^{N}\sum_{j\neq i}\frac{y_i(d_k)}{\pi_i(d_k)}\frac{y_j(d_l)}{\pi_j(d_l)}\left[\pi_{ij}(d_k,d_l) - \pi_i(d_k)\pi_j(d_l)\right] \\ &-\sum_{i=1}^{N} y_i(d_k)y_i(d_l). \end{aligned} \tag{5}$$

A proof for Proposition 4.1 follows from Wood (2008), algebraic manipulations, and noting that $\pi_{ii}(d_k,d_l) = 0$.

Expressions (2) and (5) allow us to see the conditions under which exact variances are identified. So long as all joint exposure probabilities are non-zero (that is, $\pi_{ij}(d_k) > 0$ for all i, j), unbiased estimators for $\mathrm{Var}\left[\widehat{y^T_{HT}}(d_k)\right]$ are identified for the population U. Because we only observe one potential outcome for each unit, the last term in (5) is always unidentified, and thus $\mathrm{Cov}\left[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)\right]$ is always unidentified. This is a familiar problem in estimating the randomization variance for the average treatment effect—e.g., Neyman (1923) or Freedman et al. (1998, A32-A34). If $\pi_{ij}(d_k) = 0$ for some i, j, $\mathrm{Var}\left[\widehat{y^T_{HT}}(d_k)\right]$ is unidentified. Similarly, if $\pi_{ij}(d_k,d_l) = 0$ for some i, j, then additional components of $\mathrm{Cov}\left[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)\right]$ are unidentified. Nonetheless, we can always identify estimators for $\mathrm{Var}\left[\widehat{y^T_{HT}}(d_k)\right]$ and $\mathrm{Cov}\left[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)\right]$ that are guaranteed to have nonnegative bias. Thus, we can always identify a conservative approximation to the exact variances. We take

this and related issues up in the next section.

5 Variance Estimators

We derive conservative estimators for both $\text{Var}[\widehat{y_{HT}^T}(d_k)]$ and $\text{Var}[\widehat{\tau_{HT}}(d_k,d_l)]$. The formulations in this section follow from Aronow and Samii (2013) which considers conservative variance estimation for generic sampling designs with some zero pairwise inclusion probabilities. Although not necessarily unbiased, the estimators we present here are guaranteed to have a nonnegative bias relative to the randomization distributions of the estimators.

Given $\pi_{ij}(d_k) > 0$ for all i, j, the Horvitz-Thompson estimator for $\text{Var}[\widehat{y_{HT}^T}(d_k)]$ is

$$\begin{aligned}\widehat{\text{Var}}[\widehat{y_{HT}^T}(d_k)] = &\sum_{i\in U} \mathbf{I}(D_i = d_k)[1-\pi_i(d_k)]\left[\frac{y_i(d_k)}{\pi_i(d_k)}\right]^2 \\ &+ \sum_{i\in U}\sum_{j\in U\setminus i} \mathbf{I}(D_i = d_k)\mathbf{I}(D_j = d_k) \\ &\qquad \times \frac{\pi_{ij}(d_k) - \pi_i(d_k)\pi_j(d_k)}{\pi_{ij}(d_k)} \frac{y_i(d_k)}{\pi_i(d_k)} \frac{y_j(d_k)}{\pi_j(d_k)}. \end{aligned} \tag{6}$$

Lemma 5.1. *If $\pi_{ij}(d_k) > 0$ for all i, j, then* $\text{E}[\widehat{\text{Var}}[\widehat{y_{HT}^T}(d_k)]] = \text{Var}[\widehat{y_{HT}^T}(d_k)]$.

Lemma 5.1 follows from unbiasedness of the Horvitz-Thomspon estimator for measurable designs. Then an unbiased estimator for the variance of $\widehat{\mu_{HT}}(d_k)$ is $\widehat{\text{Var}}[\widehat{\mu_{HT}}(d_k)] = (1/N^2)\widehat{\text{Var}}[\widehat{y_{HT}^T}(d_k)]$.

In the case where $\pi_{ij}(d_k) = 0$ for some i, j, the Horvitz-Thompson estimator of $\text{Var}[\widehat{y_{HT}^T}(d_k)]$ is not unbiased, but its bias is readily characterized.

Proposition 5.1. *If $\pi_{ij}(d_k) = 0$ for some i, j, then* $\text{E}[\widehat{\text{Var}}[\widehat{y_{HT}^T}(d_k)]] = \text{Var}[\widehat{y_{HT}^T}(d_k)] + A$, *where*

$$A = \sum_{i\in U} \sum_{j\in\{U\setminus i:\pi_{ij}(d_k)=0\}} y_i(d_k)y_j(d_k).$$

A proof for Lemma 5.1 follows from Aronow and Samii (2013, Proposition 1).

Note that $\widehat{\text{Var}}[\widehat{\mu_{HT}}(d_k)]$ is guaranteed to have nonnegative bias when $y_i(d_k)y_j(d_k) \geq 0$ for all i, j with $\pi_{ij}(d_k) = 0$. The bias will be small when the terms in the sum tend to offset each other, as when the relevant $y_i(d_k)$ and $y_j(d_k)$ values are centered on 0 and have low correlation with each other. (This notation requires that we assume that $0/0 = 0$.)

Another option is to use the following correction term (derived via Young's inequality),

$$\widehat{A_2}(d_k) = \sum_{i \in U} \sum_{j \in \{U \setminus i : \pi_{ij}(d_k) = 0\}} \left[\frac{\mathbf{I}(D_i = d_k) y_i(d_k)^2}{2\pi_i(d_k)} + \frac{\mathbf{I}(D_j = d_k) y_j(d_k)^2}{2\pi_j(d_k)} \right],$$

noting that $\widehat{A_2}(d_k) = 0$ if $\pi_{ij}(d_k) > 0$ for all i, j.

