Statistical Science
2023, Vol. 0, No. 00, 1–19
https://doi.org/10.1214/23-STS883

Experimental Design in Marketplaces

Patrick Bajari, Brian Burdick, Guido W. Imbens, Lorenzo Masoero, James McQueen, Thomas S. Richardson and Ido M. Rosen

Abstract. Classical Randomized Controlled Trials (RCTs), or A/B tests, are designed to draw causal inferences about a population of units, for example, individuals, plots of land or visits to a website. A key assumption underlying a standard RCT is the absence of interactions between units, or the *stable unit treatment value assumption* (*Ann. Statist.* **6** (1978) 34–58). Modern experimentation, however, is often conducted in settings characterized by complex interactions between units. Such interactions can invalidate the standard estimators and make classical experimental designs ineffective. Although the presence of interference forces us to make untestable assumptions on the nature of the interactions even under randomization, sophisticated experimental designs can ameliorate the dependence on such assumptions. In this manuscript, we review the recent and rapidly growing literature on novel experimental designs for these settings. One key feature common to many of these designs is the presence of multiple layers of randomization within the same experiment. We discuss a novel experimental design, called *Multiple Randomization Designs* or MRDs, that provides a general framework for such experiments. Through these complex designs, we can study questions about causal effects in the presence of interference that cannot be answered by classical RCTs.

Key words and phrases: Experimental design, causal inference, online experimentation, multiple randomization designs, two-sided marketplaces.

1. INTRODUCTION

Randomized Controlled Trials (RCTs) have been an essential tool to obtain credible causal estimates since the seminal work of Neyman (1923/1990) and Fisher (1937). Their ability to obtain unbiased estimates of causal effects under weak assumptions, relative to those required for nonexperimental, observational studies, has led to their widespread adoption in a variety of fields. In biomedical settings, RCTs are a critical component of the drug approval process by the Food and Drug Administration.

In the last few decades, however, RCTs have expanded far beyond biomedical settings. Rather, hundreds of thousands of RCTs are performed in online settings every year (Kohavi et al. (2009)), Gupta et al. (2019). The units of these online experiments are often heterogenous agents acting and interacting strategically to further their objectives, rather than passive subjects. Interactions between units that invalidate the few assumptions required for the standard analysis of classical RCTs are intrinsic to these cases. For example, in ride-share companies such as Uber and Lyft, a driver picking up a particular rider affects nearby drivers and riders in the same marketplace because that driver is no longer available to pick up other riders in the short term. In short-term rental marketplaces, a rental choice made by one renter changes the choice set available to other renters. On Ebay, changing the auction format for one set of items can affect the desirability of other items. As a result of these interactions, the standard analyses of classical RCTs where the experimenter randomizes units into a treatment and control group and compares average outcomes for the two groups are no longer valid. Formally, the Stable Unit Treatment Value Assumption

Patrick Bajari is Vice President, Amazon, Seattle, WA 98109, USA (e-mail: bajari@amazon.com). Brian Burdick was Director of Research at Core-AI at Amazon while doing this work. Guido W. Imbens is Professor of Economics, Graduate School of Business and Department of Economics, Stanford University, SIEPR, NBER, Stanford, CA 94305, USA (e-mail: imbens@stanford.edu). Lorenzo Masoero is Research Scientist, Amazon, Seattle, WA 98109, USA (e-mail: masoerl@amazon.com). James McQueen is Principal Scientist, Amazon, Seattle, WA 98109, USA (e-mail: jmcq@amazon.com). Thomas S. Richardson is Professor of Statistics, University of Washington, Seattle, WA 98195, USA (e-mail: thomasr@u.washington.edu). Ido M. Rosen is Sr Principal Scientist, Core AI, Amazon, Seattle, WA 98109, USA (e-mail: ido@uchicago.edu).

(or SUTVA (Rubin (1978))), which states that the treatment assigned to one unit does not impact the outcomes observed for other units, fails to hold. Moreover, not only are many standard analyses no longer valid in that case, but the data from such an experiment do not even allow the experimenter to assess whether the no-interference assumptions on which the analyses rely are violated.

A key challenge in the absence of SUTVA is that randomization no longer frees researchers from the need to make substantive, untestable, assumptions. In particular, assumptions on the extent of the interference are necessary for designs and estimation methods that allow us to consistently estimate average effects of the interventions.

Motivated by settings in modern marketplaces where the presence of interactions is unavoidable, but their magnitude typically unknown, researchers have recently proposed a number of novel experimental designs, some of them building on older proposals in traditional experimental settings, to answer three questions: first, how do we test for the presence of interactions; second, how do we estimate the magnitude of these interaction effects; third, how do we estimate average treatment effects that account for the presence of these effects. These designs often exploit information on the structure on the interactions implied by the particular marketplace to motivate the key assumptions. For example, in some marketplaces the interactions for a particular unit are mediated solely by prices. In others, the interference may be mediated by the fraction of units treated in a peer group of that given unit.

In this paper, we review this new and rapidly growing literature, we discuss some of the open challenges and we review a novel class of experimental designs in the presence of interference, *multiple randomization designs* [MRDs] (Bajari et al. (2021), Johari, Li and Weintraub (2020)). MRDs are intended for settings where the treatment can be assigned to pairs (or tuples) of units from different populations, for example, buyers and sellers, drivers and riders or renters and rental properties. These MRDs can be thought of as generalizing many of the aforementioned new experimental designs for interference.

Another partially overlapping new class of experimental designs focuses on settings with a distinction between the population of units to which the treatments are applied, and the population of units for which the outcome is measured (Zigler and Papadogeorgou (2021)). In traditional experiments, these two populations are identical, say individuals, or plots of land. In other settings, they may be different. In Zigler and Papadogeorgou (2021), the treatments are indicators for pollution abatement associated with power plants, and the outcomes are measures of health associated with hospitals that are down wind.

2. RANDOMIZED EXPERIMENTS WITHOUT INTERFERENCE

To set the stage, let us start by briefly recalling the standard analysis of classical randomized controlled trials (RCTs) or A/B tests without interference across experimental units. Consider a finite population of N units. In traditional settings, these would be individuals, plots of land or animals. They could also be firms, countries, shopping trips or visits to a website. In a classical RCT, or A/B test, we think of each unit i in the population as being characterized by two potential outcomes, $Y_i(C)$ and $Y_i(T)$, describing the outcomes that would be obtained for unit i if it were exposed to the control or active treatment, respectively. Here the notation, with the potential outcomes for unit i only depending on the treatment for that unit already captures the SUTVA condition (Rubin (1978)). The outcomes may be, for example, an individual's spending, or the crop yield of a plot of land. We start by randomly assigning each unit i to a binary treatment denoted by $W_i \in \{C, T\}$. The randomly selected subset of the population of units with $W_i = T$ receives the experimental treatment, and the remainder of the population (units assigned $W_i = C$) receives the control treatment. A benchmark estimand is the average effect of the treatment on Y in the population, the *average causal effect*:

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(T) - Y_i(C)\big) = \overline{Y}(T) - \overline{Y}(C),$$

where $\overline{Y}(w) = \sum_{i=1}^{N} Y_i(w)/N$ for $w \in \{C, T\}$. τ is a population quantity, not directly measurable, because every unit is either assigned to treatment or control, but not both. Paul Holland refers to this as the "fundamental problem of causal inference" (p. 947, Holland (1986)). Under SUTVA, assuming the number of treated and control units is strictly positive, the difference in average outcome by treatment group is an unbiased estimate of τ:

$$\hat{\tau} = \overline{Y}_T - \overline{Y}_C,$$

where N_w is the number of units in treatment group $w \in \{C, T\}$, $\overline{Y}_w = \sum_{i=1}^{N} 1\{W_i = w\}Y_i/N_w$. The variance of $\hat{\tau}$ (conditional on N_C, N_T) is simple to characterize (Neyman (1923/1990), Cochran (1977)):

$$\mathbb{V} = \frac{S_C^2}{N_C} + \frac{S_T^2}{N_T} - \frac{S_{CT}^2}{N},$$

where for $w \in \{C, T\}$,

$$S_w^2 = \frac{1}{N-1}\sum_{i=1}^{N}\big(Y_i(w) - \overline{Y}(w)\big)^2,$$

and

$$S_{CT}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\big(Y_i(T) - Y_i(C) - \tau\big)^2.$$

The standard (although conservative) estimator for this variance is

$$\widehat{\mathbb{V}} = \frac{s_C^2}{N_C} + \frac{s_T^2}{N_T},$$

where, for $w \in \{C, T\}$,

$$s_w^2 = \frac{1}{N_w - 1} \sum_{i:W_i=w} \big(Y_i(w) - \overline{Y}_w\big)^2.$$

See Fisher (1937), Neyman (1923/1990), Holland (1986), Wu and Hamada (2011), Imbens and Rubin (2015) for general discussion of RCTs.

3. TYPES OF INTERFERENCE

In recent years, RCTs and generalizations thereof have been widely adopted by online technology companies, and have become the standard tool to test new features and measure the effectiveness of new policies (Kohavi et al. (2009)). Standard RCTs are designed for, and work well in, settings where the "no-interference" assumption or SUTVA holds, although some challenges remain (Gupta et al. (2019)). However, in contrast to traditional biomedical settings, in many such experiments it is likely that the no-interference assumptions fails to hold and that interference is present. To account for the presence of interference in the potential outcomes framework, we let unit i's potential outcomes be indexed not only by the treatment assignment for that unit, w_i, but also by the assignments of other units. In the most general case, the potential outcomes are now indexed by the full N-vector of assignments $\boldsymbol{w} = [w_1, \ldots, w_N]^\top$, $Y_i(\boldsymbol{w})$. Without any restrictions on the extent of the interference, it is difficult to make progress because we have only observations for a single treatment vector $\boldsymbol{w}$. We therefore need to impose some structure on the potential outcomes, or, in other words, postulate some model, that limits the dependence of the potential outcomes on the full assignment vector. This is true even in experimental settings where we have full control over the assignment and can choose the assignment distribution.

In the remainder of this section, we discuss some restrictions on interference that have received particular attention in the literature so far. Later we discuss specific experimental designs that can account for these types of interference. We should note that the cases below do not form an exhaustive list of types of interference.

The type of restrictions that have been analyzed can be thought of as belonging to two classes. The first class assumes that the potential outcomes for unit i depend only on a subset of the elements of $\boldsymbol{w}$, implicitly defining *exposure mappings* (Aronow and Samii (2017)). Only treatment assignments for a subset of units in the population can affect the outcomes for unit i. Ogburn and VanderWeele (2014) refer to this as *direct interference*. For example, consider an experiment where customers are offered free cancellations on some randomly selected rental properties. The fact that one rental property in a particular town comes with a free cancellation offer may affect rentals of other properties in the same town, but it is plausible that it does not affect rentals in other towns. In educational experiments, it is often reasonable to assume that interference is limited to students within the same classroom or within the same school, but does not extend beyond that.

The second class of restrictions allows for the possibility that treatments for *all* other units affect the outcome for unit i. In this setting, the interference is typically assumed to flow through the outcomes, rather than directly through the treatments. For example, suppose that the treatment is a vaccine. The outcome for individual i depends on the treatment for a different individual j only if individual j is affected by his or her own treatment. It does not matter for individual i whether individual j is treated or not, unless that the treatment for individual j affects the disease status of individual j. Ogburn and VanderWeele (2014) refer to this type of interference as *interference by contagion*. Identification given this type of interference requires additional assumptions. For example, if the researcher has a measure of the distance between unit i and j, it may be reasonable to assume that the magnitude of the spillover effects declines monotonically with the distance.

3.1 Cluster Interference

Cluster interference is the most commonly studied type of interference (Hudgens and Halloran (2008), Manski (1993), Rosenbaum (2007), Ugander et al. (2013), VanderWeele, Tchetgen and Halloran (2014), Ogburn and VanderWeele (2014), Papadogeorgou, Mealli and Zigler (2019)). It is the leading example of the first class, direct interference, where the dependence of the potential outcomes is restricted to a subset of the assignment vector. In a clustered setting, the key assumption is that the population of units can be partitioned *ex ante* in clusters, groups or subpopulations, such that the outcome for a unit in a specific cluster may be impacted only by the treatment received by other units within the same cluster. In other words, interference is unrestricted within each cluster, but assumed absent between units in different clusters. Common examples of settings in which these dynamics are present include education (e.g., outcomes for a given student may depend on treatments for other students in the same class), or labor market interventions such as job training programs (e.g., outcomes for a particular individual may be affected by training status for other individuals in the same labor market segment).

Formally, let $B_i \in \{1, \ldots, B\}$ denote the cluster (block) that unit i belongs to. The key assumption for cluster interference is that

$$Y_i(\mathbf{w}) = Y_i(\mathbf{w}'),$$

for all $\mathbf{w}$ and $\mathbf{w}'$ such that $w_j = w'_j$ for all units j in the same cluster as unit i, that is for all j such that $B_i = B_j$. Cluster interference rules out types of interference that go viral, where seeding one, or a small number of units, can lead to effects cascading throughout the whole population.

