Network exposure to multiple universes

Johan Ugander
Cornell University
jhu5@cornell.edu

Brian Karrer
Facebook
briankarrer@fb.com

Lars Backstrom
Facebook
lars@fb.com

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Abstract

A/B tests are a standard approach for evaluating the effect of a product or feature on a set of individuals – to estimate the so-called unit-level average treatment effect. Here we propose a methodology for computing the unit-level average treatment effect of A/B tests under social interference. We characterize formal conditions under which users are considered to be network exposed to an experiment, along with a graph randomization scheme based on vertex clustering that yields non-vanishing probabilities of network exposure for users. We provide algorithms for efficiently computing these exposure probabilities, and show how a Horvitz-Thompson estimator can provide an unbiased effect estimate. We develop an understanding of the variance of this estimator through sufficient conditions on its asymptotic smallness, and construct an example in which our randomization scheme requires a non-trivial optimal vertex cluster size of 3 vertices.

1 Introduction

Social products and services – from fax machines and cell phones to online social networks – inherently exhibit 'network effects' with regard to their value to users. The value of these products to a user is inherently non-local, intimately linked to the value derived by the user's social ties also using the product. Yet randomized experiments (or 'A/B tests'), the standard machinery of the Rubin causal model [10], critically assume what is known as the 'Stable Unit Treatment Value Assumption' (SUTVA), that each individual's response is affected only by their own treatment and not by the treatment of any other individual.

Under ordinary randomized trials where SUTVA can be assumed – for example when a search engine A/B tests the effect of their color scheme upon the visitation time of their users – the users being treated respond just as they would if the entire population were treated, and the individuals in the control population respond just as they would if the entire population were in control. In this manner it is possible to observe behavior from two universes, 'Universe A' and 'Universe B', at the same time even though they are fundamentally counterfactual. Under social interference it is much more difficult to simultaneously observe the habits of the counterfactual social universes A and B. In this work, we adopt a framework recently characterized by Aronow and Samii for causal inference without this SUTVA assumption [2], and attune it to the challenges of interference on social networks.

Our underlying goal is to compute the average treatment effect between the universe where 'everyone has the service', the treatment universe, and the universe where 'no one has the service', the control universe. Let $\vec{z} \in \{0,1\}^n$ be the treatment assignment vector, where $z_i = 1$ means that user i is subject to a treatment intervention and $z_i = 0$ means they are in the control. Let $Y_i(\vec{z}) \in \mathbb{R}$ be the potential outcome of user i under the treatment assignment vector $\vec{z}$. The fundamental quantity we are interested in is the average treatment effect, τ, between the two diametrically opposite universes $\vec{z} = \vec{0}$ and $\vec{z} = \vec{1}$,

$$\tau(\vec{z} = \vec{1}, \vec{z} = \vec{0}) = \frac{1}{n}\sum_{i=1}^{n} \Big(Y_i(\vec{z} = \vec{1}) - Y_i(\vec{z} = \vec{0}) \Big), \tag{1}$$

and the clear difficulty in evaluating the above expression is that unlike ordinary A/B testing, no two users can ever truly be in opposing universes at the same time.

To get beyond this difficulty, we define a user as 'network exposed' to the treatment if the assignment vector of users to treatment and control, $\vec{z}$, is such that the response of the user is the same as in the treatment universe (when $\vec{z} = \vec{1}$). Similarly, a user is 'network exposed' to the control if the response of the user is the same as in the control universe ($\vec{z} = \vec{0}$). For example, one basic condition for network exposure to the treatment would be if a user is treated and all of their neighbors are treated. The definition of such situations is fundamentally a modeling decision by the experimenter, and in this work we introduce several families of exposure conditions, conditions that specify in which sets of assignment vectors a user is assumed to be 'network exposed' to the treatment and control universes, providing several characterizations of the continuum between the two universes. Choosing network exposure conditions is crucial because they specify when we can observe the potential outcome of a user as if they were in the treatment or control universe, without actually placing all users into the treatment or control universe.

For a single fixed exposure condition, it may be possible to design a randomization scheme that produces a sample where users enter the condition with equal probability, as has been the goal of recent work by Backstrom and Kleinberg [3]. When considering multiple exposure conditions, however, and when attempting to analyze these conditions using the entire graph, such sampling is not applicable. Thus, we develop a generic graph randomization scheme based on graph clustering, *clustered graph randomization*, and show how it is possible to precisely determine the non-uniform probabilities of entering network exposure conditions under such randomization. Using inverse probability weighting [8], we are then able to derive an unbiased estimator of the average treatment effect τ under any network exposure for which we can explicitly compute probabilities.

We motivate the power of clustered graph randomization by furnishing a proposition covering formal sufficient conditions under which clustered graph randomization will have asymptotically small variance: if the graph is sparse and the sizes of all the graph clusters are $O(1)$ in n, we show that the estimator variance is $O(1/n)$. To conclude, we study a simple example where the minimum variance is indeed achieved by a cluster size c_{opt} that is nontrivially $O(1)$, $c_{opt} = 3$.

