

Estimation of Monotone Treatment Effects in Network Experiments

David Choi

August 20, 2014

Abstract

Randomized experiments on social networks are a trending research topic. Such experiments pose statistical challenges due to the possibility of interference between units. We propose a new method for estimating attributable treatment effects under interference. The method does not require partial interference, but instead uses an identifying assumption that is similar to requiring nonnegative treatment effects. Pre-treatment network observations can be used to customize the test statistic, so as to increase power without making assumptions on the data generating process. The inversion of the test statistic is a combinatorial optimization problem which has a tractable relaxation, yielding conservative estimates of the attributable effect.

Keywords: causal inference, attributable effect, interference, randomized experiments, network data, Facebook, graph cut

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](#)], education [[Sweet et al., 2013](#)], and online sharing [[Kramer et al., 2014](#)].

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 instance, [Kramer et al., 2014] describes a recent experiment that was conducted using Facebook, a social network website. Facebook users often write posts, and read their “News Feed,” a listing of posts tailored to each user. For a random subset of users, posts were less made likely to appear in their News Feed if they contained words from a predetermined list¹. Over time, these users were observed to use those same words less frequently than other users (the “control group”) when writing new posts of their own, suggesting that they had been influenced by their News Feed. To complicate matters, however, the new posts could then appear in other News Feeds, potentially exposing control users. While Facebook could have blocked these posts from appearing in other News Feeds, interference would not be eliminated by doing so, as it is impossible to know what content – perhaps including content shared outside of Facebook – would have been created and viewed in the absence of treatment.

We propose a new estimation method for these types of experiments, requiring 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 approach will be applicable on network data exhibiting clusters of similar outcomes, potentially having been caused by contagion or by social propagation of the treatment effect. We illustrate such an effect with simulated experiments in a spatial setting and on a subset of the Facebook network.

The outline of the paper is as follows. Section 2 discusses related work. Section 3 gives the problem setup. Sections 4 and 5 describe the general approach and computational methodology. Section 6 describes simulated experiments for spatial and social network settings, and discusses connections to some recently published experiments. Extensions and technical details are contained in the appendix.

¹This created significant negative publicity for Facebook [Bailey, 2014]

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, Eckles et al., 2014, Aronow and Samii, 2012, Ogburn and VanderWeele, 2014, Manski, 2013]. These models generally require fine-grained knowledge of the network dynamics, i.e., who influences whom. As a result, they may not be suitable when the underlying social mechanisms are not well understood.

3 Setup and notation

Let N denote the number of units in the experiment. Let the randomization be performed by sampling L units with replacement, where a unit is assigned to treatment if it is sampled at

least once. Let $X = (X_1, \dots, X_N)$ denote a count-valued vector, where X_i denotes the number of times the i th unit was sampled for treatment. Let $Y = (Y_1, \dots, Y_N)$ denote the vector of observed outcomes. Let $Y^0 = (Y_1^0, \dots, Y_N^0)$ denote the vector of counterfactual outcomes if none of the units had received treatment.

We will require the following assumption on the treatment effect:

Assumption 1 (Monotonicity). $Y_i^0 \leq Y_i$, for $i = 1, \dots, N$.

This assumption might not be appropriate for some applications; for example, police interventions might displace crime, so that crime rates decrease in some areas but increase in others. On the other hand, a vaccination program 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 Y^0 :

$$A = \sum_{i=1}^N (Y_i - Y_i^0). \quad (1)$$

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.

Let G denote a network of observed pre-treatment social interactions between the units. This snapshot of observed interactions might be only a crude proxy for the actual social dynamics. Hence, we will not use G to make explicit assumptions on the influence between units. Instead, G will be used to choose a test statistic. Our motivation is robustness to model error. If G turns out to be a poor proxy, the method will lose power but not correctness, so that any significant findings will still be valid.

