Reasoning about Interference Between Units

Jake Bowers [*] Mark Fredrickson [†] Costas Panagopoulos [‡]

January 25, 2012

Abstract

This paper demonstrates that, while statements about specific patterns of interference are required for causal inference, there is no single general "no interference assumption" that is required in order to make meaningful statements about causal relations between potential outcomes. A statement about "no interference between units" has long been seen as a fundamental and untestable assumption that is a precondition for meaningful counterfactual causal inference (Cox (1958, p. 19); Rubin (1980, 1986); Brady (2008)). This paper is a proof of concept: We show that "no interference" need not constrain creative researchers who have interesting questions about interference. In so doing, we offer researchers the ability to ask and formalize questions about how treatment given to some units may come to influence outcomes for other units: for example, how treatment may spillover from treated units to control units. We further show that statistical inference about these causal effects is possible, and that the procedures for producing p-values and confidence intervals about causally defined parameters have expected operating characteristics. Finally, we offer some advice, conceptualization, and notation about how to specify the models that represent ideas about how units and treatments might interfere.

The conceptual and methodological framework we develop here is particularly applicable to social networks, but may be usefully deployed whenever a researcher wonders about interference between units. Interference between units need not be an untestable assumption. Rather, interference is an opportunity to ask meaningful questions about theoretically interesting phenomena.

Key Words: Causal effect; Interference; Randomized experiment; Randomization inference; Fisher's Sharp Null Hypothesis; SUTVA

[*] Assistant Professor, Dept of Political Science, University of Illinois @ Urbana-Champaign (`jwbowers@illinois.edu`). *Acknowledgements:* Part of this work funded by NSF Grants SES-0753168 and SES-0753164 (joint with Ben Hansen, Dept of Statistics, University of Michigan). Thanks are due to the participants in the Student-Faculty Workshop in the Dept of Political Science at the University of Illinois at Urbana-Champaign and at the Experiments in Governance and Politics Workshop at MIT (EGAP-6). We especially appreciate the in-depth comments provided by Cara Wong.

[†] Ph.D. Student in Political Science, MS in Statistics, University of Illinois @ Urbana-Champaign.

[‡] Assistant Professor, Dept of Political Science, Fordham University.

1 Introduction

Randomization guarantees unconfounded comparisons and meaningful statistical inference about causal effects. Yet, when those randomly assigned a control condition inadvertently receive some of the treatment simple average treatment effects become difficult to conceptualize and may systematically under- or over-state the unobserved average treatment effect on the treated (see, e.g. Chen, Humphreys and Modi 2010). By common assumption, there is no treatment effect on the controls. Because aggregating over unobserved and observed potential outcomes allows the use of averages for statistical inference about causally defined quantities (Neyman 1923 [1990]), "no interference between units" is thus often described as a foundational precondition for credible causal inference (Cox (1958, p. 19); Rubin (1980, 1986); Brady (2008)). Averaging potential outcomes, however, is only one way that one might use what we know to learn about what we do not know.

Fisher (1935) proposed the use of a sharp null hypothesis test of no effects for this purpose. Whereas Neyman confronted the problem of causal inference by changing the focus of causal inference from units to averages, Fisher approached the problem by specifying unit level hypotheses. Whereas Neyman focused on estimation, Fisher focused on testing. In this paper we use Neyman's conceptualization of causal inference as comparison of unobserved counterfactuals, especially as it has been developed by Rubin (see for example Rubin 2005; Little and Rubin 2000), but show that Fisher's approach allows us to ask questions about interference among units and produce statistical inferences about substantively meaningful quantities parameterizing those questions. Our approach builds on the work of Rosenbaum, as he has done the most to combine the Neyman and Rubin model of causal effects with Fisher's approach (see for example Chap 2 in both Rosenbaum 2010, 2002).

In fact, in this paper we show that "no interference" is not an *assumption* in Fisher's framework, but rather an *implication* that need not always hold.[1] In fact, we already know that Fisher's hypothesis test can be shown to either enable detection of interference (Aronow 2010) or directly enable testing a hypothesis of a particular kind of no effects under unspecified interference (Rosenbaum 2007). We add to this past research showing that one may both directly hypothesize about interference between

[1] We thank Ben Hansen for this terminology. It may also be helpful to think of assumptions as *axioms* and implications as *theorems*.

units and may test such hypotheses.[2]

By posing hypotheses about interference and assessing them in Fisher's framework for statistical inference as developed by Rosenbaum (as linked to Neyman's framework for causal inference as developed by Rubin), our effort is different from, yet complements past and current efforts that have mostly aimed at credible statistical inference *despite* interference or at the decomposition of average treatment effects into parts that are "indirect" (spilled over or otherwise operating via interference) or "direct" (i.e. not operating via interference).[3] There are many variants of this approach, yet they all involve a decomposition of average treatment effects into parts arguably due to interference and parts not due to interference (with often very clever designs making such decompositions meaningful and the variance calculations feasible and valid).

Two close precursors of our method are found in two quite different papers by Rosenbaum (2007) and Hong and Raudenbush (2006). Rosenbaum (2007) enables the production of confidence intervals for causal effects without assuming anything in particular about the form of interference between units. The key to his approach is the idea that the randomization distribution of certain distribution-free rank based test statistics can be calculated without knowing the distribution of outcomes — i.e. can be calculated before the experiment has been run, when no unit at all has received treatment. Rosenbaum (2007) thus successfully enables randomization-justified confidence intervals about causal effects without requiring assumptions about interference. Our aim here, however, is more akin to Hong and Raudenbush (2006). They used a multilevel model to estimate the size of interference effects as they occurred between students nested within schools. We want

[2] Our RItools package for R (Bowers, Fredrickson and Hansen 2010) allows for fast, flexible randomization based inference. All computation in this paper is carried out using a publicly available, open-source, beta version of RItools. Appendix Appendix A provides a brief overview of how one might actually execute our methodological proposals in current software.

[3] For only a few examples of this approach see McConnell, Sinclair and Green (2010); Sinclair (2011); Nickerson (2008, 2011); Hudgens and Halloran (2008); Sobel (2006); Tchetgen and VanderWeele (2010); VanderWeele (2008*a*,*b*, 2009, 2010); VanderWeele and Hernan (2011); Miguel and Kremer (2004); Chen, Humphreys and Modi (2010); Ichino and Schündeln (2011).

to enable statistical inference about particular substantively relevant and theoretically motivated hypotheses about interference and causal effects simultaneously. Hong and Raudenbush (2006) also provide precedent for some of our work here in collapsing aspects of the interference into a scalar valued function. We are not required to collapse the possible avenues of interference in this way, but, in this, our first foray into asking questions about interference, such a simplification makes life much easier. Chen, Humphreys and Modi (2010) provide another model in which Fisher's randomization inference for the sharp null of no effects is used; their sophisticated design assigns units to indirect and direct effects in a controlled manner, and then a Neyman style randomization inference is used to estimate average treatment effects to complement their Fisher style tests of the sharp null of no effects.

This paper is a proof of concept that statistical inference about causally defined quantities is possible when said quantities are defined to allow for interference between units. Other, perhaps gold standard, approaches which begin at the design stage have been developed involving layered or multilevel randomization such that one may, in essence, directly assign units to "indirect" or spillover effects rather than direct effects (McConnell, Sinclair and Green 2010; Hudgens and Halloran 2008; Chen, Humphreys and Modi 2010; Ichino and Schündeln 2011). Our ideas here would complement those designs in that, as we hope will become clear, we would allow scholars to more precisely specify mechanisms for both direct and indirect effects and to assess the support in the data for such mechanisms. We do not offer design advice in this paper.

Finally, as a paper written by social scientists rather than by statisticians, this contribution is not agnostic about the role of substantive theory in the enterprise of statistical inference about causal effects. That is, this paper considers interference between units not as an assumption to be supported with argument and evidence, or a nuisance to be detected and adjusted for, but as an implication of social and political processes to be reasoned about and tested. The conceptual framework and technology which allows us to engage so directly with interference is a consequence of Fisher's sharp null hypothesis (Fisher 1935) and subsequent developments linking Fisher's original ideas with contemporary formal frameworks for conceptualizing causal effects and establishing statistical inferences. Our demonstrations here are meant as a proof of concept: statistical inference about causally meaningful quantities is possible even if we hypothesize about interference directly. A consequence of this demonstration is that it shows that social scientific theory may contribute directly to statistical inference via the specification of meaningful hypotheses whether or not they are about interference.

1.1 Definitions: Assumptions, Implications and Interference

Before diving into the body of the paper, let us quickly elucidate a few important terms.

Two units can be said to *interfere* with each other when the potential outcomes of one unit depend on treatment assigned to another unit. It follows that another definition (which is equivalent) is to say that two units interfere when the potential outcomes of one unit depend on the potential outcomes of another unit (since potential outcomes are defined as reactions to treatment assignment).

An *assumption* is a statement about a truth without proof. An *implication* is a condition following from other decisions including assumptions. In statistics, we can often bring evidence to bear regarding statements of assumptions, but, like hypotheses, assumptions are never proved. The line between assumption and hypothesis can be blurry. For example some linear model analyses assume $y \sim N(\mathbf{X}\beta, \sigma^2)$. Other, linear model uses assume only that the sample size is large enough and observations are independent enough that Central Limit Theorem applies — and this assumption implies the same sampling distribution for $\hat{\beta}$ as that implied by the assumption about the Normality of y: $\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$ (Achen 1982). A hypothesis test for some $H_0 : \beta = \beta_0$ may indicate that $\hat{\beta}$ would be very surprising from the perspective of the hypothesis (so surprising that we might want to reject the hypothesis as entirely implausible). Yet even such information would not address claims about the iid Normality of y or the speed of convergence of the sampling distribution to the Normal.

In this paper, we consider an assumption to be a condition necessary to create a valid test. If a condition is not necessary, then it is part of a hypothesis. For example, in what follows we will assume that treatment is randomly assigned according to the reports of the principal investigator of a study, and from this assumption, implications about the shape of the reference distribution will follow. We will show that Fisher's framework allows one to directly assess hypotheses which imply interference between units.

1.2 Roadmap

We organize this paper around an applied example and a simulation study. To fix ideas and develop notation we first analyze a small ($n = 8$) field experiment in which pairs of US cities were randomly assigned to a get-out-the-vote newspaper advertisement (Panagopoulos 2006; Bowers and Panagopoulos 2011). Then, to explore the operating characteristics of our proposal and to engage with some interesting questions arising about models of hypotheses, we create a medium-sized ($n = 100$) simulated social network experiment which allows us to show that we can recover "true" treatment and spillover effects. We also use this simulation to explore the implications of using models that do not exactly reproduce the data — i.e. "wrong" models. We close with a discussion of how this framework enables the opportunity to ask new questions and more directly engage with theory in an experimental context. We also provide an appendix in which we provide some mathematical language in which to formalize such questions. Such a language becomes important because a common impediment to reasoning about interference has been the sheer number of possibilities that one must engage. Toward that end, we will echo Hong and Raudenbush (2006) in advocating the imposition of structure on the problem:

> Without imposing further structure, the sheer number of causal effects per subject undermines any attempt to summarize evidence in a readily interpretable way. Moreover, a shift in the treatment assignment of any subject changes the potential outcome of any other subject, making it difficult to conceive of average causal effects or to frame interesting questions for policy. (Hong and Raudenbush 2006, page 902)

2 Statistical Inference about Causal Effects in a Small Field Experiment

In a voter mobilization field experiment, Panagopoulos (2006) randomly assigned newspaper advertisements within four pairs of similar cities during the 2005 Mayoral elections. We are accustomed to thinking about the causal effect of such a turnout inducement in terms of a comparison of two partially observed quantities: the turnout, or potential outcome, we would expect to see for city i if that city were treated, $Z_i = 1$, often written $y_{Z=1,i} \equiv y_{i,1}$ and the turnout we would expect if treatment were withheld, $y_{Z=0,i} \equiv y_{i,0}$. If treatment had a causal effect for city i in this experiment then turnout after advertisements would be higher than turnout without advertisements: $y_{i,1} > y_{i,0}$. If treatment had no effect then city i would display the same turnout regardless of treatment condition $y_{i,1} = y_{i,0}$. Treatment could have been assigned differently, but the observed outcome would be identical.

To formalize the idea that treatment may have been assigned within pairs differently than it was, consider the set of all possible vectors of treatment assignment, $\boldsymbol{\Omega}$. Following Rosenbaum (2002, Chap 2) we call an assignment vector drawn from this set $\mathbf{z}$ to distinguish it from our observed random draw $\mathbf{Z} = \{0, 1, 0, 1, 0, 1, 0, 1\}$. So, our shorthand of writing $y_{i,1}$ is really saying that $y_{i,\mathbf{z}=\{0,1,0,1,0,1,0,1\}} = y_{i,\mathbf{z}=\{0,1,1,0,1,0,1,0\}} = \ldots = y_{i,\text{any other }\mathbf{z}}$. For example, if i = "Sioux City", then, when we write $y_{i,1}$ we are saying that the counterfactual outcome under treatment for Sioux City would be the same regardless of the configuration of assignments to any other set of cities. Or, potential outcomes for Sioux City do not depend on the treatment assigned to any other unit. So, notice that the notation that we use to express concepts about counterfactual causal inference implies something about no interference: $y_{i,z_i=1,\mathbf{z}_{-i}} = y_{i,z_i=1,\mathbf{z}'_{-i}}$ for all $\mathbf{z}, \mathbf{z}' \in \boldsymbol{\Omega}$, $\mathbf{z} \neq \mathbf{z}'$. This manner of writing causal effects encodes a decision to ignore the treatment assignment status of other cities. How sensible is this decision?

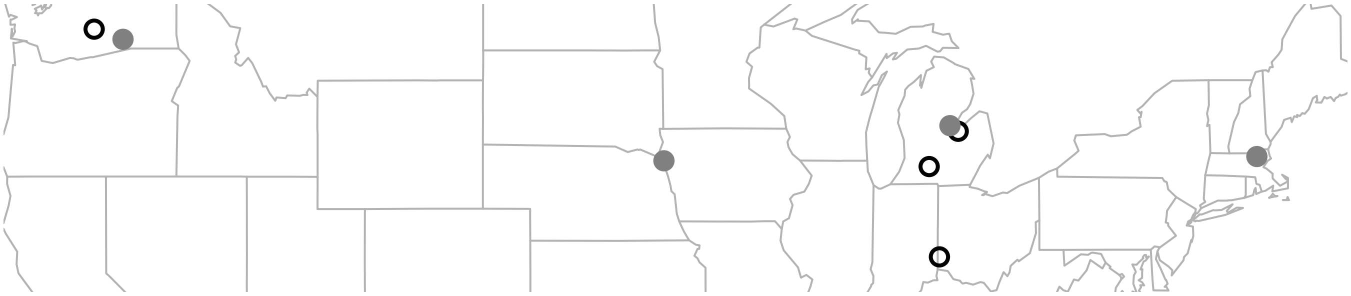


Figure 1: Locations of cities in the US newspapers field experiment. Control cities plotted as open black circles. Treated cities plotted as filled dark gray circles.

Figure 1 shows the locations of the cities chosen for the study. Sioux City is the gray dot alone in the center of the map. Unless communication between Sioux City and other cities regarding political advertisements is quite high, it seems reasonable to suppose that treatment assigned to other cities did not interfere with potential outcomes in Sioux City (or Oxford, OH or Lowell, MA). However, we can see two places on the map where one might imagine newspaper advertisements could spillover between cities: specifically the pairs Yakima-Richland in Washington and Midland-Saginaw in Michigan where the dots are very close. Perhaps the idea of no interference is sensible for some cities in this study and not for others.

To formalize the intuition developed from the map in Figure 1, we develop a mechanism for causal inference in the presence of interference rooted in the work of Fisher (1935) and Rosenbaum (2010, 2002). In the following sections we explain what we mean when we say that one may reason about "no effects" from an intervention in the presence of possible interference. We then show how this same logic can extend to reasoning about the functional forms and parameters of effects of more substantively interesting models than the simple sharp null.