This work occupies a mediating perch between recent work from the statistical literature on causal inference under interference [1, 2, 12], as well as recent work from the computer science literature on network bucket testing [3, 9]. Our contribution extends upon the ordinary inference literature by developing exposure models and randomization schemes particularly suited for experiments on large social graphs, where previous approaches are intractable. Meanwhile, our contribution also connects to existing work on network bucket testing by contributing an exposure framework for the full graph and a randomization scheme that is capable of considering multiple exposure conditions at once, a necessity for true concurrent causal experimentation.

Section 2 describes our models of network exposure. Section 3 presents our graph clustered randomization scheme, along with an algorithm for efficiently computing exposure probabilities, and sufficient conditions for the estimator variance to be small. In Section 4, we show how homophily can plan a key role in the priorities of our randomization by analyzing a simple example of inference under homophily using clustered randomization, with the surprising property that the minimal variance estimator is achieved by a graph clustering with strictly intermediate cluster size.

2 Exposure models

For A/B randomized experiments, the *treatment condition* of an individual decides whether or not they are subject to an intervention. This typically takes two values: 'treatment' or 'control'. In most randomized experiments, the experimenter has explicit control over how to randomize the treatment conditions, and generally individuals are assigned independently. Meanwhile, the *exposure condition* of an individual determines how they experience the intervention in full conjunction with how the world experiences the intervention. Without SUTVA, at worst each of the 2^n possible values of $\vec{z}$ define an exposure condition for each user. Aronow and Samii call this "arbitrary exposure" [2], and there would be no meaningful way to analyze experiments under arbitrary exposure.

Consider the potential outcomes for user i. In the "arbitrary exposure" case, $Y_i(\vec{z})$ is completely different for every possible $\vec{z}$. This means that we will never be able to observe either $Y_i(\vec{z})$ for

$\vec{z} = \vec{1}$ or $\vec{z} = \vec{0}$ without putting all users into the treatment or control universes. Therefore we cannot estimate the average treatment effect through an experiment without further assumptions. Here, we escape this conclusion by assuming a many-to-one relationship between the treatment assignment vector $\vec{z}$ and the potential outcomes. We define the set d_i^k of $\vec{z}$'s that map a user i into a particular potential outcome (with outcomes indexed by k) for that user as an "exposure condition". Our interest is in the sets d_i^1 and d_i^0 that we define to include $\vec{z} = \vec{1}$ and $\vec{z} = \vec{0}$ respectively. In this way, we are assuming that for all $\vec{z}_1 \in d_i^1$, $Y_i(\vec{z} = \vec{z}_1) = Y_i(\vec{z} = \vec{1})$, and for all $\vec{z}_0 \in d_i^0$, $Y_i(\vec{z} = \vec{z}_0) = Y_i(\vec{z} = \vec{0})$.[1] Note that it is possible that $\vec{z} = \vec{1}$ and $\vec{z} = \vec{0}$ belong to the same exposure condition and that $d_i^1 = d_i^0$. This corresponds to a treatment that has no effects.

We define an *exposure model* for user i as a set of exposure conditions that completely partition the possible assignment vectors $\vec{z}$. The set of all such models, across all users, is the exposure model for an experiment. For our purposes though, it is unnecessary to entirely specify an exposure model to determine the average treatment effect between the extreme universes. We only care about the exposure conditions d_i^1 and d_i^0 for which each user i will experience network exposure to the treatment or control universe[2].

Of course, the *true* exposure conditions d_i^1 and d_i^0 for each user are not known to the experimenter a priori, and analyzing the results of an experiment requires choosing such conditions in our framework. If the wrong exposure conditions are chosen by the experimenter, what happens to the estimate of the average treatment effect? The conclusion is that if users are responding in ways that do not correspond to $\vec{z} = \vec{1}$ and $\vec{z} = \vec{0}$, we will be introducing bias into the average treatment effect. The magnitude of this bias depends on how close the potential outcomes actually observed are to the potential outcomes at $\vec{z} = \vec{1}$ and $\vec{z} = \vec{0}$ that we wanted to observe. It may even be favorable to allow such bias in order to lower variance in the results of the experiment.

We now introduce local exposure conditions, where the relevant part of the treatment assignment vector only depends on the immediate graph neighborhood of a vertex. We consider absolute and fractional conditions on the number of treated neighbors. Note we are not asserting that these possible exposure conditions are the *actual* exposure conditions with respect to the actual potential outcomes in an experiment, but rather that they provide useful abstractions for the analysis of an experiment, where again the degree of bias introduced depends on how well the exposure conditions approximate belonging to the counterfactual universes.

Exposure condition 1 (Full neighborhood exposure) *A vertex experiences full neighborhood exposure to a treatment condition if they and all their neighbors receive that treatment condition.*

Exposure condition 2 (Absolute k-neighborhood exposure) *A vertex of degree r, where $r \geq k$, experiences absolute k-neighborhood exposure to a treatment condition if they and $\geq k$ of their neighbors receive that treatment condition.*

Exposure condition 3 (Fractional q-neighborhood exposure) *A vertex of degree r experiences fractional q-neighborhood exposure to a treatment condition if they and $\geq qr$ of their neighbors receive that treatment condition.*

The k-absolute and q-fractional neighborhood exposures can be considered relaxations of the full neighborhood exposure for vertex i in that they require less neighbors of i to have a fixed treatment condition for i to be considered as belonging to that exposure condition. In fact, the set of assignment vectors that correspond to k-absolute and q-fractional neighborhood exposures are each nested under the parameters k and q respectively. Increasing k or q decreases the set of assignment vectors until reaching full neighborhood exposure for vertex i.