4 General Approach

We present a one-sided confidence interval to upper bound $\sum_i Y_i^0$, thus inducing a lower bound on the attributable effect $A = \sum_i (Y_i - Y_i^0)$. Our approach is to treat Y^0 as a fixed, unknown parameter, and then to induce a confidence set for this parameter by considering the set of all hypotheses for Y^0 that are not rejected by a test statistic. This is known as inverting a test statistic.

Let $W(X; Y^0)$ denote a test statistic of X that is parameterized by the unknown Y^0 . Let $w_\alpha(Y^0)$ denote the α -quantile of its distribution, defined by

$$\mathbb{P}(W(X; Y^0) \leq w_\alpha(Y^0)) = \alpha. \quad (2)$$

While Y^0 is unknown, we know two constraints on its value. First, we know that $Y^0 \leq Y$, by Assumption 1. Second, we know that $W(X; Y^0) \leq w_\alpha(Y^0)$ with probability α , by (2). Hence, to upper bound $\sum_i Y_i^0$ with probability α , we can find the largest vector satisfying these constraints. That is, we can solve the optimization problem

$$\begin{aligned} & \max_{y \in \mathbb{R}^N} \sum_{i=1}^N y_i & (3) \\ \text{such that } & W(X; y) \leq w_\alpha(y) \\ & y \leq Y. \end{aligned}$$

Thus (3) considers all non-rejected hypotheses to find a one-sided confidence interval for $\sum_i Y_i^0$.

Example 1 gives a test statistic W_{basic} that is adapted from [Rosenbaum, 2001], which takes large values when treatments are co-located with high outcome units. For this statistic, the optimization problem (3) is known to be easily computable if the outcomes are binary-valued. However, the statistic does not use the network observations G and does not identify clusters of outcomes. We will use it as a baseline to which a new statistic may be compared.

Example 1 (Binomial Distribution). Let W_{basic} equal

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

Suppose that $W = W_{\text{basic}}$ and consider (3) for the case of binary-valued outcomes. Since y is binary, the critical value $w_\alpha(y)$ is constant over $\{y : \sum_i y_i = C\}$, and equals that of a Binomial($L, C/N$) distribution. Let $w_\alpha(C)$ denote this value, and let $\nu = \sum_i Y_i$. The solution to (3) is given by $\nu - \delta$, for the minimum value of $\delta \geq 0$ satisfying $W(X; Y) - \delta \leq w_\alpha(\nu - \delta)$, as derived in [Rosenbaum, 2001, Appendix] with slight modification.

5 Using the observed network G

We define a new statistic W_{spill} that takes large values when treatments are co-located with clusters of high outcome units. As a result, W_{spill} will have increased power to detect treatment effects that spill over from the treated units to their neighbors in the observed network G . However, exact inversion of W_{spill} will not be computationally feasible, so we will instead solve a relaxation of (3) yielding a more conservative estimate.

5.1 A Clustering Test Statistic

Let W_{spill} be given by

$$W_{\text{spill}}(X; y) = \frac{1}{L} X^T K y, \tag{4}$$

where $K \in \mathbb{R}^{N \times N}$ has nonnegative elements and columns summing to one.

The matrix K is based on the observed network G . Let D_{ij} denote the distance between

units i and j in G . Given constants $d_{\max} \geq 0, c > 0$, let K equal

$$K_{ij} = \begin{cases} \frac{1}{Z_j} \exp(-D_{ij}/c) & \text{if } D_{ij} \leq d_{\max} \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where Z_j denotes a normalizing constant

$$Z_j = \sum_{i: D_{ij} \leq d_{\max}} \exp(-D_{ij}/c),$$

chosen so that the columns sum to one. For this choice of K , the statistic $W_{\text{spill}}(X; Y)$ will be large if treatments cause nearby units to have high outcome values.

