A Randomization Test for the Spillover Effect in Two-Stage Observational Data

by

Tyler Mansfield

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Arts

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alan Hubbard, Chair
Assistant Professor Sam Pimentel
Professor Peng Ding

Spring 2023

The thesis of Tyler Mansfield, titled A Randomization Test for the Spillover Effect in Two-Stage Observational Data, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

A Randomization Test for the Spillover Effect in Two-Stage Observational Data

Abstract

A Randomization Test for the Spillover Effect in Two-Stage Observational Data

by

Tyler Mansfield

Master of Arts in Biostatistics

University of California, Berkeley

Professor Alan Hubbard, Chair

The spillover effect refers to the impact of treatment on those units not directly assigned to the intervention. While the classic Neyman-Rubin causal model assumes that no such effect exists, countless examples exist in practice where such an assumption may either be obviously erroneous or in need of validation before proceeding with classic methods. We present a matched randomization test for detecting the presence of arbitrary spillover effects in two-stage observational data. We show this test is valid by theoretical proof and also by simulation. We conclude by discussing extensions to effect estimation and other future avenues of research.

Contents

List of Figures

Acknowledgments

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Chapter 1

Introduction

1.1 Motivation

One of the most fundamental assumptions in the classic Neyman-Rubin causal model is the stable unit treatment value assumption, often termed SUTVA [1]. One key aspect of this assumption declares that there is no interference between units, e.g. the effect of treatment on one unit is not affected by the treatment status of other units in the study. While this assumption may have validity in many scenarios, there are countless other research questions in causal inference for which this assumption is unrealistic. One classic example is vaccine efficacy since the likelihood of an individual contracting an illness is affected by the vaccination status of both the individual and the individuals with whom they associate. This idea has been termed the 'spillover effect', which broadly refers to how the outcome of a single unit is affected by the treatment status of all other units. As noted by Hudgens and Halloran [2], there are times when this effect is a nuisance, while there are other circumstances when this spillover effect may be the parameter of interest.

Various researchers have extended the classic Neymen-Rubin model to allow for general interference patterns between units. A common experimental design is that of the two-stage study, where units are organized into clusters and interference is only expected within-cluster but not between. In the vaccine example, this might include the assumption that an individual's contraction of an illness is dependent on the vaccination status of those in their household/community, but not on those outside this cluster.

Work has been done to create model-free randomization methods to test the null hypothesis of no spillover effect in two-stage designs [3] [4]. However, these methods operate under the assumption of randomization at both the "access to treatment" cluster-level and "assignment to treatment" individual-level. This leaves a critical gap for the observational setting in which an entire cluster's access to treatment or an individual's treatment status may be non-randomized. Methods, such as the one we develop here, will allow researchers to identify whether or not a spillover effect is present in observational data. Questions such as "Does an individual's intervention status affect (either positively or negatively) those

who did not receive this intervention?" arise in countless fields such as medicine, education, sociology, and politics.

In this thesis, we will first briefly review prior research efforts related to identifying spillover effects within the framework of causal inference. In Chapter 2, we will build upon the framework of Basse [4] by constructing a matched randomization method for testing the presence of spillover effects in two-stage observational data. Chapter 3 will demonstrate the efficacy and power of these methods through simulation. Chapter 4 will conclude with a discussion of limitations and potential future work.

1.2 Prior Work

Spillover Effects

The notion of a "spillover effect" is rather broad and can take many forms. When viewed from the lens of a randomized experiment, this effect refers to the unintended effects or influences of an intervention on individuals that are not directly targeted by the intervention. This definition extends easily to the observational data setting with the only modification being that individuals choose to participate in some intervention as opposed to being randomly assigned. In the classical context of causal inference, the spillover effect can pose challenges in accurately estimating the causal effects of an intervention, as it can lead to biases and confounders that may affect the validity of causal inference [5]. In practice, the spillover effect can be the result of many potential mechanisms such as social contagion, network effects, or supply limitations of an intervention.

Some of the earliest spillover effect research efforts began by seeking to recover the direct causal effect of an intervention in the presence of interference such as in Sobel [5] and Hong [6]. The framework of Hudgens and Halloran [2] decomposed the causal effect of treatment into the "direct causal effect" resulting from a unit's treatment and the "indirect (or spillover) effect" resulting from the treatment status of other units. This distinction allowed for this spillover effect to be directly quantified and various estimators were constructed that gave researchers the ability to measure such an effect.

One challenge of the work described above is that arbitrary interference patterns often make the task of effect identification and estimation infeasible. Additional assumptions are frequently required before any valid inference can be made, such as the stratified interference assumption [2] [7]. In this thesis, our randomization test will seek to preserve the most general formulation of interference (within our framework of two-stage observational studies) so that we can detect spillover patterns regardless of the mechanism or form. The Discussion section will show what additional gains can be made on this methodology if assumptions about the form of interference can be reasonably assumed.

Two-stage Observational Studies

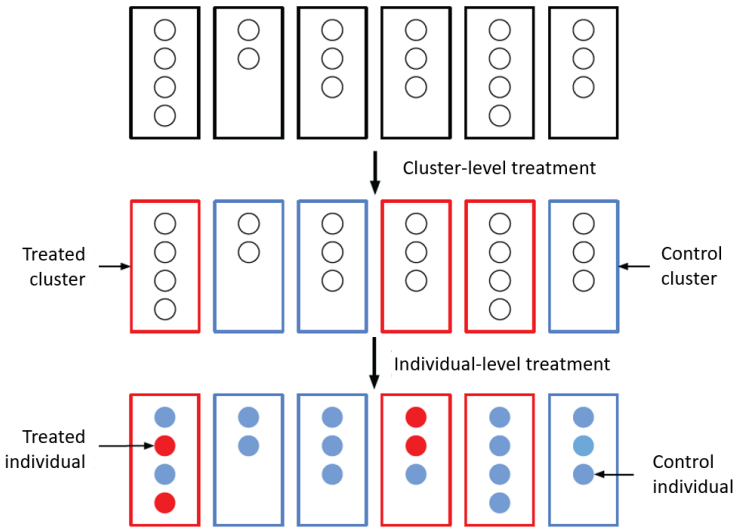

Figure 1.1: A Two-Stage Observational Design

Rectangles represent clusters and circles represent individual units within the cluster. Clusters receive either treatment (such as access to a medical intervention for units in-cluster) or a control condition (no access). Each individual in a treated cluster selects an additional individual-level treatment (whether or not they use the treatment). The second row of this figure shows a possible allocation of cluster-level treatment, and the third row shows a possible allocation of individual-level treatment. Figure by Aurelien Bibaut.

Consider two-stage observational studies in which treatments are assigned to clusters of individuals, and then units within treated clusters also make an individual decision about whether or not to participate in treatment. (Figure 1.1) This set-up mirrors several real-world scenarios, such as a unique medical treatment only available in certain communities or an educational intervention offered to students at a subset of all schools.

Under a randomized framework, Hudgens and Halloran [2] noted that the spillover effect could be assessed by comparing control units in control clusters to control units in treated clusters, since the no interference assumption guarantees that these two groups should be equivalent. Athey et al [3] and Basse [4] proposed randomization methods to assess the null hypothesis of no interference in this randomized two-stage setting. In these methods, cluster and individual treatment assignments were permuted and control units in treated and control clusters were compared. However, both methods relied on the fundamental assumption of randomization. when extending this method to observational data, a challenge arises for randomization inference: if we permute the cluster-level treatment assignments, how do we know which units in the original control clusters would have adopted treatment had they been given access to the treatment? This is a fundamental challenge that this method will

seek to address.

Barkley et al. [8] proposed weighting methods for this observational two-stage setting. However, one notorious drawback to weighting methods is the high variance and instability of effect estimation when propensity scores are close to either 0 or 1. By building upon Basse's interference framework [4], we will be able to construct a design-based randomization test for interference that addresses the shortcomings of weighting methods.