[1] If this strikes the reader as too restrictive a definition of "exposure condition", consider instead partitioning the space of potential outcomes (rather that partitioning the space of assignment vectors) using small ϵ-sized bins, and define the "exposure conditions" as all assignment vectors that produce a potential outcome in that ϵ bin. In cases where no other potential outcomes correspond to the outcomes for $\vec{z} = \vec{0}$ or $\vec{z} = \vec{1}$, it may be more appropriate to manage bias using ϵ distances in the space of potential outcomes in this way.

[2] If one was to assume functional relationships between the potential outcomes in different exposure conditions then other exposure conditions besides d_i^1 and d_i^0 could be relevant to our computations.

It is natural to consider heterogeneous values k or q that are different for each user. One might imagine that some users are more susceptible to their neighbors than others, but we limit our discussion to exposure conditions homogeneous across users as much as possible. Unfortunately, we are required to consider a mild heterogeneity in the case of k-neighborhood exposure where some vertices have degree $r<k$. For these vertices when considering k-neighborhood exposure, we consider full neighborhood exposure instead. Fractional exposure is free from these parsimony problems.

Full neighborhood exposure is clearly only an approximation of full immersion in a universe. Beyond local exposure conditions, it may be fruitful to consider exposure condition with global dependence. As one approach, consider individuals as exposed to a treatment only if they are sufficiently surrounded by sufficiently many treated neighbors who are in turn also surrounded by sufficiently many treated neighbors, and so on. This recursive definition may initially appear intractable, but we note here briefly that such recursive exposure is exactly characterized by the concept of heterogeneous $\mathbf{k}$-cores [6] on the induced graph of treated individuals. We do not use this definition in what follows, so we do not pursue it further here. Instead, we elaborate on 'core exposure' in an appendix.

Other exposure conditions may very well prove perfectly relevant to some applications. In particular, we draw attention to the mesoscale idea of placing absolute or fraction conditions on the population of vertices within k hops, where $k = 1$ yield the neighborhood exposure conditions above. We also note that on social networks with very high degree, for many applications it may be more relevant to define the exposure conditions in terms of a lower degree network that considers only strong ties.

3 Estimation and randomization

Using the concept of network exposure, we can now consider estimating the average treatment effect τ between the two counterfactual universes using a randomized experiment. Recall that $\vec{z}$ is the treatment assignment vector of an experiment. To randomize the experiment, let $\vec{z}$ be drawn from Z, a random vector that takes values on $\{0,1\}^n$, the range of $\vec{z}$. The distribution of Z over $\{0,1\}^n$ given by $\Pr(Z = \vec{z})$ is what defines our randomization scheme, and it is also exactly what determines the relevant probabilities of network exposure. For a user i, $\Pr(Z \in d_1^i)$ is the probability of network exposure to treatment and $\Pr(Z \in d_0^i)$ is the probability of network exposure to control.

In general, these probabilities will be different for each user and each treatment condition, and knowing these probabilities makes it possible to correct for allocation bias during randomization. In particular, it becomes possible to use the Horvitz-Thompson estimator, $\hat{\tau}$, to obtain an unbiased estimate of τ, here given by

$$\hat{\tau}(Z) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i(Z)\mathbf{1}[Z \in d_0^i]}{\Pr(Z \in d_0^i)} - \frac{Y_i(Z)\mathbf{1}[Z \in d_1^i]}{\Pr(Z \in d_1^i)}\right), \tag{2}$$

where $\mathbf{1}[x]$ is the indicator function. Taking the expectation over Z clearly gives τ, though note that this does assume that the exposure conditions are not misspecified.

Let us examine the exposure probabilities for the simplest network exposure condition, full neighborhood exposure, and under the simplest randomization scheme, independent vertex randomization. If all vertices are treated independently with probability p then the probability of full neighborhood exposure to treatment for a user i of degree r_i is simply given by $\Pr(Z \in d_1^i) = p^{r_i+1}$, and the probability of full neighborhood exposure to control is given by $\Pr(Z \in d_0^i) = (1-p)^{r_i+1}$. This calls plain attention to the main challenge of network exposure: the chance that a vertex with high degree manages to reach full neighborhood exposure, or anywhere near it, can be astronomically small. Intuitively, such small exposure probabilities will dramatically increase the variance of the Horvitz-Thompson estimator, and it indicates the necessity of using more intelligent randomization.

A useful randomization scheme would work to reduce the variance of this Horvitz-Thompson estimator. Minimizing variance requires manipulating the exposure probabilities, which are tightly constrained by the connectivity of the underlying graph and the properties of the network exposure conditions being considered. Rather than attack this variance directly, we study the variance of a tractable and scalable class of randomization schemes based on graph clustering, where we are able to provide sufficient conditions for the design of graph clusters that provide asymptotic guarantees about the estimator variance.

