Biometrics ,

Randomizing over randomized experiments to test for network interference

Summary: We propose an experimental design for testing whether the stable unit value assumption holds, by comparing two different estimates of the total treatment effect obtained through two different assignment strategies: a completely randomized assignment and a cluster-based randomized assignment. We provide a methodology for obtaining these two estimates simultaneously and provide theoretical guarantees for rejecting the null hypothesis that the stable unit value assumption holds without specifying a model of interference. We provide a discussion on how to apply our methodology to large internet experimentation platforms. Finally, we illustrate the proposed multilevel design to a live experiment on the LinkedIn platform.

Key words: Causal inference; Network interference; Spillovers; SUTVA; Randomized designs.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Causal inference (Imbens and Rubin, 2015), and in particular the Neyman-Rubin potential outcomes approach (Rubin, 1974), often relies on a fundamental assumption that the outcome of each unit in the population depends only on the intervention they receive. This assumption is known as the Stable Unit Value Assumption (SUTVA) (Cox, 1958; Rubin, 1990). Unfortunately, in many scenarios, SUTVA does not hold and many fundamental results no longer hold true. Instances of causal questions where interference is present are numerous, including education policy (Hong and Raudenbush, 2012), and viral marketing campaigns (Aral and Walker, 2011; Eckles et al., 2016). In each of these examples, the difference-in-means estimator under Bernoulli randomized assignments is no longer guaranteed to be an unbiased estimator of the Average Treatment Effect (ATE).

The research community has made significant efforts to extend the theory of causal inference to scenarios where SUTVA does not hold. A popular approach to minimizing the effects of interference, cluster-based randomized designs, have been extensively studied, spanning from the early work of (Cornfield, 1978; COMMIT, 1991; Donner and Klar, 2004) to more recent contributions by (Aronow et al., 2013; Eckles et al., 2014). Multi-level designs where treatment is applied with different proportions across the population (Hudgens and Halloran, 2008; Tchetgen and VanderWeele, 2012) are also an important tenet of the literature on improving causal estimates under interference, having been applied to vaccination trials in (Datta et al., 1999) and more recently voter-mobilization campaigns (Sinclair et al., 2012). A recent branch of the literature has developed around various assignment strategies and estimators, beyond cluster-based randomized design or multi-level designs with interesting guarantees under specific models of interference (Backstrom and Kleinberg, 2011; Katzir et al., 2012; Toulis and Kao, 2013; Manski, 2013; Aronow and Samii, 2013; Ugander et al., 2013; Choi, 2014; Basse and Airoldi, 2015; Gui et al., 2015).

Whilst mitigating interference is the end goal of the field of causal inference with network interference, a precursor to that question is to detect whether or not SUTVA holds in the experiments we run. Rosenbaum (Rosenbaum, 2007) was the first to state two sharp null hypotheses which imply SUTVA does not hold. We can obtain exact distribution of network parameters under these restricted null hypotheses. A more recent continuation of this work (Aronow, 2012; Athey et al., 2015) explicitly tackles testing for the non-sharp null that SUTVA holds, by considering an approximate distribution of network effect parameters under SUTVA.

The main focus of our work is to develop a randomized experimental design which allows the experimenter to test for whether SUTVA holds. Thus, our work differs from prior work, which focuses on applying post-experiment analysis methods to existing experimental designs instead of designing new ones. The design that we introduce makes no assumptions on the interference model between units or a graph between units, and has a known bound on the Type I error rate. Most importantly, the proposed design is non-intrusive: it allows the experimenter to analyse the experiment in a classical way, for example to conduct an analysis of the treatment effect with the standard assumptions.

In Section 2, we discuss the theoretical framework of the experimental design and provide certain guarantees on its validity. In Section 3, we discuss some hurdles that we must overcome when applying this framework to major online experimentation platforms. Finally, in Section 4, we present some of the results obtained for an experiment launched in August 2016 on LinkedIn's experimentation platform using our suggested framework. For more details on the experimental implementation and results, we refer the reader to Saveski et al. (2017)

2. Theoretical Framework

2.1 *Two assignment strategies.*

Let $G = (V, E)$ be a graph of N units which we can assign to two possible intervention buckets: treatment ($Z_i = 1$) and control ($Z_i = 0$). Without making the SUTVA assumption, we define each unit's potential outcome as a function of the entire assignment vector $\mathbf{Z} \in \{0,1\}^G$ of units to treatment buckets: $Y_i(\mathbf{Z})$. The causal estimand of interest is the *Total Treatment Effect* (TTE) given by:

$$\mu := \frac{1}{N} \sum_{i \in G} Y_i(\mathbf{Z} = \vec{1}) - Y_i(\mathbf{Z} = \vec{0})$$

We briefly review the notation and main results for two popular experimental designs: the completely randomized (CR) design and the cluster-based randomized (CBR) design. For any vector $\vec{u} \in \mathbb{R}^n$, we let $\overline{u} = \frac{1}{n} \sum_{i=1}^n u_i$ and $\sigma^2(\vec{u}) = \frac{1}{n-1} \sum_{i=1}^n (u_i - \bar{u})^2$. In a *completely randomized* experiment, we sample the assignment vector $\mathbf{Z}$ 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 defined as:

$$\hat{\mu}_{cr} := \overline{Y_t} - \overline{Y_c},$$

where $Y_t := \{Y_i : Z_i = 1\}$ and $Y_c := \{Y_i : Z_i = 0\}$. In a *cluster-based randomized* assignment, we randomize over clusters of units in the graph, rather than individual units. We suppose that each unit in G is assigned to one of m clusters. We sample the cluster assignment vector z uniformly at random from $\{v \in \{0,1\}^m : \sum v_i = m_t\}$, assigning units in cluster j to the corresponding treatment: $Z_i = 1 \Leftrightarrow z_j = 1$ if $i \in j$, 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. We let Y' be the vector of *aggregated* potential outcomes, defined as $Y'_j := \sum_{i \in j} Y_i$, the sum of all outcomes within that cluster. We let Y'_t be the aggregated outcomes of the

treated clusters $Y'_t \coloneqq \{Y'_j : z_j = 1\}$ and Y'_c the aggregated outcomes of the control clusters $Y'_c \coloneqq \{Y'_j : z_j = 0\}$. The aggregated difference-in-means estimator is given by:

$$\hat{\mu}_{cbr} \coloneqq \frac{m}{N}\left(\overline{Y'_t} - \overline{Y'_c}\right).$$