Chapter 2

Methodology

2.1 Notation

Consider a sample of N clusters from a population of clusters. Each cluster i has an associated vector $(W_i, \mathbf{L}_i)$ where W_i is a cluster-level treatment indicator and $\mathbf{L}_i$ is the observed multidimensional vector of cluster-level covariates. Let $\theta_i = P(W_i = 1|\mathbf{L}_i)$ be our cluster-level propensity score. Each cluster i also contains n_i individuals, represented by a length-n_i vector $\mathbf{Z}_i$ and matrix $\mathbf{X}_i$ with n_i rows, giving individual-level treatment indicators and observed covariates respectively. When $W_i = 0$, then we will have $\mathbf{Z}_i = \mathbf{0}$ (no one in a control cluster receives treatment) and Ω_i represents the set of all possible values of $\mathbf{Z}_i$ when $W_i = 1$. Conditional on $W_i = 1$ we denote the individual-level propensity score by $\pi_i(\mathbf{z}_i) = P(\mathbf{Z_i} = \mathbf{z_i}|W_i = 1, \mathbf{L}_i, \mathbf{X_i})$. Note that here we allow for interference in treatment status, meaning one unit's probability of treatment is not necessarily independent of whether other units in the cluster receive treatment.

For the purpose of this thesis, we will assume that we are in the scenario with no unmeasured confounding, meaning that θ_i and $\pi_i(\mathbf{z}_i)$ are correctly identified. See the Discussion section for comments related to confounding and sensitivity.

Defining spillover effects requires us to modify the notation classically used for potential outcomes. For cluster i, define length-n_i vectors $\mathbf{Y_i}(\mathbf{0}, 0)$ and $\mathbf{Y_i}(\mathbf{z_i}, 1)$ for all $\mathbf{z_i} \in \{\mathbf{z_i} : \pi(\mathbf{z_i}) > 0\}$. These potential outcomes depend both on the cluster-level treatment (the final index) and the individual-level treatment values for all individuals in cluster i, which allows for arbitrary spillover structures within clusters, albeit with no spillovers across clusters as in Sobel [5]. To test for spillover effects, the following null hypothesis is used:

$$Y_{ij}(\mathbf{0}, 0) = Y_{ij}(\mathbf{z}_i, 1) \text{ for all } \mathbf{z}_i \text{ such that } z_{ij} = 0 \tag{2.1}$$

In words, there are no spillover effects of any kind for any individual under the null, although individuals may be affected by their own treatments.

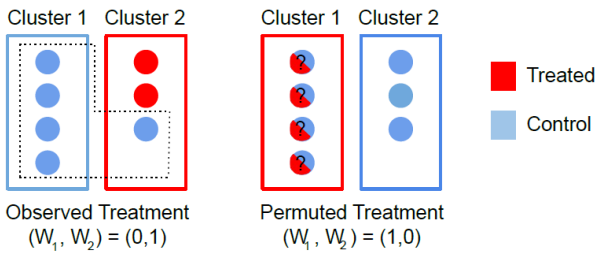

Figure 2.1: Permutation Challenge for Observational Data

Two paired clusters are shown under two different cluster-level treatment assignments. On the left are the observed values of treatment; on the right is a permuted version. In the observed data on the left, since all individual-level treatments are known, it is clear which individuals in each cluster are controls and should have the potential to contribute to the test statistic (those contained in the dotted line). On the right, it is no longer clear which individuals in Cluster 1 would have opted to remain in the control group and should contribute to the test statistic. Figure by Aurelien Bibaut.

2.2 Design

To construct a design for this observational data, we will be adopting the framework of a cluster-level matched pairs experiment, which is a recommended experimental design choice when working with two-stage settings in practice [9]. Pairs of clusters are first matched exactly on the cluster-level propensity scores. To examine the presence of spillover, comparing control units in treated clusters to control units in matched control clusters seems like a logical next step, but a problem arises for randomization inference. In order to compute the randomization distribution of a test statistic under a null hypothesis, the value of the test statistic must be known under each possible treatment assignment, a property called imputability. If we permute the cluster-level assignment between matched clusters, it is not known which individuals in control clusters would have adopted treatment had their clusters been treated. (See Figure 2.1) Thus, unlike in a randomized cluster experiment, the control units in the treated cluster may be fundamentally different than the controls in the control cluster in a non-random way.

To address this issue, we will generate a virtual treatment vector $\mathbf{Z}_i^*$ for each control cluster during the randomization test which describes how individual-level treatment might have occurred if cluster i had received cluster-level treatment. The key is to generate this virtual treatment vector according to the propensity score distribution so that

$$P(\mathbf{Z}_i^* = \mathbf{z}_i | W_i = 0, \mathbf{L}_i, \mathbf{X_i}) = \pi_i(\mathbf{z}_i)$$
$$\stackrel{\text{def}}{=} P(\mathbf{Z_i} = \mathbf{z_i} | W_i = 1, \mathbf{L}_i, \mathbf{X_i})$$

We extend the definition of $\mathbf{Z}_i^*$ to treated clusters i, in which we let $\mathbf{Z}_i^* = \mathbf{Z}_i$. This virtual treatment replicates the individual-level treatment assignment procedure for the control clusters in an identical manner as the treated clusters. Thus, we can be assured that pseudo-control units (those with $z_{ij}^* = 0$ in control clusters) are comparable to the observed control units in treated clusters (those with $z_{ij}^* = 0$ in treated clusters). The following section will give the theoretical justification for this argument and show how we can use these cluster matches and virtual treatment assignments in our hypothesis test for testing the existence of spillover effects.

2.3 Theoretical Framework

In this section, we will give some intuition behind the main theoretical tools needed to justify the validity of our main results.

Imputable Test Statistic

The first key idea to develop is that of an imputable test statistic. The following is the rigorous definition of an imputable test statistic given by Basse [4]:

Definition 1. *A test statistic $T(\mathbf{Z}|\mathbf{Y})$ is imputable with respect to a null hypothesis H_0 if for all $\mathbf{Z}, \mathbf{Z}'$ for which $pr(\mathbf{Z}) > 0$ and $pr(\mathbf{Z}') > 0$,*

$$T(\mathbf{Z}'|\mathbf{Y}(\mathbf{Z}')) = T(\mathbf{Z}'|\mathbf{Y}(\mathbf{Z}))$$

Intuitively, a test statistic is imputable with respect to a null hypothesis if the value of the test statistic $T(\mathbf{Z}|\mathbf{Y})$ can be imputed for every possible counterfactual assignment vector $\mathbf{Z}'$, using only outcomes $\mathbf{Y}$ observed under $\mathbf{Z}$. This property allows us to construct the correct sampling distribution of the test statistic. To illustrate this concept, we give an example:

> **Example 1** Consider the classic Neyman-Rubin framework with the standard sharp null hypothesis
>
> $$Y_i(0) = Y_i(1) \text{ for all } i \tag{2.2}$$
>
> As a concrete example of a test statistic, consider the difference-in-means test statistic.

$$T(\mathbf{Z}'|Y(\mathbf{Z})) = Ave\{Y_i(\mathbf{Z})|Z_i' = 1\} - Ave\{Y_i(\mathbf{Z})|Z_i' = 0\}$$

where the observed test statistic is $T(\mathbf{Z^{obs}}|Y(\mathbf{Z}^{obs}))$ and the permutation distribution of test statistic is generated by considering $T(\mathbf{Z}'|Y(\mathbf{Z}^{obs}))$ for various values of $\mathbf{Z}'$.

Under the classic potential outcomes framework, each unit only has two potential outcomes. Because the null hypothesis assumes that these two potential outcomes are equal, then given any set of observed outcomes, we can impute all of the potential outcomes for every unit from the observed data. Consequently, our test statistic (which is a function of potential outcomes) will be a function of known values regardless of which treatment vector permutations we consider when constructing the randomization distribution. This will be true for any test statistic we select in addition to the difference-in-means test statistic we chose.