3.1 Clustered graph randomization

The basic idea behind clustered graph randomization is to assign connected vertices to the same treatment condition more often then would happen with independent assignment. By doing so, we are essentially increasing the expected number of users who are network exposed at the cost of increased correlations between users' exposure conditions.

We assign each user to one of n_c clusters $C_1, \ldots, C_{n_c}$, such that connections are generally internal to these clusters, and then randomize treatment independently at the cluster level. The efficient creation of these clusters is not our focus here. For the creation of the clusters, we defer to the broad literature on graph clustering and community detection algorithms [7]. Under such randomization, a vertex's probability of exposure to local neighborhood exposure conditions now depends on the treatment conditions of the clusters to which it is connected. Let $N_i \subset V$ denote the vertices that i is connected to in the graph, and let $S_i = \{C_j : (i \cup N_i) \cap C_j \neq \emptyset\}$ denote the set of clusters to which i has a connection, as well as the cluster containing i itself.

For full neighborhood exposure, the probabilities of network exposure to treatment becomes $\Pr(Z \in d_i^0) = p^{|S_i|}$ and to control becomes $\Pr(Z \in d_i^1) = (1-p)^{|S_i|}$. If the graph clusters are relatively distinct in the network, $|S_i|$ will be much smaller than the degree of i, which serves to dramatically increase the assignment probability. We now show how computing the exposure probabilities for absolute and fractional neighborhood exposure conditions is tractable as well.

Consider the challenge of computing the probability that vertex i with degree r_i is treated and more than k of its neighboring vertices are treated under independent cluster randomization. This applies when considering both absolute and fractional neighborhood exposures. First, let us reindex the clusters such that if i is connected to $|S_i| = \ell$ clusters, i itself resides on cluster ℓ, and we let $j = 1, \ldots, \ell-1$ denote the other connected clusters. Let $w_{i1}, \ldots, w_{i\ell}$ be the number of connections i has to each cluster, and let the Bernoulli(p) random variables $X_1, \ldots, X_\ell$ denote the independent coin tosses associated with each cluster. Then:

$$\Pr[Z \in d_i^1] = \Pr\left[X_\ell = 1, \textstyle\sum_{j=1}^{\ell} w_{ij} X_j \geq k\right] = \Pr\left[X_\ell = 1\right] \cdot \Pr\left[\textstyle\sum_{j=1}^{\ell-1} w_{ij} X_j \geq k - w_{i\ell}\right],$$
$$\Pr[Z \in d_i^0] = \Pr\left[X_\ell = 0, \textstyle\sum_{j=1}^{\ell} w_{ij} X_j \leq r_i - k\right] = \Pr\left[X_\ell = 0\right] \cdot \Pr\left[\textstyle\sum_{j=1}^{\ell-1} w_{ij} X_j \leq r_i - k\right].$$

Here the random quantity $\sum_j w_{ij} X_j$ obeys a weighted equivalent of a Poisson-binomial distribution, and the probabilities in question can be computed explicitly using a dynamic program defined by the following recursion

$$\Pr\left[\textstyle\sum_{j=1}^{\ell} w_j X_j \geq T\right] = p \Pr\left[\textstyle\sum_{j=1}^{\ell-1} w_{ij} X_j \geq T - w_{i\ell}\right] + (1-p) \Pr\left[\textstyle\sum_{j=1}^{\ell-1} w_{ij} X_j \geq T\right],$$

Note that T is bounded for bounded degree graphs, making this is a strongly polynomial dynamic program with runtime $O(T\ell)$. We formalize this computation into the following proposition.

Proposition 1 *The probability that vertex i is treated and $\geq$ k neighboring vertices are treated under independent cluster randomization is given by* $\Pr[Z \in d_i^1] = pf(\ell-1, k-w_{i\ell}; p, \vec{w})$ *where*

$$\begin{aligned} f(0, T; p, \vec{w}_i) &= 1 - p\mathbf{1}[T < w_{i0}], \\ f(j, T; p, \vec{w}_i) &= pf(j-1, T - w_{ij}; p, \vec{w}_i) + (1-p) f(j-1, T; p, \vec{w}_i). \end{aligned}$$

The probability that vertex i is in control and $\geq$ k neighboring vertices are in control under independent cluster randomization is given by $\Pr[Z \in d_i^0] = (1-p)[1 - f(\ell-1, r_i-k+1; p, \vec{w})]$.

Recall that these partial neighborhood exposure conditions (absolute and fractional) are nested. In fact, for a given vertex i the recursion can be used to derive the probability for every possible threshold value under consideration in one straight-forward doubled for-loop.

3.2 Clustered randomization and estimator variance

There are two ingredients behind exposure probabilities: the probability distribution over treatment assignments and the specification of the exposure conditions. The randomization scheme is certainly under the experimenter's control and hence we can adjust these probabilities, but when randomizing for network exposure, the probability of network exposure can only be controlled indirectly.