When SUTVA holds, $\hat{\mu}_{cr}$ is an unbiased estimator of the total treatment effect μ under a completely randomized assignment: $E_Z[\hat{\mu}_{cr}] = \mu$. Similarly, when SUTVA holds, $\hat{\mu}_{cbr}$ is an unbiased estimator of μ under a cluster-based randomized assignment. However, when SUTVA does not hold, these result are no longer guaranteed and we expect the estimate of the total treatment effect to be different under each design when interference is present.

2.2 *Randomizing over randomized assignments.*

[Figure 1 about here.]

If it were possible to apply both the completely randomized and cluster-based randomized designs, we could test for interference by comparing the two estimates from each assignment strategy: if the two estimates are significantly different, there is interference. Unfortunately, just as we cannot assign both treatment and control to each unit, we cannot apply both assignment designs on G. We solve this problem by first randomly assigning units to treatment *arms* and within each treatment arm, applying a specific assignment strategy of units to treatment *buckets*. In order to maintain some of the graph structure intact within each treatment arm without sacrificing covariate balance or introducing bias, we suggest to use a cluster-based randomized design to assign units to treatment arms. Once units are assigned to treatment arms, we suggest to apply within each treatment arm either a cluster-based randomized design or a completely randomized design. See Figure 1 for an illustration.

We now set up the formal notation. We cluster graph G into m clusters $\mathcal{J}$. Let m_{cr} and m_{cbr} be the number of clusters we wish to assign to treatment arms *cr* and *cbr* respectively and let n_{cr} and $n_{cbr} = N - n_{cr}$ be the resulting number of units assigned to each arm.

Let $\mathbf{W} \in \{0,1\}^N$ be the assignment vector of units to treatment arm *cr* ($W_i = 1$) and treatment arm *cbr* ($W_i = 0$) and let $\omega \in \{0,1\}^{\mathcal{J}}$ be the corresponding cluster-to-treatment-arm assignment vector. In treatment arm *cr*, let $n_{cr,t}$ and $n_{cr,c}$ be the number of units that we wish to assign to treatment and control respectively. In treatment arm *cbr*, let $m_{cbr,t}$ and $m_{cbr,c}$ be the number of clusters that we wish to assign to treatment and control respectively. Let $\mathbf{Z} \in \{0,1\}^N$ be the assignment vector of units to treatment and control, composed of two parts $\mathbf{Z}_{cr} \in \{0,1\}^{n_1}$ for units in treatment arm *cr* and $\mathbf{Z}_{cbr} \in \{0,1\}^{n_{cbr}}$ for units in treatment arm *cbr*.

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

Though we could imagine re-clustering the graph for step (*iii*), a simpler option from an analytical and methodological perspective is to re-use the same clustering used in step (*i*). We define the two estimates of the causal effect for this experiment as follows:

$$\hat{\mu}_{cr}(\mathbf{W}, \mathbf{Z}) \coloneqq \overline{Y}_{cr,t} - \overline{Y}_{cr,c}, \tag{1}$$

$$\hat{\mu}_{cbr}(\mathbf{W}, \mathbf{Z}) \coloneqq \frac{m_{cbr}}{n_{cbr}} \left(\overline{Y'}_{cbr,t} - \overline{Y'}_{cbr,c} \right), \tag{2}$$

where $Y_{cr,t} \coloneqq \{Y_i : W_i = 1 \wedge Z_i = 1\}$, $Y_{cr,c} \coloneqq \{Y_i : W_i = 1 \wedge Z_i = 0\}$, $Y'_{cbr,t} \coloneqq \{Y'_j : \omega_j = 0 \wedge z_j = 1\}$, $Y'_{cbr,c} \coloneqq \{Y'_j : \omega_j = 0 \wedge z_j = 0\}$.

2.3 *Theoretical guarantees.*

If SUTVA does not hold, the moments of $\hat{\mu}_{cr}$ and $\hat{\mu}_{cbr}$ (cf. Eq. 1–2) are unknown without specifying a model of interference. However, assuming that n_{cr}, n_{cbr}, $n_{cr,t}$, $n_{cr,c}$, $m_{cbr,t}$ and

$m_{cbr,c}$ are constants and not random variables, we can easily compute the first and second-order moment of each estimator under SUTVA. Thus, we require the clustering of the graph to be balanced such that for any cluster $j \in \mathcal{J}$, $|j| = \frac{N}{M} = \frac{n_{cr}}{m_{cr}} = \frac{n_{cbr}}{m_{cbr}}$. Under this assumption, the following theorem holds:

