Reasoning about Interference Between Units

Jake Bowers [*] Mark Fredrickson [†]

October 28, 2011

Abstract

When an intervention has been randomly assigned, can we entertain specific hypotheses about situations in which treatment assigned to one unit may influence the outcomes of another unit? Most work in statistics would answer this question with a loud, "No." A statement about "no interference between units" has long been seen as a fundamental and untestable assumption that is a precondition for meaningful causal inference ((Cox 1958, p. 19), Rubin (1980, 1986); Brady (2008)).

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. 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 questions about how treatments 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; 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.

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

1 Introduction

Imagine a social scientist fields an experiment in which some people are given information that theory suggests ought to change their behavior: perhaps by telling them about the voting behavior of their neighbors the social costs of not voting rise (Gerber, Green and Larimer 2008). If treatment compels behavioral change on the outcome of interest, then it would also be natural to anticipate that people would talk with one another about the treatment. The control group and the treatment group will both have experienced the treatment, and unless experiencing the treatment via email forwarded from a friend is very different from experiencing the treatment via email sent from a researcher, we would be unable to detect any differences in average outcome in a cross-sectional comparison of two groups even if every person in both groups changed their behavior.

Since Neyman (1923 [1990]) we have known that average treatment effects can reflect an unobservable comparison of potential outcomes when treatments have been randomized. Yet, although randomization guarantees unconfounded comparisons and valid statistical inference about causal effects, when treatment spills over to the control group, simple average treatment effects become difficult to conceptualize and uniquely identify. Because aggregating over unobserved and observed potential outcomes allows the use of averages for statistical inference about causally defined quantities, "no interference between units" is thus often described as a foundational precondition for credible causal inference. Yet, aggregating over potential outcomes is only one way to overcome what Holland (1986) called "the fundamental problem of causal inference." Fisher (1935) proposed the use of a sharp null hypothesis test of no effects for this purpose. This hypothesis test made no assumptions about inference about units as presented by Fisher and, in fact, under certain conditions can be shown to either enable detection of interference (Aronow 2010) or directly enable testing the null of a particular kind of no effects under unspecified interference (Rosenbaum 2007).[1] Briefly, whereas Neyman confronted the problem of causal inference by changing the focus from units to aggregates (and specifically averages and sums), Fisher approached the problem by specifying hypotheses about each unit individually. In this paper we show that Fisher's approach allows us to ask questions about interference between units and produce statistical inferences about substantively meaningful quantities parameterizing those questions.

In fact, in this paper we show that "no interference" is not an *assumption* in Fisher's framework, but rather than *implication* that need not always hold.[2] An *assumption* is a statement about a truth without proof. In statistics, we can often bring evidence to bear regarding statements of assumptions, but, like hypotheses, assumptions are never proved but rather supported or, not-disproved. An

[1] There is another approach involving direct imputation of the potential outcomes as realizations of a random process using, usually Bayesian, models that one may find in most of the work on principal stratification .

[2] We thank Ben Hansen for this terminology.

implication is a condition following from other decisions including assumptions. If one set of decisions about how to assess hypotheses about causal effects involves units for which treatment does not spillover, then we say that these hypotheses imply that units do not interfere. However, the fact that one set of hypotheses implies no interference does not necessarily lead to the fact that another set of hypotheses must imply no interference. In fact, we will show that Fisher's framework allows one to directly assess hypotheses which imply interference between units just as one may produce confidence intervals implying no interference. We elaborate on these points below. This is our contribution: we show that it is possible to directly hypothesize about interference between units and to test such hypotheses.[3] Our contribution is possible in part because of advances in computing, and also in part because of advances in methodology extending both Neyman's and Fisher's frameworks to apply to our current sets of experimental designs.

By posing hypotheses about interference and assessing them in Fisher's framework for statistical inference (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 about decomposing average treatment effects into parts that are "indirect" (spilled over or otherwise operating via interference) or "direct" (i.e. not operating via interference) (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). 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 meaningful).

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) in which we want to enable statistical inference about particular hypotheses about interference and causal effects simultaneously. And Hong and Raudenbush (2006) also provide precedent for some of our work here by assessing treatment effects by collapsing aspects

[3] And our RItools software for R (Bowers, Fredrickson and Hansen 2010) will include all of these capabilities soon so that researchers can write down and assess their own ideas about interference on their own data and designs.

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, it makes life much easier (as we'll show later in the 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 differs from previous work in considering 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 (or, as we will see, via the specification of meaningful functions which generate hypotheses).

1.1 Roadmap

We organize this paper around an applied example and a simulation study. To fix ideas and develop notation we first analysis a field experiment in which four 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 simulated social network experiment which allows us to show that we can recover "true" treatment and spillover effects. We then pull back from the data to propose a more general conception of interference between units drawing on isomorphisms between graphs, networks, and matrices. That is, we show that one may ask new questions of datasets and designs, and we try to provide a mathematical language in which to formalize such questions (as well as software to interpret and assess the relationships between questions and data).

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. Table 1 shows the design of the study as well as baseline and post-treatment outcomes.

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

			Turnout	
City	Pair	Treatment	Baseline	Outcome
i	b	$\mathbf{Z}$	$\mathbf{x}$	$\mathbf{R}$
Saginaw	1	0	17	16
Sioux City	1	1	21	22
Battle Creek	2	0	13	14
Midland	2	1	12	7
Oxford	3	0	26	23
Lowell	3	1	25	27
Yakima	4	0	48	58
Richland	4	1	41	61

Table 1: Design and outcomes in the Newspapers Experiment. The Treatment column shows treatment with the newspaper ads as 1 and lack of treatment as 0. Panagopoulos (2006) provides more detail on the design of the experiment. For example, 21% of registered voters voted in the baseline election in Sioux City and 22% voted in the post-treatment election. Bold letters are vectors. Uppercase letters are random quantities and lowercase letters are fixed.

expect to see for city i if that city were treated, $Z = 1$, often written $r_{Z=1,i} \equiv r_{1i}$ and the turnout we would expect if treatment were withheld, $r_{Z=0,i} \equiv r_{0i}$. If treatment had a causal effect for city i in this experiment then turnout after advertisements would be higher than turnout without advertisements: $r_{1i} > r_{0i}$. If treatment had no effect then city i would display the same turnout regardless of treatment condition $r_{1i} = r_{0i}$.

Now, we only observe one set of treatment assignments. Different cities could have been assigned treatment had the randomizer chosen differently. 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, $\mathbf{\Omega}$. An assignment vector drawn from this set would be written $\mathbf{z}$ to distinguish it from our observed random draw $\mathbf{Z} = \{0, 1, 0, 1, 0, 1, 0, 1\}$. So, our shorthand of writing $r_{1,\text{Sioux City}}$ is really saying that $r_{\mathbf{z}=\{0,1,0,1,0,1,0,1\},\text{Sioux City}} = r_{\mathbf{z}=\{0,1,1,0,1,0,1,0\},\text{Sioux City}} = \ldots$. That is, we are saying that the counterfactual outcome under treatment for Sioux City would be the same regardless of configuration of assignment to any other city. So, notice that the notation that we use to express concepts about counterfactual causal inference implies something about no interference: $r_{z_{\text{Sioux City}}=1,\mathbf{z}_{-\text{Sioux City}},\text{Sioux City}} = r_{z_{\text{Sioux City}}=1,\mathbf{z}'_{-\text{Sioux City}},\text{Sioux City}}$ for all $\mathbf{z}, \mathbf{z}' \in \mathbf{\Omega}$.

This manner of conceptualizing causal effects implies that we can ignore the treatment assignment status of other cities. This idea — that treatment assignments to other cities did not interfere with potential outcomes in Sioux City — would be sensible if advertisements in newspapers in other cities were not viewed in Sioux City or if the act of withholding advertisements in other cities did not matter for turnout in Sioux City. How sensible is this idea in this study?

Figure 1 shows the locations of the cities chosen for the study. Sioux City is the gray dot

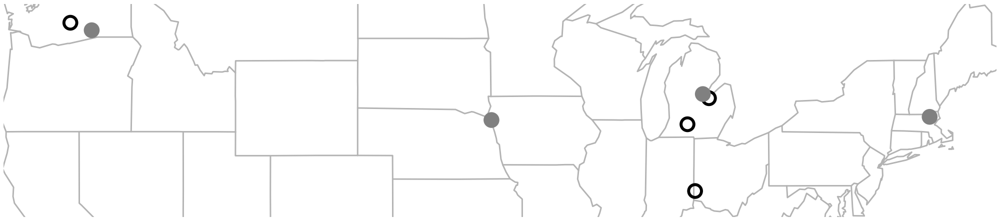


Figure 1: Locations of cities in the US newspapers field experiment. Control cities plotted as open black circles. Treated cities plotted as filled dark gray circles.

alone in the center of the map. Unless there are many people traveling between Sioux City and other cities (or emailing the advertisements to people in the other 7 cities), it seems reasonable to suppose that treatment assigned to other cities did not interfere with potential outcomes in Sioux City. However, we can see two places on the map where one might imagine newspaper advertisements could spillover between cities: specifically Yakima and Richland in Washington and Midland and Saginaw in Michigan. So, perhaps the idea of no interference is sensible for some cities in this study and not for others.