The variance of the Horvitz-Thompson estimator under interference has been studied by Aronow and Samii, where they also present a conservative estimator of the variance itself, along with several

variance reduction schemes. Estimating the variance under their approach require knowledge of joint exposure conditions such as the probability that vertex i is network exposed to treatment/control and vertex j is network exposed to treatment/control. This is the probability that the random vector Z is in the exposure condition for vertex i and for vertex j simultaneously, e.g. $\Pr(Z \in (d_i^1 \cap d_j^1))$ for joint network exposure to treatment. Note that there is nothing fundamentally different about this probability computation over the single vertex case, besides for the fact that the intersection of the two sets can be empty. Aronow and Samii observe that an empty intersection does make it impossible to derive an unbiased estimate of the variance, but it does not bias the effect estimator itself.

The variance of the estimator where $\hat{Y}^x(Z) = 1/n \sum_i (Y_i(Z)\mathbf{1}[Z \in d_i^x]/\Pr(Z \in d_i^x))$ is given by

$$\text{Var}(\hat{\tau}(Z)) = \left[\text{Var}(\hat{Y}^1(Z)) + \text{Var}(\hat{Y}^0(Z)) - 2\text{Cov}(\hat{Y}^1(Z), \hat{Y}^0(Z))\right]. \tag{3}$$

Assuming the exposure conditions are properly specified, namely assuming that $Y_i(\vec{z})$ is constant for all $\vec{z} \in d_i^x$, we can introduce the notation $Y_i(d_i^x) := Y_i(\vec{z} \in d_i^x)$. Using the further notation $\pi_i^x := Pr[Z \in d_i^x]$ and $\pi_{ij}^{xy} := Pr[Z \in (d_i^x \cup d_j^y)]$ we obtain

$$\text{Var}[\hat{Y}^x(Z)] \quad = \quad \frac{1}{n^2}\left[\sum_{i=1}^{n} \frac{1-\pi_i^x}{\pi_i^x} Y_i(d_i^x)^2 + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\pi_{ij}^{xx} - \pi_i^x \pi_j^x}{\pi_i^x \pi_j^x} Y_i(d_i^x) Y_j(d_j^x)\right],$$

and

$$\text{Cov}(\hat{Y}^1(Z), \hat{Y}^0(Z)) = \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\pi_{ij}^{10} - \pi_i^1 \pi_j^0}{\pi_i^1 \pi_j^0} Y_i(d_i^1) Y_j(d_j^0) - \sum_{i=1}^{n} Y_i(d_i^1) Y_i(d_i^0)\right].$$

For neighborhood exposure conditions, we can write sufficient conditions under which the variance of the estimator is asymptotically $O(1/n)$ in n for clustered graph randomization.

Proposition 2 *Assume the potential outcomes are all $O(1)$ in n. If G is sparse with degrees bounded by $O(1)$ and the size of each cluster is $O(1)$, then the variance of the Horvitz-Thompson ATE for full, k-neighborhood, and q-fractional neighborhood exposure under clustered graph randomization is $O(1/n)$.*

Proof. Assume G is sparse and the size of each cluster is $O(1)$. All of the single sums are clearly $O(n)$: π_i^x is $\mathrm{O}(1)$ since all vertices have bounded degree. For the double sums, note that $\pi_{ij}^{xx} = \pi_i^x \pi_j^x$ if and only if i and j have no common cluster neighbors, $|S_i \cap S_j| = 0$. Whenever $|S_i \cap S_j| > 0$, $\pi_{ij}^{xx} > \pi_i^x \pi_j^x$ for full, k-neighborhood, and q-fractional neighborhood exposure. Further, we have that $\pi_{ij}^{10} < \pi_i^1 \pi_j^0$ if $|S_i \cap S_j| > 0$ and $\pi_{ij}^{10} = \pi_i^1 \pi_j^0$ otherwise.

Terms of the double sums are zero whenever $\pi_{ij} = \pi_i \pi_j$ and when the terms are not zero ($|S_i \cap S_j| > 0$), the terms are all positive and bounded above $O(1)$ due to the bounded degrees. Then a single vertex i at most connects to $O(1)$ clusters and therefore $|S_i| = O(1)$. Then $|S_i \cap S_j| > 0$ for every vertex j in each of i's cluster neighbors. Since the sizes of all clusters are $O(1)$, there are $O(1)$ such vertices j for each cluster in S_i. Thus for each vertex, at most $O(1)$ of the terms in the double sum are positive, making the total variance $O(1/n)$. ■

4 Homophily and estimator variance

So far we have primarily considered the conditions under which a user *qualifies* as being network exposed, but have not focused much on what happens when the user *is* network exposed. The variance of the estimator depends not just on the exposure probabilities but also on the actual potential outcomes. Further, these potential outcomes can be expected to be related to the graph structure from which we have generated clusters. One reason for such a relationship is homophily [11, 4]: network neighbors might be expected to respond in similar ways to network exposure.