Proposition 5.2.

$$\text{E}\left[\widehat{\text{Var}}[\widehat{y_{HT}^T}(d_k)] + \widehat{A_2}(d_k)\right] \geq \text{Var}[\widehat{y_{HT}^T}(d_k)],$$

A proof for Propositon 5.2 follows directly from Aronow and Samii (2013, Corollary 2). Then let $\widehat{\text{Var}}_A[\widehat{\mu_{HT}}(d_k)] = (1/N^2)\left[\widehat{\text{Var}}[\widehat{y_{HT}^T}(d_k)] + \widehat{A_2}(d_k)\right]$. $\widehat{\text{Var}}_A[\widehat{\mu_{HT}}(d_k)]$ then provides a conservative estimator for the variance of the estimated average of potential outcomes under exposure d_k.

As discussed above, $\text{Cov}[\widehat{y_{HT}^T}(d_k), \widehat{y_{HT}^T}(d_l)]$ is unidentified, which is to say that there exist no unbiased or consistent estimators for this quantity. However, we can compute an approximation that is guaranteed to have expectation less than or equal to the true covariance, providing a conservative (here, nonnegatively biased) estimator for $\text{Var}(\widehat{\tau_{HT}}(d_k, d_l))$. For the case where $\pi_{ij}(d_k, d_l) > 0$ for all i, j such that $i \neq j$, we propose the Horvitz-

Thompson-type estimator for the covariance

$$\widehat{\mathrm{Cov}}[\widehat{y_{HT}^T}(d_k), \widehat{y_{HT}^T}(d_l)] = \sum_{i \in U} \sum_{j \in U \setminus i} \frac{\mathbf{I}(D_i = d_k)\mathbf{I}(D_j = d_l)}{\pi_{ij}(d_k, d_l)} \frac{y_i(d_k)}{\pi_i(d_k)} \frac{y_j(d_l)}{\pi_j(d_l)} \times [\pi_{ij}(d_k, d_l) - \pi_i(d_k)\pi_j(d_l)] - \sum_{i \in U} \left[\frac{\mathbf{I}(D_i = d_k) y_i(d_k)^2}{2\pi_i(d_k)} + \frac{\mathbf{I}(D_i = d_l) y_i(d_l)^2}{2\pi_i(d_l)} \right]. \quad (7)$$

Proposition 5.3. *If $\pi_{ij}(d_k, d_l) > 0$ for all i, j such that $i \neq j$,*

$$\mathrm{E}\left[\widehat{\mathrm{Cov}}[\widehat{y_{HT}^T}(d_k), \widehat{y_{HT}^T}(d_l)]\right] \leq \mathrm{Cov}[\widehat{y_{HT}^T}(d_k), \widehat{y_{HT}^T}(d_l)],$$

A proof for Proposition 5.3 follows from noting that the term on the second line in expression (7) has expected value less than or equal to the quantity in the last line of expression (5), again via Young's inequality. See Aronow and Samii (2013, Proposition 2) for greater detail.

$\widehat{\mathrm{Cov}}[\widehat{y_{HT}^T}(d_k), \widehat{y_{HT}^T}(d_l)]$ is exactly unbiased if, for all $i \in U$, $y_i(d_l) = y_i(d_k)$, implying no effect associated with condition l relative to condition k.

Proposition 5.4. *If $\pi_{ij}(d_k, d_l) > 0$ for all i, j such that $i \neq j$ and for all $i \in U$, $y_i(d_l) = y_i(d_k)$*

$$\mathrm{E}\left[\widehat{\mathrm{Cov}}[\widehat{y_{HT}^T}(d_k), \widehat{y_{HT}^T}(d_l)]\right] = \mathrm{Cov}[\widehat{y_{HT}^T}(d_k), \widehat{y_{HT}^T}(d_l)],$$

A proof follows from Aronow and Samii (2013, Corollary 1).

For the case where $\pi_{ij}(d_k, d_l) = 0$ for some i, j and k, l, we can refine the expression for

the covariance given in (5) to

$$\begin{aligned}\widehat{\mathrm{Cov}}[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)] = &\sum_{i\in U}\sum_{j\in\{U\setminus i:\pi_{ij}(d_k,d_l)>0\}}\frac{y_i(d_k)}{\pi_i(d_k)}\frac{y_j(d_l)}{\pi_j(d_l)}\\ &\times[\pi_{ij}(d_k,d_l)-\pi_i(d_k)\pi_j(d_l)]\\ &-\sum_{i\in U}\sum_{j\in\{U:\pi_{ij}(d_k,d_l)=0\}} y_i(d_k)y_j(d_l), \end{aligned} \tag{8}$$

where the term on the last line subsumes the term on the last line in expression (5). This leads us to propose a more general estimator for the covariance

$$\begin{aligned}\widehat{\mathrm{Cov}}_A[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)] = &\sum_{i\in U}\sum_{j\in\{U\setminus i:\pi_{ij}(d_k,d_l)>0\}}\frac{\mathbf{I}(D_i=d_k)\mathbf{I}(D_j=d_l)}{\pi_{ij}(d_k,d_l)}\\ &\times\frac{y_i(d_k)}{\pi_i(d_k)}\frac{y_j(d_l)}{\pi_j(d_l)}\\ &\times[\pi_{ij}(d_k,d_l)-\pi_i(d_k)\pi_j(d_l)]\\ &-\sum_{i\in U}\sum_{j\in\{U:\pi_{ij}(d_k,d_l)=0\}}\left[\frac{\mathbf{I}(D_i=d_k)y_i(d_k)^2}{2\pi_i(d_k)}\right.\\ &\left.+\frac{\mathbf{I}(D_j=d_l)y_j(d_l)^2}{2\pi_j(d_l)}\right]. \end{aligned} \tag{9}$$

Proposition 5.5.

$$\mathrm{E}\left[\widehat{\mathrm{Cov}}_A[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)]\right] \leq \mathrm{Cov}[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)],$$

A proof again follows from the fact the term in the last line in (9) has expected value no greater than the term in the last line of (8) by Young's inequality.