Before we conceptualize a restricted mixture of interfering and non-interfering causal effects for the cities in this particular study, however, let us first be clear about how one might assess hypotheses about no effects in the presence of unrestricted interference.

3 Some formalities and definitions about causal inference in the presence of interference

The foundation of Fisher's statistical inference for causal effects is the null hypothesis of no effects. Although we will build on this foundation to produce confidence intervals and confidence sets summarizing hypotheses about effects, we begin with the hypothesis of no effects.

3.1 What does "no effects" mean

3.1.1 "No effects" without interference

The map above suggested that treatment assigned to other cities probably did not interfere with potential outcomes in Sioux City. We might then wonder whether treatment had no effects, and formalize this hunch by saying that $H_0 : r_{1,\text{Sioux City}} = r_{0,\text{Sioux City}}$ — responses of Sioux City if treated would be the same as responses if not treated. If we wanted to ask this question of all cities we would formalize this by $H_0 : r_{1,i} = r_{0,i}$ for all i — all cities would act the same under treatment and control — the treatment has no effect. If we imagined no interference, then we would write: $R_i = Z_i r_{1i} + (1 - Z_i) r_{0i}$ as the identity linking potential outcomes to observed outcomes. And our hypothesis would imply that $R_i = r_{0i}$ (by substituting r_{0i} for r_{1i} and simplifying). So, we could test this hypothesis using our observed outcomes.[4]

3.1.2 "No effects" with interference

Now, if we want to consider interference, the situation becomes more complex. 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: $\{r_{1010}, r_{1001}, r_{0110}, r_{0101}\}$. 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: $H_0 : r_{1010} = r_{1001} = r_{0110} = r_{0101}$ for all cities. Another hypothesis of no effects states $H_0 : r_{1010} = r_{1001} = r_{0110} = r_{0101} = r_{0000}$ where r_{0000} is the response when the intervention was not applied to any units. We can write these hypotheses more generally for any vector of treatment assignments, $\mathbf{z}$. Following Rosenbaum (2007) we write $r_{i,b,\mathbf{z}}$ for city i in pair b to represent the potential turnout for one possible assignment configuration across all of the cities. To hypothesize that treatment had no effect might lead us to write $H_0 : r_{i,b,\mathbf{z}} = r_{i,b,\mathbf{z}'}$ for all $\mathbf{z}, \mathbf{z}' \in \mathbf{\Omega}$, where $b = 1, \ldots, B$, $i = 1, \ldots, I$. Rosenbaum (2007) calls this kind of hypothesis of "no effects" a hypothesis of "no primary effects." Rosenbaum (2007) contrasts "no primary effects" to the situation in which, although a configuration of treatment is chosen, no treatment is assigned anywhere in the set of cities, and thus cities display a potential outcome to this hypothetical "uniformity trial", $\tilde{r}_{i,b}$ and he writes $\tilde{H}_0 : r_{i,b,\mathbf{z}} = \tilde{r}_{i,b}$ as the hypothesis of what he calls "no effects" — i.e. that the potential outcomes in response to a given treatment assignment vector are those that would have been observed if treatment had been entirely with-held from all cities. We might write the potential response to the uniformity trial in our set of four cities as $r_{i,b,\mathbf{z}=\mathbf{0}}$, $r_{i,b,0000}$ (we use this notation rather than $\tilde{r}$ for the rest of this paper).

3.2 Testing the hypothesis of no effects under interference

We can easily test either hypothesis of no effects. Now, let us restrict attention, just for a moment, to only two cities in one pair (just to simplify notation). For city 1 our observed outcomes relate to possible potential outcomes in principle via the following equation:

$$R_1 = Z_1(Z_2 r_{11} + (1 - Z_2) r_{10}) + (1 - Z_1)(Z_2 r_{01} + (1 - Z_2) r_{00}) \tag{1}$$

Now, under the hypothesis of no effects, $H_0 : r_{11} = r_{10} = r_{01} = r_{00}$ where r_{00} is the response to

[4] This is a very abbreviated introduction to Fisher's sharp null hypothesis. See (Rosenbaum 2010, Chap2) for a textbook discussion. Bowers and Panagopoulos (2011) introduce and explain these ideas in the context of this field experiment.

the uniformity trial. So,

$$R_1 = Z_1(Z_2 r_{00} + (1 - Z_2) r_{00}) + (1 - Z_1)(Z_2 r_{00} + (1 - Z_2) r_{00}) \tag{2}$$

$$= Z_1 r_{00} + (1 - Z_1) r_{00} \tag{3}$$

$$= r_{00} \tag{4}$$

And under the hypothesis of no primary effects, $H_0 : r_{10} = r_{01}$ (we exclude the "both treated" and "both control" possibilities) so we might just write $r_{10} = r_{01} = r_*$.

$$R_1 = Z_1(1 - Z_2) r_* + (1 - Z_1) Z_2 r_* \tag{5}$$

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

$$= Z_1 r_* + (1 - Z_1) r_* = r_* \tag{6}$$

Notice that both of these hypotheses imply that what we observe is what we would observe in the putative world of the hypothesis. And this formulation carries over for all i. We develop a matrix formulation to engage with the question of deriving the relationship between hypotheses and observed outcomes for arbitrarily large sample sizes later.

Imagine that we summarized the results of our experiment with a difference in mean turnout, say $t(Z_i, R_i) = \sum_{i=1}^{n} Z_i \frac{R_i}{m} - \sum_{i=1}^{n} (1 - Z_i) \frac{R_i}{n-m}$, where m is the number of treated units, n is the total sample size. We use $t()$ here to emphasize that it is not the mean difference which matters here, but rather that we have an effect increasing function of treatment assignments and observed outcomes — a test statistic. If we calculate $t(\mathbf{z}, \mathbf{r}_{i,b,\mathbf{z}} = \mathbf{R} | \mathbf{z} \in \mathbf{\Omega})$ we can trace out the distribution of $t()$ under the null of no effects and the null of no primary effects. If many of the possible mean differences are close to (or less than) our observed mean difference, then we might say that our result is not surprising from the perspective of these hypotheses. If our result is extreme compared to the others, then we say that it would be surprising to observe our result under the hypothesis.

For example, using paired mean differences as $t()$, we find a one-sided p-value of $p = 0.38$ when we consider the null hypothesis of no effects (no primary effects and no primary effects). If our hypotheses were true, we would not be that surprised to see a mean difference as large as or larger than our observed value.

So, what did we do? We stated a hypothesis in which all of the potential outcomes implied by unrestricted interference are equal (to each other, and/or to a "uniformity trial"). We then asked what this hypothesis implies for what we observe. A test of this hypothesis suggested that our data are not surprising from the perspective of the null — we cannot reject the null at any conventional level of significance. Rejecting this hypothesis could suggest that some other pattern of responses to treatment are detectable in this dataset. Not rejecting this hypothesis is not confirming: an observed value which may be surprising with $n = 100$ may not surprise when $n = 2$.

However, notice that we have just executed what is often the most important and fundamental step in experimental data analysis without making a particular assumption about interference: the comparison of what we observe with the hypothesis that what we observe is merely due to chance and that the treatment had no effect at all. And we did not rely on any assumption of no interference.[5]

3.3 From "no effects" to "effects"

Scholarly imaginations tend not to stop at hypotheses of no effects. After all, the motivation for a study tends to involve some expectations about sign and magnitude if not rough value range for a treatment effect: In a study of turnout in a low salience election as influenced by newspaper ads, past experience, literature, and theory of two party elections, all might combine to lead us to expect values of effects for cities that are definitely less than 10 percentage points of turnout, but also probably more than 0 percentage points of turnout. Here we formalize thinking about effects in the context of interference between units. Considering effects in the presence of interference is more complicated than considering no effects.

Effects with no interference First, consider the simple situation with no interference. If we could hypothesize about no effects by writing $H_0 : r_{i,b,Z_i=1} = r_{i,b,Z_i=0}$ then we could express an expectation about a 7 percentage point difference in turnout for city i with $H_0 : r_{i,b,Z_i=1} = r_{i,b,Z_i=0} + 7$. Notice, hypothesis about effects tells us how potential responses in the absence of treatment would turn into potential responses to control. In this case, 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 following a particular model. For example, if past literature, experience, and theory suggest that we ought to investigate the idea that treatment with advertisements provides the same turnout boost in all cities in our pool, we might write $h(r_{i,b,Z_i=0}) = r_{i,b,Z_i=0} + \tau = r_{i,b,Z_i=1}$. Given our identity linking observed turnout to potential turnout under no interference, $R_{i,b} = Z_{i,b}r_{i,b1} + (1 - Z_{i,b})r_{i,b0}$ and a hypothesis about a specific value for τ, $H_0 : \tau = \tau_0$, one can make a test for this hypothesis under this model using the same process as we did for the null of no effects.

