Biometrika (2019), **106**, 4, *pp.* 929–940
Printed in Great Britain
doi: 10.1093/biomet/asz047
Advance Access publication 30 September 2019

Testing for arbitrary interference on experimentation platforms

By J. POUGET-ABADIE

Google Research, 111 8th Avenue, New York, New York 10011, U.S.A.
jeanpa@google.com

G. SAINT-JACQUES, M. SAVESKI

Massachusetts Institute of Technology, 100 Main Street, Cambridge, Massachusetts 02139, U.S.A.
gsaintja@mit.edu msaveski@mit.edu

W. DUAN, S. GHOSH, Y. XU

LinkedIn, 1000 W. Maude Avenue, Sunnyvale, California 94043, U.S.A.
wduan@linkedin.com sghosh@linkedin.com yaxu@linkedin.com

and E. M. AIROLDI

Fox School of Business, Temple University, 1810 Liacouras Walk, Philadelphia, Pennsylvania 19122, U.S.A.
airoldi@temple.edu

Summary

Experimentation platforms are essential to large modern technology companies, as they are used to carry out many randomized experiments daily. The classic assumption of no interference among users, under which the outcome for one user does not depend on the treatment assigned to other users, is rarely tenable on such platforms. Here, we introduce an experimental design strategy for testing whether this assumption holds. Our approach is in the spirit of the Durbin–Wu–Hausman test for endogeneity in econometrics, where multiple estimators return the same estimate if and only if the null hypothesis holds. The design that we introduce makes no assumptions on the interference model between units, nor on the network among the units, and has a sharp bound on the variance and an implied analytical bound on the Type I error rate. We discuss how to apply the proposed design strategy to large experimentation platforms, and we illustrate it in the context of an experiment on the LinkedIn platform.

Some key words: Arbitrary interference; Causal inference; Potential outcome; Violation of the stable unit treatment value assumption.

1. Introduction

Applications of the potential outcomes approach to causal inference (Rubin, 1974; Imbens & Rubin, 2015) often rely on the assumption of no interference among the units of analysis:

the outcome of each unit does not depend on the treatment assigned to other units. This fact is formalized by the stable unit treatment value assumption (Rubin, 1990, p. 475). However, in applications where social information among the units of analysis is recorded this is often untenable, and classical results in causal inference may no longer hold true (Karwa & Airoldi, 2018). Examples of causal analyses where interference is present are numerous and include education policy (Hong & Raudenbush, 2006), viral marketing campaigns (Aral & Walker, 2011; Eckles et al., 2016) and social networks and healthcare (Shakya et al., 2017). In these examples, for instance, the difference-in-means estimator computed under Bernoulli randomization is no longer guaranteed to be an unbiased estimator of the average treatment effect.

Significant efforts to extend the theory of causal inference to scenarios where the stable unit treatment value assumption does not hold have been made. A popular approach to minimizing the effects of interference, cluster-based randomized designs, have been extensively studied, spanning from the early work (Cornfield, 1978; COMMIT, 1991; Donner & Klar, 2004) to more recent contributions (Ugander et al., 2013; Eckles et al., 2017). Multilevel designs where treatment is applied with different proportions across the population, known as randomized saturation designs (Hudgens & Halloran, 2008; Tchetgen & VanderWeele, 2012), are also an important tenet of the literature on improving causal estimates under interference, having been applied to vaccination trials (Datta et al., 1999) and more recently to voter mobilization campaigns (Sinclair et al., 2012). See Baird et al. (2014) for more references. More recent literature has developed around various assignment strategies and estimators, beyond cluster-based randomized design or multilevel designs, with some guarantees under specific models of interference (Backstrom & Kleinberg, 2011; Katzir et al., 2012; Manski, 2013; Toulis & Kao, 2013; Basse & Airoldi, 2018; Gui et al., 2015; Choi, 2017).

Although dealing with interference for the purpose of estimating causal effects is often important in causal inference problems where interference is due to a network, a more fundamental need is to be able to detect interference in a given experiment. Rosenbaum (2007) was the first to formulate two sharp null hypotheses that imply that the stable unit treatment value assumption does not hold. Under these restricted null hypotheses, the exact distribution of network parameters is known. More recent work (Aronow, 2012; Athey et al., 2018) explicitly tackles testing for the nonsharp null that the stable unit treatment value assumption holds, by considering the distribution of network effect parameters for a subpopulation of the graph under this assumption. In contrast to these post-experiment analysis methods, this paper suggests an experimental design test for interference.

