




## Estimation of Monotone Treatment Effects in Network Experiments

David Choi

**To cite this article:** David Choi (2016): Estimation of Monotone Treatment Effects in Network Experiments, Journal of the American Statistical Association, DOI: [10.1080/01621459.2016.1194845](https://doi.org/10.1080/01621459.2016.1194845)


**To link to this article:** <http://dx.doi.org/10.1080/01621459.2016.1194845>

 View supplementary material 

 Accepted author version posted online: 17 Jun 2016.  
Published online: 17 Jun 2016.

 Submit your article to this journal 

 Article views: 6

 View related articles 

 View Crossmark data 

# Estimation of Monotone Treatment Effects in Network Experiments

David Choi

## Abstract

Randomized experiments on social networks pose statistical challenges, due to the possibility of interference between units. We propose new methods for finding confidence intervals on the attributable treatment effect in such settings. The methods do not require partial interference, but instead require an identifying assumption that is similar to requiring nonnegative treatment effects. Network or spatial information can be used to customize the test statistic; in principle, this can increase power without making assumptions on the data generating process.

**Keywords:** causal inference, attributable effect, interference, randomized experiments, network data, Facebook, peer effects

## 1 Introduction

Spillover effects, social influence, and the sharing of information are widely believed to be important mechanisms for social and economic systems. To better understand them, researchers may collect network data on relationships between units. In some cases, the data may come from a randomized experiment; past examples include studies in viral marketing [Aral and Walker, 2011], voting behavior [Bond et al., 2012, Nickerson, 2008], online sharing [Kramer et al., 2014], education [Sweet et al., 2013], and health [Miguel and Kremer, 2004].

In such experiments, the outcomes tend to be social in nature, and the treatment of one individual may influence others. This phenomenon, known as interference, often complicates the analysis. For example, [Bond et al., 2012] describes an experiment that was conducted using Facebook, a

social network website. On the day of the 2010 US midterm Congressional elections, participants received a banner advertisement on Facebook which encouraged them to vote, with the option to self-report that they had voted by clicking on an “I voted” button. This advertisement was customized for each recipient, so that it displayed the total number of users who had already viewed the advertisement and clicked “I voted”; for a random subset, the advertisement also displayed the profile pictures of up to six of the recipient’s Facebook friends who had already self-reported. The self-reported voting rate for the treatment group (those receiving profile pictures) was 2.08% higher than for the other participants, a difference large enough to reject a sharp null of zero effect. Since the content of the advertisement for each viewer depended on the actions of previous viewers, the presence of peer effects was ensured by the experiment design. Additionally, participants may have influenced each other through conversations caused by viewing the advertisement. Due to this interference, rigorous estimates of the effect size do not necessarily follow from rejection of the sharp null.

We propose a new approach for these types of experiments, which is based on an identifying assumption that the treatment effect is monotone. This is slightly weaker than requiring the treatment to not have negative effects, either directly or indirectly, on the outcome of any unit. Aside from this assumption, the interference will be allowed to take arbitrary and unknown form. Specifically, we do not assume partial interference or a correctly specified model of social influence.

The outline of the paper is as follows. Section 2 surveys related works. The basic problem formulation is given in Section 3. Three methods for interval estimation are presented in Section 4. These methods are demonstrated using data and simulation examples in Section 5. Section 6 discusses practical issues and future directions. Further technical details of the methods are presented in the appendices.

## 2 Related Work

Early discussion of interference in the potential outcomes framework is attributed to [Rubin, 1990, Halloran and Struchiner, 1995]. Current methods can be broadly divided between those which use a distribution-free rank statistic, and those which add identifying assumptions.

Distribution-free rank statistics are considered in [Rosenbaum, 2007, Luo et al., 2012]. In this approach, no assumptions are made on the interference, so that the estimates are highly robust. However, estimation is limited to rank-based quantities, i.e., on whether the treatment caused an overall shift in the ranks of the treated population when ordering the units by outcome. For non-rank quantities of interest, such as the average outcome under a counterfactual treatment, it appears that additional assumptions are required.

The most common identifying assumption is that the units form groups (such as households or villages) that do not interfere with each other; this is termed partial interference [Sobel, 2006]. The paper [Hudgens and Halloran, 2008] derives unbiased point estimates under partial interference, and variance bounds on the estimation error under a stronger condition termed stratified interference. Asymptotically normal estimates are given in [Liu and Hudgens, 2013], again assuming stratified interference, and finite sample error bounds are derived in [Tchetgen and VanderWeele, 2012]. For settings where partial interference does not apply, more general exposure models have been investigated by [Toulis and Kao, 2013, Ugander et al., 2013, Aronow and Samii, 2012], [Ogburn and VanderWeele, 2014, Manski, 2013], with rigorous results if one can model all of the network dynamics, such as who influences whom. As a result, they may not be suitable when the underlying social mechanisms are not well understood.

Two recent papers study inference methods that do not involve confidence intervals. The paper [Eckles et al., 2014] studies biased point estimation of treatment effects under weaker assumptions than partial or fully modeled interference, and is similar in spirit to this present work. The recent paper [Athey et al., 2015] (and also the related paper [Aronow, 2012]) gives exact tests for sharp

null hypotheses beyond the generic null of no effect and no interference.

### 3 Setup and notation

Let  $N$  denote the number of units in the experiment. Let treatments be assigned by sampling  $L$  units without replacement, and let  $X = (X_1, \dots, X_N)$  encode the treatment assignment, where  $X_i = 1$  if the  $i$ th unit was selected for treatment and  $X_i = 0$  otherwise. Let  $Y = (Y_1, \dots, Y_N)$  denote the observed outcomes, and let  $\theta = (\theta_1, \dots, \theta_N)$  denote the counterfactual outcomes under “full control”, i.e., if none of the units had received treatment and  $X_i = 0$  for all  $i$ .

As previously mentioned, we do not require an assumption of partial interference to hold. Instead, we require the following assumption on the treatment effect:

**Assumption 1** (Monotonicity).  $\theta_i \leq Y_i$ , for all  $i = 1, \dots, N$ .

This assumption might not be appropriate for some applications; for example, police interventions might displace crime, so that crime rates would decrease in some areas but increase in others. On the other hand, a vaccination program via “herd immunity” might have a strictly beneficial effect on the risk of infection.

Let  $A$  denote the attributable effect of the treatment, defined to be the total difference between  $Y$  and  $\theta$ :

$$A = \sum_{i=1}^N (Y_i - \theta_i). \quad (1)$$

Our definition for  $A$  generalizes that of [Rosenbaum, 2001] to allow for interference; if no interference is present, the two definitions are equivalent. Our inferential goal is a one-sided confidence interval lower bounding  $A$ . If this lower bound on  $A$  is large, it implies that the observed treatment had a large effect on the outcomes.

For each pair of units  $i$  and  $j$ , let  $d_{ij}$  denote an observed distance between them. For example,  $d_{ij}$  could represent geographic distance between units, or the shortest-path distance on a network

of observed pre-treatment social interactions. In Section 4.3, we will allow  $d_{ij}$  to be used as a crude proxy for the actual social dynamics underlying the experiment. Unlike previous work,  $d_{ij}$  will not be used as a model for the data generating process, but rather to customize a test statistic so as to increase its power to detect spillover effects. Our motivation is robustness to model error. Even if the choice of test statistic is based on erroneous beliefs, inversion of the statistic produces a confidence interval with the correct coverage probabilities, so that any significant findings will still be valid.

## 4 Constructing a Confidence Interval for $A$

In this section, we present three methods for estimating one-sided confidence intervals that upper bound  $\sum_i \theta_i$ , which by (1) is equivalent to a lower bound on the attributable effect  $A$ . In Section 4.1, a t-test based asymptotic confidence interval is presented for count-valued outcomes, i.e., when  $\theta$  and  $Y$  are nonnegative integers. In Section 4.2, a non-asymptotic estimate is presented for the special case of binary outcomes, which is then extended in Section 4.3 to utilize observed distances between the units.

### 4.1 T-test Based Asymptotic Confidence Interval

Suppose that the entries of  $\theta$  are actually observed for the  $N - L$  untreated units. Assuming that these units are sampled without replacement, it is well known [Thompson, 2012] that an unbiased point estimate for  $\bar{\theta} = N^{-1} \sum_i \theta_i$  is given by the sample average  $\hat{\theta}$ ,

$$\hat{\theta} = \frac{1}{N - L} \sum_{i: X_i=0} \theta_i.$$

Under certain conditions,  $\hat{\theta}$  is asymptotically normal and a  $(1 - \alpha)$  confidence upper bound for  $\bar{\theta}$  is asymptotically given by

$$\hat{\theta} + t_\alpha \sqrt{\left(\frac{L}{N}\right) \frac{\hat{\sigma}^2}{N - L}}, \quad (2)$$

where  $\hat{\sigma}^2$  is the estimated variance,

$$\hat{\sigma}^2 = \frac{1}{N - L - 1} \sum_{i: X_i=0} (\theta_i - \hat{\theta})^2,$$

and where  $t_\alpha$  is the  $\alpha$ -critical value of a  $t$  distribution with  $N - L - 1$  degrees of freedom.