3.2 Network Interference Through Treatments

A generalization of the cluster set up that still fits into the first class of direct interference relies on the existence of an underlying network in the population. Suppose a symmetric binary adjacency matrix $\boldsymbol{A} \in \{0, 1\}^{N \times N}$ encodes links between units, where unit j is a neighbor of unit i if $\mathbf{A}_{ij} = 1$. We can use this graph structure to restrict interference by imposing that the potential outcome for unit i depends only on the treatment assignment of unit i's connections, in other words the treatments for units j such that $A_{ij} = 1$; see Aronow (2012), Athey, Eckles and Imbens (2018), Basse, Feller and Toulis (2019). Formally,

$$Y_i(\mathbf{w}) = Y_i(\mathbf{w}'),$$

for all $\mathbf{w}$ and $\mathbf{w}'$ such that $w_j = w'_j$ for all units j linked to unit i, that is, for all j such that $\mathbf{A}_{ij} = 1$. In the special case where $\mathbf{A}$ has a block structure, where $A_{ij} = 1$ and $A_{ik} = 1$ implies $A_{kj} = 1$, this setting is identical to the cluster setting.

In network settings, this type of first-order interference, where only treatments of friends affect the outcomes for a particular unit can be generalized to higher-order interference. For example, one can allow that the outcome for unit i may depend not only on the treatment assignment for the units that unit i is connected to, but also to the units they are connected to, the friends-of-friends. In other words, the outcome for unit i can be affected by treatment assignment for all units j who either are friends with i, or who have friends in common with i, such that $\sum_{j'} A_{ij'} A_{jj'} \geq 1$. See Bond et al. (2012) for an example where allowing individual j to express whether they voted or not can affect whether i votes, even if they are not directly connected. In practice, establishing the presence of such higher order interference is challenging. Although exact finite sample randomization tests are available (see Athey, Eckles and Imbens (2018)), in practice their power to detect such effects is limited because often the second-order networks are so large that there is insufficient variation in average treatments in the set of friends-of-friends.

In this setting, it may be desirable to assign treatment in such a way that there are (approximately) random samples of individuals for whom all of their neighbors in the network are treated. Backstrom and Kleinberg (2011) describe a treatment assignment method that they call "bucket testing" that achieves this while minimizing the total number of treated individuals.

3.3 Network Interference Through Outcomes

A third type of spillover assumption also uses networks to restrict interference, but in a way that units far away from unit i can still affect the potential outcomes for that unit. This allows for cascading effects, or viral effects, also called contagion by Ogburn and VanderWeele (2014). Differently from the earlier direct interference setting discussed above, here the potential outcome for unit i depends on neighbors' realized outcomes, rather than their treatment assignment. Indirectly this means that the potential outcomes for unit i depend on the treatment status for all units that unit i is connected though, whether these connections are first order or of arbitrarily higher order. Ogburn and VanderWeele (2014) present an example of a vaccine in a single possibly large population of connected individuals. Treating an individual j may affect the outcome of individual i even if these two individuals are far away if treating individual j affects their outcome, and affects the outcomes for the individuals on a path between j and i. The larger the distance between i and j, that is, the larger the number of stops along the shortest path between i and j, the smaller the probability that treating unit j has an effect on the outcome for unit i.

3.4 Decoupling Treatment Units and Outcome Units

In the discussion thus far, and in much of the traditional experimental design literature, the starting point is a single set of N units, to which both treatments could be applied and for which potential outcomes were defined. This is the natural viewpoint when starting with a classical analysis that assumes SUTVA: individuals can each separately be assigned to receive a new medical treatment or not, and their health status can be measured, or plots of land can be assigned to a particular fertilizer and crop yields can be measured.

However, this set up is not always appropriate. When SUTVA is relaxed, there is potentially a larger set of treatments that can affect a given unit's outcome, and thus there is no longer a simple one-to-one correspondence between the units on which the treatments are defined and the units on which the outcomes are measured. More generally, the set of treatable units and the set of outcomes may even be entirely distinct and have different cardinality. Thus, for example, consider a study of the effect of vaccinating parents on children, where the children are ineligible to receive the vaccine. In another example, in the spatial analysis in Pollmann (2020) treatments are the presence of restaurants in particular locations, and the outcomes are shopping trips associated with individuals.

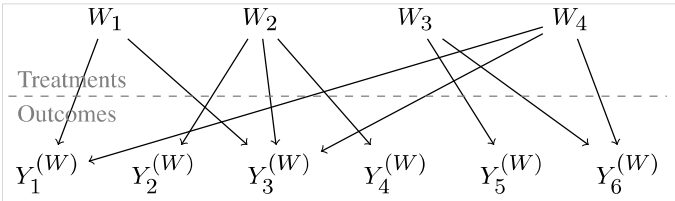

FIG. 1. *Bipartite graph representation of a general causal model design with four treatments affecting six outcome units.*

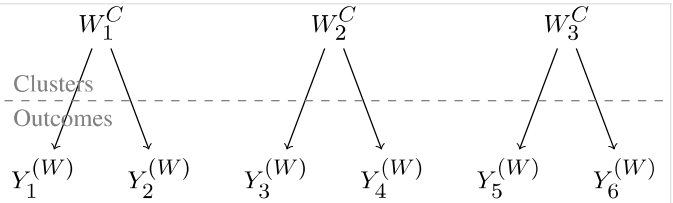

FIG. 2. *Bipartite graph representation* ($I = 3$) *of a clustered experiment. Clusters* $i \in \{1, \ldots, I\}$ *have cluster-specific treatment indicators* $W_i^C \in \{0, 1\}$.

Similarly, treatments can be at the teacher level, with outcomes measured at the student level, where each student is taught by a number of teachers.

Zigler and Papadogeorgou (2021) introduce a formal framework for describing a class of such settings, based on bipartite graphs; see also Pouget-Abadie et al. (2019). Specifically, consider a set of treatment units $\mathcal{P}$, with treatment indicators, $\{W_i, i \in \mathcal{P}\}$, with possibly binary treatments $W_i \in \{0, 1\}$, and a set of outcome units $\mathcal{Q}$ with observed responses $\{Y_j, j \in \mathcal{Q}\}$, together with a bipartite graph with vertex sets $\mathcal{P}$ and $\mathcal{Q}$.

There is an edge in the graph between treatment i and outcome j if and only if the potential outcome for unit j, $Y_j(\boldsymbol{w})$, where $\boldsymbol{w}$ is the $|\mathcal{P}|$-vector of treatments, functionally depends on its ith element, the assignment w_i for treatment unit i. Figure 1 illustrates this set up.

Zigler and Papadogeorgou (2021) consider an example relating to air pollution. The treatment units are power stations, which may be fitted ($W_i = 1$) or not ($W_i = 0$) with a device to reduce emissions of particulate matter. The outcome units are hospitalization rates for cardiovascular disease by zip codes. The outcome for zip code j may be affected by the treatment at multiple power plants, but not necessarily by all. Similarly, the treatment at powerplant i may affect outcomes at multiple zip codes. A key assumption, motivated by the physical processes, is that a given power station j potentially affects only hospitalization rates in zip codes that are downwind and within a given distance from that power station.

3.5 Equilibrium Interference

A fourth type of interference concerns anonymous interactions through marketplaces. The outcome for unit i may depend on the treatment assigned to unit i, but also on a second variable p whose value is partly determined by the set of potential outcomes and treatments for all units. For example, suppose that the potential outcomes are demand for a particular product. The treatment may make the product more attractive to customers, for example, through a discount or because it makes the product more visible through placement. Treating some individuals (giving them a discount) but not others may lead to an increase in overall demand. That in turn may lead to an increase in the price p of the product, which in turn may affect the demand for individuals in the control group. Heckman, Lochner and Taber (1998) discusses the possibilities of such equilibrium effects. Crépon et al. (2013) illustrate empirically that these equilibrium effects are present in an experimental evaluation of labor market programs. In their experiment, there are two levels of randomization. First, labor market segments (e.g., geographic areas) are randomly assigned to a share of treated individuals. Second, the unemployed individuals in each labor market segment are randomly assigned to a labor market program given the preassigned fraction treated. They find that in labor market segments where a larger share of unemployed individuals is treated the average effect of the treatment is smaller. Recently, Wager and Xu (2021), Munro, Wager and Xu (2021) have developed new experimental designs intended to deal with equilibrium effects.

4. CLUSTERED EXPERIMENTS

In contexts where cluster interference is assumed, a distinction is made between *direct* and *indirect* causal effects. A direct effect is the impact on unit i of unit i being treated *versus* untreated with the exposure of the other units within the group held constant. An indirect effect is the impact on an untreated unit i when the exposure level to treatment changes for other units (e.g., 50% of the other individuals are exposed versus 20%). *Herd immunity*, where unvaccinated individuals are less likely to become infected if they are in a community with a high vaccination rate, is a prominent example of an indirect effect.

In a setting with cluster interference, one natural approach is to do a *clustered experiment* where all units within a cluster are assigned to the same treatment, and to define all causal estimands at the cluster level. We can illustrate this using the bipartite graph setup from Zigler and Papadogeorgou (2021) as in Figure 2.

In the presence of cluster interference, effects within groups are not in general identified without further assumptions. Rosenbaum (2007) shows that if the exposure level is constant across all groups, the null hypothesis of "no effect" (both direct effect and indirect effects are zero) and "no direct effect" are indistinguishable, for

example, if every unit benefits equally, as long as some units are treated. Rosenbaum (2007) generates confidence statements for the total effect by inverting tests which are valid under either null hypothesis. In this setup, there are $B > 2$ blocks each containing $I > 2$ individuals. Further, within each block J individuals, $1 \leq J < I$, are selected at random to receive treatment. The randomization assignment $\boldsymbol{W}$ is now a matrix with B rows (one per block) and I columns (one for each of the individuals in each group), with W_{bi} indicating the treatment assignment (1 if assigned and 0 otherwise) of the ith unit in block b. It is assumed that the outcome $Y_{bi}(\boldsymbol{W})$ for unit i in block b depends on the random assignment $\boldsymbol{W}$ but only on the other units within a given block (i.e., the assignments along the same row).

Rosenbaum (2007) provides two approaches to producing valid confidence statements (i.e., tests which under *either* the null hypothesis of no effect *or* of no direct effect maintain the desired α level). First, he considers the Mann–Whitney/Wilcoxon statistic:

$$\Xi = \sum_{b=1}^{B} \sum_{i=1}^{I} \sum_{j=1}^{I} W_{bi}(1 - W_{bj}) 1_{Y_{bi}(\boldsymbol{W}) > Y_{bj}(\boldsymbol{W})}. \tag{4.1}$$

Thus, Ξ counts the number of pairs of units in the same block for which a treated unit had a higher response than a control; that is, Ξ is the sum of B Mann–Whitney statistics, one for each block, and $\Xi + BJ(J+1)/2$ is the sum of B Wilcoxon rank sum statistics. Then consider the same statistic $\widetilde{\Xi}$ but in a uniformity trial (also called an A/A test) where units are randomized to treatment groups but no intervention is applied. Under either null hypothesis, the distribution of $\widetilde{\Xi}$ is known (in particular it is the sum of B independent Mann–Whitney statistics under the null). Thus one may construct a confidence statement for $F = \Xi - \widetilde{\Xi}$ by showing that $\operatorname{pr}(F \geq \Xi - K_\alpha + 1) = 1 - \alpha$ where K_α is the appropriate critical value from the distribution of $\widetilde{\Xi}$ under the null. This implies the $1 - \alpha$ confidence statement that $V \geq 2(\Xi - K_\alpha + 1)/\{BJ(I-J)\}$ where $V = 2F/\{BJ(I-J)\}$ so that $V \in [-2, 2]$ and takes on value 0 when the (sharp) null of no effect is true.

As a second approach, Rosenbaum applies the same technique to the statistic of Mathisen (1943), Gart (1963), Gastwirth (1968), which counts the number of responses observed under treatment that exceed the median observed under control. Define the function $h(\cdot, \cdot)$ to count the number of times a treated response in block b exceeds the kth-order statistic of control responses in block b. Then write

$$H = \sum_{b=1}^{B} h\big(\boldsymbol{W}_b, Y_b(\boldsymbol{W})\big), \tag{4.2}$$

so that H counts the total number of times across all blocks that a treated response exceeded the kth-order statistic of control responses in the same block. Define $\widetilde{H}$ analogously except for a uniformity trial (i.e., an A/A test) note that H is observable in the experiment and $\widetilde{H}$ is unobservable but the distribution of $\widetilde{H}$ is known under the null. For $S = H - \widetilde{H}$, which is also unobservable, Rosenbaum (2007) shows how to construct a confidence statement by proving that $\operatorname{pr}(S \geq H - D_\alpha + 1) = 1 - \alpha$ thereby creating a $1 - \alpha$ confidence statement about S using H and D_α, which is the appropriate critical value from the (known) distribution of $\widetilde{H}$.