Section 5.2 will require the following basic identities. Because the columns of K sum to one, it holds that

$$\mathbb{E}W_{\text{spill}}(X; y) = \frac{1}{N} \sum_{i=1}^N y_i,$$

so that $\mathbb{E}W_{\text{spill}}(X; Y^0)$ equals the average of Y^0 . It can be seen that $W_{\text{spill}}(X; y)$ equals the average of L samples drawn with replacement from the vector Ky . Let $W_{\text{spill}}(i; y)$ denote $W_{\text{spill}}(X; y)$ when $L = 1$. It holds that

$$\mathbb{E}W_{\text{spill}}(X; y) = \mathbb{E}W_{\text{spill}}(i; y) \quad (6)$$

$$\text{Var } W_{\text{spill}}(X; y) = \frac{1}{L} \left(\mathbb{E} [W_{\text{spill}}(i; y)^2] - [\mathbb{E}W_{\text{spill}}(i; y)]^2 \right). \quad (7)$$

5.2 Conservative Estimation of $\sum_i Y_i^0$

For $W = W_{\text{spill}}$, the solution of of the optimization problem (3) is computationally hard. We present a conservative approximation of (3) that yields a larger confidence interval for A . The main steps of the approximation are to bound the critical value $w_\alpha(y)$ using a simpler expression, and to enclose the feasible region of (3) by linear inequalities. In this section, we derive the

approximation assuming binary outcomes. Non-binary outcomes and improvements to the method are discussed in Section 5.3.

By Chebychev's inequality, it holds for any choice of W that

$$\mathbb{P} \left(\frac{W(X; Y^0) - \mathbb{E}W(X; Y^0)}{(\text{Var } W(X; Y^0))^{1/2}} \geq \alpha^{-1/2} \right) \leq \alpha. \quad (8)$$

This is a highly conservative bound, but we use it here for simplicity and defer improvements for later discussion. Analogous to (3), a one-sided $(1 - \alpha)$ confidence interval for $\mathbb{E}W(X; Y^0)$ is given by

$$\begin{aligned} & \max_{y \in \{0,1\}^N} \mathbb{E}W(X; y) & (9) \\ \text{such that } & \frac{W(X; y) - \mathbb{E}W(X; y)}{(\text{Var } W(X; y))^{1/2}} \leq \alpha^{-1/2} \\ & y \leq Y. \end{aligned}$$

Since $\mathbb{E}W_{\text{spill}}(X; Y^0)$ equals the average of Y^0 , solving (9) for $W = W_{\text{spill}}$ yields an upper bound for $\sum_i Y_i^0$. To rewrite this problem with a smaller number of decision variables, let $m(y) \in \mathbb{R}^3$ denote the vector given by

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

Let $\mathcal{M} = \{m(y) : y \text{ binary}, y \leq Y\}$ denote the set of all achievable values for $m(y)$. Equating terms and using (6)-(7), the optimization problem (9) can be restated as

$$\begin{aligned} & \max_{m \in \mathbb{R}^3} m_1 & (10) \\ \text{such that } & \frac{m_2 - m_1}{(m_3 - m_1^2)^{1/2}} \leq (\alpha L)^{-1/2} \\ & m \in \mathcal{M}. \end{aligned}$$

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 $m(y) \in \mathcal{M}$:

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

Given a set $\Lambda \subset \mathbb{R}^3$, let \mathcal{P} 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} contains \mathcal{M} . Hence the following optimization problem upper bounds (10), yielding a conservative confidence interval:

$$\begin{aligned} & \max_{m \in \mathbb{R}^3} m_1 & (11) \\ \text{such that } & \frac{m_2 - m_1}{(m_3 - m_1^2)^{1/2}} \leq (\alpha L)^{-1/2} \\ & \lambda^T m \leq f^*(\lambda), \quad \forall \lambda \in \Lambda. \end{aligned}$$

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 (11), we must compute $f^*(\lambda)$ for all $\lambda \in \Lambda$. For $W = W_{\text{spill}}$, it holds by (4) and the following identities,

$$\mathbb{E}W_{\text{spill}}(i; y) = \frac{\mathbf{1}^T K y}{N} \quad \text{and} \quad \mathbb{E}[W_{\text{spill}}(i; y)^2] = \frac{y^T K^T K y}{N},$$

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

$$f^*(\lambda) = \max_{y \in \{0,1\}^N} \lambda_1 \frac{\mathbf{1}^T K y}{N} + \lambda_2 \frac{X^T K y}{L} + \lambda_3 \frac{y^T K^T K y}{N}, \quad (12)$$

such that $y \leq Y$.