2.1 How can we write a hypothesis about 'no effects and no interference'?

The map above suggested that treatment assigned to other cities probably did not interfere with potential outcomes in Sioux City. If we wanted to consider the hypothesis of "no effects" for Sioux City and we did not want to consider the possibility of interference for Sioux City's potential outcome, then we could write the following hypothesis: $H_0 : y_{\text{Sioux City},1} = y_{\text{Sioux City},0}$. Sioux City's outcomes if treated would be the same as outcomes if not treated regardless of the outcomes of other cities. If we wanted to consider this hypothesis for all cities i we might write $H_0 : y_{i,1} = y_{i,0}$ for all i. These hypotheses imply no interference, from which it follows that we only consider two potential outcomes for each city. And only one potential outcome $((y_{i,1}, y_{i,0}))$ is revealed to us by treatment assignment (Z) such that $Y_i = Z_i y_{i,1} + (1 - Z_i) y_{i,0}$ is an identity linking potential outcomes to observed outcomes (Y). By substituting $y_{i,0}$ for $y_{i,1}$ and simplifying we see that $(H_0, Y_i) \Rightarrow Y_i = y_{i,0}$. That is, the hypothesis and the identity linking observed to potential outcomes together tell us how our observed outcomes would look under the hypothesis. Here, the implication of this pair is that what we observe is what we would observe in the absence of treatment effects, $Y_i = y_{i,0}$.[4]

[4] This is a very abbreviated introduction to Fisher's sharp null hypothesis. See Rosenbaum (2010, Chap2) for a textbook discussion. Keele, McConnaughy and White (2008) introduce and explain these ideas for political scientists in the context of this field experiment.

2.2 How can we write a hypothesis about 'no effects and interference'?

For clarity in our discussion let us restrict attention to the first two pairs of cities. If cities may interfere both within and across assignment blocks, then each of the four cities has four potential outcomes: $\{y_{i,1010}, y_{i,1001}, y_{i,0110}, y_{i,0101}\}$. That is, we imagine that a given city, i, would respond with $y_{i,1010}$ when city 1 and city 3 are assigned treatment, but city 2 and 4 are assigned control. A strong hypothesis of "no effects" would state that turnout is insensitive to treatment assignment across all four cities — the alternative being that any difference in treatment assigned in the system would cause differences in potential outcomes — so we might write, $H_0 : y_{i,1010} = y_{i,1001} = y_{i,0110} = y_{i,0101}$ for all cities. Another strong hypothesis of no effects states $H_0 : y_{i.1010} = y_{i,1001} = y_{i,0110} = y_{i,0101} = y_{i,0000}$ where $y_{i,0000}$ is the response when the intervention was not applied to any units. Rosenbaum (2007) calls $y_{i,0000}$ the potential response to the "uniformity trial", referring to a methodology used to calibrate variance calculations in agricultural experiments by assigning treatment but not applying it. We can write these hypotheses more generally for any vector of treatment assignments, $\mathbf{z}$. We write $y_{i,\mathbf{z}}$ for city i to represent the potential turnout for one possible assignment configuration across all of the cities. Using this notation, our hypothesis of no effects is $H_0 : y_{i,\mathbf{z}} = y_{i,\mathbf{z}'}$ for all $\mathbf{z}, \mathbf{z}' \in \boldsymbol{\Omega}$, where $i = 1, \ldots, I$. The potential outcomes in response to a given treatment assignment vector are those that would have been observed if treatment had been entirely withheld from all cities. We write the potential response to the uniformity trial in our set of four cities as $y_{i,b,\mathbf{z}=\mathbf{0}}$ or $y_{i,b,0000}$.[5]

2.3 Testing the hypothesis of no effects under interference

Having written a hypothesis in which we do not restrict interference between units, we now show how to test either hypothesis of no effects. Let us further restrict attention to only two cities. For city 1, our observed outcome relates to possible potential outcomes via the following equation:

$$Y_1 = Z_1(Z_2 y_{1,11} + (1 - Z_2)y_{1,10}) + (1 - Z_1)(Z_2 y_{1,01} + (1 - Z_2)y_{1,00}) \tag{1}$$

, where Z_2 is 0 or 1 depending on the treatment assigned to unit 2, and Z_1 records treatment assigned

[5] When we analyze the pair-assigned city data, we make use of the paired-assignment mechanism. Yet, we surpress the b subscripts from now on since most of the discussion is not particular to paired or blocked experiments.

to unit 1. The hypothesis of no effects for city 1, $H_0 : y_{1,11} = y_{1,10} = y_{1,01} = y_{1,00}$, implies that

$$\begin{aligned} Y_1 =& Z_1(Z_2 y_{1,00} + (1 - Z_2) y_{1,00}) + (1 - Z_1)(Z_2 y_{1,00} + (1 - Z_2) y_{1,00}) && (2) \\ =& Z_1 y_{1,00} + (1 - Z_1) y_{1,00} && (3) \\ =& y_{1,00}. && (4) \end{aligned}$$

And under the hypothesis of no primary effects, $H_0 : y_{1,10} = y_{1,01}$ (we exclude the "all treated" and "all control", or "uniformity trial", possibilities) so we might just write $y_{1,10} = y_{1,01} = y_{1,*}$ and so,

$$Y_1 = Z_1(1 - Z_2) y_{1,*} + (1 - Z_1) Z_2 y_{1,*}. \quad (5)$$

Now, $Z_1 = 1 - Z_2$ and $Z_1 \in \{0, 1\}$ by design so we can simplify such that

$$Y_1 = Z_1 y_{1,*} + (1 - Z_1) y_{1,*} = y_{1,*} \quad (6)$$

Both of these hypotheses imply that what we observe, Y_1, is what we would observe in the putative world of the hypothesis. There is nothing special about this derivation that prevents it from applying to any number of units or more complex hypotheses (as we shall see).

Once the hypothesis has been formalized and the link between the causal quantities of interest and the observed outcomes has been deduced, the next step is to look at our data from the perspective of the hypothesis. For example, if the treatment assignment had no effect, then the mean of the outcomes in the treated group should be the same as the mean of the outcomes in the control groups such that we would expect no difference between those means. This kind of reasoning suggests using a test statistic, $t(Z, Y)$ to summarize the relationship observed between treated and control units. Here, for example, we might use the mean difference, $t(Z_i, Y_i) = \sum_{i=1}^{n} Z_i \frac{Y_i}{m} - \sum_{i=1}^{n} (1 - Z_i) \frac{Y_i}{n-m}$, where m and n are the fixed number of treated units and total sample size respectively. We use $t()$ to emphasize that it is not the particular function which matters, but rather that we have a test statistic that is a function of treatment assignments and observed outcomes. Calculating $t(\mathbf{z}, \mathbf{y}_{i,\mathbf{z}} = \mathbf{Y} \mid \forall \mathbf{z} \in \mathbf{\Omega})$ we can trace out the distribution of $t()$ under the null of no effects and/or the null of no primary effects (both hypotheses lead to the same null distribution of the test statistic). That is, for each possible way to assign treatment given our design, we calculate the value of the test statistic that would be implied by the hypothesis. The distribution of these calculations indicates the natural variability in the test statistic under the null hypothesis: any differences between different $\mathbf{z}$'s drawn from $\mathbf{\Omega}$ can only tell us about chance variation, not about systematic relationships between treatment and outcomes (at least, this is the fiction we entertain when we do hypothesis tests in general). We follow the norm in defining p-values in relation to these reference distributions (See, e.g. Cox 2006; Rosenbaum 2002). If our observed test statistic appears extreme compared to the distribution generated by re-running the experiment under the hypothesis, then we might say that the hypothesis is very surprising or implausible.[6]

Summary We stated a hypothesis in which all of the potential outcomes implied by unrestricted interference are equal (to each other and to a "uniformity trial"). We deduced what this hypothesis implied for what we observe. We tested this hypothesis by computing a test statistic over all possible randomizations — all possible realizations of the experiment consistent with the design. For each randomization, we adjusted the data in accordance with our hypothesis: in this case of the models of no effects, the hypothesis implies no adjustment (i.e. that $Y_i = y_{i,0}$). In the next section we will need to make actual adjustments to observed outcomes in order to test hypothesis about particular effects. If the treatment really had no effects, then differences from randomization to randomization in the test statistic reveal only the natural uncertainty in our study, allowing us to create a distribution representing this uncertainty to which we can refer when asking about the support our data bring to the hypothesis. Comparing the observed responses to the reference distribution quantifies how surprising our data would be in the world of the hypothesis: we could further decide to reject a hypothesis if the data looked extremely surprising (if we had prespecified a "surprisingness" threshold, often

[6] Although most common functions relating treatment and outcomes may be useful test statistics, only "effect increasing" functions of treatment and outcomes have been shown to produce unbiased tests of no effects against one-sided alternatives (Rosenbaum 2002, Propositions 4 and 5). This requirement that a test statistic be effect increasing means that we would prefer our test statistics to take on higher values as the treatment effect increases. Common test statistics with this property include Wilcoxon's signed rank statistic (which is basically the sum of the ranks of the treated units), differences of means, or other such functions with the general form involving sums of the outcomes of treated units. That is, a good test statistic is one which maps onto the size of the causal effect that we are considering in our models of effects (defined a bit later in the paper).

called α). Rejecting this hypothesis could suggest that some other pattern of responses to treatment are detectable in this dataset. Not rejecting the hypothesis does not mean accepting it.

Applying this test to the newspapers data, using paired mean differences as $t()$ with $\mathbf{\Omega}$ defined by a pair-randomized assignment, we find a one-sided $p = 0.38$ when we consider the null hypothesis of no effects (no effects and no primary effects). From the perspective of the null hypothesis, we would not be that surprised to see a mean difference as large as, or larger than, our observed value. Notice that the sharp null of no effects required no specific statements about patterns of interference for a valid test.[7]

2.4 From "no effects" to "effects"

Asking questions about "no effects" is only the beginning. The motivation for a study tends to involve some expectations about sign and magnitude if not rough value range for a treatment effect. Past experience, literature, and theory of two party elections, all might lead us to expect values of effects for cities exposed to turnout-encouraging advertisements that are definitely less than 10 percentage points of turnout, but also possibly more than 0 percentage points of turnout.

First, consider the simple situation with no interference. If we could hypothesize about no effects by writing $H_0 : y_{i,Z_i=1} = y_{i,Z_i=0}$ then we could formalize a question about a 7 percentage point difference in turnout for city i with $H_0 : y_{i,Z_i=1} = y_{i,Z_i=0} + 7$. A hypothesis about effects is a hunch about how potential responses in the absence of treatment would turn into potential responses to control. In this example, for one city, we entertain the idea that advertisements could add 7 points of turnout to that city as compared to the situation in which that city did not receive advertisements. More generally, imagine a function of potential outcomes to control and possibly other parameters and variables, $h()$, which transforms potential outcomes to control into potential outcomes to treatment. For example, if past literature, experience, and theory suggest that we ought

[7] We also did not rely on assumptions about large-samples, linearity, heteroskedasticy, or Normality. All of the code to replicate every analysis in this paper will be available at [insert web link]. This fact, that the test of the sharp null of no effects requires no specific statement about interference has been known and is made use of by Aronow (2010) and Rosenbaum (2007). Our description above is the most detailed exposition that we have seen to date.

to investigate the idea that treatment with advertisements provides the same turnout boost in all treated cities in our pool, we might write $h(y_{i,Z_i=0}) = y_{i,Z_i=0} + \tau = y_{i,Z_i=1}$. Given our identity linking observed turnout to potential turnout under no interference, $Y_i = Z_i y_{i1} + (1 - Z_i) y_{i0}$, one can make a test for a specific value, of τ, $H_0 : \tau = \tau_0$ under this model using the same process as we did for the null of no effects. The only difference here is that, now $Y_i \neq y_{i,0}$ but rather, when we substitute for $y_{i,1}$ in the identity and solve for $y_{i,0}$ we get $y_{i,0} = Y_i - Z_i \tau_0$, or equivalently:

$$Y_i = y_{i,0} + Z_i \tau_0. \tag{7}$$

Thus, we can now compare $t(\mathbf{Z}, \mathbf{Y} - \mathbf{Z}\tau_0)$ to its distribution across the many ways that treatment could have been assigned as generated by $t(\mathbf{z}, \mathbf{Y} - \mathbf{z}\tau_0 \mid \forall \mathbf{z} \in \mathbf{\Omega})$.[8]

In this simplest case, we consider $0 \leq \tau_0 \leq 10$. If we define "surprising" by $\alpha = 1/8 = .125$ and we discover that our observed values are not so surprising from the perspective of hypotheses as long as $\tau_0 \leq 5$. However, this model makes it strange to see our observed mean differences in turnout for $\tau_0 > 5$ (for example $p = 0.1875$ for $\tau_0 = 5$ but $p = 0.125$ for $\tau_0 = 5.1$). Figure 2 plots the p-values for the range of hypotheses we just tested.[9]

It is one thing to reject the sharp null, but it is much more substantively interesting to know that hypotheses about τ greater than 5 points of turnout in the model $y_{i,1} = y_{i,0} + \tau$ would make our data look very surprising. In order to move beyond the null of no effects, we did have to specify something about the structure of the effects: we considered that it was scientifically useful to say that $h(y_{i,0}) = y_{i,1} = y_{i,0} + \tau$ — perhaps this hypothesis generator follows from theory, literature, or experience even if it is not realistic as a mechanism of treatment effects. Notice that we do not assume that the hypotheses are true, but we can assess the extent to which our data look unlikely

[8] See Rosenbaum (2010, Chap 2), Keele, McConnaughy and White (2012), and Bowers and Panagopoulos (2011) for more in-depth exposition of the basics of inverting Fisher's hypothesis test to produce a confidence interval.

[9] While a 95% confidence interval is conventional, not all designs and data allow for such an α-level. For the newspapers data, there are $2^4 = 16$ possible assignments of treatment due to the blocking scheme. Therefore, even the most extreme one-sided p-value can be no smaller than $\frac{1}{16} = .0625$. The widest meaningful confidence interval here is thus $1 - (2 * (1/16)) \approx 88\%$

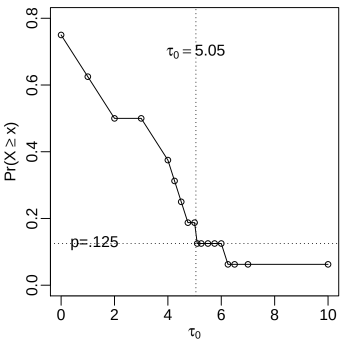


Figure 2: One-sided p-values for hypotheses about τ_0 in the model $y_{i,1} = y_{i,0} + \tau$. The vertical dotted line at $\tau_0 = 5.05$ shows the break between $\tau_0 = 5$ and $\tau_0 = 5.1$.

from the perspective of the hypotheses. The hypotheses are a lens through which we can see our data, or a question we can ask, and the p-values which define plausible ranges of hypotheses are like the answers we give to the optometrist when she places different lenses over our eyes asking “which is better?”. It may be scientifically useful to view our data from perspectives that are not realistic: knowing how we react to strange lenses tells the optometrist information useful to assessing our eyes. Fisher’s testing framework enables us to use our data and design to reflect back on even strange perspectives in a formal manner.

As we are building on the basic mechanism of the previous section, we made statements about the presence or lack of interference between units. As a consequence of the necessary structure, the hypotheses that we used here *implied* no interference among units. Yet, the fact that these models and hypotheses implied no interference does not mean that we cannot ask questions about interference (or that we cannot build lenses which allow us to see the data from perspectives involving interference between units). We make this development in the next section.

2.5 Hypotheses about effects and interference

If we can write $y_{i,1} = h(y_{i,0}) = y_{i,0} + \tau$ then we can write other functions of potential outcomes in the control condition which likewise generate a list or vector of $y_{i,1}$ implied by our hypotheses. Although we might wonder about interference of all possible kinds, the map in Figure 1 suggests two sets of cities as particularly plausible candidates for interference: the pair of Yakima and Richland in Washington State, and Saginaw and Midland in Michigan.[10] This section demonstrates both the flexibility of Fisher's approach to statistical inference, and highlights a way in which social science theory can be crucial for statistical inference — by precise specification of unit-level hypotheses.

