The Annals of Applied Statistics
2017, Vol. 11, No. 4, 1912–1947
https://doi.org/10.1214/16-AOAS1005

ESTIMATING AVERAGE CAUSAL EFFECTS UNDER GENERAL INTERFERENCE, WITH APPLICATION TO A SOCIAL NETWORK EXPERIMENT

BY PETER M. ARONOW AND CYRUS SAMII

Yale University and New York University

This paper presents a randomization-based framework for estimating causal effects under interference between units motivated by challenges that arise in analyzing experiments on social networks. 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. 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 evaluate empirical performance in realistic settings with a naturalistic simulation using social network data from American schools. We then present results from a field experiment on the spread of anti-conflict norms and behavior among school students.

1. Introduction. We develop methods for analyzing an experiment in which treatments are applied to individuals in a social network and causal effects are hypothesized to transmit to peers through the network. 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. Other applications share structural similarities to the social network case. For example, 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

Received January 2016; revised November 2016.

Key words and phrases. Interference, potential outcomes, causal inference, randomization inference, SUTVA, networks.

cases, we need methods to estimate effects of both direct and indirect exposure to a treatment. Moreover, researchers may be interested in understanding how such indirect effects vary depending on individuals' characteristics.

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. For the school field experiment, program participation was randomly assigned, but encouragement to support anti-conflict norms could come from direct participation in the program as well as indirect exposure via social network peers that participated in the program.

We start with theoretical results for an estimation framework that consists of three components: (i) the experimental (or quasi-experimental) "design," which characterizes precisely the probability distribution of treatments *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, non-negatively 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. We assess finite-sample empirical performance with a naturalistic simulation on real-world school social network data. The results demonstrate the reliability of the proposed methods in a realistic sample.

We then present our analysis of a field experiment on the effects of a program meant to promote anti-conflict norms and behavior among middle school students. In the experiment, schools were first randomly assigned to host the anti-conflict program and then sets of students within the host schools were randomly assigned to participate directly in the program. The goal was to understand how attitudinal and behavioral effects on participants might transmit through their social network and affect their peers' behavior.

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 (2012) extend Hudgens and Halloran's results, providing conservative variance estimators, a framework for finite sample inference with binary outcomes, and extensions to observational studies. Liu and Hudgens (2014) develop asymptotic results for such two-stage designs. 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, estimation 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. Bowers, Fredrickson and Panagopolous (2013) apply exact tests to evaluate parameters in models of spillover processes. Such a testing approach differs in its aims from ours, which focuses on estimating averages of potentially heterogenous unit-level causal effects.

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-induced 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, then 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 $i = 1, \ldots, N$ on which a randomized experiment is performed. Define a treatment assignment vector, $\mathbf{z} = (z_1, \ldots, z_N)'$, where $z_i \in \{1, \ldots, M\}$ 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 $\mathbf{Z} = (Z_1, \ldots, Z_N)'$ is a random vector with support Ω and $\Pr(\mathbf{Z} = \mathbf{z}) = p_{\mathbf{z}}$. Our analysis below focuses on the case where the design is known in the following sense: $\Pr(\mathbf{Z} = \mathbf{z})$ for all $\mathbf{z} \in \Omega$ is known.

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, $f(\cdot)$. In applications we consider below, θ_i is unit i's row in a network adjacency matrix. More complex exposure mappings could take in θ_i's that encode other traits of units and their peers—not only network ties, but also differences in age, gender, or other unit-level characteristics. Or, θ_i could encode not only first-degree peer connections, but also second-degree connections, third-degree, and so on. The codomain Δ contains all of the possible treatment-induced exposures that might be induced in the experiment. The contents of Δ depend on the nature of interference. These exposures may be represented as vectors, discrete classes, or scalar values. 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 amounts to using information about *treatment assignments*, which come from the experiment's design, to estimate effects defined in terms of *treatment-induced 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_i's).

To make things more concrete, consider some examples of exposure mappings. The Neyman–Rubin causal model typically considers inference under an exposure mapping in which we set $\Delta = \{1, \ldots, M\}$ 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 [Holland (1986), Imbens and Rubin (2015), Rubin (1978), Neyman (1990)]. An exposure mapping that allowed for completely arbitrary interference 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 to estimate average exposure-specific effects (although other types of causal effects may admit well-behaved estimators). 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. 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, $|\Delta|$ would equal the largest number of assignment possibilities for any group, and $f(\cdot)$ would map each assignment possibility for units in unit i's group to a separate exposure condition. The types of effects that Hudgens and Halloran (2008) construct are ones that average over these exposures for each unit. Below we discuss implications of using an exposure model that does not fully account for interference, drawing connections back to the estimators in Hudgens and Halloran (2008). In the simulation study below and in the application, we provide more examples of exposure mappings. Finally, our characterization of exposures is "reduced form" in that it does not distinguish between the channels through which interference occurs [Ogburn and VanderWeele (2014)]. The exposure mapping does not distinguish between effects that emanate directly from treatments being assigned to peers or that are mediated by changes in peers' outcomes [Eckles, Karrer and Ugander (2014), pages 8–9; Manski (1995), Chapter 7].

Units' probabilities of falling into one or another exposure condition are crucial for the estimation strategy that we develop below. Define the exposure that unit i receives as $D_i = f(\mathbf{Z}, \theta_i)$, 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 $\Delta = \{d_1, \ldots, d_K\}$. Then for each unit, i, we have a vector of probabilities, $(\pi_i(d_1), \ldots, \pi_i(d_K))' = \boldsymbol{\pi}_i$. Invoking Imbens's (2000) *generalized propensity score*, we call $\boldsymbol{\pi}_i$ 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 $\{d_1, \ldots, d_K\}$. We have

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

Thus 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 assignments 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

$$\begin{aligned}\mathbf{I}_k &= \big[\mathbf{I}\big(f(\mathbf{z}, \theta_i) = d_k\big)\big]_{\substack{\mathbf{z} \in \Omega \\ i = 1, \ldots, N}} \\ &= \begin{bmatrix} \mathbf{I}\big(f(\mathbf{z}_1, \theta_1) = d_k\big) & \mathbf{I}\big(f(\mathbf{z}_2, \theta_1) = d_k\big) & \cdots & \mathbf{I}\big(f(\mathbf{z}_{|\Omega|}, \theta_1) = d_k\big) \\ \mathbf{I}\big(f(\mathbf{z}_1, \theta_2) = d_k\big) & \mathbf{I}\big(f(\mathbf{z}_2, \theta_2) = d_k\big) & \cdots & \mathbf{I}\big(f(\mathbf{z}_{|\Omega|}, \theta_2) = d_k\big) \\ \vdots & \vdots & \ddots & \\ \mathbf{I}\big(f(\mathbf{z}_1, \theta_N) = d_k\big) & \mathbf{I}\big(f(\mathbf{z}_2, \theta_N) = d_k\big) & & \mathbf{I}\big(f(\mathbf{z}_{|\Omega|}, \theta_N) = d_k\big) \end{bmatrix},\end{aligned}$$

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 $\mathbf{P} = \operatorname{diag}(p_{\mathbf{z}_1}, p_{\mathbf{z}_2}, \ldots, p_{\mathbf{z}_{|\Omega|}})$. Then

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

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 nonsymmetric $N \times N$ matrix

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

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 $\boldsymbol{\pi}_i$'s and the joint probability matrices exactly. One may nonetheless approximate the $\boldsymbol{\pi}_i$'s 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 $k = 1, \ldots, K$ exposure conditions. Then an estimator for $\mathbf{I}_k \mathbf{P} \mathbf{I}_k'$ that incorporates mild additive smoothing to ensure nonzero marginal probability estimates is $(\hat{\mathbf{I}}_k \hat{\mathbf{I}}_k' + \iota_N)/(R+1)$, where ι_N is an $N \times N$ identity matrix. Similarly, an estimator for $\mathbf{I}_k \mathbf{P} \mathbf{I}_l'$, which does not admit additive smoothing due to zero joint inclusion probabilities, is $(\hat{\mathbf{I}}_k \hat{\mathbf{I}}_l')/R$.

PROPOSITION 3.1. *As* $R \to \infty$,

$$(\hat{\mathbf{I}}_k \hat{\mathbf{I}}_k' + \iota_N)/(R+1) \xrightarrow{\text{a.s.}} \mathbf{I}_k \mathbf{P} \mathbf{I}_k', \quad \textit{and} \quad (\hat{\mathbf{I}}_k \hat{\mathbf{I}}_l')/R \xrightarrow{\text{a.s.}} \mathbf{I}_k \mathbf{P} \mathbf{I}_l'.$$

All proofs appear in the Appendix. Rates of convergence of $\widehat{\mathbf{I}}_k \widehat{\mathbf{I}}_k'/R$ are discussed in Fattorini (2006) and Aronow (2013). Below we give guidance on selecting a value of R based on a bound on the relative bias for an estimator of a target quantity.

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 nonparametrically identified. While our focus is on estimating exposure-specific 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 of unit-level, exposure-specific causal effects 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 nonzero 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 $\pi_i(d_k) > 0$.] In the most general terms, each $\mathbf{z} \in \Omega$ can generate a potential outcome for unit i. We label the randomization potential outcome of unit i associated with $\mathbf{z}$ as $y_i^r(\mathbf{z})$. These randomization potential outcomes are fixed for all units in the population and do not depend on the value of randomized treatment, $\mathbf{Z}$. A condition that we use in our analysis below is that the exposure mapping fully characterizes interference.