In Hudgens and Halloran (2008), an additional level of randomization is allowed: the groups (blocks) are assigned to two different regimes: a high-exposure regime, parametrized by ϕ, and a low-exposure regime parametrized by ψ (e.g., $\phi = 50\%$ of the units in the group exposed to treatment, and $\psi = 20\%$ exposed). Thus, the authors employ a two stage (or double) randomization mechanism where first the group is randomized to be within a specific regime and then the units within the group are randomized according to that regime. See Imai, Jiang and Malani (2021) for an example of such an experimental design.

The main result of Hudgens and Halloran (2008) is to provide, under few assumptions, unbiased estimates and conservative variance estimators for indirect and the direct effect, already considered in Rosenbaum (2007) and Halloran and Struchiner (1991), as well as other effects (called total and overall effects). Analogously to Rosenbaum (2007), the individual direct effect is defined as the difference in potential outcomes for an individual for a fixed regime ϕ between the individual being in treatment ($W_{bi} = 1$) versus control ($W_{bi} = 0$). The indirect individual effect is defined to be the difference in potential outcomes for an individual in control ($W_{bi} = 0$) in the high-treatment regime ϕ versus the low-treatment regime ψ. The total effect is then simply given by the sum of the direct and indirect effects. Finally, the overall effect is defined to be the difference between the average response under regime ϕ and average response under a regime ψ.

Specifically, for the overall effect, let $Y_{bi}(\boldsymbol{w}_b)$ be the potential outcome for individual i in block b under randomization for the bth block, and let $\operatorname{pr}_\phi(\boldsymbol{W}_b = \boldsymbol{w}_b)$ be the probability of the particular randomization for the block under regime ϕ. Then define the individual average effect under regime ϕ as

$$\overline{Y}_{bi}(\phi) \equiv \sum_{\boldsymbol{w}_b \in \Omega_{n_b}} Y_{bi}(\boldsymbol{w}_b) \operatorname{pr}_\phi(\boldsymbol{W}_b = \boldsymbol{w}_b),$$

where Ω_{n_b} is the space of possible randomization assignments of the n_b individuals in block b compatible with ϕ. Define the average of this over individuals in the block as $\overline{Y}_b(\phi) \equiv \sum_{i=1}^{n_b} \overline{Y}_{bi}(\phi)/n_b$. Similarly, define the average over blocks under regime ϕ to be $\overline{Y}(\phi) = \sum_{b=1}^{B} \overline{Y}_b(\phi)/B$.

Define the average outcomes under regime ψ similarly, then the overall effect is obtained as

$$\tau^O = \overline{Y}(\phi) - \overline{Y}(\psi),$$

which is the difference in expected outcomes under the two regimes, averaged over individuals in each block and averaged over blocks. Corresponding definitions of τ^D (direct effect), τ^I (indirect effect) and τ^T (total effect) can be obtained by replicating the procedure outlined above. Hudgens and Halloran (2008) provide unbiased estimators of these estimands. For brevity, we here outline the approach to obtain an unbiased estimate of the overall population average effect $\overline{\mathrm{CE}}^O$, but analogous results are obtained for the indirect, direct and total effects. To estimate $\overline{\mathrm{CE}}^O$, compute the average effect in each block, then average over blocks assigned to the same regime, and finally take the difference between the estimates for the two regimes. Formally, let $S_b = 1$ if block b was randomized to regime ψ and $S_b = 0$ if block b was randomized to regime ϕ. Then, if $S_b = 1$, $\widehat{Y}_b(\psi) \equiv \sum_{i=1}^{n_b} Y_{bi}(\boldsymbol{w}_b)/n_b$. Taking the average across blocks, which were randomized to ψ, gives

$$\widehat{Y}(\psi) = \frac{\sum_{b=1}^{B} \widehat{Y}_b(\psi) 1_{S_b=1}}{\sum_{b=1}^{B} 1_{S_b=1}}.$$

Defining similarly $\widehat{Y}(\phi)$, now summing over blocks with $S_b = 0$, the following is an unbiased estimate of τ^O:

$$\widehat{\tau}^O \equiv \widehat{Y}(\phi) - \widehat{Y}(\psi).$$

The second theoretical contribution is to provide conservative estimates of the variance of these quantities under an additional assumption that the potential outcomes only depend on the *fraction* of exposed individuals in the block and not the specific individuals. For example, the outcome for individual i in block b who receives the treatment is the same regardless of which $k-1$ other individuals are selected for treatment. This reduces potential outcomes per individual to n_b, rather than 2^{n_b} in the complete interference case or 1 in the no-interference case. Under this assumption, the variance estimator given by

$$\widehat{\mathrm{Var}}\big(\widehat{\tau}^O(\phi,\psi)\big) \equiv \frac{\widehat{\sigma}^2_M(\phi)}{N - B_\psi} + \frac{\widehat{\sigma}^2_M(\psi)}{B_\psi},$$

with

$$\widehat{\sigma}^2_M(\psi) \equiv \sum_{b=1}^{B} \frac{(\widehat{Y}_b(\psi) - \widehat{Y}(\psi))^2 S_b}{B_\psi}.$$

Here, $B_\psi = \sum_{b=1}^{B} S_b$ is the number of blocks allocated to regime ψ and $\widehat{\sigma}^2_M(\phi)$ is defined analogously. Then

$$E\big[\widehat{\mathrm{Var}}\big(\widehat{\tau}^O(\phi,\psi)\big)\big] \geq \mathrm{Var}\big(\widehat{\tau}^O(\phi,\psi)\big).$$

Equipped with an unbiased estimate and a conservative variance estimator, we can construct confidence intervals or perform hypothesis testing.

5. NETWORK EXPERIMENTS

The framework of Hudgens and Halloran (2008) is further generalized in Ugander et al. (2013), where instead of considering explicit disjoint clusters or groups (such as classrooms or households) the authors encode relationships between randomization units through a graph $G = (V, E)$. Vertices V are the randomization units and edges E between vertices indicate some relationship between the randomization units, which could cause violations of SUTVA. In the case that a graph can be completely decomposed into many connected components, this may be seen as similar to the clustered experiments of Rosenbaum (2007) and Hudgens and Halloran (2008). When the graph does not decompose into connected components, the authors propose an approach where exposure probabilities are computed from the graph G and then the treatment effect is estimated via application of the estimator from Horvitz and Thompson (1952) (H-T). A clustering approach where *groups* of nodes are collectively randomly exposed (or not) to an intervention is then selected to reduce the variance of the H-T estimator. More formally, they define an *exposure condition*, which states that there is an equivalence set Ω_i^x of random allocations $\boldsymbol{W}$ for which the outcome for individual i, $Y_i(\boldsymbol{W})$ is the same for all elements in the set. Namely, Ω_i^0 is the set of randomizations such that $Y_i(\boldsymbol{w}_0) = Y_i(\mathbf{0})$ for all $\boldsymbol{w}_0 \in \Omega_i^0$, where $\mathbf{0}$ is the randomization where no one receives an intervention. The set Ω_i^1 is defined analogously.

To make progress, the authors describe different *network exposure* conditions, which define these sets. The two main examples are "*Absolute k-neighborhood exposure*," which states that the equivalence sets Ω_i^0 and Ω_i^1 depend only on the individual unit i and the k neighbors receiving the treatment condition. That is, a fixed constant k is assumed to exist and a unit i is considered treated as long as its k neighbors are. "*Fractional q-neighborhood exposure*," states that Ω_i^0 and Ω_i^1 depend only on i and $q \cdot d$ neighbors receiving the treatment condition with d being the degree of the node i. That is, a fixed proportion q exists such that as long as $q \cdot d$ neighbors of i are are exposed then i is considered exposed. Under both of these assumptions, these exposure condition sets can be defined precisely by the underlying graph. Consequently, the exposure probabilities $\mathrm{pr}(\boldsymbol{W} \in \Omega_i^0)$ and $\mathrm{pr}(\boldsymbol{W} \in \Omega_i^1)$ can be computed from the graph G given a randomization distribution $\mathrm{pr}(\boldsymbol{W} = \boldsymbol{w})$. Then it is possible to estimate an unbiased average treatment effect by

$$\widehat{\tau}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i(\boldsymbol{W}) 1_{\boldsymbol{W}\in\Omega_i^1}}{\mathrm{pr}(\boldsymbol{W} \in \Omega_i^1)} - \frac{Y_i(\boldsymbol{W}) 1_{\boldsymbol{W}\in\Omega_i^0}}{\mathrm{pr}(\boldsymbol{W} \in \Omega_i^0)}\right).$$

In many of these examples of group or cluster randomized experiments, the cause of the spillover is some form of secondary actor or agency, which creates the interference. For example, in the schools case it could be the

teacher whose behavior changes due to the intervention on some students or in-class activities causing students to interact with each other. In other cases, physical proximity between units caused by being in a community is responsible for the interactions. As Hudgens and Halloran (2008) demonstrated, taking advantage of the second level of agency by double randomizing enables estimating the indirect and direct effects, which were unidentifiable in the case examined by Rosenbaum (2007). However, the group randomized approach in Hudgens and Halloran (2008) is still restrictive in design: each unit is assigned to precisely one group, the exposure level for each group cannot be 0% or 100%, and it is assumed that interference only occurs within the group. There are many settings, such as infectious diseases or marketplaces, where the pattern of interaction connects all units, and thus precludes dividing the population into groups that do not interfere.

6. CROSSOVER DESIGNS OR SWITCHBACK EXPERIMENTS

Crossover or *switchback* designs are a prominent alternative to simple RCTs with a long history in the econometric and statistical literature, including early applications in agricultural experiments. See Cochran (1939), Cochran and Cox (1948) for early discussions and Brown (1980), Cook and DeMets (2007) for a modern exposition. Traditionally, the main idea of switchback designs is to expose sequentially the same physical unit (e.g., a cow) to a random treatment (e.g., different types of feed). By measuring the unit's response to different treatments over time, one can directly estimate an average (over time) of the unit's individual level causal effect. In cases with substantial heterogeneity between units, this can lead to large gains in precision. Although traditionally the primary goal of such designs is to improve precision over standard RCTs where units receive the same treatment over time by leveraging within-unit across-time comparisons, in modern settings these designs have been adopted to reduce the effect of interference across units. Consider a setting with a population of customers that can be partitioned into market segments. Treating some customers (e.g., riders in an rideshare company) may affect other customers in the same market segment. In that case, using the market segment as the unit of analysis may be an attractive way to avoid biases arising from such interactions. However, this may leave the researcher with too few units (market segments) to do an experiment with sufficient power. In that case, a crossover experiment where the treatments change within a market segment over time may be a much more effective way to get precise estimates that are not contaminated by spillovers.

An important contribution in crossover designs that goes in this direction is Bojinov, Simchi-Levi and Zhao (2020).

The authors develop a framework for the optimal design and analysis of switchback experiments that explicitly allows for *carryover effects*. These are causal effects on units' future outcomes arising from past exposure to the active treatment. For example, if in a rideshare application there is a change in the matching algorithm it may take riders and drivers some time to adjust to the new environment. Thus, comparing times when the market segment is exposed to the new algorithm versus the old algorithm may lead to different results if the algorithms are switched every hour versus every week. Bojinov, Simchi-Levi and Zhao (2020) consider a class of experiments where units can be switched between treatment and control status at various time points. The optimal design question is the choice of the number and timing of these switch points. On the one hand, increasing the number of switch points can increase the bias from carryover effects. This suggests having just a single switch point so that the researcher can still exploit within unit comparison while minimizing carryover effects. However, in the absence of carryover effects, increasing the number of switch points so that the researcher can compare outcomes for the same physical unit at points in time that are close can improve the variance of the estimators, if there are smooth changes in the potential outcomes over time. The authors balance the possible bias arising from having many potential switch points so that carryover effects are prominent, and the variance that increases if there are few potential switch points. They establish formal results on the optimal design of switchback experiments, and propose a data-driven procedure for estimation and inference. Through their proposal, they derive optimal designs for switchback experiments, leading to the lowest variance among the most popular class assignment mechanisms. This is crucial as causal estimators from switchback experiments typically have large variances and can be impractical even in medium-sized experiments.

A related design corresponds to the case where units can only switch from the control group to the treatment group, but not backwards, in *staggered adoption designs* (Athey and Imbens (2022)) or *stepped-wedge* designs (Hemming et al. (2015)). This restriction often arises from context, where even short-term exposure to a new treatment persistently changes subsequent behavior. Xiong et al. (2019) study optimal designs in such settings, which often involve complex dynamic treatment effects.

A third, related, design that has a long tradition, especially in agricultural and industrial settings, is the *split-plot design* (Yates (1935), Brandt (1938), Jones and Nachtsheim (2009)). In such designs, there are multiple primary units (i.e., plots of land) that each consist of multiple units (subplots) that can separately be assigned to a treatment and for whom we can measure the outcome. Unlike in the crossover designs, there is no sequence to

the units within the primary units. The concern is that even in the absence of any treatment, the outcomes for the units that are part of the same primary unit are correlated. Taking such correlations into account by balancing assignments within primary units can lead to more effective experimental designs relative to completely randomized designs. Note that in the split-plot designs the concern is typically not about estimating spillovers and interference for such effects, which are our primary concerns. However, if there is concern regarding the presence of spillovers within the primary units, such designs could be helpful for inferring their magnitude. Again, these designs can be viewed as special cases of the MRDs discussed below.