THEOREM 1: *If SUTVA holds, the first and second order moments of the difference-in-difference-in-means estimator are given by:*

$$E_{\mathbf{W},\mathbf{Z}}[\hat{\mu}_{cr} - \hat{\mu}_{cbr}] = 0 \tag{3}$$

$$\begin{aligned} var_{\mathbf{W},\mathbf{Z}}[\hat{\mu}_{cr} - \hat{\mu}_{cbr}] = \lambda_1 \cdot \left(\tfrac{S_1}{n_{cr,t}} + \tfrac{S_0}{n_{cr,c}} - \tfrac{S_{10}}{n_{cr}}\right) + \tfrac{m_{cbr}^2}{n_{cbr}^2}\left(\tfrac{S'_1}{m_{cbr,t}} + \tfrac{S'_0}{m_{cbr,c}} - \tfrac{S'_{10}}{m_{cbr}}\right) \\ + \tfrac{M}{n_{cr}n_{cbr}} S'_{10} + \lambda_2 \cdot \left(\tfrac{\overline{Y^2(1)}}{n_{cr,t}} + \tfrac{\overline{Y^2(0)}}{n_{cr,c}} - \tfrac{\overline{(Y(1)-Y(0))^2}}{n_{cr}}\right) \\ - \lambda_2 \cdot \tfrac{m_{cr}m_{cbr}}{n_{cr}n_{cbr}} \cdot \left(\tfrac{\overline{(Y')^2(1)}}{n_{cr,t}} + \tfrac{\overline{(Y')^2(0)}}{n_{cr,c}} - \tfrac{\overline{(Y'(1)-Y'(0))^2}}{n_{cr}}\right) \end{aligned} \tag{4}$$

where we let $S_1 := \sigma^2(Y_i(1) : i \in G)$ *and* $S_0 := \sigma^2(Y_i(0) : i \in G)$ *be the variance of the potential outcomes and let* $S_{10} := \sigma^2(Y_i(1) - Y_i(0) : i \in G)$ *be the variance of the difference in potential outcomes. Similarly, we defined the variance of the* aggregated *potential outcomes:* $S'_1 := \sigma^2(Y'_j(1) : j \in \mathcal{J})$ *and* $S'_0 := \sigma^2(Y_j(0) : j \in \mathcal{J})$*, as well as the difference in* aggregated *potential outcomes:* $S'_{10} := \sigma^2(Y_j(1) - Y_j(0) : j \in \mathcal{J})$*. Furthermore, let* $\overline{Y^2(1)} := \overline{(Y_i^2(1))}_{i\in G}$, $\overline{(Y')^2(1)} := \overline{(Y'_j(1)^2)}_{j\in\mathcal{J}}$, $\overline{Y^2(0)} := \overline{(Y_i^2(0))}_{i\in G}$ *and* $\overline{(Y')^2(0)} := \overline{(Y'_j(0)^2)}_{j\in\mathcal{J}}$*. Finally, let* $\lambda_1 := \frac{N-1}{N}\frac{n_{cr}}{n_{cr}-1}\frac{M}{M-1}$ *and* $\lambda_2 := \frac{n_{cr}n_{cbr}}{n_{cr}(n_{cr}-1)(M-1)}$.

Note that $\lambda_1 \approx 1$ and that $\lambda_2 << \lambda_1$. The following corollary is a direct application of Chebyshev's inequality.

COROLLARY 1: *Let* $\hat{\sigma}^2 \in \mathbb{R}_+$ *be any computable quantity from the experimental data which upper-bounds the true variance:* $\hat{\sigma}^2 \geqslant var_{\mathbf{W},\mathbf{Z}}[\Delta]$*. Suppose that we reject the null if and only if:*

$$\frac{|\hat{\mu}_{cr} - \hat{\mu}_{cbr}|}{\sqrt{\hat{\sigma}^2}} \geqslant \frac{1}{\sqrt{\alpha}}. \tag{5}$$

Then if SUTVA holds, we reject the null (incorrectly) with probability no greater than α*.*

A proof of Theorem 1 and Corollary 1 is included in the supplementary material. Note that this results hold for any clustering, for any graph, and for any model of intereference. This is not surprising because the theorem states a result on the Type I error of our test, under which we can assume SUTVA. Another way of rejecting the null is to approximate the test statistic $T := \frac{\hat{\mu}_{cr} - \hat{\mu}_{cbr}}{\sqrt{\hat{\sigma}^2}}$ by a normal distribution $\mathcal{N}(0, 1)$. In this case we obtain the following conservative $(1 - \alpha) \times 100\%$ confidence intervals:

$$CI^{1-\alpha}(T) = \left(T - z_{\frac{\alpha}{2}}, T + z_{1-\frac{\alpha}{2}}\right) \tag{6}$$

where $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ are the $\frac{\alpha}{2}$ quantile of the standard normal distribution.

2.4 *Variance estimators*

The main assumption made in Corollary 1 is that the quantity $\hat{\sigma}^2$, computable from observable data, is an upper-bound of the unknown theoretical variance of our estimator σ^2. In this section, we discuss solutions to finding the smallest possible upper-bound $\hat{\sigma}^2$ of the theoretical variance.

One common quantity studied in the causal inference literature is the variance assuming Fisher's Null hypothesis of no treatment effect: $\forall i, Y_i(1) = Y_i(0)$. Note that Fisher's Null states that each units potential outcome under all treatment assignment is the same, which implies the null hypothesis that SUTVA holds, but the converse is not true (Rosenbaum, 2007). However, it is sometimes reasonable to assume that the variance of our estimator under Fisher's null is a good proxy for the theoretical variance under SUTVA. For example, for a completely randomized assigment and treatment effect given by $Y_i(1) = (1+\rho)Y_i(0)+\epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \kappa^2) \perp\!\!\!\perp Y_i(0)$, then

$$\frac{var_1 - var_0}{var_0} = (2\rho + \rho^2)\frac{N_c}{N} + \frac{\kappa^2}{S}\left(\frac{N_c}{N}\right)^2. \quad (7)$$

In other words, for a small treatment effect $\rho = 5\%$, $\kappa^2 << S$ and $\frac{N_c}{N} = \frac{1}{2}$, the variance will change by only 6%. See the supplementary materials for more details. Under Fisher's null, the theoretical formula for the variance becomes computable from the observed data. Let $S := \sigma^2(Y)$ be the variance of all observed potential outcomes, and $S' := \sigma^2(Y')$ be the variance of all observed *aggregated* outcomes. Furthermore, $\overline{Y^2} := \overline{(Y_i^2)_{i\in G}}$ and $\overline{(Y')^2} := \overline{((Y'_j)^2)_{j\in\mathcal{J}}}$.