The only difference here is that, now $R_{i,b} \neq r_{0,i,b}$ but rather, when we substitute for $r_{1,i,b}$ in the identity and solve for $r_{0,i,b}$ we get $r_{0,i,b} = R_{i,b} - Z_{i,b}\tau_0$.[6] Thus, we can now compare $t(\mathbf{Z}, \mathbf{R} - \mathbf{Z}\tau_0)$ to it's distribution across the many ways that treatment could have been assigned as generated by $t(\mathbf{z}, \mathbf{R} - \mathbf{z}\tau_0 | \mathbf{z} \in \mathbf{\Omega})$.

In this simplest case, we consider $0 \leq \tau_0 \leq 10$. If we define "surprising" by $\alpha = .125$ (i.e. we

[5] We also did not rely on assumptions about large-samples, linearity, heteroskedasticy, Normality. All of the code to replicate every analysis in this paper will be available at http:/xxx.

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

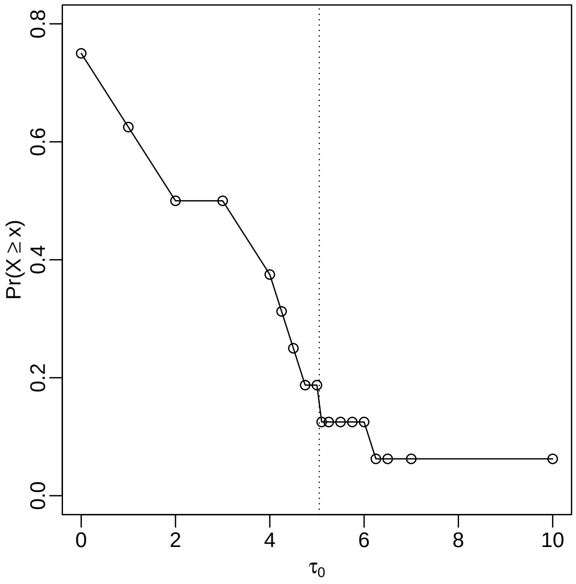


Figure 2: One-sided p-values for hypotheses about τ_0 in the model which implies $r_{bi,0} = R_{i,b} - Z_{i,b}\tau_0$. Dotted line at $\tau_0 = 5.05$ shows break between $\tau_0 = 5$ and $\tau_0 = 5.1$.

desire a one-sided $87.5 \approx 88\%$ confidence interval) we discover that our observed values are not so surprising from the perspective of hypotheses where $\tau_0 \leq 5$. However, it would be 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.[7]

It is one thing to know that we can or cannot reject the sharp null, but it is much more substantively interesting to know that effects greater than 5 points of turnout would make our observed data implausible. In order to move beyond the null of no effects, we did have to specify something about the structure of the effects: we pretended that it was scientifically useful to say that $h(r_{bi,0}) = r_{bi,1} = r_{bi,0} + \tau$ — perhaps this hypothesis generator follows from some past theory or literature or experience even if it is not immediately 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 strange from the perspective of the hypotheses — that is, the hypotheses are a kind of lens through which we can see our data, or a question we can ask, and the p-values which define ranges

[7] In large samples, we consider p-values to be continuous, but in finite samples, there are similarly a finite number of possible p-values. In our example, there are $2^4 = 16$ possible assignments of treatment due to the blocking scheme (i.e. within each blocked pair, we can consider treatment the flip of a fair coin). Therefore, even the most extreme one-sided p-value can be no smaller than $\frac{1}{16} = .0625$. Similarly, precise 95% confidence intervals may not be possible. For this example using a one tailed tests, one must choose from 100*(15/16)=93.75% (i.e. , 100*(7/8)=87.5% or smaller, but still discrete, confidence intervals.

of hypotheses are like the answers we give to the optometrist when she places different lenses over our eyes asking “which is better, this? or this?”. 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. And Fisher’s testing framework enables us to use our data and design to reflect back on the perspective in a formal manner.

However, the hypotheses and hypothesis generator, or model of effects, which we used here did imply 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).

4 Assessing hypotheses about interference in a Small Field Experiment

If we can write $r_{1i} = h(r_{bi,0}) = r_{bi,0} + \tau$ then we can write other functions of potential outcomes in the control condition which likewise generate a list of $r_{bi,1}$ implied by our hypotheses.[8] To make concrete how one might specify hypotheses about interference between units, we continue to use the eight city field experiment and then complicate our hypotheses later when we consider an experiment run on a social network.

Although we might wonder about interference of all possible kinds, a glance at the map in Figure 1 allows two sets of cities suggest themselves as particularly plausible candidates for interference: the pair of Yakima and Richland in Washington State (interference within pair), and Saginaw and Midland in Michigan (interference across pair).[9].

Of course in any given design allows the set of potential hypotheses is essentially infinite. Just as theory, and other substantive concerns, must guide and constrain choice of research design amid the infinitely many ways to observe the world, so too, must theory help scholars select among the possible hypotheses to assess (or models to estimate). Earlier in the paper we commented that common assumptions required of causal inference, like SUTVA, play a role in Fisher’s randomization inference as implications of hypotheses to be rejected rather than assumptions to justified. We demonstrated a version of this claim with tests of no effects. Now, this section demonstrates this claim by considering some simple hypotheses about interference among units. In doing so it shows both the flexibility of Fisher’s approach to statistical inference, and highlights the way in which social science theory can be crucial for statistical inference — by specification of hypotheses.

If we are going to entertain hypotheses about spill-over we need to be more specific: for the purposes of this example, let us presume that treatment might “spill-over” from a treated unit onto nearby control units. For simplicity and to introduce notation let us first consider the situation of one-way interference: perhaps we imagine that news in the larger Yakima, WA is likely to reach the

[8] See for example, (Bowers and Panagopoulos 2011, § 3.3).

[9] Saginaw is paired with Sioux City and Midland is paired with Battle Creek

smaller Richland, but not vice-versa. One might formalize this simple case for two cities with as follows.

First, define the potential outcomes for the two cities: The potential outcomes for any city i given our mini-design (one pair, one binary treatment assigned) are written in a general form as $r_{i,Z_i=1,Z_j=0} = r_{i,\mathbf{Z}=\{1,0\}} = r_{i,10}$, $r_{i,01}$, $r_{i,11}$,$r_{i,00}$. These last two potential outcomes, when both units are treated and when both units are not treated are never directly observable in our design, whereas the first two are partially observable. Yet, including the full range of possible potential outcomes will be useful as we will see.

Second, write the function that links potential outcomes to observed outcomes via treatment assignment:

$$R_i = Z_i(Z_j r_{i,11} + (1 - Z_j)r_{i,10}) + (1 - Z_i)(Z_j r_{i,01} + (1 - Z_j)r_{i,00}) \tag{7}$$

Equation 7 shows that we can relate what we observe to the pattern of treatment assignments in the two cities, i and j. We saw this equation before (equation 1) when we considered hypotheses about no effects in the presence of interference. In the next steps we will simplify this identity based on our hypotheses.

Third, express our ideas about spill-over and treatment effects and other relations among potential outcomes. In the no interference case, our hypothesis generators $h()$ were functions that specified how we thought control responses would turn into treated responses. That is, the treatment effects were a function of potential treated and potential control responses with control responses being the baseline against which comparisons would be made: for example that $r_{i,1} = r_{i,0} + \tau$. Now, when we have interference in this simple case, we have two different potential responses to control: $r_{i,00}$ — potential response when no city receives any treatment (the "uniformity trial") — and $r_{i,01}$ — the potential response when city i does not receive treatment but city j does receive treatment (we might call this the "only spill-over" 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. At minimum "no effects" in a situation with spillover is the situation in which the intervention was not fielded at all.

In this example, for simplicity, 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 $r_{i,10} = r_{i,11}$). This leads us to simplify equation 7:

$$R_i = Z_i r_{i,10} + (1 - Z_i)(Z_j r_{i,01} + (1 - Z_j)r_{i,00}) \tag{8}$$

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

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

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

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

$$R_i = Z_i \left(\left(1 - Z_j\right) (\tau + r_{i,\{0,0\}}) + Z_j \left(\tau + r_{i,\{0,0\}}\right) \right) + (1 - Z_i) \left(\left(1 - Z_j\right) r_{i,\{0,0\}} + Z_j \left(w\tau + r_{i,\{0,0\}}\right) \right) \tag{10}$$

which simplifies to

$$R_i = r_{i,\{0,0\}} + \tau Z_i + w\tau Z_j - w\tau Z_i Z_j. \tag{11}$$

Equation 11 allows us to recover $r_{i,00}$ under certain hypotheses about w and τ:

$$r_{i,\{0,0\}} = R_i - \tau Z_i - w\tau Z_j + w\tau Z_i Z_j = R_i - \tau Z_i - w\tau Z_j. \tag{12}$$

Notice that when $\tau = 0$, $r_{i,00} = R_i$ as implied by Fisher's sharp null hypothesis, when $w = 0$, we have the adjustment implied by the simple constant, additive effects model, $R_i - Z_i\tau$. 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.