CONDITION 1 (Properly specified exposure mapping). For all $i \in \{1, \ldots, N\}$ and $\mathbf{z}, \mathbf{z}' \in \Omega$ such that $f(\mathbf{z}, \theta_i) = f(\mathbf{z}', \theta_i)$, $y_i^r(\mathbf{z}) = y_i^r(\mathbf{z}')$.

Given Condition 1, each unit i has $|\Delta| = K$ potential outcomes, which we can write in terms of the exposure conditions as $(y_i(d_1), \ldots, y_i(d_K))$, where $y_i(d_k) = y_i^r(\mathbf{z}), \forall i \in \{1, \ldots, N\}$, $k \in \{1, \ldots, K\}$, and $\mathbf{z} \in \Omega$ such that $f(\mathbf{z}, \theta_i) = d_k$.

Let Y_i be the observed outcome for unit i. We assume the following consistency condition that relates the observed data to potential outcomes under the exposure model [VanderWeele (2009)].

CONDITION 2 (Consistent potential outcomes).

$$Y_i = \sum_{k=1}^{K} \mathbf{I}(D_i = d_k) y_i(d_k), \qquad \forall i \in \{1, \ldots, N\}.$$

Although SUTVA is violated at the level of treatment assignment (i.e., the individual z_i values), Conditions 1 and 2 restore a form of SUTVA with respect to exposures in a manner that is conceptually similar to Hudgens and Halloran's (2008) stratified interference assumption. Below we examine implications of violating Conditions 1 and 2—that is, implications of misspecifying the exposure mapping. Throughout, unless otherwise noted, we will be assuming Conditions 1 and 2.

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. This allows us to define an average causal effect of being in exposure condition d_k as compared to being in condition d_l as

$$\tau(d_k, d_l) = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k) - \frac{1}{N}\sum_{i=1}^{N} y_i(d_l).$$

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 $(y_1(d_k), \ldots, y_N(d_k))$, with the sampling probabilities known exactly. By Horvitz and Thompson (1952), a design-based estimator for $y^T(d_k)$ is the inverse probability weighted estimator

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

Below, we consider variance-reducing refinements to this estimator. We start with an analysis of $\widehat{y^T_{\mathrm{HT}}}(d_k)$ because it very clearly reveals first-order issues for estimation under interference. Estimator 1 is unbiased, and its variance is characterized in Lemma 4.1.

LEMMA 4.1.

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

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

Above we indicated that one can approximate $\mathbf{I}_k\mathbf{P}\mathbf{I}_k'$ with $(\hat{\mathbf{I}}_k\hat{\mathbf{I}}_k' + \iota_N)/(R+1)$, which has diagonal elements,

$$\widehat{\pi}_i(d_k) = \frac{X_i + 1}{R+1},$$

where $X_i = \sum_{r=1}^{R} \mathbf{I}_r(f(\mathbf{z}_r, \theta_i) = d_k)$ and $r = 1, \ldots, R$ indexes the replicates. Define the Horvitz–Thompson estimator that uses the $\widehat{\pi}_i(d_k)$ estimates:

$$\widehat{y^T_{\mathrm{HT},R}}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{Y_i}{\widehat{\pi}_i(d_k)}.$$

Following Fattorini (2006), the following proposition provides guidance on choosing R in terms of a bound on the relative bias for estimating $\widehat{y^T}(d_k)$.

PROPOSITION 4.1. *The relative bias for* $\widehat{y^T_{\mathrm{HT},R}}(d_k)$ *is bounded as*

$$\left| \frac{\mathrm{E}[\widehat{y^T_{\mathrm{HT},R}}(d_k)] - y^T(d_k)}{y^T(d_k)} \right| \leq \big(1 - \pi_0(d_k)\big)^{R+1},$$

where $\pi_0(d_k) = \min_i \{\pi_i(d_k)\}$.

For a relative bias target of b and given some approximation of $\pi_0(d_k)$, the bound implies selecting a number of replicates $R \geq \log(b)/\log(1 - \pi_0(d_k)) - 1$. Thus, for $b = 0.005$ and $\pi_0(d_k) = 0.0005$, this would imply at least 10,593 replicates. Given the apparent computational feasibility of producing enough replicates so as to render relative biases negligible, from here on our analysis assumes that we are working with $\mathbf{I}_k \mathbf{PI}'_k$ and $\mathbf{I}_k \mathbf{PI}'_l$.

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

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

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 4.2.

$$\mathrm{E}\big[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)\big] = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k) - \frac{1}{N}\sum_{i=1}^{N} y_i(d_l), \tag{4}$$

$$\begin{aligned}\mathrm{Var}\big(\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)\big) = \frac{1}{N^2}\{&\mathrm{Var}\big[\widehat{y^T_{\mathrm{HT}}}(d_k)\big] + \mathrm{Var}\big[\widehat{y^T_{\mathrm{HT}}}(d_l)\big] \\ &- 2\,\mathrm{Cov}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big]\},\end{aligned} \tag{5}$$

where

$$\begin{aligned}\operatorname{Cov}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big] &= \sum_{i=1}^{N}\sum_{j\neq i}\frac{y_i(d_k)}{\pi_i(d_k)}\frac{y_j(d_l)}{\pi_j(d_l)}\big[\pi_{ij}(d_k,d_l)-\pi_i(d_k)\pi_j(d_l)\big]\\ &\quad -\sum_{i=1}^{N} y_i(d_k)y_i(d_l).\end{aligned} \tag{6}$$

Expressions (2) and (6) allow us to see the conditions under which exact variances are identified. As long as all joint exposure probabilities are nonzero [that is, $\pi_{ij}(d_k) > 0$ for all i, j], unbiased estimators for $\operatorname{Var}[\widehat{y^T_{\mathrm{HT}}}(d_k)]$ are identified for the population U. Because we only observe one potential outcome for each unit, the last sum in (6) is always unidentified, and thus $\operatorname{Cov}[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)]$ is always unidentified. This is a familiar problem in estimating the randomization variance for the average treatment effect—for example, Neyman (1990) or Freedman, Pisani and Purves (1998), A32–A34. If $\pi_{ij}(d_k) = 0$ for some i, j, then $\operatorname{Var}[\widehat{y^T_{\mathrm{HT}}}(d_k)]$ is unidentified. Similarly, if $\pi_{ij}(d_k, d_l) = 0$ for some i, j, then additional components of $\operatorname{Cov}[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)]$ are unidentified. Nonetheless, we can always identify estimators for $\operatorname{Var}[\widehat{y^T_{\mathrm{HT}}}(d_k)]$ and $\operatorname{Cov}[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)]$ that are guaranteed to have non-negative bias. Thus, we can always identify a conservative approximation to the exact variances. We discuss this in the next section.

5. Variance estimators. We derive conservative estimators for both $\operatorname{Var}[\widehat{y^T_{\mathrm{HT}}}(d_k)]$ and $\operatorname{Var}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]$. 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 non-negative bias relative to the randomization distributions of the estimators.

Given $\pi_{ij}(d_k) > 0$ for all i, j, the Horvitz–Thompson estimator for $\operatorname{Var}[\widehat{y^T_{\mathrm{HT}}}(d_k)]$ is

$$\begin{aligned}\widehat{\operatorname{Var}}\big[\widehat{y^T_{\mathrm{HT}}}(d_k)\big] &= \sum_{i\in U}\mathbf{I}(D_i = d_k)\big[1-\pi_i(d_k)\big]\left[\frac{Y_i}{\pi_i(d_k)}\right]^2\\ &\quad +\sum_{i\in U}\sum_{j\in U\setminus i}\mathbf{I}(D_i = d_k)\mathbf{I}(D_j = d_k)\\ &\quad \times\frac{\pi_{ij}(d_k)-\pi_i(d_k)\pi_j(d_k)}{\pi_{ij}(d_k)}\frac{Y_i}{\pi_i(d_k)}\frac{Y_j}{\pi_j(d_k)}.\end{aligned} \tag{7}$$

LEMMA 5.1. *If* $\pi_{ij}(d_k) > 0$ *for all* i, j, *then* $\mathrm{E}[\widehat{\mathrm{Var}}[\widehat{y^T_{\mathrm{HT}}}(d_k)]] = \mathrm{Var}[\widehat{y^T_{\mathrm{HT}}}(d_k)]$.

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

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

PROPOSITION 5.1. *If* $\pi_{ij}(d_k) = 0$ *for some* i, j, *then* $\mathrm{E}[\widehat{\mathrm{Var}}[\widehat{y^T_{\mathrm{HT}}}(d_k)]] = \mathrm{Var}[\widehat{y^T_{\mathrm{HT}}}(d_k)] + A$, *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 Proposition 5.1 follows from Aronow and Samii (2013), Proposition 1, reproduced in the Appendix below.

Note that $\widehat{\mathrm{Var}}[\widehat{\mu_{\mathrm{HT}}}(d_k)]$ is guaranteed to have non-negative 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 define $0/0 = 0$.)