In our setting,  $\theta$  is not actually observed, and hence (2) cannot be evaluated. Let us assume that Assumption 1 holds, and also that  $\theta$  is restricted to the set of nonnegative integers, so that  $0 \leq \theta \leq Y$  and  $\theta \in \mathbb{Z}^N$ . Then an upper bound to the unknown value of (2) can be found by solving the following optimization problem:

$$\begin{aligned} \max_{\theta \in \mathbb{Z}^N} \quad & \hat{\theta} + t_\alpha \sqrt{\left(\frac{L}{N}\right) \frac{\hat{\sigma}^2}{N - L}} \\ \text{such that} \quad & 0 \leq \theta_i \leq Y_i \text{ for all } i, \end{aligned} \tag{3}$$

which equals the highest value of (2) over all possible values of  $\theta$ . A polynomial-time solution method for this optimization problem is described in Appendix A.

**Example 1.** *It may seem counterintuitive that (3) may be maximized by  $\theta$  smaller than  $Y$ . To illustrate that this may be possible, let  $L = 20$ ,  $N = 25$ , and let the entries of  $Y$  equal (10, 10, 10, 11, 11) for the untreated units. Using (2) while letting  $\theta = Y$  gives a 95% upper bound of 10.9. On the other hand, letting  $\theta$  equal (0, 10, 10, 11, 11) for the untreated units gives an upper bound of 12.4, achieving the optimal value of (3).*

**Example 2.** *To illustrate that anti-conservative intervals may occur if (3) is not used, let  $X_i = 0$  for  $i = 1, \dots, 5$ , and let  $\theta_1, \dots, \theta_5$  and  $Y_1, \dots, Y_5$  be i.i.d. distributed as*

$$\theta_i = \begin{cases} 0 & \text{w.p. } 1/3 \\ 10 & \text{w.p. } 1/3 \\ 22 & \text{w.p. } 1/3 \end{cases} \quad \text{and} \quad Y_i = \begin{cases} 10 & \text{if } \theta_i = 0 \\ \theta_i & \text{otherwise.} \end{cases}$$

*This approximates our setting for large  $N$  and  $L/N \approx 1$ . In 10,000 simulations, using (2) with  $\alpha = 0.95$  while substituting  $Y$  as a proxy for the unobserved  $\theta$  resulted in a confidence interval that included  $\bar{\theta}$  only 87% of the time. Using (2) with the correct values for  $\theta$  resulted in 93% coverage, and using (3) resulted in 100% coverage.*

As with any t-test, by using (3) we are implicitly assuming that  $\hat{\theta}$  satisfies a central limit theorem. Equivalently, we may instead state that one of two alternatives must be true: either (3) gives a correct confidence interval, or the  $\alpha$ -quantile of  $\hat{\theta}$  (after studentization) is greater than  $t_\alpha$ , which for large  $N - L$  and  $L$  roughly equates to  $\theta$  having heavy tails.<sup>1</sup>

We remark that bootstrapping the untreated entries in  $Y$  will not compute a confidence interval for  $\hat{\theta}$ , since in general  $\theta \neq Y$ . However, the bootstrap may be still useful as a diagnostic check, testing whether (2) is valid for the point hypothesis  $\theta = Y$ .

## 4.2 Non-asymptotic Confidence Interval for Binary Outcomes

For binary-valued outcomes, a non-asymptotic one-sided confidence interval for  $\sum_i \theta_i$  can be computed. This can be done by a process known as “inverting a test statistic”<sup>2</sup>. Let  $W(X; \theta)$  denote a test statistic of  $X$  that is parameterized by the unknown  $\theta$ . Let  $w_\alpha(\theta)$  denote the  $\alpha$ -quantile of  $W(X; \theta)$ , defined by

$$\mathbb{P}(W(X; \theta) \leq w_\alpha(\theta)) = \alpha. \quad (4)$$

While  $\theta$  is unknown, we know two constraints on its value. First, we know that  $\theta \leq Y$ , by Assumption 1. Second, we know that  $W(X; \theta) \leq w_\alpha(\theta)$  with probability  $\alpha$ , by (4). Hence, to upper bound  $\sum_i \theta_i$  with probability  $\alpha$ , we can find the  $\theta$  which maximizes  $\sum_i \theta_i$  while satisfying these

<sup>1</sup>for example, [Bloznelis, 1999, Th. 1.1] implies that  $(N^{-1} \sum_i |\theta_i^3|) \cdot (N^{-1} \sum_i (\theta_i - \bar{\theta})^2)^{-3/2}$  must be large.

<sup>2</sup>We remark that inverting a test statistic to produce a confidence interval can potentially result in unstable behavior when the underlying assumptions are violated [Gelman, 2011]. While we do not recommend our methods when Assumption 1 is violated, they do not suffer from this behavior. This is because (5) will always have at least one feasible solution,  $\theta = 0$ .



constraints. That is, we can solve the optimization problem

$$\begin{aligned} \max_{\theta \in \{0,1\}^N} \sum_{i=1}^N \theta_i \\ \text{such that } W(X; \theta) \leq w_\alpha(\theta) \\ \theta_i \leq Y_i \text{ for all } i. \end{aligned} \tag{5}$$

It can be seen that (5) includes all non-rejected hypotheses, thus finding a one-sided confidence interval for  $\sum_i \theta_i$ .

We will use the test statistic  $W_{\text{basic}}$ , defined as

$$W_{\text{basic}}(X; \theta) = \sum_{i=1}^N X_i \theta_i.$$

It can be seen that  $W_{\text{basic}}(X; \theta)$  is generated by sampling  $L$  entries from  $\theta$  without replacement, so that  $W_{\text{basic}}(X; \theta)$  is a Hypergeometric( $\sum_i \theta_i, N - \sum_i \theta_i, L$ ) random variable. As a result, the optimization problem (5) is easily computable for  $W = W_{\text{basic}}$ , and we describe a solution method in Appendix B. This method was originally presented in [Rosenbaum, 2001, Appendix], but for the case of no interference.

**Weaker Assumption** We present a weaker assumption than Assumption 1, which may be applicable when the treatment effect is not strictly nonnegative:

**Assumption 2** (Aggregate Monotonicity for the Untreated).  $\sum_{i: X_i=0} \theta_i \leq \sum_{i: X_i=0} Y_i$ .

Unlike Assumption 1, which requires the treatment effect to be nonnegative for every individual, Assumption 2 only restricts the sum of the treatment effect over those units which did not receive treatment.

To upper bound  $\sum_i \theta_i$  under Assumption 2, we can solve a modification of (5),

$$\begin{aligned} \max_{\theta \in \{0,1\}^N} \sum_{i=1}^N \theta_i \\ \text{such that } W(X; \theta) \leq w_\alpha(\theta) \\ \sum_{i: X_i=0} \theta_i \leq \sum_{i: X_i=0} Y_i, \end{aligned} \quad (6)$$

where we have replaced the constraint  $\theta \leq Y$  by Assumption 2. Details of the solution method for  $W = W_{\text{basic}}$  are given in Appendix B.

### 4.3 Using observed distances between units

We extend the approach of Section 4.2 to handle a new statistic  $W_{\text{spill}}$ , which utilizes the observed distances  $\{d_{ij}\}_{i,j=1}^N$ . This statistic will have power to detect treatment effects that spill over from treated units to their untreated neighbors.

Let  $W_{\text{spill}}$  be given by

$$\begin{aligned} W_{\text{spill}}(X; \theta) &= \frac{1}{L} W_{\text{basic}}(\tilde{X}; \theta) \\ &= \frac{1}{L} \sum_{i=1}^N \tilde{X}_i \theta_i, \end{aligned}$$

where  $\tilde{X}$  is a smoothed version of  $X$ , so that each entry in  $\tilde{X}$  is a weighted average of nearby entries in  $X$ . More precisely, let  $\tilde{X}$  equal

$$\tilde{X} = X^T K,$$

where the smoothing matrix  $K \in \mathbb{R}^{N \times N}$  is given by

$$K_{ij} = \begin{cases} \frac{1}{Z_j} \exp(-d_{ij}^2 / \sigma_K^2) & \text{if } d_{ij} \leq d_{\max, K} \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

where  $d_{ij}$  denotes the distance between units  $i$  and  $j$ , as discussed in Section 3; where  $d_{\max,K} \geq 0, \sigma_K > 0$  are shape parameters; and where  $Z_j$  denotes a normalizing constant

$$Z_j = \sum_{i: d_{ij} \leq d_{\max,K}} \exp(-d_{ij}^2 / \sigma_K^2),$$

chosen so that the columns sum to one, making each element of  $\tilde{X}$  a weighted average of elements in  $X$ .

Because each entry of  $\tilde{X}$  is a weighted average, units that are close to treated units will have high values in  $\tilde{X}$ , even if they are not treated themselves. This will give  $W_{\text{spill}}$  power to detect spillovers. However, unlike  $W_{\text{basic}}$ , exact solution of (5) is not computationally feasible for  $W = W_{\text{spill}}$ . In Appendix C, (25) gives a relaxation of (5) that can be efficiently solved when the outcomes are binary-valued, yielding an asymptotically conservative estimate of  $A$  under Assumption 1.