Hypotheses about spill-over demand specificity: for the purposes of this example, let us presume that treatment might "spill-over" from a treated unit onto nearby control units but not every control unit is susceptible to spill-over. A reasonable theory would be that news in the larger Yakima, WA is likely to reach the smaller Richland, but not vice-versa.[11] This model specifically excludes treated cities interfering with other treated cities or control cities exterting any influence on other units, though such models would be possible and perhaps substantively useful in other contexts.

We formalize this simple case for two cities and then show how the same logic applies to the entire design. Much of the formalities were done already in § 2.3 so we do not repeat them here. First and second, define the potential outcomes for the two cities (Yakima, j, and Richland, i,) and link those potential outcomes to observed outcomes as shown in equation 1: $Y_i = Z_i\Big(Z_j y_{i,11} + (1 - Z_i) y_{i,10}\Big) + (1 - Z_i)\Big(Z_j y_{i,01} + (1 - Z_j) y_{i,00}\Big)$. Third, express our questions about relations among potential outcomes as a function that transforms observed data, Y_i, to the uniformity trial, $y_{i,00}$. In the no interference case, our hypothesis generators $h()$ were functions that specified how we thought control responses would turn into treated responses. The treatment effects were an additive function of potential treated and potential control responses, with control responses being the baseline against which comparisons would be made. With interference in this simple case, we have two different potential responses to control: $y_{i,00}$ — potential response when no city receives any treatment — and $y_{i,01}$ — the potential response when city i does not receive treatment but city j does receive

[10] This experiment involving first blocking cities into pairs and then randomizing treatment within pair. The treated cities were (in order of pair), Sioux City, Midland, Lowell, and Richland while the control cities were (also in order of pair) Saginaw, Battle Creek, Oxford and Yakima.

[11] Notice that we can entertain this hypothesis even though Yakima was the control and Richland the treated city in the actual study. The hypothesis might not be plausible given the actual administration of the study, but it is reasonable a priori and it might be the kind of hypothesis specified before the study is fielded.

treatment (we might call this the "only spill-over" potential response). Although other conceptions may be possible, in this paper we pursue the idea that the baseline of comparison is the uniformity trial: that is, we want our randomization distribution to represent how cities might have responded when no intervention was applied in the system at all. A minimal conception of "no effects" in a situation with spillover is the situation in which the intervention was not fielded at all.

In this example, we want to investigate the idea that treatment effects may spill-over from a treated unit to a control unit but not vice-versa and that treated units would not interfere with each other (implying that $y_{i,10} = y_{i,11}$). This leads us to simplify equation 1:

$$Y_i = Z_i y_{i,10} + (1 - Z_i)\left(Z_j y_{i,01} + (1 - Z_j) y_{i,00}\right). \tag{8}$$

Now, we must specify how we think $y_{i,00}$ becomes $y_{i,10}$ and $y_{i,01}$. In this case, we write a simple constant and additive treatment effect for cases where $Z_i = 1$, $y_{i,00} + \tau$, but adding a weighted additive spill-over effect (i.e. a proportion of the treatment effect) when $Z_i = 0$, $y_{i,00} + w\tau$, such that our hypothesis generator is

$$h(y_{i,00}, Z_i) = Z_i(y_{i,00} + \tau) + (1 - Z_i)(y_{i,00} + w\tau). \tag{9}$$

Equation 9 gives us $y_{i,10} = y_{i,00} + \tau$ when $Z_i = 1$ and $y_{i,01} = y_{i,00} + w\tau$ when $Z_i = 0$. Here we introduce a weight $w \in [0, 1]$, parameterizing how much of the effect, τ, spills over.

Equations 8 and 9 imply that we can write observed outcomes in terms of hypothesized effects:

$$Y_i = Z_i\left((1 - Z_j)(\tau + y_{i,00}) + Z_j(\tau + y_{i,00})\right) + (1 - Z_i)\left((1 - Z_j) y_{i,00} + Z_j(w\tau + y_{i,00})\right) \tag{10}$$

which simplifies to

$$Y_i = y_{i,00} + \tau Z_i + w\tau Z_j - w\tau Z_i Z_j. \tag{11}$$

Equation 11 allows us to recover $y_{i,00}$ such that

$$y_{i,00} = Y_i - \tau Z_i - w\tau Z_j + w\tau Z_i Z_j = Y_i - \tau Z_i - w\tau Z_j. \tag{12}$$

When $\tau = 0$, we have $y_{i,00} = Y_i$ as implied by Fisher's sharp null hypothesis. When $w = 0$, we have the adjustment implied by the simple constant, additive effects model, $Y_i - Z_i\tau$. The term, $w\tau Z_i Z_j$ is always zero in our design in which $Z_i + Z_j = 1$: at least one of Z_i and Z_j are always 0.

Thus far, we have developed our hypothesis for Richland — a city we presume could receive some spill-over from Yakima. Our hypothesis for Yakima involves the hunch that citizens in Yakima are probably not reading Richland newspapers, but the treatment might act on Yakima as specified by the constant effects model as developed in the no interference cases: $h(y_{j,00}) = h(y_{j,01}) = y_{j00} + \tau = y_{i,11} = y_{i,10}$. For Yakima, the observed outcome identity simplifies even further that it did for Richland such that

$$Y_j = Z_j y_{j,10} + (1 - Z_j) y_{j,00}. \tag{13}$$

We can adjust Y_j to reflect hypotheses about τ just as we did under the no interference model to recover the potential outcome under the uniformity trial, $y_{j,00} = Y_j - Z_j\tau$.

With $y_{j,00} = Y_j - Z_j\tau$ and $y_{i,00} = Y_i - Z_i\tau - Z_j w\tau$, we can assess hypotheses about treatment effects, τ, given a pre-specified spill-over effect, w, or vice-versa, or even assess hypotheses about the two parameters jointly. If we had more than two cities we could set w to be a function of the distance between the cities (or another measure of connectedness). Considering two-way, symmetric interference between two cities k and l leads each city to have the same form of adjustment (shown here for city k only): $y_{k00} = Y_k - \tau Z_k - w\tau Z_l$. Taken together we have a small system of models which formalize some informal theory about how treatment might matter for treated and control cities and about relationships among the cities:

$$h(y_{i,00}) = \begin{cases} Z_i(y_{i,00} + \tau) + (1 - Z_i)(y_{i,00}) & \text{for } i \in \{ \text{ Yakima, Oxford, Lowell, Battle Creek, Sioux City } \} \\ Z_i(y_{i,00} + \tau) + (1 - Z_i)(y_{i,00} + w\tau) & \text{for } i \in \{ \text{ Richland, Midland, Saginaw } \} \end{cases} . \tag{14}$$

That is, we want to consider (a) one-way interference from Yakima to Richland, (b) two-way interference between the two neighboring Michigan cities of Midland and Saginaw, and (c) no interference for the other cities in the dataset. Equation 14 implies the following adjustments to observed outcomes, Y_i to enable the generation of the reference distributions arising from the hypotheses about τ and w that govern equation 14:

$$y_{i,00} = \begin{cases} Y_i - \tau Z_i & \text{when } i \in \{ \text{ Yakima, Oxford, Lowell, Battle Creek, Sioux City } \} \\ Y_i - \tau Z_i - w\tau Z_j & \text{when } i\text{=Richland, } j\text{=Yakima} \\ Y_i - \tau Z_i - w\tau Z_j & \text{when } i\text{=Midland, } j\text{=Saginaw} \\ Y_i - \tau Z_i - w\tau Z_j & \text{when } i\text{=Saginaw, } j\text{=Midland} \end{cases} \tag{15}$$

Equation 15 can be read as saying that the potential outcome to control for city i when no cities receive treatment, $y_{i,00}$, can be recovered by modifying what we observe, Y_i, with a hypothesis about what we do not observe. Specifically, we consider the idea that for four cities, the effect of advertising is simply τ percentage points of turnout. But for other cities we consider the idea that, if specific other cities received the treatment, some proportion of that treatment, w, would be experienced by the city in the control condition.

Models with an interference parameter and a causal effect parameter seem to arise naturally when we consider interference between units. The hypothesis generator introduced in equation 14 involved a parameter w in order to enable us to engage briefly with such the questions. Statistical inference in the presence of parameters like w depends on one's perspective on w. If w is a fixed feature of the design, inference may proceed as done in the previous paragraph setting such parameters at fixed values, or one may consider w as a kind of tuning parameter, and values for it could be chosen using a power analysis or cross-validation. If w is not fixed but is considered a nuisance parameter, then one may produce confidence intervals either by (1) assessing a given hypothesis about τ over the range of w, keeping the hypothesis about τ with the largest p-value (Barnard 1947; Silvapulle 1996) or (2) producing a confidence interval for w and adjusting the largest p-value from a set of tests about a given τ_0 over the range of w in the confidence interval (Berger and Boos 1994; Nolen and Hudgens 2010). Nolen and Hudgens (2010) show that either solution will maintain the correct coverage of the resulting confidence intervals about treatment effects, although using the largest p-value is apt to make those confidence intervals overly conservative.

Yet, parameters like w need not be a nuisance. In this case, we wrote our model so that w could represent the extent of spillover. The approaches to w as a substantive parameter to be estimated involve, in essence, exploring a 2-dimensional slice through the set of possible hypotheses. In fact, we could easily produce a confidence region for w_0, τ_0 pairs which would encode the evidence our data bring against such joint hypotheses.

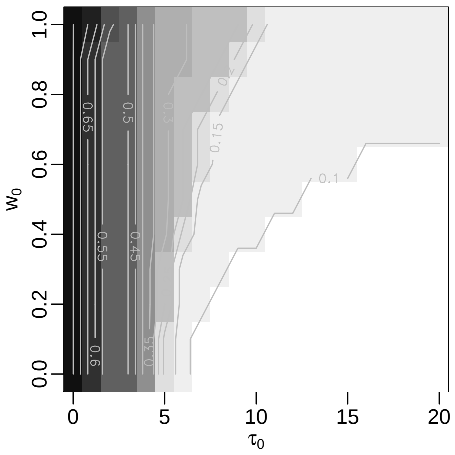


Figure 3: Tests of joint hypotheses about w and τ generated by the model in equation 14. The plot shows areas delimited by the one-sided p-values for the tests of each joint hypothesis. Lower p-values are plotted in lighter color. Higher p-values are plotted in darker colors.

Figure 3 is one representation of such a confidence region using our 8 city data. Equation 14 provides observable implications for our hypotheses, and thus we can test these hypotheses about treatment effects τ and spillover amount w. As before, we use the mean difference test statistic and sweep it over all $\mathbf{z} \in \Omega$ to construct the null hypothesis distributions. For each hypothesized pair of (τ_0, w_0), we can ask, "are the observed responses unlikely from the perspective of the model?" The shading of the plot shows the one-sided p-values from assessing these pairs. The lower bound of the one-sided region is, of course, unbounded, since in this case since we do not consider $\tau_0 < 0$ or $w_0 < 0$. In the absence of spillover (when $w_0 = 0$), hypotheses in the form of equation 14 are rejected for $\tau_0 > 5$ at $\alpha = .125$ — thereby recovering the confidence interval for the non-interference hypotheses.

The figure shows that, as we begin to entertain hypotheses about some positive amount of spillover, the confidence interval expands. This is sensible: if, when treatment is assigned to one unit, most of that treatment is also experienced by another unit, then we have less information available about the treatment effect than we would have if the two units had been independent. Consider the extreme case in which treatment assigned to one unit is fully experienced by the relevant control unit — then we would not have enough information to calculate a treatment effect at all. In this small dataset, we see little to distinguish among different hypotheses about w except when treatment effects are hypothesized to be quite high. For example, at $H_0 : \tau_0 = 6$, we could reject hypotheses about $w_0 < .2$ at $\alpha = .125$ — i.e. when the treatment effect is large, we cannot reject hypotheses in which there is moderate to severe spillover but we can reject hypotheses about low to no spillover.

This simple illustration is merely an example of how far one may push Fisher's framework. It is not an argument in favor of a particular set of hypotheses in this application, though we hope readers will find our choice of a geographically informed model plausible. The main point is that one may reason directly about interference and such reasoning, if formalized, can produce hypotheses about both causal effects and structural features of the effects. The data can provide evidence against such hypotheses. Multi-dimensional hypotheses may be tested to produce substantively interesting and useful confidence regions. The region tells the analyst both about what kinds of values might be implausible under a theoretically informed model and also about the amount of information available to make such plausibility assessments. In the next section we demonstrate these same properties for a much larger simulated experiment with a more intricate network connecting units, but as a simulated network, it is an experiment in which we have less theoretically useful qualitative information about each unit to aid our reasoning about interference.

3 Assessing Machines and Models

An advantage of our approach is that it allows models of the substance of the causal process to have direct implications for data. Of course, the flexibility of this approach is both a blessing and curse: Models are choices, and choices must be justified. Although we cannot say how to choose the true model generating the observed outcomes, we might be able to assess a few interesting models, and, by comparing them, allow them to shed light on the relationships between our theories and our observations.

In this section we use a simulation study to illustrate how one might choose and assess a model that involves treatment effects that depend on network attributes. First, we suggest comparing the size of a test (the probability that it rejects a correct null hypothesis) with its level (the pre-specified probability of rejecting a correct null hypothesis). Rosenbaum (2010, Glossary) calls the level of a test the "promise" of a given error rate. So, this assessment can be thought of as asking whether a given hypothesis testing procedure fulfills (size) its promises (level). Second, we show that confidence regions and intervals contain the "true" parameter values (set by us in advance of the simulation). Third, we offer some graphical methods for inspecting the implications of the models used — both in terms of how the models can make treated subjects and control subjects look alike (when the model is correct) and in terms of how different models may intersect in certain sets of hypotheses.

3.1 The Simulation Setup

Our simply randomized experiment involves a known and fixed social network in which half the units are assigned to treatment. Figure 4 shows this network as a collection of nodes (units) and lines (connections) with colored shapes indicating treatment assignment status. Our simulations use a moderately large sample, $n = 100$, in order to demonstrate the feasibility of this approach when $n > 8$.

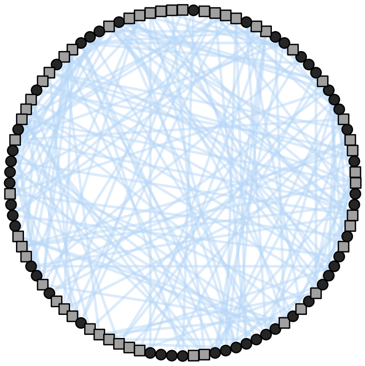


Figure 4: An undirected, random, social network assigned to treatment (circle) or control (square). Probability of a connection between any two units is 0.04 for all units.

We can summarize networks with matrices. In this case, for example, four units in this network have the following $n \times n$ adjacency matrix: $\mathbf{S} = \begin{array}{cccccc} & 6 & 7 & 8 & 9 & 10 \\ 6 & 0 & 0 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & 0 & 1 \\ 8 & 0 & 0 & 0 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 & 0 \\ 10 & 0 & 1 & 0 & 0 & 0 \end{array}$. This matrix shows that unit 7 is connected to unit 10 (and because we have undirected connections, unit 10 is also connected to unit 7). The other units are not connected to each other or units 7 and 10. Summarizing network connections as a matrix allows us to flexibly summarize aspects of the network for focused testing.[12]

In this network, the number of connections ranges from 0 to 12; the number of connections to assigned treatment units range from 0 to 7, with 50% of units connected to between 1 and 3 units. The outcomes in the absence of any intervention in the network are generated simply as draws from a Uniform distribution: $\mathbf{y_0} \sim U(30, 70)$.[13] Following Hong and Raudenbush (2006), we represent network dependence by a scalar function (i.e. we here consider the situation where the effect of the network can be represented by some function summarizing the network with a single number for each unit). Specifically, we consider the situation where the network effect is represented by a function of the number of directly connected treated units. Note that this simplification, while convenient, is not a crucial part of our proof of concept.

We generate outcomes to follow two different models, each of which displays network related dependence in relationships among potential outcomes. As in the previous sections, these models describe the relationship between uniformity outcomes and observed outcomes. Both of our generating models exhibit relatively complex dependence of treatment effects on network attributes. What happens if we assess a model that ignores the network when the actual outcomes depend strongly on the network? As an illustrative response to this question, we contrast these models with the constant additive effects model (which implies no network effects) as defined in Equation 7.