2. METHODOLOGY

2.1. *Complete randomization and cluster-based randomization*

Consider N experimental units and a possible intervention on these units. Each unit's potential outcome Y_i is a function of the entire assignment vector $Z \in (0, 1)^N$ of units to one of two possible cases. If $Z_i = 1$, unit i is treated and given the intervention. Otherwise $Z_i = 0$, indicating that unit i is placed in control. The causal estimand of interest is the total treatment effect:

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

For any vector $u \in \mathbb{R}^N$, let $\bar{u} = N^{-1} \sum_{i=1}^{N} u_i$ and $\sigma^2(u) = (N - 1)^{-1} \sum_{i=1}^{N} (u_i - \bar{u})^2$. The total treatment effect can be rewritten as $\tau = \bar{Y(1)} - \bar{Y(0)}$. Two popular experimental designs, the completely randomized design and the cluster-based randomized design, provide unbiased

estimates of the total treatment effect. In a completely randomized experiment, the assignment vector Z is sampled uniformly at random from the set $\{z \in (0,1)^N : \sum z_i = n_t\}$, where n_t is the number of units assigned to treatment and $n_c = N - n_t$ is the number of units assigned to control. The difference-in-means estimator is $\hat{\tau}_{cr} = \bar{Y}_t - \bar{Y}_c$, where $Y_t = (Y_i : Z_i = 1)$ is the outcome vector of all units in treatment and $Y_c = (Y_i : Z_i = 0)$ is the outcome vector of all units in control. Let $S_t = \sigma^2\{Y(1)\}$, $S_c = \sigma^2\{Y(0)\}$ be the variances of the two potential outcomes under the stable unit treatment value assumption, and let $S_{tc} = \sigma^2\{Y(1) - Y(0)\}$ be the variance of the differences of the potential outcomes. The following result is widely known.

LEMMA 1. *When the stable unit treatment value assumption holds, the expectation and variance of the difference-in-means estimator $\hat{\tau}_{cr}$ under a completely randomized design are*

$$E_Z\left(\hat{\tau}_{cr}\right) = \tau, \quad \sigma^2_{cr} = \mathrm{var}_Z\left(\hat{\tau}_{cr}\right) = \frac{S_t}{n_t} + \frac{S_c}{n_c} - \frac{S_{tc}}{N}.$$

In a cluster-based randomized assignment, the randomization is over clusters of units, rather than individual units. Supposing that each experimental unit is assigned to one of M clusters, the cluster assignment vector z is sampled uniformly at random from $\{v \in (0,1)^M : \sum v_i = m_t\}$, assigning units in cluster $\mathcal{C}_j$ to the corresponding treatment: $Z_i = 1 \Leftrightarrow z_j = 1$ if $i \in \mathcal{C}_j$, where m_t is the number of clusters assigned to treatment and $m_c = M - m_t$ is the number of clusters assigned to control. Let Y^+ be the vector of aggregated potential outcomes, defined as $Y^+_j = \sum_{i \in \mathcal{C}_j} Y_i$, the sum of all outcomes within cluster $\mathcal{C}_j$. The Horvitz–Thompson estimator is defined as $\hat{\tau}_{cbr} = M/N\left(\bar{Y^+_t} - \bar{Y^+_c}\right)$, where $Y^+_t = (Y^+_j : z_j = 1)$ is the cluster-level outcome vector of all treated clusters and $Y^+_c = (Y^+_j : z_j = 0)$ is the cluster-level outcome vector of all clusters in the control bucket. Let $S^+_t = \sigma^2\{Y^+(1)\}$, $S^+_0 = \sigma^2\{Y^+(0)\}$ be the variance of the two aggregated potential outcomes under the stable unit treatment value assumption, and $S^+_{tc} = \sigma^2\{Y^+(1) - Y^+(0)\}$ be the variance of the difference of the aggregated potential outcomes. The following result is also widely known.

LEMMA 2. *When the stable unit treatment value assumption holds, the expectation and variance of the Horvitz–Thompson estimator $\hat{\tau}_{cbr}$ under a cluster-based randomized design are*

$$E_Z\left(\hat{\tau}_{cbr}\right) = \tau, \quad \sigma^2_{cbr} = \mathrm{var}_Z\left(\hat{\tau}_{cbr}\right) = \frac{M^2}{N^2}\left(\frac{S^+_t}{m_t} + \frac{S^+_c}{m_c} - \frac{S^+_{tc}}{M}\right).$$

Lemma 2 does not require the clusters to be of equal size. However, this assumption will be required for the hierarchical design presented in § 2.2. When the stable unit treatment value assumption holds, $\hat{\tau}_{cr}$ and $\hat{\tau}_{cbr}$ are unbiased estimators of the total treatment effect under their respective randomized designs. However, when the stable unit treatment value assumption does not hold, these results are no longer guaranteed and the estimate of the total treatment effect is expected to be different under each design when interference is present.

Assume a network over the experimental units, such that in the immediate neighbourhood $\mathcal{N}_i$ of unit i are the units likely to interfere with it, and the following model of potential outcomes:

$$Y_i(Z) = \alpha + \beta Z_i + \gamma \rho_i + \epsilon_i \quad (i = 1, \ldots, N), \tag{1}$$

where $\rho_i = |\mathcal{N}_i|^{-1} \sum_{j \in \mathcal{N}_i} Z_j$ is the average number of treated neighbours in unit i's neighbourhood $\mathcal{N}_i$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is a noise factor, with $\epsilon_i \perp\!\!\!\perp \rho_i$. Interference is present if and only if $\gamma \neq 0$.

Hence, β is often interpreted as a direct treatment effect parameter, while γ is often interpreted as an interference effect parameter. Under this parameterized model of potential outcomes with interference, the total treatment effect is $\tau = \beta + \gamma$.

