Network A/B Testing: From Sampling to Estimation

Huan Gui[†] Ya Xu[‡] Anmol Bhasin[‡] Jiawei Han[†]
[†] University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA
[‡] LinkedIn Corporation, Mountain View, CA 94043 USA
[†]{huangui2, hanj}@illinois.edu [‡]{yaxu, abhasin}@linkedin.com

ABSTRACT

A/B testing, also known as bucket testing, split testing, or controlled experiment, is a standard way to evaluate user engagement or satisfaction from a new service, feature, or product. It is widely used in online websites, including social network sites such as Facebook, LinkedIn, and Twitter to make data-driven decisions. The goal of A/B testing is to estimate the treatment effects of a new change, which becomes intricate when users are interacting, *i.e.*, the treatment effects of a user may spill over to other users via underlying social connections. When conducting these online controlled experiments, it is a common practice to make the Stable Unit Treatment Value Assumption (SUTVA) that each individual's response is affected by their own treatment only. Though this assumption simplifies the estimation of treatment effects, it does not hold when network interference is present, and may even lead to wrong conclusion.

In this paper, we study the problem of network A/B testing in real networks, which have substantially different characteristics from the simulated random networks studied in previous works. We first examine the existence of network effects in a recent online experiment conducted at LinkedIn; Secondly, we propose an efficient and effective estimator for Average Treatment Effect (ATE) considering the interference between users in real online experiments; Finally, we apply our method in both simulations and a real world online experiment. The simulation results show that our estimator achieves better performance with respect to both bias and variance reduction. The real world online experiment not only demonstrates that large-scale network A/B test is feasible but also further validates many of our observations in the simulation studies.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Experimental design

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741081.

Keywords

A/B testing; controlled experiments; network A/ B testing; design of experiments; peer effects; network effects; balanced graph partition.

1. INTRODUCTION

A/B testing, also called controlled experimentation, is widely used in many consumer facing web technology companies to guide product development and data-driven decisions, including Amazon, eBay, Etsy, Facebook, Google, Groupon, LinkedIn, Microsoft, Netflix and Yahoo. It has become the gold standard for testing out new product strategies and approaches [14, 13].

The theory of A/B test is simple and dates back to Sir Ronald A. Fisher's experiments at the Rothamsted Agricultural Experimental Station in England in the 1920s [29]. Rubin causal model [23], a standard machinery of testing framework, is usually adopted in conducting and analyzing A/B tests. A key assumption made in Rubin causal model is the *Stable Unit Treatment Value Assumption* (SUTVA), which states that the behavior of each user in the experiment depends only on their own treatment and not on the treatments of others.

However, in a social network setting, a user's behavior is likely impacted by that of his/her social neighborhood [12, 27, 28]. In most cases, a user would find a new feature more valuable and hence more likely to adopt it if more of his/her neighbors adopt it. For example, video chat is a useless feature unless one's friends use it too. The phenomenon that an individual's behavior has a non-trivial effect on his/her social neighborhood is called *network effects*, also known as *social interactions*, *peer influence*, or *social interference* [2, 8]. In an A/B experiment, this implies that if the treatment has a significant impact on a user, the effect would spill over to his/her social circles, regardless whether his/her neighbors are in treatment or control.

To understand how network effects introduces challenges for A/B testing, consider the *Average Treatment Effect* (ATE), a primary quantity of interest defined as the difference of the average outcomes between applying treatment to the *entire* user population and applying control to the *entire* user population. Define $\mathcal{U}$ as the set of users and $|\mathcal{U}| = N$, where N is the total number of users. Let $\boldsymbol{Z} \in \mathcal{M}^N$, a vector of length N, as the experiment assignment vector for all users, where $\mathcal{M}$ is the set of variants. Without loss of generality, for the rest of the paper, we assume the experiment only has two variants that $\mathcal{M} = \{0, 1\}$, where 1 stands for treatment and 0 for control. We let $Y_i(\boldsymbol{Z} = \mathbf{z})$ be the response func-