Combining expressions, we obtain a conservative variance estimator for

$\text{Var}\,(\widehat{\tau_{HT}}(d_k,d_l))$ as

$$\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)] = \frac{1}{N^2}\left\{\widehat{\text{Var}}[\widehat{y^T_{HT}}(d_k)] + \widehat{A_2}(d_k) + \widehat{\text{Var}}[\widehat{y^T_{HT}}(d_l)] + \widehat{A_2}(d_l) \right. \\ \left. -2\widehat{\text{Cov}}_A[\widehat{y^T_{HT}}(d_k),\widehat{y^T_{HT}}(d_l)]\right\}. \tag{10}$$

Proposition 5.6.

$$\text{E}\left[\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)]\right] \geq \text{Var}\,[\widehat{\tau_{HT}}(d_k,d_l)],$$

The result follows from Proposition 5.2, Proposition 5.5, and linearity of expectations.

6 Asymptotics and Intervals

Consider a sequence of nested populations indexed by size N, (U_N). To define a notion of asymptotic growth, we let N tend to infinity, allowing for the design and exposure mapping as applied to each U_N to vary for each N (Brewer, 1979; Isaki and Fuller, 1982). Consistency and the asymptotic validity of Wald-type confidence intervals will then follow from restrictions on the growth process of the design and exposure mapping.

6.1 Consistency

We first establish conditions for the estimator $\widehat{\tau_{HT}}(d_k,d_l)$ to converge to $\tau(d_k,d_l)$ as N grows. We will show that, under two regularity conditions, $\widehat{\tau_{HT}}(d_k,d_l) - \tau(d_k,d_l) \xrightarrow{p} 0$ as $N \to \infty$.

Condition 1 (Boundedness of potential outcomes and exposure probabilities.)**.** *The ratio of each potential outcome to its exposure probability is bounded, so that, for all values i and d_k, $|y_i(d_k)|/\pi_i(d_k) \leq c < \infty$.*

Condition 1 can be relaxed, though Condition 2 would likely need to be strengthened accordingly.

We will also make an assumption about the amount of dependence in exposure conditions in the population. Define a pairwise dependency indicator g_{ij} such that if $g_{ij} = 0$, then $D_i \perp\!\!\!\perp D_j$, else let $g_{ij} = 1$.

Condition 2 (Restrictions on pairwise dependence.)**.** $\sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij} = o(N^2)$.

Condition 2 entails that, as N grows, the amount of pairwise clustering in exposure conditions induced by the design and exposure mapping is limited in scope.

Proposition 6.1. *Given Conditions 1 and 2,* $\widehat{\tau_{HT}}(d_k, d_l) - \tau(d_k, d_l) \xrightarrow{p} 0$ *as* $N \to \infty$.

Proof. We follow the logic of Robinson (1982). $\widehat{\mu_{HT}}(d_k)$ is unbiased for $\mu(d_k)$, and thus we need only consider the variance. Substituting from Equation (2), $N^2 \mathrm{Var}(\widehat{\mu_{HT}}(d_k)) \leq c^2 N + c^2 \sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij}$. Consistency of $\widehat{\mu_{HT}}(d_k)$ for $\mu(d_k)$ is therefore ensured when $\sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij} = o(N^2)$, as this implies that $\widehat{\mu_{HT}}(d_k) - \mu_{HT}(d_k) \xrightarrow{p} 0$. Consistency of $\widehat{\tau_{HT}}(d_k, d_l)$ for $\tau(d_k, d_l)$ follows by Slutsky's Theorem. □

6.2 Confidence Intervals

We now establish conditions for the asymptotic validity of Wald-type confidence intervals under stricter conditions on the asymptotic growth process. Consistency for the variance estimators, asymptotic normality, and therefore confidence intervals, follow straightforwardly when the amount of dependence across units in the population is limited.

We shall assume that Condition 1 holds, but will strengthen Condition 2 to ensure that dependence across exposures is limited in scope. Unlike Condition 2, we will exploit joint independence of observations rather than pairwise independence. Define a binary dependency indicator h_{ij} over all pairs $(i, j) \in \{1, 2, ..., N\} \times \{1, 2, ..., N\}$, that satisfies the

following: for any pair of disjoint sets Γ_1 and $\Gamma_2 \in \{1,...,N\}$ such that there exists no pair (i,j) with $h_{ij} = 1$ and either (i) $i \in \Gamma_1$ and $j \notin \Gamma_2$, (ii) $j \in \Gamma_1$ and $i \notin \Gamma_2$, (iii) $i \notin \Gamma_1$ and $j \in \Gamma_2$, or (iv) $j \notin \Gamma_1$ and $i \in \Gamma_2$, $\{D_i, i \in \Gamma_1\}$ and $\{D_i, i \in \Gamma_2\}$ are independent.

Condition 3 (Local dependence.)**.** *For each N, there exists an m such that for* $i \in 1,...,N$, $\sum_{j=1}^{N} h_{ij} \leq m$, *where* $m = o(N^{1/3})$.