LEMMA 3. *Under the model of interference in* (1)*, the expectations of* $\hat{\tau}_{cr}$ *and* $\hat{\tau}_{cbr}$ *under their respective randomized designs are*

$$E_{Z,\epsilon}(\hat{\tau}_{cr}) = \beta - \gamma(N-1)^{-1}, \quad E_{Z,\epsilon}(\hat{\tau}_{cbr}) = \beta + \gamma(\rho_{\mathcal{C}} M - 1)(M-1)^{-1},$$

where $\rho_{\mathcal{C}} = N^{-1}\sum_{i=1}^{N} |\mathcal{N}_i \cap \mathcal{C}(i)|/|\mathcal{N}_i|$ *is the average number of each unit's neighbours also present in its cluster, with* $\mathcal{C}(i)$ *being the units belonging to unit i's cluster.*

Lemma 3 states that when interference is present, neither estimator is unbiased for the total treatment effect, and, crucially, they do not have the same expected value.

2.2. *A hierarchical randomization strategy*

If it were possible to apply both the completely randomized and cluster-based randomized designs to all experimental units, testing for interference could be done by comparing the two estimates from each assignment strategy. If the two estimates were significantly different, then there is evidence that interference is present.

Unfortunately, just as both treatment and control cannot be assigned to each unit, both assignment designs cannot be given to all experimental units. To solve this problem, we randomly assign units to treatment arms and, within each treatment arm, apply one experimental strategy for assigning units to either intervention. In order to maintain some of the interference structure intact within each treatment arm without sacrificing covariate balance or introducing bias, we suggest using a cluster-based randomized design to assign units to treatment arms. Once clusters of units are assigned to one of two treatment arms, we suggest applying within each treatment arm either a cluster-based randomized design or a completely randomized design.

The experimental units are grouped into M balanced clusters, such that each cluster has the same number of units. Let m_{cr} be the number of clusters to be assigned to treatment arm *cr* and m_{cbr} be the number of clusters to be assigned to treatment arm *cbr*. Let n_{cr} and $n_{cbr} = N - n_{cr}$ be the resulting number of units assigned to each arm. Let $\omega \in (0,1)^M$ be the cluster-to-treatment-arm assignment vector and $W \in (0,1)^N$ be the corresponding unit-to-treatment-arm assignment vector: $W_i = 1$ if unit i is assigned to treatment arm *cr* and $W_i = 0$ if unit i is assigned to treatment arm *cbr*.

In treatment arm *cr*, let $n_{cr,t}$ be the number of units to be assigned to treatment and $n_{cr,c}$ be the number of unit to be assigned to control. Similarly, in treatment arm *cbr*, let $m_{cbr,t}$ and $m_{cbr,c}$ be the number of clusters that are assigned to treatment and control, respectively. Let $Z \in (0,1)^N$ be the assignment vector of units to treatment and control, composed of the two parts $Z_{cr} \in (0,1)^{n_{cr}}$ for units in treatment arm *cr* and $Z_{cbr} \in (0,1)^{n_{cbr}}$ for units in treatment arm *cbr*.

The hierarchical design is as follows: sample W in a cluster-based randomized way. Conditioned on W, sample Z_{cr} using a completely randomized assignment to assign units in treatment arm *cr* to treatment and control. Conditioned on W, sample Z_{cbr} using a cluster-based randomized assignment to assign units in treatment arm *cbr* to treatment and control. The resulting assignment vector Z of units to treatment and control is such that $Z_{cr} \perp\!\!\!\perp Z_{cbr} \mid W$.

Though the graph could be reclustered for the third step, a simpler option from an analytical and methodological perspective is to reuse the same clustering used in the first step. The two estimates of the causal effect for this experiment as well as the difference-in-differences estimator

Δ are defined as follows:

$$\hat{\tau}_{cr} = \bar{Y}_{cr,t} - \bar{Y}_{cr,c}, \quad \hat{\tau}_{cbr} = \frac{m_{cbr}}{n_{cbr}}\big(\bar{Y^+_{cbr,t}} - \bar{Y^+_{cbr,c}}\big), \quad \Delta = \hat{\tau}_{cr} - \hat{\tau}_{cbr},$$

where $Y_{cr,t} = (Y_i : W_i = 1 \wedge Z_i = 1)$ is the vector of outcomes of units in treatment arm cr that are treated. Similarly, $Y_{cr,c} = (Y_i : W_i = 1 \wedge Z_i = 0)$, $Y^+_{cbr,t} = (Y^+_j : \omega_j = 0 \wedge z_j = 1)$ and $Y^+_{cbr,c} = (Y^+_j : \omega_j = 0 \wedge z_j = 0)$. In the spirit of the Durbin–Hausman–Wu test, we could have chosen to compare different estimators or different designs altogether, a comparison which is left to future work.