## 5 Data and Simulation Examples

In this section, we present data and simulation examples to demonstrate the performance of the methods described in the previous section. In Section 5.1, the estimator (3) is used to analyze a primary school deworming experiment presented in [Miguel and Kremer, 2004]. In Section 5.2, the Facebook election experiment of [Bond et al., 2012] is analyzed using the test statistic  $W_{\text{basic}}$ . In Section 5.3, simulated experiments are used to evaluate the performance of the test statistic  $W_{\text{spill}}$ .

### 5.1 Analysis of [Miguel and Kremer, 2004]

[Miguel and Kremer, 2004] describes a primary school deworming project that was carried out in 1998 in Busia, Kenya, in order to reduce the number of infections by parasitic worms in young children. We restrict analysis to  $N = 50$  schools in a high infection area of Busia, which were divided into 2 equal-sized groups. Schools in group 1 received free deworming treatments beginning

in 1998, while group 2 did not. Students were surveyed in 1999, and substantially fewer infections were found in the treatment-eligible pupils, with 141 and 506 infections in groups 1 and 2 respectively. It is believed that the number of infections in each schools was affected not only by its own treatment status, but also that of other schools as well. This is because students that received the deworming treatment were susceptible to re-infection by infected students.

To demonstrate the estimator given by (3) on this experiment, we will assume that treatment was assigned by sampling without replacement<sup>3</sup>, and that all missing values in the data are ignorable. We also assume that the deworming treatment never increases the risk of infection, either to its direct recipient or to others. Under these assumptions, we solve a variant of (3) as discussed in Appendix A. The resulting confidence interval is that that with 95% confidence, the number of infections that would have occurred if all schools received deworming lies in the interval  $[0, 347]$ , and the number of infections that would have occurred if no schools received deworming lies in the interval  $[829, \infty)$ .

For purposes of comparison, a naive t-test (i.e., assuming that  $\theta = Y$ ) gives a confidence interval of  $[205, 347]$  for the number of infections if all schools received deworming. Since this interval assumes SUTVA, the desired coverage probability may not hold under Assumption 1. We note that the upper bound of 347 matches that given by (3), so only the lower bound of 205 is suspect in this instance. In general, either bound of the t-test could be suspect; as suggested by Example 2, it is possible for the upper bound of a naive t-test to also lack the correct coverage probability.

Since no spatial information is used by (3), the method does not have power to detect spillover effects. Section 6 discusses preliminary work towards incorporating spatial information to detect spillovers for this dataset, which may result in less conservative confidence intervals.

---

<sup>3</sup>Groups 1, 2, and 3 (with group 3 excluded from the 1999 survey) were actually assigned by dividing the schools into administrative subunits, listing them in alphabetical order, and assigning every third school to the same group.

## 5.2 Election Day Facebook Experiment

Using the reported counts for each treatment/outcome combination for the Facebook experiment of [Bond et al., 2012], we may estimate the attributable effect  $A$  by solving (5) or (6) for  $W = W_{\text{basic}}$ . In both cases, the resulting 95% confidence interval for  $A$  equals  $[1200426, \infty)$ , implying that the usage of profile pictures caused at least 1,200,426 users to click “I voted”, when they would not have done so otherwise. This is 2.00% of the treated population, so that our lower bound on  $A$  is approximately equal to the observed voting rate of 2.08%.

As the solutions to (5) and (6) are the same, our estimate of  $A$  is valid under either Assumption 1 or Assumption 2. Possibly, some individuals may have been discouraged from voting by seeing the profile picture of a Facebook friend (for example, perhaps due to a negative relationship), which would violate Assumption 1. Assumption 2 allows for this possibility, since no restrictions are made on the effects of treatment on the treated.

For comparison, an exact binomial test assuming no interference returns an estimate of  $[1200575, 1297686]$  for  $A$ . Although slightly less conservative, the lower bound is very similar that of (5) or (6). Although Example 2 demonstrates that coverage can be violated if interference is ignored, an interesting direction for future work might be to understand when such violations are most likely to occur in practice.

## 5.3 Simulated Study

In settings where spillover effects are large, the statistic  $W_{\text{spill}}$  may outperform  $W_{\text{basic}}$  by identifying clusters of outcomes that were caused by the treatment. To demonstrate this behavior, we ran simulations in which treatments resulted in higher probabilities of positive outcomes not only for the treated units, but also for those nearby as well. We explored a range of scenarios, varying the number of treatments and their spatial separation, the spillover radius of the treatment effect, the counterfactual  $\sum_i \theta_i$ , and also the choice of kernel matrix  $K$ . We found that estimates using  $W_{\text{spill}}$

were most accurate and robust to choice of  $K$  when the treatments resulted in many well-separated clusters of positive outcomes; in particular, increasing the number of treatments or their potency could actually decrease accuracy, by causing clusters to lose separation and run into each other.

**Description of Simulated Experiments** In each simulation,  $N$  units were placed on a uniformly spaced  $\sqrt{N} \times \sqrt{N}$  grid. Sampling with replacement was used to select units  $j_1, \dots, j_L$  for treatment, and auxiliary binary variables  $Z_1, \dots, Z_L$  were generated with distribution Bernoulli(1/2). For  $i = 1, \dots, N$ , each counterfactual outcome  $\theta_i$  was a Bernoulli( $p_0$ ) random variable, and each observed outcome  $Y_i$  equaled 1 if  $\theta_i = 1$ , and otherwise equaled a Bernoulli( $P_i$ ) random variable, where the probability  $P_i$  of having outcome  $Y_i = 1$  due to treatment was given by

$$P_i = 1 - \prod_{\ell=1}^L (1 - h(i, j_\ell))^{Z_\ell}, \quad (8)$$

where  $h$  denotes a truncated gaussian,

$$h(i, j) = \begin{cases} 0 & \text{if } d_{ij} > d_{\max, h} \\ \min(1, C \exp\{-d_{ij}^2 / \sigma_h^2\}) & \text{otherwise,} \end{cases} \quad (9)$$

where  $d_{ij}$  denotes distance (in sup norm) between units  $i$  and  $j$  on the grid, and where  $d_{\max, h}$ ,  $C$ , and  $\sigma_h$  are shape parameters. In words, (8)-(9) imply that each treatment  $\ell$  has no effect if  $Z_\ell = 0$ , and otherwise has an area effect that is independent of other treatments, i.e., each treatment  $\ell$  for which  $Z_\ell = 1$  has probability  $h(i, j_\ell)$  of independently causing unit  $i$  to have outcome  $Y_i = 1$ .

For each experiment, interval estimation using  $W_{\text{spill}}$  was computed by solving (25), which is a relaxation of (5) as discussed in Appendix C. In all simulations where the spillover effects were large, we note that  $W_{\text{basic}}$  and (3) gave nearly vacuous estimates, since they cannot detect spillovers.

**Simulation Results** Figure 1a shows estimation performance as a function of the generative  $h$  and the assumed kernel  $K$ . To construct this figure, 7 different choices for  $h$  were used, in which  $\sigma_h$  and  $C$  were adjusted so that the degree of localization of the treatment effect was varied while  $A$  was kept constant in expectation. These choices for  $h$  are shown in Figure 1b, with examples of the simulated outcomes shown in Figure 2. The assumed kernel  $K$  was varied by ranging the bandwidth parameter  $\sigma_K$  used in (7) from  $\sigma_h/3$  to  $6\sigma_h$ . In all cases, performance eventually decreased for large  $\sigma_K$ , suggesting that the choice of  $K$  should reflect knowledge about the anticipated treatment effect. For localized effects (i.e., small  $\sigma_h$ ), the estimates were more accurate, and allowed for the bandwidth of  $K$  to be chosen many times larger than  $\sigma_h$ . For diffuse effects (i.e., large  $\sigma_h$ ), estimates were highly conservative and more sensitive to the choice of  $K$ . These results suggest that estimation using  $W_{\text{spill}}$  may require spatial separation between treated units, so that the effects can be localized to their source.

Figure 3 shows average estimation performance as a function of the number of treatments  $L$ , and also their spatial density  $L/N$ , which was controlled by varying the grid size  $N$ . Increasing the number of treatments while also increasing  $N$  to keeping the spatial density constant (“outfill asymptotics”) improved accuracy. On the other hand, increasing spatial density while keeping  $L$  constant worsened the accuracy, due to reduced spatial separation between clusters of outcomes. As a result, increasing  $L$  while keeping  $N$  fixed (i.e., “infill asymptotics”) could either increase or decrease accuracy, since both  $L$  and  $L/N$  are changing simultaneously. Figure 4 shows examples of the simulated outcomes.