Condition 3 is equivalent to assuming that dependencies across exposures can be represented by a dependency graph such that the maximal degree of each unit tends to be limited relative to N. Condition 3 will allow us to straightforwardly invoke a central limit theorem for random fields as derived via Stein's Method (Chen and Shao, 2004, Example 2.4.1). The slow rate of growth of m arises to ensure that our variance estimators will converge at a sufficiently fast rate. Note that Condition 3 subsumes Condition 2, as $\sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij} = o(N^{4/3})$ when Condition 3 holds.

It is illustrative to consider settings where Condition 3 holds. For Bernoulli-randomized designs, Condition 3 would hold if interference were characterized by first-order dependence on a graph connecting units and network degrees were bounded above by some value m. So long as m does not grow at a rate faster than $o(N^{1/3})$, the asymptotic scaling would be applicable. Condition 3 also generalizes the partial interference setting considered by, e.g., Sobel (2006) and Hudgens (2012) given finite subpopulations across which interference is localized (in this case, m would be the size of the largest subpopulation). However, Condition 3 would be violated if changing the treatment assigned to one unit would affect the exposure received by all N units.

Condition 4 (Nonzero limiting variance.)**.** $N\mathrm{Var}[\widehat{\tau_{HT}}(d_k, d_l)] \xrightarrow{p} c$, *where* $c > 0$.

Convergence of $N\mathrm{Var}[\widehat{\tau_{HT}}(d_k, d_l)]$ to a nonnegative constant is generally ensured by Conditions 1 and 3, sufficient for root-n consistency of $\widehat{\tau_{HT}}(d_k, d_l)$. Condition 4 is a mild reg-

ularity condition that ensures that this constant is positive, and rules out degenerate cases (e.g., all outcomes are zero).

Proposition 6.2. *Given Conditions 1, 3 and 4, Wald-type intervals constructed as*

$$\widehat{\tau_{HT}}(d_k,d_l) \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)]}$$

will tend to cover $\tau_{HT}(d_k,d_l)$ *at least* $100(1-\alpha)\%$ *of the time for large N.*

Proof. We follow a proof technique similar to that of Aronow et al. (2015) to establish convergence of $N\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)]$, though for a considerably more general setting. By Proposition 5.6, $\text{E}[N\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)] \geq N\text{Var}[\widehat{\tau_{HT}}(d_k,d_l)]$. Thus, by Chebyshev's Inequality, $\text{Var}[N\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)] \xrightarrow{p} 0$ is sufficient to establish convergence of $N\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)]$ to a value greater than or equal to $N\text{Var}[\widehat{\tau_{HT}}(d_k,d_l)]$, which is itself nonzero by Condition 4. Denote $a_{ij}(D_i,D_j)$ as the sum of the elements in $\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)]$ that incorporate observations i and j. Note that all $a_{ij}(D_i,D_j)$ are bounded above by some finite constant by Condition 1.

$$\begin{aligned}\text{Var}[N\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)]] &\leq N^{-2}\text{Var}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} h_{ij}a_{ij}(D_i,D_j)\right] \\ &= N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\text{Cov}\left[h_{ij}a_{ij}(D_i,D_j), h_{kl}a_{kl}(D_k,D_l)\right].\end{aligned}$$

Note that $\text{Cov}[h_{ij}a_{ij}(D_i,D_j), h_{kl}a_{kl}(D_k,D_l)] \neq 0$ if and only if $h_{ij}=1$ and $h_{kl}=1$ and either: $h_{ik}=1$, $h_{il}=1$, $h_{jk}=1$, or $h_{jl}=1$. By Condition 3, given $m \ll N$, each of these four conditions is satisfied by fewer than Nm^3 of the elements of the quadruple summation, and the number of elements in their union is at most $4Nm^3$. Thus, $\text{Var}[N\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)]] = o(N^{-2}\times N\times(N^{1/3})^3) = o(1)$.

Define

$$t = \frac{\widehat{\tau_{HT}}(d_k, d_l) - \tau_{HT}(d_k, d_l)}{\sqrt{\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k, d_l)]}} = \frac{\widehat{\tau_{HT}}(d_k, d_l) - \tau_{HT}(d_k, d_l)}{\sqrt{\text{Var}[\widehat{\tau_{HT}}(d_k, d_l)]}} \left(\frac{\text{Var}[\widehat{\tau_{HT}}(d_k, d_l)]}{\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k, d_l)]} \right)^{1/2}.$$

Under Conditions 1, 3 and 4, then, by Chen and Shao (2004, Theorem 2.7), $\frac{\widehat{\tau_{HT}}(d_k,d_l) - \tau_{HT}(d_k,d_l)}{\sqrt{\text{Var}[\widehat{\tau_{HT}}(d_k,d_l)]}}$ is asymptotically $\text{N}(0,1)$, while $(\text{Var}[\widehat{\tau_{HT}}(d_k,d_l)]/\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)])^{1/2}$ converges in probability to a quantity in $(0,1]$. By Slutsky's Theorem, t is asymptotically normal and Wald-type confidence intervals constructed as $\widehat{\tau_{HT}}(d_k,d_l) \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)]}$ will tend to cover $\tau_{HT}(d_k,d_l)$ at least $100(1-\alpha)\%$ of the time as $N \to \infty$. □

7 Refinements

The mean and difference-in-means estimators presented thus far are unbiased by sample theoretic arguments, and we have derived conservative variance estimators. However, we may wish to improve efficiency by incorporating auxiliary covariate information. In addition, by analogy to results from the unequal probability sampling literature, ratio approximations of the Horvitz-Thompson estimator may significantly reduce mean squared error with little cost in terms of bias (Särndal et al., 1992, pp. 181-184). We discuss such refinements here.