THEOREM 2: *Under Fisher's null hypothesis of no treatment effect,*

$$var_{\mathbf{W},\mathbf{Z}}[\Delta] = \lambda_1 \cdot \frac{n_{cr}}{n_{cr,t}n_{cr,c}}S + \frac{m_{cbr}^2}{n_{cbr}^2}\frac{m_{cbr}}{m_{cbr,t}m_{cbr,c}}S' + \lambda_2 \cdot \frac{n_{cr}}{n_{cr,t}n_{cr,c}}\left(\overline{Y^2} - \frac{m_{cr}m_{cbr}}{n_{cr}n_{cbr}}\overline{(Y')^2}\right) \quad (8)$$

A proof is included in the supplementary material. It is not always the case that assuming Fisher's null of no treatment effect is reasonable. Another approach which we can take is to compute an empirical approximation of the variance, which upper-bounds the true expectation. Note that the condition of Corollary 1 will be met only in expectation.

THEOREM 3: *We consider the following variance estimator, computable from the observed data:*

$$\hat{\sigma}^2 := \frac{\hat{S}_{cr,t}}{n_{cr,t}} + \frac{\hat{S}_{cr,c}}{n_{cr,c}} + \frac{m_{cbr}^2}{n_{cbr}^2}\left(\frac{\hat{S}'_{cbr,t}}{m_{cbr,t}} + \frac{\hat{S}'_{cbr,c}}{m_{cbr,c}}\right) \quad (9)$$

If SUTVA holds, then the previous quantity upper-bounds the theoretical variance of the Δ estimator in expectation:

$$\mathbb{E}_{\mathbf{W},\mathbf{Z}}\left[\hat{\sigma}^2\right] \geqslant \sigma^2.$$

Furthermore, in the case of a constant treatment effect, $\exists\tau \in ,\ \forall i, Y_i(1) = Y_i(0) + \tau$, the above inequality becomes tight: $E_{\mathbf{W},\mathbf{Z}}[\hat{\sigma}^2] = \sigma^2$.

where we introduced the following empirical variance quantities in each treatment arm and treatment bucket: $\hat{S}_{cr,t} := \sigma^2(Y_i : W_i = 1 \wedge Z_i = 1)$, $\hat{S}_{cr,c} := \sigma^2(Y_i : W_i = 1 \wedge Z_i = 0)$,

$\hat{S}'_{cbr,t} := \sigma^2(Y'_j : \omega_j = 0 \wedge z_j = 1)$, and $\hat{S}'_{cbr,c} := \sigma^2(Y'_j : \omega_j = 0 \wedge z_j = 0)$. A proof can be found in the supplementary material. In other words, the intuitive idea of summing the normalized variances of the potential outcomes in each treatment bucket of each treatment arm, results in an upper-bound in expectation of the variance of the difference-in-difference-in-mean estimator under our suggested hierarchical assignment. Due to its simplicity and lack of strong assumptions, we use the empirical upper-bound of Theorem 3 in our experiments.

2.5 *Understanding the type II error rate.*

To paraphrase the result stated in Corollary 1, if we set our rejection region to $\{T \geqslant \frac{1}{\sqrt{\alpha}}\}$, then if $\hat{\sigma}^2 \geqslant \sigma^2$, the probability under the null of rejecting the null is lower than α. Computing and controlling the Type I error is straightforward because we can make the assumption that SUTVA holds. The same is not true of the type II error rate, where we must posit a model for the interference between units. We assume here a common linear model of interference and compute the resulting type II error of our test under this model. This linear model, sometimes referred to as the Distributional Interactions with Reference Groups (DIRG), assumes the following structure for the outcomes

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 \rho_i + \epsilon_i,$$

where $\rho_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} Z_j$ is the average number of treated friends in unit i's neighbhorhood and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\epsilon \perp\!\!\!\perp \rho_i$.

We now provide some intuition as to what might happen when computing the Δ estimator under this model. Suppose that the graph clustering $\mathcal{J}$ is a perfect clustering (i.e cuts no edges of the graph), then the average treatment effect estimated in the completely randomized treatment arm is approximately β_1 and $\beta_1 + \beta_2$ in the cluster-based randomized treatment arm, such that $\Delta = \beta_2$ under a perfect clustering. Since there is interference only when $\beta_2 \neq 0$, the greater the interference the greater our chances of rejecting the null. In this

sense, our test for interference is sensible. In fact, we can compute the expectation of the Δ estimator under any clustering, as given by the following theorem:

THEOREM 4: *Fix a clustering C. Let $\rho := \frac{1}{N}\sum_{i\in G}\frac{|N(i)\cap C(i)|}{|N(i)|}$ be the average fraction of a unit's neighbors contained in the cluster. Then,*

$$E\left[\Delta\right] = \left(\frac{m_{cbr}}{m_{cbr}-1}\left(\rho - \frac{1}{m_{cbr}}\right) - \frac{m_{cr}}{n_{cr}(n_{cr}-1)}\right)\beta_2$$

A proof is included in the supplementary material. Note that when m_{cbr} and n_{cr} are large: $E[\Delta] \approx \rho \cdot \beta_2$. With the variance remaining the same, this results corroborates the intuitive idea that a clustering of the graph which cuts fewers edges of the graph will have a greater chance of rejecting the null under this model of interference.