3.1.1 A model with treatment effect dependent on network relationships.

Our first generating model specifies that the number of treated connections have a non-linear and non-monotonic relationship with a constant, additive treatment effect, τ. We set $\tau = 10$ — a large

[12] The use of linear algebra to simplify and represent our ideas about graphs (such as networks) and potential outcomes is expanded upon in the Web Appendix online. That document is primarily useful for those wishing a more general way to write the relationships between potential outcomes and potentially interferring network/graph structures.

[13] One might think of this kind of outcome as a proportion if the units were aggregates (like towns). The procedures for statistical inference used here require no assumptions about the stochastic processes producing outcomes. Thus, we choose a uniform distribution of this sort to link to common political science data types, not because the methods here depend on this distribution.

effect, 1 standard deviation, on the scale of the outcome. The function linking number of treated connections to outcomes in the group assigned to treatment is a cubic polynomial with parameters a, b, and c fixed at -.5, 1, and 7 respectively:

$$f(s, a, b, c) = (a/c)s^3 - as^2 + bs \tag{16}$$

The inner product of the vector of treatment assignments, $s = \mathbf{Z}^T\mathbf{S}$, is a vector containing the number of treated connections for each unit in the study.

Equation 17 shows how observed outcomes are generated additively from the uniformity trial using different versions of the linking function in Equation 16: units assigned to control receive only $\tau = 10\times$ the number of treated directly connected other units while the those assigned to treatment get some non-linearly changing proportion of τ depending on the number of directly connected treated other units.

$$\mathbf{Y} = \mathbf{y_0} + \mathbf{Z}\tau f(\mathbf{Z}^T\mathbf{S}, -.5, 1, 7) + (1 - \mathbf{Z})\tau\mathbf{Z}^T\mathbf{S} \tag{17}$$

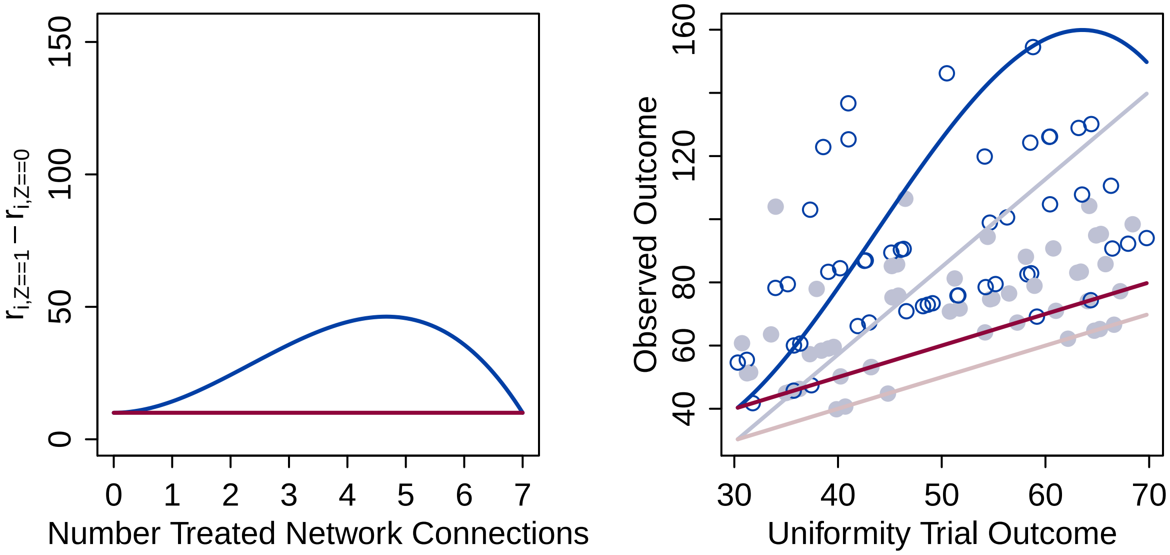


Figure 5: Implications of models graphed by number of treated connections (left) and Uniformity trial (right). The left panel shows treatment effect (a difference in potential outcomes) implied by either $y_{1i} = y_{0i} + \tau$ (the flat line) or the model in equation 17 (the curved line). The right panel shows the same functions as applied to the $n = 100$ draws from $\mathbf{y_0} \sim \mathcal{U}(30, 70)$ representing the uniformity trial. The high blue lines (one curved, one straight) arise from equation 17. The less steeply sloping red lines at the bottom of the panel show the implications of the model of constant treatment effects.

Figure 5 shows how equations 16 and 17 combine with treatment assignment and network characteristics to (1) transform the potential outcomes under the uniformity trial potential outcomes to treatment (the left panel) and (2) transform potential outcomes under the uniformity trial to observed outcomes (the right panel). The idea of the model in equation 17 is that the number of connected treated units amplifies treatment in the treatment group up to a point after which the saturation of connected treated units begins to have a negative effect. In contrast, spillover onto the controls is additive and monotonic in the numbers of connected treated units.

3.1.2 A model with a treatment effect and a spillover parameter.

Although the one-parameter model allows for treatment effects to vary nonlinearly with network attributes, it does not easily allow us to answer questions about the existence of such effects per se. To illustrate how one might specify a model in which a no-interference model is nested within a model of interference we specify a two parameter model. The functions linking number of treated connections to outcomes, equations 18 and 19, are simple nonlinear functions: $g(s)$ approaches 1 as $|s| \to \infty$ and so can be understood as a proportion of the treatment effect derived from treated neighbors.; $f(s)$ approaches 0 under the same conditions. Both are bounded between 0 and 1.

$$g(s) = 1 - \frac{1}{1+s} \tag{18}$$

$$f(s) = \frac{1}{1+s} \tag{19}$$

Like equation 17 for the first model, equation 20 shows how observed outcomes are generated additively from the uniformity trial. For this model, the story involves both a network spillover effect (τ_2) and a direct treatment effect (τ_1). Units in the control group do not get the additive direct effect of τ_1 but get some proportion of τ_1 depending on τ_2 and a function of the number of directly connected treated units.

$$\mathbf{Y} = \mathbf{y_0} + (\mathbf{Z}(\tau_1 + \tau_1\tau_2 g(\mathbf{Z}^T\mathbf{S})) + (1 - \mathbf{Z}) * (\tau_1\tau_2 g(\mathbf{Z}^T\mathbf{S}))) \tag{20}$$

When $\tau_2 = 0$, this model of effects reduces to the constant additive effects model. Thus, one can compare both the results of the constant additive effects model (our naive/simple model of choice in this paper) to the more complex model which generates the data.

Figure 6 shows the implications of this model as compared to the constant additive effects model. Potential responses to treatment would always be larger than potential responses to control, but

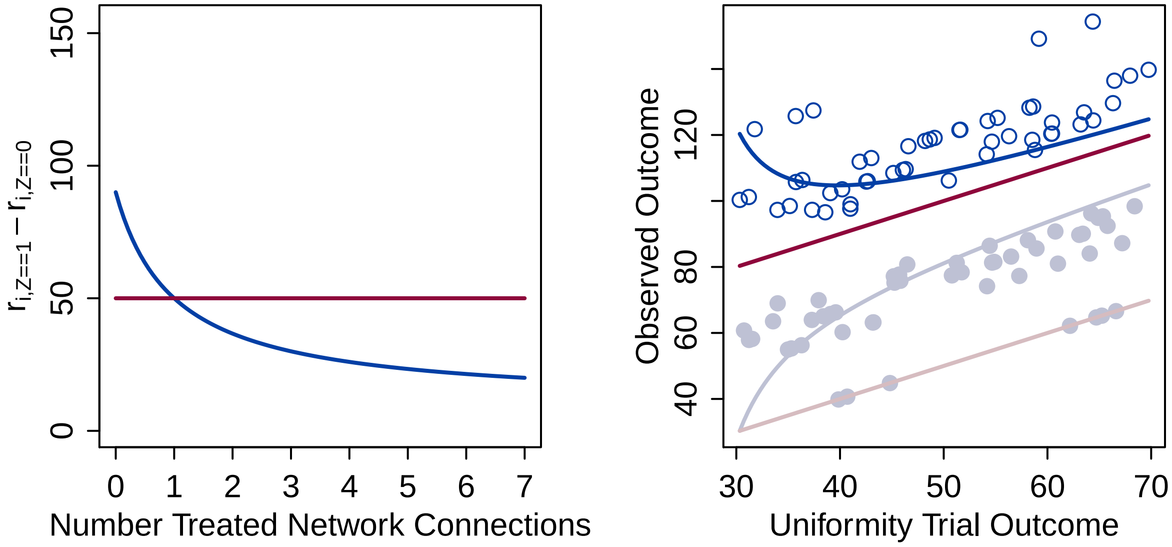


Figure 6: Implications of models graphed by number of treated connections (left) and Uniformity trial (right). The left panel shows treatment effect (a difference in potential outcomes) implied by either $y_{1i} = y_{0i} + \tau$ (the flat line) or the model in equation 20 (the curved line). The right panel shows the same functions as applied to the $n = 100$ draws from $\mathbf{y_0} \sim \mathcal{U}(30, 70)$ representing the uniformity trial. The curved blue lines arise from equation 20. The straight red lines show the implications of the model of constant treatment effects.

this difference would decrease as the network local to a unit saturates with other treated units. The marginal effect of the network diminishes for both treated and control units, although the marginal effects change rapidly when the unit is connected with few other treated units.

3.2 Right Model, Right Parameter: The Size of Tests

When we generate data with a known model with known parameters, "true" null hypotheses should be rejected no more than $100\alpha\%$ of the time at level α. That is, when we have the right model and the right parameter, we still will reject right hypotheses, but we want to do so in a controlled manner and rarely. Although knowledge that the size of the test is the same or less than the level is nice, it is really only a minimal condition. We do not tend to know either the model or the parameter. Thus, we also here present the results of using models that ignore network characteristics when the true outcomes are generated by the model presented above (and the results of using the network models when the true outcomes come from the constant effects model).

Our simulation study uses 1000 repetitions of the following algorithm, based on the fixed uniformity trial outcomes generated by $\mathbf{y_0} \sim \mathcal{U}(30, 70)$.

1. Draw a vector of treatment assignments from the set of possible assignments consistent with the design (here, simply, having 50 1s and 50 0s.). Consider this vector the "temporary real" assignment.

2. Generate a set of observed outcomes from the fixed uniformity trial following the "temporary real" assignment and the "right" model (i.e., equation 17 or 20).

3. Assess hypotheses about the parameters of the true model and about the parameters of the wrong model (where the wrong model is the constant additive effects model in both cases). For example, for the two parameter model, we assess all the integer hypotheses for which $10 \le \tau_1 \le 50$ and $10 \le \tau_2 \le 50$ (in all, about $50 * 50 = 2500$ hypotheses total).

Size is calculated from the proportion of "true" hypotheses rejected in step 3 at the range of α levels.

Figure 7 shows in the right hand column the proportion of correct hypotheses rejected for every $0 \le \alpha \le 1$ level. We can see that when our model has strong support in the data the test fulfills its promises whether we are testing a true τ or a true joint $\{\tau_1, \tau_2\}$. When our model has little support in the data, the test rejects the true hypotheses nearly always as shown in the left and middle columns of the figure.

3.3 Interpreting Models of Interference

When we test a range of hypotheses about parameters made meaningful by a given model of effects, including interference, how should we interpret the results?

Figure 8 uses two-sided p-values to summarize the evidence against hypotheses generated from these models. The curve on the left panel intersects with the horizontal lines at $p = .05$ and $p = .1$ — the boundaries of 95% and 90% confidence intervals. We can see that the correct value is at the center of the interval: the p-values quantify the idea that our data would not be surprising if τ were 10 and our model were that defined by equation 17. The right panel plots two-sided p-values for pairs of hypotheses about τ_1 and τ_2. The correct value of $(\tau_1 = 50, \tau_2 = .8)$ is in the middle of the region. We show contour lines at $p = .05$ and $p = .1$ to illuminate the boundaries of a two-dimensional confidence region. In this model, it is τ_1 which really drives differences between those assigned treatment and control. The parameter governing the network effects contributes

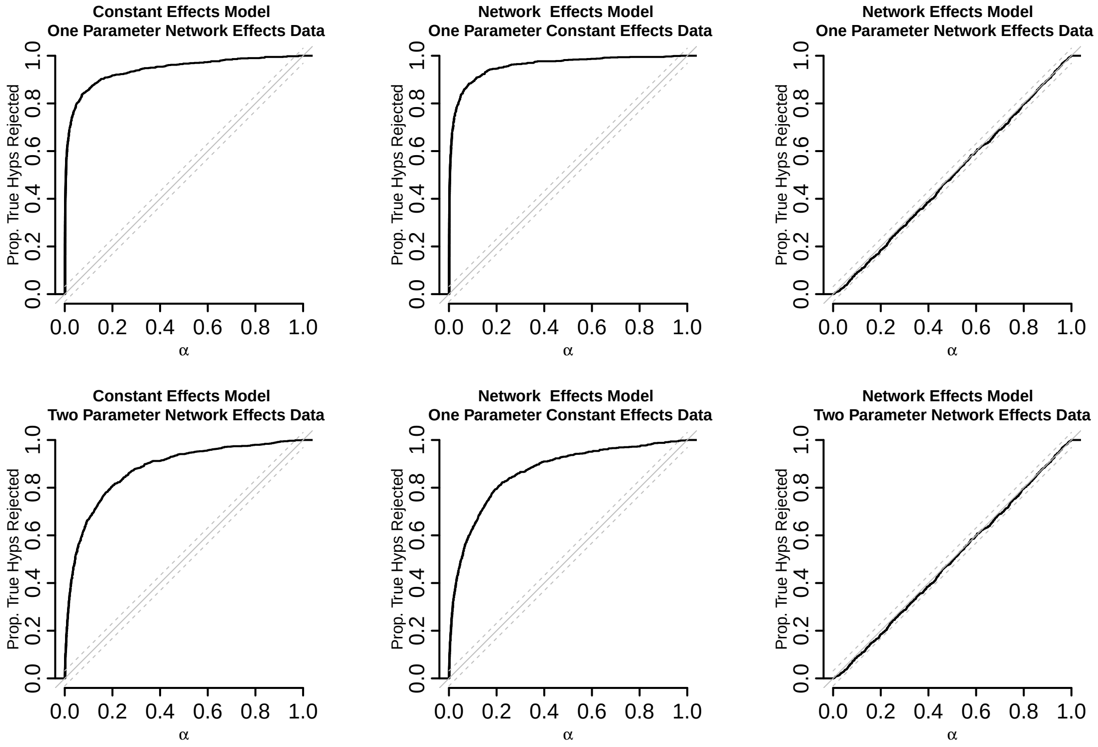


Figure 7: For simulated data (n = 100, over 1000 simulations) with network moderated treatment effects, the proportion of true null hypotheses rejected at level α for all α. The panels labeled "Network Effects Model …Network Effects Data" show the correct models assessing the true parameters. The other panels (in which the models are not "right") reject the "truth" always, whereas the "right" models reject just about 100α % of the time (within the simulation error range — shown by the dashed gray lines.)

relatively little to this difference (compared to τ_1). So, we see that the region is in the shape of a line: many hypotheses about interference are compatible with these data as long as they are paired with a relatively small range of τ_1. That is, one interprets confidence regions just as one would interpret confidence intervals — as containers for hypotheses that are difficult to reject. As we well know, such containers (intervals or regions) do not tell us the location of the truth, but rather tell us what we can learn from a given design. We can learn that our design does not support precise answers to questions of particular forms (as we learn here for τ_2) or that our design supports precise answers (as we learn here for τ_1 and τ).