7. EQUILIBRIUM EXPERIMENTS

Another interesting recent line of work has adopted the lens of *equilibrium effects* to address the problem of experimental design in the presence of cross-unit interference in marketplaces (Wager and Xu (2021)), Munro, Wager and Xu (2021). In this setting, naive standard RCTs can fail to provide unbiased estimates of causal effects of interest. To illustrate this consider, for example, the effect of tuition subsidy on college attendance (Heckman, Lochner and Taber (1998)). Suppose students choose to enroll in college partially because of the college wage premium (the increase in earnings due to attending college). An increase in the number of students attending college may lead to increased competition for a limited number of high wage jobs and, therefore, reduce the wage premium. Thus, a tuition subsidy directly reduces college cost, but may also indirectly reduce the incentives to attend college. In this context, consider estimating the effect of subsidies on college enrollment through a simple RCT that assumes SUTVA, with a small fraction of treated individuals. In such a setting, the equilibrium effects would be modest because few individuals are treated. As a result, the RCT would overestimate the actual effect of a uniform policy, because the experimental design fails to incorporate the overall equilibrium effects on the college wage premium that would be present if all individuals received the tuition subsidy.

Crépon et al. (2013) provide an empirical illustration of this problem. They designed a set of experiments where unemployed individuals were randomly assigned to labor market programs intended to help them find jobs. In different labor market segments, different shares of unemployed individuals were assigned to the treatment group. They found that in labor market segments with a larger fraction of treated individuals the treatment effect was smaller, consistent with the notion that the program made unemployed individuals more attractive to firms, but that it did not change the number of vacancies available.

The strategy proposed to overcome this issue is to carefully analyze how interference affects agents' behavior when the market segment has reached equilibrium. In turn, this analysis can guide experimental design and inform how estimation should be performed. The key assumption made here is that the interference across units can be captured via an intermediate outcome, such as a price, whose value is determined by an equilibrium condition.

In Wager and Xu (2021), the authors consider a centralized marketplace in which available demand is randomly allocated to a set of available suppliers, and the goal is to identify the optimal payment for the supply side (e.g., the tuition subsidy), that is, the scheme that maximizes the overall utility in the marketplace. Here, it is reasonable to imagine that interference is at play: if the platform doubles transaction payments for a random half of suppliers, these suppliers will be more inclined to participate and reduce the amount of demand available to the remaining suppliers, and thus, reduce their incentives to participate. To design an experiment that is able to capture this effect, the idea is to leverage the structure of the marketplace and selectively perturb the per-transaction payment available to the ith supplier by means of random noise. In turn, this informs the experimenter about the supplier's sensitivity to "local" changes in prices that marginally deviate from the market equilibrium. This marginal response is not directly useful to inform optimal policy design, since it does not take into account the shift in market equilibrium that would be observed if all suppliers were to receive the same payment change. The authors show, however, that when the number of suppliers is large, under additional mild assumptions, a simple mean-field model can be employed to propagate the findings of these "local perturbations" and form an estimate of the gradient of the platform's utility with respect to the payment schemes. These gradient estimates can then be used to optimize the payment scheme via generic stochastic first-order optimization methods, such as stochastic gradient descent.

In Munro, Wager and Xu (2021), the authors consider a similar setting, where agents in the market are characterized by an individual demand and supply curve that drives their behavior. The experimenter is interested in assessing the effect of some intervention (e.g., introduction of a subsidy) on the agents' preferences in equilibrium, when cross-unit interference is present. The authors analyze equilibrium effects assuming that this cross-unit interference is "restricted," in the sense that the agents' behavior only depends on whether they are exposed to treatment or control (through a random assignment mechanism that the experimenter gets to choose, for example, whether the individual enjoys a subsidy or not), and the market equilibrium prices, but not directly on other agents' treatment assignment. Under relatively mild assumptions, an experimental design in which one jointly randomizes over (i)

treatment assignments and (ii) equilibrium prices can be used to consistently estimate causal effects in the presence of interference in this setting.

8. MULTIPLE RANDOMIZATION DESIGNS

In this section, we review a novel class of experimental designs, *Multiple Randomization Designs* (MRDs), recently introduced in Bajari et al. (2021) and Johari, Li and Weintraub (2020). As we show, this is a rich class of designs that can be thought of as a generalization of many of the designs discussed in the previous sections.

Many experiments in modern settings involve multiple populations, with both outcomes and treatment assignments indexed by members of each population. For example, in marketplaces outcomes may be indexed by buyers and sellers, in rental marketplaces, outcomes may be indexed by renters and properties, in social media they may be indexed by content creators and subscribers, and in other settings we may have services and customers, movies and viewers, or market segments and time periods. In this paper, we use the terminology buyers and sellers to make the discussion specific, but the applications are more general. These settings are challenging for traditional experimental designs because members of both populations often act strategically in their interactions with each other. This leads to responses to the treatment assignment that are characterized by interference or spillovers between two different buyer–seller pairs. Here, we discuss experimental designs, where, by letting the buyer–seller pair be the experimental unit, and indexing the treatment at the pair level, we are able to to assess the presence and magnitude of such interference within a single experiment. We discuss the design as well as the analyses of such experiments.

To show the information content of such experiments, consider a marketplace with eight sellers and five buyers. We assign the pair corresponding to buyer i and seller j to a binary treatment, $W_{ij} \in \{\mathrm{C}, \mathrm{T}\}$. This leads to the assignment being a matrix, rather than a vector as in a traditional experiment. An example of a MRD treatment assignment matrix in this setting is shown in (8.1).

$$
\boldsymbol{W} = \overbrace{\begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array}}^{j \quad \text{(Sellers)}}
\begin{pmatrix}
\mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} \\
\mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} \\
\mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{T} & \mathrm{T} & \mathrm{T} & \mathrm{T} \\
\mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{T} & \mathrm{T} & \mathrm{T} & \mathrm{T} \\
\mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{T} & \mathrm{T} & \mathrm{T} & \mathrm{T}
\end{pmatrix}
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}
\Big\} i \ \text{(Buyers)}
\tag{8.1}
$$

In comparison, a conventional buyer randomized experiment where the buyer is the unit of analysis and assignment would have an assignment matrix $\boldsymbol{W}$ with all columns identical and rows consisting of either all T or all C. Symmetrically, a seller randomized experiment would correspond to an assignment matrix with all rows identical and columns consisting of either all T or all C.

There are two key advantages to allowing the experimental design to be a choice for the distribution of a matrix $\boldsymbol{W}$ jointly randomizing across buyers and sellers, rather than a choice for the distribution of a vector, randomizing only at the level of buyers or at the level of sellers.

First, by allowing for a richer set of potential assignment matrices $\boldsymbol{W}$, MRDs can be more efficient than conventional RCTs in answering standard questions. Specifically, under some conditions, designs where the fraction of treated buyer–seller pairs is the same in all columns and all rows are more efficient than either a standard buyer or a seller experiment. This insight is related to the motivation underlying stratification, Latin squares and factorial designs in agricultural experiments. An example of the use of MRDs for this purpose is in crossover designs. Here, the two populations that we can randomize over are units and time-periods. By randomizing both over units and time periods within the same experiment, one will in general get more precise estimates of the average causal effects than by randomizing only over units or only over time periods if there is heterogeneity by units or by time periods.

Second, and this is the more important of the two advantages, MRDs can generate information about spillovers and interference that cannot be learned from simple, single randomization designs such as standard RCTs. Specifically, MRDs allow for tests for the presence of spillovers and estimation of their magnitude by generating subpopulations of pairs of buyers and sellers that are *ex ante* comparable because of the assignment mechanism, but that *ex post* are different in their assignments, despite all being exposed to the control treatment.

Consider the example treatment assignment matrix in (8.1). There, the buyer–seller pairs, which receive the control treatment can be divided into three subsets. In the top right, we have the blue C buyer-seller pairs (buyers 1–2, and sellers 5–8). These pairs are exposed to the control treatment, but these sellers are exposed to the active treatment in their interactions with other buyers (buyers 3–5). In the bottom left, we have the green C buyer-seller pairs (buyers 3–5, and sellers 1–4). These pairs are again exposed to the control treatment, but now the buyers are exposed to the active treatment when interacting with other sellers (sellers 5–8). Finally, in the top left we have the red C buyer-seller pairs (buyers 1–2, and sellers 1–4). This set consist of pairs where both the buyers and the sellers are always in the control group. *Ex ante* these three sets of pairs are comparable because of the randomization. However, *ex post* they differ systematically in terms

of their general exposure. Comparing outcomes for these three sets of controls pairs is informative about the extent and nature of spillovers, and allows us to correct for spillovers under some assumptions on the structure of the interference.

The general MRD framework allows for much richer experiments than the simple one in (8.1). In principle, any distribution of assignments for the matrix $\boldsymbol{W}$ is now an experimental design where a standard RCT corresponds to cases where either the rows or columns of $\boldsymbol{W}$ are all identical. As an illustration, suppose we are concerned that there are both direct effects of the treatment on buyer-seller pair (i, j) from the treatment W_{ij}, but also indirect effects on the buyer–seller pair (i, j) from treatments W_{km} for *other* pairs (k, m). Not necessarily indirect effects for a generic pair (k, m), but possibly for pairs with either $k = i$ or $m = j$. To assess the presence and magnitude of, as well as disentangle, these direct and indirect effects consider the following two-stage experimental design. We first randomize sellers into two groups $(X_j^{\mathrm{S}} \in \{B, S\})$, for all sellers j. For the first group of sellers, those with $X_j^{\mathrm{S}} = B$, we conduct a buyer experiment (with assignment for buyer i equal to $W_i^{\mathrm{B}} \in \{\mathrm{C}, \mathrm{T}\}$), and conduct a seller experiment for the second set of sellers, those with $X_j^{\mathrm{S}} = S$ (with assignment for seller j equal to $W_j^{\mathrm{S}} \in \{\mathrm{C}, \mathrm{T}\}$). An example of the assignment matrix under such a design is given in (8.2). Interpreting and analyzing such experiments requires careful consideration of the types and mechanisms for interference and spillovers.

(8.2)

	Buyer Exp.			Seller Exp.					W_i^{B}	i (Buyers)
(Sellers) $j \to$	1	2	3	4	5	6	7	8		
$X_j^{\mathrm{S}} \to$	B	B	B	S	S	S	S	S		
$W_j^{\mathrm{S}} \to$				T	C	T	C	C	$\downarrow$	$\downarrow$
$\boldsymbol{W} =$	C	C	C	T	C	T	C	C	C	1
	C	C	C	T	C	T	C	C	C	2
	T	T	T	T	C	T	C	C	T	3
	C	C	C	T	C	T	C	C	C	4
	T	T	T	T	C	T	C	C	T	5

Another example of a MRD is important in the context of clustering where it has been considered in Holtz et al. (2020). Suppose we have a population of items, either physical items, or listings, which may be purchased by individuals. The researcher is interested in the effect of some discount or informational treatment, but there is concern that a simple randomized experiment may be subject to bias as a result of interference. One alternative to a standard randomized experiment is a cluster-randomized experiment described above. Another alternative is to construct the clusters, but then randomly assign the clusters to one of two sets of clusters. In the first set a standard clustered experiment will be conducted, which by construction will be free from bias arising from within-cluster interference. In the second set of clusters a completely randomized experiment will be conducted, which will be subject to bias from spillovers within clusters. Like the other MRDs, this will allow the researcher to assess the presence and magnitude of within-cluster bias. An example of an assignment matrix for such an experiment is given below.

(8.3)

	Individual			1	2	3	4	5	6	7	8
	Cluster										
	Listing										
	1	i	A	C	C	C	C	C	C	C	C
	2	i	A	C	C	C	C	C	C	C	C
	3	ii	A	T	T	T	T	T	T	T	T
	4	ii	A	T	T	T	T	T	T	T	T
$\boldsymbol{W} =$	5	iii	B	C	C	C	C	C	C	C	C
	6	iii	B	T	T	T	T	T	T	T	T
	7	iv	B	C	C	C	C	C	C	C	C
	8	iv	B	T	T	T	T	T	T	T	T
	9	iv	B	T	T	T	T	T	T	T	T
	10	iv	B	C	C	C	C	C	C	C	C

There are I buyers, indexed by $i \in \mathbb{I} := \{1, \ldots, I\}$. There are J sellers, indexed by $j \in \mathbb{J} := \{1, \ldots, J\}$. Both I and J may be large, although under random assignment it is possible to derive finite sample results that are valid irrespective of the absolute and relative magnitude of I and J.

Over a fixed period of time, say a week or a month, there is for each buyer-seller pair a measure of engagement, which we use as the outcome, denoted by Y_{ij}. The outcome could be the amount of money spent by a buyer with a particular seller, or an indicator that buyer and seller have interacted.