Furthermore, our null hypothesis gives us that $Y_i(\mathbf{Z}') = Y_i(\mathbf{Z})$ for all i and any $\mathbf{Z}'$, $\mathbf{Z}$. This implies that $T(\mathbf{Z}'|Y(\mathbf{Z}')) = T(\mathbf{Z}'|Y(\mathbf{Z}))$ for any test statistic T, satisfying Definition 1.

In our framework, we now have more than 2 potential outcomes for each unit. Given our particular null hypothesis (2.1), not every test statistic will be imputable. This is because our null hypothesis does not allow us to determine all of the potential outcomes from the observed data. Specifically, suppose a unit is observed under treatment. In that case, the null hypothesis of "no interference" (2.1) gives us no ability to impute how this unit would have responded if this unit were a control. Thus, our test statistic can never be a function of this unit's potential outcome under individual-level control.

Conditioning Events

In order to account for restrictions such as this and to encapsulate our design proposed in Section 2.2, we will make use of Basse's concept of a conditioning event $\mathcal{C}$. When we perform our randomization test, we will first generate our pseudo-treatment vectors $\mathbf{Z}_i^*$ and select only a subset of units to consider in our test statistic. Then, we will perform cluster-level matching exactly on the propensity score. Finally, we will incorporate this information into our test and perform randomization inference where we only allow cluster-level treatment permutations where exactly one cluster in each pair is treated.

These steps listed above are known as our conditioning event $\mathcal{C}$ and are drawn from a random distribution $m(\mathcal{C}|\mathbf{Z^{obs}}, \mathbf{W^{obs}})$ that is conditional on our observed cluster and individual level assignments. This distribution is also implicitly conditional on known values such as the cluster-level covariates and propensity scores since those factors are used to perform matching.

One of the helpful aspects of Basse's work is that the distribution from which we draw the conditioning event is allowed to be conditional on the observed assignment vector. This means we can use the observed data in the observational study to perform matching, generate the pseudo-assignments, and dictate which cluster-level treatment permutations to consider without compromising the validity of the randomization test. The technical details are omitted here and can be found in the appendix with the proof for Proposition 1.

2.4 Results

We now can propose the following testing procedure for testing the null hypothesis as in (2.1).

Proposition 1 (Randomization Test for Spillover Effect in Two-Stage Observational Data)**.** *Consider the following testing procedure:*

1. *In all control clusters, simulate a virtual treatment assignment* $\mathbf{z}_i^*$ *where* $P(\mathbf{Z}_i^* = \mathbf{z}_i^*) = \pi(\mathbf{z}_i^*)$*. Set* $\mathbf{z}_i^* = \mathbf{z}_i$ *for any treated clusters.*

2. *Discard any clusters where all units are treated/pseudo-treated (where* $\mathbf{z}_i^* = \mathbf{1}$*)*

3. *Match remaining clusters exactly on propensity scores. Let* $\mathcal{W}$ *be the event that exactly one cluster in each pair receives cluster-level treatment.*

4. *In each cluster, choose one unit from the subset of pseudo-control units (where* $z_{ij}^* = 0$*) with uniform probability. Let* $\mathcal{U}$ *be the set of ordered pairs* (i, j) *where if* $(i, j) \in \mathcal{U}$*, then the* j*th unit was selected from cluster* i*.*

5. *Compute the observed test statistic (the difference between the selected units in treated clusters and control clusters)*

$$T(\mathbf{W}^{obs}|\mathbf{Y}(\mathbf{Z}^*, \mathbf{W}^{obs}), \mathcal{W}, \mathcal{U}) = \frac{1}{m} \sum_{\substack{(i,j) \in \mathcal{U} \\ W_i^{obs}=1}} Y_{ij} - \frac{1}{m} \sum_{\substack{(i,j) \in \mathcal{U} \\ W_i^{obs}=0}} Y_{ij}$$

   where m *is the number of cluster matches.*

6. *Compute the distribution of the test statistic induced by permutations of the cluster-level treatment assignment* $\mathbf{W}$ *(conditional on the cluster-level matching event* $\mathcal{W}$*)*

7. *Compute the p-value*

Steps 1-7 outline a procedure that is valid for testing the null hypothesis of no spillover effect as in (2.1)

The proof for this proposition is in the appendix and primarily shows that this proposition is an application of Basse's work on conditioning events [4].

While the above methodology is valid and the most straightforward extension of Basse's work, there are three steps that each result in discarding portions of our data and consequently may lead to a loss of power. The first is the generation of the virtual treatment vector $\mathbf{Z}_i^*$ in control clusters, which disqualifies any unit that is selected as being virtually treated. The second loss is in the matching step which discards any clusters without a match. The third loss is in the selection of focal units which only gives us one focal unit per cluster. This final loss may be quite significant if the cluster sizes are large. While some discarding

of data is inevitable in matched pairs analyses, we will positively benefit by leveraging as much data as we can in the test.

Generally, it has been recommended to re-run the above method repeatedly as a naïve way to marginalize over the random draws that discard data. However, if possible, it is better to modify the conditioning event and test statistic to incorporate more information while keeping the test valid. The following proposition is a modified procedure that increases the number of focal units chosen, changes the test statistic, and results in greater power to detect deviations from the null.

Proposition 2 (Using Cluster-Level Means for Increased Power)**.** *Consider the following modified version of the procedure given in Proposition 1.*

- *Perform steps 1-3 as previously stated.*
- *When selecting focal units, let $\mathcal{A}$ be the set of all clusters that were matched. Then let $\mathcal{U}$ be the set of all units that have $z^*_{ij} = 0$ (psuedo-controls) and where $i \in \mathcal{A}$.*
- *Our new test statistic is now the difference in matched cluster-level means using the outcomes of all of the psuedo-controls in each cluster:*

$$T(\mathbf{W}^{obs}|\mathbf{Y}(\mathbf{Z}^*, \mathbf{W}^{obs}), \mathcal{W}, \mathcal{U}) = \frac{1}{m} \sum_{\substack{i \in \mathcal{A} \\ W_i^{obs}=1}} Mean[Y_{ij}|z^*_{ij} = 0] - \frac{1}{m} \sum_{\substack{i \in \mathcal{A} \\ W_i^{obs}=0}} Mean[Y_{ij}|z^*_{ij} = 0]$$

  where $Mean[Y_{ij}|z^*_{ij} = 0]$ is the mean outcome of all units in cluster i that have $z^*_{ij} = 0$.

- *Perform steps 6-7 as previously stated.*

This outlines a procedure that is valid for testing the null hypothesis of no spillover effect as in (2.1)

The proof of this proposition is a direct extension of the proof of Proposition 1. The selection of focal units is now more straightforward since we select all pseudo-controls from each cluster instead of a single unit. The argument for the imputability of this test statistic and the validity of the permutation test as an accurate calculation of the p-value follows equivalently from the proof of Proposition 1. We show in the subsequent section that this test yields power greater than or equal to Proposition 1.

Chapter 3

Data Simulation

3.1 Set-Up

To empirically verify and further analyze the findings in Propositions 1 and 2, we formulated several simulation scenarios to test this methodology's ability to identify a spillover effect in observed data. We began by using the same simulation set-up as used in Barkley [8] with the only modifications being those necessary to match our framework (specifically, the addition of a cluster-level treatment assignment). We also created two additional simulation set-ups by modifying the outcome generation while keeping everything else identical.

To generate each dataset, the following steps were carried out for each of the $i = 1, 2, ...N = 125$ clusters:

1. The number of individual units n_i in the cluster i was simulated such that $P(n_i = 8) = 0.4$, $P(n_i = 22) = 0.35$, and $P(n_i = 40) = 0.25$.