Knowing the type II error rate can help us determine which clustering of the graph is most appropriate. The selection of hyper-parameters 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 varies depending on the choice of variance estimator $\hat{\sigma}^2_{\mathcal{J}}$ for a clustering $\mathcal{J}$:

$$\max_{M,\mathcal{J}} \frac{\rho}{\sqrt{\hat{\sigma}^2_{\mathcal{J}}}},$$

where $\mathcal{J}$ is composed of M *balanced* clusters. We discuss a reasonable heuristic in Section 4.2 to solving this optimization program, conjectured to be NP-hard.

3. Practical considerations.

3.1 *Stratification to achieve covariate balance.*

If the chosen number of clusters is small, we run the risk of having strong covariate imbalances in the sampled treatment arm assignment vector $\mathbf{W}$. In this case, we recommend using a stratified treatment arm assignment. Suppose that each graph cluster $j \in \mathcal{J}$ is assigned to one of S strata. We let $G(s)$ be the nodes in the graph which belong to strata s. Within each

strata $s \in [1, S]$, we assign $m_{cr}(s)$ clusters completely at random to treatment arm cr and $m_{cbr}(s)$ clusters to treatment arm 2, denoted by vector $\mathbf{W}(s)$, sampled uniformly at random from vectors $\{\vec{v} \in \{0,1\}^{G(s)} : \sum_j v_j = m_{cr}(s)\}$.

Let $\mathbf{Z}_{cr}(s)$ be the assignment of units in treatment arm cr to treatment buckets within strata s. If $G_{cr}(s)$ is the subgraph of G in strata s assigned to treatment arm cr, $\mathbf{Z}_{cr}(s)$ is chosen uniformly at random from the vectors $\{\vec{u} \in \{0,1\}^{G_{cr}(s)} : \sum_i u_i = n_{cr,t}(s)\}$. Similarly, let $z_{cbr}(s)$ be the assignment of clusters in treatment arm cbr to the treatment buckets within strata s. If $\mathcal{J}_{cbr}(s)$ is the subset of clusters in strata s assigned to treatment arm cbr, $z_{cbr}(s)$ is chosen uniformly at random from the vectors $\{\vec{v} \in \{0,1\}^{\mathcal{J}_{cbr}(s)} : \sum_j v_j = m_{cbr,t}(s)\}$.

We now extend the results given in the previous sections. Let $n_{cbr}(s)$ be the total number of units assigned to treatment arm 2 and $m_{cbr}(s)$ be the total number of clusters assigned to treatment arm cbr within strata s. We can extend the previous estimators of the average treatment effect under stratification. Let $Y_{cr,t}(s) \coloneqq \{Y_i : Z_i = 1 \wedge i \in G_{cr}(s)\}$, $Y_{cr,c}(s) \coloneqq \{Y_i : Z_i = 0 \wedge i \in G_{cr}(s)\}$, $Y'_{cbr,t}(s) \coloneqq \{Y'_j : z_j = 1 \wedge j \in \mathcal{J}_{cbr}(s)\}$ and $Y'_{cbr,c}(s) \coloneqq \{Y'_j : z_j = 0 \wedge j \in \mathcal{J}_{cbr}(s)\}$. The stratified estimators are given by:

$$\hat{\mu}_{cr}(s) \coloneqq \overline{Y_{cr,t}(s)} - \overline{Y_{cr,c}(s)}$$

$$\hat{\mu}_{cbr}(s) \coloneqq \frac{m_{cbr}(s)}{n_{cbr}(s)} \left(\overline{Y'_{cbr,t}(s)} - \overline{Y'_{cbr,c}(s)}\right)$$

We can therefore define the stratified Δ estimator as follows:

$$\Delta(s) \coloneqq \hat{\mu}_{cr}(s) - \hat{\mu}_{cbr}(s)$$

Furthermore, for every strata s, we can define the following quantity $\hat{\sigma}^2(s)$ that upper-bounds $var_{\mathbf{W}(s),\mathbf{Z}(s)}[\Delta(s)]$ in expectation.

$$\hat{\sigma}^2(s) \coloneqq \frac{\hat{S}_{cr,t}(s)}{n_{cr,t}(s)} + \frac{\hat{S}_{cr,c}(s)}{n_{cr,c}(s)} + \frac{m^2_{cbr}(s)}{n^2_{cbr}(s)} \left(\frac{\hat{S}'_{cbr,t}(s)}{m_{cbr,t}(s)} + \frac{\hat{S}'_{cbr,c}(s)}{m_{cbr,c}(s)}\right) \tag{10}$$

The overall Δ estimator can be expressed as an appropriately weighted average of the $\Delta(s)$. We can also define an empirical upper-bound $\hat{\sigma}^2$:

$$\Delta := \sum_{s\in[1,S]} \frac{m(s)}{M}\Delta(s)$$

$$\hat{\sigma}^2 := \sum_{s\in[1,S]} \left(\frac{m(s)}{M}\right)^2 \hat{\sigma}^2(s)$$

The results of the previous sections can be extended straightforwardly. Notably, $E_{\mathbf{W},\mathbf{Z}}[\hat{\sigma}^2] \geqslant var_{\mathbf{W},\mathbf{Z}}[\Delta]$, and the Corollary 1 still holds with the above definitions.

3.2 *Bernoulli assignments.*

The completely randomized assignment is a well-understood assignment mechanism, which 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 (BR) 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 $O(1/N^2)$ terms.