tion of user i given $\boldsymbol{Z} = \mathbf{z}$. The response can be any user metrics the experiment tries to optimize, such as number of pageviews or clicks. The ATE can be expressed as

$$\delta(\mathbf{1}, \mathbf{0}) = \frac{1}{N} \cdot \sum_i \mathbb{E}\big[Y_i(\boldsymbol{Z} = \mathbf{1}) - Y_i(\boldsymbol{Z} = \mathbf{0})\big], \quad (1)$$

where $\mathbf{1}$, and $\mathbf{0}$ are the treatment assignment where all users receive the control variant and treatment variant respectively.

Of course, a user can only receive one treatment at a time in reality. So in classical A/B experiments, we randomly select N_0 users to receive the control variant, and N_1 users to receive treatment. The ATE can then be estimated by

$$\hat{\delta} = \frac{1}{N_1} \cdot \sum_{\{i;\mathbf{z}_i=1\}} Y_i(\boldsymbol{Z} = \mathbf{z}) - \frac{1}{N_0} \cdot \sum_{\{i;\mathbf{z}_i=0\}} Y_i(\boldsymbol{Z} = \mathbf{z}), \quad (2)$$

where $\mathbf{z}_i$ is the experiment assignment of user i. Under SUTVA and the Law of Large Numbers [11], we have $\hat{\delta} \to \delta(\mathbf{1}, \mathbf{0})$ when $N_1, N_0 \to \infty$. However, when there are network effects, STUVA is no longer valid, and the convergence does not hold any more.

This poses a special challenge for running A/B tests in many online social and professional networks like Facebook, Twitter and LinkedIn. Many features tested there through A/B experiments are likely to have network effects. For example, a better recommendation algorithm in treatment for the People You May Know module on LinkedIn encourages a user to send more invitations. However, users who receive such invitations can be in the control variant and when they visit LinkedIn to accept the invitation they may discover more people they know. If the primary metric of interest is the total number of invitations sent, we would see a positive gain in both the treatment and the control groups. The ATE estimated ignoring network effects would be biased and not fully capturing the benefit of the new algorithm. Such bias exists in testing almost any features that involve social interactions, which is truly ubiquitous in a social network environment.

In this paper, we define the problem of A/B testing when there are network effects as *Network A/B Testing*. Specifically, we study the problem in large real social networks, which have substantially different properties and characteristics than simulated random networks studied in previous works. Our work aims at bridging the gap among theoretical analysis on casual analysis in the Statistical literature, recent works on network bucket testing [3, 8, 28], and real-world applications of online controlled experiments. Our main contributions are as follows:

1. As far as we know, we are the first to extensively study the problem in real social networks.
2. We propose a simple yet efficient network sampling algorithm that is able to overcome an important challenge presented in sampling real social networks.
3. We propose a new estimation model that not only generalizes the existing methods, but also produces smaller bias and variance in our extensive simulation studies.
4. We are the first to study how the overall traffic split between treatment and control impacts the performance of various ATE estimators.
5. We run a large-scale network A/B test with a real application at LinkedIn. The results from the experiment not only help us compare the various models but also further validate many of the observations from the simulation studies.

The rest of the paper is organized as follows. Related work is discussed in Section 2. In Section 3 we take a close look at possible network effects in a recent A/B experiment conducted at LinkedIn to motivate our study. Section 4 is where we introduce our framework for network A/B testing and propose our new network sampling and estimation methods. Extensive simulations and online experimental results are included in Section 5. Section 6 concludes our study, summarizing our approach and motivating future work in this area.

2. RELATED WORK

In this section, we will introduce some related work on network A/B testing, mainly including two parts, interference analysis in Statistics, and network bucket testing in Computer Science.