For nonnegative K and λ_3 , (12) can be transformed into a canonical optimization problem of finding an “ s - t min cut” in a graph. The transformation, described in Appendix A, 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 Λ A good choice of Λ will result in a small feasible region for (11). For each (m_1, m_2) , the constraint $m_2 - m_1 < (\alpha L)^{-1/2} (m_3 - m_1^2)^{1/2}$ is satisfied by large m_3 . To eliminate such values from the feasible region, a reasonable choice for Λ is to discretize the upper hemisphere $\{\lambda \in \mathbb{R}^3 : \|\lambda\|_2 = 1, \lambda_3 \geq 0\}$, so that for each (m_1, m_2) , the largest value of m_3 such that $(m_1, m_2, m_3) \in \mathcal{P}$ is controlled.

It may be practical to add constraints to Λ iteratively, yielding an improving sequence of upper bounds. Such algorithms are known as cutting-plane optimization methods. For such a method to be effective, a good heuristic would be needed for selecting new constraints. We leave this for future work.

Other choices of K Other choices are possible for K . For example, D might be based on node attributes found by fitting G to a latent space or mixed membership model [Airoldi et al., 2008, Hoff et al., 2002]. This may carry benefits by “denoising” G , without requiring strong assumptions on the network formation process. The network G could itself be chosen to reflect structural assumptions. For example, to encode partial interference one might chose G to be a collection of cliques. More generally, given any conjectured model θ in which binary

outcomes in Y are caused by specific elements in X , the matrix K may be chosen to equal $K_{ij} = \mathbb{P}_\theta(Y_i \text{ caused by } X_j | Y_i = 1)$.

5.3 Extensions

The following extensions to (11) are known. The technical details have been deferred to Appendix B, in order to simplify the discussion.

1. **Reducing conservativeness** Chebychev's inequality gives a very conservative approximation to the critical value of the test statistic. Because $W_{\text{spill}}(X; Y^0)$ is a sample average, a normal approximation may yield a better estimate of its critical value. That is, it may hold that

$$\mathbb{P} \left(\frac{W_{\text{spill}}(X; Y^0) - \mathbb{E}W_{\text{spill}}(X; Y^0)}{(\text{Var } W_{\text{spill}}(X; Y^0))^{1/2}} \geq z_\alpha \right) \approx \alpha,$$

where z_α is the upper critical value of a standard normal. Using this approximation leads to the following optimization problem

$$\begin{aligned} & \max_{m \in \mathbb{R}^3} m_1 & (13) \\ \text{such that } & \frac{m_2 - m_1}{(m_3 - m_1^2)^{1/2}} \leq z_\alpha L^{-1/2} \\ & \lambda^T m \leq f^*(\lambda), \quad \forall \lambda \in \Lambda, \end{aligned}$$

which may yield more accurate estimates of $\mathbb{E}W(X; Y^0)$.

In Appendix B, we give a Berry-Esseen bound on the error of the normal approximation, and show that this error term may be incorporated into the optimization problem of (13) if desired. A similar refinement is also possible using Bernstein's inequality.

2. **Non-binary outcomes** While (11) assumes binary valued outcomes, it is easily modified to accommodate bounded real-valued outcomes, as described in Appendix B.

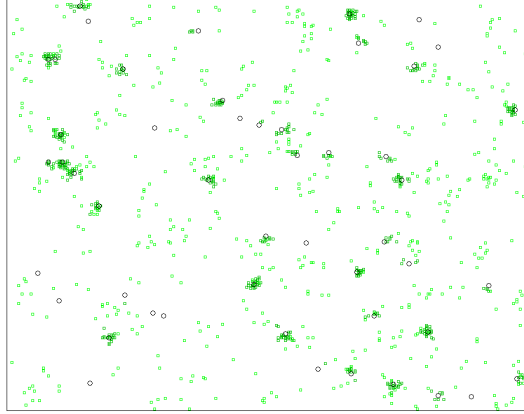


Figure 1: Simulated experiment with 90,000 units placed on a 300×300 grid. Markers show the locations of the 50 treated units (circles), and 1027 units with outcome one (shaded dots). The true value of A was 463; using (13), the estimated 95% confidence lower bound on A was 345.