Figure 5 shows the average estimation performance in cases where  $d_{\max,h}$  ranged from 0 (no spillovers) to a distance of 2. The estimated lower bound on  $A$  was produced either by inverting  $W_{\text{basic}}$ , or by inverting  $W_{\text{spill}}$  with  $d_{\max,K}$  ranging from 0 to 5; the parameter  $d_{\max,K}$  can be interpreted as an assumption on the maximum distance between a treated unit and its spillover. When no spillovers were present, estimation was most accurate using  $W_{\text{basic}}$ ; on average, the estimated lower bound on  $A$  was 93% of the true value. Estimation using  $W_{\text{spill}}$  with no spillovers was less accurate,

ranging from 63% of the true value when  $d_{\max,K} = 0$  to the trivial lower bound of zero when  $d_{\max,K} = 2$ . When spillovers were present, estimation using  $W_{\text{basic}}$  was degraded; this is to be expected, given that  $W_{\text{basic}}$  can only detect direct effects, which comprise only a portion of the total effects. For these cases, estimation was most accurate when  $d_{\max,K}$  was matched to  $d_{\max,h}$ , with better performance when  $d_{\max,K}$  was chosen too large rather than too small. These results suggest that caution should be exercised when  $d_{\max,h}$  is small, and also that careful experiment design may be necessary in such settings. In Section 6, we discuss a possible design mechanism allowing treatments with small values of  $d_{\max,h}$  to result in larger clustered effects, such as those shown in Figure 2.

As a general remark, we observed that the coverage rates for the estimated 95% one-sided confidence intervals were conservatively high. The highest frequency of violated confidence intervals was 3%, which occurred when  $L = 10, L/N = 0.04$ . Over all of the simulations, only 0.1% of them resulted in a confidence interval which did not cover the true value of  $A$ .

## 6 Discussion

**Applicability of  $W_{\text{spill}}$**  The simulations of Section 5.3 are stylized, and are mainly meant to show that it is possible to use information about the social mechanism to improve estimates, without encoding this information as a formal assumption on the generative model. An approach of this type may be attractive whenever an observed network is believed to be only a crude proxy for truth, since it avoids anti-conservative estimates so long as Assumption 1 holds.

However, the results also suggest that inverting  $W_{\text{spill}}$  may require the following in practice:

1. The treatments should result in a large number of well-separated clusters of outcomes. If spillovers are non-existent or very small,  $W_{\text{basic}}$  should be used instead.
2. The kernel smoothing matrix  $K$  should be at least somewhat matched to the form of the



spillovers.

Given these requirements, and also given that  $W_{\text{spill}}$  was not demonstrated on a real data set, we believe  $W_{\text{spill}}$  should be regarded more as proof-of-concept, rather than recommended practice.

How practical are these requirements? We would not expect the effects of a single physical treatment, such as a coupon or advertisement, to resemble the simulations, in which an average of 12.5 outcomes were caused per treatment. However, the condition  $X_i = 1$  need not represent a single physical treatment. Instead, it could mean administering the physical treatment to a subset of units in the vicinity of  $i$ . For example, the condition  $X_i = 1$  could signify that some percentage of all units within some distance to  $i$  (or belonging to the same region as  $i$ ) receive the physical treatment. If the bandwidth of  $K$  is chosen with the treatment vicinities in mind, it may be possible to design experiments in which the outcomes tend to be approximately clustered according to  $K$ .

Assumption 1 allows for a good deal of flexibility in the allocation of the physical treatment. For example, if a unit belonged to multiple vicinities that were selected for treatment, the experiment protocol could give the unit a higher probability of receiving the physical treatment, or limit the unit to the same probability as those units in a single treatment vicinity, or even disqualify the unit from treatment altogether, as all three design options are allowed under Assumption 1.

**General Usage** A one-sided confidence interval for  $A$ , if it is not vacuously conservative, may help in determining whether an experimental treatment had a practically significant effect. In returning only a lower bound on  $A$ , we are taking a conservative approach to the possibility of errors in the network or spatial information encoded by  $\{d_{ij}\}_{i,j=1}^N$ .

Alternatively, one might consider testing the hypothesis that  $A = \sum_i (\theta_i - Y_i)$  equals zero. However, under Assumption 1,  $A$  can equal zero only if  $\theta = Y$ , meaning that the treatment must have zero effect on each individual unit. As a definition of “no effect”, this is far more restrictive than the hypothesis of zero average treatment effect, which allows for individual outcomes to change under treatment so long as the totals remain the same. For this reason, we recommend that sig-

nificance tests should not assume Assumption 1. When interference is present, a better choice for significance testing might be to use the rank-based methods of [Rosenbaum, 2007].

While we have focused on estimation of the attributable effect  $A$ , our methods can sometimes also be applied to estimate a version of the average treatment effect, which we define as follows. Let  $\theta^{\text{ft}}$  denote the counterfactual outcomes under full treatment, i.e., the outcome if all units were treated and  $X_i = 1$  for all  $i$ . Let  $\theta^{\text{fc}} \equiv \theta$  denote the counterfactual under full control. One definition for the average treatment effect is

$$ATE = \frac{1}{N} \sum_{i=1}^N (\theta_i^{\text{ft}} - \theta_i^{\text{fc}}),$$

which is the difference in outcomes between full treatment and full control, averaged over all units. As an example, in Section 5.1 (and with further details in Appendix A), we report an upper bound on  $\sum_i \theta_i^{\text{ft}}$  and a lower bound on  $\sum_i \theta_i^{\text{fc}}$  using (3) for the data of [Miguel and Kremer, 2004], thus inducing a lower bound of 482 on the average treatment effect.

In principle,  $W_{\text{basic}}$  and  $W_{\text{spill}}$  can also be adapted to estimate  $\theta^{\text{ft}}$  instead of  $\theta^{\text{fc}}$ . For binary outcomes, it can be seen that solving (5) for  $W_{\text{basic}}$  with  $1 - X$  in place of  $X$  and  $1 - Y$  in place of  $Y$  is equivalent to estimating a upper bound on  $1 - \sum_i \theta_i^{\text{ft}}$ , which gives a lower bound on  $\sum_i \theta_i^{\text{ft}}$ . While (25) for  $W_{\text{spill}}$  can also be solved with  $X$  and  $Y$  transformed in the same manner, the runtime will be prohibitively large if  $\sum_i (1 - Y_i) \gg \sum_i Y_i$ , as was the case in the simulations. As a result, the performance of the relaxation (25) under this transformation has not been investigated.

**Future directions and further analysis of [Miguel and Kremer, 2004]** As a possible direction for future work, we are investigating how the method of (3) might be applied to the “effective treatment” estimator discussed in [Eckles et al., 2014, Sec. 2.4.3]. This estimator, also discussed in [Aronow and Samii, 2012], was shown in [Eckles et al., 2014, Thm 2.2] to reduce bias under Assumption 1, but currently requires a correctly specified exposure model to compute a confidence interval. As this is a very strong assumption, a conservative estimate similar to (3) may be of interest.

We describe a special case of this estimator for which (3) can be seen to apply, in the context of the deworming experiment of [Miguel and Kremer, 2004]. We grouped 48 of the 50 schools into 16 triplets by order of distance, i.e., the closest three schools were grouped together, then the closest three out of the remaining schools, and so forth. The final 2 schools were removed from the analysis. We declared that a group of schools was treated if at least 2 schools in the group were treated (i.e., if they received the deworming treatment). The treated schools in the treated groups were declared to be selected. In this manner, 18 schools belonging to 8 treated groups were selected. Conditioned on the number of treated groups, and the number of selected schools in each group, the distribution of the 18 selected schools equals a two-stage sample [Thompson, 2012], in which the treated groups are selected by sampling without replacement, and then the selected schools are sampled within the treated groups. It follows by arguments similar to Section 4.1 that the average number of observed infections for the 18 selected schools is a conservatively biased point estimate for the per-school infections under full treatment. This value equaled 3.8, implying an point estimate of 182 for the total number of infections under full treatment. This is a 33% reduction from the point estimate of 270 that would result from an assumption of no interference, i.e., if all 24 treated schools were averaged.

To compute a confidence interval, in principle the method of (3) can be applied to the selected schools, using the estimated variance of a two stage sample in place of  $\hat{\sigma}$ . This resulted in a 95% confidence upper bound of 297 on the number of infections under full treatment, which is less than the estimate of 347 found in Section 5.1. While the small sample size of 8 groups likely invalidates the central limit theorem requirements of (3), the result suggests that at least that the proposed approach will not be vacuously conservative, and may be applicable in a larger experiment, such as [Bond et al., 2012]. Also, we observe that the point estimate is reminiscent of a U-statistic, since it can be written as a function of all  $\binom{N}{3}$  school triplets and their respective treatments. This suggests further possibilities for new estimators.

In this preliminary analysis, the spatial information in [Miguel and Kremer, 2004] was used

to remove treated schools from consideration if they were far from other treated schools. This improved the point estimate because such schools were more susceptible to reinfection. This is quite different from the simulations, where well-separated treatments gave the best estimates. We conjecture that both types of settings can arise in practice.

## Acknowledgements

The author would like to thank the reviewers for their feedback and suggestions which greatly improved the paper. This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.

## Appendices

### A T-test Based Asymptotic Confidence Interval

**Solution of (3)** It can be seen that the objective function of (3) is a function of  $\hat{\theta}$  and  $\hat{\sigma}^2$ , and is increasing in the latter argument. Hence, the optimal  $\theta$  will maximize  $\hat{\sigma}^2$  over some level set of  $\hat{\theta}$ , which is equivalent to solving

$$\begin{aligned} \max_{\theta \in \mathbb{Z}^N} \quad & \sum_{i: X_i=0} \theta_i^2 \\ \text{such that} \quad & \sum_{i: X_i=0} \theta_i = c \\ & 0 \leq \theta_i \leq Y_i \text{ for all } i, \end{aligned} \tag{10}$$

for some value of  $c$ . Since  $c$  must be an integer between 0 and  $\sum_{i: X_i=0} Y_i$ , we can solve (10) for all possible values of  $c$ , and then choose the solution that maximizes (3).