Formally, we consider a binary intervention that can be assigned at the level of pair (i, j) of a buyer i and a seller j, denoted by $W_{ij} \in \{\mathrm{C}, \mathrm{T}\}$, with $\boldsymbol{W}$ the $I \times J$ matrix with element W_{ij}. There may also be buyer, seller or buyer-seller-specific characteristics that are correlated with the outcomes. These characteristics can improve the precision of the experiments, but we ignore their presence for the moment.

An example assignment matrix is shown in Equation (8.2), where the columns correspond to the J sellers, and

the rows correspond to the I buyers.

(8.4)
$$\begin{array}{c} \text{(Sellers) } j \\ \text{(Buyers) } i \end{array} \quad \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \boldsymbol{W} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} & \mathrm{C} \\ \mathrm{C} & \mathrm{T} & \mathrm{T} & \mathrm{T} & \mathrm{C} & \mathrm{C} & \mathrm{T} & \mathrm{T} \\ \mathrm{C} & \mathrm{T} & \mathrm{C} & \mathrm{C} & \mathrm{T} & \mathrm{C} & \mathrm{C} & \mathrm{T} \\ \mathrm{T} & \mathrm{C} & \mathrm{C} & \mathrm{T} & \mathrm{T} & \mathrm{C} & \mathrm{C} & \mathrm{T} \\ \mathrm{T} & \mathrm{T} & \mathrm{C} & \mathrm{C} & \mathrm{T} & \mathrm{C} & \mathrm{C} & \mathrm{T} \end{pmatrix}, \end{array}$$

We refer to a distribution $p : \mathbb{W} \mapsto [0, 1]$ as an *experimental design*, where $\mathbb{W}$ is the set of values that the matrix $\boldsymbol{W}$ can take on. Many conventional RCTs impose substantial restrictions on these assignment matrices. More specifically, conventional designs restrict all elements within rows (or within columns) of the matrix to be identical, with variation only between rows (or columns). As discussed earlier, experimental designs that allow for variation in the treatment both within rows and columns can be more efficient in answering questions that can already be answered using conventional designs; and, more importantly, they can be informative about the presence and magnitude of spillover effects, thus answering questions that conventional designs cannot address. (Whenever the distinction matters and is not clear from context, we refer to a draw of an assignment matrix from $\mathbb{W}$ as $\boldsymbol{w}$, and to $\boldsymbol{W}$ as a matrix-valued random variable.)

The buyer–seller pair (i, j) is the unit of observation as well as the unit of analysis. In principle each potential outcome can depend on the full matrix $\boldsymbol{W}$, $Y_{ij}(\boldsymbol{W})$, with $\boldsymbol{Y}(\boldsymbol{W})$ denoting the full $I \times J$ matrix of potential outcomes for a given $I \times J$ matrix of assignments $\boldsymbol{W}$. Particular interesting values for the assignment matrix are $\boldsymbol{W} = \mathbf{T}$, with typical element $T_{ij} = \mathrm{T}$ for all $i \in \mathbb{I}$ and $j \in \mathbb{J}$, corresponding to all pairs/interactions being exposed to the new treatment, and $\boldsymbol{W} = \mathbf{C}$, with typical element $C_{ij} = \mathrm{C}$ for all $i \in \mathbb{I}$ and $j \in \mathbb{J}$, corresponding to all pairs being exposed to the control treatment. The realized outcomes correspond to the potential outcomes evaluated at the actual assignment: $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{W})$.

In a buyer (or seller) experiment, the potential outcomes are typically the sum over all sellers (buyers) of the pair-specific outcome:

$$Y_i^{\mathrm{B}}(\boldsymbol{W}) = \sum_{j=1}^{J} Y_{ij}(\boldsymbol{W}), \quad \text{and}$$

$$Y_j^{\mathrm{S}}(\boldsymbol{W}) = \sum_{i=1}^{I} Y_{ij}(\boldsymbol{W}).$$

In conventional RCTs the unit that we randomize over is the buyer (or the seller), and the standard estimand is in that case the the effect of universal policies, that is, exposing all interactions to the treatment versus exposing none of the interactions, or the difference between the average potential outcomes given treatment and control:

$$\tau^{\mathrm{B}} = \frac{1}{I} \sum_{i=1}^{I} \big(Y_i^{\mathrm{B}}(\mathbf{T}) - Y_i^{\mathrm{B}}(\mathbf{C}) \big).$$

In a MRD set up the corresponding estimand is the average difference between $Y_{ij}(\mathbf{T}) - Y_{ij}(\mathbf{C})$,

$$\tau = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \big(Y_{ij}(\mathbf{T}) - Y_{ij}(\mathbf{C}) \big). \tag{8.5}$$

These estimands are not directly comparable because for the buyer (seller) experiments we sum over all pairs, and so the estimands are measured on a different scale. To make the estimands identical, it is useful to focus on lifts, the difference in average outcomes scaled by the average control outcome. For MRD experiments, this would lead to

$$\theta = \frac{\frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij}(\mathbf{T}) - Y_{ij}(\mathbf{C}))}{\frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij}(\mathbf{C})},$$

which is identical to the corresponding estimand for a conventional buyer (seller) experiment.

9. EXPERIMENTS IN DOUBLE-SIDED MARKETPLACES

In this section, we discuss the benefits of adopting MRDs for the estimation of the direct and indirect (spillover) causal effects in a specific context. We use an example of an experiment in a video streaming service to illustrate the role of modelling assumptions on the nature of the interference. Key is that once we have interference, even randomization cannot completely free us from the need to make untestable assumptions.

9.1 Experimental Setup

Consider a video streaming service where viewers $i = 1, \ldots, I$ can choose to watch streaming content provided by content creators (creators for short) $j = 1, \ldots, J$. When they choose to watch a video, viewers are shown advertisements (ads). We are interested in estimating the effect of doubling, for all viewers and for all content creators, the number of ads shown before the content is shown (Pre-Roll-Ads; from hereon, PRAs) on a metric Y_{ij}, measured at the viewer-creator level. For the sake of concreteness, let Y_{ij} denote the amount of time (*e.g.*, the number of minutes) that viewer i spends watching the content of creator j during a fixed period of time (e.g., a week, or a month). Experimenters can expose each viewer/creator pair (i, j) to either the default policy (a single PRA) or to the new policy (two PRAs). At the discretion of the creators, additional ads can be displayed throughout a stream (Mid-Roll-Ads; from hereon, MRAs).

Creators observe, but do not control, how many PRAs each viewer sees prior to seeing the content. In contrast, creators do decide how many MRAs will be displayed during the video. Each viewer watching the content in principle sees the same number of MRAs, rolled out at the exact same time for every viewer on the stream, that is, 120 seconds into the content, although viewers of course may choose to stop watching the video at any moment in time. A key feature of this set up is that although the creator has complete control over the timing and the number of MRAs, they are indirectly constrained by the fact that viewers may not return if they are shown too many MRAs. This is further complicated by the possibility that the response of the viewer to the number and timing of the MRAs may be affected by the number of PRAs, which are not controlled by the content creator, and which may vary by viewer.

9.1.1 *Potential outcomes and estimands of interest.* Given the potential outcomes $Y_{ij}(\boldsymbol{w})$, we are interested in the effect of switching all viewers from one to two PRAs for content from any of the creators, τ in Equation (8.5). To make progress in our analysis, we now consider assumptions about the potential outcomes $Y_{ij}(\mathbf{T})$ and $Y_{ij}(\mathbf{C})$ in Equation (8.5) that impose limits on the structure of the spillovers.

ASSUMPTION 9.1. Assume that the entire population gets to see only one PRA—that is, the service adopts $\mathbf{C}$. The time that viewers would spend watching creators' contents is drawn i.i.d. across viewers and creators from a "base" distribution F_0:

$$\begin{aligned} Y_{ij}(\mathbf{C}) &\overset{\text{iid}}{\sim} F_0, \\ \mathbb{E}\big[Y_{ij}(\mathbf{C})\big] &= \mu_0 \in \mathbb{R}, \quad \text{and} \\ \operatorname{Var}\big(Y_{ij}(\mathbf{C})\big) &= \sigma_0^2 < \infty. \end{aligned} \tag{9.1}$$

While the results in Bajari et al. (2021) on which this example builds do not require a distributional assumption as in Equation (9.1), Assumption 9.1 is helpful for the simulations discussed in Section 9.5. In practice, we expect the distribution F_0 to induce a very sparse matrix $\boldsymbol{Y}(\mathbf{C})$: that is, most streams are watched by a few viewers, and most viewers watch only a few streams. To capture this behavior, many options are possible. The simplest is to assume that F_0 is a mixture distribution with a mass point at zero, where for some $\pi \in (0, 1)$ and distribution $F_{0,+}$ with support on $\mathbb{R}_+$, so that for all y,

$$F_0(y) = (1-\pi)\mathbf{1}_{\{y=0\}} + \pi F_{0,+}(y). \tag{9.2}$$

Another generalization that may be important in practice is to relax the i.i.d. assumption and allow for correlation between outcomes for different creators for the same viewer and for different viewers for the same creator.

9.1.2 *Potential outcomes under treatment, full exposure.* We next posit assumptions to relate potential outcomes $Y_{ij}(\boldsymbol{T})$ to potential outcomes $Y_{ij}(\mathbf{C})$—*i.e.*, to model the relationship between what happens in the absence of experimentation to what would happen if the treatment policy were rolled out where some viewers were exposed to two PRAs for some creators.

In our setting, interference relates to the fact that by exposing a viewer/creator pair (i, j) to assignment W_{ij} we might affect, directly or indirectly, the outcome $Y_{i'j'}$ for some other viewer/creator pair (i', j'). Here, we call a *direct effect* a change in outcome for a viewer/creator pair (i, j) caused *directly* by a change in exposure for that pair, W_{ij}. Conversely, an *indirect effect* (spillover) is a change in the outcome for viewer/creator pair (i, j) caused *indirectly* by the change in exposure for another pair $(i', j') \neq (i, j)$.

When full exposure $\mathbf{T}$ is implemented, so that the new policy of showing two PRAs is rolled out to all viewer/creator pairs, two things happen. First, a direct effect of the increased ad exposure. That is, the viewers' experience is affected due to the increase in PRAs. We expect this to cause fewer streams being watched overall. We capture this effect via a "dispersion" parameter, $\delta > 0$. In turn, this direct effect triggers a content creators (indirect) reaction: as a consequence of the viewers' dispersion, some content creators may choose to reduce their MRAs, to reduce the impact of the increased PRAs on the viewer behavior. We capture this effect via a "premium" parameter $\alpha > 0$.

ASSUMPTION 9.2. Conditionally on $\boldsymbol{Y}(\mathbf{C})$, potential outcomes $Y_{ij}(\mathbf{T})$ are a stochastic function of $Y_{ij}(\mathbf{C})$:

$$Y_{ij}(\mathbf{T}) | \boldsymbol{Y}(\mathbf{C}) \sim Y_{ij}(\mathbf{C})(1 + A_{ij} - \Delta_{ij}), \tag{9.3}$$

with $A_{ij} \overset{\text{iid}}{\sim} F_A$, $\Delta_{ij} \overset{\text{iid}}{\sim} F_\Delta$, independent of $Y_{ij}(\mathbf{C})$, and $\mathbb{E}[A_{ij}] = \alpha$, $\mathbb{E}[\Delta_{ij}] = \delta$. For both random variables we assume finite second moments.

With these assumptions in place, it follows by the law of large numbers that as $I \times J \to \infty$, the average treatment effect of switching all viewer/creator pairs from one to two PRAs is

$$\tau \xrightarrow{p} \mu_0(1 + \alpha - \delta) - \mu_0 = \mu_0(\alpha - \delta). \tag{9.4}$$

Given this model for the outcomes and the treatment effects, what can different experimental designs tell us?

9.2 Viewer-Randomized Experiments

One option is for the experimenter to run a viewer-randomized experiment. In the simplest case, let $W_i^v \sim \text{Bernoulli}(p^v)$ for some $p^v \in (0, 1)$, and let $W_{ij} = W_i^v$ for all j. That is, regardless of the creator j, viewer i will always see the same number of PRAs. We expect this intervention to have a direct effect on viewers, causing dispersion: viewers who are randomly selected to always see

two PRAs (viewers with $i : W_i^v = 1$) reduce their engagement and they spend less time watching (any) content. In turn, we expect two indirect effects to be triggered.

First, content creators will react: content creators see that some of their viewers are exposed to more PRAs, and they are aware of the risk of losing such viewers. As a consequence, some content creators might strategically reduce MRAs in response. The reason for this response is that the higher number of PRAs seen on average will increase the revenue-per-viewer of content creators, and since more PRAs lead to viewers abandoning the content creator, they might choose to be conservative and reduce the MRAs (e.g., by targeting the same level of total average Ads-per-viewer before the experiment). This is a negative externality for the creators. This effect will be small if p^v is small and there are few viewers with multiple PRAs, but it will be increasing with p^v, and might be substantial if p^v is large.

Second, if content creators adapt their behavior, all viewers will see fewer MRAs. Crucially for the spillovers, this includes viewers assigned to the control group, who were not exposed to additional PRAs. This might lead these viewers in the control group to enjoy the content more, and not interrupt it mid-way—that is, spend more time watching the stream. Therefore, the experience of the control viewers is affected by the exposure of the treatment viewers. This is a positive externality for the viewers.