6 Simulation Results

6.1 A Spatial Example

A simulated experiment with monotone interference was run involving $N = 90,000$ units, where the observed network was a 300×300 grid. $L = 50$ units were randomly sampled for treatment. Outcomes were binary valued, with 1027 units (1.1%) having outcome one. In Figure 1, we see that the treated units were sparse and well-separated, with many of them surrounded by clusters of outcome one caused by spillover effects.

The true value of A was 463, meaning that the observed treatment caused 463 units to change their outcome from 0 to 1. Table 1 shows one-sided 95% confidence intervals yielding a lower bound on A , computed either by inversion of W_{basic} , or by inversion of W_{spill} . The inversion of W_{spill} was computed using either a Chebychev inequality or normal approximation, as given by (11) or (13), with K given by (5) using $(c = 3, d_{\text{max}} = \infty)$ which matched the parameters of the simulation. This yielded the lower bounds $A \geq 307$ and $A \geq 345$. In comparison, inversion of

Test statistic	One-sided 95% interval on A
W_{basic} (Ex. 1)	$[5, \infty)$
W_{spill} w/ Chebychev inequality (11)	$[307, \infty)$
W_{spill} w/ normal approximation (13)	$[345, \infty)$
Actual value of A	463

Table 1: Results for simulated spatial experiment. Inversion of W_{basic} , which does not consider spillover effects, is compared to inversion of W_{spill} .

W_{basic} returned $A \geq 5$, which is much less informative. The reason for the poor performance of W_{basic} is that it does not consider spillovers.

To investigate robustness to the choice of K , the bandwidth parameter c was ranged from 0.5 to 75; at $c = 75$, the bandwidth is 150 which is a sizeable fraction of the entire 300×300 grid. The resulting lower bound estimates for A are shown in Figure 2. The results are significantly better than those attained using W_{basic} for a wide range of bandwidths. This may be due to the large spatial separation between clusters.

6.2 Facebook graph example

Motivated by large-scale experiments that were conducted by Facebook [Bond et al., 2012, Kramer et al., 2014], a second simulation was run using a Facebook dataset that is available at <http://snap.stanford.edu/data/egonets-Facebook.html>. The network is composed of 10 survey participants and their immediate Facebook friends, totaling $N = 4039$ units. $L = 100$ of the units were randomly sampled for treatment. Simulated outcomes were binary, with 745 units (18.4%) observed to have outcome one.

We found estimation to be more difficult than in the previous example. Every unit was connected to at least one of the survey participants, who in turn acted as hubs neighboring as much as 20% of the entire network. As a result, clustering of outcomes is much less apparent, since random walks on the network mix quickly to a stationary distribution. As the true Facebook network is private information, we do not know if our data is representative of experiments on

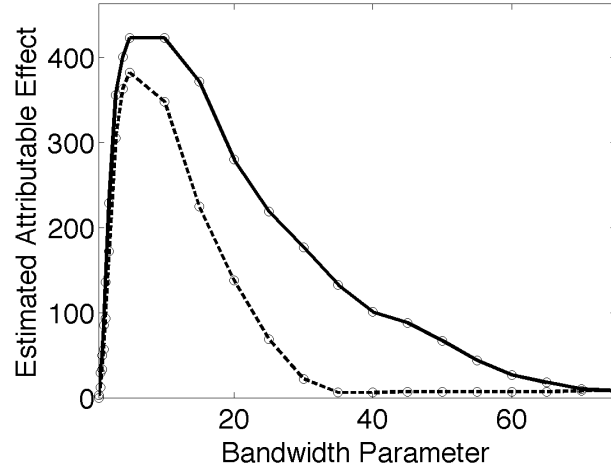


Figure 2: Estimation results in spatial example, with K chosen by varying the bandwidth parameter c from 0.5 to 75. Solid line: Normal approximation. Dotted line: Chebychev inequality. The true value of A was 463.