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

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

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

PROPOSITION 5.2.

$$\mathrm{E}\big[\widehat{\mathrm{Var}}\big[\widehat{y^T_{\mathrm{HT}}}(d_k)\big] + \widehat{A_2}(d_k)\big] \geq \mathrm{Var}\big[\widehat{y^T_{\mathrm{HT}}}(d_k)\big].$$

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

As discussed above, $\mathrm{Cov}[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)]$ 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, non-negatively biased) estimator for $\mathrm{Var}(\widehat{\tau_{\mathrm{HT}}}(d_k, d_l))$. 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

$$\begin{aligned}\widehat{\mathrm{Cov}}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big] = \sum_{i\in U}\sum_{j\in U\setminus i}\Bigg[&\frac{\mathbf{I}(D_i = d_k)\mathbf{I}(D_j = d_l)}{\pi_{ij}(d_k, d_l)}\frac{Y_i}{\pi_i(d_k)}\frac{Y_j}{\pi_j(d_l)}\\ &\times\big[\pi_{ij}(d_k, d_l) - \pi_i(d_k)\pi_j(d_l)\big]\Bigg]\\ -\sum_{i\in U}\Bigg[&\frac{\mathbf{I}(D_i = d_k)Y_i^2}{2\pi_i(d_k)} + \frac{\mathbf{I}(D_i = d_l)Y_i^2}{2\pi_i(d_l)}\Bigg].\end{aligned} \tag{8}$$

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

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

A proof for Proposition 5.3 follows from noting that the term on the second line in expression (8) has expected value less than or equal to the quantity in the last line of expression (6), again via Young's inequality; see Aronow and Samii [(2013), Proposition 2, reproduced in the Appendix below] for greater detail.

$\widehat{\mathrm{Cov}}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big]$ 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 $i \in U$, $y_i(d_l) = y_i(d_k)$, then*

$$\mathrm{E}\big[\widehat{\mathrm{Cov}}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big]\big] = \mathrm{Cov}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big].$$

A proof follows from Aronow and Samii (2013), Corollary 1, reproduced in the Appendix below.

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 (6) to

$$\begin{aligned}\mathrm{Cov}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big] = \sum_{i\in U}\sum_{j\in\{U\setminus i:\pi_{ij}(d_k,d_l)>0\}}\Bigg[&\frac{y_i(d_k)}{\pi_i(d_k)}\frac{y_j(d_l)}{\pi_j(d_l)}\\ &\times\big[\pi_{ij}(d_k, d_l) - \pi_i(d_k)\pi_j(d_l)\big]\Bigg]\\ -\sum_{i\in U}\sum_{j\in\{U:\pi_{ij}(d_k,d_l)=0\}}&y_i(d_k)y_j(d_l),\end{aligned} \tag{9}$$

where the term on the last line subsumes the term on the last line in expression (6). This leads us to propose a more general estimator for the covariance

$$
\begin{aligned}
&\widehat{\mathrm{Cov}_A}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big] \\
&\quad = \sum_{i\in U} \sum_{j\in\{U\setminus i:\pi_{ij}(d_k,d_l)>0\}} \Bigg[\frac{\mathbf{I}(D_i = d_k)\mathbf{I}(D_j = d_l)}{\pi_{ij}(d_k,d_l)} \frac{Y_i}{\pi_i(d_k)} \frac{Y_j}{\pi_j(d_l)} \\
&\qquad \times \big[\pi_{ij}(d_k,d_l) - \pi_i(d_k)\pi_j(d_l)\big]\Bigg] \\
&\qquad - \sum_{i\in U} \sum_{j\in\{U:\pi_{ij}(d_k,d_l)=0\}} \Bigg[\frac{\mathbf{I}(D_i = d_k)Y_i^2}{2\pi_i(d_k)} + \frac{\mathbf{I}(D_j = d_l)Y_j^2}{2\pi_j(d_l)} \Bigg].
\end{aligned} \tag{10}
$$

PROPOSITION 5.5.

$$
\mathrm{E}\big[\widehat{\mathrm{Cov}_A}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big]\big] \le \mathrm{Cov}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big].
$$

A proof again follows from the fact the term in the last line in (10) has expected value no greater than the term in the last line of (9) by Young's inequality.

Based on the variance expressions and correction terms defined above, we obtain a conservative variance estimator for $\mathrm{Var}(\widehat{\tau_{\mathrm{HT}}}(d_k,d_l))$ as

$$
\begin{aligned}
\widehat{\mathrm{Var}}\big[\widehat{\tau_{\mathrm{HT}}}(d_k,d_l)\big] = \frac{1}{N^2}\{&\widehat{\mathrm{Var}}\big[\widehat{y^T_{\mathrm{HT}}}(d_k)\big] + \widehat{A_2}(d_k) + \widehat{\mathrm{Var}}\big[\widehat{y^T_{\mathrm{HT}}}(d_l)\big] + \widehat{A_2}(d_l) \\
&- 2\widehat{\mathrm{Cov}_A}\big[\widehat{y^T_{\mathrm{HT}}}(d_k), \widehat{y^T_{\mathrm{HT}}}(d_l)\big]\}.
\end{aligned} \tag{11}
$$

PROPOSITION 5.6.

$$
\mathrm{E}\big[\widehat{\mathrm{Var}}\big[\widehat{\tau_{\mathrm{HT}}}(d_k,d_l)\big]\big] \ge \mathrm{Var}\big[\widehat{\tau_{\mathrm{HT}}}(d_k,d_l)\big].
$$

The result follows from Proposition 5.2, Proposition 5.5, and the 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 experimental design to be reapplied anew to each U_N, subject to the conditions defined below [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 $\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)$ to converge to $\tau(d_k, d_l)$ as N grows. We will show that, under two regularity conditions, $\widehat{\tau_{\mathrm{HT}}}(d_k, d_l) - \tau(d_k, d_l) \xrightarrow{p} 0$ as $N \to \infty$.

CONDITION 3 (Boundedness of potential outcomes and exposure probabilities). Potential outcomes and exposure probabilities are bounded so that, for all values i and d_k, $|y_i(d_k)| \leq c_1 < \infty$ and $|1/\pi_i(d_k)| \leq c_2 < \infty$.

Condition 3 can be relaxed, though Condition 4 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 4 (Restrictions on pairwise dependence). $\sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij} = o(N^2)$.

Condition 4 entails that, as N grows, the amount of pairwise clustering in exposure conditions induced by the design and exposure mapping is limited in scope. As units are added to the sample, the number of new nonzero entries in the expanding pairwise correlation matrix of exposures should be limited by the order condition.

PROPOSITION 6.1. *Given Conditions 3 and 4,* $\widehat{\tau_{\mathrm{HT}}}(d_k, d_l) - \tau(d_k, d_l) \xrightarrow{p} 0$ *as* $N \to \infty$.

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 asymptotic validity of confidence intervals, follow straightforwardly when the amount of dependence across units in the population is limited.

We shall assume that Condition 3 holds, but will strengthen Condition 4 to ensure that dependence across exposures is limited in scope. Unlike Condition 4, we will exploit joint independence of observations rather than pairwise independence. Define a binary dependency indicator h_{ij} over all pairs $(i, j) \in \{1, 2, \ldots, N\} \times \{1, 2, \ldots, N\}$, where h_{ij} satisfies the following: for any pair of disjoint sets Γ_1 and $\Gamma_2 \subseteq \{1, \ldots, N\}$ 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 5 (Local dependence). There exists a finite constant m such that, for all U_N, $i \in 1, \ldots, N$, $\sum_{j=1}^{N} h_{ij} \leq m$.

Condition 5 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 5 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 authors thank Betsy Ogburn for the suggestion of the use of Stein's method in this setting.) Finite m ensures that our variance estimators will converge at a sufficiently fast rate. Note that Condition 5 subsumes Condition 4, as $\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij} = O(N)$ when Condition 5 holds.

It is illustrative to consider settings where Condition 5 holds. For Bernoulli-randomized designs, Condition 5 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. Condition 5 also generalizes the partial interference setting considered by, for example, Sobel (2006) and Liu and Hudgens (2014) given finite subpopulations across which interference is localized (in this case, m would be the size of the largest subpopulation). However, Condition 5 would be violated if changing the treatment assigned to one unit would affect the exposure received by all N units. In comparing Conditions 4 and 5, note that Condition 4 is a restriction on the order of growth of pairwise dependencies, while Condition 5 requires local dependence. The latter condition is more restrictive, as it imposes conditions on all higher-order joint inclusion probabilities. It is possible that Condition 4 could hold, but Condition 5 would be violated if, for example, there exists a single unit for which the number of associated pairwise dependencies tended to infinity in N.

CONDITION 6 (Nonzero limiting variance.). $N\,\mathrm{Var}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)] \xrightarrow{p} c$, where $c > 0$.

Convergence of $N\,\mathrm{Var}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]$ to a non-negative constant is generally ensured by Conditions 3 and 5, sufficient for root-n consistency of $\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)$. Condition 6 is a mild regularity condition that ensures that this constant is positive, and rules out degenerate cases (e.g., all outcomes are zero).