THEOREM 5: *Let CR be the completely randomized assignment, assigning exactly N_{cr} units to treatment and $N_{cbr} := N - N_{cr}$ to control. Let BR be the re-randomized Bernoulli assignment, assigning units to treatment with probability $p := N_{cr}/N$ and to control with probability $1-p = N_{cbr}/N$. For all $N \geqslant 2$ and for all $N_{cr} \in [1, N-1]$ such that $p^N+(1-p)^N \leqslant 1/N^2$, we have the following upper-bound:*

$$|var_{\mathbf{Z}\sim BR}[\hat{\tau}] - var_{\mathbf{Z}\sim CR}[\hat{\tau}]| \leqslant 5\left(\frac{S_{cr}}{N_{cr}^2} + \frac{S_0}{N_{cbr}^2}\right)$$

A proof is included in the supplementary material. Note that we did not seek to optimize the constant term, which can be easily improved. For the analysis, we considered a re-

randomized Bernoulli assignment scheme, which rejects assignments where all units are assigned to treatment ($\vec{\mathbf{Z}} = \vec{0}$) or to control ($\vec{\mathbf{Z}} = \vec{1}$).

3.3 *Simultaneous experiments and sub-sampling.*

Online experimentation platforms often need to run multiple experiments simultaneously. They also regularly change the treatment assignment of their users as products get rolled out or replaced. As a result, the population of interest is randomly divided up into *segments*. When an experiment launches, a segment of the population is chosen at random and a Bernoulli randomized assignment is assigned only to the units in that population, leaving the units in other segments available for other experiments. Formally, there is an extra layer of randomization $\mathbf{X}$ which decides which units are included in the population of interest and will be assigned to either treatment bucket. Whilst we were able to maintain this sub-sampling step in the first treatment arm, we decided against sub-sampling in the second treatment arm, so as to not weaken the effect of the cluster-based randomization.

Formally, we begin by sampling $\mathbf{W}$ and assigning units to treatment arms. For the units assigned to treatment arm cr, we sample $\mathbf{X}$ uniformly at random from all vectors in $\{0,1\}^{n_{cr}}$. We assign units such that $W_i = X_i = 1$ to treatment using a Bernoulli randomized assignment. For units assigned to the second treatment arm, we directly sample the vector z to assign clusters in $\mathcal{J}_{cbr}$ to their treatment bucket, without a sub-sampling step. We redefine $Y_{cr,t} \coloneqq \{Y_i : W_i = 1 \wedge X_i = 1 \wedge Z_i = 1\}$, $Y_{cr,c} \coloneqq \{Y_i : W_i = 1 \wedge X_i = 1 \wedge Z_i = 0\}$, $\hat{S}_{1,t} \coloneqq \sigma^2(Y_{cr,t})$, and $\hat{S}_{cr,c} \coloneqq \sigma^2(Y_{cr,c})$. We let $n_{cr,t}$ and $n_{cr,c}$ be the number of units assigned to treatment and control in treatment arm cr post-subsampling. Taking into account these new definitions, the results of the previous sections still holds. More details can be found in the supplementary material.

As an aside which we leave to future work, testing multiple models simultaneously raises the issue of cross-experiment interference, which is not often the center of attention in the

causal inference with network interference literature. Whilst bounding the type I error of our test assumes SUTVA holds such that Corollary 1 remain valid, the interference from the assignment of units to other models can assuredly impact the type II error of our test.

4. Experimental results on LinkedIn's platform.

4.1 *Experimental set-up*

Today's major Internet companies (e.g. Google (Tang et al., 2010), Microsoft, (Kohavi et al., 2013), Facebook (Bakshy et al., 2014), LinkedIn (Xu et al., 2015)) rely heavily on experimentation to understand the effect of each product decision, from minor UI changes to major product launches. Due to their extensive reliance on randomized experiments, these companies have each built mature experimentation platforms. It is an open question how many of the experiments run on these platforms violate SUTVA. By collaborating with the team in charge of LinkedIn's experimentation platform, we were able to apply the previous theoretical framework to test for interference in one of LinkedIn's many randomized experiments.

Users on LinkedIn can interact with content posted by their friends 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, LinkedIn feed team continually suggests new feed ranking algorithms and seeks to determine the impact of each one on key user metrics through randomized experiments: time spent on the site, engagement with content on feed, original content creation etc. Experimentation on feed ranking algorithms is a typical case where interference between units is present. If a user receives a "good" treatment, they will interact more with the content on their feed which impact the feed of their connections. We seek to understand whether or not these network effects are negligible. To run the experiment, we (i) clustered the LinkedIn graph into

balanced clusters (cf. Section 4.2) (ii) stratified the clusters by blocking on cluster covariates (cf. Section 4.3) (iii) assigned a set number 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 sub-population of units is first (iv) sampled at random (cf. Section 3.3) and then (v) assigned to treatment and control using a Bernoulli randomized assignment (cf. Section 3.2).

Before applying treatment to units assigned to treatment, we ran covariate balance checks and measured outcomes, and again 2 weeks after treatment was applied. The number of units per bucket and per treatment arm was in the order of several million.

4.2 *Clustering the graph.*

The main challenge of implementing the proposed test for interference is clustering the graph into clusters of equal size. Due to the scale of the LinkedIn graph—millions of nodes and billions of edges—we considered only parallelizable streaming algorithms that explicitly enforce balance. We 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 (reLDG) (Nishimura and Ugander, 2013) to work best. To cluster the full LinkedIn graph, we ran the parallel version of reLDG on 300 Hadoop nodes for 30 iterations. We set the number of clusters to $k = \{1000, 3000, 5000, 7000, 10000\}$ and a leniency of 1% for the balance constraint, to slightly sacrifice balance for better clustering quality. In the experiment, we used the clustering obtained by setting $k = 3000$ as it compromises between maximizing the fraction of edges within clusters (28.28%) and minimizing pre-treatment variance. See (Saveski et al., 2017) for more details.