3. Theory

3.1. *Type I error*

In this section we consider the expectation and the variance of the Δ estimator under the assumption of no interference in order to construct a statistical test for whether interference is present. To simplify the analysis, $n_{cr}, n_{cbr}, n_{cr,t}, n_{cr,c}, m_{cbr,t}$ and $m_{cbr,c}$ must be constants, which implies that the clustering of the graph must be balanced. In other words, for any cluster $\mathcal{C}_j \in \mathcal{C}$, $|\mathcal{C}_j| = N/M = n_{cr}/m_{cr} = n_{cbr}/m_{cbr}$. Recall that $S^+_{tc} = \sigma^2\{Y^+(1) - Y^+(0)\}$ is the variance of the differences of the cluster-level outcomes, that σ^2_{cr} is the variance of the difference-in-means estimator under a completely randomized assignment, and σ^2_{cbr} is the variance of the Horvitz–Thompson estimator under a cluster-based randomized assignment.

Theorem 1. *If the stable unit treatment value assumption holds, and every cluster is the same size, then the expectation and variance of the difference-in-differences estimator are*

$$E_{W,Z}(\Delta) = 0, \quad \mathrm{var}_{W,Z}(\Delta) = \sigma^2_{cr} + \sigma^2_{cbr} + \frac{M}{n_{cr}n_{cbr}}S^+_{tc} + O\left(\frac{M^2}{n_{cr}N^2}\sigma^2_{cr}\right).$$

The following corollary is a direct application of Chebyshev's inequality.

Corollary 1. *Let the null hypothesis be that the stable unit treatment value assumption holds and let $\hat{\sigma}^2 \in \mathbb{R}_+$ be any computable quantity from the experimental data which upper-bounds the true variance: $\hat{\sigma}^2 \geqslant \mathrm{var}_{W,Z}(\Delta)$. Suppose that we reject the null if and only if $|\Delta| \geqslant \alpha^{-1/2}\sqrt{\hat{\sigma}^2}$; then if the null hypothesis holds, we reject the null incorrectly with probability no greater than α.*

This result holds for any balanced clustering and for any model of interference because the Type I error assumes the null hypothesis. Another way of rejecting the null is to approximate the test statistic $T = (\hat{\mu}_{cr} - \hat{\mu}_{cbr})(\hat{\sigma}^2)^{-1/2}$ by a normal distribution. In this case, a conservative $(1 - \alpha) \times 100\%$ confidence interval is $(T - z_{\alpha/2}, T + z_{1-\alpha/2})$, where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the $\alpha/2$-quantiles of the standard normal distribution.

3.2. *Variance estimators*

Corollary 1 makes the assumption that $\hat{\sigma}^2$, computable from observable data, is an upper bound of the unknown theoretical variance of the estimator σ^2. We discuss two solutions to finding the smallest possible upper bound $\hat{\sigma}^2$ of the theoretical variance.

The following variance estimator is inspired from Neymann's conservative variance estimator, which upper-bounds the variance of the difference-in-means estimator under a completely

randomized assignment in expectation. Consider the following empirical variance quantities in each treatment bucket of each treatment arm: $\hat{S}_{cr,t} = \sigma^2(Y_i : W_i = 1 \wedge Z_i = 1)$ is the variance of the observed outcomes of the treated units in the completely randomized treatment arm, $\hat{S}_{cr,c} = \sigma^2(Y_i : W_i = 1 \wedge Z_i = 0)$ is the variance of the observed outcomes of the control units in the completely randomized treatment arm. Similarly, let $\hat{S}^+_{cbr,t} = \sigma^2(Y^+_j : \omega_j = 0 \wedge z_j = 1)$ and $\hat{S}^+_{cbr,c} = \sigma^2(Y^+_j : \omega_j = 0 \wedge z_j = 0)$ in the cluster-based randomized arm.

THEOREM 2. *Let $\hat{\sigma}^2$ be the variance estimator defined by*

$$\hat{\sigma}^2 = \frac{\hat{S}_{cr,t}}{n_{cr,t}} + \frac{\hat{S}_{cr,c}}{n_{cr,c}} + \frac{m^2_{cbr}}{n^2_{cbr}}\left(\frac{\hat{S}^+_{cbr,t}}{m_{cbr,t}} + \frac{\hat{S}^+_{cbr,c}}{m_{cbr,c}}\right). \tag{2}$$

If the null hypothesis holds, then $\hat{\sigma}^2$ upper-bounds the theoretical variance of the difference-in-differences estimator Δ in expectation: $E_{W,Z}(\hat{\sigma}^2) \geqslant \mathrm{var}_{W,Z}(\Delta)$. Furthermore, in the case of a constant treatment effect, there exists $\tau \in \mathbb{R}$ such that for all i, $Y_i(1) = Y_i(0) + \tau$, the inequality becomes tight: $E_{W,Z}(\hat{\sigma}^2) = \mathrm{var}_{W,Z}(\Delta)$.

The condition of Corollary 1 will be met only in expectation. This is often deemed sufficient in the literature (Imbens & Rubin, 2015). Due to its simplicity, we use this empirical upper bound in the analysis of our experiments on LinkedIn.

Another common upper bound is obtained by assuming Fisher's null hypothesis of no treatment effect, which posits that $Y_i(1) = Y_i(0)$ for all units i. In particular, this implies that there is no interference. The converse is not true (Rosenbaum, 2007). Under Fisher's null hypothesis, the theoretical formula for the variance is computable from the observed data. Let $S = \sigma^2(Y)$ be the variance of all observed outcomes, and $S^+ = \sigma^2(Y^+)$ be the variance of all observed aggregated outcomes.