PROPOSITION 6.2. *Given Conditions 3, 5, and 6, Wald-type intervals constructed as*

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

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

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, Swensson and Wretman (1992), pages 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, Swensson and Wretman (1992), Chapter 6]. Covariance adjustment of this variety can reduce the randomization variance of the estimated exposure means and average causal effects without compromising unbiasedness. In addition, the difference estimator addresses the problem of location noninvariance that afflicts Horvitz–Thompson-type estimators [Fuller (2009), pages 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 $g(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)) \to \mathbb{R}$, 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 $\operatorname{Cov}[g(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)), Z_i] = 0$ as a sufficient condition for unbiasedness. Define

$$\begin{aligned}\widehat{y_G^T}(d_k) = &\sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{Y_i}{\pi_i(d_k)} - \sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{g(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k))}{\pi_i(d_k)} \\ &+ \sum_{i=1}^{N} g\big(\mathbf{x_i}, \boldsymbol{\xi_i}(d_k)\big),\end{aligned} \tag{12}$$

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

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

Define $\varepsilon_i(d_k) = Y_i - g(\mathbf{x_i}, \boldsymbol{\xi_i}(d_k))$ for cases with $D_i = d_k$. Then, by substitution,

$$\widehat{y_G^T}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{\varepsilon_i(d_k)}{\pi_i(d_k)} + \sum_{i=1}^{N} g\big(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k)\big). \tag{13}$$

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 $y_i(d_k)$ as long as $g(\mathbf{x_i}, \boldsymbol{\xi}_\mathbf{i}(d_k))$ 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, Swensson and Wretman (1992), pages 225–239]. We may use weighted least squares to es-

timate 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 $g(\cdot)$, it is possible to estimate optimal parameter vectors [Särndal, Swensson and Wretman (1992), pages 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,

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

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

$$\widehat{y_R^T}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{Y_i - g(\mathbf{x_i}, \widehat{\boldsymbol{\xi}}(d_k))}{\pi_i(d_k)} + \sum_{i=1}^{N} g\big(\mathbf{x_i}, \widehat{\boldsymbol{\xi}}(d_k)\big). \tag{14}$$

Estimation proceeds as above using $\widehat{y_R^T}(d_k)$ in place of $\widehat{y_{\mathrm{HT}}^T}(d_k)$ to estimate $y^T(d_k)$. Under weak regularity conditions on $g(\cdot)$, a variance estimator based on a Taylor linearization of $\widehat{y_R^T}(d_k)$ is consistent [Särndal, Swensson and Wretman (1992), pages 236–237]. The linearized variance estimator can be computed by substituting the residuals, $e_i = y_i(d_k) - g(\mathbf{x_i}, \widehat{\boldsymbol{\xi}}(d_k))$, for the $y_i(d_k)$ terms in constructing the variance estimator given in expression (11).

7.2. *Hajek ratio estimation via weighted least squares.* The Hájek (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 designed to resolve. The high variance of $\widehat{\mu_{\mathrm{HT}}}(d_k)$ is often driven by the fact that some randomizations may yield units with exceptionally 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

$$\widehat{\mu_H}(d_k) = \frac{\sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{Y_i}{\pi_i(d_k)}}{\sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{1}{\pi_i(d_k)}}. \tag{15}$$

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, Swensson and Wretman (1992), page 176,

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

Under Conditions 3 and 4, 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, Swensson and Wretman (1992), pages 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 (11).

8. Misspecification. Recall Condition 1, which states that the exposure mapping fully characterizes interference. Here we examine what happens when this assumption fails, for example, there is interference between units that is not fully characterized by the exposure mapping. By "misspecification" of the exposure mapping, we refer to the situation in which the condition $D_i = d_k$ may be consistent with multiple potential outcomes for some i. As in Section 4, we have randomization potential outcomes for unit i as $y_i^r(\mathbf{z})$ for all $\mathbf{z} \in \Omega$.

CONDITION 7 (Misspecification). There exists some $i \in \{1, \ldots, N\}$ and $\mathbf{z}, \mathbf{z}' \in \Omega$ such that $f(\mathbf{z}, \theta_i) = f(\mathbf{z}', \theta_i)$ and $y_i^r(\mathbf{z}) \neq y_i^r(\mathbf{z}')$. Then $Y_i = \sum_{\mathbf{z}\in\Omega}\mathbf{I}(\mathbf{Z} = \mathbf{z})y_i^r(\mathbf{z}), \forall i \in \{1, \ldots, N\}$.

The following proposition shows the implications of misspecification for the potential outcome population total estimator given in expression (1).

PROPOSITION 8.1. *Define* $\widehat{y_{\mathrm{HT}}^T}(d_k)$ *as above, but suppose Condition* 7 *instead of Conditions* 1 *and* 2. *Then*

$$\mathrm{E}\big[\widehat{y_{\mathrm{HT}}^T}(d_k)\big] = \sum_{i=1}^{N}\sum_{\mathbf{z}: f(\mathbf{z},\theta_i)=d_k} w_{i,\mathbf{z}}\, y_i^r(\mathbf{z}), \tag{16}$$

where $w_{i,\mathbf{z}} = p_{\mathbf{z}}/\pi_i(d_k)$.

Under Condition 7, the estimator $\widehat{\mu_{\mathrm{HT}}}(d_k) = (1/N)\widehat{y_{\mathrm{HT}}^T}(d_k)$ is unbiased for the population mean of what Hudgens and Halloran (2008), page 834, refer to as the "individual average potential outcome" given $D_i = d_k$. The causal effect estimate given in (3), which compares mean outcomes given exposures d_k versus d_l, is a difference in population means of individual average randomization potential outcomes given different restrictions on the set of treatments implied in constructing exposures d_k and d_l.

COROLLARY 8.1. *Under Condition 7,*

$$\mathrm{E}\big[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)\big] = \frac{1}{N}\sum_{i=1}^{N}\Bigg[\sum_{\mathbf{z}: f(\mathbf{z},\theta_i)=d_k} w_{i,\mathbf{z}} y_i^r(\mathbf{z}) - \sum_{\mathbf{z}': f(\mathbf{z}',\theta_i)=d_l} w_{i,\mathbf{z}'} y_i^r(\mathbf{z}')\Bigg].$$

Inference for such an effect would not follow immediately from the results above. However, under partial interference, inference would follow from the results of Liu and Hudgens (2014).

9. 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 realistic 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 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 Bramoullé, Djebbari and Fortin (2009), Chung, Lanza and Loken (2008), Goel and Salganik (2010), Goodreau, Kitts and Morris (2009), Goodreau (2007), Handcock, Raftery and Tantrum (2007), and Hunter, Goodreau and Handcock (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 school network. Indirect effects are transmitted only within a subject's school. This simulated experiment resembles various studies of network persuasion campaigns [Aral and Walker (2011), Chen, Humphreys and Modi (2010), Paluck (2011)], including the field experiment that we analyze below.

We define the exposure mapping as a function $f(\mathbf{z}, \theta_i)$ such that the parameter, θ_i, is a column vector equal to the transpose of subject i's row in a network adjacency matrix (modified such that we have zeroes on the diagonal). The inner product, $\mathbf{z}'\theta_i$, counts the number of subject 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 exposure mapping was selected to mimic the one used in the application studied in the next section. A contrast of mean outcomes under d_{10} versus d_{00} isolates the effect of direct exposure in the absence of any interaction with indirect exposure,

whereas the d_{11}–d_{00} contrast yields an effect that incorporates such interactive effects. The d_{01}–d_{00} contrast isolates the effect of indirect exposure in the absence of any interaction with direct exposure.

This experiment is repeated independently across the 144 school classes included in Add Health, with an average class size of 626 students. We constructed the school network graphs as undirected graphs where a link between two students was assigned if either student nominated the other as a friend in the Add Health survey. Students could nominate up to 5 male and 5 female friends. To ensure that our effect estimates all refer to the same underlying population, we dropped subjects that reported zero friendship ties. For the resulting sample, 42% of students have network degree in the 1 to 5 range, 40% in the 6 to 10 range, 18% in the 11 to 20 range, and 1% greater than 20, with a maximum degree of 39. To give an idea of the range of exposure probabilities, for the student with degree of 39, the probability of isolated direct exposure was 0.00067. In Figure 3 of the Appendix, we display the cumulative distribution functions for the four exposure probabilities. About 3% of students have an exposure probability of less than 0.01 for the direct + indirect exposure condition, 0.5% for isolated direct exposure, and then there were no cases with probabilities less than 0.01 for either the indirect- or no-exposure conditions.