So, this is the implication of our hypothesis for Richland — a city we presume could receive some spill-over from Yakima. Our idea is that people in the smaller nearby city of Richland are more likely to read Yakima newspapers but that people in Yakima probably are not reading Richland newspapers, so the hypothesis generator or model of effects for Yakima would merely be the constant effects model just posited. Say, Yakima, is city j (as it is in our development of the testing framework for Richland). For Yakima we presume no interference and a simple constant, additive effects model such that:

$$h(r_{j,00}) = h(r_{j,01}) = r_{j00} + \tau = r_{i,11} = r_{i,10}$$

So, for Yakima, the observed outcome identity simplifies even further that it did for Richland:

$$R_j = Z_j r_{j,10} + (1 - Z_j) r_{j,00} \tag{13}$$

And we can adjust R_j to reflect hypotheses about τ as we did under the no interference model: $r_{j,00} = R_j - Z_j\tau$.

With $r_{j,00} = R_j - Z_j\tau$ and $r_{i,00} = R_i - Z_i\tau - Z_j w\tau$ we can thus 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 the traffic between them). 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): $r_{k00} = R_k - \tau Z_k - w\tau Z_l$.

Equation 15 summarizes the adjustments implied by our different hypotheses for this field experiment in which we wanted to consider one-way interference from Yakima to Richland, two-way interference between the two neighboring Michigan cities, and no interference for the other cities in the dataset. The hypothesis generator would then be:

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

Equation 14 implies the following adjustments to observed outcomes, R_i to enable the generation of the randomization distribution representing the hypotheses implied by equation 14:

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

Equation 14 can be read as saying that the potential outcome to control for city i when no cities receive treatment, $r_{i,00}$, can be recovered by modifying what we observe, R_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 (and is the same across all cities). But for other cities we consider the idea that, if specific other cities received the treatment, some of that treatment (the proportion, w of it) would be experienced by the city in the control condition.

Equation 14 provides observable implications for our hypotheses, and thus we can test these hypotheses about treatment effects τ and spillover amount w. Figure 3 shows the two-dimensional one-sided confidence region bounded by those tests.

The hypothesis generator introduced in equation 14 involved a parameter w in order to enable us to engage briefly with the question about such non-causal parameters. 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 of the type demonstrated in the previous section. If w is not fixed but is considered a nuisance parameter, then one may draw 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) produce 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). Either solution will maintain the correct coverage of the resulting confidence intervals about treatment effects, although the merely using the largest p-value is apt to make those confidence intervals overly conservative.

Parameters like w, however, need not be a nuisance. Rather, in this case, w represents the extent of spillover (assumed the same in both the one-way and the two-way cases for simplicity). The approaches to w as a nuisance parameter to be estimated involve, in essence, exploring a 2-dimensional slice through the set of possible hypotheses (in comparison to the standard 1-dimensional slice that we have been representing with confidence intervals with upper and lower boundaries). So, in fact, we could easily produce a confidence region for w_0, τ_0 pairs which would encode the evidence our data brings against such joint hypotheses.

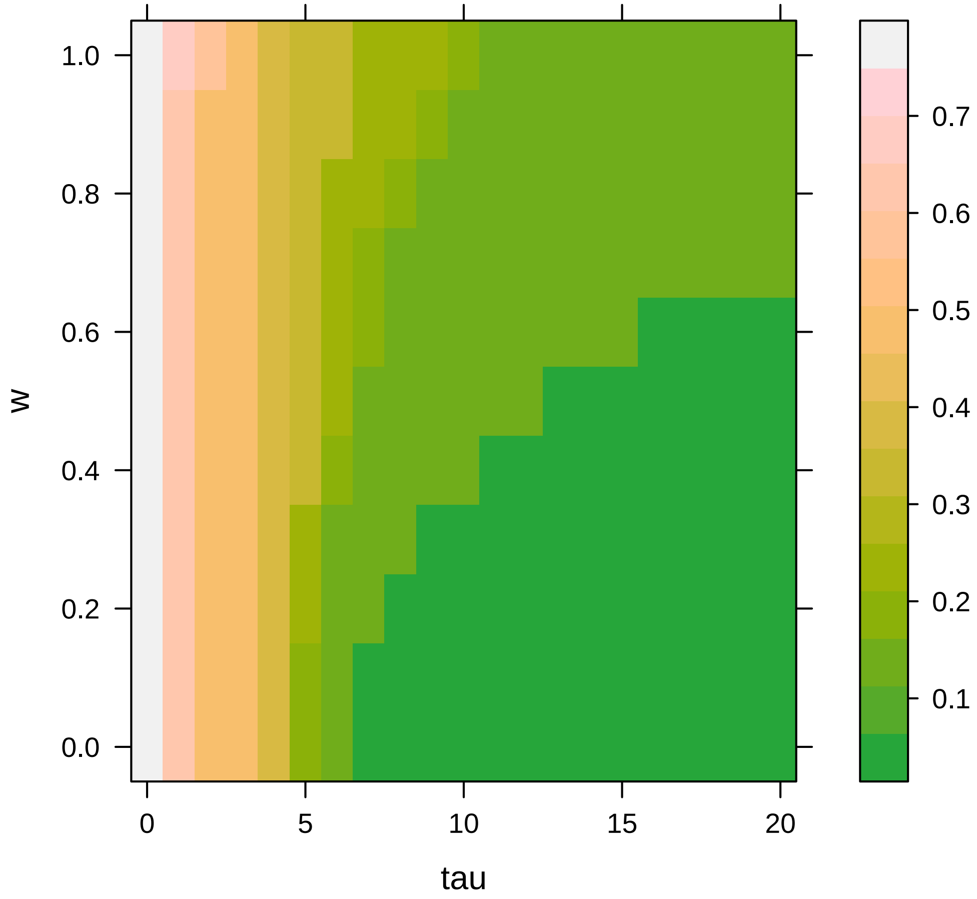


Figure 3: Tests of joint hypotheses about w and τ generated by the model in equation 14. The plot shows areas delimited by the one-sided p-values for the tests of each joint hypothesis. Lower p-values are plotted in darker green. Higher p-values are plotted in lighter colors. (See the key at the right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

Figure 3 is one representation of such a confidence region using our 8 city data. The shading of the plot shows the one-sided p-values from assessing the different combinations of w_0 and τ_0. The lower bound of the one-sided intervals is, of course, unbounded, since in this case since we do not consider $\tau_0 < 0$. 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 constant effects hypothesis generator. As we begin to entertain hypotheses about some positive amount of spill-over, 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 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 can reject hypotheses about w which are less than $w_0 = .2$ 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. The main point is that one may reason directly about interference and such reasoning, if formalized, may produce hypotheses about both causal effects and structural or mechanistic features of the effects. The data can provide evidence against such hypotheses, and, in fact 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 the maintained set of hypotheses and also about the amount of information available to make such plausibility assessments.

4.1 "No interference" is a specific hypothesis to be assessed in Fisher's framework not an untestable assumption

So we have a proof of concept regarding the link between foundational assumptions of causal inference (such as "no interference") and the Fisher hypothesis testing framework. Moreover, this exercise has demonstrated the flexibility of the method regarding the ability to link social science substantive theory with implications for counterfactuals and with observed data under this framework.

5 Assessing Machines and Models

We showed that hypotheses about arbitrary patterns of interference are possible to articulate, link to a well-established mode of statistical inference, and possible to execute. However, more questions arise: Does our machine for statistical inference have good operating characteristics? If we have spillover of a known type and a primary treatment effect of known amount, would our method recover it? How should we formalize and test patterns of interference in larger samples? Here we use a simulated social network to test our method and also to show how these ideas may be profitably applied to the study of social networks. In the next section we engage with the question writing down and restricting attention to what may appear to be infinitely many possible hypotheses. In this section, we engage with the question of choice of models.

Social scientists reason using models. An advantage of our approach here is that it allows models on the substance of the causal process to have direct implications for data — rather than requiring separate data models and theoretical models. Yet the question must always arise about whether and how a particular model (interesting or not, useful or not) is supported by the data. We engage with this question in two ways. First, in a narrow technical sense, questions about how a method behaves when the model is wrong (because the parameters are wrong) are questions about the operating characteristics of the method. For example, a well operating statistical method ought to reject a true null rarely (and, in fact, no more than the promise made by the stated α significance level). That is, a good statistical inference machine should have (1) correct coverage of confidence intervals and (2) it also ought to reject false nulls rather quickly as they diverge from the truth [i.e. be powerful]. In a series of simulations, we show that our proposal has good operating characteristics: when we assess hypotheses about "correct" models it would not mislead researchers into rejecting the null too quickly, and it does reject hypotheses that are far from the truth. Notice the qualifier here: about "correct" models. This is what classical frequentist operating characteristics are meant to reflect on: at minimum, when we ask the right question, we should not get misleading answers.

Second, and more difficult, are questions about how to assess wrong models. There are more ways for a model to be wrong than right. And, in the end, "all models are wrong."[cite] So, we engage with this question by assessing hypotheses using incorrect and correct models in simulations where we know the true model and true parameters. We further compare 1 parameter models and 2 parameter models in which one parameter is interpretable as a "spillover effect."