To solve (10), let  $n = N - L$  and let  $i_1, \dots, i_n$  sort the elements of  $\{Y_i : X_i = 0\}$  in descending order. It can be seen that (10) is maximized by letting  $\theta_{i_1} = \min\{c, Y_{i_1}\}$ , and following the recursion

$$\theta_{i_j} = \min\left\{c - \sum_{k=1}^{j-1} \theta_{i_k}, Y_{i_j}\right\}, \quad j = 2, \dots, n, \quad (11)$$

so that the entries of  $\theta$  corresponding to the untreated units are “filled up” in decreasing order of  $Y$ , i.e.,  $\theta_{i_j} = 0$  unless  $\theta_{i_k} = Y_{i_k}$  for  $k = 1, \dots, j - 1$ .

**Variant of (3) used in [Miguel and Kremer, 2004]** To estimate the number of infections that would occur if all of the schools were treated, we define  $Y, X$ , and  $\theta$  as follows. Let  $Y_i$  denote the number of infections observed in school  $i$ . Reversing the definition of  $X$ , let  $X_i = 0$  denotes that school  $i$  receives the deworming treatment. Let  $\theta$  denote the counterfactual outcomes that would occur if  $X_i = 0$  for all  $i$ . With  $Y, X$ , and  $\theta$  thus defined, Assumption 1, which states that  $\theta \leq Y$ , means that treating all of the schools would not increase the infection counts over the observed values. A 95% confidence upper bound on  $\bar{\theta}$  can be found by solving (3).

To estimate the number of infections that would occur if none of the schools were treated, let  $Y$  be defined as before; let  $X_i = 1$  denote that school  $i$  receives deworming treatment; and let  $\theta$  denote the counterfactual outcome that would occur if no schools receive treatment. In place of Assumption 1, we assume that  $\theta_i \geq Y_i$ , meaning that treating no schools would not reduce the infection counts below the observed values, and also that  $\theta_i \leq S_i$ , where  $S_i$  is the total number of students at school  $i$  that were measured in the 1999 survey. By similar reasoning as (3), in order to lower bound  $\bar{\theta}$  we can solve

$$\begin{aligned} \min_{\theta \in \mathbb{Z}^N} \quad & \hat{\theta} - t_\alpha \sqrt{\left(\frac{L}{N}\right) \frac{\hat{\sigma}^2}{N - L}} \\ \text{such that} \quad & Y_i \leq \theta_i \leq S_i \text{ for all } i, \end{aligned} \quad (12)$$

where  $\hat{\theta}$  and  $\hat{\sigma}^2$  are defined as before. Similar to (3), the optimal  $\theta$  must maximize  $\hat{\sigma}^2$  along a level

set of  $\hat{\theta}$ , so that

$$\begin{aligned} & \max_{\theta \in \mathbb{Z}^N} \sum_{i: X_i=0} \theta_i^2 \\ & \text{such that } \sum_{i: X_i=0} \theta_i = c \\ & Y_i \leq \theta_i \leq S_i \text{ for all } i \end{aligned} \tag{13}$$

can be solved for different values of  $c$  to find the optimal  $\theta$ .

The optimization problem (13) can be formulated and solved as a dynamic programming problem. Generically, a simplified version of a dynamic program involves choosing a sequence of discrete decision variables  $u_1, \dots, u_T$ , so as to control a sequence of state variables  $s_0, \dots, s_T$ , where the initial state  $s_0$  is given and  $s_t = f_t(s_{t-1}, u_t)$  for  $t = 1, \dots, T$  and some set of functions  $f_1, \dots, f_T$  which model the state dynamics. A reward  $g(u_t)$  is paid for each decision, and a final reward  $G(s_T)$  is paid based on the final state. The goal is to choose  $u_1, \dots, u_T$  to maximize  $G(s_T) + \sum_t g(u_t)$ , thereby steering towards a high reward final state while also maintaining high rewards for each decision. A canonical algorithm to solve this problem is value iteration [Bertsekas et al., 1995], which is also called backwards induction or Bellman's equation.

To formulate (13) as a dynamic programming problem, let  $T = n$  and let the decisions  $u_1, \dots, u_T$  equal  $\theta_1, \dots, \theta_n$ . Let  $g(u_t) = u_t^2$ , so that  $\sum_t g(u_t)$  equals the objective of (13). Let  $s_0 = 0$ , and let  $s_t = s_{t-1} + u_t$ , so that  $s_T = \sum_t u_t$ , which equals  $\sum_{i: X_i=0} \theta_i$ . Let the final reward  $G(s_T)$  equal 0 if  $s_T = c$ , and  $-\infty$  otherwise, thus enforcing the constraint that  $\sum_{i: X_i=0} \theta_i = c$ .

## B Estimation Using $W_{\text{basic}}$

**Solution of (5) for  $W_{\text{basic}}$**  For  $W = W_{\text{basic}}$ , the  $\alpha$ -level critical value of  $W$  is a function of  $\sum_i \theta_i$ , since  $W$  is a Hypergeometric( $\sum_i \theta_i, N - \sum_i \theta_i, L$ ) random variable. Let  $w_\alpha(\sum_i \theta_i)$  denote the  $\alpha$ -level

critical value of  $W$ . It follows that (5) can be rewritten as

$$\begin{aligned} & \max_{\theta \in \{0,1\}^N} \sum_{i=1}^N \theta_i \\ \text{such that } & \sum_{i:X_i=1} \theta_i \leq w_\alpha \left( \sum_{i=1}^N \theta_i \right) \end{aligned} \quad (14)$$

$$\sum_{i:X_i=1} \theta_i \leq \sum_{i:X_i=1} Y_i \quad (15)$$

$$\sum_{i:X_i=0} \theta_i \leq \sum_{i:X_i=0} Y_i, \quad (16)$$

where (15) and (16) are consequences of  $\theta \leq Y$ . This optimization problem depends only the quantities  $\sum_{i:X_i=1} \theta_i$  and  $\sum_{i:X_i=0} \theta_i$ . As these quantities are integer valued and bounded above and below, their optimal values can be easily found by exhaustive search.

**Solution of (6) for  $W_{\text{basic}}$**  For  $W = W_{\text{basic}}$ , the optimization problem (6) can be rewritten as above, but with constraint (15) removed. This removes the upper bound on  $\sum_{i:X_i=1} \theta_i$ . However, since  $\sum_{i:X_i=1} \theta_i \leq \sum_i X_i$ , an upper bound still exists, so the optimal solution may be found by exhaustive search as before.

## C Estimation Using $W_{\text{spill}}$

For  $W = W_{\text{spill}}$ , the solution of of the optimization problem (5) is computationally hard. We present a conservative approximation of (5) that yields a larger confidence interval for  $A$ . The main steps of the approximation are to bound the critical value  $w_\alpha(\theta)$  using a simpler expression, and to enclose the feasible region of (5) by linear inequalities.

**Preliminaries** We will require the following basic identities. It can be seen that  $W_{\text{spill}}(X; \theta)$  equals the average of  $L$  samples drawn without replacement from the vector  $K\theta$ . Because the columns of

$K$  sum to one, it holds that

$$\mathbb{E}W_{\text{spill}}(X; \theta) = \frac{1}{N} \sum_{i=1}^N \theta_i, \quad (17)$$

where we note that the expectation  $\mathbb{E} \equiv \mathbb{E}_X$  is taken over the random treatment  $X$ .

Let  $u$  denote a unit sampled uniformly from  $1, \dots, N$ . Let  $1_u \in \{0, 1\}^N$  denote the indicator function returning 1 for unit  $u$  and 0 elsewhere. It follows that  $W_{\text{spill}}(1_u; \theta)$  is equal in distribution to  $W_{\text{spill}}(X; \theta)$  for  $L = 1$ . For all  $L$ , it holds that

$$\mathbb{E}W_{\text{spill}}(X; \theta) = \mathbb{E}_u W_{\text{spill}}(1_u; \theta) \quad (18)$$

$$\text{Var } W_{\text{spill}}(X; \theta) = \frac{N-L}{L(N-1)} \left( \mathbb{E}_u [W_{\text{spill}}(1_u; \theta)^2] - [\mathbb{E}_u W_{\text{spill}}(1_u; \theta)]^2 \right), \quad (19)$$

where (19) follows from basic properties of simple random sampling [Thompson, 2012, Eq. 2.5].