If potential outcomes were homogeneous or exchangeable across the vertices of the graph, a situation corresponding to a complete absence of homophily in the potential outcomes, a plausible strategy for reducing variance would be to cut the graph into two maximally dense halves so as to maximize exposure probabilities, and to consider a randomization scheme that placed one cluster

into treatment and the other into control. For partitioning the United States, this would effectively be achieved by randomizing the population as two large clusters partitioned along the longitudinal population median. Meanwhile, with homophily, this seemingly reasonable strategy would be crucially ill-fated: many east-west confounders would contribute considerably to the variance of the estimator, occurring exclusively in the treatment population or exclusively in the control population. Examples of confounders include differences in social norms, differences in time zone and differences in datacenter latency, the latter two being very important for online real-time experiments.[3]

The presence of these two forcing terms – maximizing exposure probabilities suggests using only a few large clusters while homophily suggest using many small clusters – suggests that an intermediate cluster size may minimize variance. The following example is meant to illustrate this subtlety.

In our example, we are interested in measuring the average treatment effect on a graph G consisting of a single cycle with $2n$ vertices. We consider the simple full neighborhood exposure model, where we are interested in the average treatment effect between d_i^1, where a vertex is treated and both of their neighbors are treated, and d_i^0, where a vertex is not treated and neither of their neighbors are treated. For the fixed response model, we introduce a highly polarized response where the vertices are split by a balanced division of the cycle, where one half of the cycle responds $Y_i(d_i^1) = a$ to the network exposure condition and the second half responds $Y_i(d_i^1) = b$. Meanwhile, all $2n$ vertices respond $Y_i(d_i^0) = 0$ to the network control exposure condition d_i^0. Importantly, we consider the orientation of the partition to be unknown to the experiment designer.

The challenge now is to design clusters for randomization in this example. If the orientation of the polarization divide were known, it would be possible to bisect the cycle orthogonal to this response divide, creating two clusters that were evenly balanced with respect to the polarization. But with the orientation of the polarization unknown, the best one can hope to do with two clusters is to randomize uniformly over the $2n$ possible optimal graph bisections. At the other end of the spectrum, it is not clear that randomizing each vertex independently into treatment and control who produce meaningful exposure populations with enough vertices adjacent to two identically treated neighbors. We therefore consider: is it possible that the variance is minimized by an intermediate scale clustering?

Let us consider the variance of the Horvitz-Thompson average treatment effect estimator from (3) above. Since all vertices respond zero to the control condition in our example, as long as the exposure probability for the control condition is strictly positive then both $\text{Var}(\hat{Y}_{HT}(d_0))$ and $\text{Cov}(\hat{Y}_{HT}(d_1), \hat{Y}_{HT}(d_0))$ are zero. Since our calculations will rely only on probabilities π_i^1 for the network exposure to treatment condition, we omit the superscript. The variance reduces to:

$$\text{Var}[\hat{\tau}(Z)] = \frac{a^2+b^2}{4n^2}\left[\sum_{i=1}^{n}\frac{1-\pi_i}{\pi_i} + \sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j}\right] + \frac{ab}{4n^2}\sum_{i=1}^{n}\sum_{j=i}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j}. \tag{4}$$

Notice that the terms of the double sums are only non-zero for vertex pairs where $\pi_{ij} \neq \pi_i\pi_j$.

Optimal bisection

The variance of the Horvitz-Thompson estimator when randomizing uniformly over the $2n$ bisections that form two macro-clusters is easy to characterize directly from the definition of variance. The estimator has only $2n$ possible outcomes, with equal probability. It is an unbiased estimator, with expected value $\mu = (a+b)/2$. A straightforward asymptotic calculation shows us:

$$\text{Var}[\hat{\tau}_{HT}(d_1, d_0)] \quad = \quad \frac{1}{2n}\left[(a-\mu)^2 + (b-\mu)^2 + 2\sum_{k=0}^{n-2}\left(\frac{k}{n-2}a + \frac{n-2-k}{n-2}b - \mu\right)^2\right]$$

This computes to $\text{Var}(\hat{\tau}(Z)) = \frac{(a-b)^2}{12}\left[1 + O\left(1/n\right)\right]$. Note that if $a = b$, this variance is zero, but we are interested in the case of polarization, when $a \neq b$.

Individual randomization

On the other end of the spectrum, consider each vertex as a cluster. The probability of being exposed and both of one's neighbors being exposed is equal to the probability of seeing three independent

[3]Even if we do have covariate balance across the split of the vertices into two groups, this certainly does not imply that the potential outcomes must be homogeneous or exchangeable.

coins come up heads. When the randomization is balanced (e.g. $p = 1/2$), we obtain $\pi_i = 1/8, \forall i$. Note that the co-assignment probabilities depend on whether vertices i and j are neighbors or share a neighbor. From this we derive $\pi_{ij} = 1/16$ if $|i - j| = 1$ and $\pi_{ij} = 1/32$ if $|i - j| = 2$, and if $|i - j| > 2$, the probabilities are independent.

We can evaluate the variance using the simplification (4) given earlier. For the second double sum, this adjacency condition contributes that only $O(1)$ (i, j) pairs, so the second double sum is asymptotically dominated. Thus we obtain $\text{Var}(\hat{\tau}(Z)) = (15/4)(a^2 + b^2)\frac{1}{n} + O(1/n^2)$.