4.3 *Stratifying the clusters.*

As suggested in Section 3.1, we assigned each cluster to one of S strata in order to reduce variance of the estimator, and to ensure balance of cluster-level covariates. Each cluster is described by four covariates: number of edges within the cluster, 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). We experimented with different stratification strategies (Saveski et al., 2017) and found stratification using balanced k-means clustering (Malinen and Fränti, 2014) to work best. We first ran balanced clustering to group the clusters into equally-sized strata and then sampled clusters from each stratum to place them in the different treatment arms and treatment buckets.

4.4 *Experimental results*

We implemented our experimental design on the LinkedIn experimental platform, in August 2016. We considered a subset of the LinkedIn graph, containing several million users. The primary outcome variables of interest is related to a user's engagement with their LinkedIn feed. We measured this outcome prior to launching the experiment and 2 weeks after treatment was applied. These results are reported in Table 1. The experiment failed to reject the null. To reject at 5% (resp. 10%) using Corollary 1, the rejection zone is $\{T \geqslant 4.47\}$ (resp. $\{T \geqslant 3.36\}$) which we did not come close to. However, under the assumption that our test statistic is normally distributed, the post-treatment test statistic for the outcome variable measuring the levels of engagement with the feed decreased. In general, the treatment effect within each arm was not as significant as expected. With roughly 2/3 of edges cut, and a pre-treatment variance upper-bound equal to .08, we would have needed an interference parameter value $\beta_2 = 0.48$ to achieve statistical significance at 5% under the DIRG model of outcomes (cf. Theorem 4). Yet, under the DIRG model of interference, an easy upper-bound for β_2 is 0.17 from Table 1. A more disruptive product launch, with a greater treatment

effect, might be better able to identify interference effects on the LinkedIn experimentation platform. Furthermore, the majority of the variance came from the cluster-based randomized arm which was expected. A better clustering method which minimizes edges cut *and* focuses on minimizing the pre-treatment variance might lead to a statistically significant result.

[Table 1 about here.]

References

Aral, S. and Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science* **57,** 1623–1639.

Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research* **41,** 3–16.

Aronow, P. M., Middleton, J. A., et al. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference* **1,** 135–154.

Aronow, P. M. and Samii, C. (2013). Estimating average causal effects under interference between units. *arXiv preprint arXiv:1305.6156* .

Athey, S., Eckles, D., and Imbens, G. W. (2015). Exact p-values for network interference. Technical report, National Bureau of Economic Research.

Backstrom, L. and Kleinberg, J. (2011). Network bucket testing. In *Proceedings of the 20th international conference on World wide web*, pages 615–624. ACM.

Bakshy, E., Eckles, D., and Bernstein, M. S. (2014). Designing and deploying online field experiments. In *Proceedings of the 23rd international conference on World wide web*, pages 283–292. ACM.

Basse, G. W. and Airoldi, E. M. (2015). Optimal design of experiments in the presence of network-correlated outcomes. *arXiv preprint arXiv:1507.00803* .

Choi, D. S. (2014). Estimation of monotone treatment effects in network experiments. *arXiv preprint arXiv:1408.4102* .

COMMIT (1991). Community intervention trial for smoking cessation (commit): summary of design and intervention. *Journal of the National Cancer Institute* **83,** 1620–1628.

Cornfield, J. (1978). Symposium on chd prevention trials: Design issues in testing life style intervention randomization by group: A formal analysis. *American journal of epidemiology* **108,** 100–102.

Cox, D. R. (1958). Planning of experiments.

Datta, S., Halloran, M. E., and Longini, I. M. (1999). Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: randomization by individual versus household. *Biometrics* **55,** 792–798.

Donner, A. and Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health* **94,** 416–422.

Eckles, D., Karrer, B., and Ugander, J. (2014). Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530* .

Eckles, D., Kizilcec, R. F., and Bakshy, E. (2016). Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences* **113,** 7316–7322.

Gui, H., Xu, Y., Bhasin, A., and Han, J. (2015). Network a/b testing: From sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web*, pages 399–409. ACM.

Hong, G. and Raudenbush, S. W. (2012). Evaluating kindergarten retention policy. *Journal of the American Statistical Association* .

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association* pages 832–842.

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

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

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., and Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1168–1176. ACM.

Malinen, M. I. and Fränti, P. (2014). Balanced k-means for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 32–41. Springer.

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

Nishimura, J. and Ugander, J. (2013). Restreaming graph partitioning: simple versatile algorithms for advanced balancing. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1106–1114. ACM.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association* **102,**.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66,** 688.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference* **25,** 279–292.

Saveski, M., Pouget-Abadie, J., Saint-Jacques, G., Duan, W., Ghosh, S., Xu, Y., and Airoldi, E. (2017). Detecting network effects: Randomizing over randomized experiments.

Manuscript .

Sinclair, B., McConnell, M., and Green, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science* **56,** 1055–1069.

Tang, D., Agarwal, A., O'Brien, D., and Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 17–26. ACM.

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

Toulis, P. and Kao, E. (2013). Estimation of Causal Peer Influence Effects. In *International Conference on Machine Learning*.

Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337.

Xu, Y., Chen, N., Fernandez, A., Sinno, O., and Bhasin, A. (2015). From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2227–2236. ACM.