As a result a comparison of average outcomes for treated viewers and control viewers in a simple viewer experiment will be biased for the average effect τ. In this particular case, with $p^v < 1$, the content creators will reduce the MRAs less than they would do if all viewers were exposed to the PRAs. Consequently, both viewers in treatment and control will be streaming with more interruptions than they would do if all viewers were exposed to the treatment. In addition, because the control viewers find their average MRAs reduced by the creator response, they will be streaming more than they would do if none of the viewers were exposed to the treatment. Thus, the comparison between treated and control viewers will overestimate (in absolute value) the negative effect of the treatment, and may suggest that the increase in PRAs has a bigger effect on engagement than it actually will have.

9.2.1 *Data generating mechanism.* We next posit a simple model for the effect of a viewer randomized experiment on potential outcomes, when we expect interference to act as we described above. Let $\boldsymbol{W}^{(\mathrm{vr})}$ be the (random) matrix of random assignments in a viewer randomized experiment.

ASSUMPTION 9.3. In a simple viewer experiment, if viewer i is in control—$\boldsymbol{W}_{ij}^{(\mathrm{vr})} = 0$ for all j—then:

$$Y_{ij}(\boldsymbol{W}^{(\mathrm{vr})}) | \{\boldsymbol{Y}(\mathbf{C}), \boldsymbol{W}_{ij}^{(\mathrm{vr})} = 0\} \stackrel{d}{=} Y_{ij}(\mathbf{C})(1 + A_{ij}),$$

If viewer i is in treatment, $\boldsymbol{W}_{ij}^{(\mathrm{vr})} = 1$, then for all j:

$$Y_{ij}(\boldsymbol{W}^{(\mathrm{vr})}) | \{\boldsymbol{Y}(\mathbf{C}), \boldsymbol{W}_{ij}^{(\mathrm{vr})} = 1\}$$
$$\stackrel{d}{=} Y_{ij}(\mathbf{C})(1 + A_{ij} - \Delta_{ij}),$$

where $A_{ij} \overset{\text{iid}}{\sim} F_A$, $\Delta_{ij} \overset{\text{iid}}{\sim} F_\Delta$.

Notice that the notation in the simple model in Assumption 9.3 does not capture that the indirect effect is likely to be a function of p^v, the fraction of treated viewers. A simple way to incorporate this dependency would be to let $\alpha = \alpha(p^v)$ depend on the randomization proportion p^v. For ease of exposition, we do not consider this dependency here.

Consider the usual difference-in-means estimator: letting $\mathcal{I}_w^{(\mathrm{vr})} = \{(i,j) : W_{ij} = w\}$,

$$\hat{\tau}^{(\mathrm{vr})} = \sum_{(i,j)\in\mathcal{I}_1^{(\mathrm{vr})}} \frac{Y_{ij}(\boldsymbol{W}^{(\mathrm{vr})})}{|\mathcal{I}_1^{(\mathrm{vr})}|} - \sum_{(i,j)\in\mathcal{I}_0^{(\mathrm{vr})}} \frac{Y_{ij}(\boldsymbol{W}^{(\mathrm{vr})})}{|\mathcal{I}_0^{(\mathrm{vr})}|}.$$

Then, by independence of $Y_{ij}(\mathbf{C})$, A_{ij}, and Δ_{ij},

$$\mathbb{E}\big[\hat{\tau}^{(\mathrm{vr})}\big] = (1 + \alpha - \delta)\mu_0 - (1 + \alpha)\mu_0 = -\delta\mu_0,$$

which differs from the target τ by $\alpha\mu_0$. This simple analysis suggests a viewer-randomized experiment would capture the target ATE τ only when $\alpha = 0$ ($\mu_0 = 0$ is not an interesting case if the outcome is a nonnegative variable).

9.3 Content-Creator-Randomized Experiments

Another option is for the experimenter to run a content-creator-randomized experiment. In the simplest case, let $W_j^c \sim \text{Bernoulli}(p^c)$ for some $p^c \in (0,1)$, and let $W_{ij} = W_j^c$ for all i. That is, the content of creator j contains the same number of PRAs for all viewers.

The direct effect of a creator-randomized experiment is that creators assigned to display more PRAs (creators with $j : W_j^c = 1$) become comparatively less appealing. As a result, after realizing that a content creator in the treatment group displays more PRAs than other content creators, viewers might get annoyed or bored, and might switch to substitute streams from creators which are in the control group, or reduce the amount of content they stream.

As a consequence of these direct effects, the intervention triggers two indirect effects. First, as a consequence of viewers changing their preferences towards creators in the control group, some creators in the treatment group may respond strategically and choose to reduce their MRAs to make their content more appealing. Second, content creators in the control group may increase their MRAs in response to the increase in the number of viewers they attract as a result of them having fewer PRAs than the treatment group. The result of both these indirect effects is that the decrease in engagement for the treated content creators is partly offset via their response by reducing MRAs.

9.3.1 *Data generating mechanism.* Similarly to Assumption 9.3, we assume that we can capture the way in which potential outcomes for the treated and control groups behave in a content-creator-randomized experiment, where we model the interference as described above. Let $\boldsymbol{W}^{(\text{ccr})}$ be the matrix of random assignments in a content creator experiment.

ASSUMPTION 9.4. In a content-creator experiment, if viewer/content pair (i, j) is in control — $\boldsymbol{W}_{ij}^{(\text{ccr})} = 0$,

$$Y_{ij}(\boldsymbol{W}^{(\text{ccr})}) | \{\boldsymbol{W}_{ij}^{(\text{ccr})} = 0, \boldsymbol{Y}(\mathbf{C})\} \stackrel{d}{=} Y_{ij}(\mathbf{C})(1 + \eta_{ij}).$$

Here $\eta_{ij} = \frac{S_{ij}}{\pi} \frac{p^c}{1-p^c}(\Gamma_i - \alpha)$. $S_{ij} \sim \text{Bernoulli}(\pi)$ denotes whether viewer i would switch over to view content j if that content was in control due to other content being treated. Here, π is assumed fixed across viewers, creators. $\frac{p^c}{1-p^c}(\Gamma_i - \alpha)$ is the switchover "bonus" obtained by creators in control: for every viewer i, a fraction p^c of the J content creators are treated, and for every treated content creator $(\Gamma_i - A_{ij}) \times 100\%$ of the time viewer i will switch over to a control content creator. This time gets "distributed" across all the streams in control, on average $J(1 - p^c)$.

If creator j is in treatment, $\forall i$, $\boldsymbol{W}_{ij}^{(\text{ccr})} = 1$, then:

$$\begin{aligned} &Y_{ij}(\boldsymbol{W}^{(\text{ccr})}) | \{\boldsymbol{W}_{ij}^{(\text{ccr})} = 1, \boldsymbol{Y}(\mathbf{C})\} \\ &\quad \stackrel{d}{=} Y_{ij}(\mathbf{C})(1 + A_{ij} - \Gamma_i - \Delta_{ij}), \end{aligned}$$

with $A_{ij} \overset{\text{i.i.d.}}{\sim} F_A$, $\Gamma_i \overset{\text{i.i.d.}}{\sim} F_\Gamma$ with mean γ, what we call the switchover parameter. Here, $\Gamma_i \in (0, 1)$ is a random fraction representing the "elasticity" of viewer i to the policy change (from control to treatment) when watching (any) streaming j quantifying the fraction of time they would switch to content from other creators if stream j was treated.

Letting, for $w \in \{0, 1\}$, $\mathcal{I}_w^{(\text{ccr})} = \{(i, j) : W_{ij} = w\}$, the usual difference-in-means estimator is

$$\hat{\tau}^{(\text{ccr})} = \sum_{(i,j) \in \mathcal{I}_1^{(\text{ccr})}} \frac{Y_{ij}}{|\mathcal{I}_1^{(\text{ccr})}|} - \sum_{(i,j) \in \mathcal{I}_0^{(\text{ccr})}} \frac{Y_{ij}}{|\mathcal{I}_0^{(\text{ccr})}|}.$$

Then

$$\begin{aligned} &\mathbb{E}[\hat{\tau}^{(\text{ccr})}] \\ &\quad = \left[(1 + \alpha - \gamma - \delta) - \left(1 + \frac{p^c}{1 - p^c}(\gamma - \alpha)\right)\right]\mu_0 \\ &\quad = \frac{1}{1 - p^c}(\alpha - \gamma)\mu_0 - \delta\mu_0, \end{aligned}$$

which differs from the target $\tau = \mu_0(\alpha - \delta)$ by

$$\mathbb{E}[\hat{\tau}^{(\text{ccr})}] - \tau = \left(\frac{\alpha p^c - \gamma}{1 - p^c}\right)\mu_0.$$

In this case, there are two sources for the bias. It depends on the viewer elasticity with respect to the policy change γ and on the fraction of treated creators, p^c. Viewers prefer streams that are not treated because of a switchover option that would not be present if the service were to roll out the intervention to everyone.

Assumption 9.3 and Assumption 9.4 are similar to the restricted interference assumption of Bajari et al. (2021). Their restriction implies that we can encode the interference via a distributional "shift" with respect to what would have happened in the absence of an intervention, that is, the case in which we would have applied a treatment arm to the entire population. We emphasize that while the simple Assumptions 9.3 and 9.4 might fail to capture the full complexity of dynamics (e.g., switchovers might depend on randomization proportions p^c, or treating viewers might change the sparsity pattern of the observations, etc.), they provide a natural starting point to model the presence of spillovers, and the pitfalls associated with conventional single randomization designs.

9.4 Double Randomized Experiments

A third option is for the experimenter to implement a simple MRD, specifically, a double randomized experiment, as discussed in Section 8. We randomly split the viewers into two groups, $W_i^v \sim \text{Bernoulli}(p^v)$ for some $p^v \in (0, 1)$, and we randomly split the creators into two groups, $W_j^c \sim \text{Bernoulli}(p^c)$ for some $p^c \in (0, 1)$. The assignment for viewer/creator pair (i, j) is then $W_{ij} = W_i^v W_j^c$ for all i, j (this nests the viewer experiment if $p^v = 1$ and the content-creator experiment if $p^c = 1$). We let $\boldsymbol{W}^{(\text{dbr})}$ denote the corresponding binary assignment matrix. See Figure 3 for a bipartite graph version of this assignment mechanism.

In this case, remember from Section 8 that we partition the space of outcomes into four (not two) groups:

- consistent controls (c): viewer-creator pairs (i, j) for which only one PRA is displayed, and where viewer i never experiences more than one PRA on other streams, and other viewers watching the stream of creator j never experience more than one PRA.
- inconsistent viewers (iv): pairs (i, j) for which only one PRA is displayed, and other viewers watching the stream of creator j never experience more than one PRA. However, viewer i sees more than one PRA for other creators' streams.
- inconsistent creator (ic): pairs (i, j) for which only one PRA is displayed, and other viewers watching the stream of creator j never experience more than one PRA. However, other viewers watching the same stream of creator j see more than one PRA for that same stream.
- treatment (t): pairs (i, j) for which two PRAs are displayed. By construction, viewer i *does* get to see only

one PRA for other creators' streams, and some other viewer interested in watching the stream of creator j only get to see one PRA.

Direct effects of a double randomized experiment mirror previously discussed viewer and content-creator randomized experiments. A viewer with $W_i^v = 0$ will not experience any direct-effect from taking part in the experiment. However, a viewer with $W_i^v = 1$ may reduce overall streaming and may partially switch toward content that shows fewer PRAs.

Indirect effects of a double randomized experiment also mirror previously discussed viewer and content-creator randomized experiments. For content creator j with $W_j^c = 1$, some viewers i get to see substitute streams j' with fewer PRAs. Hence, j might reduce their MRAs to make their content more appealing to i. For all viewers, that is, either $W_i^v = 0$ or $W_i^v = 1$, if content creators j for which $W_j^c = 1$ reduce their MRAs as discussed above, this will make viewer i more likely to watch that content over content for which $W_i^c = 0$—which do not update their MRA policy.