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

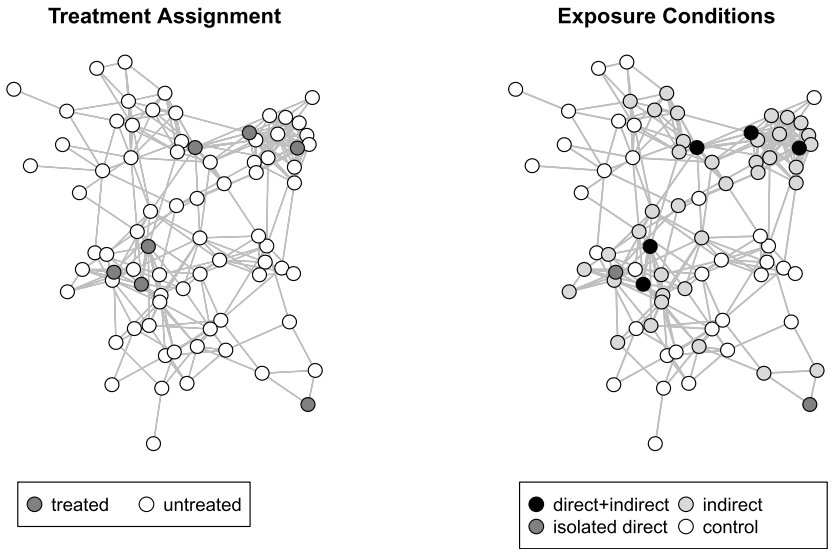

FIG. 1. *Illustration of a treatment assignment* (*left*) *and then treatment-induced exposures* (*right*) *for one of the school classes in the study. Each dot is a student, and each line represents an undirected friendship tie.*

this case its network degree, and the 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, 0.25, 0.5, 0.75, 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 (11);
- The Hajek ratio estimator given in expression (15), with the associated linearized variance estimator;
- The weighted least squares (WLS) estimator given in expression (14), 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's (1985) 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.

TABLE 1
Results from 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.

The Horvitz–Thompson estimator exhibits higher variability than the Hajek and WLS estimators, although the differences are not very pronounced, perhaps owing to the small number of cases with very small exposure probabilities. 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.

10. Analysis of a social network field experiment. In this section we analyze a field experiment on the promotion of anti-conflict norms and behavior among American middle school students. The experiment sought to shed light on how such a program might affect attitudes and behaviors of participant youth and also, crucially, to understand how these effects transmit to participants' social network peers. Full details and a richer analysis of the experiment are given in Paluck, Shepherd and Aronow (2016). The experiment involved two levels of randomization. First, 28 of 56 schools were randomly selected to host the anti-conflict program, via block randomization. Within all schools, a group of between 40 to 64 students were nonrandomly selected as eligible to participate in the program. Within each school hosting the program, half of the eligible students were then block randomized to participate in the program, with blocking on gender, grade, and a measure of network closure. Every two weeks over the course of the school year, the program had participants attend meetings with program staff during which they discussed social conflicts and patterns of exclusion at their school and formulated behavioral strategies to help friends and other students. At the beginning of the school year, the research team measured students' social networks, asking students to nominate up to 10 students in their school that they had chosen to spend time with, face to face or online, in the last two weeks. These nominations were used to construct an undirected adjacency graph so that students were considered "peers" if either student nominated one another. In Figure 2, we present an illustrative graph of one of these networks. As expected, students of the same grade and gender are more likely to associate with one another. At the end of the school year, the research team implemented a survey to measure behaviors and attitudes that reflected conflict-related norms. In the current analysis, we focus on one particular behavior: (self-reports of) wearing a wristband issued to students through the program that was meant to reflect a student's public endorsement of anti-conflict norms.

Given this design, let $z_i = 0, 1$ be an indicator for whether student i is assigned to participate in the program and let $\mathbf{z}$ be the vector of student-level assignments in the school. Let $s_i = 0, 1$ be an indicator for whether subject i's school hosts the program. Finally, as in the simulation study above, let θ_i be a column vector equal to the transpose of student i's row in the school network adjacency matrix (with zeroes on the diagonal), in which case $\mathbf{z}'\theta_i$ is again the number of subject i's peers that are assigned to participate in the program. Then we define the exposure mapping as follows:

$$f(\mathbf{z}, \theta_i) = \begin{cases} d_{111} \text{ (Direct + Indirect Exposure)} : & z_i \mathbf{I}(\mathbf{z}'\theta_i > 0) s_i = 1, \\ d_{101} \text{ (Isolated Direct Exposure)} : & z_i \mathbf{I}(\mathbf{z}'\theta_i = 0) s_i = 1, \\ d_{011} \text{ (Indirect Exposure)} : & (1 - z_i) \mathbf{I}(\mathbf{z}'\theta_i > 0) s_i = 1, \\ d_{001} \text{ (School Exposure)} : & (1 - z_i) \mathbf{I}(\mathbf{z}'\theta_i = 0) s_i = 1, \\ d_{000} \text{ (No Exposure)} : & (1 - s_i) = 1. \end{cases}$$

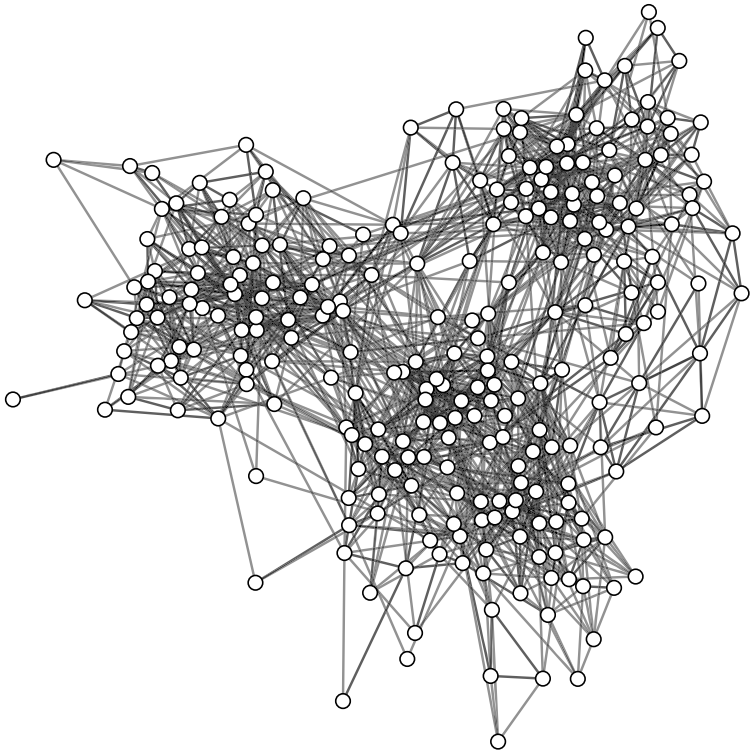

FIG. 2. *Example of a proximity network for one school in the social network field experiment. Network edges were measured using student nomination data in the first survey.*

Three features emerge from examination of the exposure mapping. First, the exposure mapping reflects four different sources of exposure to the program: being in a school with the program (School), having a peer who was a participation student (Indirect), and being a participating student (Direct). Second, only students selected as "eligible" have a nonzero possibility of being in all exposure conditions. Thus, we limit the present analysis to the set of students with nonzero probabilities of exposure ($N = 2{,}050$). [Paluck, Shepherd and Aronow (2016) examine effects for members of the ineligible subpopulation who nonetheless have nonzero probability of indirect exposure.] Third, the conditions for our asymptotic results hold in the number of schools. This exposure model provides a parsimonious characterization of first-order peer effects and school-wide climate effects, which were the primary effects of interest for Paluck, Shepherd and Aronow (2016) when designing the experiment. If other types of peer effects are present, analysis under this exposure model estimates treatment-regime-specific aggregates that average over those other effects, as described in the section on misspecification above.

TABLE 2
Social network experiment results: *effects of exposures on probability of wearing a program wristband*

Estimator	Estimand	Estimate	S.E.	95% CI
HT	$\tau(d_{001}, d_{000})$	0.057	0.062	$(-0.065, 0.179)$
	$\tau(d_{011}, d_{000})$	0.154	0.029	$(0.097, 0.211)$
	$\tau(d_{101}, d_{000})$	0.305	0.141	$(0.029, 0.581)$
	$\tau(d_{111}, d_{000})$	0.299	0.020	$(0.260, 0.338)$
Hajek	$\tau(d_{001}, d_{000})$	0.058	0.064	$(-0.067, 0.183)$
	$\tau(d_{011}, d_{000})$	0.154	0.037	$(0.081, 0.227)$
	$\tau(d_{101}, d_{000})$	0.292	0.123	$(0.051, 0.533)$
	$\tau(d_{111}, d_{000})$	0.307	0.049	$(0.211, 0.403)$
WLS	$\tau(d_{001}, d_{000})$	0.056	0.066	$(-0.072, 0.186)$
	$\tau(d_{011}, d_{000})$	0.156	0.037	$(0.083, 0.229)$
	$\tau(d_{101}, d_{000})$	0.295	0.124	$(0.050, 0.536)$
	$\tau(d_{111}, d_{000})$	0.306	0.049	$(0.212, 0.404)$

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.
S.E. = Estimated standard error; CI = Normal approximation confidence interval.

Table 2 presents Horvitz–Thompson (HT), Hajek, and weighted least squares (WLS) estimates of the effects of different exposure conditions relative to the no exposure condition. The WLS estimates control for a subject's network degree (as in the simulation study above). The outcome of interest, $y_i \in \{0, 1\}$, is a binary indicator for whether the subject wore a program wristband. The effect estimates characterize, for eligible students, the average increase in the probability of wearing a wristband relative to the average in the no exposure condition. (The average for eligible students in the no exposure condition was essentially zero, at 0.000.)