**Approximation of (5)** By Chebychev's inequality, it holds for any choice of  $W$  that

$$\mathbb{P} \left( \frac{W(X; \theta) - \mathbb{E}W(X; \theta)}{(\text{Var } W(X; \theta))^{1/2}} \geq \alpha^{-1/2} \right) \leq \alpha. \quad (20)$$

This is a highly conservative bound, but we use it here for simplicity and defer improvements for later discussion. Analogous to (5), a one-sided  $(1 - \alpha)$  confidence interval for  $\sum_i \theta_i$  is given by

$$\begin{aligned} & \max_{\theta \in \{0,1\}^N} \frac{1}{N} \sum_{i=1}^N \theta_i \\ & \text{such that } \frac{W(X; \theta) - \mathbb{E}W(X; \theta)}{(\text{Var } W(X; \theta))^{1/2}} \leq \alpha^{-1/2} \\ & \theta_i \leq Y_i \text{ for all } i. \end{aligned} \quad (21)$$

To rewrite this problem with a smaller number of decision variables, let  $m(y) \in \mathbb{R}^3$  denote the vector given by

$$m_1(\theta) = \mathbb{E}_u W(1_u; \theta), \quad m_2(\theta) = W(X; \theta), \quad \text{and} \quad m_3(\theta) = \mathbb{E} [W(1_u; \theta)^2].$$



Let  $\mathcal{M} = \{m(\theta) : \theta \leq Y\}$  denote the set of all achievable values for  $m(\theta)$ . Equating terms and using (17)-(19), the optimization problem (21) can be restated as

$$\begin{aligned} & \max_{m \in \mathbb{R}^3} m_1 \\ & \text{such that } \frac{m_2 - m_1}{(m_3 - m_1^2)^{1/2}} \leq \left( \alpha L \frac{N-1}{N-L} \right)^{-1/2} \\ & m \in \mathcal{M}. \end{aligned} \tag{22}$$

While this optimization problem has only 3 decision variables, it is hard to optimize because the constraint  $m \in \mathcal{M}$  is difficult to check. As a relaxation, we will replace the constraint  $m \in \mathcal{M}$  by a weaker constraint  $m \in \mathcal{P}$ , where  $\mathcal{P}$  is a polyhedron that contains  $\mathcal{M}$ , and which can be represented by a tractable number of linear inequalities. Let  $f^*(\lambda)$  denote the maximum inner product between  $\lambda \in \mathbb{R}^3$  and  $m(\theta) \in \mathcal{M}$ :

$$f^*(\lambda) = \max_{\theta \in \{0,1\}^N} \lambda^T m(\theta) \quad \text{such that } \theta \leq Y.$$

Given a set  $\Lambda \subset \mathbb{R}^3$ , let  $\mathcal{P}_\Lambda$  denote the set  $\{m : \lambda^T m \leq f^*(\lambda) \text{ for all } \lambda \in \Lambda\}$ . Since  $\lambda^T m \leq f^*(\lambda)$  for all  $m \in \mathcal{M}$ , it follows that  $\mathcal{P}_\Lambda$  contains  $\mathcal{M}$ . Hence the following optimization problem upper bounds (22), yielding a conservative confidence interval:

$$\begin{aligned} & \max_{m \in \mathbb{R}^3} m_1 \\ & \text{such that } \frac{m_2 - m_1}{(m_3 - m_1^2)^{1/2}} \leq \left( \alpha L \frac{N-1}{N-L} \right)^{-1/2} \\ & \lambda^T m \leq f^*(\lambda), \quad \forall \lambda \in \Lambda. \end{aligned} \tag{23}$$

This optimization problem is low dimensional. As a result, it can be practically solved by a grid-based search over the feasible region, provided that  $f^*(\lambda)$  is known for all  $\lambda \in \Lambda$ .

**Computation of  $f^*(\lambda)$**  To solve (23), we must compute  $f^*(\lambda)$  for all  $\lambda \in \Lambda$ . For  $W = W_{\text{spill}}$ , it holds by the following identities,

$$\mathbb{E}_u W_{\text{spill}}(1_u; \theta) = \frac{\mathbf{1}^T K \theta}{N}, \quad \mathbb{E}_u [W_{\text{spill}}(1_u; \theta)^2] = \frac{\theta^T K^T K \theta}{N}, \quad \text{and} \quad W(X; \theta) = \frac{X^T K \theta}{L},$$

that we may write  $f^*(\lambda)$  as

$$f^*(\lambda) = \max_{\theta \in \{0,1\}^N} \lambda_1 \frac{\mathbf{1}^T K \theta}{N} + \lambda_2 \frac{X^T K \theta}{L} + \lambda_3 \frac{\theta^T K^T K \theta}{N}, \quad (24)$$

such that  $\theta_i \leq Y_i$  for all  $i$ .

For nonnegative  $K$  and  $\lambda_3$ , (24) can be transformed into a canonical optimization problem of finding an “ $s$ - $t$  min cut” in a graph. The transformation, described in Appendix D, was originally proposed in [Greig et al., 1989] for image denoising. After the transformation, the min cut problem can be solved by linear programming or the Ford-Fulkerson algorithm, which runs in  $O(n^3)$  time where  $n = \sum_i Y_i$ . [Papadimitriou and Steiglitz, 1998]

**Selection of  $\Lambda$**  Figure 6 gives a geometric picture of the role of  $\mathcal{M}$  and  $\mathcal{P}_\Lambda$  in determining the feasible region of (23). The set  $\Lambda$  must satisfy  $\lambda_3 \geq 0$  for all  $\lambda \in \Lambda$ , since  $f^*(\lambda)$  cannot be efficiently computed otherwise. By definition, each half-space  $H_\lambda = \{m : \lambda^T m \leq f^*(\lambda)\}$  equals a supporting hyperplane of the set  $\mathcal{M}$  in the direction  $\frac{\lambda}{\|\lambda\|}$ . This implies that  $H_\lambda = H_{c\lambda}$  when  $c$  is a positive scalar. As a result, a reasonable strategy is to choose  $\Lambda$  to cover the allowable directions  $\{\lambda : \|\lambda\| = 1, \lambda_3 \geq 0\}$  as densely as possible, so that  $\mathcal{P}_\Lambda$  approximates the convex hull of  $\mathcal{M}$  in those directions.

**Reducing conservativeness** Chebychev’s inequality gives a very conservative approximation to the critical value of the test statistic. Because  $W_{\text{spill}}(X; \theta)$  is a sample average, a normal approxi-

mation may yield a better estimate of its critical value. That is, it may hold that

$$\mathbb{P} \left( \frac{W_{\text{spill}}(X; \theta) - \mathbb{E} W_{\text{spill}}(X; \theta)}{(\text{Var } W_{\text{spill}}(X; \theta))^{1/2}} \geq z_\alpha \right) \approx \alpha,$$

where  $z_\alpha$  is the upper critical value of a standard normal. Using this approximation leads to the following optimization problem

$$\begin{aligned} & \max_{m \in \mathbb{R}^3} m_1 \\ & \text{such that } \frac{m_2 - m_1}{(m_3 - m_1^2)^{1/2}} \leq z_\alpha L^{-1/2} \\ & \lambda^T m \leq f^*(\lambda), \forall \lambda \in \Lambda. \end{aligned} \tag{25}$$

**Summary of method** Given binary observations  $Y$ , treatment assignment  $X$ , and distances  $\{d_{ij}\}_{i,j=1}^N$ , the method entails the following steps:

1. Choose a smoothing matrix  $K$ , for example by choosing values of  $d_{\max, K}$  and  $\sigma_K$ .
2. Choose a set  $\Lambda \subset \mathbb{R}^3$  such that  $\lambda_3 \geq 0$  for all  $\lambda \in \Lambda$ . This will ultimately induce the set  $\mathcal{P}$  which relaxes the actual feasible region.
3. For each  $\lambda \in \Lambda$ , compute  $f^*(\lambda)$  by solving (24). The solution of (24) is discussed in Appendix D.
4. Solve (23) or (25) to the desired level of precision. This is done by discretizing the feasible region of (23) or (25) along a grid, and checking every grid point. Because the objective is linear and the feasible region is 3-dimensional, the number of grid points that must be checked increases cubically with the desired precision. The best solution is an upper bound on  $\sum_i \theta_i$ , up to the precision of the grid search.

## D Transformation of $f^*(\lambda)$ to min-cut problem

Given a nonnegative matrix  $A \in \mathbb{R}^{d \times d}$  with zero diagonal, and  $s, t \in 1, \dots, d$ , the s-t min cut problem is

$$\min_{x \in \{0,1\}^d} \sum_{i \neq j} A_{ij} x_i (1 - x_j) \quad (26)$$

such that  $x_s = 1, x_t = 0$ .

The interpretation of (26) is that  $A$  denotes a weighted adjacency matrix of a network, and  $x$  divides the nodes  $1, \dots, d$  into two groups, with  $s$  and  $t$  in separate groups, so as to minimize the sum of the weighted edges that are “cut” by the division. This problem is polynomially solvable by the Ford-Fulkerson algorithm and also by linear programming [Papadimitriou and Steiglitz, 1998].