3.4 Comparing Wrong Models

The methodology presented here does fulfill the minimal criteria that we tend to expect from our statistical procedures: Figure 7 shows that the size is less than or equal to the level of the test (and by implication the truth is inside a confidence interval). And we have also discussed how one

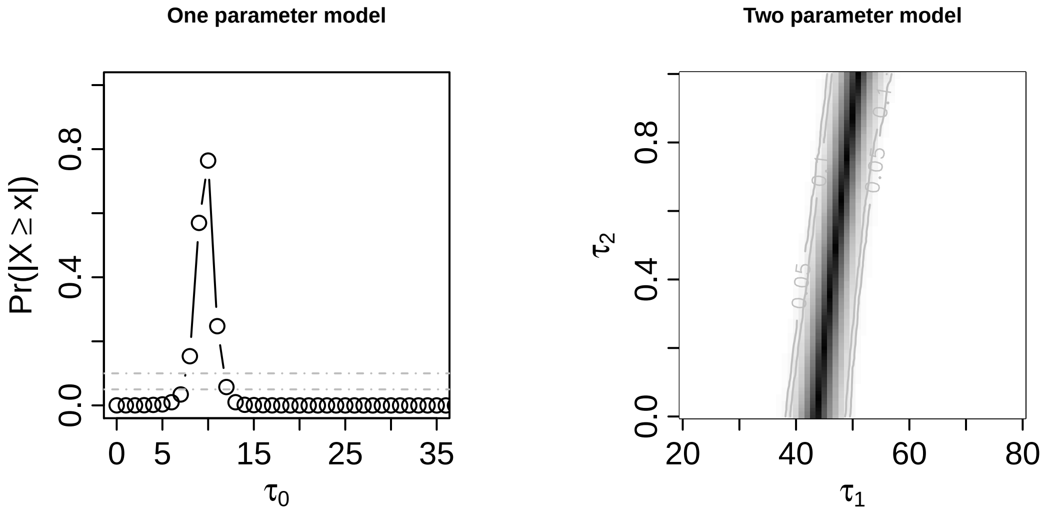


Figure 8: The left panel shows two-sided p-values (on the y-axis) plotted for hypotheses about τ as defined in equation 17 with $\tau = 10$. The right panel shows two-sided p-values (on the z-axis as colored regions and contour lines) for hypotheses about τ_1 and τ_2 as defined in equation 20 with $\tau_1 = 50, \tau_2 = .8$.

might interpret the results of using such models with results shown in Figure 8 (in the fortuitious circumstance that the model agrees very strongly with the data).

The comparison of "right" to "wrong" wrong models is somewhat unfair, of course. For example, the meaning of τ differs between the two one-parameter models. In the constant effects model it refers to an overall constant and additive shift in the potential outcomes. In the nonlinear network effects model, τ is still an additive difference between the treated and controls, but it is moderated by the network attributes of the person. The model determines the meaning of the parameter. Perhaps we should not have even named the parameter with the same symbol in both models. Yet, we did so in part because we wanted to lay the groundwork for an important point about models of effects: two models may be isomorphic at least for certain hypotheses. Recall that the ingredients of testing specific, sharp null hypotheses are: (1) hypothesis generating functions or models of effects, $h()$ (e.g. $h(y_{i,0}) = y_{i,1} = y_{i,0} + \tau$) which pose a question in terms of counter-factuals and give causal meaning to one or more parameters like τ and (2) observed outcome identity equations (e.g. $Y_i = Z_i y_{i,1} + (1 - Z_i) y_{i,0}$) which link potential outcomes to observed outcomes via treatment assignment. Pairs of models and identities allow tests of hypotheses about parameters because the pairs, if correct, have observable implications: here, for example, if $H_0 : \tau_0 = 10$ is correct and the model is correct, then we can remove the effect of the treatment from Y_i and recover $y_{i,0}$: $y_{i,0} = Y_i - Z_i\tau$ with τ set to 10. Probability statements about H_0 arise from applying a test statistic to the adjusted outcomes to compare the observed test statistic $t(Z_i, Y_i - Z_i\tau)$ to the distribution which we would observe if the experiment were repeated. It is conceivable that there are certain hypotheses from the constant effects model which imply the same adjustment to outcomes as the polynomial network effects model: that is, even if the constant effects model is "wrong" when applied to the data generated by a network effects process, certain comparisons of treated to control groups may be identical for some hypotheses and test statistics. And, the distributions of responses to treatment and control may be brought into alignment via multiple mechanisms. In this short section we consider two ways to compare models: first in regards to a concept of the distance between models for a given unit and second in regards to the differences in distributions across units implied by their adjustments. In both cases we emphasize graphical methods.

3.4.1 Competing models, complementary answers

All models used in this paper are functions that map the potential responses under the uniformity trial to potential responses to treatment and control. The known design of the study, the models and the identity equation linking potential outcomes to observed outcomes allow us to test hypotheses. We can think of the model of potential outcomes and the observed outcomes identity as combining to produce a function that converts responses in the uniformity trial into observed responses. For example, let f be a model with possibly vector valued parameters $\boldsymbol{\theta}_f$ and let g be a different model with parameters $\boldsymbol{\theta}_g$:

$$\mathbf{Y}_f = f(\mathbf{y_0}, \mathbf{Z}, \boldsymbol{\theta}_f), \tag{21}$$

$$\mathbf{Y}_g = g(\mathbf{y_0}, \mathbf{Z}, \boldsymbol{\theta}_g). \tag{22}$$

Under the right set of parameters $(\boldsymbol{\theta}_g, \boldsymbol{\theta}_g)$ different models will map the uniformity trial to the same observed data, i.e. $\mathbf{Y}_f = \mathbf{Y}_g$. The trivial example is the sharp null of no effects. All models share this hypothesis for at least one vector of parameters. In the case of the additive models used in the previous simulations, these models reduce to the sharp null when $\tau_i = 0, \forall i$. At this value of τ, the model implies no adjustment to the observed data. In the multiplicative model, $\mathbf{Y} = \delta\mathbf{Z}\mathbf{y_0} + (1 - Z)\mathbf{y_0}$, the sharp null of no effects is represented by $\delta = 1$.

While all models, suitably parameterized, can represent the sharp null hypothesis of no effects, it is more likely that two models will only agree in adjusted outcomes for certain values of their parameters. Consider the trivial case of a uniformity trial where all units have unit outcome: $\mathbf{y_0} = \mathbf{1}$. For these data, the additive model will imply the same adjustment as the multiplicative model when $(\tau = 1, \delta = 2)$: $\mathbf{Y}_{\tau=1} = \mathbf{1} + \mathbf{Z} = 2\mathbf{Z}\mathbf{1} + (1 - \mathbf{Z})\mathbf{1} = \mathbf{Y}_{\delta=1}$.

For more complex designs and models, these relationships between models may not be immediately apparent, but can be deduced by applying the models to the data and comparing the discrepancy via a loss function. For example, we might define D to be the Euclidean distance between the predictions made by two different models such that

$$D = \sqrt{\Big(f(\mathbf{y_0}, \mathbf{Z}, \boldsymbol{\theta}_f) - g(\mathbf{y_0}, \mathbf{Z}, \boldsymbol{\theta}_g)\Big)^T \Big(f(\mathbf{y_0}, \mathbf{Z}, \boldsymbol{\theta}_f) - g(\mathbf{y_0}, \mathbf{Z}, \boldsymbol{\theta}_g)\Big)}. \tag{23}$$

Notice that this comparison is made *at the unit level*, rather than at the level of the distribution. Two models may imply similar changes in distributions of outcomes, but have very different unit level implications. For example, a given $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_f$ may imply the same shift in the mean of the outcome distributions but imply different ranks assigned to the individual units.

Let us unpack this idea of unit-level distance between models. Consider two matrices A and B containing the adjustments implied by two models, also A and B. Each one is $p \times n$ where p is the number of different hypotheses specified (assume, for now, that we are only comparing two models, each with a single parameter). If we subtract the first column of A from the first column of B we have, for unit 1, the differences in adjusted outcomes implied by the different hypotheses for model A versus model B. The sum of the squares of these differences is a measure of the overall difference between model A and model B for unit 1 (given the hypotheses). The square root of this sum of squares is an Euclidean distance, or straight line distance, in the units of the outcome between the two points in n-dimensional space where n refers to the total number of units. If we had only two units, and the two models of interest implied $\{10, 11\}$ and $\{20, 30\}$ for the two units respectively, then we would ask about the distance between the point (10,11) and the point (20,30) — which would be $\sqrt{(10 - 20)^2 + (11 - 30)^2} = 21.47$ in the same units that we measured. If we had more units, then our sum inside the square root bracket would get longer, but would still return a single number.

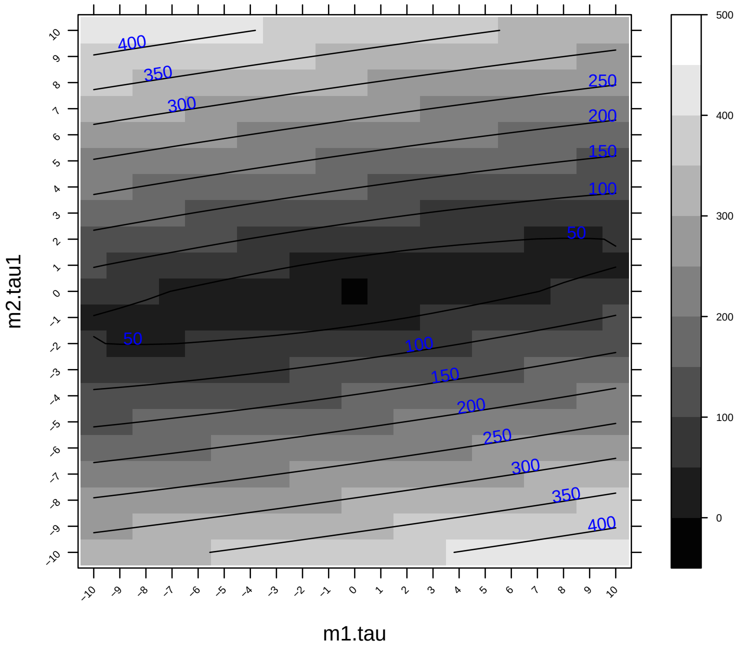


Figure 9: Euclidean distance between the constant additive effects model (m1) and the one parameter interference model (m2). The data arise from the interference model from equation 17 with $\tau = 10$ and the network is as shown in figure 4. When effects are small, the two models imply very similar adjustments to the data.

Figure 9 shows the distance between the constant additive model and the one parameter interference model (17) for a range of hypotheses relevant to each model (here the hypotheses happen to be the same but need not be the same) as applied to data generated by the polynomial network interference model shown in figure 4. As we would expect, both models produce the same predictions for all units when $\tau = 0$ (as shown by the darkest colored square at the 0,0 grid location). The dark shades at the center of the plot (where treatment effects are small) depicts a situation where the two models more or less agree. As treatment effects increase in size the two models imply ever more distinctive patterns in the outcomes. As they attempt to recover the uniformity trial, each model has to adjust the outcomes more severely as the treatment effect becomes larger, and these adjustments differ from each other more and more.

Figure 10 shows the relationship between the two parameter network effects model and the one parameter constant effects model nested within it. When $\tau_2 = 0$ (the panel labeled "m2.tau2=0" in the lower left hand part of the plot), the two models agree exactly at all hypotheses (the black

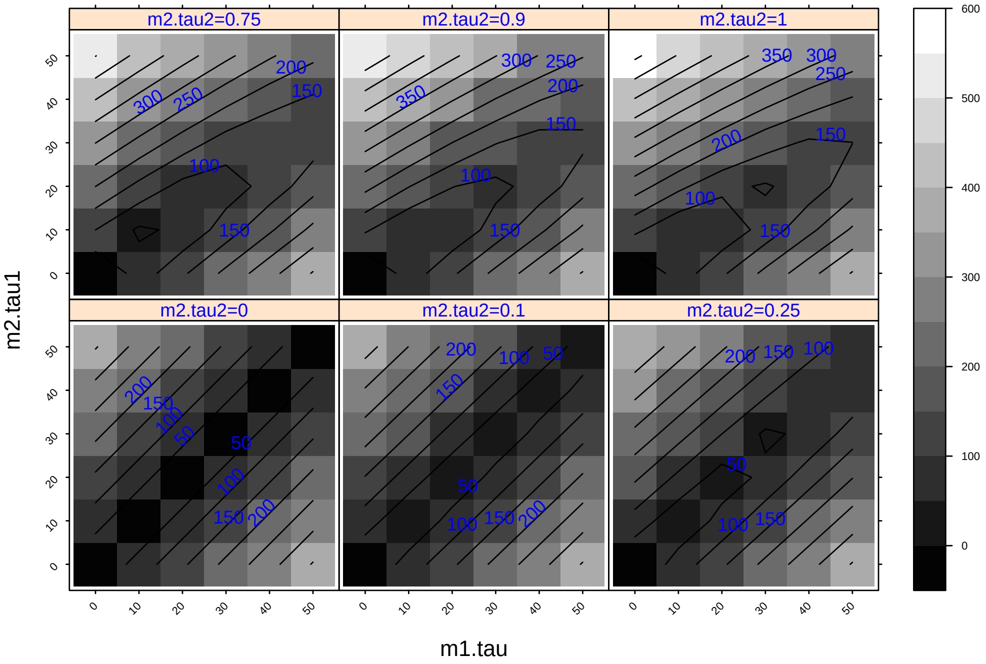


Figure 10: Euclidean distance between the constant additive effects model (m1) and the two parameter interference model (m2) used in the simulation. The data arise from the interference model from equation 20 with $\tau_1 = 50, \tau_2 = .8$ and the network is as shown in figure 4. The two parameter model is the same as the constant effects model when $\tau_2 = 0$ ('m2.tau2=0') and when $\tau = \tau_1$ ('m1.tau'='m2.tau1').

squares on the diagonal show this). In that plot, the models differ only when the hypotheses differ (for example, the panel labeled 'm2.tau2=0' shows that two constant effects models will differ from each other if one is using an effect of 10 and the other an effect of 20). The models also agree exactly, regardless of the value of τ_2 when both treatment effects are zero (the black squares at the bottom left corner of each panel).

When models imply the same adjustments to the data, the *p*-value for both hypotheses will be the same if we use the same test statistic. Such a result would indicate that the data are equally surprising from the perspective of both models at these hypotheses. Consequently, models that make similar predictions at specific hypotheses provide opportunities to interpret results under either theory. Even though the interference model performs better overall on the interference-model generated data, sometimes a simpler model may be useful especially if it is compared with one or two other models.

3.4.2 Competing models, contrasting "fit"

Although we do not have a well-defined concept of model "fit" in this framework, we still might ask whether a given model implies adjustments to the data which do, in fact, bring the control and treated groups into alignment. One method for assessing models in this way was suggested by Rosenbaum (2009, Chap 2) in which distributions of outcomes after and before application of models were displayed using boxplots. We build on that method here. Although we do not know the "true" model, it may be useful to see how the application of a given model moves the distributions of the control and treated groups toward each other (indicating successful removal of a difference between the two groups).

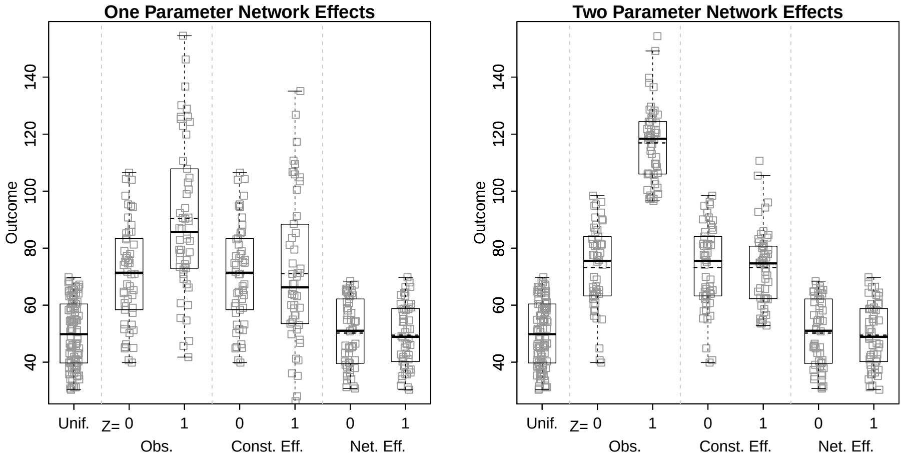


Figure 11: Adjusting observed data using the least surprising hypotheses for each model in the simulations (the one parameter simulation is the left panel, the two parameter simulation is the right panel). Solid horizontal lines with boxes are group medians, dashed horizontal lines are group means. Uniformity trial is on the far left in each panel. Observed outcomes ('Obs.') are plotted next to the uniformity trial ('Unif.'), followed by a "wrong" model ('Const. Eff') and finally the "right" model ('Net. Eff') is shown at the far right column within each panel.