Facebook, or an artifact of our collection method.

Table 2 shows the true values and estimated confidence intervals for A , by either inversion of W_{basic} or W_{spill} . The matrix K is given by (5) with ($c = \infty, d_{\text{max}} = 1$), producing uniform kernels supported on the treated unit and its immediate neighbors. Inversion of W_{spill} leads to significantly improved results over W_{basic} , although less accurate than the estimates of Table 1.

Increasing d_{max} to 2 yielded poor results, comparable to inversion of W_{basic} . This is perhaps due to the fact that every node was connected to at least one hub, so that kernels with support extending beyond immediate neighbors no longer corresponded to localized clusters.

6.3 Discussion

In the experiments conducted by Facebook [Bond et al., 2012, Kramer et al., 2014], the p -value was computed for the null hypothesis that the treatment had no effect. In both experiments, the null hypothesis was rejected, implying that with high confidence, the treatment had a non-zero effect. If no interference was present, this p -value would also correspond to a confidence interval

Test statistic	One-sided 95% interval on A
W_{basic} (Ex. 1)	$[12, \infty)$
W_{spill} w/ Chebychev inequality (11)	$[38, \infty)$
W_{spill} w/ normal approximation (13)	$[161, \infty)$
Actual value of A	697

Table 2: Results for simulated Facebook graph experiment. Inversion of W_{basic} , which does not consider spillover effects, is compared to inversion of W_{spill} .

on the average treatment effect. This confidence interval was reported by Facebook as the effect size.

Given the possibility of interference, the reported effect sizes may have been inaccurate for both experiments. In settings where Assumption 1 is unreasonable, using [Rosenbaum, 2007] to estimate the effect on rank may be the best option. If one is willing to entertain Assumption 1, it may be of interest to estimate non-rank quantities by inverting either W_{basic} or W_{spill} . If clustering is exhibited, inverting W_{spill} may yield larger estimates for the effect size. This would be particularly relevant for [Kramer et al., 2014], where the reported effect size was small.

Inversion of W_{spill} offers a principled way of leveraging crude proxies to actual social mechanisms, such as Facebook or geographic data, as side information to increase the power of a test statistic to detect spillovers. In this manner, one could use online social network data to aid in the analysis of non-virtual experiments, where the treatment is not an alteration to a webpage, and where network elicitation would otherwise be difficult and expensive.

Appendices

A 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

$$\min_{x \in \{0,1\}^d} \sum_{i \neq j} A_{ij} x_i (1 - x_j) \quad (14)$$

$$\text{such that } x_s = 1, x_t = 0.$$

The interpretation of (14) 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 (14), we observe that

$$f^*(\lambda) = \max_{y \in \{0,1\}^N} \lambda_1 \frac{\mathbf{1}^T K y}{N} + \lambda_2 \frac{X^T K y}{L} + \lambda_3 \frac{y^T K^T K y}{N},$$
$$\text{such that } y \leq Y,$$

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

$$\begin{aligned} x^T Mx + b^T x &= - \sum_{i,j} (M_{ij}x_i(1-x_j) - M_{ij}x_i) + \sum_i b_i x_i \\ &= - \sum_{i \neq j} M_{ij}x_i(1-x_j) + \sum_i x_i \left(b_i + \sum_j M_{ij} \right). \end{aligned} \quad (15)$$

Let $\gamma_i = b_i + \sum_j M_{ij}$. Then maximizing (15) is equivalent to

$$\max_{x \in \{0,1\}^d} - \sum_{i \neq j} M_{ij}x_i(1-x_j) - \sum_{i:\gamma_i \geq 0} |\gamma_i|(1-x_i) + \sum_{i:\gamma_i < 0} |\gamma_i|x_i. \quad (16)$$

Let $s = d + 1$, $t = d + 2$, and let $x_s = 1$, $x_t = 0$. We can rewrite (16) 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 (14) for some nonnegative $A \in \mathbb{R}^{d+2 \times d+2}$ with zero diagonal.