7.1 Covariance Adjustment

Auxiliary covariate information may help to improve efficiency. A first method of covariance adjustment is based on the so-called "difference estimator" (Raj, 1965; Särndal et al., 1992, Ch. 6). Covariance adjustment of this variety can reduce the randomization variance of the estimated exposure means and average causal effects without compromis-

ing unbiasedness. In addition, the difference estimator addresses the problem of location non-invariance that afflicts Horvitz-Thompson-type estimators (Fuller, 2009, 9-10). The estimator requires prior knowledge of how outcomes relate to covariates, perhaps obtained from analysis of auxiliary datasets.

Assume an auxiliary covariate vector $\mathbf{x_i}$ is observed for each i. We have some predefined function $g\left(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\right) \to \mathbb{R}$, where $\boldsymbol{\xi}_\mathbf{i}$ is a parameter vector. Ideally $g(\cdot)$ is calibrated on auxiliary data to produce values that approximate $y_i(d_k)$. We assume $\mathrm{Cov}\left[g\left(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\right), Z_i\right] = 0$ as a sufficient condition for unbiasedness. Define

$$\widehat{y_G^T}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{y_i(d_k)}{\pi_i(d_k)} - \sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{g\left(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\right)}{\pi_i(d_k)} + \sum_{i=1}^{N} g\left(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\right), \tag{11}$$

which is unbiased for $y^T(d_k)$ by

$$\mathrm{E}\left[-\sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{g\left(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\right)}{\pi_i(d_k)} + \sum_{i=1}^{N} g\left(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\right)\right] = 0.$$

Define $\varepsilon_i(d_k) = y_i(d_k) - g\left(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\right)$. Then, by substitution,

$$\widehat{y_G^T}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{\varepsilon_i(d_k)}{\pi_i(d_k)} + \sum_{i=1}^{N} g\left(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\right). \tag{12}$$

Estimation proceeds as above using $\widehat{y_G^T}(d_k)$ in place of $\widehat{y^T}(d_k)$ to estimate $y^T(d_k)$. Middleton and Aronow (2011) and Aronow and Middleton (2013) demonstrate that $\widehat{y_G^T}(d_k)$ is location invariant. Variance estimation proceeds as in Section 5, using $\varepsilon_i(d_k)$ in place of $y_i(d_k)$ so long as $g\left(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\right)$ is fixed.

An approximation to the difference estimator is given by regression adjustment using the data at hand. Regression can be thought of as a way to automate selection of the parameters in the difference estimator. In doing so, unbiasedness is compromised although the regression estimator is typically consistent (Särndal et al., 1992, pp. 225-239). We may use

weighted least squares to estimate a sensible parameter vector. For some common experimental designs, the least squares criterion will be optimal (Lin, 2013), and weighting by $1/\pi_i(d_k)$ ensures that the regression proceeds on a sample representative of the population of potential outcomes. With additional details on $\mathbf{I}_k$ and $g(.)$, it is possible to estimate optimal parameter vectors (Särndal et al., 1992, 219-244), though such values will typically be close to those produced by the weighted least squares estimator (barring unusual and extreme forms of clustering).

Define an estimated parameter vector associated with exposure condition d_k

$$\widehat{\xi}(d_k) = \arg\min_{\xi(d_k)} \sum_{i:D_i=d_k} \frac{1}{\pi_i(d_k)} \left[y_i(d_k) - g\left(\mathbf{x_i}, \xi(d_k)\right)\right]^2,$$

where $g(.)$ is the specification for the regression of $y_i(d_k)$ on $\mathbf{I}(D_i = d_k)$ and $\mathbf{x_i}$. Then the regression estimator for the total is

$$\widehat{y_R^T}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{y_i(d_k) - g\left(\mathbf{x_i}, \widehat{\xi}(d_k)\right)}{\pi_i(d_k)} + \sum_{i=1}^{N} g\left(\mathbf{x_i}, \widehat{\xi}(d_k)\right). \tag{13}$$

Estimation proceeds as above using $\widehat{y_R^T}(d_k)$ in place of $\widehat{y_{HT}^T}(d_k)$ to estimate $y^T(d_k)$. Under weak regularity conditions on $g(.)$, a variance estimator based on a Taylor linearization of $\widehat{y_R^T}(d_k)$ is consistent (Särndal et al., 1992, 236-237). The linearized variance estimator can be computed by substituting the residuals, $e_i = y_i(d_k) - g(\mathbf{x_i}, \widehat{\xi}(d_k))$, for the $y_i(d_k)$ terms in constructing the variance estimator given in Expression (10).

7.2 Hajek Ratio Estimation Via Weighted Least Squares

The Hajek (1971) ratio estimator is a refinement of the standard Horvitz-Thompson estimator that often facilitates efficiency gains at the cost of some finite N bias and complications in variance estimation. Let us first consider the problem that the Hajek estimator is de-

signed to resolve. The high variance of $\widehat{\mu_{HT}}(d_k)$ is often driven by the fact that some randomizations may yield an unusually large or small number of units or, depending on the nature of $\mathbf{I}_k$, an unusually large or small number of units with high values of the weights $1/\pi_i(d_k)$. The Hajek refinement allows the denominator of the estimator to vary according to the sum of the weights $1/\pi_i(d_k)$, thus shrinking the magnitude of the estimator when its value is large, and raising the magnitude of the estimator when its value is small. The Hajek ratio estimator is

$$\widehat{\mu_H}(d_k) = \frac{\sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{y_i(d_k)}{\pi_i(d_k)}}{\sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{1}{\pi_i(d_k)}}. \tag{14}$$

Note that $\mathrm{E}[\sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{1}{\pi_i(d_k)}] = N$, so that the Hajek estimator is the ratio of two unbiased estimators. It is well known that the ratio of two unbiased estimators is not an unbiased estimator of the ratio. However, the bias will tend to be small relative to the estimator's sampling variability, and we may place bounds on its magnitude.