THEOREM 3. *Under the null hypothesis of no treatment effect, if all clusters are the same size,*

$$\mathrm{var}_{W,Z}(\Delta) = \frac{n_{cr}}{n_{cr}-1}\frac{M}{M-1}\frac{n_{cr}}{n_{cr,t}n_{cr,c}}S + \left\{1 - \frac{m_{cbr}}{N(n_{cr}-1)}\right\}\frac{m_{cbr}}{m_{cbr,t}m_{cbr,c}}S^+.$$

3.3. *Type II error*

To paraphrase the result stated in Corollary 1, if the rejection region is $\{T \geqslant \alpha^{-1/2}\}$ and $\hat{\sigma}^2 \geqslant \mathrm{var}_{W,Z}(\Delta)$, then the probability of falsely rejecting the null is lower than α. Computing the Type I error is straightforward because the stable unit treatment value assumption holds. The same is not true of the Type II error rate, where a model for the interference between units must be assumed.

In § 2.1 we looked at the expectation of both the $\hat{\tau}_{cr}$ and $\hat{\tau}_{cbr}$ estimators for a completely randomized and a cluster-based randomized assignment, respectively, under a linear model of interference in (1). We complete this analysis by giving the Type II error of the test under this same model of interference. Recall that $\rho_{\mathcal{C}} = N^{-1}\sum_{i=1}^{N} |\mathcal{N}_i \cap \mathcal{C}(i)||\mathcal{N}_i|^{-1}$ is the average fraction of a unit's neighbours contained within its cluster, as originally defined in Lemma 3.

THEOREM 4. *If all clusters are the same size, then under the linear model of interference defined in (1), the expectation of the Δ estimator under the suggested hierarchical design is $E_{W,Z}(\Delta) = \gamma\left[\rho_{\mathcal{C}} + O(n_{cr}^{-1} + m_{cbr}^{-1})\right]$.*

A proof is included in the Supplementary Material, which also includes the computation of the variance of the Δ estimator under the above interference model. The result of Theorem 4 is intuitive: when m_{cbr} and n_{cr} are large, the stronger the interference, with high $|\gamma|$, and the better the clustering, with high $\rho_{\mathcal{C}}$, the larger the expected difference between the two treatment arms.

Knowing the Type II error rate can help determine which clustering of the units is most appropriate. The selection of hyperparameters in clustering algorithms, including the number of clusters to set, can be informed by minimizing the Type II error under plausible models of interference. The optimization program $\max_{M,\mathcal{C}} \rho_{\mathcal{C}}(\hat{\sigma}^2_{\mathcal{C}})^{-1/2}$ depends on the choice of variance estimator $\hat{\sigma}^2_{\mathcal{C}}$ for a clustering $\mathcal{C}$, where $\mathcal{C}$ is composed of M balanced clusters. We discuss a reasonable heuristic in § 6.2 to solving this optimization program, conjectured to be NP hard.

4. Variations of the hierarchical design

4.1. *Bernoulli randomization*

Completely randomized assignment is a well-understood assignment mechanism that avoids degenerate cases, where all units are assigned to treatment and control. However, experimentation platforms at major internet companies are rarely set up to run completely randomized experiments. Instead, these platforms run Bernoulli randomized assignments, which for large sample sizes are intuitively equivalent to completely randomized assignments. We provide a formal explanation for why running a Bernoulli randomized assignment does not affect the validity of our test in practice: the variance of the difference-in-means estimator under the Bernoulli randomized mechanism and the completely randomized mechanism are equivalent up to $O(N^{-2})$ terms.

Theorem 5. *Let* CR *be the completely randomized assignment, assigning exactly n_t units to treatment and $n_c = N - n_t$ to control. Let* BR *be the corresponding re-randomized Bernoulli assignment, assigning units to treatment with probability $p = n_t/N$ and to control with probability $1-p = n_c/N$. For all $N \geqslant 2$ such that $p^N + (1-p)^N \leqslant N^{-2}$, the following upper bound holds:*

$$\left|\mathrm{var}_{Z\sim \mathrm{BR}}(\hat{\tau}) - \mathrm{var}_{Z\sim \mathrm{CR}}(\hat{\tau})\right| \leqslant 5\left[\frac{\sigma^2\{Y(1)\}}{n_t^2} + \frac{\sigma^2\{Y(0)\}}{n_c^2}\right].$$

A proof, which considers a re-randomized Bernoulli assignment scheme that rejects assignments where all units are assigned to treatment or to control, is included in the Supplementary Material.