2. Two cluster-level covariates were generated with $L_{i1} \sim N(6, 1)$ and $L_{i2} \sim N(0, 0.75)$.

3. Cluster-level propensity scores were generated with

$$\begin{aligned} P(\theta_i = 0.3|L_{i1}) &= 0.25 + 0.25\mathbb{1}[L_{i1} \leq 5.5], \\ P(\theta_i = 0.5|L_{i1}) &= 0.25 + 0.25\mathbb{1}[5.5 < L_{i1} < 6.5], \text{ and} \\ P(\theta_i = 0.7|L_{i1}) &= 0.25 + 0.25\mathbb{1}[L_{i1} \geq 6.5]. \end{aligned}$$

   These propensity scores were used to generate the cluster-level treatment assignment W_i where $P(W_i = 1) = \theta_i$.

4. Covariates for each individual $j = 1, ..., n_i$ in cluster i were simulated to be $X_{ij1} \sim N(40, 5)$ and $X_{ij2} \sim N(L_{i1}, 0.2)$.

5. Treatment status Z_{ij} for each individual j in a treated cluster i was simulated from a Bernoulli distribution with

$$Pr(Z_{ij} = 1|W_i = 1, \mathbf{L}_i, \mathbf{X}_{ij}) = \mathcal{L}^{-1}(0.75 - 0.015X_{ij1} - 0.025X_{ij2} + L_{i2})$$

where $\mathcal{L}$ is the logit link function. All units in non-treated clusters ($W_i = 0$) were untreated.

6. The outcome for each individual Y_{ij} varied across each simulation, but each made use of the following quantity: $g(Z_{i,-j}) = (n_i - 1)^{-1} \sum_{j' \neq j} Z_{ij'}$, which intuitively can be thought of as the proportion of units in cluster i besides unit j which received treatment.

   a) Simulation A (Same as in Barkley [8]):

   $$P(Y_{ij} = 1 | \mathbf{Z_i}, \mathbf{X}_{ij}) = \mathcal{L}^{-1}(0.1 - 0.05X_{ij1} + 0.5X_{ij2} - 0.5Z_{ij} + \lambda g(Z_{i,-j})[0.2(1 - Z_{ij}) - 0.05Z_{ij}])$$

   where $\lambda \in \mathbb{R}$. (Barkley [8] corresponds to $\lambda = 1$).

   b) Simulation B (Vaccination interpretation):

   $$P(Y_{ij} = 1 | \mathbf{Z_i}, \mathbf{X}_{ij}) = \mathcal{L}^{-1}(0.1 - 0.05X_{ij1} + 0.5X_{ij2} - 0.5Z_{ij}) \times [1 - \lambda(1 - g(Z_{i,-j}))]$$

   where $\lambda \in [0, 1]$.

   c) Simulation C (Continuous outcome):

   $$Y_{ij} = 0.1 - 0.05X_{ij1} + 0.5X_{ij2} - 0.5Z_{ij} + \lambda g(Z_{i,-j})[0.2 \times \mathbb{1}[Z_{ij} = 0] - 0.05 \times \mathbb{1}[Z_{ij} = 1]] + \varepsilon$$

   where $\varepsilon \overset{\text{i.i.d.}}{\sim} N(0, 0.3^2)$ and $\lambda \in \mathbb{R}$.

In each simulation, we included a parameter λ whose magnitude dictated the strength of the spillover effect, with $\lambda = 0$ corresponding to no spillover.

While Simulation A matches previous work [8], we found its generation of the outcome less interpretable. In this set-up, being treated directly decreases the odds of a unit experiencing the outcome. For those units that are treated, the odds of experiencing the outcome are further reduced by having a greater proportion of other units in the cluster treated. However, if a unit is a control, the trend reverses and the odds of the outcome occurring increase if the proportion of treated units in the cluster increases.

The outcome in Simulation B seeks to be a more intuitive data-generating process motivated by the application of vaccinations. The quantity inside the inverse logit can be viewed as the unit's baseline probability of illness given the unit's vaccination status and covariates. However, the true probability of illness can be reduced by a spillover effect from others in the cluster also being vaccinated. We assume that $\lambda \in [0, 1]$ dictates the largest proportional reduction in probability if everyone else in the cluster is treated. For instance, if $\lambda = 0.60$, then a unit's probability of illness would reduce by 60% if all other units in the cluster were

treated. This reduction scales linearly with the proportion of others that received treatment. Thus if $\lambda = 0.60$ and only one-fourth of the other units in the cluster received treatment, then the overall probability of illness would be the baseline probability reduced by $0.60 \times \frac{1}{4} = 15\%$.

These first two simulations involve binary outcomes, so Simulation C modifies the original setup to include a continuous outcome.

3.2 Results

Validity of Test

Propositions 1 and 2 claim that our methods create valid hypothesis tests. Thus we must show that if the null hypothesis is true (i.e. $\lambda = 0$) we have the property that $P(pval \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$. Graphically, this means that the CDF of the p-value (our random variable of interest) must not be greater than the line $f(x) = x$.

In order to test the finite-sample performance of this test, we simulated 1000 datasets for various values of λ under each of our three simulation settings. After generating each dataset as specified previously, we carried out each of the steps in Propositions 1 and 2 which resulted in a p-value. Matching was performed exactly without replacement by cluster-level propensity scores. Results are shown in Figure (3.1).

We see that for all simulations and both propositions, the empirical CDF of the p-value for no spillover effect ($\lambda = 0$) falls at or below the black dashed line $f(x) = x$. This shows that our test is in fact valid. Notably, for Simulations A and B, the p-value distribution is sub-uniform when using the method in Proposition 1. The reason that these distributions are sub-uniform is a result of the fact that many permutations of the cluster-level assignment vector result in an identical test statistic to the one we observed. More specifically, when the selected focal units from matched clusters experience the same binary outcome, the permutation of the cluster-level assignment does not affect the test statistic. Thus when many matched clusters have focal units with the same outcome, there are few permutations that actually change the test statistic. When the outcome is continuous as in Simulation C or we use the mean outcome of all of the pseudo-controls as in Proposition 2, the distribution of the p-value for no spillover effect converges to the uniform distribution.

Power of Test

For every simulation, as the magnitude of the spillover increased, the p-value distribution became more right-skewed. This indicates a greater power for the test to identify the spillover effect.

One interesting thing to note is that our method in Proposition 1 had difficulty identifying a spillover effect at all in the simulation used by Barkley [8] (Simulation A, $\lambda = 1$). However, when we use the method in Proposition 2, our resulting p-value distribution has significantly