There are two parts of related work on interference, one is based on group-level interference analysis, where there is interference within each group, and no interference across groups [10, 22, 24, 26]; the other one is based on unit-level interference analysis, where interference between any two units may be non-trivial [2, 8, 16, 27, 28]. [22] explores methods of inverting distribution-free randomization tests for Fisher's sharp null hypothesis based on group-level interference analysis. [24] considers the potential bias for ATE estimation when SUTVA is assumed, and further defines several causal estimands including direct and indirect effects that might be identifiable. This work is further extended by [10] by defining direct, indirect, total and overall causal effects, and relationship between these estimands are established. Hudgens *et al.* [10] also propose unbiased estimators of the proposed estimands by a two-stage randomization procedure experimental design that first perform randomization at the group level, then at the individual level within groups. [26] gives conservative variance estimators, and proposes a new framework of finite sample inference analysis and inverse probability weighting estimators considering interference.

Regarding arbitrary interference of known form between units, [2] proposes randomization based methods to estimate ATE, and further refines the estimator by covariance adjustment, besides analysis of conservative estimators of the ATE estimators. [16] explores the identification of users' responses under different constraints and assumptions, considering that social interference will influence users' behaviors. To quantify the causal effect of peer influence, [27] introduces new randomization based causal estimand of peer influence, by extending the potential responses with social interference. [8, 28] focus on estimating the ATE in networks, in which a new randomization scheme is proposed, called graph clustering randomization. Extensive simulations are conducted on random networks. We further this work by studying the ATE estimation on *real* networks, which have drastically different topologies and structures.

Another line of research that is quite related to our work is network bucket testing [3, 12], in which the new feature will take effect only if some minimal number of treated users' social neighbors are also in treatment. [3] proposes a walk-

based sampling method to generate the core set of users which are internally well-connected while approximately uniform over the whole network. This work is generalized by [12], which introduces a general framework for network A/B testing, which first generate the core set based on different algorithms, and then gives out the corresponding variance bound based on the core set generation function.

Our work takes both the sampling (design) and estimation (analysis) into consideration, by proposing a realistic sampling method that works well in real networks and a new estimation framework that generalizes the existing ones.

3. NETWORK EFFECTS IN REAL EXPERIMENT

There is a line of research studying social interference, also known as information diffusion [6, 9, 21]. Whenever information can be propagated from a user to his/her social neighborhood, there is likely to be network effects. In particular, we are interested in studying how such information propagates between the treatment and control groups in an A/B test setting. With that in mind, we first take a close look at a real A/B experiment recently conducted at LinkedIn on homepage Feeds. By incorporating components that are specific to one's social neighborhood, we show using a linear model that there are significant network effects present in this experiment.

3.1 Feed Experiment

Feed is an important part of LinkedIn homepage experience. It provides users with stories that they may be interested in reading, updates from their network (such as a job change), and more. One can click on a feed, or interact with it through social gestures such as "like", "comment" or "share".

The Feed team strives to surface the most relevant items to users by constantly testing new recommendation algorithms. One important evaluation criterion is whether the new algorithm has improved the total number of user interactions with the feeds. In a recent experiment, users are randomly split into two equal groups, one receives the production algorithm (control) and the other receives a new experimental algorithm (treatment) that is supposed to provide more relevant feeds.

Because the sampling is uniformly random, users in treatment have neighbors in both treatment and control. Therefore, users in the control group will not only receive the set of recommended feeds from the control recommendation algorithm, but also the set of feeds shared/liked from their neighbors in treatment group, and vice versa. If the new algorithm is indeed better, and users are interacting more with the more relevant content discovered in their feeds, those good content pieces will leak into the control group and hence makes the control algorithm look better that it actually is.

In the following section, we will show that such concerns of information "leaking" is not vacuous and there is indeed significant network effects present in this feed experiment. We verify by extending the classical framework used to estimate ATE to incorporate components that are specific to one's social neighborhood.