The HT, Hajek, and WLS results all mostly agree. They suggest that being in a program school but being a nonparticipant with no participant peers (d_{001}) has negligible effects for eligible students: our point estimate suggests about a six percentage point increase in the probability of wearing a wrist band, although the 95% confidence interval has a lower bound of about -7 percentage points. However, effects for eligible students with either indirect or direct exposure are substantially larger. The effect of indirect exposure (d_{011}) is about a 15 to 16 percentage point increase in the probability of wearing a wrist band (95% confidence interval between about 8 and 23 percentage points). The effect of direct exposure, whether or not it is accompanied by indirect exposure (d_{101} or d_{111}), is about a 30 percentage point increase in the probability of wearing a wrist band (95% confidence interval between about 5 and 50 percentage points for the d_{101} condition and about 21 and 40 percentage points for the d_{111} condition).

Thus, the program is seen as having substantial direct but also indirect effects on subject's willingness to endorse anti-conflict norms by wearing a program wristband. These indirect effects mean that one would drastically underestimate the effect of the program if one performed a naive analysis that simply compared participant and nonparticipant individuals in schools hosting the program. Moreover, an analysis that failed to account for indirect effects might understate the cost-effectiveness of the program: substantial increases in school-level expressions of commitment to anti-conflict norms would not require administering the program to everyone.

11. Conclusion. This paper proposes an analytical framework for causal inference under interference and applies it to the analysis of experiments on social networks. As discussed in the Introduction, the framework can be applied to other settings where interference is considered to be important. 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 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, and they 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 estimands and methods presented here may be useful in evaluating alternative experimental designs for estimating causal effects in the presence of interference [Airoldi (2016), Baird et al. (2016), Eckles, Karrer and Ugander (2014), Toulis and Kao (2013), Ugander et al. (2013)]. Finally, the framework is readily applicable to deriving estimators for estimands other than the average unit-level effect of exposures.

APPENDIX A: PROOFS

A.1. Proof of Proposition 3.1. 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 result follows from the strong law of large numbers.

A.2. Proof of Lemma 4.1. To show unbiasedness, by Condition 2 we have

$$\begin{aligned}\mathrm{E}\big[\widehat{y_{\mathrm{HT}}^{T}}(d_k)\big] &= \mathrm{E}\left[\sum_{i=1}^{N}\mathbf{I}(D_i = d_k)\frac{Y_i}{\pi_i(d_k)}\right] \\ &= \sum_{i=1}^{N}\mathrm{E}\big[\mathbf{I}(D_i = d_k)\big]\frac{y_i(d_k)}{\pi_i(d_k)} = \sum_{i=1}^{N} y_i(d_k).\end{aligned}$$

The variance expression follows from the fact that $\widehat{y_{\mathrm{HT}}^{T}}(d_k)$ is a sum of correlated random variables.

A.3. Proof of Proposition 4.1. We have

$$\begin{aligned}\mathrm{E}\big[\widehat{y_{\mathrm{HT},R}^{T}}(d_k)\big] &= \sum_{i=1}^{N} y_i(d_k)\pi_i(d_k)\mathrm{E}\left[\frac{R+1}{X_i+1}\right] \\ &= \sum_{i=1}^{N} y_i(d_k)\big[1-\big(1-\pi_i(d_k)\big)^{R+1}\big] \\ &= y^{T}(d_k) - \sum_{i=1}^{N} y_i(d_k)\big(1-\pi_i(d_k)\big)^{R+1}.\end{aligned}$$

And so

$$\begin{aligned}\big|\mathrm{E}\big[\widehat{y_{\mathrm{HT},R}^{T}}(d_k)\big] - y^{T}(d_k)\big| &= \left|\sum_{i=1}^{N} y_i(d_k)\big(1-\pi_i(d_k)\big)^{R+1}\right| \\ &\le \big|y^{T}(d_k)\big|\big(1-\pi_0(d_k)\big)^{R+1}.\end{aligned}$$

A.4. Proof of Proposition 4.2. Results (4) and (5) follow from Lemma 4.1 and properties of the variance operator. For the covariance term (6), first note that $\pi_{ii}(d_k, d_l) = 0$. Then following Wood (2008),

$$\begin{aligned}\mathrm{Cov}\big[\widehat{y_{\mathrm{HT}}^{T}}(d_k), \widehat{y_{\mathrm{HT}}^{T}}(d_l)\big] &= \sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Cov}\big[\mathbf{I}(D_i = d_k), \mathbf{I}(D_j = d_l)\big]\frac{y_i(d_k)}{\pi_i(d_k)}\frac{y_j(d_l)}{\pi_j(d_l)} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{y_i(d_k)}{\pi_i(d_k)}\frac{y_j(d_l)}{\pi_j(d_l)}\big[\pi_{ij}(d_k, d_l) - \pi_i(d_k)\pi_j(d_l)\big].\end{aligned}$$

A.5. Key results for Propositions 5.1, 5.2, 5.3, and 5.4. We reproduce key results from Aronow and Samii (2013) for the conservative variance corrections. We do so in the general case of the Horvitz–Thompson estimator for a population total. Suppose a population U indexed by $1, \ldots, k, \ldots, N$ and a sampling design such that the probability of inclusion in the sample for unit k is given by π_k, and the joint inclusion probability for units k and l is given by π_{kl}.

The Horvitz–Thompson estimator of a population total is given by

$$\hat{t} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} I_k \frac{y_k}{\pi_k},$$

where $I_k \in \{0, 1\}$ is unit k's inclusion indicator, the only stochastic component of the expression, with $\mathrm{E}(I_k) = \pi_k$, the inclusion probability, and s and U refer to the sample and the population, respectively. Define $\mathrm{E}(I_k I_l) = \pi_{kl}$, the probability that both units k and l from U are included in the sample. Since $I_k I_k = I_k$, $\mathrm{E}(I_k I_k) = \pi_{kk} = \pi_k$ by construction. The variance of the Horvitz–Thompson estimator for the total is given by

$$\begin{aligned}\mathrm{Var}(\hat{t}) &= \sum_{k \in U} \sum_{l \in U} \mathrm{Cov}(I_k, I_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \\ &= \sum_{k \in U} \mathrm{Var}(I_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in U \setminus k} \mathrm{Cov}(I_k, I_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.\end{aligned}$$

Under a measurable design, two conditions obtain: (1) $\pi_k > 0$ and π_k is known for all $k \in U$ and (2) $\pi_{kl} > 0$ and π_{kl} is known for all $k, l \in U$. Nonmeasurable designs include those for which either of the two conditions for a measurable design do not hold. We label a sample from a measurable design, s^M, and an unbiased estimator for $\mathrm{Var}(\hat{t})$ on s^M is given by

$$\widehat{\mathrm{Var}}(\hat{t}) = \sum_{k \in s^M} \sum_{l \in s^M} \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_{k \in U} \sum_{l \in U} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$

where the only stochastic part of the latter expression is $I_k I_l$, and unbiasedness is due to $\mathrm{E}(I_k I_l) = \pi_{kl}$.

Suppose a nonmeasurable design for which $\pi_{kl} = 0$ for some units $k, l \in U$. We label a sample from such a nonmeasurable design as s^0. Because I_k is a Bernoulli random variable with probability π_k, $\mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$ for $k \neq l$, and $\mathrm{Cov}(I_k, I_k) = \mathrm{Var}(I_k) = \pi_k(1 - \pi_k)$. Then we can re-express the variance above as

$$\begin{aligned}\mathrm{Var}(\hat{t}) &= \sum_{k \in U} \pi_k(1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in U \setminus k} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \\ &= \sum_{k \in U} \pi_k(1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} > 0\}} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}\end{aligned}$$

$$\underbrace{-\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} y_k y_l}_{A}.$$

For k and l such that $\pi_{kl}=0$, the sampling design will never provide information on the component of the variance labeled as A above, since we will never observe y_k and y_l together.

When $\widehat{\mathrm{Var}}(\hat{t})$ is applied to s^0, the result is unbiased for $\mathrm{Var}(\hat{t})+A$. We state this formally as follows.

PROPOSITION A.1 [Aronow and Samii (2013), Proposition 1]. *When s^0 refers to a sample from a design with some $\pi_{kl}=0$, we have*

$$\mathrm{E}\big[\widehat{\mathrm{Var}}(\hat{t})\big]=\mathrm{Var}(\hat{t})+\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} y_k y_l=\mathrm{Var}(\hat{t})+A.$$

PROOF. The result follows from

$$\begin{aligned}
&\mathrm{E}\Bigg[\sum_{k\in s^0}\sum_{l\in s^0}\frac{\mathrm{Cov}(I_k,I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\Bigg]\\
&\quad=\mathrm{E}\Bigg[\sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l\frac{\mathrm{Cov}(I_k,I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\Bigg]\\
&\quad=\sum_{k\in U}\mathrm{Var}(I_k)\left(\frac{y_k}{\pi_k}\right)^2+\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}>0\}}\mathrm{Cov}(I_k,I_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\\
&\quad=\mathrm{Var}(\hat{t})+\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} y_k y_l\\
&\quad=\mathrm{Var}(\hat{t})+A.
\end{aligned}$$

□

Now, consider the following variance estimator:

$$\begin{aligned}
\widehat{\mathrm{Var}_C}(\hat{t})&=\sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l\frac{\mathrm{Cov}(I_k,I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\\
&\quad+\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}\left(I_k\frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k}+I_l\frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l}\right),
\end{aligned}$$

where a_{kl}, b_{kl} are positive real numbers such that $\frac{1}{a_{kl}}+\frac{1}{b_{kl}}=1$ for all pairs k,l with $\pi_{kl}=0$.