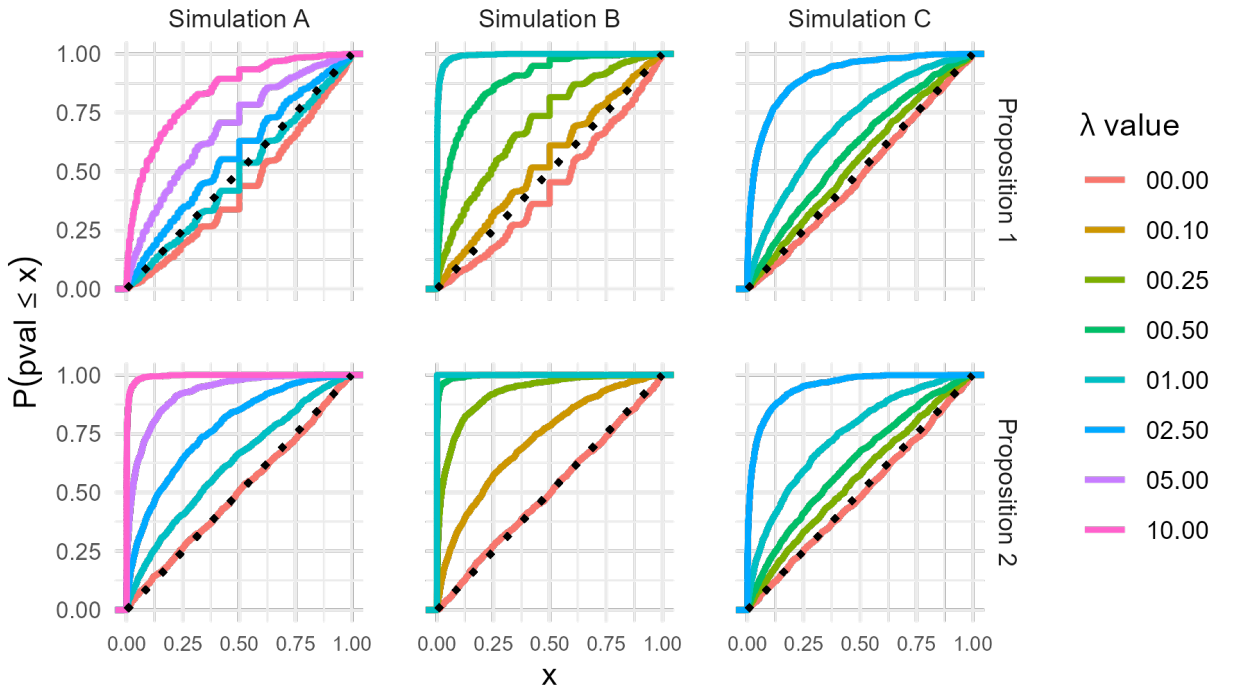

Figure 3.1: Empirical CDF of the p-value from the methods in Propositions 1 and 2

In each simulation setting, the strength of the interference structure was determined by a parameter λ, with $\lambda = 0$ indicating no spillover effect. A valid hypothesis test requires that the true CDF of the p-value lies on or below the dotted line $f(x) = x$ when the null hypothesis of no interference holds.

diverged from the uniform distribution. In fact, for every $\lambda > 0$ in simulations A and B, using the method in Proposition 2 resulted in a significantly greater right-skew in the p-value distribution than Proposition 1. Interestingly, the increase of power between the two propositions is much smaller for the continuous outcome of Simulation C.

Using a Single Dataset

For the following discussion, we will restrict our attention to Proposition 2. As we ran the simulations above, there were essentially two components of randomness present in our set-up and hypothesis test:

1. The generation of the sample dataset from the underlying population
2. The generation of the virtual treatment assignments $\mathbf{z}_i^*$

3. The matching of clusters, since the only requirement for two clusters being matched is having identical cluster-level propensity scores

In practice, we often only have a single dataset to work with (we cannot repeatedly generate additional samples), but we hypothetically could repeat the procedure in Proposition 2 over and over to generate a distribution of p-values. One advantage of doing this is that it "averages" over the randomness due to the generation of $\mathbf{z}_i^*$ and matching. The distribution of p-values may be useful to understand the strength of the evidence against the null hypothesis. As noted in Basee [4], the fact that we now have a distribution of p-values does not compromise the validity of the test, but it does raise challenges for the interpretation and sensitivity of the results.

To illustrate this point, we generate 40 data sets using the Simulation B set-up for various λ values (8 datasets for each of the 5 values of λ). For each dataset, we repeat the procedure in Proposition 2 repeatedly and obtain a distribution of p-values. The results are shown on the left of Figure (3.2).

For the $\lambda = 0$ scenario, we saw previously that the distribution of a single p-value is uniform for repeated draws from the data-generating distribution. However, once we've drawn a fixed dataset, repeatedly sampling virtual treatment vectors $\mathbf{z}_i^*$ and performing all possible matches no longer creates a uniform distribution of p-values. This is because there may be a difference between control units in control clusters and control units in treated clusters by random chance. However, our test still seeks to determine if any difference is statistically significant. When $\lambda = 0$, the center and spread of the p-value distribution changes for each dataset, but in general no spillover signal is detected amidst other noise in the data. For larger values of λ, our p-value is significant regardless of which units are selected for virtual treatment and which clusters are matched. A common metric to look at is the proportion of p-values that are less than some threshold α, which is shown on the right of Figure (3.2).

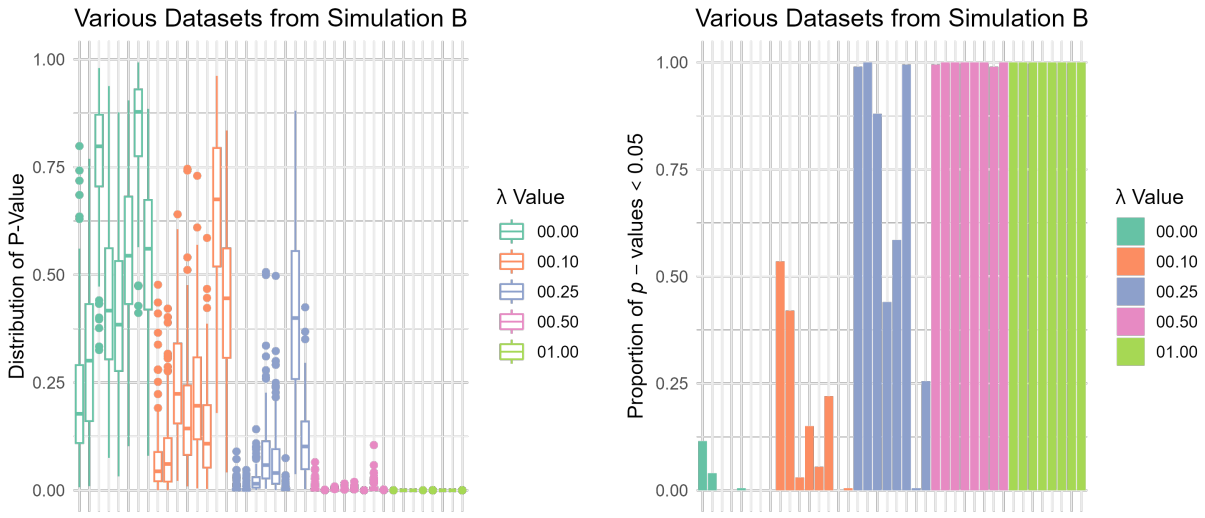

Figure 3.2: Various p-value distributions for fixed Simulation B datasets and repeated applications of Proposition 2

On the left, each boxplot represents the resulting p-value distribution for a single dataset. The same datasets are represented on the right with the bars showing the proportion of p-values that are less than $\alpha = 0.05$. We rarely get significant p-values when the null is true ($\lambda = 0$). For small amounts of spillover, our ability to detect the relevant signal varies, while the spillover signal is always detectable for larger values of λ

.

Chapter 4

Discussion

4.1 Estimation

In the previous sections, we have shown that the procedures in Propositions 1 and 2 are valid for testing the null hypothesis of no interference. One of the most natural extensions of a hypothesis test is to use it for estimation since the two procedures are dual problems. However, in our case, the corresponding statistical estimand isn't directly apparent. Under the framework we have built, there is only one counterfactual outcome for each unit under the scenario of cluster-level control, $Y_{ij}(\mathbf{0}, 0)$. However, there are up to 2^{n_i} counterfactual outcomes for each unit under the scenario of cluster-level treatment, $Y_{ij}(\mathbf{z}_i, 1)$ for all $\mathbf{z}_i$ with $\pi(\mathbf{z}_i) > 0$.