To simplify analysis, we adopt the local interference assumptions of Bajari et al. (2021). Let $\boldsymbol{S}^{(\mathrm{dbr})}$ be the matrix of types induced by the double randomized experiment,

$$S_{ij}^{(\mathrm{dbr})} = \begin{cases} \mathrm{c} & \text{if } W_i^v = 0 \text{ and } W_j^c = 0, \\ \mathrm{iv} & \text{if } W_i^v = 1 \text{ and } W_j^c = 0, \\ \mathrm{ic} & \text{if } W_i^v = 0 \text{ and } W_j^c = 1, \\ \mathrm{t} & \text{if } W_i^v = 1 \text{ and } W_j^c = 1. \end{cases} \tag{9.5}$$

ASSUMPTION 9.5. Under a simple double-randomized experiment as described above, the following holds:

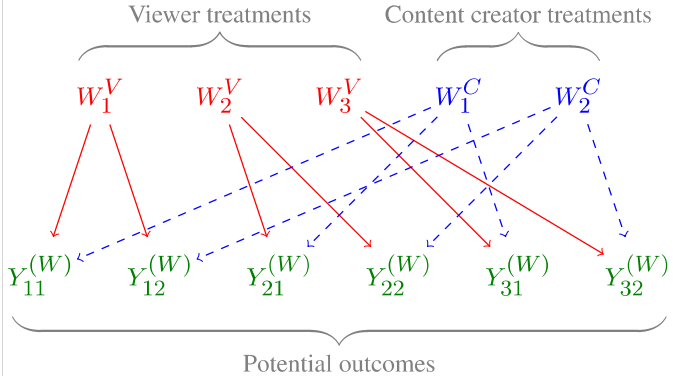

FIG. 3. *Bipartite graph representation ($I = 3$, $J = 2$) of a simple double randomization design. Viewers $i \in \{1, \ldots, I\}$ have treatment indicators $W_i^V \in \{0, 1\}$. Content creators $j \in \{1, \ldots, J\}$ have treatment indicators $W_j^C \in \{0, 1\}$. Treatment assignment for each (viewer, creator) pair (i, j) is $W_{ij} = W_i^V W_j^C$ so that it is treated iff both treatment indicators are 1. Potential outcome for pair (i, j) is $Y_{ij}^{(W)} = Y_{ij}(S_{ij}^{\mathrm{dbr}}) = Y_{ij}(type(W_i^V, W_j^C))$, where 'type' is given by equation (9.5).*

- if $S_{ij}^{(\mathrm{dbr})} = \mathrm{c}$, the experience of the viewer/creator tuple coincides with $\mathbf{C}$:

$$Y_{ij}(\boldsymbol{W}^{(\mathrm{dbr})}) | \{S_{ij}^{(\mathrm{dbr})} = \mathrm{c}, \boldsymbol{Y}(\mathbf{C})\} \stackrel{d}{=} Y_{ij}(\mathbf{C}).$$

- if $S_{ij}^{(\mathrm{dbr})} = \mathrm{iv}$ ($W_i^v = 1$ and $W_j^c = 0$), the experience of the tuple (ij) coincides with that of an untreated pair in a content creator experiment:

$$Y_{ij}(\boldsymbol{W}^{(\mathrm{dbr})}) | \{S_{ij}^{(\mathrm{dbr})} = \mathrm{iv}, \boldsymbol{Y}(\mathbf{C})\} \stackrel{d}{=} Y_{ij}(\mathbf{C}) + \eta_{ij}\mu_0.$$

- if $S_{ij}^{(\mathrm{dbr})} = \mathrm{ic}$ ($W_i^v = 0$ and $W_j^c = 1$), the experience of the pair (ij) coincides with that of an untreated pair in a viewer creator experiment:

$$Y_{ij}(\boldsymbol{W}^{(\mathrm{dbr})}) | \{S_{ij}^{(\mathrm{dbr})} = \mathrm{ic}, \boldsymbol{Y}(\mathbf{C})\} \stackrel{d}{=} (1 + A_{ij})Y_{ij}(\mathbf{C}).$$

- if $S_{ij}^{(\mathrm{dbr})} = \mathrm{t}$ ($W_i^v = 1$ and $W_j^c = 1$),

$$Y_{ij}(\boldsymbol{W}^{(\mathrm{dbr})}) | \{S_{ij}^{(\mathrm{dbr})} = \mathrm{t}, \boldsymbol{Y}(\mathbf{C})\}$$
$$\stackrel{d}{=} (1 - \Delta_{ij} - \Gamma_i - \Delta_{ij} + A_{ij})Y_{ij}(\mathbf{C}).$$

Notice that in Assumption 9.5 we ignore "second-order" feedbacks, for example, the fact that an untreated content j might become much more popular overall, and that would impact via, for example, the search feed how likely it is for viewer i to watch it.

Let $\mathcal{I}_s^{(\mathrm{dbr})} = \{(ij) : S_{ij}^{(\mathrm{dbr})} = s\}$. For types s, s',

$$\hat{\tau}_{s,s'} = \sum_{(ij)\in\mathcal{I}_s^{(\mathrm{dbr})}} \frac{Y_{ij}(\boldsymbol{W}^{(\mathrm{dbr})})}{|\mathcal{I}_s^{(\mathrm{dbr})}|} - \sum_{(ij)\in\mathcal{I}_{s'}^{(\mathrm{dbr})}} \frac{Y_{ij}(\boldsymbol{W}^{(\mathrm{dbr})})}{|\mathcal{I}_{s'}^{(\mathrm{dbr})}|}.$$

We can make all the $\binom{4}{2}$ binary comparisons for the types. These are listed in Table 1.

The effects α, δ, γ, see Table 2 for their interpretation, can be identified from pairwise comparisons. First, let

$$\hat{\mu}_0 = \hat{Y}_{\mathrm{c}} = \frac{1}{|\mathcal{I}_{\mathrm{c}}|} \sum_{(ij)\in\mathcal{I}_{\mathrm{c}}} Y_{ij}(\boldsymbol{W}^{(\mathrm{dbr})}).$$

TABLE 1
Pairwise comparisons in the simple double randomized experiment

$\mathbb{E}\left[\hat{\tau}_{s,s'}\right]$	c	iv	ic
iv	$\frac{p_c}{1-p_c}(\gamma - \alpha)\mu_0$		
ic	$\alpha\mu_0$	$\frac{\alpha - p_c\gamma}{1-p_c}\mu_0$	
t	$(\alpha - \delta - \gamma)\mu_0$	$\left[-\delta + \frac{\alpha-\gamma}{1-p_c}\right]\mu_0$	$(-\delta - \gamma)\mu_0$

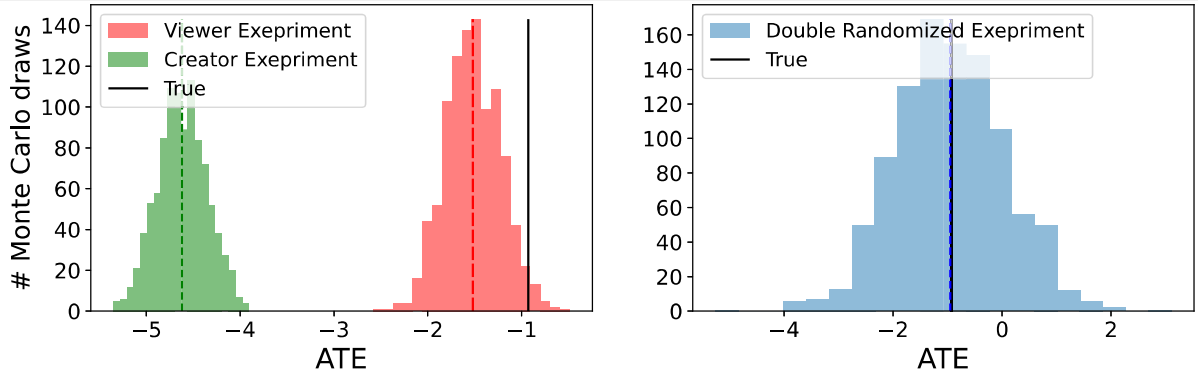

FIG. 4. *Estimation of the ATE with standard and double randomized experiments.*

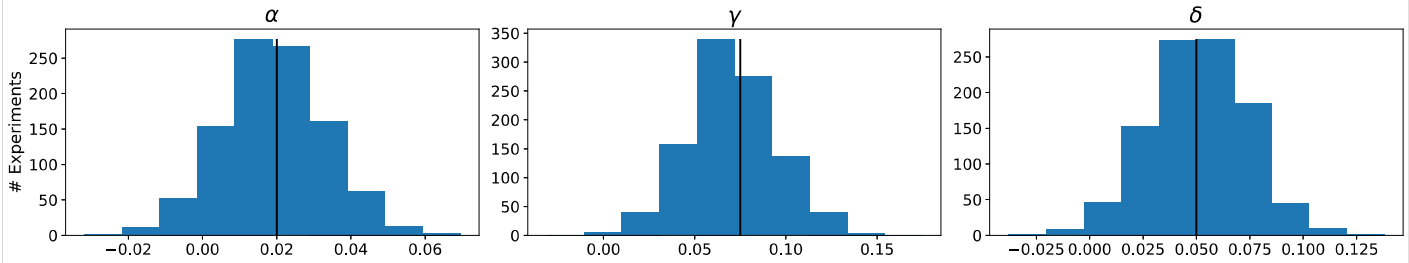

FIG. 5. *Estimation of parameters α, γ, δ. In each subplot, the solid black line identifies the underlying true value of the parameter.*

To identify α, we can use the fact that $\mathbb{E}[\hat{\tau}_{\text{ic,c}}] = \alpha\mu_0$ and let $\hat{\alpha} = \hat{\tau}_{\text{ic,c}}/\hat{\mu}_0$. To identify γ, we can use the estimate $\hat{\alpha}$ and simply solve for γ in the equation defining $\hat{\tau}_{\text{iv,c}}$, that is, $\hat{\gamma} = \hat{\alpha} + (\hat{\tau}_{\text{iv,c}})/(\frac{p_c}{1-p_c}\hat{\mu}_0)$.

With these two estimates in hand, δ can be directly estimated using, for example, $\hat{\tau}_{\text{t,ic}}$.

Because we have six entrants in the set of treatment effects $\tau_{s,s'}$, we can actually test some of the modeling assumptions, or relax some of the assumptions underlying the model.

9.5 Quantifying Spillover Effects: A Simulation Study

To demonstrate our framework, we here run a simulation study. We draw the potential outcomes in control, $\boldsymbol{Y}(\mathbf{C})$ from a mixture distribution as in Equation (9.2), with $\pi = 0.10$, $I = 4000$ viewers, $J = 100$ content creators. The distribution $F_{0,+}$ is a gamma distribution with mean 300 minutes/month per active content, and standard deviation 50.

To quantify the effectiveness of the different experiments, we reran experiments by keeping potential outcomes fixed, and re-allocating units (viewers in a viewer experiment, creators in a creator experiment, or viewer-creators pairs in a double randomized experiment) over $N = 1000$ Monte Carlo reruns. In our simulation, we let A_{ij} be i.i.d. draws from a beta distribution with mean 0.02 and standard deviation 0.1. Similarly, Δ_{ij} are i.i.d. draws from a beta distribution with mean 0.05 and standard deviation 0.1. Last, Γ_i are i.i.d. draws from a beta distribution with mean 0.075 and standard deviation 0.1. Results are shown in Figure 4. As expected, the simple randomization strategies fail to correctly identify the ATE, while the double randomization scheme allows an unbiased estimate for the ATE to be obtained. As a byproduct of this, we are also able to estimate the parameters α, γ, δ, see Figure 5.

TABLE 2
Dynamics in multisided intervention

Parameter	*Interpretation*
α	*Premium parameter*: extra time viewers spend watching a stream when MRAs are reduced
δ	*Dispersion parameter*: decrease in time viewers spend watching a stream when PRAs are increased
γ	*Switchover parameter*: fraction of time viewers spend watching substitute streams with fewer PRAs

10. CONCLUSIONS

We have discussed the recent literature on experimental designs in the presence of interference. These designs include a new class of experimental designs that is intended to allow the researcher to learn about spillovers in certain settings. The key feature of the settings we consider is that we have multiple populations and can assign treatments to pairs (or tuples) with each tuple element corresponding to the identity of a member of each population. This allows

for much richer designs than the conventional designs. We demonstrated how such designs can lead to more precise inferences about standard estimands such as the overall average effect of the intervention and that they can generate information about spillovers that conventional designs cannot reveal. We also propose methods for estimation and inference.

ACKNOWLEDGMENTS

We are grateful for discussions with Susan Athey, John Geweke, Matt Taddy, and Johan Ugander and for comments by participants in the following conferences: CODE@MIT, NABE (2019), ASSA (2020) and ICSDS (2022). This research was carried out as part of Imbens' and Richardson's consulting relationship with Amazon. Brian Burdick contributed while working for Amazon.[1]

REFERENCES

ARONOW, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociol. Methods Res.* **41** 3–16. MR3190698 https://doi.org/10.1177/0049124112437535

ARONOW, P. M. and SAMII, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.* **11** 1912–1947. MR3743283 https://doi.org/10.1214/16-AOAS1005

ATHEY, S., ECKLES, D. and IMBENS, G. W. (2018). Exact p-values for network interference. *J. Amer. Statist. Assoc.* **113** 230–240. MR3803460 https://doi.org/10.1080/01621459.2016.1241178

ATHEY, S. and IMBENS, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *J. Econometrics* **226** 62–79. MR4348786 https://doi.org/10.1016/j.jeconom.2020.10.012

BACKSTROM, L. and KLEINBERG, J. (2011). Network bucket testing. In *Proceedings of the* 20*th International Conference on World Wide Web* 615–624.

BAJARI, P., BURDICK, B., IMBENS, G. W., MASOERO, L., MCQUEEN, J., RICHARDSON, T. and ROSEN, I. M. (2021). Multiple randomization designs. arXiv preprint. Available at arXiv:2112.13495.