By Hartley and Ross (1954) and Särndal et al. (1992, 176),

$$\left|\mathrm{E}[\widehat{\mu_H}(d_k)] - \mu(d_k)\right| \leq \sqrt{\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{1}{\pi_i(d_k)}\right) \mathrm{Var}\left(\widehat{\mu_H}(d_k)\right)}$$

Under Conditions 1 and 2, both variances will converge to zero, and the bias ratio will converge to zero. Practically speaking, the Hajek estimator can be computed using weighted least squares, with covariance adjustment through weighted least squares residualization. Variance estimation proceeds via Taylor linearization (Särndal et al., 1992, 172-176). A linearized variance estimator can be computed by substituting the residuals, $u_i = y_i(d_k) - \widehat{\mu_H}(d_k)$, for the $y_i(d_k)$ terms in constructing the variance estimator given in Expression (10).

8 A naturalistic simulation with social network data

We use a naturalistic simulation to illustrate how our framework may be applied and also to study operating characteristics of the proposed estimators in a finite sample. We estimate direct and indirect effects of an experiment with individuals linked in a complex, undirected social network. We use friendship network data from American high school classes collected through the National Longitudinal Study of Adolescent Health (Add Health).The richness of these data makes Add Health a canonical dataset for methodological research related to social networks, as with Bramoulle et al. (2009), Chung et al. (2008), Goel and Salganik (2010), Goodreau et al. (2009), Goodreau (2007), Handcock et al. (2007), and Hunter et al. (2008). We simulate experiments in which a treatment, $\mathbf{Z}$, is randomly assigned without replacement and with uniform probability to $1/10$ of individuals in a high school network. Indirect effects are transmitted only within a subject's high school. This simulated experiment resembles various studies of network persuasion campaigns (Chen et al., 2010; Aral and Walker, 2011; Paluck, 2011).

We define the exposure mapping as function $f(\mathbf{z}, \theta_i)$ such that the parameter, θ_i, equals subject i's row in a network adjacency matrix (modified such that we have zeroes on the diagonal). The cross product, $\mathbf{z}'\theta_i$, counts the number of subjects i's peers assigned to treatment. We use a simple exposure mapping that captures direct and indirect effects of the treatment, with indirect effects being transmitted to a subject's immediate peers:

$$f(\mathbf{z}, \theta_i) = \begin{cases} d_{11} \text{ (Direct + Indirect Exposure)} : & z_i \mathbf{I}(\mathbf{z}'\theta_i > 0) = 1 \\ d_{10} \text{ (Isolated Direct Exposure)} : & z_i \mathbf{I}(\mathbf{z}'\theta_i = 0) = 1 \\ d_{01} \text{ (Indirect Exposure)} : & (1 - z_i)\mathbf{I}(\mathbf{z}'\theta_i > 0) = 1 \\ d_{00} \text{ (No Exposure)} : & (1 - z_i)\mathbf{I}(\mathbf{z}'\theta_i = 0) = 1 \end{cases}$$

where each unit falls into exactly one of the four exposure conditions. This experiment is

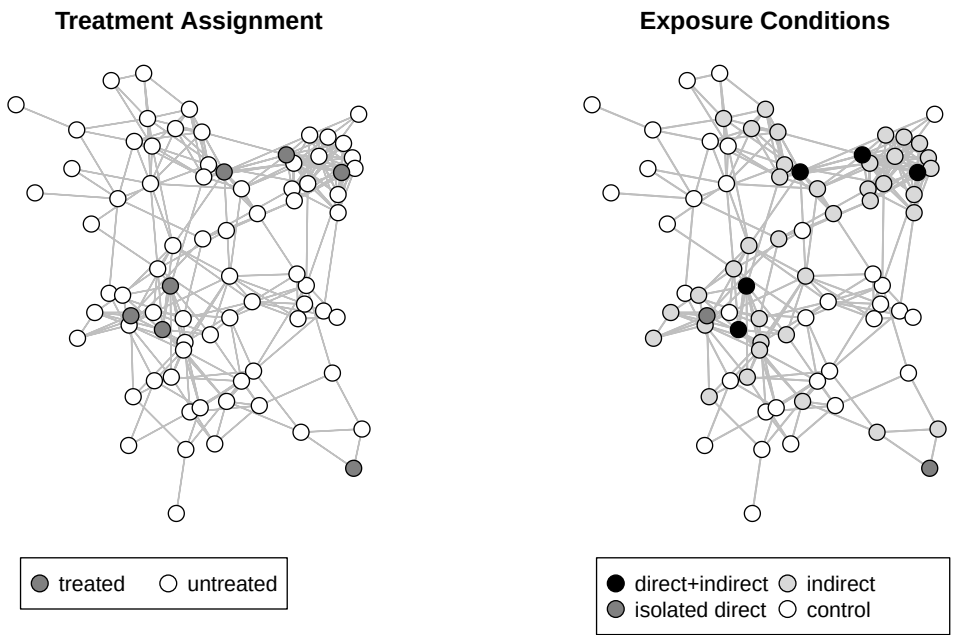

Figure 1: Illustration of a treatment assignment (left) and then treatment-induced exposures (right) for one of the high school classes in the study. Each dot is a student, and each line represents an undirected friendship tie.

repeated independently across the 144 high school classes included in Add Health, with an average class size of 626 students. To ensure that our effect estimates all refer to the same underlying population, we dropped subjects that reported zero friendship ties). We chose this exposure mapping because of its parsimony; the analyst is free to choose more complex mappings.