B Extensions

As the extensions apply only to W_{spill} , we let $W = W_{\text{spill}}$ in this appendix so as to lighten the notation.

Normal approximation of critical value If z_α is used in (13), replacing the value $\alpha^{-1/2}$ given by Chebychev's inequality, the actual confidence level of the test will be $\tilde{\alpha}$, given by

$$\tilde{\alpha} = \mathbb{P} \left(\frac{W(X; Y^0) - \mathbb{E}W(X; Y^0)}{(\text{Var } W(X; Y^0))^{1/2}} \geq z_\alpha \right).$$

For binary outcomes, Lemma 1 gives a bound on $|\tilde{\alpha} - \alpha|$:

Lemma 1. *Let $\sum_{j=1}^n K_{ij} \leq b$ for all $i = 1, \dots, n$. If $0 \leq Y^0 \leq 1$, then*

$$|\tilde{\alpha} - \alpha| \leq 3 \left(\frac{b}{\mathbb{E}W(X; Y^0)L} \right)^{1/2}.$$

Proof. By a Berry-Esseen theorem [Durrett, 2010, Thm 3.4.9], it holds that

$$|\tilde{\alpha} - \alpha| \leq 3\gamma L^{-1/2},$$

where γ is the skewness of $W(i; Y^0)$. Under the assumptions of the lemma, $0 \leq KY^0 \leq b$. Maximum skewness is attained when KY^0 only takes values 0 and b . This implies $\gamma \leq b^{1/2}/\mathbb{E}W(i; Y^0)^{1/2}$, proving the lemma. \square

If desired, the result of Lemma 1 can be directly incorporated into the optimization problem, as an improved bound on the critical value. This is done by replacing the critical value $\alpha^{-1/2}$ not with z_α , but rather by the expression

$$\min \left(\alpha^{-1/2}, z_{\alpha-3(b/L)^{1/2}} \right),$$

As (11) is optimized by a grid-based search over the feasible region, no additional computational cost is incurred by this modification.

Bernstein approximation of critical value Assuming that $0 \leq KY^0 \leq b$, Bernstein's inequality implies

$$\mathbb{P}(W(X; y) - \mathbb{E}W(X; y) > t) \leq \exp \left(-\frac{Lt^2/2}{\text{Var } W(i; y) + bt/3} \right)$$

Substituting $W(X; y) - \mathbb{E}W(X; y)$ for t , this implies with probability $1 - \alpha$ that

$$\exp\left(-\frac{L(W(X; y) - \mathbb{E}W(X; y))^2/2}{\text{Var } W(X; y) + b(W(X; y) - \mathbb{E}W(X; y))/3}\right) \geq \alpha$$

If desired, we may incorporate this constraint into the optimization problem given by (11), as an improved critical value. This is done by replacing the constraint $(m_2 - m_1)/(m_3 - m_1^2)^{1/2} \leq (\alpha L)^{1/2}$ with

$$\exp\left(-\frac{L(m_2 - m_1)^2/2}{m_3 - m_1^2 + b(m_2 - m_1)/3}\right) \geq \alpha.$$

As (11) is optimized by a grid-based search over the feasible region, no additional computational cost is incurred by this modification.

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 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 (12) 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

[Airoldi et al., 2008] Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

- [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 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.
- [Bailey, 2014] Bailey, B. (2014). Facebook runs into uproar over experiments that tested emotional reactions. http://www.mercurynews.com/business/ci_26064438/facebook-runs-into-uproar.
- [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.
- [Durrett, 2010] Durrett, R. (2010). *Probability: theory and examples*. Cambridge university press.
- [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*.
- [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.
- [Hoff et al., 2002] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

- [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.
- [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.
- [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*.