4.2. *Stratification and subsampling*

One practical concern with our suggested hierarchical design is that if the chosen number of clusters is small, possibly much smaller than the number of units, we run the risk of having strong covariate imbalances between the two treatment arms. In this case, we recommend using a stratified treatment arm assignment. Let each graph cluster be assigned to one of L strata. Within each stratum s, we assign clusters completely at random to treatment arm *cr* and treatment arm *cbr*. Within each stratum s, units in treatment arm *cr* are assigned completely at random to treatment and control, while in treatment arm *cbr*, clusters are assigned completely at random to treatment and control. Let $\hat{\tau}_{cr}(s)$, $\hat{\tau}_{cbr}(s)$ and $\Delta(s)$ be the restrictions of $\hat{\tau}_{cr}$, $\hat{\tau}_{cbr}$ and Δ, respectively, to stratum s. Let $M(s)$ be the total number of clusters in stratum s, and M be the total number of clusters. The stratified Δ' estimator and its empirical variance upper bound estimator can be

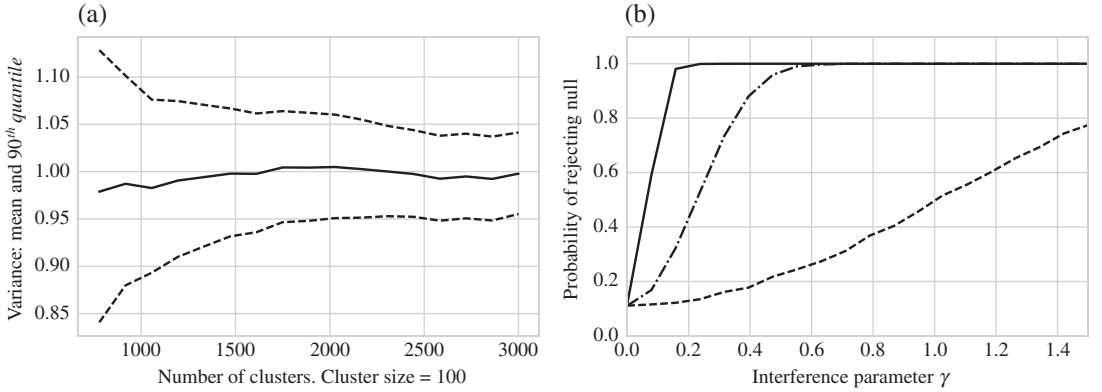

Fig. 1. (a) The expectation (solid line) and 10th and 90th quantiles (dashed lines) of the ratio of the empirical upper bound estimator $\hat{\sigma}^2$ over the true variance $\mathrm{var}_{W,Z}(\Delta)$. (b) Probability of rejecting the null of no interference under the linear interference model in (1) for $\rho_C = 0.4$ (solid line), 0.2 (dot-dashed line) and 0.05 (dashed line).

expressed as a weighted average of $\Delta(s)$ and $\hat{\sigma}^2(s)$,

$$\Delta' = \sum_{s=1}^{L} \frac{M(s)}{M} \Delta(s), \qquad \hat{\sigma}'^2 = \sum_{s=1}^{L} \left\{ \frac{M(s)}{M} \right\}^2 \hat{\sigma}^2(s), \tag{3}$$

where $\hat{\sigma}^2(s)$ is the empirical upper bound of $\mathrm{var}_{W(s),Z(s)}\{\Delta(s)\}$ suggested in (2).

An additional constraint for our suggested design is that online experimentation platforms often need to run multiple experiments simultaneously, with multiple values for treatment and control. As a result, each experiment runs within a segment of the population chosen completely at random, leaving the other units available for other experiments. Since this subsampling might negatively affect the quality of the clustering in the cluster-based randomized treatment arm, we decided against subsampling at the unit level prior to the experiment. In other words, we recommend subsampling at the cluster level, prior to the assignment to treatment arms, rather than at the unit level, when deciding which units to include in the experiment.

5. SIMULATION STUDY

The effect of the clustering on the Type I and Type II errors of the test for interference can be better understood through simulation. From Theorem 2, the bound on the Type I error holds under the assumption that the empirical upper bound for the variance upper-bounds the theoretical variance for the realized assignment $\hat{\sigma}^2(Z) \geqslant \mathrm{var}_Z(\Delta)$, a property that is only guaranteed in expectation. In a first simulation study, we examined how often it held in practice for different numbers of clusters, from 500 to 3000, of fixed size, 100 units. Figure 1(a) reports the expectation and the 10th and 90th quantiles of the ratio of the empirical upper bound $\hat{\sigma}^2$ over the true variance $\mathrm{var}_{W,Z}(\Delta)$. For each point on the x-axis, 5×10^5 simulations were run. The upper bound holds tightly in expectation, but is not an upper bound almost surely. Despite the diminishing returns on the reduction of the confidence intervals from increasing the number of clusters, for a number of clusters greater than 2000, the ratio is in the $(0.95, 1.05)$ range more than 90% of the time.

In a second simulation study, we quantified the Type II error of our test under the linear interference model in (1), fixing the value of the constant parameter to 0 and the direct treatment effect parameter to 1, and choosing as a graph a block model with 40 balanced clusters of 1000

units each. The probability of an edge existing between two units of the graph is a constant cluster-level probability, set consecutively to $(0.01, 0.31)$, $(0.15, 0.45)$ and $(0.3, 0.6)$, denoting the intracluster and intercluster probabilities. These three tuples correspond to values of the graph cut parameter ρ_C equal to $(0.05, 0.2, 0.4)$, respectively, defined in Lemma 3. The higher ρ_C, the fewer edges of the graph are cut. The value of the interference parameter was varied from 0 to 1.4. Figure 1(b) reports the probability of rejecting the null under 1000 simulations. Even with a low value of the graph cut parameter, $\rho_C = 0.05$, the test of interference correctly rejects the null under levels of interference that are of similar magnitude to the direct effect 75% of the time. Furthermore, if $\rho_C \geqslant 0.4$, the test correctly rejects the null for levels of interference that are at least $1/5$th of the magnitude of the direct treatment effect 99.9% of the time.