Figure 1 illustrates a treatment assignment and corresponding treatment-induced exposures under this mapping. The figure illustrates two key issues that our methods address. First is the connection between a unit's underlying traits, in this case its network degree, and propensity to fall into one or another exposure condition. The second is the irregular clustering that occurs in exposure conditions. Such irregular clustering is precisely what one must address in deriving variance estimates and intervals for estimated effects.

We use as our outcome a variable in the dataset that records the number of after-school activities in which each student participates. This variable defines the $y_i(d_{00})$ values—that

is, potential outcomes under the "control" exposure. This makes our simulation naturalistic not only in the networks that define the interference patterns, but also in the outcome data. The variable exhibits a high degree of right skew, with mean 2.14, standard deviation 2.64, and 0, .25, .5, .75, and 1 quantiles of 0, 1, 2, 3, and 33, respectively. We consider a simple "dilated effects" scenario (Rosenbaum, 1999) where potential outcomes are such that $y_i(d_{11}) = 2 \times y_i(d_{00}), y_i(d_{10}) = 1.5 \times y_i(d_{00}), y_i(d_{01}) = 1.25 \times y_i(d_{00})$. We run 500 simulated replications of the experiment, applying five estimators in each scenario:

- The Horvitz-Thompson estimator for the causal effect given in expression (3), with the associated conservative variance estimator, given in expression (10);
- The Hajek ratio estimator given in expression (14), with the associated linearized variance estimator;
- The weighted least squares (WLS) estimator given in expression (13), adjusting for network degree as the sole covariate, with the associated linearized variance estimator;
- An ordinary least squares (OLS) estimator that regresses the outcome on indicator variables for the exposure conditions, adjusting for network degree as a covariate, with MacKinnon and White (1985)'s finite sample adjusted "HC2" heteroskedasticity consistent variance estimator;
- A simple difference in sample means (DSM) for the exposure conditions, also with the HC2 estimator.

With respect to point estimates, the Horvitz-Thompson estimator is unbiased but possibly unstable, while the Hajek and WLS estimators are consistent and expected to be more stable. The DSM estimator is expected to be biased because it totally ignores relationships between exposure probabilities and outcomes. The OLS estimator controls for network

degree, and so this will remove bias due to correlation between exposure probabilities and outcomes. However, OLS is known to be biased in its aggregation of unit-level heterogeneity in causal effects (Angrist and Krueger, 1999). With respect to standard error estimates and confidence intervals, the variance estimators for the Horvitz-Thompson, Hajek, and WLS estimators are expected to be conservative though informative. The variance estimators for OLS and DSM may be anti-conservative because they ignore the clustering in exposure conditions.

Table 1 shows results of the simulation study, which conform to expectations. The Horvitz-Thompson, Hajek, and WLS estimators exhibit no perceivable bias. The Horvitz-Thompson estimator exhibits higher variability than the Hajek and WLS estimators. The OLS estimator and DSM estimator are heavily biased when considered relative to the variability of the effect estimates. The bias in OLS is expected because unit-level causal effects, defined in terms of differences, are heterogenous from unit to unit when underlying potential outcomes are based on dilated effects. Thus OLS will suffer from an aggregation bias in addition to any biases due to inadequate conditioning on network degree. The standard error estimates for the Horvitz-Thompson, Hajek, and WLS estimators are informative but conservative, resulting in empirical coverage rates that exceed nominal levels. The intervals for the OLS and DSM variance estimators badly undercover, primarily due to the bias in the point estimates rather than understatement of variability.

9 Conclusion

This paper proposes an analytical framework for causal inference under interference. The framework integrates (i) an experimental design that defines the probability distribution for treatments assigned, (ii) an exposure mapping that relates treatments assigned to exposures received, and (iii) an estimand chosen to make use of an experimental design to

Table 1: Results from high school friends' network simulated experiment

Estimator	Estimand	Bias	S.D.	RMSE	Mean S.E.	95% CI Coverage	90% CI Coverage
HT	$\tau(d_{01}, d_{00})$	0.00	0.04	0.04	0.05	0.960	0.924
	$\tau(d_{10}, d_{00})$	0.00	0.10	0.10	0.19	0.986	0.970
	$\tau(d_{11}, d_{00})$	0.00	0.13	0.13	0.28	0.990	0.970
Hajek	$\tau(d_{01}, d_{00})$	0.00	0.03	0.03	0.03	0.968	0.916
	$\tau(d_{10}, d_{00})$	0.00	0.07	0.07	0.13	0.992	0.970
	$\tau(d_{11}, d_{00})$	0.00	0.12	0.12	0.25	0.986	0.970
WLS	$\tau(d_{01}, d_{00})$	0.00	0.03	0.03	0.03	0.970	0.928
	$\tau(d_{10}, d_{00})$	0.00	0.07	0.07	0.12	0.992	0.968
	$\tau(d_{11}, d_{00})$	0.00	0.11	0.11	0.25	0.988	0.950
OLS	$\tau(d_{01}, d_{00})$	-0.02	0.03	0.03	0.02	0.842	0.768
	$\tau(d_{10}, d_{00})$	-0.08	0.06	0.10	0.07	0.706	0.576
	$\tau(d_{11}, d_{00})$	0.12	0.09	0.15	0.09	0.660	0.530
DSM	$\tau(d_{01}, d_{00})$	0.42	0.02	0.42	0.02	0.000	0.000
	$\tau(d_{10}, d_{00})$	-0.08	0.06	0.10	0.07	0.726	0.614
	$\tau(d_{11}, d_{00})$	0.56	0.09	0.57	0.09	0.000	0.000

HT = Horvitz-Thompson estimator with conservative variance estimator.

Hajek = Hajek estimator with linearized variance estimator.