Figure 11 presents this graphical indication of model performance. This plot shows the distributions of outcomes in the uniformity trial, the "observed" data generated from the different simulation models, and the adjustments implied by the least surprising hypotheses for each model. Recall that each hypothesis implies a different adjustment to the outcomes. Here we take the hypothesis that is as close as possible to the center of the confidence intervals that we presented earlier, that is, the hypothesis with the maximum observed p-value. Such a hypothesis is least surprising or hardest to reject and represents one way to think about a point estimate in this framework.[14]

[14] This way of thinking about point estimates is rough and useful for graphs but does not have theoretical grounding. The Hodges-Lehmann point estimate is the more common and well developed method for producing point estimates from hypothesis tests (Rosenbaum 1993; Hodges and Lehmann

The model comparison for the one parameter network effect simulation data occupies the left plot. We see the uniformity trial data (known here since we are using a simulation) at left. To the right within that plot we see the treatment-vs-control comparison for the simulated observed data: the treated group is higher in distribution than the control group. Next, we show the results of applying the constant effects model using the mean difference test statistic. As expected, this model and test statistic equalizes the means of the two groups by shifting the treated group down. The distance between the original mean of the treated group and the mean of the treated group after shifting it down to the control group is the least surprising hypothesis. We still see that the variances of the two groups are very disimilar as are their medians. The right most column in this panel shows the results of applying the correct model to the observed data. Now, the control and treated groups look like two random samples from the uniformity trial.

The right panel shows the comparison for the two parameter simulation data. Again, the constant effects model acts by moving the treated group to align means with the control group whereas the networks effect model involves shifting both the control and treated groups to be equal to each other.

In practice, of course, researchers will not have access to the uniformity trial, but researchers can still compare observed and model adjusted distributions for any dataset using this technique. Comparing multiple “wrong” models in this way would illuminate the ways in which the models leave aspects of the distributions of the experimental groups unexplained/unaligned.

4 Where do models come from?

When one is familiar with reporting average treatment effects, sitting down to specify a model of specific, unit-level effects may cause an experience not unlike the moment of sitting down to

1964, 1963). We focus on least surprising hypothesis rather than Hodges-Lehmann point estimates because a different closed-form solution for the HL-estimate arises from each combination of model of effects and test statistic (the HL estimate for the difference of means test statistic with a constant effects model, for example, is simply the observed difference in means. But this simplicity is not guaranteed for other models and test statistics). In our experience, the value of the parameters with the maximum p-value (or the median of the set of maximum p-values) tends to be the same as the HL-estimate or very close.

confront a blank page at the start of a writing project. How should one start? The blank page scares us because it is a glance at infinity: nearly any combination of words may be written down. Since writing is a high stakes activity for academics, the knowledge of impending judgment combined with the realization of the nearly infinite possibilities can paralyze. Yet, even if the blank page is scary, we either figure out how to cope or leave the writing life for another profession.

We not only face the need to make decisions when we write, but also when we plan an experiments and analysis. The average treatment effect may be a comforting default, but, of course, in the presence of interference without some clever design or model by which the average is decomposed it is of no use.

Fisher's sharp null hypothesis encourages us to confront infinity with specificity in more or less the same way that we do so in planning, design, and writing. After all, it seems overwhelming to consider all of the possible ways that any given treatment could have an effect on all of the units in a study. And allowing ourselves to think about such effects while allowing also for interference between units may appear to court insanity. Yet, recall that each and every researcher is always confronting infinity during research design, data collection, and data analysis. Research involves engagement with details, and if the devil lies not in the details, at least infinity hides there. Thus, the fact that we must make decisions in the face of infinity is something common: we use past decisions (i.e. "The Literature") or current observations or past or current theory to help constrain the general boundaries of a research project. And we are well used to justifying our current decisions. We are always and everywhere making certain choices. It is to help us make scientifically interesting choices that we read thousands of pages in graduate school, for example. Formalizing certain putatively scientifically interesting choices to enable discussion and criticism in the form of testable hypotheses ought to enhance scientific communication and research accumulation.

The social organization of science is, in fact, designed to help us carve narrow paths through the enormous thicket of decisions that always face us. No scholar can claim to have made an "ideal" set of decisions, just as no writer can claim to have written the "best" paper. The question is never whether a given paper is best in an absolute sense, but whether the large set of decisions by which the researcher winnowed down the infinite set of possibilities allows us to understand something new and useful about the world.

From this perspective, Fisher's sharp null is no different. We should not fear it. In fact, the method we developed in this paper merely gives us more possibilities for scientifically useful, justified, decision making. The fact that one scholar assesses one set of hypotheses does not preclude others from assessing another set — and may even require the assessment of a future set of hypotheses from the questions raised in the original research. Readers will want to know why this set of hypotheses were assessed and not a few obvious others, yet, again, this is no different from explaining the decisions behind the design and administration of the experiment itself.

As we write models, we can certainly go too far. For 100 units, we could create a model with 100 different effects. Testing every possible combination of those effects would give us a 2^{100} hypothesis vector and a 100 dimensional space of hypotheses tested (for a binary treatment). Canvassing a collection of such hypotheses is intractable. Moreover, it is not clear that such an effort would be fruitful. As Rosenbaum notes in his discussion of this very topic, "…it is straightforward to make valid statistical inferences that are so complex, so faithful to the minute detail of reality, that they are unintelligible and of no practical use whatsoever" (Rosenbaum 2010, 45). So we must simplify with our models. Referring to our 100 dimensional space as $2I$, and a parameter θ, Rosenbaum further explains:

> In this sense, a 1-dimensional model for the 2I dimensional effect, such as the constant effect model, may be understood as an attempt to glean insight into the 2I dimensional confidence set for θ while recognizing that any 1-dimensional model, indeed any intelligible model, is to some degree an oversimplification. Understanding of θ is often aided by contrasting several intelligible models, rather than discarding them. Arguably, the joint consideration of three 1-dimensional models for the 2I-dimensional parameter θ provides more humanly accessible insight into θ than would a 2I-dimensional confidence set. (Rosenbaum 2010, page 45)

Writing several simple models may be superior to a single complex model. Selecting which simple models and just how simple is an opportunity to engage with theory. Explaining how simple models glean insight into complex realities is an opportunity for future work.

Where do hypotheses and models come from? This question is well beyond the scope of this paper although it is important (Clarke and Primo 2012). As a statistical methodology paper, we have chosen to focus on a few models of effects and empirical examples in the hope that our proof of concept stimulates others to apply these ideas to their own work. One aspect of this process which we elided in the interests of space and time is the question of the role of theory in the specification of sets of hypotheses. It is obvious that specific statements about counterfactuals which produce predictions for all units appears to be something which formal theory of all kinds is well situated to help provide. Thus, although there is much work to be done on the statistical methodology side of this approach, perhaps the most profound impact of this work will be to offer a new way to allow theory to speak with data, a new way to ask interesting questions.

5 Conclusion

When treatments given to one unit may change the potential outcomes for another unit, the consequences of ignoring interference may be serious. Imagine a development project aiming to assess a policy as applied to different villages in need of aid. If members of control villages communicate with members of treated villages, then we will have trouble advising policy makers about whether the policy should be rolled out at a large scale. We have long known, in fact, that the average treatment effect is not even well identified or meaningful under interference (Cox 1958).

So far attempts to enable statistical inference about treatment effects with interference have taken for granted the average treatment effect framework and worked to partition the average into parts attributable to interference and parts attributable to direct experience with the treatment. In this paper, we propose a different approach based on asking direct questions about specific forms of interference. Fisher's test of the sharp null is still meaningful even when each unit may have many potential outcomes due to interference. Additionally, Fisher's framework allows detection of interference (Aronow 2010), and under certain conditions, allows the creation of intervals for hypotheses about treatment effects without requiring specific statements about the form of interference (Rosenbaum 2007). Our paper contributes to this literature by showing how one may directly specify and assess hypotheses about theorized forms of interference. We also show that one may produce confidence sets that illuminate the information contained in a dataset regarding different combinations of hypotheses about interference and treatment effects. This form of statistical inference does not require asymptotic justifications or assumptions about the stochastic processes generating outcomes.[15] This approach is not difficult to use. In appendix Appendix A we provide a few code snippets to illustrate the simple relationship between a formal model of effects and a statistical test of relevant hypotheses.

[15] While not required, such additional assumptions can profitably speed computation. For example, in our simulation, we employed a test statistic and Normal approximation suggested in (Hansen and Bowers 2009). When applied to data generated by the method of random assignment used in the simulation, the statistic rapidly converges to a Normal distribution. While the asymptotic justification would have been inappropriate for the newspaper example (with $n = 8$), we considered it compelling in our simulation ($n = 100$). More information on computation can be found in Appendix Appendix A. The materials to completely reproduce the results in this paper as well as supplementary analyses not presented here may be found in the reproduction compendium found at [insert web link].

Appendix A Code Examples

Here we provide snippets of R code meant to illustrate and make concrete the workflow involved in implementing our ideas.

To test a set of hypotheses about the parameters of a given model we require: (1) An $n \times 1$ vector denoting assignment for each unit to treatment and control ($\mathbf{Z}$); (2) An $n \times 1$ vector recording membership in any strata or blocks ($\mathbf{B}$); (3) An $n \times 1$ observed outcome vector ($\mathbf{Y}$); (4) The inverse model of effects (i.e. the function which recovers the uniformity trial by removing the hypothesized effects from the vector of observed outcomes); (5) One or more vectors of particular values for the parameters that are worth investigating (i.e. if the model involves τ then we need a vector of possible τ_0 that we desire to test); (6) A test statistic that is a function of outcomes, treatment, and blocking.

Here, for example, we assess the constant effects model from § 2.4 as follows:

Generating the Hypothetical Randomization Distributions from the Constant Effects Model

```
constant.additive.inverse.model ← function(ys, z, b, tau){ ys - (z*tau) }
tau0s ← list(tau=seq(0,10,1))
nsims ← 1000
tau0.dist ← pRD(data=Y, ##the outcome
                treatment=Z, ##the treatments
                test.stat=mean.difference, ##the test statistic
                moe=constant.additive.inverse.model, ## the inverse model
                parameters=tau0s, ## the hypotheses
                blocks=B, ## Block membership
                samples=nsims ) ## Number of samples from possible set
```

We include the `samples=nsims` argument here although the software actual returns the exact, enumerated, distributions since the total size of the possible assignments set, $\mathbf{\Omega}$, is less than 100 (16 in this case).

Having generated the randomization distributions that characterize the different hypotheses, we can calculate p-values and invert the tests to create intervals.

P-values and Intervals for Hypotheses from the Constant Effects Model

```
## A one-sided interval
one.sided.lower.ps ← na.omit(p.values(tau0.dist, lower.p.value))
high.ci.bound ←
   max(one.sided.lower.ps[one.sided.lower.ps[,"p"]>.125,"tau"])
## A two-sided interval
two.sided.ps ← na.omit(p.values(tau0.dist,general.two.sided.p.value))
two.sided.ci ← range(two.sided.ps[two.sided.ps[,"p"]>.125,"tau"])
```

To assess the hypotheses that were specific to named cities in § 2.5 we used the following model:

Testing Hypotheses about Named Interference

```
named.spillover ← function(r, z, b, tau, w) {
  names(z) ← names(r)
  r.adj ← r - z * tau
  r.adj["Richland"] ← r.adj["Richland"] - w * tau * z["Yakima"]
  r.adj["Midland"]  ← r.adj["Midland"]  - w * tau * z["Saginaw"]
  r.adj["Saginaw"]  ← r.adj["Saginaw"]  - w * tau * z["Midland"]
  return(r.adj)
}
theparams ← list(tau = seq(0, 20, 1), w = seq(0,1,.1))
```

```
news.spill.dist ← pRD(data=R,
                      treatment=Z,
                      test.stat=mean.difference,
                      moe=named.spillover,
                      parameters = theparams,
                      blocks = B,
                      samples = nsims )
```

The two parameter inverse model from § 3 relied on an $n \times n$ network adjacency matrix, $\mathbf{S}$, and was coded as follows:

Testing Hypotheses from the Two Parameter Network Effects Model

```
link1←function(x){ 1 - ( 1 / (1+x) ) }
link2←function(x){     ( 1 / (1+x) ) }
make.inv.moe.2parm ← function(S,linkT,linkC) {
  function(ys, z, blocks, tau1, tau2) {
   ys - (z *    ( tau1 +  ( tau1*tau2  * linkT( as.vector( z %*% S ) ) )
      )
      + (1-z) * ( tau1 * tau2  * linkC( as.vector( z %*% S ) ) ) )
  }
}
inv.moe.2parm←make.inv.moe.2parm(S,linkT=link2,linkC=link1)
theparams.2parm←list(tau1 = seq(20,80,1), tau2 = seq(0,1,.01))
net.spill.true ← pRD(Y.2parm, ## Y generated by the 2 parm model
                     Z, ## vector of treatment assignemnts
                     mean.diff.noblocks, ## mann.whitney.u,
                     inv.moe.2parm, ##
                     blocks = B, ## constant vector to communicate no
                        blocking.
                     parameters = theparams.2parm,
                     samples=thesims )
```

Appendix B A General Representation of Interference Effects

Here we propose a general representation of interference effects which enables us to reason about datasets and experiments of any design or size.

Appendix B.1 The complete interference case

We begin by developing a way to write down the observational identity (i.e. the equation relating observed outcomes to potential outcomes) without any restrictions on the potential outcomes. Later we will consider how to prune or constrain this equation to reflect both the facts of design, outside knowledge about outcomes, and hypotheses about effects and interference. In the same way that the notation for potential outcomes allowed us to formalize our reasoning about counterfactual causation, so too will a notation for sets of potential outcomes and interacting assignments help us reason about and specify questions we want to ask of a given design. We make use of the isomorphism between graphs, networks, and matrices to accomplish our task.

In the most general terms we can think of any set of units (an experimental pool for example), as a "complete graph":

Figure 12 shows such a graph. Here we have $n = 3$, and thus $2^3 = 8$ potential outcomes per unit. A complete graph has $n(n-1)/2$ edges (or $2n(n-1)/2 = n(n-1)$ possible unidirectional paths for interference). So, figure 12 has 6 paths of possible interference. Notice that each unit here depends on all the other units and influences all the other units in turn whether or not the unit is assigned treatment or control.

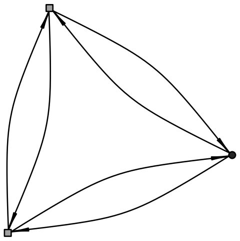


Figure 12: A Simulated Network and field experiment: treatment (circles) and control (squares). Without further assumptions, treatment or control assigned to any unit may influence any other unit. The edges have arrows to show that influence may be directional.

The vector of possible potential outcomes for unit 1, $\mathbf{y}_{1,.}$, given the graph in figure 12 and no further assumptions in is, lexicographic order:

$$\mathbf{y}_{1,.} = \{y_{1,\{111\}}, y_{1,\{110\}}, y_{1,\{101\}}, y_{1,\{100\}}, y_{1,\{011\}}, y_{1,\{010\}}, y_{1,\{001\}}, y_{1,\{000\}}\} \tag{24}$$

If an arrow does not connect a unit i to another unit j, this we can write $y_{j,Z_i,\mathbf{Z}_{(-i)}} = y_{j,Z'_i,\mathbf{Z}_{(-i)}}$ for any $Z_i \neq Z'_i$. Since we are only considering the case of binary treatment here, this general statement of equality can be simplified to say, $y_{j,Z_i=1,\mathbf{Z}_{(-i)}} = y_{j,Z_i=0,\mathbf{Z}_{(-i)}}$. That is, for a given vector of treatment assignments to j and every unit but i, unit j would show the same response whether unit i is treated or not. Equalities of this form are implied by such pruning of the complete graph. That is, we set potential outcomes equal to each other when we take away edges in the graph.

Before we begin to prune the complete graph, let us ask what the complete graph implies for the relationship between what we observe for unit 1, Y_1, and the potential outcomes shown in equation 24: what is the observed outcome identity equation implied here?