3.2 Presence of Network Effect

Two-sample t-test is the most commonly used framework to analyze online experiment [7]. Suppose we are interested in some response metric Y (e.g. pageviews per user). The null hypothesis is that treatment and control have the same mean for this metric (i.e. ATE $= 0$) and the alternative is that they do not (i.e. ATE $\neq 0$) . The t-test is based on the t-statistic

$$\frac{\hat{\delta}}{\sqrt{\text{var}\left(\hat{\delta}\right)}}, \tag{3}$$

and under SUTVA, $\hat{\delta}$ defined in (2) is an unbiased estimator of the ATE and the t-statistic is a normalized version of that estimator. This framework is in fact equivalent to the following linear model

$$Y_i(\boldsymbol{Z}) = \alpha + \beta \boldsymbol{Z}_i \tag{4}$$

where $\boldsymbol{Z}_i \in \{0, 1\}$ is user i's experiment assignment and the least square estimator for β turns out to be $\hat{\delta}$.

In order to take into consideration of possible network effects, we introduce two additional components to the linear model above, social interference and homophily [15, 17]. The social interference component aims at capturing the "spillover" treatment effects from one's neighbors, and is therefore modeled based on the total number of treated neighbors, *i.e.*, $\boldsymbol{A}_{\cdot i}^\top \boldsymbol{Z}$, where $\mathbf{A}$ is the adjacency matrix, and $\boldsymbol{A}_{\cdot i}$ is the ith column of $\mathbf{A}$. The homophily refers to the observation that one's social network is homogeneous considering different sociodemographic, behavioral, and intrapersonal traits, *i.e.*, people are more likely to connect with others who are similar (birds of a feather). To this end, a user's behavior can be partially explained by the behavior of his/her neighbors. Therefore, homophily is approximated by the average behavior of i's neighborhood, *i.e.*, $\boldsymbol{A}_{\cdot i}^\top \boldsymbol{Y}/\mathbf{D}_{ii}$, where $\mathbf{D}$ is the diagonal matrix that $\mathbf{D}_{ii} = \sum_{j=1}^{N} A_{ij}$. Putting the three pieces together, we have

$$Y_i(\boldsymbol{Z}) = \alpha + \beta Z_i + \gamma \boldsymbol{A}_{\cdot i}^\top \boldsymbol{Z} + \eta \boldsymbol{A}_{\cdot i}^\top \boldsymbol{Y}/D_{ii}. \tag{5}$$

Note that linear additive models have been widely used in other works on causal analysis [15, 27]. The three parameters β, γ and η aim at capturing treatment effects, network effects and homophily respectively. Under this simple linear model, we can estimate the size of each effect and test whether each effect is statistically significant. The results are shown in Table 1 where we can see that users' responses are positively correlated with treatment effects (*primary effects*), network effects, and homophily.

Metric	β	γ	η
# of interactions	0.0486	0.1252	0.0626

Table 1: The linear regression results of the feed experiment. All the p-value are $< 10^{-16}$, and thus omitted.

It is important to point out that this experiment is conducted with fully uniform random sampling, and hence the number of treated neighbors and the neighborhood size are highly correlated. This means that the network effects we measure here could be confounded with one's popularity in

the network. To control for that, we have also tried a model that includes one's neighborhood size as the fourth component, as follows:

$$Y_i(\boldsymbol{Z}) = \alpha + \beta Z_i + \gamma \boldsymbol{A}_{\cdot i}^\top \boldsymbol{Z} + \eta \boldsymbol{A}_{\cdot i}^\top \boldsymbol{Y}/D_{ii} + \kappa D_{ii}.$$

The coefficient γ for the network effects stays significantly positive. Also note that because the sampling is done uniformly random, every user in the experiment has about the same percentage of neighbors in treatment. This means that we cannot expect to remove such confounding effects by using the percent of treated neighbors instead of the absolute number.

4. NETWORK A/B TESTING

The framework of A/B testing usually involves two parts: sampling, which determines who gets what treatment; and estimation, which provides a framework to compute the ATE.