Supplementary Material

The supplementary web appendix have been made available at the following website

`https://jean.pouget-abadie.com/suppl.pdf`

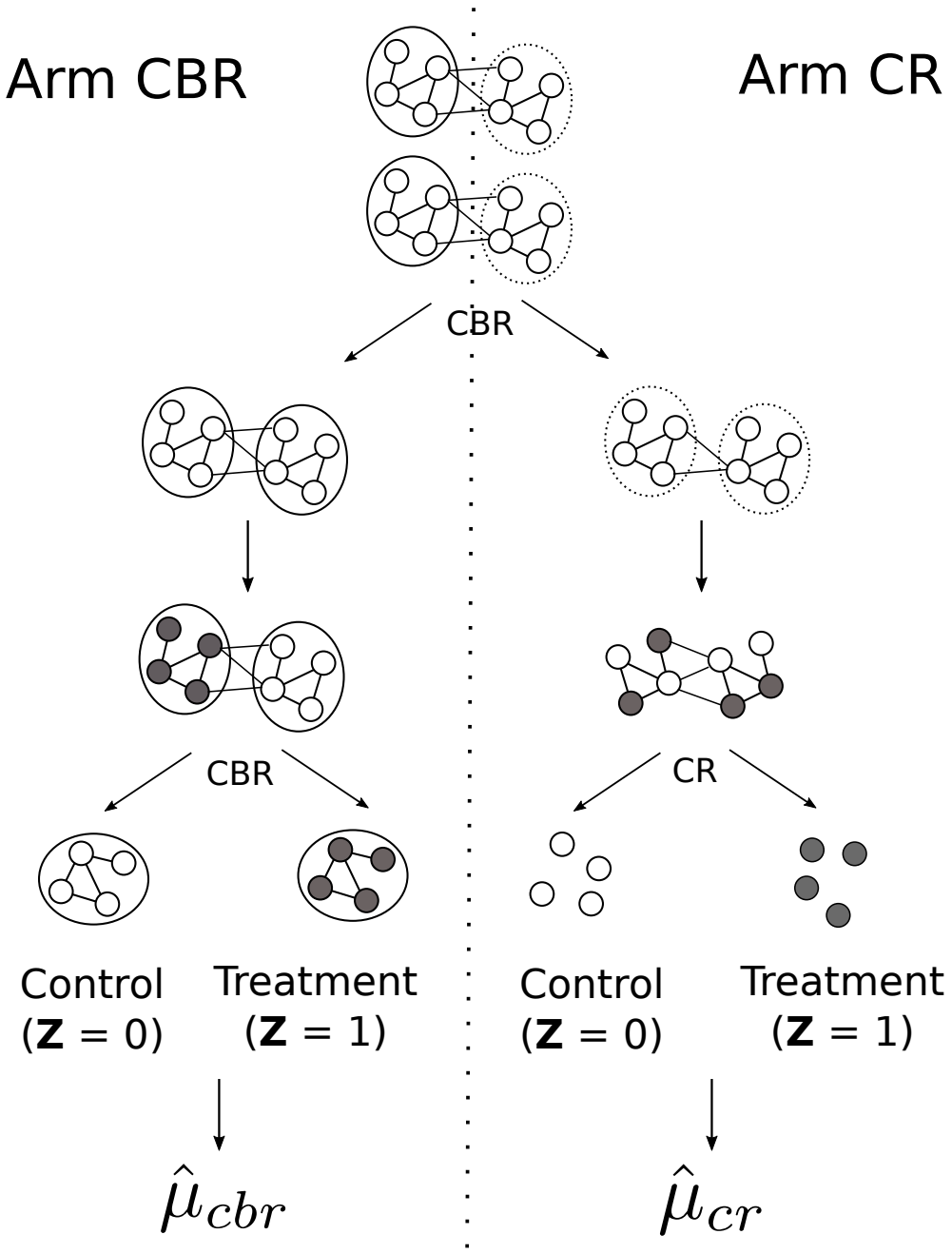

Figure 1. Suggested hierarchical assignment.

Statistic	Formula	Pre-treatment	Post-treatment
BR delta	$\sum_s \frac{m(s)}{M} \cdot \left(\overline{Y_{br,t}(s)} - \overline{Y_{br,c}(s)}\right)$	-0.0261	0.0432
CBR delta	$\sum_s \frac{m(s)}{M} \cdot \frac{m_{cbr}(s)}{n_{cbr}(s)} \cdot \left(\overline{Y'_{cbr,t}(s)} - \overline{Y'_{cbr,c}(s)}\right)$	0.0638	0.1653
Delta of Deltas	$\sum_s \frac{m(s)}{M} \cdot \Delta(s)$	-0.0899	-0.1221
Upp. bound of BR std	$\sqrt{\sum_s \left(\frac{m(s)}{M}\right)^2 \cdot \left(\frac{\hat{S}_{br,t}(s)}{n_{br,t}(s)} + \frac{\hat{S}_{br,c}(s)}{n_{br,c}(s)}\right)}$	0.0096	0.0098
Upp. bound of CBR std	$\sqrt{\sum_s \left(\frac{m(s)}{M}\right)^2 \cdot \left(\frac{m_{cbr}(s)}{n_{cbr}(s)}\right)^2 \cdot \left(\frac{\hat{S}_{cbr,t}(s)}{m_{cbr,t}(s)} + \frac{\hat{S}_{cbr,c}(s)}{m_{cbr,c}(s)}\right)}$	0.0805	0.0848
Upp. bound of std	$\sqrt{\sum_s \left(\frac{m(s)}{M}\right)^2 \cdot \hat{\sigma}^2(s)}$	0.0811	0.0856
Test statistic		-0.0111	-0.0143
2-tailed p-value		0.2670	$\mathbf{0.1530}$

Table 1

Results of experiments on subset of the LinkedIn graph. BR, Bernoulli randomized; CBR, cluster-based randomized. The final row displays the two-tailed p-value of the test statistic T under assumption of normality $T \sim \mathcal{N}(0,1)$. Outcomes have been multiplied by a constant to avoid disclosing raw numbers.