In scalar form we might write this identity as follows:

$$
\begin{aligned}
Y_1 = & Z_3\Big(Z_2\big(Z_1 y_{1,111} + (1 - Z_1) y_{1,011}\big) + (1 - Z_2)\big(Z_1 y_{1,101} + (1 - Z_1) y_{1,001}\big)\Big) + \\
& (1 - Z_3)\Big(Z_2\big(Z_1 y_{1,110} + (1 - Z_1) y_{1,010}\big) + (1 - Z_2)\big(Z_1 y_{1,100} + (1 - Z_1) y_{1,000}\big)\Big)
\end{aligned}
\tag{25}
$$

Notice that equation 25 specifies the circumstances under which what we observe for unit 1, Y_1, represents any of the potential outcomes possible from the complete graph and no further restrictions. For example, it says that we would observe $y_{1,\mathbf{Z}=\{1,1,1\}}$ when $Z_1 = Z_2 = Z_3 = 1$, or $\mathbf{Z} = \{1, 1, 1\}$. We can write this identity more cleanly using matrices. The matrix representation also allows us to write this equation for any sample size (whereas the scalar form would get incredibly messy very quickly). The matrix representation collects all of the potential outcomes into a $2 \times (2^n)/2 = 2^{n-1}$ matrix that we call $\boldsymbol{\rho}$. For $n = 3$, we might write $\boldsymbol{\rho}$ for a unit i as follows:

$$
\boldsymbol{\rho}_i = \begin{pmatrix} y_{i,111} & y_{i,110} & y_{i,101} & y_{i,100} \\ y_{i,011} & y_{i,010} & y_{i,001} & y_{i,000} \end{pmatrix}
\tag{26}
$$

Equation 25 multiplies each of the entries in $\boldsymbol{\rho}_i$ by the corresponding collections of treatment assigned to each unit. If we collect those $\boldsymbol{\zeta} = \{Z_i, (1 - Z_i)\}$ into a $2 \times 2^{n-1}$ matrix, $\mathcal{Z}$, we can write the observed outcome identity equation very succinctly for binary treatments as

$$
Y_i = \mathbf{1}_{(1\times 2)} \cdot (\mathcal{Z}_i \times \boldsymbol{\rho}_i) \cdot \mathbf{1}_{(2^{n-1}\times 1)},
\tag{27}
$$

where $\mathcal{Z}_i$, represents the Kronecker product, written $\otimes$, of all of the vectors representing the treatment possibilities for the units in the study, $\mathcal{Z}_i = \bigotimes_j^n \boldsymbol{\zeta}_j = \boldsymbol{\zeta}_1 \otimes \boldsymbol{\zeta}_3 \otimes \boldsymbol{\zeta}_2 = \{Z_1, (1 - Z_1)\} \otimes \{Z_2, (1 - Z_2)\} \otimes \{Z_3, (1 - Z_3)\}$. The terms $\mathbf{1}$ are merely vectors of 1s, which collapse the result of $(\mathcal{Z}_i \times \boldsymbol{\rho}_i)$ into a single equation.

Here we write out equation 27 showing the full matrices (but doubly transposed to fit on the page) for $n = 3$:

$$
Y_i = \begin{pmatrix} 1 & 1 \end{pmatrix} \cdot \left[\begin{pmatrix} Z_1 Z_2 Z_3 & (1 - Z_1) Z_2 Z_3 \\ Z_1 Z_2 (1 - Z_3) & (1 - Z_1) Z_2 (1 - Z_3) \\ Z_1 (1 - Z_2) Z_3 & (1 - Z_1)(1 - Z_2) Z_3 \\ Z_1 (1 - Z_2)(1 - Z_3) & (1 - Z_1)(1 - Z_2)(1 - Z_3) \end{pmatrix} \times \begin{pmatrix} y_{i,111} & y_{i,011} \\ y_{i,110} & y_{i,010} \\ y_{i,101} & y_{i,001} \\ y_{i,100} & y_{i,000} \end{pmatrix} \right]^T \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}
\tag{28}
$$

Since a priori all units in the study have the same relation between potential outcomes, treatment assignments, and observed outcomes, we can create the $n \times 1$ vector containing the equations for all of the units in the study, $\mathbf{Y}$, simply by multiplying Y_i by $\mathbf{1}_{(n\times 1)}$, such that $\mathbf{Y} = Y_i \times \mathbf{1}_{(n\times 1)}$.

Appendix B.1.1 Summary

We have shown that with only knowledge about (1) the size of the experimental pool and (2) the number of unique possible treatments (here set to 2), we can have a compact notation for the possible potential outcomes, treatment assignments, and the identity linking potential outcomes and treatment assignments to observed outcomes. When n is large, these matrices become too large to generate in software (let alone to write down their entries with a pencil), yet having this framework now allows us to represent restrictions on this case for more realistic experimental designs and empirical structures; which in turn will allow us to specify and test hypotheses about treatment effects and interference.

Appendix B.2 The Pruned Graph

No real study entertains hypotheses about 2^n potential outcomes in any detailed manner. Even with $n = 40$ we would have $1.1 \cdot 10^{12}$ possible potential outcomes! Even if we want to hypothesize directly about interference, we do not want to specify patterns of hypotheses for so many possibilities. In a series of steps here we show how one may (and must) reduce the set of potential outcomes considered. First, one may use information from the design of the study itself. Second, one may have a good idea about subsets of units which ought to be seen as not interferring with units in other subsets: For example, Sioux City, Lowell, and Oxford in the newspapers example were so geographically distant from the other cities that we felt comfortable claiming no interference for these cities. Third, the particular hypotheses that one desires to consider may involve further simplifications: For example, in the social network example, we collapsed set of potential outcomes even further (in fact, we could collapse them to only two potential outcomes and scalar functions of network characteristics since the particular patterns did not matter). There is no requirement to collapse the potential outcomes down to only two pieces, but fewer makes our exposition here more clear.

Appendix B.2.1 Pruning by Design

Most of the potential outcomes listed in lists such as equation 24 will never occur in any real design.[16] For example consider again the $n = 40$ case, such a design would involve assigning exactly 20 to treatment. Thus, rather than 2^n outcomes we have $\binom{40}{20} = 1.378 \cdot 10^{11}$ which has 0.13 as many entries as the original set. Of course, in that case, we still have too many potential outcomes to consider based only on how treatment was assigned.[17]

What does this mean for the core of the equation relating potential outcomes to observed outcomes $((\mathcal{Z}_i \times \boldsymbol{\rho}_i)$)? It means that the matrices of assignments, $\mathcal{Z}_i$ and potential outcomes $\boldsymbol{\rho}_i$ are smaller — reflecting now the actually possible assignments rather than all possible n-tuples.

Appendix B.2.2 Pruning by Knowledge of Structure

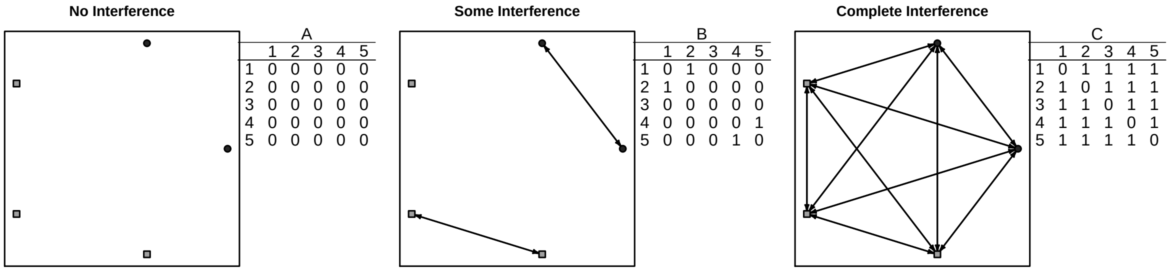
We say “knowledge” here to distinguish it from “hypotheses about structure” although, of course, we could include such structural statements as hypotheses. However, in many applications there are subsets and groupings or even types of interference which are just not credible or would never be interesting. Representing such incredible (i.e. not even worth hypothesizing about) relations prunes the complete graph even more.

Figure 13 shows three plots representing certain structural presumptions about interference and the related adjacency matrices for the case of $n = 5$.

Usually we have some idea about the groups of units within which interference is apt to occur,

[16] That vector can be thought of as all of the possible size 3 subsets of the 2-tuple $\{0, 1\}$.

[17] When an experiment uses blocking or pairing the set of possibilities may reduce even more dramatically. For example, if we had organized 40 units into 20 pairs, then the set of possible treatment assignments in which exactly one unit in each pair is treated would have about 1000000 elements. In the Newspapers study the total possible treatment assignments are 16 compared to 70 for the unpaired case.



A

	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0

B

	1	2	3	4	5
1	0	1	0	0	0
2	1	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	1
5	0	0	0	1	0

C

	1	2	3	4	5
1	0	1	1	1	1
2	1	0	1	1	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	0

Figure 13: Graphs and corresponding adjacency matrices representing different interference/connectedness structures.

or are willing to make some other decision which simplifies the "Everything is related to everything" statement represented by the complete interference graph.

Notice, in fact, that the adjacency matrices (or graphs) tell us specific things about the relations among potential outcomes. In particular, the 0s on the off-diagonal elements of those graphs tell us that certain sets of potential outcomes can be made equal. To make this more clear, let us think about what kinds of restrictions on the complete graph are implied by the graph in the central panel. We have reproduced the adjacency matrix here with one change — we have made the diagonal contain 1s. We'll explain why soon.

$$\mathbf{B} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \tag{29}$$

The restrictions on the potential outcomes for unit 1 are those listed in the first column of $\mathbf{B}$. In that column we have 3 zeros in positions $\{(3,1),(4,1),(5,1)\}$. These zeros imply the following equality: $y_{1,\mathbf{Z}_{\{3,4,5\}},\mathbf{Z}_{-(\{3,4,5\})}} = y_{1,\mathbf{Z}'_{\{3,4,5\}},\mathbf{Z}_{-(\{3,4,5\})}}$ for all $\mathbf{Z}_{\{3,4,5\}} \neq \mathbf{Z}'_{\{3,4,5\}}$. That is, any set of potential outcomes for the unit which are the same in all entries *except for those reflecting assignment to any combination of units 3,4, and 5* can be considered the same.

The complete graph for binary treatment with $n = 5$ with no further information would imply $2^5 = 32$ potential outcomes for each unit. The design of the study would reduce this number to $\binom{5}{2} = 10$. And, now stating restrictions on the possibilities for interference (such as noticing that one of our units was just too isolated (perhaps by geography) to interfer or be interferred with, leaves us with the following sets of potential outcomes: for the isolated unit 3 we have only 2 potential outcomes $\{y_{3,\{.,.,0,.,.\}}, y_{3,\{.,.,1,.,.\}}\}$ and for the other units (which interact with only one other unit) we have 4 potential outcomes $\{y_{i,\{0,0,.,.,.\}}, y_{i,\{0,1,.,.,.\}}, y_{i,\{i,0,.,.,.\}}, y_{i,\{1,1,.,.,.\}}\}$ for $i \in \{1,2,4,5\}$.

Now, the matrix encoding possible interference, B, does tell us exactly how many potential outcomes are available for hypotheses, but we cannot use it simply via some matrix multiplication to simplify $(\mathcal{Z}_i \times \boldsymbol{\rho}_i)$. After all B is $n \times n$ and $(\mathcal{Z}_i \times \boldsymbol{\rho}_i)$ is $2 \times |\boldsymbol{\Omega}|$ where $|\boldsymbol{\Omega}|$ is the size of the $\boldsymbol{\Omega}$ matrix in terms of the numbers of $\mathbf{z}$ vectors it contains. In our simple $n = 5$ and $n_t = 2$ case, $|\boldsymbol{\Omega}| = \binom{5}{2} = 10$. One way to write down this operation uses the following algorithm:

Define a function $\text{Pos}(\mathbf{M}, s)$ which returns the positions of the scalar number s in the matrix $\mathbf{M}$. So,

$$
\begin{aligned}
\text{Pos}(\mathbf{B},0) =\{&\{1,3\},\{1,4\},\{1,5\},\\
&\{2,3\},\{2,4\},\{2,5\},\\
&\{3,1\},\{3,2\},\{3,4\},\{3,5\},\\
&\{4,1\},\{4,2\},\{4,3\},\\
&\{5,1\},\{5,2\},\{5,3\}\}
\end{aligned}
\tag{30}
$$

Now $\mathbf{B}$ is $n \times n$ and rows and columns hold the units in the same order (from $1 \dots n$).

Now, consider all pairs of vectors of treatment assignments, $\mathbf{Z}, \mathbf{Z}'$ written in partitioned form focusing on unit j as $\mathbf{Z} = \{Z_j, \mathbf{Z}_{(-j)}\}$ and $\mathbf{Z}' = \{Z'_j, \mathbf{Z}'_{(-j)}\}$. Algorithmn 1 shows how we would infer the relations between pruning the graph and the set of possible potential outcomes.

```
input  : An adjacency matrix, $\mathbf{B}$, with 1s on the diagonal indicating connections with 1 and
         lack of connection with 0. Two vectors of treatment assignments, $\mathbf{Z}$ and $\mathbf{Z}'$. In the
         simple case, these are of length $\binom{n}{k}$.
output : Two vectors of treatment assignments, $\mathbf{Z}$ and $\mathbf{Z}'$ either unchanged or set to be equal
         by replacing a numeric element with a symbol.
if $\mathbf{Z} \neq \mathbf{Z}'$ such that $Z_j \neq Z'_j$ and $\mathbf{Z}_{(-j)} = \mathbf{Z}'_{(-j)}$ and $\mathbf{B}_{j,i} = 0$ then
    Set $Z_j = .$ such that $\mathbf{Z} = \{Z_j = ., \mathbf{Z}_{(-j)}\}$ and $\mathbf{Z}' = \{Z'_j = ., \mathbf{Z}'_{(-j)}\}$ and thus $\mathbf{Z} = \mathbf{Z}'$
else
    do nothing
end
```

Algorithm 1: An algorithmic representation for how an adjacency matrix restricts potential outcomes for a unit i.

So, if $\mathbf{B}_{3,1} = 0$ then, for unit 1, we would set equal any potential outcomes which differ only in the third element (indicating a difference of treatment to unit 3). So, at this point we have 2 potential outcomes to consider for unit 3 and 4 for each of the other units. What hypotheses might we care to assess?

Appendix B.2.3 Specifying and testing hypotheses involving interference between units

Given restrictions of design and structure (often geography but it could represent other kinds of knowledge). We tend to have a small set of potential outcomes on which we can focus. How should we write down hypotheses that we desire to assess?

Often, we are only interested in hypotheses in which units do not interfere and we write: $y_{i,Z_i=1,\mathbf{Z}_{(-i)}} = y_{i,Z_i=1,\mathbf{Z}'_{(-i)}}$ and $y_{i,Z_i=0,\mathbf{Z}_{(-i)}} = y_{i,Z_i=0,\mathbf{Z}'_{(-i)}}$ for all $\mathbf{Z} \neq \mathbf{Z}'$. That is, the essence of entertaining ideas about "no interference" is to drastically prune the set of potential outcomes.

However, imagine we had some claims to assess involving consideration of interference — either because we want to assess hypotheses about treatment effects in the presence of interference or because we want to assess hypotheses about the interference process itself. In the $n = 5$ example above, we have the opportunity to make such hypotheses about units 1,2,4 and 5 (assuming that 3 is so isolated that hypotheses about interference with it would be uninteresting). Imagine, again for simplicity, the constant and additive treatment effect hypothesis generator for unit 3 such that

$y_{3,\{.,.,1,.,.\}} = y_{3,\{.,.,0,.,.\}} + \tau$ or $y_{3,Z_3=1,\mathbf{Z}_{(-3)}} = y_{3,Z_3=0,\mathbf{Z}_{(-3)}} + \tau$ for any $\mathbf{Z}_{(-3)}$. So, control response turns into treatment response by the addition of a constant for unit 3 (according to this theory that we desire to assess/this question we desire to ask).