BASSE, G. W., FELLER, A. and TOULIS, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106** 487–494. MR3949317 https://doi.org/10.1093/biomet/asy072

BOJINOV, I., SIMCHI-LEVI, D. and ZHAO, J. (2020). Design and analysis of switchback experiments. Available at SSRN 3684168.

BOND, R. M., FARISS, C. J., JONES, J. J., KRAMER, A. D., MARLOW, C., SETTLE, J. E. and FOWLER, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature* **489** 295–298.

BRANDT, A. (1938). Tests of significance in reversal or switchback trials. *Iowa Agric. Home Econ. Exp. Stat. Res. Bull.* **21** 1.

BROWN, B. W. JR. (1980). The crossover experiment for clinical trials. *Biometrics* 69–79.

COCHRAN, W. (1939). Long-term agricultural experiments. *Suppl. J. R. Stat. Soc.* **6** 104–148.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0474575

COCHRAN, W. G. and COX, G. M. (1948). *Experimental Designs*. Wiley, New York, NY.

COOK, T. D. and DEMETS, D. L. (2007). *Introduction to Statistical Methods for Clinical Trials*. CRC Press/CRC, Boca Raton.

CRÉPON, B., DUFLO, E., GURGAND, M., RATHELOT, R. and ZAMORA, P. (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Q. J. Econ.* **128** 531–580.

FISHER, R. A. (1937). *The Design of Experiments*. Oliver & Boyd, Edinburgh, London.

GART, J. J. (1963). A median test with sequential application. *Biometrika* **50** 55–62. MR0156424 https://doi.org/10.1093/biomet/50.1-2.55

GASTWIRTH, J. L. (1968). The first-median test: A two-sided version of the control median test. *J. Amer. Statist. Assoc.* **63** 692–706. MR0240933

GUPTA, S., KOHAVI, R., TANG, D., XU, Y., ANDERSEN, R., BAKSHY, E., CARDIN, N., CHANDRAN, S., CHEN, N. et al. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explor. Newsl.* **21** 20–35.

HALLORAN, M. E., and STRUCHINER, C. J. (1991). Study Designs for Dependent Happenings. *Epidemiology.* **2** 331–338.

HECKMAN, J. J., LOCHNER, L. and TABER, C. (1998). General-equilibrium treatment effects: A study of tuition policy. *Amer. Econ. Rev.* **88** 381–386.

HEMMING, K., HAINES, T. P., CHILTON, P. J., GIRLING, A. J. and LILFORD, R. J. (2015). The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *BMJ* **350**.

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

HOLTZ, D., LOBEL, R., LISKOVICH, I. and ARAL, S. (2020). Reducing interference bias in online marketplace pricing experiments. arXiv preprint. Available at arXiv:2004.12489.

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 https://doi.org/10.1198/016214508000000292

IMAI, K., JIANG, Z. and MALANI, A. (2021). Causal inference with interference and noncompliance in two-stage randomized experiments. *J. Amer. Statist. Assoc.* **116** 632–644. MR4270009 https://doi.org/10.1080/01621459.2020.1775612

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge Univ. Press, Cambridge.

JOHARI, R., LI, H. and WEINTRAUB, G. (2020). Experimental design in two-sided platforms: An analysis of bias. arXiv preprint. Available at arXiv:2002.05670.

JONES, B. and NACHTSHEIM, C. J. (2009). Split-plot designs: What, why, and how. *J. Qual. Technol.* **41** 340–361.

KOHAVI, R., CROOK, T., LONGBOTHAM, R., FRASCA, B., HENNE, R., FERRES, J. L. and MELAMED, T. (2009). Online experimentation at Microsoft. *Data Mining Case Stud.* 11.

MANSKI, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Rev. Econ. Stud.* **60** 531–542. MR1236836 https://doi.org/10.2307/2298123

MATHISEN, H. C. (1943). A method of testing the hypothesis that two samples are from the same population. *Ann. Math. Stat.* **14** 188–194. MR0009285 https://doi.org/10.1214/aoms/1177731460

MUNRO, E., WAGER, S. and XU, K. (2021). Treatment effects in market equilibrium. arXiv preprint. Available at arXiv:2109.11647.

[1]Please send correspondence to imbens@stanford.edu or thomasr@u.washington.edu.

NEYMAN, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472.

OGBURN, E. L. and VANDERWEELE, T. J. (2014). Causal diagrams for interference. *Statist. Sci.* **29** 559–578. MR3300359 https://doi.org/10.1214/14-STS501

PAPADOGEORGOU, G., MEALLI, F. and ZIGLER, C. M. (2019). Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics* **75** 778–787. MR4012083 https://doi.org/10.1111/biom.13049.

POLLMANN, M. (2020). Causal inference for spatial treatments. arXiv preprint, arXiv:2011.00373.

POUGET-ABADIE, J., AYDIN, K., SCHUDY, W., BRODERSEN, K. and MIRROKNI, V. (2019). Variance reduction in bipartite experiments through correlation clustering. *Adv. Neural Inf. Process. Syst.* **32**.

ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537 https://doi.org/10.1198/016214506000001112.

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

UGANDER, J., KARRER, B., BACKSTROM, L. and KLEINBERG, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD* '13 329–337. Association for Computing Machinery, New York, NY, USA.

VANDERWEELE, T. J., TCHETGEN, E. J. T. and HALLORAN, M. E. (2014). Interference and sensitivity analysis. *Statist. Sci.* **29** 687–706. MR3300366 https://doi.org/10.1214/14-STS479.

WAGER, S. and XU, K. (2021). Experimenting in equilibrium. *Manage. Sci.* https://doi.org/10.1287/mnsc.2020.3844.

WU, C. J. and HAMADA, M. S. (2011). *Experiments*: *Planning*, *Analysis*, *and Optimization* **552**. Wiley, New York.

XIONG, R., ATHEY, S., BAYATI, M. and IMBENS, G. W. (2019). Optimal experimental design for staggered rollouts. http://dx.doi.org/10.2139/ssrn.3483934.

YATES, F. (1935). Complex experiments. *Suppl. J. R. Stat. Soc.* **2** 181–247.

ZIGLER, C. M. and PAPADOGEORGOU, G. (2021). Bipartite causal inference with interference. *Statist. Sci.* **36** 109–123. MR4194206 https://doi.org/10.1214/19-STS749.

HALLORAN, E. M. and STRUCHINER, J. (1991). Study Designs for Dependent Happenings. *Epidemiology* **2** 331-338. https://doi.org/10.1097/00001648-199109000-00004.

THE ORIGINAL REFERENCE LIST

The list of entries below corresponds to the original Reference section of your article. The bibliography section on previous page was retrieved from MathSciNet applying an automated procedure. Please check both lists and indicate those entries which lead to mistaken sources in automatically generated Reference list.

P. M. Aronow. A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16, 2012.

P. M. Aronow and C. Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.

S. Athey, D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.

S. Athey and G. W. Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79, 2022.

L. Backstrom and J. Kleinberg. Network bucket testing. In *Proceedings of the 20th International Conference on World Wide Web*, pages 615–624, 2011.

P. Bajari, B. Burdick, G. W. Imbens, L. Masoero, J. McQueen, T. Richardson, and I. M. Rosen. Multiple randomization designs. *arXiv preprint arXiv:2112.13495*, 2021.

G. W. Basse, A. Feller, and P. Toulis. Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494, 2019.

I. Bojinov, D. Simchi-Levi, and J. Zhao. Design and analysis of switchback experiments. *Available at SSRN 3684168*, 2020.

R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

A. Brandt. Tests of significance in reversal or switchback trials. *Iowa Agriculture and Home Economics Experiment Station Research Bulletin*, 21(234):1, 1938.

B. W. Brown Jr. The crossover experiment for clinical trials. *Biometrics*, pages 69–79, 1980.

W. Cochran. Long-term agricultural experiments. *Supplement to the Journal of the Royal Statistical Society*, 6(2):104–148, 1939.

W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, third edition, September 1977.

W. G. Cochran and G. M. Cox. *Experimental Designs*. Wiley, 1948.

T. D. Cook and D. L. DeMets. *Introduction to Statistical Methods for Clinical Trials*. Chapman and Hall/CRC, 2007.

B. Crépon, E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora. Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *The Quarterly Journal of Economics*, 128(2):531–580, 2013.

R. A. Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

J. J. Gart. A median test with sequential application. *Biometrika*, 50(1/2):55–62, 1963.

J. L. Gastwirth. The first-median test: A two-sided version of the control median test. *Journal of the American Statistical Association*, 63(322):692–706, 1968.

S. Gupta, R. Kohavi, D. Tang, Y. Xu, R. Andersen, E. Bakshy, N. Cardin, S. Chandran, N. Chen, D. Coey, et al. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, 21(1):20–35, 2019.

J. J. Heckman, L. Lochner, and C. Taber. General-equilibrium treatment effects: A study of tuition policy. *The American Economic Review*, 88(2):381–386, 1998.

K. Hemming, T. P. Haines, P. J. Chilton, A. J. Girling, and R. J. Lilford. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, 350, 2015.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–970, 1986.

D. Holtz, R. Lobel, I. Liskovich, and S. Aral. Reducing interference bias in online marketplace pricing experiments. *arXiv preprint arXiv:2004.12489*, 2020.

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

M. Hudgens and E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, pages 832–842, 2008.

K. Imai, Z. Jiang, and A. Malani. Causal inference with interference and noncompliance in two-stage randomized experiments. *Journal of the American Statistical Association*, 116(534):632–644, 2021.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

R. Johari, H. Li, and G. Weintraub. Experimental design in two-sided platforms: An analysis of bias. *arXiv preprint arXiv:2002.05670*, 2020.

B. Jones and C. J. Nachtsheim. Split-plot designs: What, why, and how. *Journal of Quality Technology*, 41(4):340–361, 2009.

R. Kohavi, T. Crook, R. Longbotham, B. Frasca, R. Henne, J. L. Ferres, and T. Melamed. Online experimentation at Microsoft. *Data Mining Case Studies*, page 11, 2009.

C. Manski. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3):531–542, 1993.

H. C. Mathisen. A method of testing the hypothesis that two samples are from the same population. *The Annals of Mathematical Statistics*, 14(2):188–194, 1943.

E. Munro, S. Wager, and K. Xu. Treatment effects in market equilibrium. *arXiv preprint arXiv:2109.11647*, 2021.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1923/1990.

E. L. Ogburn and T. J. VanderWeele. Causal diagrams for interference. *Statistical Science*, 29(4):559–578, 2014.

G. Papadogeorgou, F. Mealli, and C. M. Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787, 2019.

M. Pollmann. Causal inference for spatial treatments. *arXiv preprint arXiv:2011.00373*, 2020.

J. Pouget-Abadie, K. Aydin, W. Schudy, K. Brodersen, and V. Mirrokni. Variance reduction in bipartite experiments through correlation clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 329–337, New York, NY, USA, 2013. Association for Computing Machinery.

T. J. VanderWeele, E. J. T. Tchetgen, and M. E. Halloran. Interference and sensitivity analysis. *Statistical Science*, 29(4):687, 2014.

S. Wager and K. Xu. Experimenting in equilibrium. *Management Science*, 2021.

C. J. Wu and M. S. Hamada. *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons, 2011.

R. Xiong, S. Athey, M. Bayati, and G. W. Imbens. Optimal experimental design for staggered rollouts. *Available at SSRN*, 2019.

F. Yates. Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, 2(2):181–247, 1935.

C. M. Zigler and G. Papadogeorgou. Bipartite causal inference with interference. *Statistical Science*, 36(1):109, 2021.

META DATA IN THE PDF FILE

Following information will be included as pdf file Document Properties:

Title : Experimental Design in Marketplaces
Author : Patrick Bajari, Brian Burdick, Guido W. Imbens, Lorenzo Masoero, James McQueen, Thomas S. Richardson, Ido M. Rosen
Subject : Statistical Science, 2023, Vol. 0, No. 00, 1-19
Keywords: Experimental design, causal inference, online experimentation, multiple randomization designs, two-sided marketplaces
Affiliation:

THE LIST OF URI ADDRESSES

Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are indicated as ERROR. Please check and update the list where necessary. The e-mail addresses are not checked – they are listed just for your information. More information can be found in the support page:
http://www.e-publications.org/ims/support/urihelp.html.

200 https://imstat.org/journals-and-publications/statistical-science/ [2:pp.1,1] OK
301 https://www.imstat.org [2:pp.1,1] Moved Permanently // https://www.imstat.org/
--- mailto:imbens@stanford.edu [4:pp.1,1,17,17] Check skip
--- mailto:thomasr@u.washington.edu [4:pp.1,1,17,17] Check skip
301 http://arxiv.org/abs/arXiv:2112.13495 [2:pp.17,17] Moved Permanently
301 http://arxiv.org/abs/arXiv:2004.12489 [2:pp.18,18] Moved Permanently
301 http://arxiv.org/abs/arXiv:2002.05670 [2:pp.18,18] Moved Permanently
301 http://arxiv.org/abs/arXiv:2109.11647 [2:pp.18,18] Moved Permanently
301 http://arxiv.org/abs/arXiv:2011.00373 [2:pp.18,18] Moved Permanently