PROPOSITION A.2 [Aronow and Samii (2013), Proposition 2].

$$\mathrm{E}\big[\widehat{\mathrm{Var}_C}(\hat{t})\big]\geq\mathrm{Var}(\hat{t}).$$

PROOF. By Young's inequality,

$$\frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} \geq |y_k||y_l|,$$

if $\frac{1}{a_{kl}} + \frac{1}{b_{kl}} = 1$. Define A^* such that

$$A^* = \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} \frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} \geq \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} |y_k||y_l|$$
$$\geq \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} y_k y_l = A$$

and

$$A^* \geq \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} |y_k||y_l| \geq \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} -y_k y_l = -A.$$

Therefore,

$$\operatorname{Var}(\hat{t}) + A + A^* \geq \operatorname{Var}(\hat{t}).$$

The associated Horvitz–Thompson estimator of A^* would be

$$\widehat{A^*} = \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} \left(I_k \frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k} + I_l \frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l} \right),$$

which is unbiased by $\mathrm{E}(I_k) = \pi_k$ and $\mathrm{E}(I_l) = \pi_l$.

Since $\mathrm{E}[\widehat{A^*}] = A^*$, by Proposition A.1,

$$\mathrm{E}\left[\sum_{k \in s^0} \sum_{l \in s^0} \frac{\operatorname{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \widehat{A^*} \right] = \operatorname{Var}(\hat{t}) + A + A^*,$$

$$\mathrm{E}\big[\widehat{\operatorname{Var}_C}(\hat{t})\big] \geq \operatorname{Var}(\hat{t}).$$

Substituting terms,

$$\mathrm{E}\Bigg[\sum_{k \in U} \sum_{l \in \{U : \pi_{kl} > 0\}} I_k I_l \frac{\operatorname{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$
$$+ \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} \left(I_k \frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k} + I_l \frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l} \right) \Bigg] \geq \operatorname{Var}(\hat{t}).$$

□

This estimator is unbiased under a special condition:

COROLLARY A.1 [Aronow and Samii (2013), Corollary 1]. *If, for all pairs k, l such that $\pi_{kl} = 0$,* (i) $|y_k|^{a_{kl}} = |y_l|^{b_{kl}}$ *and* (ii) $-y_k y_l = |y_k||y_l|$,

$$\mathrm{E}\big[\widehat{\operatorname{Var}_C}(\hat{t})\big] = \operatorname{Var}(\hat{t}).$$

PROOF. By (i), (ii), and Young's inequality,

$$\frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} = |y_k||y_l| = -y_k y_l.$$

Therefore,

$$\begin{aligned} A^* &= \sum_{k\in U} \sum_{l\in\{U\setminus k:\pi_{kl}=0\}} \frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} \\ &= \sum_{k\in U} \sum_{l\in\{U\setminus k:\pi_{kl}=0\}} |y_k||y_l| = \sum_{k\in U} \sum_{l\in\{U\setminus k:\pi_{kl}=0\}} -y_k y_l = -A. \end{aligned}$$

It follows that

$$\operatorname{Var}(\hat{t}) + A + A^* = \operatorname{Var}(\hat{t})$$

and

$$\mathrm{E}\big[\widehat{\operatorname{Var}_C}(\hat{t})\big] = \operatorname{Var}(\hat{t}). \qquad \square$$

In general, it would be difficult to assign optimal values of a_{kl} and b_{kl} for all pairs k, l such that $\pi_{kl} = 0$. Instead, we examine one intuitive case, assigning all $a_{kl} = b_{kl} = 2$:

$$\begin{aligned} \widehat{\operatorname{Var}_{C2}}(\hat{t}) = &\sum_{k\in U} \sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l \frac{\operatorname{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \\ &+ \sum_{k\in U} \sum_{l\in\{U\setminus k:\pi_{kl}=0\}} \left(I_k \frac{y_k^2}{2\pi_k} + I_l \frac{y_l^2}{2\pi_l} \right). \end{aligned}$$

As a special case of $\widehat{\operatorname{Var}_C}(\hat{t})$, $\widehat{\operatorname{Var}_{C2}}(\hat{t})$ is also conservative:

COROLLARY A.2 [Aronow and Samii (2013), Corollary 2].

$$\mathrm{E}\big[\widehat{\operatorname{Var}_{C2}}(\hat{t})\big] \geq \operatorname{Var}(\hat{t}).$$

PROOF. For all pairs k, l such that $\pi_{kl} = 0$, $\frac{1}{a_{kl}} + \frac{1}{b_{kl}} = \frac{1}{2} + \frac{1}{2} = 1$. Proposition A.2 therefore holds. $\square$

A.6. Proof of Proposition 6.1. We follow the logic of Robinson (1982). $\widehat{\mu_{\mathrm{HT}}}(d_k)$ is unbiased for $\mu(d_k)$, and thus we need only consider the variance. Condition 3 implies that, for all values i and d_k, $|y_i(d_k)|/\pi_i(d_k) \leq c_3 < \infty$. Substituting from equation (2), $N^2 \operatorname{Var}(\widehat{\mu_{\mathrm{HT}}}(d_k)) \leq c_3^2 N + c_3^2 \sum_{i=1}^N \sum_{j=1}^N g_{ij}$. Consistency of $\widehat{\mu_{\mathrm{HT}}}(d_k)$ 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 $\widehat{\mu_{\mathrm{HT}}}(d_k) - \mu_{\mathrm{HT}}(d_k) \xrightarrow{p} 0$. Consistency of $\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)$ for $\tau(d_k, d_l)$ follows by Slutsky's theorem.

A.7. Proof of Proposition 6.2. We follow a proof technique similar to that of Aronow, Samii and Assenova (2015) to establish convergence of $N\widehat{\mathrm{Var}}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]$, though for a considerably more general setting. By Proposition 5.6, $\mathrm{E}[N\widehat{\mathrm{Var}}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]] \geq N\,\mathrm{Var}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]$. Thus, by Chebyshev's inequality, $\mathrm{Var}[N\widehat{\mathrm{Var}}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]] \xrightarrow{p} 0$ is sufficient to establish convergence of $N\widehat{\mathrm{Var}}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]$ to a value greater than or equal to $N\,\mathrm{Var}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]$, which is itself nonzero by Condition 6. Denote $a_{ij}(D_i, D_j)$ as the sum of the elements in $\widehat{\mathrm{Var}}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]$ that incorporate observations i and j. Note that all $a_{ij}(D_i, D_j)$ are bounded above by some finite constant by Condition 3:

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

Note that $\mathrm{Cov}[h_{ij}a_{ij}(D_i, D_j), h_{kl}a_{kl}(D_k, D_l)] \neq 0$ 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 5, 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, $\mathrm{Var}[N\widehat{\mathrm{Var}}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]] = O(N^{-2} \times N) = O(N^{-1})$.

Define

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

Under Conditions 3, 5, and 6, then, by Chen and Shao [(2004), Theorem 2.7], $\frac{\widehat{\tau_{\mathrm{HT}}}(d_k, d_l) - \tau_{\mathrm{HT}}(d_k, d_l)}{\sqrt{\mathrm{Var}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]}}$ is asymptotically $\mathrm{N}(0, 1)$, while $(\mathrm{Var}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]/\widehat{\mathrm{Var}}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)])^{1/2}$ converges in probability to a quantity in $(0, 1]$. By Slutsky's theorem, t is asymptotically normal and Wald-type confidence intervals constructed as $\widehat{\tau_{\mathrm{HT}}}(d_k, d_l) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\widehat{\tau_{\mathrm{HT}}}(d_k, d_l)]}$ will tend to cover $\tau_{\mathrm{HT}}(d_k, d_l)$ at least $100(1-\alpha)\%$ of the time as $N \to \infty$.