Using the Sharp Null Hypothesis

One approach we could take for estimation, which parallels the framework of this thesis, is to use a perspective motivated by the sharp null hypothesis. In this case, we could assume there is some constant spillover effect τ for control units in treated clusters. At this point, we would seek to find the values of τ such that the following null hypothesis was not rejected:

$$Y_{ij}(\mathbf{z}_i, 1) - Y_{ij}(\mathbf{0}, 0) = \tau \text{ for all } \mathbf{z}_i \text{ such that } z_{ij} = 0 \tag{4.1}$$

However, if there was a spillover effect, in most practical applications it would be unreasonable to expect it to be constant. In most applications, we would likely expect there to be no spillover effect for being a control unit in a treated cluster if all units in the cluster still chose to adopt control, i.e. $Y_{ij}(\mathbf{0}, 1) = Y_{ij}(\mathbf{0}, 0)$.

Following the literature, we note that often we need to make some assumptions about the form of interference in order to obtain estimates and inference [7]. For instance, Hudgens and Halloran assumed that the outcome of one unit depended on the treatment assignment of other units only through the number of those who are assigned to the treatment condition

within the same cluster. In other words, what mattered was solely the number of units assigned treatment condition rather than which units were assigned to the treatment condition [2]. Such an assumption could lead to the following null hypothesis:

$$Y_{ij}(\mathbf{z}_i, 1) - Y_{ij}(\mathbf{0}, 0) = \tau \times f(\mathbf{z}_i) \text{ for all } \mathbf{z}_i \text{ such that } z_{ij} = 0 \tag{4.2}$$

where $f(\mathbf{z}_i)$ could be any function of the assignment vector, such as the number or proportion of units receiving treatment in the given cluster. This formulation would allow us to obtain a summary of how spillover behaves according to some plausible mechanism.

The advantage of using these forms of estimation is that almost all the work in this thesis can be naturally extended to these scenarios. In fact, the null hypothesis of (4.2) could be tested by modifying the test statistic used in Proposition 1 to be

$$T(\mathbf{W}'|\mathbf{Y}(\mathbf{Z}^*, \mathbf{W}^{obs}), \mathcal{W}, \mathcal{U}) = \frac{1}{m} \sum_{\substack{(i,j) \in \mathcal{U} \\ W'_i = 1}} Y^*_{ij} - \frac{1}{m} \sum_{\substack{(i,j) \in \mathcal{U} \\ W'_i = 0}} Y^*_{ij}$$

where Y^*_{ij} is the outcome of interest shifted for those units in observed treated clusters, defined by $Y^*_{ij} = Y_{ij} - \big(\mathbb{1}(W^{obs}_i = 1) \times \tau \times f(\mathbf{z}^*_i)\big)$. Note how choosing $f(\mathbf{z}^*_i) = 1$ returns us to testing the constant treatment effect and choosing $\tau = 0$ returns the above test statistic to our original set-up.

Using the Weak Null Hypothesis

In some cases, it may be too presumptuous to assume forms such as those in (4.2). Following the example of Imai [7], we can aggregate together the $Y_{ij}(\mathbf{z}_i, 1)$ potential outcomes of interest into a single potential outcome and use this to compare to $Y_{ij}(\mathbf{0}, 0)$. Let $\mathbf{z}_{i,-j} = (z_{i,1}, ..., z_{i,j-1}, z_{i,j+1}, ..., z_{i,n_i})$ represent the $(n_i - 1)$ dimensional subvector of $\mathbf{z_i}$ with the entry for unit j removed and $\mathcal{Z}_{i,-j} = \{(z_{i,1}, ..., z_{i,j-1}, z_{i,j+1}, ..., z_{i,n_i}) | z_{i,j'} \in \{0, 1\}$ for $j' \in 1, ..., j-1, j+1, ..., n_i\}$ be the set of all possible values of the assignment vector $\mathbf{z}_{i,-j}$. We define an individual's aggregated control-unit treated-cluster potential outcome to be

$$\bar{Y}_{ij}(z_{ij} = 0, W_i = 1) = \sum_{\mathbf{z}'_{i,-j} \in \mathcal{Z}_{i,-j}} Y_{ij}(z_{ij} = 0, \mathbf{z}_{i,-j} = \mathbf{z}'_{i,-j}, W_i = 1) \times P(\mathbf{z}_{i,-j} = \mathbf{z}'_{i,-j} | z_{ij} = 0, W_i = 1, L_i, \mathbf{X}_i)$$

Since we only have a single potential outcome when a unit is in a control cluster, we analogously define $\bar{Y}_{ij}(z_{ij} = 0, W_i = 0)$ to be equal to $Y_{ij}(\mathbf{0}, 0)$

This leads us to the natural estimand of spillover defined as

$$\tau = E\left[\bar{Y}_{ij}(z_{ij} = 0, W_i = 1) - \bar{Y}_{ij}(z_{ij} = 0, W_i = 0)\right] \tag{4.3}$$

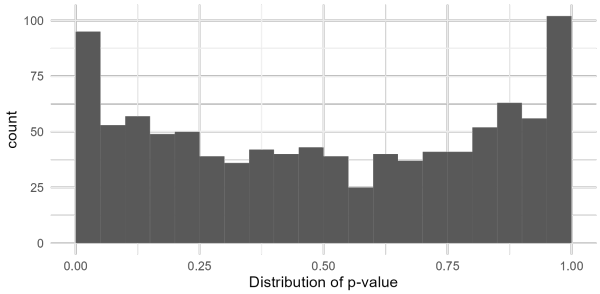

Figure 4.1: Empirical CDF of p-value when using inexact covariate matching

The p-value distribution was generated in a manner identical to the methodology in Chapter 3 for Simulation C and $\lambda = 0$ with the exception that nearest-neighbor matching was done on X_{i1} within exact propensity score matches. While the p-value distribution should be uniform under the null hypothesis, inexact matching adds a bias when correction methods are not employed.

where we will have $\tau = 0$ under the null hypothesis (2.1). This is equivalent to Imai's defintion of $SEY(0)$ [7]. Imai's paper focuses solely on the estimation of this effect for randomized experiments, but natural extensions can be made by substituting propensity scores in for randomized treatment probabilities. Additional work would need to be done to determine if the methods in this paper with an appropriate test statistic and conditioning mechanism could also yield an unbiased estimate of τ.

4.2 Challenges with Inexact Matching

One of the largest practical challenges of this method is that it requires clusters to be matched exactly on propensity scores. In a true observational study, exact matching is likely infeasible. Inexact matching will naturally add bias to our test unless the procedures are properly adjusted.

To illustrate this challenge, consider the set-up of the simulation in Chapter 3. We originally only matched on propensity scores, but suppose that within a subset of clusters with the exact same propensity score, we additionally matched on our cluster-level continuous covariate L_{i1}. We would likely choose to match on this cluster-level covariate because we assume it is predictive of the outcome, as is L_{i1} in our simulation, and will help improve the test's precision. However, when we do this, we found that the distribution of the p-values under the null hypothesis is no longer uniform, but rather exhibits a U-shape as shown in Figure 4.1. Why does this happen?

Under the null hypothesis, the distribution of L_{i1} should be identical between control clusters and treated clusters with the same propensity score. However, because we are

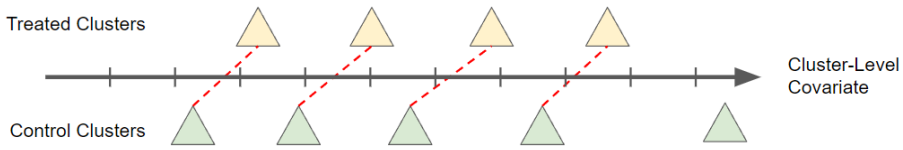

Figure 4.2: Illustration of Induced Bias from Inexact Matching

The above figure represents a hypothetical scenario when trying to match 4 treated clusters to a subset of 5 control clusters using nearest-neighbor without replacement. While the optimal solution is shown, every control cluster has a lower cluster-level covariate value than its matched treated cluster. The fact that matching was done without replacement forced many of the matches to be chosen in this manner.

working in a finite-sample scenario, the distributions between the two groups may vary by random chance. In addition, the task of matching is an inherently dependent process when done without replacement (as is common). By this, we mean that if two treated clusters have the same ideal "equivalent" control cluster, only one treated cluster may match with the ideal control cluster while the other treated cluster must find another match.