5.1 The Simulation Setup

For our simulation, we imagine a simply randomized experiment over a known and fixed social network in which half the units are assigned to treatment. Figure 4 shows this network as a collection of vertices/nodes and edges/lines 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 $n > 8$.

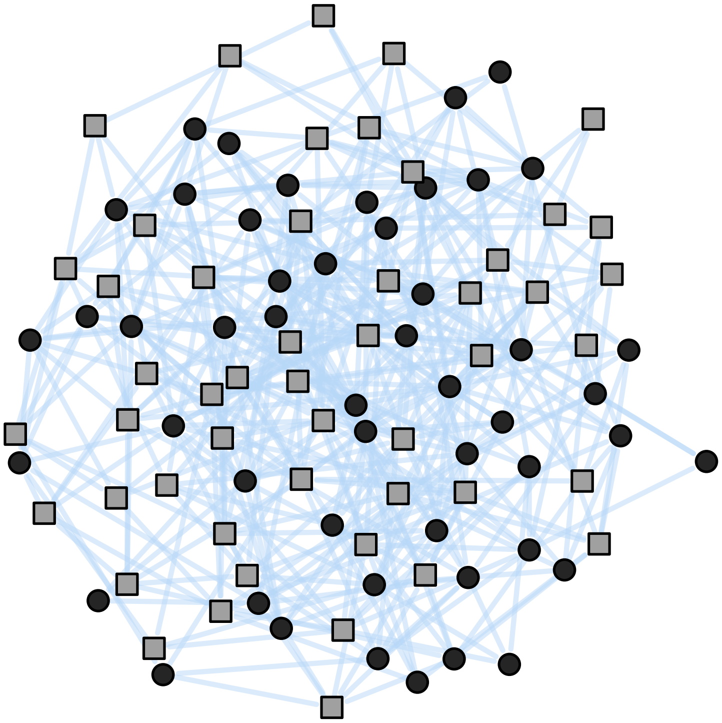


Figure 4: An undirected social network assigned to treatment (circle) or control (square). Probability of a connection between any two units is 0.08 for all units (so this is a random network).

We can summarize networks with matrices. In this case, for example, the first 4 units in this network have the following adjacency matrix:

```
  1 2 3 4
1 0 0 1 0
2 0 0 0 0
3 1 0 0 0
4 0 0 0 0
```

This matrix shows that unit 1 is connected to unit 3 (and because we have undirected connections, unit 3 is also connected to unit 1). Units 2 and 4 are not connected to each other or units 1 and 3.

In this network, the number of connections ranges from 3 to 16. And the number of connections to treated units range from 0 to 10 with 50% of units connected to between 3 and 5 units. The outcomes in the absence of any intervention in the network, $\mathbf{r_0}$ — the responses to the "uniformity trial" — are generated simply as draws from a Normal distribution, $N(n/2, (n/10)^2)$ such that it ranges from about 30 to 78. We consider two "right" models chosen to display network related dependence in relationships among potential outcomes. And, 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). Specifically, we consider the situation in which it is the number of directly connected treated units which matters for treatment effects.

First, consider a 1 parameter model in which the number of treated connections has a non-linear and non-monotonic relationship with a constant, additive treatment effect. The function linking number of treated connections to outcomes is a cubic polynomial:

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

Equation 17 shows how observed outcomes are generated additively from the uniformity trial using different versions of the linking function in the treatment and control groups (i.e. the treatment and control groups experience network effects differently although both functions are from the same family). We call the $n \times n$ network adjacency matrix $\mathbf{S}$ and so the inner product of the vector of treatment assignments $\mathbf{Z}^T\mathbf{S}$ is the number of treated connections.

$$\mathbf{R} = \mathbf{r_0} + \mathbf{Z}\tau f(\mathbf{Z}^T\mathbf{S}, -2, 10, 7) + (1 - \mathbf{Z})\tau f(\mathbf{Z}^T\mathbf{S}, 0, 5, 10) \tag{17}$$

We consider the situation where $\tau = 50$ (this is a very large effect on the scale of the outcome). 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 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. The candidate "wrong" model is the model of constant treatments in which turnout is shifted up a constant amount in all treated units — and in which network effects and the network is ignored.

A second "right" model involves two parameters, τ_1 and τ_2. The function linking number of treated connections to outcomes, equation 18, is a simple nonlinear function which approaches 1 as $x \rightarrow \infty$ and so can be understood as a proportion of the treatment effect.

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

Equation 19 shows how observed outcomes are generated additively from the uniformity trial using different versions of the linking function in the treatment and control groups (i.e. the treatment and control groups experience network effects differently although both functions are from the same family). Only here, the story involves both a network spillover effect, τ_2 and an additive treatment

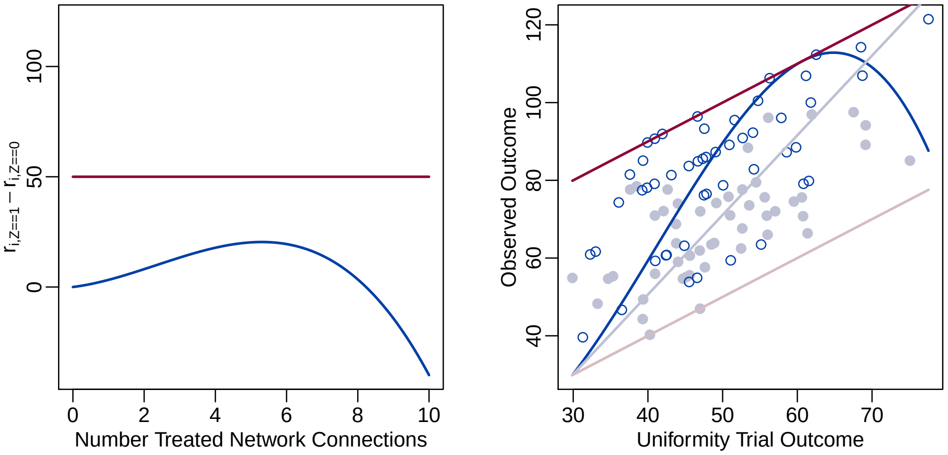


Figure 5: Comparing implications of nonlinear network moderated treatment effects. Left panel shows implications of two models for unit level differences in potential outcomes. The constant additive effects model, a red straight line, $r_{1i} = r_{0i} + \tau$, ignores the network (here $\tau = 50$). The cubic polynomial (a curved blue line) is a function of number of treated units. As the network local to treated unit saturates with other treated units, the response to treatment if treated ($r_{i,Z=1,...}$) goes below the response to treatment if control ($r_{i,Z=0,...}$). The right panel shows the same functions, only now we see the simulated treated (dark blue, open circles) and control (light blue, solid circles) observations. Observed outcomes (measured on the y-axis) were generated by application of the models depicted by the dark and light blue cubic polynomial model curves for the treated and control groups respectively to the $n = 100$ draws from $N(n/2, (n/2)^2)$ representing the uniformity trial. The red lines show the implications of the model of constant treatment effects. Here, τ is the vertical distance between the light red 45 degree line (showing $r_{0i} = r_{i,0...0}$) and the dark red line (showing $r_{1i} = r_{0i} + \tau$).

effect τ_1. Units in the control group do not get the direct effect of τ_1 but get some proportion of τ_2 depending on the number of directly connected treated units.

$$\mathbf{R} = \mathbf{r_0} + (\mathbf{Z}(\tau_1 + \tau_1\tau_2 g(\mathbf{Z}^T\mathbf{S})) + (1 - \mathbf{Z}) * (\tau_2 g(\mathbf{Z}^T\mathbf{S}))) \tag{19}$$

The benefit of this model is that when $\tau_2 = 0$, the model of effects is merely the constant additive effects model. That is, we can not only assess hypotheses about τ_1 and τ_2, but we can compare those effects to those from a constant additive effects model (which may be a theoretically useful comparison).