To transform  $f^*(\lambda)$  into the form of (26), we observe that

$$f^*(\lambda) = \max_{\theta \in \{0,1\}^N} \lambda_1 \frac{\mathbf{1}^T K \theta}{N} + \lambda_2 \frac{X^T K \theta}{L} + \lambda_3 \frac{\theta^T K^T K \theta}{N},$$

such that  $\theta_i \leq Y_i$  for all  $i$ ,

may be rewritten as

$$\max_{x \in \{0,1\}^d} x^T M x + b^T x + c,$$

for some  $d > 0$ ,  $b \in \mathbb{R}^d$ ,  $c \in \mathbb{R}$ , and nonnegative matrix  $M$ , where the decision variable  $x$  corresponds to the free elements in  $y$ , i.e., those in  $\{i : Y_i = 1\}$ . Following [Greig et al., 1989], we

transform this to a min-cut problem by observing that

$$\begin{aligned} x^T M x + b^T x &= - \sum_{i,j} (M_{ij} x_i (1 - x_j) - M_{ij} x_i) + \sum_i b_i x_i \\ &= - \sum_{i \neq j} M_{ij} x_i (1 - x_j) + \sum_i x_i \left( b_i + \sum_j M_{ij} \right). \end{aligned} \quad (27)$$

Let  $\gamma_i = b_i + \sum_j M_{ij}$ . Then maximizing (27) is equivalent to

$$\max_{x \in \{0,1\}^d} - \sum_{i \neq j} M_{ij} x_i (1 - x_j) - \sum_{i: \gamma_i \geq 0} |\gamma_i| (1 - x_i) + \sum_{i: \gamma_i < 0} |\gamma_i| x_i. \quad (28)$$

Let  $s = d + 1$ ,  $t = d + 2$ , and let  $x_s = 1$ ,  $x_t = 0$ . We can rewrite (28) as

$$\max_{x \in \{0,1\}^d} - \sum_{i \neq j} M_{ij} x_i (1 - x_j) - \sum_{i: \gamma_i \geq 0} |\gamma_i| (1 - x_i) x_s + \sum_{i: \gamma_i < 0} |\gamma_i| x_i (1 - x_t),$$

which can be rewritten as (26) for some nonnegative  $A \in \mathbb{R}^{d+2 \times d+2}$  with zero diagonal.

## E Non-integer and non-binary outcomes

**Using (3) with non-integer outcomes** In cases where  $\theta$  is bounded by  $0 \leq \theta \leq Y$ , but also takes continuous values, an approximate CI can be found by solving (3) with the constraint  $\theta \in \mathbb{Z}^N$  replaced by  $\theta \in \mathbb{D}^N$ , where  $\mathbb{D}$  is the set of all multiples of  $1/c$  for some integer  $c > 1$ , i.e.,  $\{i/c : i \in \mathbb{Z}\}$ . The solution can be found using dynamic programming as before (but with run-time increased by a factor of  $c$ ).

**Using  $W_{\text{spill}}$  with non-binary outcomes** Evaluation of  $f^*(\lambda)$  can be transformed to a min cut optimization problem only if its decision variables  $y$  are binary. To approximately handle non-binary outcomes where  $y_i$  is bounded below by  $a_{i \min}$ , we can constrain each decision variable  $y_i$  to

be a weighted sum of  $B$  binary decision variables  $u_{i1}, \dots, u_{iB}$  yielding a binary expansion:

$$y_i = a_{i\min} + \sum_{b=1}^B c_{ib} u_{ib}, \quad i = 1, \dots, N,$$

where  $c_{ib} = 2^{-b}(Y_i - a_{i\min})$  for  $b = 1, \dots, B$ . Then (24) can be rewritten as a optimization problem over binary decision variables  $\{u_{ib}\}_{i=1, b=1}^{N, B}$ , and can be transformed to a min cut problem.

## References

- [Aral and Walker, 2011] Aral, S. and Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639.
- [Aronow, 2012] Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16.
- [Aronow and Samii, 2012] Aronow, P. M. and Samii, C. (2012). Estimating average causal effects under general interference. In *Summer Meeting of the Society for Political Methodology*, University of North Carolina, Chapel Hill, July, pages 19–21.
- [Athey et al., 2015] Athey, S., Eckles, D., and Imbens, G. W. (2015). Exact p-values for network interference. Technical report, National Bureau of Economic Research.
- [Bertsekas et al., 1995] Bertsekas, D. P., Bertsekas, D. P., Bertsekas, D. P., and Bertsekas, D. P. (1995). *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA.
- [Bloznelis, 1999] Bloznelis, M. (1999). A berry-esseen bound for finite population student's statistic. *Annals of probability*, pages 2089–2108.
- [Bond et al., 2012] Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298.

- [Eckles et al., 2014] Eckles, D., Karrer, B., and Ugander, J. (2014). Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530*.
- [Gelman, 2011] Gelman, A. (2011). Why it doesn't make sense in general to form confidence intervals by inverting hypothesis tests. [http://andrewgelman.com/2011/08/25/why\\_it\\_doesnt\\_m/](http://andrewgelman.com/2011/08/25/why_it_doesnt_m/). Accessed: 2015-10-02.
- [Greig et al., 1989] Greig, D., Porteous, B., and Seheult, A. H. (1989). Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 271–279.
- [Halloran and Struchiner, 1995] Halloran, M. E. and Struchiner, C. J. (1995). Causal inference in infectious diseases. *Epidemiology*, pages 142–151.
- [Hudgens and Halloran, 2008] Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482).
- [Kramer et al., 2014] Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.
- [Liu and Hudgens, 2013] Liu, L. and Hudgens, M. G. (2013). Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, (just-accepted).
- [Luo et al., 2012] Luo, X., Small, D. S., Li, C.-S. R., and Rosenbaum, P. R. (2012). Inference with interference between units in an fmri experiment of motor inhibition. *Journal of the American Statistical Association*, 107(498):530–541.
- [Manski, 2013] Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.

- [Miguel and Kremer, 2004] Miguel, E. and Kremer, M. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, pages 159–217.
- [Nickerson, 2008] Nickerson, D. W. (2008). Is voting contagious? evidence from two field experiments. *American Political Science Review*, 102(01):49–57.
- [Ogburn and VanderWeele, 2014] Ogburn, E. L. and VanderWeele, T. J. (2014). Vaccines, contagion, and social networks. *arXiv preprint arXiv:1403.1241*.
- [Papadimitriou and Steiglitz, 1998] Papadimitriou, C. H. and Steiglitz, K. (1998). *Combinatorial optimization: algorithms and complexity*. Courier Dover Publications.
- [Rosenbaum, 2001] Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88(1):219–231.
- [Rosenbaum, 2007] Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477).
- [Rubin, 1990] Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.
- [Sobel, 2006] Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407.
- [Sweet et al., 2013] Sweet, T. M., Thomas, A. C., and Junker, B. W. (2013). Hierarchical network models for education research hierarchical latent space models. *Journal of Educational and Behavioral Statistics*, 38(3):295–318.
- [Tchetgen and VanderWeele, 2012] Tchetgen, E. J. T. and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75.



[Thompson, 2012] Thompson, S. K. (2012). *Sampling*. Wiley.

[Toulis and Kao, 2013] Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects.  
In *Proceedings of The 30th International Conference on Machine Learning*, pages 1489–1497.

[Ugander et al., 2013] Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013).  
Graph cluster randomization: network exposure to multiple universes. *arXiv preprint*  
*arXiv:1305.6979*.

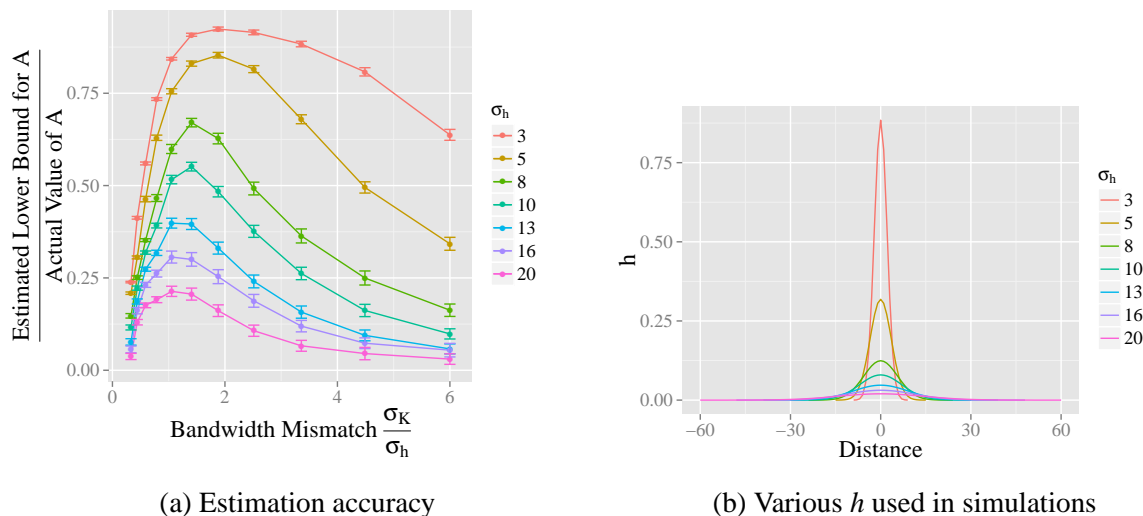


Figure 1: Average accuracy (and standard errors) of estimated lower bound for  $A$ , for various choices of spillover function  $h$  and mismatched smoothing matrix  $K$ . The spillover functions  $h$ , shown in (b), were chosen by varying the bandwidth  $\sigma_h$  while keeping  $A$  constant in expectation.  $K$  was chosen to have a mismatched bandwidth  $\sigma_K$  that was a multiple of the generative  $\sigma_h$ . 100 simulations per data point; examples of the simulations are shown in Fig. 2.