The ideal matched structure that minimizes some distance between matched pairs may induce a repeated same-direction bias between treated clusters and control clusters. As the example shown in Figure 4.2 shows, when the cluster-level covariate is strongly predictive of the outcome, then at times the ideal nearest-neighbor match shown causes every match to exhibit a bias in the same direction between the treated and control cluster. This will likely result in a p-value close to either 0 or 1 even when there is no underlying difference between units in treated or control clusters.

The situation shown in Figure 4.2 may seem like a carefully chosen catastrophic example of what may go wrong, but the U-shape in Figure 4.1 demonstrates that such a scenario likely occurs more often than one would hope, especially when the there is a small number of both treated and control clusters within a given propensity score.

The above illustration showed an induced bias for inexact matching on a continuous covariate within an exact propensity score, but the same phenomenon exists when solely matching on propensity scores when matching is done inexactly.

Solutions Within the Matching Paradigm

There are several solutions built within the matching framework that can address these challenges. The first is to impose greater restrictions upon which matches are allowed, such as a minimum distance (a caliper) between propensity scores or covariates [10]. Matching with replacement also has the potential to eliminate bias and match dependence if there are several treated units that each would be optimally matched to the same control unit (or

visa versa), though such a method would require careful consideration on how to perform randomization inference with such a set-up.

Another alternative to consider is using a regression-adjusted test statistic that incorporates any covariates used for matching [11]. Such a method attempts to correct for any bias induced by the matching process without diminishing the sample size by being overly conservative in the match process.

Solutions Outside the Matching Paradigm

Propositions 1 and 2 can be modified to create methods that do not require matching in the conditioning event, but the null test statistic distribution would not be generated by strict permutation of cluster-level assignments. More specifically, the computation of the p-value in Theorem 3 (in the appendix) would need to be done with respect to a new conditioning mechanism that would give a larger weight to more likely treatment assignments, as is done in Pimentel [12]. Beyond the framework proposed in this paper, inverse propensity score estimators have also been used to estimate spillover effects, with the natural caveats of instability for extreme propensity scores [8].

Despite these challenges, researchers often support using matched-cluster designs when appropriate in the randomized experiment setting because of the frequent efficiency gains [9]. These arguments extend to the observational setting as well and motivate the work done in this thesis.

[A section on sensitivity analysis would be good to add. Another natural question is whether/how to extend this idea to more complex interference patterns (e.g. if you observed a social network). Here is a paper that could provide some ideas about how to do this kind of extension: Jagadeesan, R., Pillai, N. S., Volfovsky, A. (2020). Designs for estimating the treatment effect in networks with interference. — Sam]

Bibliography

[1] Donald B Rubin. "Discussion of 'Randomization Analysis of Experimental Data in the Fisher Randomization Test,' by D. Basu". In: *Journal of the American Statistical Association* 75 (1980), pp. 591–593.

[2] Michael G Hudgens and M. Elizabeth Halloran. "Toward Causal Inference With Interference". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 832–842.

[3] Dean Eckles Susan Athey and Guido W. Imbens. "Exact p-Values for Network Interference". In: *Journal of the American Statistical Association* 113.521 (2018), pp. 230–240.

[4] A Feller G W Basse and P Toulis. "Randomization tests of causal effects under interference". In: *Biometrika* 106.2 (2019), pp. 487–494.

[5] Michael E Sobel. "What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference". In: *Journal of the American Statistical Association* 101.521 (2006), pp. 1398–1407.

[6] Guanglei Hong and Stephen W. Raudenbush. "Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data". In: *Journal of the American Statistical Association* 101.475 (2006), pp. 901–910.

[7] Zhichao Jiang Kosuke Imai and Anup Malani. "Causal Inference With Interference and Noncompliance in Two-Stage Randomized Experiments". In: *Journal of the American Statistical Association* 116 (2020), pp. 632–644.

[8] Brian G Barkley and Michael G Hudgens et al. "Causal inference from observational studies with clustered interference, with application to a cholera vaccine study". In: *The Annals of Applied Statistics* 14 (2020), pp. 1432–1448.

[9] Gary King Kosuke Imai and Clayton Nall. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation". In: *Statistical Science* 24 (2009), pp. 29–53.

[10] Peter C Austin. "Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies". In: *Pharm Stat* 10.2 (2011), pp. 150–161.

[11] Alberto Abadie and Guido Imbens. "Bias-Corrected Matching Estimators for Average Treatment Effects". In: *Journal of Business Economic Statistics* 29.1 (2011), pp. 1–11.

[12] Samuel Pimentel. "Covariate-adaptive randomization inference in matched designs". In: *arXiv* https://doi.org/10.48550/arXiv.2207.05019 (2022).

Appendix A

Proof of Proposition 1

Proof. We will adopt Basse's notation:

> Let $\mathbb{C} = \{(\mathcal{U}, \mathcal{Z}) : \mathcal{U} \subseteq \mathbb{U}, \mathcal{Z} \subseteq \mathbb{Z}\}$ be the space of conditioning events, where $\mathbb{U}$ denotes the power set of units, and $\mathbb{Z}$ denotes the power set of assignment vectors. For some conditioning event $\mathcal{C} = (\mathcal{U}, \mathcal{Z}) \in \mathbb{C}$, the conditioning mechanism can be decomposed, without loss of generality, as
>
> $$m(\mathcal{C}|\mathbf{Z}) = f(\mathcal{U}|\mathbf{Z})\, g(\mathcal{Z}|\mathcal{U}, \mathbf{Z})$$
>
> where f and g are distributions over $\mathcal{U}$ and $\mathcal{Z}$, respectively. Given conditioning event $\mathcal{C} = (\mathcal{U}, \mathcal{Z})$, we consider test statistics, $T(\mathbf{Z}|Y, \mathcal{C})$, that depend only on outcomes of units in $\mathcal{U}$; following terminology in Athey et al. [3], we call $\mathcal{U}$ the set of focal units.

While the set $\mathcal{U}$ denotes the set of focal units, $\mathcal{Z}$ denotes the set of possible assignment vectors to consider when performing randomization inference. As a tangible example, in the classic matched-pairs randomization test, $\mathcal{U}$ is the set of all units that are matched (some treated or control units that have no match may be discarded). In addition, we only allow permutations of our observed assignment vector $\mathbf{Z}$ where exactly one unit is treated per match, i.e.

$$\mathcal{Z} = \{\mathbf{Z}' : Z'_i + Z'_j = 1 \text{ if units i and j are matched}\}$$

This framework now allows us to construct valid hypothesis testing procedures for more general null hypotheses than the standard sharp null hypothesis (2.2). In particular, we will be making use of the following theorem developed by Basse:

Theorem 3 (Valid Hypothesis Test with Conditioning Event)**.** *Let H_0 be a null hypothesis and $T(\mathbf{Z}|Y, \mathcal{C})$ a test statistic, such that T is imputable with respect to H_0 under some conditioning mechanism $m(\mathcal{C}|\mathbf{Z})$; that is, under H_0*

$$T(\mathbf{Z}'|Y(\mathbf{Z}'),\mathcal{C}) = T(\mathbf{Z}'|Y(\mathbf{Z}),\mathcal{C}) \tag{A.1}$$

for all $\mathbf{Z},\mathbf{Z}',\mathcal{C}$ *for which* $pr(\mathbf{Z},\mathcal{C};m) > 0$ *and* $pr(\mathbf{Z}',\mathcal{C};m) > 0$. *Consider the procedure where we first draw* $\mathcal{C} \sim m(\mathcal{C}|\mathbf{Z}^{obs})$, *and then compute the conditional p-value,*

$$pval(\mathbf{Z}^{obs};\mathcal{C}) = E_Z[\mathbb{1}\{T(\mathbf{Z}|Y^{obs}) > T^{obs}\}|\mathcal{C}] \tag{A.2}$$

where $T^{obs} = T(\mathbf{Z}^{obs}|Y^{obs},\mathcal{C})$, *and the expectation is taken with respect to* $pr(\mathbf{Z}|\mathcal{C})$. *This procedure is valid at any level, that is,* $pr(pval(\mathbf{Z}^{obs};\mathcal{C}) \leq \alpha|\mathcal{C}) \leq \alpha$ *for any* $\alpha \in [0,1]$ *under* H_0.