Uniform random sampling is the most commonly used sampling method in most web facing applications. It is simple and sufficient in most cases when the sample size is large. However, as we have mentioned in the last section, uniform random sampling makes it difficult to separate out the contribution of the network effects from other confounding factors such as neighborhood size. To this end, we look for a sampling solution in Section 4.1 that is able to take into consideration the network structure. The idea is to remove information diffusion between treatment and control groups as much as we can. However, as we will see, network sampling itself poses special challenges in *real* social networks because of their sizes and structures. We propose a sampling method that is able to overcome several of these challenges in Section 4.1.2. With the new sampling algorithm, we then study the problem of estimation, where we propose an exposure model and the corresponding estimators in Section 4.2 and show that they are a generalization of the existing estimators.

4.1 Network Sampling

Sampling is the process where we decide which users get assigned into what variant. At a high level, we try to split users into treatment and control so that there is as little information flow between the variants as possible. The sampling scheme we adopt here is called *cluster randomized sampling*, which is a sampling method where clusters of users are randomized together [18, 20]. It is a two-stage procedure:

1. Partition the users into clusters;
2. Treat each cluster as a unit and randomize at the cluster level, so that all the users in the same cluster are assigned to the same variant.

The advantages of cluster randomized sampling include the ability to study the interference within the same cluster and the ability to control for "contamination" across clusters [8]. In this section, we will start with a recent proposal called ϵ-net [28]. We examine its performance in a real social network. We then propose a new sampling algorithm that can overcome an important limitation and hence reduce the bias of the ATE estimator.

4.1.1 Graph Cluster Randomization

Graph clustering, also known as community detection, graph cutting, and graph partition, is a well-researched area [1, 4, 19]. However, clustering real social networks is challenging due to their special structure and topology. [28] proposes a local clustering algorithm, ϵ-net, that overcomes some of those challenges and the idea is as follows:

1. Find k cluster centers that are at least $2\epsilon - 1$ apart;-
2. Assign the nodes to their nearest center, and break the tie randomly if needed.

To study how this clustering algorithm works on real networks, we extract a sub-network from LinkedIn's social graph by taking all the users who have ever worked at a leading internet technology company with a global presence. The sub-network (employee network) is constructed by extracting the connections among them. After removing nodes with no connections, we get a network of about 70K users. The basic statistics of this sub-network are given in Table 2. We

Nodes #	Edges #	max$\{d_i\}$	mean $\{d_i\}$	var$\{d_i\}$
7.26e4	2.88e6	3997	39.67	1.27e4

Table 2: Basic statistics of a employee network.

then apply the 3-net clustering algorithm to partition this sub-network into 100 shards. The distribution of the cluster sizes in logarithmic scale is given in Figure 1.

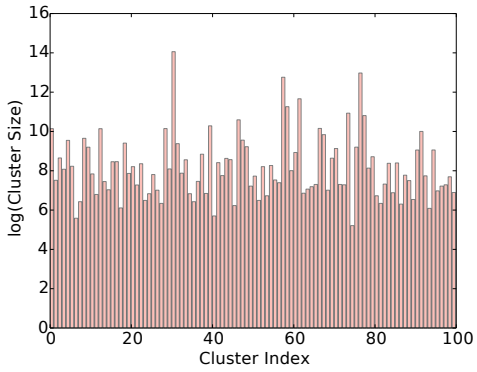


Figure 1: Cluster sizes of 3-net Network Clustering. Y axis specifies the $\log_2$ of cluster sizes.

It is not surprising to find out that the cluster sizes vary quite a lot. Intuitively, the ϵ-net algorithm constructs clusters of radius at least $\epsilon - 1$ around each center node, which implies that the size of each cluster is largely determined by the degrees of the center node. Because degrees are heterogeneous in real networks, so are the cluster sizes. But how would this influence the estimation of ATE?