Randomizing pairs of vertices

Consider now a slightly more clustered randomization, where the clusters consist of alternating pairs of adjacent vertices. There are two ways to pair the vertices of the cycle, shifted from each other by 1 vertex, but these two clusterings are asymptotically equal in their variance. With the orientation of the polarization unknown, it would be equally probable to select either of these two clusterings. Here we will simply analyze an estimator based on either one of these two clusterings.

Under paired randomization, $\pi_i(d_1) = \pi_i(d_0) = \pi = 1/4, \forall i$. The non-independent co-assignment probabilities again depend on adjacency, but now we must also consider the different possible ways in which pairs can be adjacent.

$$\begin{array}{llll} |i-j|=1: & \pi_{ij}=1/4 & (n-1) \text{ times}, & \pi_{ij}=1/8 \quad (n-1) \text{ times}, \\ |i-j|=2: & \pi_{ij}=1/8 & 2(n-2) \text{ times}, & \\ |i-j|=3: & \pi_{ij}=1/8 & (n-3) \text{ times}, & \pi_{ij}=1/16 \quad (n-3) \text{ times}. \end{array}$$

Notice that half of all distance-3 pairs have independent co-assignment probability, $\pi_{ij} = 1/16$. Now the variance of the estimator amounts to $\text{Var}(\hat{\tau}(Z)) = (10/4)(a^2 + b^2)\frac{1}{n} + O(1/n^2)$.

Randomizing $c \geq 3$ vertices

Now, consider randomizing blocks of $c \geq 3$ vertices, where c does not depend on n. The calculations for this case are expansive but straight-forward and analogous to the calculations for $c = 2$ above. For a fixed clustering (of the three shifted possibilities), the variance computes to:

$$\begin{aligned} \text{Var}(\hat{\tau}(Z)) \quad = \quad & \frac{a^2+b^2}{4n^2}\Bigg[\left(n+\frac{4n}{c}\right) + \underbrace{\frac{2n}{c}(c+2)}_{|i-j|=1} + \underbrace{\frac{2n}{c}\sum_{k=2}^{c-2}(c-k+2)}_{1<|i-j|<c-1} \\ & + \underbrace{\frac{2n}{c}3}_{|i-j|=c-1} + \underbrace{\frac{2n}{c}2}_{|i-j|=c} + \underbrace{\frac{2n}{c}}_{|i-j|=c+1}\Bigg] + O(1/n^2) \end{aligned}$$

This computes to $\text{Var}(\hat{\tau}(Z)) = \left(\frac{c}{4} + \frac{2}{c} + 1\right)(a^2 + b^2)\frac{1}{n} + O(1/n^2)$. We see that in the presence of the strong homophily of this example, the variance of the graph clustered randomization scheme is minimized by randomizing clusters of $c = 3$ vertices. Meanwhile in the absence of the homophily, represented above by $a = b$, an entirely different solution, randomizing according to an optimal bisection of the graph, minimizes the variance.

5 Discussion

In this paper we have laid out a framework for confronting A/B testing under social interference, but by no means have we "solved" this problem. There are many open questions with respect to the method proposed herein. First, it is not clear how to formulate a proper computationally tractable objective for minimizing the variance of the Horvitz-Thompson estimator. It may be possible to minimize an adversarial variance [9] or the A/A variance of the clustered graph randomization scheme for full neighborhood exposure, under the assumption of known control potential outcomes.

Open question 1 *The A/A variance of the H-T estimator of the average treatment effect for full neighborhood exposure under clustered graph randomization with independent probability p is*

$$\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Y_i(\vec{z}=\vec{0})Y_j(\vec{z}=\vec{0})\left(\frac{p^{-|S_i\cap S_j|}+(1-p)^{-|S_i\cap S_j|}}{2}+\theta(|S_i\cap S_j|)-1\right) \tag{5}$$

where $\theta(x) = 1$ if $x > 0$ and zero otherwise. Does there exist an efficient algorithm to find the cluster assignment that minimizes the above expression? Under what assumptions would such a cluster assignment be useful for A/B testing?

We note that A/A variance minimization would not be a useful strategy if the treatment is expected to be dominated by heterogeneous responses.

For additional future work, one might consider additional structure on the potential outcomes, such as modeling relationships between potential outcomes in different exposure conditions. A particularly simple example would be a linear relationship with k for k-neighborhood exposure. How could we properly take advantage of such structure to get better estimates? A deeper understanding of bias under network exposure condition misspecification would also greatly improve the framework.

Finally, a methodology is not very useful if it is not implemented. Our immediate future goal is to apply our methodology to real-world experiments, and establish that we can accurately measure responses in parallel social universes in practice.

A Appendix: Core exposure

Here we introduce a stronger, global approach to exposure, briefly referenced in the main text. Consider individuals as exposed to a treatment only if they are sufficiently surrounded by sufficiently many treated neighbors who are in turn also surrounded by sufficiently many treated neighbors, and so on. This recursive definition of network exposure can be captured by the concept of heterogeneous **k**-cores on the induced graph of treated individuals.