To show that this procedure is valid for testing the null hypothesis in (2.1), we will be employing Theorem 3. This theorem requires us to first define our conditioning mechanism, equivalent to Basse's $m(\mathcal{C}|\mathbf{Z})$. Then we must show that our test statistic is imputable with respect to our null hypothesis (2.1) under our conditioning mechanism. Finally, we must show that this permutation test accurately computes the conditional p-value as in (A.2).

First, since our matching procedure depends on our observed assignment vectors, it must be part of our conditioning event. This matching process is an extension beyond Basse's original setup. For this permutation test, we will also be permuting the cluster-level assignment vectors $\mathbf{W}$ instead of the individual-level assignments $\mathbf{Z}$. Therefore our conditioning event is now $\mathcal{C} = (\mathcal{U},\mathcal{M},\mathcal{W})$ where $\mathcal{U}$ is equivalent to Basse, $\mathcal{M}$ is our matching event (described in step 3) defined be

$$\mathcal{M} = \{(i,j) : \textit{Cluster } i \textit{ is matched to cluster } j, \textit{ with } i < j\}$$

and $\mathcal{W}$ is the set of all allowed cluster-level treatment permutations (identical to Basse's $\mathcal{Z}$, though we will only be permuting treatment on the cluster-level). Thus our conditioning mechanism will be of the form

$$m(\mathcal{C}|\mathbf{Z},\mathbf{W}) = f(\mathcal{U}|\mathbf{Z},\mathbf{W})\, g(\mathcal{M}|\mathcal{U},\mathbf{Z},\mathbf{W})\, h(\mathcal{W}|\mathcal{M},\mathcal{U},\mathbf{Z},\mathbf{W})$$

Our functions f and g are fully described by the selection of the focal units and matching described in steps 1-4 in the proposition, though we define f here in its equivalent mathematical definition. Let $\mathbf{u}$ be an indicator vector indexed identically to our treatment vector, such that $u_{i,j} = 1$ if the jth unit in the ith cluster is selected as a focal unit. Let $\vec{1}$ be the vector of all 1's. Mathematically, we have:

$$f(\mathcal{U} = \mathbf{u}|\mathbf{Z},\mathbf{W}) = \sum_{\mathbf{Z}^*} f(\mathcal{U} = \mathbf{u}|\mathbf{Z},\mathbf{Z}^*,\mathbf{W}) P(\mathbf{Z}^*|\mathbf{Z},\mathbf{W}) \tag{A.3}$$

where

$$f(\mathcal{U} = \mathbf{u}|\mathbf{Z}, \mathbf{Z}^*, \mathbf{W}) \propto \prod_{i=1}^{N} \mathbb{1}\left(\sum_{j=1}^{n_i} \mathbb{1}(u_{i,j} = 1) = \mathbb{1}\left(\mathbf{Z}_\mathbf{i}^* \neq \vec{1}\right)\right) \prod_{(i,j):u_{i,j}=1} \mathbb{1}(\mathbf{Z}_{i,j}^* = 0) \quad \text{(A.4)}$$

and

$$P(\mathbf{Z}^*|\mathbf{Z}, \mathbf{W}) = \prod_{i=1}^{N} \left(\mathbb{1}(\mathbf{W}_i = 1)\mathbb{1}(\mathbf{Z}_i^* = \mathbf{Z}_i) + \mathbb{1}(\mathbf{W}_i = 0)\pi(\mathbf{Z}_i^*)\right) \quad \text{(A.5)}$$

For some intuition to these formulas, in (A.3) we use the law of total probability to incorporate our virtual treatment assignment vector $\mathbf{Z}_i^*$. Step 1 in the proposition is given by (A.5). The two products in (A.4) describe a uniform distribution detailed by steps 2 and 4 in the proposition: selecting exactly one pseudo-control unit ($Z_{i,j}^* = 0$) from each cluster that has at least one available pseudo-treated unit to select.

The mathematical form of g will depend on the particularities of the matching scheme and the observed data, such as whether there are more control or treated clusters, whether the researcher allows for a cluster to be matched multiple times, etc. The result holds as long as the clusters are matched exactly on propensity scores.

Finally,

$$h(\mathcal{W}|\mathcal{M},\mathcal{U}, \mathbf{Z}, \mathbf{W}) = \mathbb{1}(\mathcal{W} = \{\mathbf{W}' : W_i' + W_j' = 1 \textit{ for all } \{i, j\} \in \mathcal{M}\}) \quad \text{(A.6)}$$

which is a degenerate distribution on the set of assignments for which exactly one cluster in each match receives treatment.

We now show that our test statistic is imputable, see (A.1). Under the null hypothesis (2.1), the potential outcome $Y_{i,j}$ for any unit receiving control does not depend on the cluster-level treatment status. Thus for any control unit, the observed value $Y_{i,j}$ will not change if the cluster-level treatment is altered as long as $z_{i,j} = 0$ remains constant. Since our test statistic is only a function of control units, then for any cluster-level treatment vectors $\mathbf{W}', \mathbf{W}$ we have

$$T(\mathbf{W}'|\mathbf{Y}(\mathbf{Z}, \mathbf{W}'), \mathcal{W},\mathcal{U}) = T(\mathbf{W}'|\mathbf{Y}(\mathbf{Z}, \mathbf{W}), \mathcal{W},\mathcal{U})$$

which shows imputability.

Finally, we show that this permutation test induced by permuting $\mathbf{W}$ conditional on the event $\mathcal{W}$ calculates the valid conditional p-value (A.2). For this we simply need to show that $P(\mathbf{W} = \mathbf{w}|\mathcal{C}) = P(\mathbf{W} = \mathbf{w}|\mathcal{U}, \mathcal{M}, \mathcal{W})$ is uniform in its support. Consider an arbitrary cluster-level assignment vector $\mathbf{w}$ that is in the support of our permutation test. Because we are conditioning on $\mathcal{W}$ and $\mathcal{M}$, we know that $w_i + w_j = 1$ for all $(i, j) \in \mathcal{M}$ by (A.6). Without loss of generality, let $w_i = 1$ and $w_j = 0$ for each $(i, j) \in \mathcal{M}$ since the cluster indices are arbitrary. Since each cluster-level assignment is independent, we have:

$$
\begin{aligned}
P(\mathbf{W} = \mathbf{w}|\mathcal{U}, \mathcal{M}, \mathcal{W}) &= \prod_{i=1}^{N} P(W_i = w_i|\mathcal{U}, \mathcal{M}, \mathcal{W}) \\
&\propto \prod_{(i,j)\in\mathcal{M}} P(W_i = 1 \cap W_j = 0|w_i + w_j = 1) \\
&= \prod_{(i,j)\in\mathcal{M}} \frac{P(W_i = 1 \cap W_j = 0)}{P(W_i = 1 \cap W_j = 0) + P(W_i = 0 \cap W_j = 1)} \\
&= \prod_{(i,j)\in\mathcal{M}} \frac{\theta_i(1-\theta_j)}{\theta_i(1-\theta_j) + (1-\theta_i)\theta_j} \\
&= \prod_{(i,j)\in\mathcal{M}} \frac{1}{2} \\
&\propto 1
\end{aligned}
$$

where the penultimate equality follows by the fact that we matched exactly on propensity scores.

□