According to bias analysis in [18], the ATE estimator given in (2) has the following bias

$$\mathbb{E}[\hat{\delta}] - \delta = -\frac{N_c}{N}\left[\frac{1}{|\mathcal{J}_1|}\mathrm{Cov}\left(\frac{\sum_{j\in\mathcal{J}_1}\sum_{i=1}^{n_j} Y_{ij}}{\sum_{j\in\mathcal{J}_1} n_j}, \sum_{j\in\mathcal{J}_1} n_j\right) - \frac{1}{|\mathcal{J}_0|}\mathrm{Cov}\left(\frac{\sum_{j\in\mathcal{J}_0}\sum_{i=1}^{n_j} Y_{ij}}{\sum_{j\in\mathcal{J}_0} n_j}, \sum_{j\in\mathcal{J}_0} n_j\right)\right],$$

where $\mathcal{J}_1$ ($\mathcal{J}_0$) is the set of clusters that are assigned to treatment (control), $N_c = |\mathcal{J}_0| + |\mathcal{J}_1|$ is the number of clusters, Y_{ij} is the response of i-th user in the j-th cluster, n_j is the size of cluster j, and $\text{Cov}(\cdot,\cdot)$ is the covariance function.

To get an unbiased estimator, we need to have $\mathbb{E}[\hat{\delta}]-\delta = 0$, *i.e.*,

$$\begin{aligned} \text{Cov}\left(\frac{\sum_{j\in J_1}\sum_{i=1}^{n_j} Y_{ij}}{\sum_{j\in J_1} n_j}, \sum_{j\in J_1} n_j\right) = 0, \\ \text{Cov}\left(\frac{\sum_{j\in J_0}\sum_{i=1}^{n_j} Y_{ij}}{\sum_{j\in J_0} n_j}, \sum_{j\in J_0} n_j\right) = 0. \end{aligned} \tag{6}$$

There are two scenarios under which (6) can hold. Scenario (i): users' responses Y_{ij}'s are independent of cluster size n_j; Scenario (ii): n_j's are constant. When there are network effects, we expect users in larger treatment clusters to get more social influence, and hence scenario (i) is unlikely to hold. Therefore, we need a network sampling method that is able to produce equal size clusters. Based on this observation and the large scale of real online social networks, we propose a simple yet efficient balanced graph partition algorithm, called *randomized balanced graph partition*.

4.1.2 Randomized Balanced Graph Partition

In this section, we introduce the randomized balanced graph partition algorithm with the goal of producing clusters that are of equal size and that can be computationally feasible to be applied to real, large-scale social networks.

We start with an initial feasible partition where all the clusters are of the same size. Our goal is to maintain the balance of the cluster sizes while maximize the number of edges within each cluster. To do that, we iteratively update the cluster label of each node based on the majority vote from its neighbors. The balance of the cluster sizes is achieved by swapping the cluster label of a pair of nodes, instead of simply changing them. The challenge, of course, is to decide which and how many pairs of labels should be swapped.

It is worth noting that the balanced graph partitioning problem is proven to be NP-complete even for the case with only two clusters [1]. We propose a greedy algorithm with randomization, which is effective and efficient and can be easily parallelizable. Our algorithm alternates the two steps below till convergence:

1. *Label Propagation.* Define "gain" as an increase of the number of edges between nodes of the same cluster. For every pair of nodes, switch their cluster labels in a greedy way until the gain cannot be increased any more.

2. *Random Shuffling.* Randomly select ζ% pairs of nodes, and swap their labels.

By randomly shuffling the cluster labels, we can break the local optimum, and increase the number of links within the same cluster. As shown in Table 3, we can improve the baseline label propagation by 8.9% with shuffling. In experiments, we set $\zeta = 5$[1]. Since the sub-network is not very large, we are also able to compare our proposal with

[1]The choice of ζ may be of independent interest. However it is beyond the scope of this paper.

Methods	LP	RSLP	MM
# of links in clusters ($\times 10^6$)	2.161	2.355	2.359

Table 3: Clustering results for **L**abel **P**ropagation (LP) and **R**andom **S**huffling on **L**abel **P**ropagation (RSLP), as well as one based on **M**odularity **M**aximization (MM).

Modularity Maximization [5][2], which is known to have good performance but computationally expensive for large networks. Our proposed method achieves comparable results yet is scalable and can be easily parallelized.