6. LinkedIn example

6.1. *Experimental set-up*

Google (Tang et al., 2010), Microsoft (Kohavi et al., 2013), Facebook (Bakshy et al., 2014), LinkedIn (Xu et al., 2015) and other major technology companies rely on experimentation to understand the effect of each product decision, from minor user interface changes to major product launches. Due to their extensive reliance on randomized experiments, each company has built mature experimentation platforms. It is an open question as to how many of these experiments suffer from interference effects. By collaborating with the team in charge of LinkedIn's experimentation platform, we tested for interference in several of LinkedIn's many randomized experiments.

Users on LinkedIn can interact with content posted by their neighbours through a personalized feed. Rather than presenting the content chronologically, LinkedIn strives to improve a user's feed by ranking content by relevance. In order to improve user experience, researchers at LinkedIn suggest new feed ranking algorithms and seek to determine the impact of each algorithm on user metrics through a randomized experiment. These metrics may include time spent on the site, engagement with content on the feed, and original content creation. Experimentation on feed ranking algorithms is a typical case where interference between units is present. If a user is assigned to a better feed ranking algorithm, they will interact more with their feed by liking or sharing content more. This in turn affects what her connections see on their own feed. We seek to understand whether or not these network effects are negligible.

To run the experiment we clustered the LinkedIn graph into balanced clusters, stratified the clusters by blocking on cluster covariates, and assigned a subset of clusters to treatment and to control chosen at random, comprising the second treatment arm and treatment bucket assignment. Any unit not already assigned to treatment or control was given to the main experimentation pipeline. A subpopulation of units is first sampled at random, and then assigned to treatment and control using a Bernoulli randomized assignment.

Before applying treatment to units assigned to treatment, we ran covariate balance checks and measured outcomes four and two months prior to the launch of the experiment, on the day of the launch, and again two months after the launch. The number of units per treatment arm was of the order of several million.

6.2. *Graph clustering and cluster stratification*

The main challenge of implementing the proposed test for interference is clustering the graph into clusters of equal size. Only parallelizable algorithms can operate at the scale of the LinkedIn

Table 1. *Results of experiments on a subset of the LinkedIn graph*

Statistic	Pre-treatment	Post-treatment
Δ'	−3.3	−16.4
σ'^2	8.1	8.5
Two-tailed p-value	0.68	0.048

graph. Saveski et al. (2017) performed an extensive experimental evaluation of the state-of-the-art balanced clustering algorithms and found the restreaming version of the Linear Deterministic Greedy algorithm to work best (Nishimura & Ugander, 2013). We ran the parallel version of this algorithm for 30 iterations, setting the number of clusters to $M = 3000$ and a leniency of 1% for the balance constraint, to slightly sacrifice balance for better clustering quality, as it compromised between maximizing the fraction of edges within clusters, 28.28%, and minimizing pre-treatment variance.

As suggested in § 4.2, each cluster is assigned to one of L strata in order to reduce the variance of the estimator and to ensure the balance of cluster-level covariates. Each cluster is described by the number of edges within the cluster, the number of edges with an endpoint in another cluster, and two metrics that characterize users' engagement with the LinkedIn feed, averaged over all users in the cluster. Saveski et al. (2017) found that stratification using balanced k-means clustering worked best (Malinen & Fränti, 2014).

6.3. *Results*

We launched our experimental design on LinkedIn in August 2016. Because of the nature of our intervention on the social nature of the LinkedIn feed, we expected the experiment to suffer from strong interference effects. The primary outcome was the change in a user's engagement over time, $Y_i(t) = y_i(t) - y_i(t-2)$, where $t-2$ takes place two months before date t. As a sanity check, we ran an A/A test on $Y_i(-2) = y_i(-2) - y_i(-4)$, where $t = 0$ is the month the intervention was launched, $t = -2$ takes place two months prior, and $t = -4$ takes place four months prior. As expected, we found no significant interference in the A/A test, with a p-value of 0.68 using the Gaussian assumption from Corollary 1. We then evaluated the presence of interference two months after the launch of the randomized experiment, finding a p-value of 0.048, and concluded that interference was present in the experiment, as reported in Table 1. Outcomes have been multiplied by a constant to avoid disclosing raw numbers, and Δ' and σ' are defined in (3).

ACKNOWLEDGEMENT

Saint-Jacques and Saveski contributed equally to this work while interning at LinkedIn, Sunnyvale, California. The authors wish to thank Guillaume Basse and Dean Eckles, as well as the editor, associate editor and two reviewers, for useful comments. This research was sponsored in part by the National Science Foundation and the Office of Naval Research.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs of Lemma 3, Theorems 1–5 and Corollary 1.

References

Aral, S. & Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Manag. Sci.* **57**, 1623–39.

Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociol. Meth. Res.* **41**, 3–16.