A.8. Proof of Proposition 8.1. The result follows from iterating expectations:

$$\begin{aligned}
\mathrm{E}\big[\widehat{y_{\mathrm{HT}}^{T}(d_k)}\big] &= \mathrm{E}\left[\sum_{i=1}^{N}\mathbf{I}(D_i = d_k)\frac{Y_i}{\pi_i(d_k)}\right] \\
&= \sum_{i=1}^{N}\mathrm{E}\left[\frac{\mathbf{I}(D_i = d_k)}{\pi_i(d_k)}\mathrm{E}[Y_i \mid D_i = d_k]\right]
\end{aligned}$$

$$= \sum_{i=1}^{N} \mathrm{E}[Y_i \mid D_i = d_k] = \sum_{i=1}^{N} \frac{\sum_{\mathbf{z}: f(\mathbf{z},\theta_i)=d_k} p_{\mathbf{z}} y_i^r(\mathbf{z})}{\sum_{\mathbf{z}: f(\mathbf{z},\theta_i)=d_k} p_{\mathbf{z}}}$$

$$= \sum_{i=1}^{N} \sum_{\mathbf{z}: f(\mathbf{z},\theta_i)=d_k} \frac{p_{\mathbf{z}}}{\pi_i(d_k)} y_i^r(\mathbf{z}).$$

APPENDIX B: SIMULATION STUDY EXPOSURE PROBABILITIES

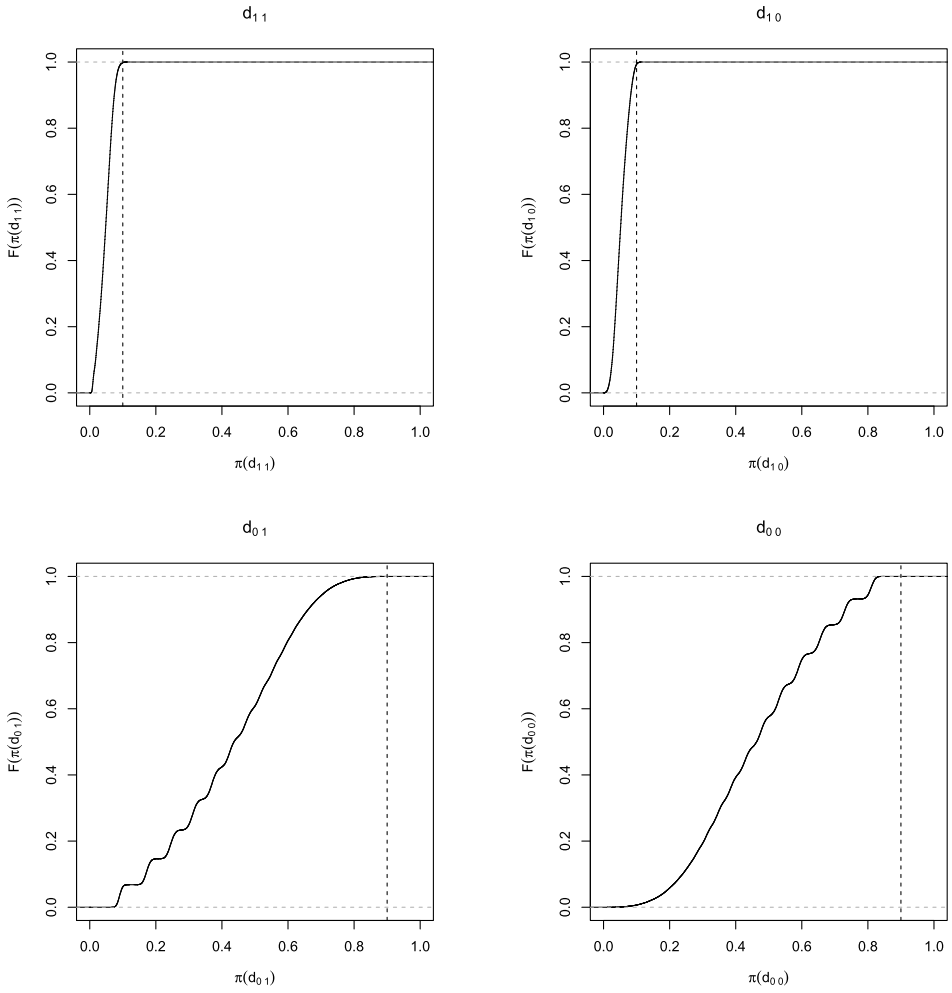

FIG. 3. *Empirical CDFs of probabilities for the four types of exposure in the simulated social network experiment.*

REFERENCES

AIROLDI, E. M. (2016). Estimating Causal Effects in the Presence of Interfering Units. Paper Presented at YINS Distinguished Lecture Series, Yale University.

ANGRIST, J. D. and KRUEGER, A. B. (1999). Empirical strategies in labor economics. In *Handbook of Labor Economics* (O. C. Ahsenfelter and D. Card, eds.) **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. *Manage. Sci.* **57** 1623–1639.

ARONOW, P. M. (2013). Model Assisted Causal Inference. Ph.D. thesis, Dept. Political Science, Yale Univ., 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** 135–144.

ARONOW, P. M. and SAMII, C. (2013). Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. *Surv. Methodol.* **39** 231–241.

ARONOW, P. M., SAMII, C. and ASSENOVA, V. A. (2015). Cluster robust variance estimation for dyadic data. *Polit. Anal.* **23** 546–577.

BAIRD, S., BOHREN, J. A., MCINTOSH, C. and OZLER, B. (2016). Designing Experiments to Measure Spillover Effects. Unpublished manuscript, George Washington Univ., Univ. Pennsylvania, Univ. California, San Diego, and World Bank.

BASU, D. (1971). An essay on the logical foundations of survey sampling. I. In *Foundations of Statistical Inference* (V. Godambe and D. A. Sprott, eds.) 203–242. Holt, Rinehart and Winston, Toronto.

BOWERS, J., FREDRICKSON, M. M. and PANAGOPOLOUS, C. (2013). Reasoning about interference between units: A general framework. *Polit. Anal.* **21** 97–124.

BRAMOULLÉ, Y., DJEBBARI, H. and FORTIN, B. (2009). Identification of peer effects through social networks. *J. Econometrics* **150** 41–55. MR2525993

BREWER, K. R. W. (1979). A class of robust sampling designs for large-scale surveys. *J. Amer. Statist. Assoc.* **74** 911–915. MR0556487

CHEN, J., HUMPHREYS, M. and MODI, V. (2010). Technology Diffusion and Social Networks: Evidence from a Field Experiment in Uganda. Manuscript, Columbia Univ., New York.

CHEN, L. H. Y. and SHAO, Q.-M. (2004). Normal approximation under local dependence. *Ann. Probab.* **32** 1985–2028. MR2073183

CHUNG, H., LANZA, S. T. and LOKEN, E. (2008). Latent transition analysis: Inference and estimation. *Stat. Med.* **27** 1834–1854. MR2420348

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

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York. MR0095561

ECKLES, D., KARRER, B. and UGANDER, J. (2014). Design and analysis of experiments in networks: Reducing bias from interference. Preprint. Available at arXiv:1404.7530 [stat.ME].

FATTORINI, L. (2006). Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika* **93** 269–278. MR2278082

FREEDMAN, D., PISANI, R. and PURVES, R. (1998). *Statistics*, 2nd 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. *Proc. Nat. Acad. Sci.* **107** 6743–6747.

GOODREAU, S. M. (2007). Advances in exponential random graph (p-star) models applied to a large social network. *Soc. Netw.* **29** 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** 103–125.

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

HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300

HARTLEY, H. O. and ROSS, A. (1954). Unbiased ratio estimators. *Nature* **174** 270–271.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472

HUNTER, D. R., GOODREAU, S. M. and HANDCOCK, M. S. (2008). Goodness of fit of social network models. *J. Amer. Statist. Assoc.* **103** 248–258. MR2394635

IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710. MR1789821

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

IMBENS, G. W. and WOOLDRIDGE, J. M. (2009). Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* **47** 5–86.

ISAKI, C. T. and FULLER, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77** 89–96. MR0648029

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420

LIU, L. and HUDGENS, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *J. Amer. Statist. Assoc.* **109** 288–301. MR3180564

MACKINNON, J. G. and WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J. Econometrics* **29** 305–325.

MANSKI, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard Univ. Press, Cambridge, MA.

MANSKI, C. F. (2013). Identification of treatment response with social interactions. *Econom. J.* **16** S1–S23. MR3030060

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, IL.

OGBURN, E. L. and VANDERWEELE, T. J. (2014). Causal diagrams for interference. *Statist. Sci.* **29** 559–578. MR3300359

PALUCK, E. L. (2011). Peer pressure against prejudice: A high school field experiment examining social network change. *J. Exp. Psychol.* **47** 350–358.

PALUCK, E. L., SHEPHERD, H. and ARONOW, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proc. Nat. Acad. Sci.* **113** 566–571.

RAJ, D. (1965). On a method of using multi-auxiliary information in sample surveys. *J. Amer. Statist. Assoc.* **60** 270–277. MR0179875

ROBINSON, P. M. (1982). On the convergence of the Horvitz–Thompson estimator. *Austral. J. Statist.* **24** 234–238. MR0678263

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

ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1990). Formal Models of Statistical Inference for Causal Effects. *J. Statist. Plann. Inference* **25** 279–292.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York. MR1140409

SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573

NEYMAN, J. S. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538

TOULIS, P. and KAO, E. (2013). Estimation of causal peer influence effects. *J. Mach. Learn. Res.* **28** 1489–1497.

UGANDER, J., KARRER, B., BACKSTROM, L. and KLEINBERG, J. (2013). Graph cluster randomization: Network exposure to multiple universes. Preprint. Available at arXiv:1305.6979 [cs.SI].

VANDERWEELE, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology* **20** 880–883.

WOOD, J. (2008). On the covariance between related Horvitz–Thompson estimators. *J. Off. Stat.* **24** 53–78.

DEPARTMENTS OF POLITICAL SCIENCE AND BIOSTATISTICS
YALE UNIVERSITY
77 PROSPECT STREET
NEW HAVEN, CONNECTICUT 06520
UNITED STATES
E-MAIL: peter.aronow@yale.edu

POLITICS DEPARTMENT
NEW YORK UNIVERSITY
19 WEST 4TH STREET
NEW YORK, NEW YORK 10012
UNITED STATES
E-MAIL: cds2083@nyu.edu