4.2 Estimation

In this section, we study the problem of how to estimate ATE. To do that, we need to first decide how we model users' exposure to network effects. As we will see, the assumptions of existing exposure models are hard to satisfy in real networks and under the new sampling scheme we proposed in Section 4.1.2.

To address this problem, we propose a new exposure model that we describe in Section 4.2.4 and the corresponding estimators. Through a linear additive model, we are able to show that our exposure model and estimator is a generalization of the existing ones.

4.2.1 Framework

Suppose $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N)'$ where $\mathbf{z}_i \in \{0,1\}$, and the probability of seeing a particular value of $\mathbf{z}$ is $p_{\mathbf{z}}$. Let $\mathcal{Z} = \{\mathbf{z} : p_{\mathbf{z}} > 0\}$, so that $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_N)'$ is a random vector with support $\mathcal{Z}$ and probability $\mathbb{P}(\mathbf{Z} = \mathbf{z}) = p_{\mathbf{z}}$. Define a unit-specific onto function that maps an assignment vector and unit specific traits $\xi_i \in \mathcal{E}$ (such as one's local neighborhood structure) to an *expected* user response $f : \mathcal{Z} \times \mathcal{E} \to \Delta$. The codomain of Δ contains all the expected treatment responses that might be induced in the experiment.

Instead of being determined by the entire treatment assignment vector, users' responses are determined by different treatment exposures, which are resulted from the interaction of sampling design ($\mathbf{Z}$) and the traits of users ξ_i. The key step of causal inference, as well as treatment effects estimation, is to specify f.

The ATE can be expressed as

$$\begin{aligned} \delta &= \frac{1}{N}\cdot\sum_{i=1}^{N} f_i(\mathbf{Z}=\mathbf{1},\xi_i) - \frac{1}{N}\cdot\sum_{i=1}^{N} f_i(\mathbf{Z}=\mathbf{0},\xi_i) \\ &= \tau_1 - \tau_0 \end{aligned} \tag{7}$$

where τ_1 is the expected response when treatment is applied *globally*, similarly for τ_0. We will start with two existing exposure models and their corresponding ATE estimators.

4.2.2 SUTVA

Under the Stable Unit Treatment Value Assumption, also known as Individual Treatment Response, f is defined as

$$f_i^S(\mathbf{Z},\xi_i) = \mathbf{I}(\mathbf{Z}_i = 1)\cdot\tau_1 + \mathbf{I}(\mathbf{Z}_i = 0)\cdot\tau_0. \tag{8}$$

In other words, (i) user i's treatment exposure depends only on his/her own treatment, regardless of the treatment as-

[2]The Modularity Maximization algorithm does not generate clusters with equal sizes, so that we constrained the maximum cluster size during each iteration.

signments of other users in the network; (ii) user i's expected response under treatment (control) is the same as if everyone were in treatment (control). It is easy to see that under this response function $f^S(\cdot)$, the ATE can be estimated based on (2).

4.2.3 Neighborhood Exposure

There are two basic Neighborhood Exposure models, one based on the percent of treated neighbors and the other based on the absolute number of treated neighbors. We assume the former in our discussion here as it is more robust to the heterogeneity of users' degrees, though the analysis for the latter is similar.

Given a threshold $\theta \in [0, 1]$, the expected response function f for the neighborhood exposure model is defined as

$$f_i^N(\mathbf{Z}, \xi_i) = \mathbf{I}(\mathbf{Z}_i = 1, \sigma_i \geq \theta) \cdot \tau_1 + \mathbf{I}(\mathbf{Z}_i = 0, \sigma_i \leq 1 - \theta) \cdot \tau_0 + \big(\mathbf{I}(\mathbf{Z}_i = 0, \sigma_i \geq 1 - \theta) + \mathbf{I}(\mathbf{Z}_i = 1, \sigma_i \leq \theta)\big) \cdot x_i(\mathbf{Z}, \xi_i),$$