Athey, S., Eckles, D. & Imbens, G. W. (2018). Exact *p*-values for network interference. *J. Am. Statist. Assoc.* **113**, 230–40.

Backstrom, L. & Kleinberg, J. (2011). Network bucket testing. In *Proc. 20th Int. Conf. on World Wide Web*, pp. 615–24.

Baird, S., Bohren, J. A., McIntosh, C. & Ozler, B. (2014). *Designing Experiments to Measure Spillover Effects*. PIER Working Paper No. 15-021.

Bakshy, E., Eckles, D. & Bernstein, M. S. (2014). Designing and deploying online field experiments. In *Proc. 23rd Int. Conf. on World Wide Web*, pp. 283–92.

Basse, G. W. & Airoldi, E. M. (2018). Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika* **105**, 849–58.

Choi, D. (2017). Estimation of monotone treatment effects in network experiments. *J. Am. Statist. Assoc.* **112**, 1147–55.

COMMIT Research Group (1991). Community intervention trial for smoking cessation (COMMIT): Summary of design and intervention. *J. Nat. Cancer Inst.* **83**, 1620–8.

Cornfield, J. (1978). Symposium on CHD prevention trials: Design issues in testing life style intervention randomization by group: A formal analysis. *Am. J. Epidemiol.* **108**, 100–2.

Datta, S., Halloran, M. E. & Longini, I. M. (1999). Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: Randomization by individual versus household. *Biometrics* **55**, 792–8.

Donner, A. & Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *Am. J. Public Health* **94**, 416–22.

Eckles, D., Karrer, B. & Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *J. Causal Inf.* **5**, 23.

Eckles, D., Kizilcec, R. F. & Bakshy, E. (2016). Estimating peer effects in networks with peer encouragement designs. *Proc. Nat. Acad. Sci.* **113**, 7316–22.

Gui, H., Xu, Y., Bhasin, A. & Han, J. (2015). Network A/B testing: From sampling to estimation. In *Proc. 24th Int. Conf. on World Wide Web*, pp. 399–409.

Hong, G. & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Am. Statist. Assoc.* **101**, 901–10.

Hudgens, M. G. & Halloran, M. E. (2008). Toward causal inference with interference. *J. Am. Statist. Assoc.* **103**, 832–42.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge University Press.

Karwa, V. & Airoldi, E. M. (2018). A systematic investigation of classical causal inference strategies under mis-specification due to network interference. *arXiv:1810.08259*.

Katzir, L., Liberty, E. & Somekh, O. (2012). Framework and algorithms for network bucket testing. In *Proc. 21st Int. Conf. on World Wide Web*, pp. 1029–36.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y. & Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proc. 19th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1168–76. Association of Computing Machinery.

Malinen, M. I. & Fränti, P. (2014). Balanced *k*-means for clustering. In *Structural, Syntactic, and Statistical Pattern Recognition. S+SSPR 2014*. Fränti P., Brown G., Loog M., Escolano F., Pelillo M. eds. Lecture Notes in Computer Science, vol 8621. Berlin: Springer.

Manski, C. F. (2013). Identification of treatment response with social interactions. *Economet. J.* **16**, S1–S23.

Nishimura, J. & Ugander, J. (2013). Restreaming graph partitioning: Simple versatile algorithms for advanced balancing. In *Proc. 19th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1106–14. Association of Computing Machinery.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *J. Am. Statist. Assoc.* **102**, 191–200.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *J. Statist. Plan. Infer.* **25**, 279–92.

Saveski, M., Pouget-Abadie, J., Saint-Jacques, G., Duan, W., Ghosh, S., Xu, Y. & Airoldi, E. M. (2017). Detecting network effects: Randomizing over randomized experiments. In *Proc. 23rd Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1027–35. Association of Computing Machinery.

Shakya, H., Stafford, D., Hughes, D., Keegan, T., Negron, R., Broome, J., Mcknight, M., Nicoll, L., Nelson, J., Iriarte, E. et al. (2017). Exploiting social influence to magnify population-level behavior change in maternal and child health: Study protocol for a randomized controlled trial of network targeting algorithms in rural Honduras. *Br. Med. J. Open* **7**, e012996.

SINCLAIR, B., MCCONNELL, M. & GREEN, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *Am. J. Polit. Sci.* **56**, 1055–69.

TANG, D., AGARWAL, A., O'BRIEN, D. & MEYER, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proc. 16th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 17–26. Association of Computing Machinery.

TCHETGEN, E. J. T. & VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Statist. Meth.: Med. Res.* **21**, 55–75.

TOULIS, P. & KAO, E. (2013). Estimation of causal peer influence effects. *Proc. Mach. Learn. Res.* **28**, 1489–97.

UGANDER, J., KARRER, B., BACKSTROM, L. & KLEINBERG, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proc. 19th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 329–37. Association of Computing Mahinery.

XU, Y., CHEN, N., FERNANDEZ, A., SINNO, O. & BHASIN, A. (2015). From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proc. 21st Int. Conf. on Knowledge Discovery and Data Mining*, pp. 2227–36. Association of Computing Machinery.

[*Received on* 11 *April* 2017. *Editorial decision on* 12 *March* 2019]