Now, there are several questions we ask about models and their relationships to statistical and causal inference. We can summarize these questions as being of two kinds. The first kind questions are about the model being correct but hypotheses about the parameters of the model being incorrect. For example, statistical inferences for our model when we ask questions about the true parameter ought to make errors in a controlled fashion (i.e. confidence intervals should have correct coverage, the Type I error rate should be controlled). The second kind of questions are about the model itself being incorrect (and thus tests of correct or incorrect hypotheses about parameters being hard to

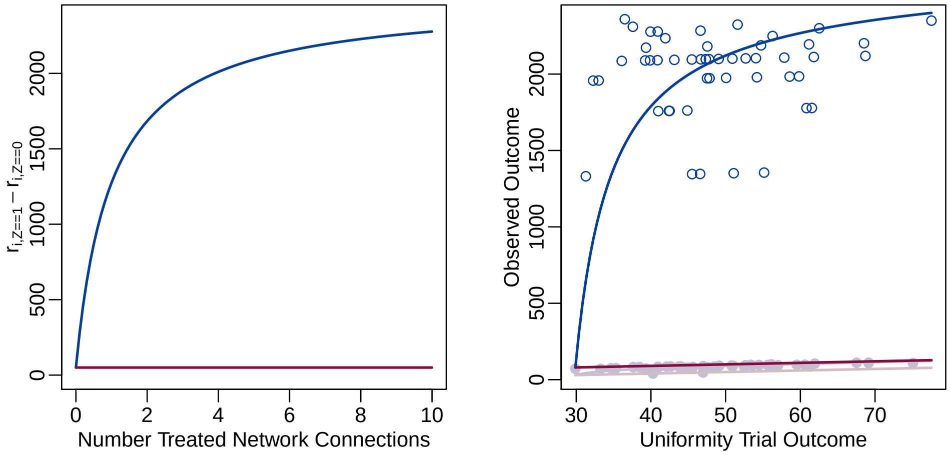


Figure 6: Comparing implications of nonlinear network moderated treatment effects for a two parameter model. Left panel shows implications of two models for unit level differences in potential outcomes. The constant additive effects model, a red straight line, $r_{1i} = r_{0i} + \tau$, ignores the network (here $\tau = 50$). The two parameter model is a function of number of treated units. As the network local to treated unit saturates with other treated units, the response to treatment if treated ($r_{i,Z=1,...}$) rapidly increases beyond the response to treatment if control ($r_{i,Z=0,...}$) and then the rate of change of this divergence slows as the numbers of connected treated units continues to increase. The right panel shows the same functions, only now we see the simulated treated (dark blue, open circles) and control (light blue, solid circles) observations. Observed outcomes (measured on the y-axis) were generated by application of the models depicted by the dark and light blue cubic polynomial model curves for the treated and control groups respectively to the $n = 100$ draws from $N(n/2, (n/2)^2)$ representing the uniformity trial. The red lines show the implications of the model of constant treatment effects. Here, τ is the vertical distance between the light red 45 degree line (showing $r_{0i} = r_{i,0...0}$) and the dark red line (showing $r_{1i} = r_{0i} + \tau$).

interpret). We engage with both sets here.

5.2 Right Model, Wrong Parameter: Coverage and Power

At minimum, we want our statistical inference procedures to do the right thing when we are correct. “True” null hypotheses should be rejected no more than $100\alpha\%$ of the time at level α. And as hypotheses about τ (and τ_1 and τ_2) diverge from the truth, we want to be able to reject those hypotheses more and more readily (i.e. we would like power to reject false null hypotheses and thus produce reasonably tight confidence intervals).

We address both of these questions using a simple simulation study using the following algorithm:

1. Draw a vector of treatment assignments, $\mathbf{Z}$ 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 uniformity trial (100 draws from a $N(n/2, (n/2)^2)$ distribution with $n = 100$) following the “right” model (either following equation 17 or 19 depending on the simulation).

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, if $\tau_1 = 30$ and $\tau_2 = 30$ we assess all the integer hypotheses for which $10 \le \tau_1 \le 50$ and $10 \le \tau_2 \le 50$ (about 50*50=2500 hypotheses).
4. Then go back to step 1: draw another "temporary real" treatment assignment.
5. Repeat about 1000 times.

5.2.1 Does our proposal keep its promises?: Type I Error Rate

Figure 7 shows in the right hand column the answer to this question (which is, simply, "yes"). If statistical inferences based on complex network models are to be trustworthy, they should encourage us to reject true null hypotheses no more than a set proportion of the time (this proportion, α is often called the Type I error rate — the rate of falsely rejecting true null hypotheses). We can see that when we ask a correct question of the data (whether the question has one or two parameters, or whether interference takes one form or another), we will not mislead with our statistical inferences.

We return to the question of asking wrong questions after we address power.

5.2.2 Does our method more quickly reject wrong parameters as they get more wronger?

A test is powerful relative to other tests if it rejects incorrect hypotheses frequently across replications of the hypothetical experiment. Here, we assessed one correct hypothesis 1000 times (with detailed results shown in Figure 7). For the one-parameter model, we assessed 99 other incorrect hypotheses 1000 times each. If our procedure is working as we would like it to, our test should encourage us to reject the incorrect hypotheses at level α more than $100\alpha\%$ of the time (across simulations). Figure 8 displays the results of these simulations for $\alpha = .05$. The second column shows the power functions of the different correct models. As desired, these two models reject the true null very rarely (well within simulation error of 5% of the time), and, as the hypotheses diverge from the truth, the proportion of p-values less than .05 goes toward 1. That is, in each case, there are some hypotheses that are just so inconsistent with the data that they rejected in nearly all of the simulations.

5.3 Wrong Models

The previous two figures also showed Type I error rate and power for constant, additive effects models when applied to the two different network effects models (the panels with "Constant Effects" in the title). In both cases, not only were the true parameters rejected in almost all simulations, but most proposed parameter values were rejected. This is especially true of the constant effects model as applied to the two parameter network effects simulated outcomes — there, every joint hypothesis was rejected at $\alpha = .05$ (i.e. the data were surprising from the perspective of nearly all parameters that one could provide this model of effects). The one parameter constant effects model

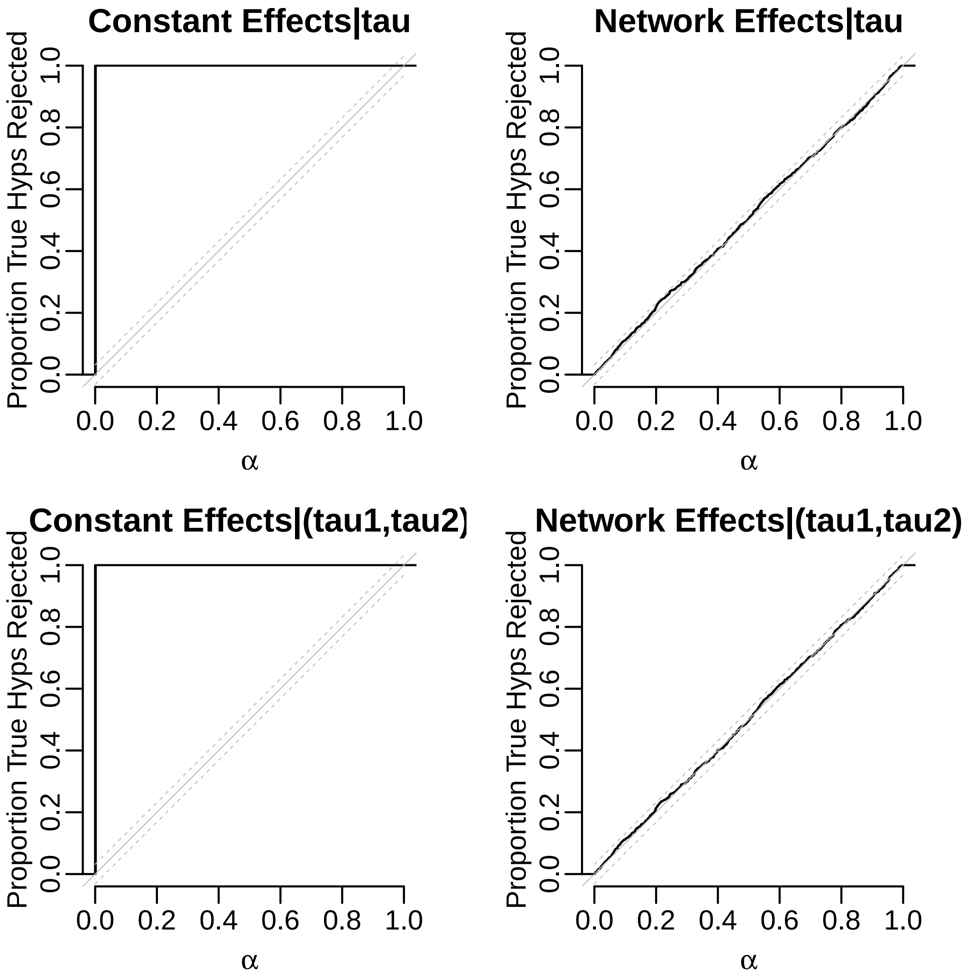


Figure 7: For simulated data (n = 100, over 1000 simulations) with network moderated treatment effects, the proportion of true null hypotheses rejected at level α for all α. The panels labeled "Network Effects" show the correct models assessing the true parameters. The panels labeled "Constant Effects" show the constant effects model (which ignores interference) assessing the true parameters. The constant effects models reject always, whereas the network effects models reject just about 100α % of the time (within the simulation error range — shown by the dashed gray lines.)

is not irrelevant to the one parameter cubic network effects model — although even the hypothesis that it rejects least often, it does so at the .05 level more often that one would like (about 10% of the time). So, varieties of the wrong question lead to consistent answers of "the data would be surprising if your hypothesis were true".

Notice that we would not have received such information had we not run these small simulation studies of the operating characteristics of our hypothesis tests under the different models. We suspect that using such simulations (as encoded in easy to use software that we have written) may be a wise part of the workflow of those desiring to ask specific and interesting (although probably not entirely "correct") questions of their data.