WLS = Least squares weighted by exposure probabilities with covariate adjustment for network degree and linearized variance estimator.

OLS = Ordinary least squares with covariate adjustment for network degree and heteroskedasticity consistent variance estimator.

DSM = Simple difference in sample means with no covariate adjustment and heteroskedasticity consistent variance estimator.

S.D. = Empirical standard deviation from simulation; RMSE = Root mean square error; S.E. = standard error estimate; CI = Normal approximation confidence interval.

answer questions of substantive interest. Using this framework, we develop methods for estimating average unit-level causal effects of exposures from a randomized experiment. Our approach combines the known randomization process with the analyst's definition of treatment exposure, thus permitting inference under clear and defensible assumptions. Importantly, the union of the design of the experiment and the exposure mapping may imply unequal probabilities of exposure and forms of dependence between units that may not be obvious ex ante.

We develop estimators based on results from the literature on unequal probability sampling rooted in the foundational insights of Horvitz and Thompson (1952). The estimators

are derived from the known sampling distribution of the "direct" treatment, $\mathbf{Z}$, and provide a basis for unbiased effect estimation and conservative variance estimation. Wald-type intervals based on a normal approximation provide a reasonable reflection of large N behavior when clustering of exposure indicator values is limited. Nonetheless, it is well known that Horvitz-Thompson-type estimators may be volatile in cases where selection probabilities vary greatly or exhibit strong inverse correlation with outcome values (Basu, 1971). Thus, we provide refinements that allow for variance control via covariance adjustment and Hajek estimation.

Our approach combines minimal assumptions about restrictions on potential outcomes with randomization-based estimators, and may be characterized as design-consistent. The framework is readily applicable to deriving estimators for estimands other than the average unit-level effect of exposures.The framework developed here represents an alternative to parametric approaches that are often employed with little substantive justification. The framework and resulting methods greatly extend the reach of randomization-based estimation of causal effects.

References

Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics. In Ahsenfelter, O. C. and Card, D., editors, *Handbook of Labor Economics*, volume 3. North Holland, Amsterdam.

Aral, S. and Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639.

Aronow, P. M. (2013). *Model Assisted Causal Inference*. PhD thesis, Department of Political Science, Yale University, New Haven, CT.

Aronow, P. M. and Middleton, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1):135–144. Manuscript, Yale University and New York University.

Aronow, P. M. and Samii, C. (2013). Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. *Survey Methodology*, 39(1):231–241.

Aronow, P. M., Samii, C., and Assenova, V. A. (2015). Cluster robust variance estimation for dyadic data. *Political Analysis*, In Press.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part i. In Godambe, V. and Sprott, D., editors, *Foundations of Statistical Inference*, page s. Holt, Rinehart, and Winston, Toronto.

Bramoulle, Y., Djebbari, H., and Fortrin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55.

Brewer, K. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, 74(368):911–915.

Chen, J., Humphreys, M., and Modi, V. (2010). Technology diffusion and social networks: Evidence from a field experiment in Uganda. Manuscript, Columbia University.

Chen, L. H. and Shao, Q. (2004). Normal approximation under local dependence. *The Annals of Probability*, 32(3A):1985–2028.

Chung, H., Lanza, S. T., and Loken, E. (2008). Latent transition analysis: Inference and estimation. *Statistics in Medicine*, 27(11):1834–1854.

Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5.

Cox, D. R. (1958). *Planning of Experiments*. Wiley.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93:269–278.

Freedman, D., Pisani, R., and Purves, R. (1998). *Statistics, 3rd Ed.* W.W. Norton, New York.

Fuller, W. A. (2009). *Sampling Statistics*. Wiley, Hoboken.

Goel, S. and Salganik, M. J. (2010). Assessing respondent-driven sampling. *Proceedings of the National Academy of Science*, 107(15):6743–6747.

Goodreau, S. M. (2007). Advances in exponential random graph (p-star) models applied to a large social network. *Social Networks*, 29(2):231–248.

Goodreau, S. M., Kitts, J. A., and Morris, M. (2009). Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescents in social networks. *Demography*, 46(1):103–125.

Hajek, J. (1971). Comment on "an essay on the logical foundations of survey sampling, part one,". In Godambe, V. and Sprott, D., editors, *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society, Series A*, 170(2):301–354.

Hartley, H. O. and Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174(4423):270–271.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hudgens, M. G. (2012). Asymptotic distribution of causal effect estimators in the presence of interference. Manuscript, University of North Carolina, Chapel Hill.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87:706–710.

Imbens, G. W. and Rubin, D. B. (2011). Causal inference in statistics and social sciences. Manuscript, Harvard University.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7(1):295–318.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.

Middleton, J. A. and Aronow, P. M. (2011). Unbiased estimation of the average treatment effect in cluster-randomized experiments. Paper Presented at the Annual Meeting of the Midwest Political Science Association, Chicago.

Neyman, J. S. (1990 [1923]). On the application of probability theory to agricultural experiments, essay on principles, section 9 (reprint). *Statistical Science*, 5(4):465–472.

Paluck, E. L. (2011). Peer pressure against prejudice: A high school field experiment examining social network change. *Journal of Experimental Psychology*, 47:350–358.

Raj, D. (1965). On a method of using multi-auxiliary information in sample surveys. *Journal of the American Statistical Association*, 60:270–277.

Robinson, P. M. (1982). On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics*, 24(2):234–238.

Rosenbaum, P. R. (1999). Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. *Biometrics*, 55(2):560–564.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

Rubin, D. B. (1990). Formal models of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3):279–292.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407.

Tchetgen-Tchetgen, E. J. and VanderWeele, T. J. (2010). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, Online early content:1–21.

Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, 24:53–78.