Now, what do we mean by "control response" turning into "treatment response" for the other putatively interferring units? Recall that the potential outcomes for those units were of the form: $\{y_{i,\{0,0,.,.,.\}}, y_{i,\{0,1,.,.,.\}}, y_{i,\{1,0,.,.,.\}}, y_{i,\{1,1,.,.,.\}}\}$ for $i \in \{1, 2, 4, 5\}$. We see two ways for unit i to have a control response in those four potential outcomes. In one way, both interferring units have control $\{0, 0\}$ and in the other way, one unit has treatment and the other control, $\{0, 1\}$ and $\{1, 0\}$. When another potentially interferring unit receives treatment, then the focal unit, i, under control may receive some spillover (or at least we may be interested in this question). So now, we use the $\{0, 0\}$ outcome as the baseline against which we compare either the direct treatment or spillover (or amplification) effects.

At this point we could write each of the three potential outcomes $y_{i,\{0,1,.,.,.\}}, y_{i,\{1,0,.,.,.\}}, y_{i,\{1,1,.,.,.\}}$ as a function of $y_{i,\{0,0,.,.,.\}}$ and some parameters. In our examples, however, we further simplified the hypotheses by saying that we were only interested in hypotheses either about direct effects or spillover effects, not amplifying effects. This decision further simplified our set of hypotheses to only two equations: (1) one for the situation in which unit i received control and the potentially interfering unit j received treatment and (2) for the situation in which unit i is assigned the treatment condition (in which we claim that $y_{i,\{1,0,.,.,.\}} = y_{i,\{1,1,.,.,.\}}$).

For example we might imagine a spillover effect when unit i is in the control condition and the potentially interferring unit j is in the treatment condition: $y_{i,Z_i=0,Z_j=1,.} = y_{i,Z_i=0,Z_j=0,.} + w\tau$ where w tells us the amount of the treatment effect that spills over. And we might also imagine a direct constant effect when unit i is treated: $y_{i,Z_i=1,Z_j=0,.} = y_{i,Z_i=1,Z_j=1,.} = y_{i,Z_i=0,Z_j=0,.} + \tau$.

One could also imagine interesting hypotheses about all three potential outcomes: perhaps one might write both $y_{i,Z_i=1,Z_j=0,.} = y_{i,Z_i=0,Z_j=0,.} + \tau$ and $y_{i,Z_i=1,Z_j=1,.} = y_{i,Z_i=0,Z_j=0,.} + a\tau$ to allow for an amplification effect (i.e. the effect of treatment is made stronger when an interfering unit is also treated).

Another approach to winnow the set of potential outcomes is to restrict attention to scalar functions of them (Hong and Raudenbush 2006). So, for example in the section on social networks we asked the question about whether (and to what extent), treatment effects might depend on the number of treated connections. In essence this kind of hypothesis (and our current framework) involves both the decision about how the function of connections ought to influence the direct treatment effect, and also a decision that we do not want to entertain hypothesis about particular combinations of potential outcomes. So, we could, in essence, think about our potential outcomes as non-interferring except in the particular way that we desired to scrutinze. That is, we could write $y_{i,Z_i=1,\mathbf{Z}_{(-i)}} = y_{i,Z_i=0,\mathbf{Z}_{(-i)}=0} + \tau + \tau w\mathbf{Z}^t\mathbf{S}$ and $y_{i,Z_i=0,\mathbf{Z}_{(-i)}} = y_{i,Z_i=0,\mathbf{Z}_{(-i)}=0} + \tau w\mathbf{Z}^t\mathbf{S}$.

Appendix B.2.4 Summary

This part of the paper has shown that (1) one may represent the complete set of potentially interferring potential outcomes in a compact form and that (2) one may begin to restrict attention to managable subsets of those outcomes using knowledge of design, information about structure, and hypotheses about effects. In general, one may use the construct of a graph or network to represent any form of interference and to allow formalization of hypotheses about treatment effects and interference. Even though the set of potential outcomes can become immense very quickly (tending to follow the law $2^{\text{number of edges}}$ — actually this much more like a logistic function that asymptotes at $|\Omega|$), we need not make untestable no-interference assumptions merely because we are overwhelmed with the size of the possibilities. Rather, we can use what we know and what we care about (from past theory and literature) to engage with manageable numbers of counter factuals in direct and substantively meaningful manners.

Appendix B.3 Applying the General Representation to the Newspapers Study

We began this paper by talking informally about the placement of cities on a map and the types of interference that the geography might imply. Such ideas led us to write a set of hypotheses:

$$h(y_{i,00}) = \begin{cases} Z_i(y_{i,00} + \tau) + (1 - Z_i)(y_{i,00}) & \text{for } i \in \{\text{ Yakima, Oxford, Lowell, Battle Creek, Sioux City }\} \\ Z_i(y_{i,00} + \tau) + (1 - Z_i)(y_{i,00} + w\tau) & \text{for } i \in \{\text{ Richland, Midland, Saginaw }\} \end{cases} . \quad (31)$$

Now we have a more general way to formalize the process of hypothesizing about interference. Let us apply it to the newspaper advertisements study.

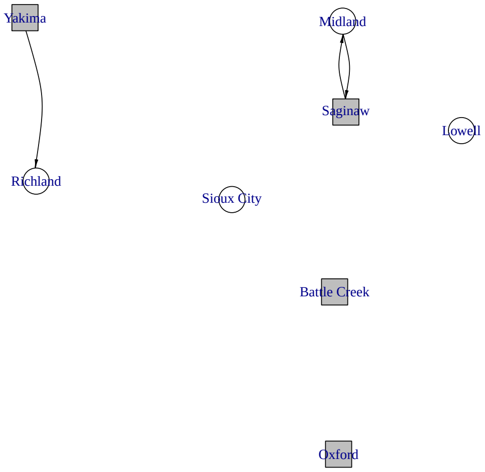


Figure 14: A directed network (or graph) representation of an interference hypothesis for the Panagopolous Newspaper study. Squares represent cities assigned to treatment. Circles are cities assigned to control. Arrows show direction of spillover: from the larger city of Yakima to the smaller city of Richland, and two way interference between Midland and Saginaw.

Figure 14 shows the cities as nodes on a graph. We know that there are $K = 16$ possible ways to assign treatment to the pairs of cities in this study, so, the complete graph would imply 16 potential outcomes for each city. A graph without any connections (encoding the idea of no interference) would imply 2 potential outcomes for each city.

We presumed, on the basis of knowledge about how local advertisements in newspapers relates to the geography of the United States that the only possible connections would be between Yakima and Richland and between Midland and Saginaw. And later we hypothesized that the interference would be one-way from Yakima to Richland, but symmetric between Midland and Saginaw. This graph encodes these statements about connections.

What potential outcomes are available for us to consider after drawing this graph? The adjacency matrix of the graph tell us that we have two potential outcomes for each of the isolated cities (or cities not plausibly interfering or interferred with). We also have two potential outcomes for Richland (but both depend on Yakima): $y_{i,Z_i=0,Z_j=1}$ and $y_{i,Z_i=1,Z_j=0}$ for i =Richland and j =Yakima. While Richland and Yakima are in the same pair, and thus only one of them may be treated at a time, Midland and Saginaw are in different pairs. So, Midland and Saginaw each have four potential outcomes to consider: $y_{i,\{11\}}$,$y_{i,\{10\}}$,$y_{i,\{01\}}$,$y_{i,\{00\}}$, where we write $\{11\}$ as shorthand for $\{Z_i = 1, Z_j = 1\}$.

For the isolated cities, we claimed (for simplicity) that we were interested in whether the hypothesis that $h(y_{i,Z_i=0,.}) = y_{i,Z_i=0,.} + \tau = y_{i,Z_i=1,.}$ could be rejected by our data, where we write $y_{i,Z_i=0,.}$ to indicate that we ignore the other potential outcomes in the network for these isolates.

Since Yakima is only a source not a destination of interference, its hypothesis is likewise $h(y_{i,Z_i=0,.}) = y_{i,Z_i=0,.} + \tau$. In this scenario, producing interference is the same as experiencing no interference under the assumption that the people of Richland do not steal the newspapers from Yakima and thereby diminish the treatment effect in Yakima [i.e. when spillover occurs with an intervention that is not renewable or is excludable, then perhaps this idea that being the source of spillover is the same as not experiencing interference is not a good one.]

Richland has two potential outcomes to consider but they both may involve interference: $y_{i,10}$,$y_{i,01}$. We wondered whether the data would exclude the idea that some treatment spilled over from Yakima to Richland, and between Midland and Saginaw, when the recipient of such spillover was in the control condition such that: $h(y_{i,Z_i=0,Z_j=1}) = y_{i,Z_i=0,Z_j=0} + w\tau$ where w is the proportion of the overall treatment effect, τ, that spills over. We also decided to assess this hypothesis about spillover in the situation in which there is no interference in the treatment condition — the idea being that direct experience of treatment drowns out any treatment leaking over from another city and also that there is no amplification of treatment.

These considerations meant that we did not need to specify hypotheses about all four potential outcomes available for Midland and Saginaw. Rather, by hypothesis, we wrote $y_{i,11} = y_{i,10} = y_{i,1.}$ and $h(y_{i,00}) = y_{i,00} + \tau = y_{i,1.}$.

We listed those hypotheses in a condensed form in equation 31. And we can now see that the equations here:

$$y_{i,00} = \begin{cases} Y_i - \tau Z_i & \text{when } i \in \{\text{ Yakima, Oxford, Lowell, Battle Creek, Sioux City }\} \\ Y_i - \tau Z_i - w\tau Z_j & \text{when } i\text{=Richland, } j\text{=Yakima} \\ Y_i - \tau Z_i - w\tau Z_j & \text{when } i\text{=Midland, } j\text{=Saginaw} \\ Y_i - \tau Z_i - w\tau Z_j & \text{when } i\text{=Saginaw, } j\text{=Midland} \end{cases} \tag{32}$$

arise from solving each observed outcome identity equation 27 (one for each type of network effects) for the potential response to the uniformity trial. And the randomization distribution against which we compare functions of observed data arises from the design of the experiment itself.

Appendix B.3.1 Workflow and Summary

In this section, we have provided a formal framework to support reasoning about treatment effects and interference effects in comparative studies of arbitrary design and size. If one can draw a graph or a network diagram (or specify an adjacency matrix) then one can know which list of potential outcomes are available for use in assessing substantively motivated hypotheses.

References

Achen, Christopher H. 1982. *Interpreting and Using Regression*. Newbury Park, CA: Sage.

Aronow, Peter M. 2010. “A General Method for Detecting Interference Between Units in Randomized Experiments.” Unpublished manuscript.

Barnard, GA. 1947. “Significance tests for 2× 2 tables.” *Biometrika* 34(1/2):123–138.

Berger, R.L. and D.D. Boos. 1994. “P Values Maximized over a Confidence Set for the Nuisance Parameter.” *Journal of the American Statistical Association* 89(427).

Bowers, Jake and Costas Panagopoulos. 2011. “Fisher’s randomization mode of statistical inference, then and now.” Unpublished manuscript.

Bowers, Jake, Mark Fredrickson and Ben Hansen. 2010. *RItools: Randomization Inference Tools*. R package version 0.1-11.
URL: *http://www.jakebowers.org/RItools.html*

Brady, Henry E. 2008. “Causation and explanation in social science.” *Oxford handbook of political methodology* pp. 217–270.

Chen, J., M. Humphreys and V. Modi. 2010. “Technology Diffusion and Social Networks: Evidence from a Field Experiment in Uganda.”.

Clarke, Kevin and David Primo. 2012. *A Model Discipline*. Oxford University Press.

Cox, D. R. 2006. *Principles of Statistical Inference*. First ed. Cambridge University Press.

Cox, David R. 1958. *The Planning of Experiments*. John Wiley.

Fisher, R.A. 1935. *The design of experiments. 1935*. Edinburgh: Oliver and Boyd.

Hansen, Ben B. and Jake Bowers. 2009. “Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign.” *Journal of the American Statistical Association* 104(487):873—885.

Hodges, J.L. and E.L. Lehmann. 1963. “Estimates of location based on rank tests.” *Ann. Math. Statist* 34:598–611.

Hodges, J.L., Jr. and E.L. Lehmann. 1964. *Basic Concepts of Probability and Statistics.* Holden-Day.

Hong, G. and S.W. Raudenbush. 2006. “Evaluating Kindergarten Retention Policy.” *Journal of the American Statistical Association* 101(475):901–910.

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

Ichino, N. and M. Schündeln. 2011. Deterring or Displacing Electoral Irregularities? Spillover Effects of Observers in a Randomized Field Experiment in Ghana. Technical report Working paper.

Keele, L., C. McConnaughy and I. White. 2012. "Strengthening the Experimenters Toolbox: Statistical Estimation of Internal Validity." *American Journal of Political Science* .

Keele, Luke, Corrine McConnaughy and Ismail White. 2008. "Statistical Inference For Experiments." Unpublished manuscript.

Little, R.J. and D.B. Rubin. 2000. "Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches." *Annual Review of Public Health* 21:121–145.

McConnell, M., B. Sinclair and D.P. Green. 2010. Detecting social networks: design and analysis of multilevel experiments. In *third annual center for experimental social science and New York University experimental political science conference*.

Miguel, E. and M. Kremer. 2004. "Worms: identifying impacts on education and health in the presence of treatment externalities." *Econometrica* 72(1):159–217.

Neyman, J. 1923 [1990]. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (1923)." *Statistical Science* 5:463–480. reprint. Transl. by Dabrowska and Speed.

Nickerson, D.W. 2008. "Is voting contagious? Evidence from two field experiments." *American Political Science Review* 102(01):49–57.

Nickerson, D.W. 2011. "Social Networks and Political Context." *Cambridge Handbook of Experimental Political Science* p. 273.

Nolen, T.L. and M. Hudgens. 2010. "Randomization-Based Inference within Principal Strata." *The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series* p. 17.

Panagopoulos, Costas. 2006. "The Impact of Newspaper Advertising on Voter Turnout: Evidence from a Field Experiment." Paper presented at the MPSA 2006.

Rosenbaum, Paul. 2009. "Design of Observational Studies." Unpublished book manuscript.

Rosenbaum, Paul R. 1993. "Hodges-Lehmann Point Estimates of Treatment Effect in Observational Studies." *Journal of the American Statistical Association* 88(424):1250–1253.
URL: *http://www.jstor.org/stable/2291264*

Rosenbaum, Paul R. 2002. *Observational Studies*. Second ed. Springer-Verlag.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. Springer.
URL: *http://www.springer.com/statistics/statistical+theory+and+methods/book/978-1-4419-1212-1*

Rosenbaum, P.R. 2007. "Interference Between Units in Randomized Experiments." *Journal of the American Statistical Association* 102(477):191–200.

Rubin, D. B. 1986. “Which ifs have causal answers? comments on “Statistics and Causal Inference”.” *Journal of the American Statistical Association* 81:961–962.

Rubin, Donald B. 1980. “Comment on “Randomization Analysis of Experimental Data: The Fisher Randomization Test”.” *Journal of the American Statistical Association* 75(371):591–593.

Rubin, Donald B. 2005. “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions.” *Journal of the American Statistical Association* 100:322–331.

Silvapulle, M.J. 1996. “A Test in the Presence of Nuisance Parameters.” *Journal of the American Statistical Association* 91(436).

Sinclair, B. 2011. Design and Analysis of Experiments in Multilevel Populations. In *Cambridge Handbook of Experimental Political Science*. Cambridge University Press p. 906.

Sobel, M.E. 2006. “What Do Randomized Studies of Housing Mobility Demonstrate?” *Journal of the American Statistical Association* 101(476):1398–1407.

Tchetgen, E.J.T. and T.J. VanderWeele. 2010. “On causal inference in the presence of interference.” *Statistical Methods in Medical Research* .

VanderWeele, T.J. 2008*a*. “Ignorability and stability assumptions in neighborhood effects research.” *Statistics in medicine* 27(11):1934–1943.

VanderWeele, T.J. 2008*b*. “Simple relations between principal stratification and direct and indirect effects.” *Statistics & Probability Letters* 78(17):2957–2962.

VanderWeele, T.J. 2009. “Marginal structural models for the estimation of direct and indirect effects.” *Epidemiology* 20(1):18.

VanderWeele, T.J. 2010. “Bias formulas for sensitivity analysis for direct and indirect effects.” *Epidemiology* 21(4):540.

VanderWeele, T.J. and M.A. Hernan. 2011. “Causal inference under multiple versions of treatment.” *COBRA Preprint Series* p. 77.