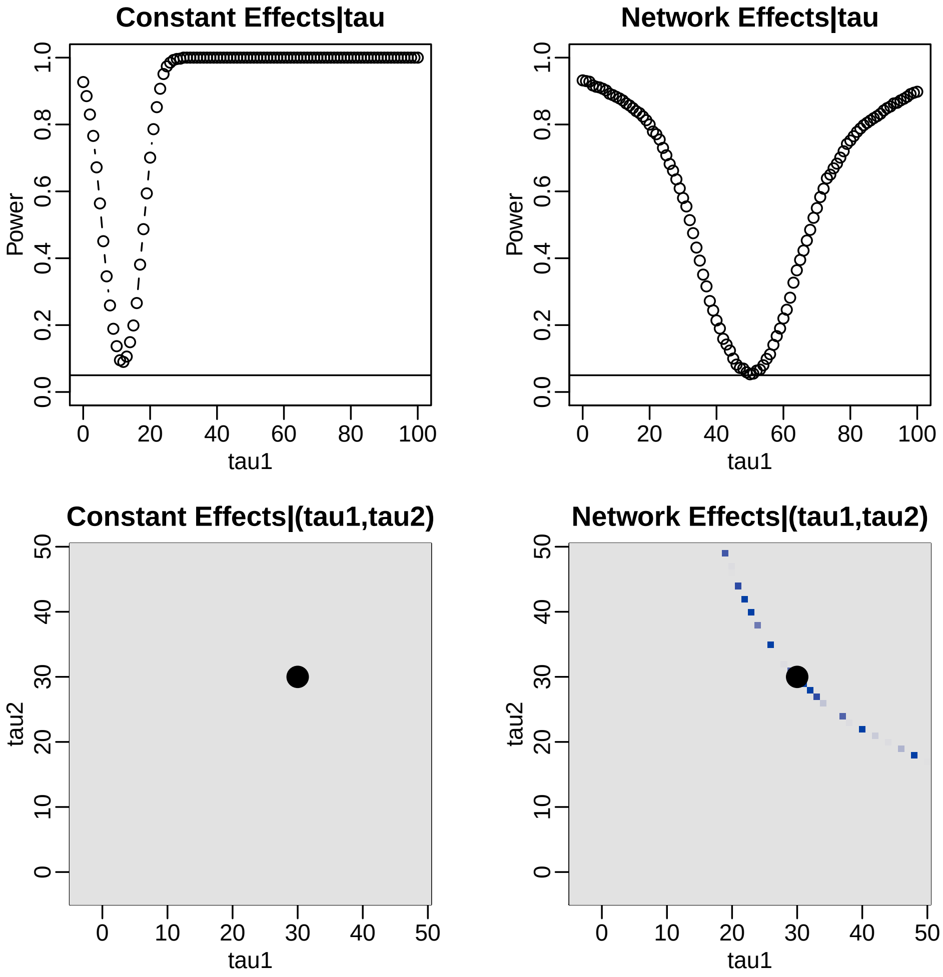


Figure 8: Power to reject hypotheses in one- and two-parameter simulated outcomes when the model is incorrect (the "Constant Effects" column) and when the model is correct (the "Network Effects" column). Power is represented as the proportion of the time a hypothesis is rejects at the $\alpha = .05$ level. In the bottom row, lighter colors show higher proportions of rejections, darker colors show lower proportions of rejections. The "Constant Effects|(tau1,tau2)" panel is entirely $p = 1$ — i.e. rejecting all hypotheses at $\alpha = .05$ always. The "truth" is $\tau = 50$, and $\tau_1 = 30, \tau_2 = 30$.

5.4 Deeper Points about Models

In the end, the truth is a treatment effect for each individual unit. For 100 units, we might have 100 different effects. We could imagine testing every possible combination of those effects: this 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 impossible. And it is not clear that such an effort would be fruitful. As Rosenbaum notes in his discussion 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 — we must ask which slices of that 100 dimensional space are scientifically interesting. And notice that if we reject a set of τ using some model, we would also reject those τ for any model which could be reduced to the simple model. So, even testing a wrong model can be useful as long as we can know how the rejections from this model relate to substance.

Refering to our 100 dimensional space as $2I$, and our τ as θ, Rosenbaum 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)

The point of this paper is to show that one may reason about interference between units not to offer the holy grail of knowing when a model is useful (Whether a model is "true" is probably a question not worth asking except in a simulation study (cite Clarke and Primo).) Now, if we are offering the possibility of new questions to ask (without specifying in particular the set of such questions because such would also be far beyond the scope of a single methodology paper), we should also offer some conceptual and notation help. After all, one can only test hypotheses that are specific enough to leave particular traces and patterns in the data after we apply the adjustments implied by them.

6 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.

6.1 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 9 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 9 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.

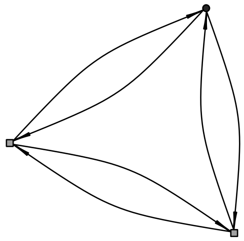


Figure 9: A Simulated Network and field experiment: treatment (circles) and control (squares). Without further assumptions, treatment or control assigned to any unit may influence any other unit. The edges have arrows to show that influence may be directional.

The vector of possible potential outcomes for unit 1, $\mathbf{r}_{1,\cdot}$, given the graph in figure 9 and no further assumptions in lexicographic order is:

$$\mathbf{r}_{1,\cdot} = \{r_{1,\{111\}}, r_{1,\{110\}}, r_{1,\{101\}}, r_{1,\{100\}}, r_{1,\{011\}}, r_{1,\{010\}}, r_{1,\{001\}}, r_{1,\{000\}}\} \tag{20}$$

If an arrow does not connect a unit i and to another unit j, this means that is meaningful to write $r_{j,Z_i,\mathbf{Z}_{(-i)}} = r_{j,Z_i',\mathbf{Z}_{(-i)}}$ 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, $r_{j,Z_i=1,\mathbf{Z}_{(-i)}} = r_{j,Z_i=0,\mathbf{Z}_{(-i)}}$. 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, R_1, and the potential outcomes shown in equation 20: what is the observed outcome identity equation implied here?

In scalar form we might write this identity as follows:

$$\begin{aligned} R_i =& Z_3\Big(Z_2\big(Z_1 r_{i,\{i,1,1\}} + (1 - Z_1) r_{i,\{0,1,1\}}\big) + (1 - Z_2)\big(Z_1 r_{i,\{i,0,1\}} + (1 - Z_1) r_{i,\{0,0,1\}}\big)\Big) + \\ & (1 - Z_3)\Big(Z_2\big(Z_1 r_{i,\{i,1,0\}} + (1 - Z_1) r_{i,\{0,1,0\}}\big) + (1 - Z_2)\big(Z_1 r_{i,\{i,0,0\}} + (1 - Z_1) r_{i,\{0,0,0\}}\big)\Big) \end{aligned} \tag{21}$$

Notice that equation 21 specifies the circumstances under which what we observe for unit 1, R_1, represents any of the potential outcomes possible from the complete graph and no further restrictions. For example, it says that we would observe $r_{\mathbf{Z}=\{1,1,1\}}$ 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:

$$\boldsymbol{\rho}_i = \begin{pmatrix} r_{i,\{1,1,1\}} & r_{i,\{1,1,0\}} & r_{i,\{1,0,1\}} & r_{i,\{1,0,0\}} \\ r_{i,\{0,1,1\}} & r_{i,\{0,1,0\}} & r_{i,\{0,0,1\}} & r_{i,\{0,0,0\}} \end{pmatrix} \tag{22}$$

Equation 21 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

$$R_i = \mathbf{1}_{(1\times2)} \cdot (\mathcal{Z}_i \times \boldsymbol{\rho}_i) \cdot \mathbf{1}_{(2^{n-1}\times1)}, \tag{23}$$

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 23 showing the full matrices (but doubly transposed to fit on the page) for $n = 3$:

$$R_i = \begin{pmatrix} 1 & 1 \end{pmatrix} \cdot \left[\begin{pmatrix} Z_1 Z_2 Z_3 & (1 - Z_1) Z_2 Z_3 \\ Z_1 Z_2 (1 - Z_3) & (1 - Z_1) Z_2 (1 - Z_3) \\ Z_1 (1 - Z_2) Z_3 & (1 - Z_1)(1 - Z_2) Z_3 \\ Z_1 (1 - Z_2)(1 - Z_3) & (1 - Z_1)(1 - Z_2)(1 - Z_3) \end{pmatrix} \times \begin{pmatrix} r_{i,\{1,1,1\}} & r_{i,\{0,1,1\}} \\ r_{i,\{1,1,0\}} & r_{i,\{0,1,0\}} \\ r_{i,\{1,0,1\}} & r_{i,\{0,0,1\}} \\ r_{i,\{1,0,0\}} & r_{i,\{0,0,0\}} \end{pmatrix} \right]^T \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{24}$$

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, $\mathbf{R}$, simply by multiplying R_i by $\mathbf{1}_{(n\times1)}$, $\mathbf{R} = R_i \times \mathbf{1}_{(n\times1)}$.

6.1.1 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.

6.2 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. Of course, we do not have to engage directly with infinity (or near infinity) in this way. 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 is easier.

6.2.1 Pruning by Design

Most of the potential outcomes listed in lists such as equation 20 will never occur in any real design.[10] 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.[11]

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.