Recall that the k-core of a graph $G = (V, E)$ is the maximal subgraph of G in which all vertices have degree at least k [5]. Similarly, the heterogeneous **k**-core of a graph $G = (V, E)$, parameterized by a vector $\mathbf{k} = (k_1, \ldots, k_{|V|})$, is the maximal subgraph $H = (V', E')$ of G in which each vertex $v_i \in V'$ has degree at least k_i [6]. Using the definition of heterogeneous **k**-core, we introduce the following natural fractional analog.

Definition 1 (Fractional q-core) *The fractional q-core is the maximal subgraph $H = (V', E')$ of $G = (V, E)$ in which each vertex $v_i \in V'$ is connected to at least a fraction q of the vertices it was connected to in G. Thus, for all $v_i \in V'$, $\deg_H(v_i) \geq q \deg_G(v_i)$. Equivalently, if r_i is the degrees of vertex i, the fractional q-core is the heterogenous **k**-core of G for $\mathbf{k} = (qr_1, \ldots, qr_{|V|})$.*

We assert without proof that this is a well defined object. Using this definition, we now define exposure conditions that are all stricter versions of corresponding earlier neighborhood conditions.

Exposure condition 4 (Component exposure) *A vertex experiences component exposure to a treatment condition if it and all of the vertices in its connected component receive that treatment condition.*

Exposure condition 5 (Absolute k-core exposure) *A vertex with degree $r \geq k$ experiences absolute k-core exposure to a treatment condition if it belongs to the k-core of the graph $G[V']$, the subgraph of G induced on the set of vertices V' that receive that treatment condition.*

Exposure condition 6 (Fractional q-core exposure) *A vertex experiences fractional q-core exposure to a treatment condition if it belongs the the fractional q-core of the graph $G[V']$, the subgraph of G induced on the set of vertices V' that receive that treatment condition.*

Component exposure is perhaps the strongest requirement for network exposure imaginable, and it is only feasible if the interference graph being studied is comprised of many disconnected components. We include it here specifically to note that the fractional q-core exposure for $q = 1$ reduces to component exposure. Again like the neighborhood exposure case, absolute core exposure requires heterogeneity in k across users for it to be a useful condition for all users. A parsimonious solution analogous to the solution for k-neighborhood exposure may be to consider heterogeneous max(degree, k)-core exposure. Fractional q-core exposure, like fractional q-neighborhood exposure, is again free from these parsimony problems.

Core exposure conditions are strictly stronger than the associated neighborhood exposure conditions above. In fact, every assignment vector in which a vertex i would be component or core exposed

corresponds to neighborhood exposure, but not vice versa. So the assignment vectors of core and component exposure are entirely contained in those of the associated neighborhood exposure.

With regards to core exposure probabilities, the dynamic program discussed earlier provides a means of computing exposure probabilities for absolute and fractional neighborhood exposure conditions. Unfortunately, computing the exact probability of k-core and fractional q-core exposure conditions is an open question, but recall that these exposure conditions were formally stricter analogs of corresponding neighborhood exposure conditions. This allows us to upper bound the core exposure probabilities, and we formalize this connection via the following proposition. Because we are generally concerned about exposure probabilities being too small, this upper bound can be useful in identifying vertices with problematically small probabilities already under neighborhood exposure.

Proposition 3 *For a vertex i, the probability of network exposure to a treatment condition under core exposure is less than or equal to the probability under the analogous neighborhood exposure:*

$$\begin{aligned} \Pr(Z \in d_i^x | \textit{k-core}) &\leq Pr(Z \in d_i^x | \textit{k-neighborhood}), \\ \Pr(Z \in d_i^x | \textit{ fractional q-core}) &\leq Pr(Z \in d_i^x | \textit{ fractional q-neighborhood}). \end{aligned}$$

References

[1] E.M. Airoldi, E. Kao, P. Toulis, and D.B. Rubin. Valid estimates of causal peer-influence effects with interfering units. *Working Paper*, August 2012.

[2] P. Aronow and C. Samii. Estimating average causal effects under general interference. *Working Paper*, September 2012.

[3] L. Backstrom and J. Kleinberg. Network bucket testing. In *Proc. 20th International World Wide Web Conference*, pages 615–624. ACM, 2011.

[4] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proc. 21st international World Wide Web Conference*, pages 519–528. ACM, 2012.

[5] B. Bollobás. *Random graphs*. Cambridge University Press, 2001, p. 150.

[6] D. Cellai, A. Lawlor, K.A. Dawson, and J.P. Gleeson. Critical phenomena in heterogeneous k-core percolation. *arXiv preprint arXiv:1209.2928*, 2012.

[7] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[8] D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *J. Am. Statistical Association*, 47(260):663–685, 1952.

[9] L. Katzir, E. Liberty, and O. Somekh. Framework and algorithms for network bucket testing. In *Proceedings of the 21st international conference on World Wide Web*, pages 1029–1036. ACM, 2012.

[10] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology; Journal of Educational Psychology*, 66(5):688, 1974.

[11] C.R. Shalizi and A.C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011.

[12] E.J.T. Tchetgen and T.J. VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.