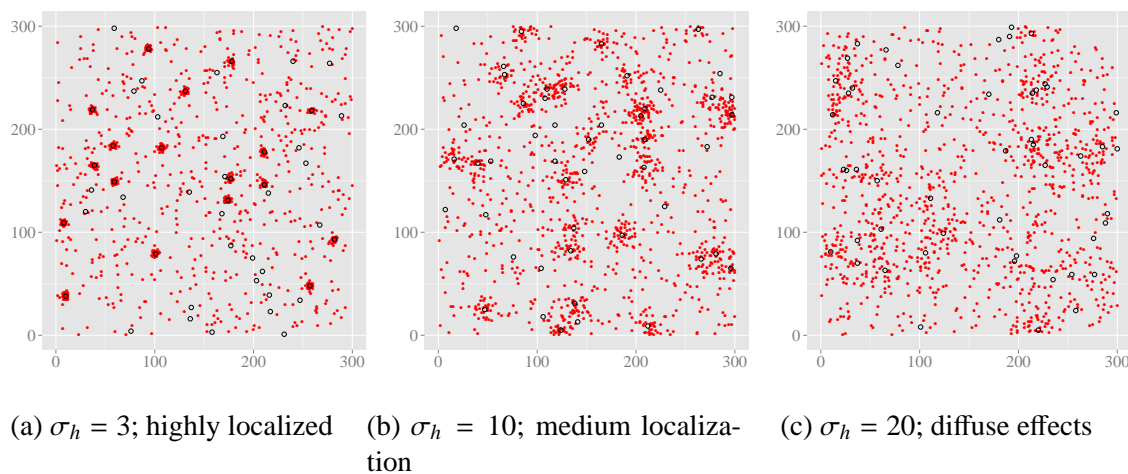


Figure 2: Examples of simulated experiments used to generate Fig. 1, in which the spillover function  $h$  was varied while the expectation of  $A$  was held constant.  $N = 90,000$  units were placed on a  $300 \times 300$  grid. Black circles denote treated units ( $L = 50$ ), red dots denote units with outcome 1. Treatment effects were large; on average, each treatment caused 12.5 outcomes, and  $\sum_i Y_i = 1225$  and  $\sum_i \theta_i = 600$  in expectation.

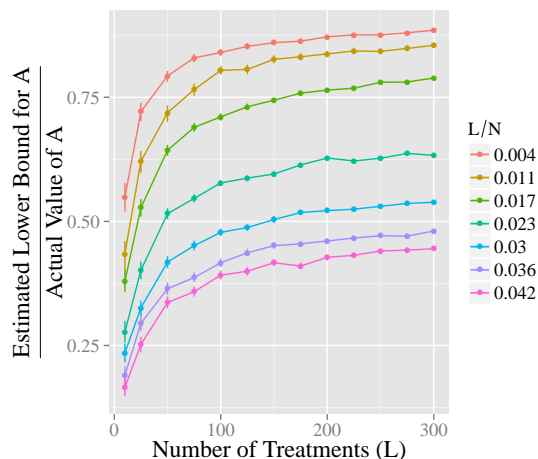


Figure 3: Average estimation accuracy (and standard errors) using  $W_{\text{spill}}$  with smoothing matrix  $K$  matched to the generative  $h$ , while varying the number of treatments  $L$  and their spatial density  $L/N$ . 400 simulations per data point; examples of the simulations are shown in Fig. 4

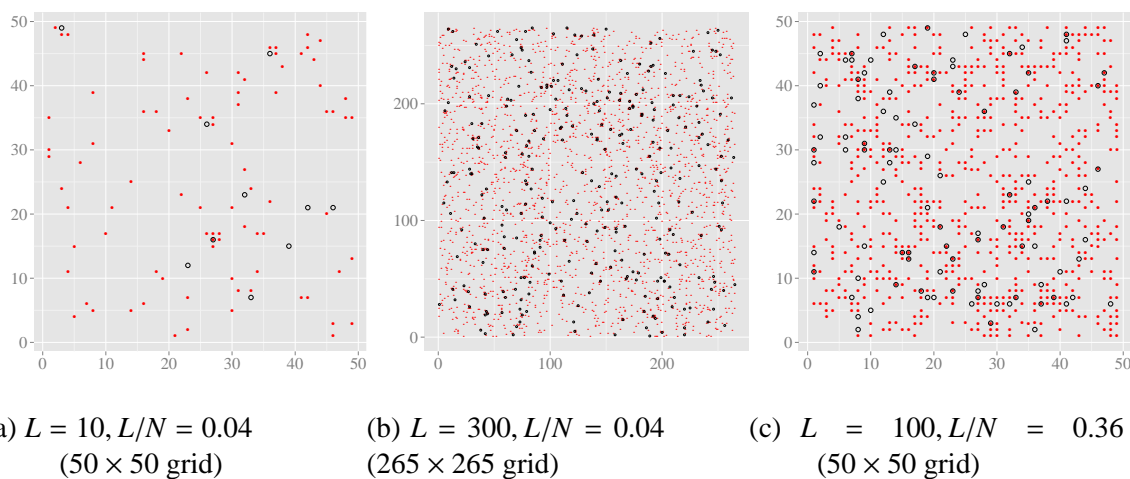


Figure 4: Examples of simulations used to generate Figure 3. (a) and (b) show low density treatments on small and large grids, while (c) shows high density treatments on a grid of equal size to (a). Estimation accuracy was best for (b), then (a), and worst for (c). Each treatment caused 1.5 outcomes on average, and  $\sum_i \theta_i / \sum_i Y_i = 0.7$  in expectation.

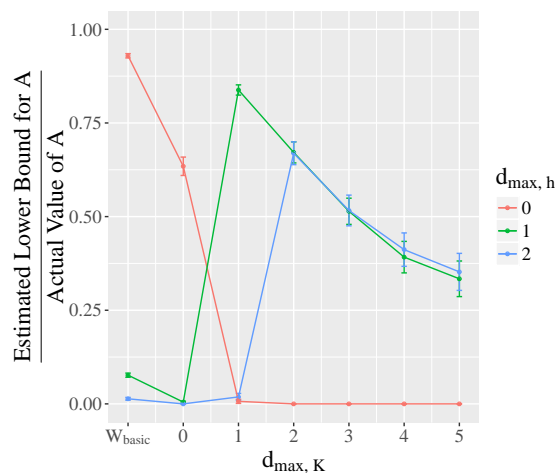


Figure 5: Estimation accuracy (average performance and standard errors) in spatial experiments in which the generative  $d_{\max, h}$  ranged from 0 (no spillovers) to 2. Estimation either used  $W_{\text{basic}}$ , or used  $W_{\text{spill}}$  with  $d_{\max, K}$  varied between 0 to 5. Experiments involved  $N = 90,000$  units placed on a  $300 \times 300$  grid, with  $\sum_i \theta_i = 600$  and  $L = 50$  treatments. Treatments caused either 0.5 outcomes on average without spillovers, or 1.5 outcomes on average if spillovers were present. 100 simulations per data point.

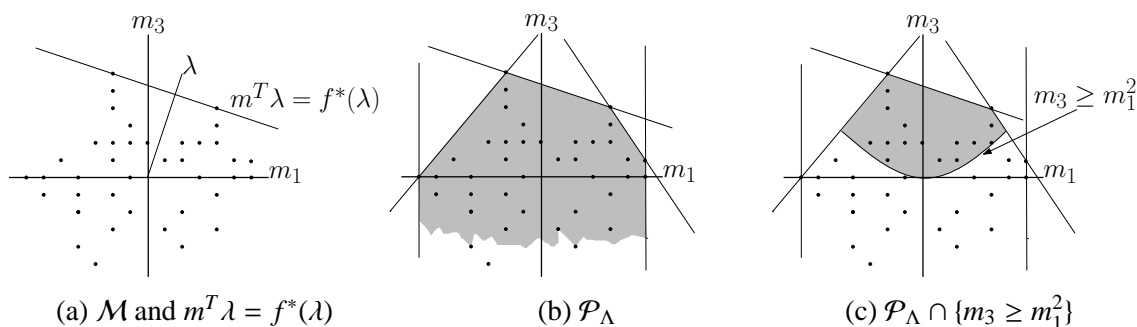


Figure 6: Cartoon depiction of (23), showing dimensions  $m_1$  and  $m_3$  only. (a) shows  $\mathcal{M}$  (as dots), and a supporting hyperplane in a direction  $\lambda$ . (b) shows  $\mathcal{P}_\Lambda$  (as shaded region), which may equal the convex hull of  $\mathcal{M}$  in all directions  $\lambda$  satisfying  $\lambda_3 \geq 0$ . (c) shows the intersection of  $\mathcal{P}_\Lambda$  and the constraint  $m_3 \geq m_1^2$ . This constraint is implicit in (23), since otherwise  $(m_3 - m_1^2)^{-1/2}$  would not be real-valued.