6.2.2 Pruning by Knowledge of Structure

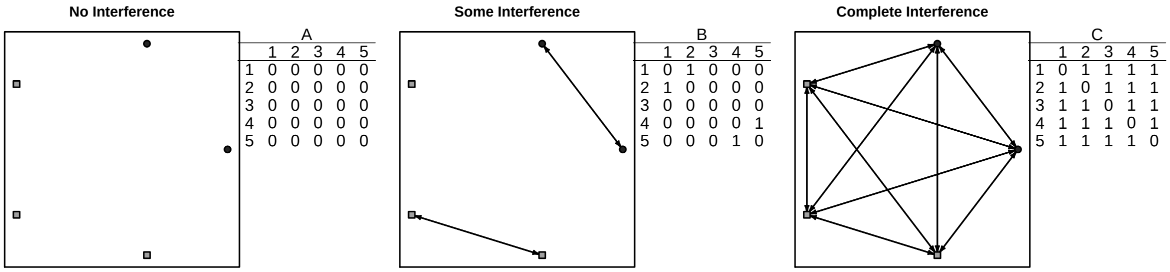
We say "knowledge" here to distinguish it from "hypotheses about structure" although, of course, we could include such structural statements as hypotheses. However, in many applications there are subsets and groupings or even types of interference which are just not credible or would never be interesting.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

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

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

the complete graph even more.

Figure 10 shows three plots representing certain structural presumptions about interference and the related adjacency matrices for the case of $n = 5$.



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 10: Graphs and corresponding adjacency matrices.

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

$$\mathbf{B} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \tag{25}$$

The restrictions on the potential outcomes for unit 1 are those listed in the first column of **B**. In that column we have 3 zeros in positions $\{(3,1),(4,1),(5,1)\}$. These zeros imply the following equality: $r_{1,\mathbf{Z}_{\{3,4,5\}},\mathbf{Z}_{-(\{3,4,5\})}} = r_{1,\mathbf{Z}'_{\{3,4,5\}},\mathbf{Z}_{-(\{3,4,5\})}}$ 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 treatmetn with $n = 5$ with no further information implied $2^5 = 32$ potential outcomes for each unit. The design of the study reduced 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

$\{r_{3,\{.,.,0,.,.\}}, r_{3,\{.,.,1,.,.\}}\}$ and for the other units (which interact with only one other unit) we have 4 potential outcomes $\{r_{i,\{0,0,.,.,.\}}, r_{i,\{0,1,.,.,.\}}, r_{i,\{i,0,.,.,.\}}, r_{i,\{1,1,.,.,.\}}\}$ 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 is as 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,

$$\begin{aligned}\text{Pos}(\mathbf{B}, 0) = &\{\{1,3\},\{1,4\},\{1,5\},\\ &\{2,3\},\{2,4\},\{2,5\},\\ &\{3,1\},\{3,2\},\{3,4\},\{3,5\},\\ &\{4,1\},\{4,2\},\{4,3\},\\ &\{5,1\},\{5,2\},\{5,3\}\}\end{aligned} \tag{26}$$

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

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 with 1s on the diagonal, $\mathbf{B}$ 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: 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 by which control responses turn into treated responses?

6.2.3 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: $r_{i,Z_i=1,\mathbf{Z}_{(-i)}} = r_{i,Z_i=1,\mathbf{Z}'_{(-i)}}$ and $r_{i,Z_i=0,\mathbf{Z}_{(-i)}} = r_{i,Z_i=0,\mathbf{Z}'_{(-i)}}$ 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 $r_{3,\{.,.,1,.,.\}} = r_{3,\{.,.,0,.,.\}} + \tau$ or $r_{3,Z_3=1,\mathbf{Z}_{(-3)}} = r_{3,Z_3=0,\mathbf{Z}_{(-3)}} + \tau$ 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: $\{r_{i,\{0,0,.,.,.\}}, r_{i,\{0,1,.,.,.\}}, r_{i,\{1,0,.,.,.\}}, r_{i,\{1,1,.,.,.\}}\}$ 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). For 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 $r_{i,\{0,1,.,.,.\}}, r_{i,\{1,0,.,.,.\}}, r_{i,\{1,1,.,.,.\}}$ as a function of $r_{i,\{0,0,.,.,.\}}$ 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 $r_{i,\{1,0,.,.,.,.\}} = r_{i,\{1,1,.,.,.,.\}}$].

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: $r_{i,Z_i=0,Z_j=1,.} = r_{i,Z_i=0,Z_j=0,.} + w\tau$ 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: $r_{i,Z_i=1,Z_j=0,.} = r_{i,Z_i=1,Z_j=1,.} = r_{i,Z_i=0,Z_j=0,.} + \tau$.

One could also imagine interesting hypotheses about all three potential outcomes: perhaps one might write both $r_{i,Z_i=1,Z_j=0,.} = r_{i,Z_i=0,Z_j=0,.} + \tau$ and $r_{i,Z_i=1,Z_j=1,.} = r_{i,Z_i=0,Z_j=0,.} + a\tau$ 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 aggregations of them. 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 $r_{i,Z_i=1,\mathbf{Z}_{(-i)}} = r_{i,Z_i=0,\mathbf{Z}_{(-i)}=0} + \tau + \sigma \mathbf{Z}^t\mathbf{S}$ and $r_{i,Z_i=0,\mathbf{Z}_{(-i)}} = r_{i,Z_i=0,\mathbf{Z}_{(-i)}=0} + \sigma \mathbf{Z}^t\mathbf{S}$.

6.2.4 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

|Ω|), 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 now and what we care about (from past theory and literature) to engage with manageable numbers of counter factuals in direct and substantively meaningful manners.

6.3 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 in equation 14. Now we have a more general way to formalize the process of hypothesizing about interference. Let us apply it to the newspaper advertisements study.

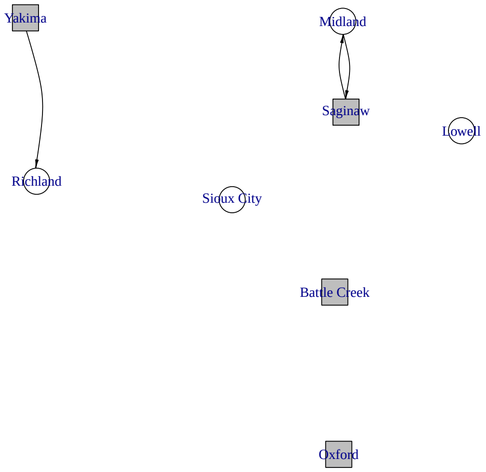


Figure 11: A directed network (or graph) representation of an interference hypothesis for the Panagopolous Newspaper study. Squares represent cities assigned to treatment. Circles are cities assigned to control. Arrows show direction of spillover: from the larger city of Yakima to the smaller city of Richland, and two way interference between Midland and Saginaw.

Figure 11 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): $r_{i,Z_i=0,Z_j=1}$ and $r_{i,Z_i=1,Z_j=0}$ 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: $r_{i,\{11\}}$,$r_{i,\{10\}}$,$r_{i,\{01\}}$,$r_{i,\{00\}}$, 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 $h(r_{i,Z_i=0,.}) = r_{i,Z_i=0,.} + \tau = r_{i,Z_i=1,.}$ could be rejected by our data, where we write $r_{i,Z_i=0,.}$ 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 $h(r_{i,Z_i=0,.}) = r_{i,Z_i=0,.} + \tau$. 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: $r_{i,10}$,$r_{i,01}$. 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: $h(r_{i,Z_i=0,Z_j=1}) = r_{i,Z_i=0,Z_j=0} + w\tau$ 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 $r_{i,11} = r_{i,10} = r_{i,1.}$ and

$h(r_{i,00}) = r_{i,00} + \tau = r_{i,1\cdot}$.

We listed those hypotheses in a condensed form in equation 14. And we can now see that the equations in equation 15 arise from solving each observed outcome identity equation 23 (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.

6.4 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.

7 Discussion and 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 know, in fact, that the average treatment effect is not even well identified or meaningful under interference.

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 part 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. We know that Fisher's test of the sharp null requires no claims about interference. And that Fisher's framework allows detection of interference (Aronow 2010) and, under certain conditions allows the creation of confidence intervals 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 hypothesis about imagined 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. We only mentioned it in passing, but it is worth noting here that the form of statistical inference used does not require that asymptotic justifications or assumptions about the stochastic processes generating outcomes (i.e. likelihood functions).

7.1 To Infinity and Beyond: Making and Justifying Choices about Hypotheses

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 confront a blank page (or screen) 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.

Notice that we not only face the need to make such decisions when we write, but also when we plan an experiment and plan analysis after the fact. 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 is it 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 appears 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 making 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 allow us to understand something new and useful about the world.

From this perspective, Fisher's sharp null is no different and should not be scary. 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.

7.2 Finally Theory

Where do hypotheses and models come from? 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.

References

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.

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

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

Gerber, A.S., D.P. Green and C.W. Larimer. 2008. “Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment.” *American Political Science Review* 102(01):33–48.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81:945–960.

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.

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*.

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 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.

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.