where $x_i(\mathbf{Z}, \xi_i)$ is an unknown function and σ_i is the percent of i's neighbors in treatment. Define $\mathcal{T}_i^\theta = \{\mathbf{Z} | \mathbf{Z}_i = 1, \sigma_i \geq \theta\}$. Then we have that $\forall\ \mathbf{Z} \in \mathcal{T}_i^\theta$, i is *neighborhood exposed to treatment*. Similarly, define $\mathcal{C}_i^\theta = \{\mathbf{Z} | \mathbf{Z}_i = 0, \sigma_i \leq 1 - \theta\}$. We have that $\forall\ \mathbf{Z} \in \mathcal{C}_i^\theta$, i is *neighborhood exposed to control*.

Under the neighborhood exposure model, we have (i) user i's treatment exposure depends on not just his/her own treatment but also his/her neighbors' treatment assignment; (ii) user i's expected response when he/she is network exposed to treatment (control) is τ_1 (τ_0), the same as if everyone were in treatment (control).

For i to be neighborhood exposed to treatment, i should be in treatment and at least θ fraction of i's neighbors are also in treatment. For users who are not neighborhood exposed to either treatment or control, their responses, $x_i(\mathbf{Z}, \xi_i)$ are considered invalid and are hence removed from estimation. Again, we can estimate τ_1 and τ_0 based on the sample means of the network-exposed users and hence the ATE as

$$\hat{\delta}^N = \frac{1}{N_1^\theta} \cdot \sum_{\{i; \mathbf{Z} \in \mathcal{T}_i^\theta\}} Y_i - \frac{1}{N_0^\theta} \cdot \sum_{\{i; \mathbf{Z} \in = \mathcal{C}_i^\theta\}} Y_i, \tag{9}$$

where N_0^θ, N_1^θ are the number of users that are network-exposed to control, treatment with threshold θ respectively.

A key step of applying Neighborhood Exposure model is to select θ, the threshold for σ_i that determines which nodes meet the condition of being neighborhood exposed. To see how θ affects the estimation, we first apply the randomized graph partition to the employee network in Table 2. After randomly assigning 50 of the 100 clusters to treatment, we have an empirical cumulative distribution function of σ_i for the nodes in treatment (Figure 2).

By setting different θ's, we have the following observations.

- If we set $\theta = 0.9$, we have $\Pr(\sigma_i \leq \theta) = 0.79$. A larger θ means a weaker assumption for the exposure model and hence smaller bias. However, about 79% observations are "invalid" and cannot be used to estimate the ATE, which leads to larger variance;
- If we set $\theta = 0.3$, we have $\Pr(\sigma_i \leq \theta) = 0.07$. 93% of the observations can be used in the estimation. However, a smaller θ means a stronger assumption and hence a larger bias.

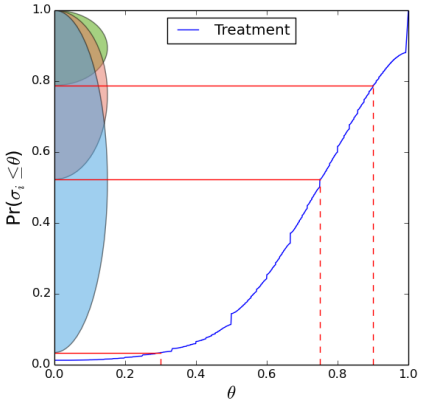


Figure 2: Empirical cumulative distribution function of σ_i for nodes in treatment.

The bias and variance trade-off presented in the choice of θ leads us to propose a new estimator that is able to take into consideration the level of exposure, and hence utilize all observations regardless whether they are network exposed. The new model also turns out to be a generalization of the exposure models and estimators discussed so far.

4.2.4 Fraction Neighborhood Exposure

In this section, we propose a new estimator based on what we call the *Fraction Neighborhood Exposure* model. We define $f(\cdot)$ the expected response function to be any function that depends only on the user's experiment assignment and the fraction of his treated neighbors σ_i:

$$f_i^F(\mathbf{Z}, \xi_i) = g(\mathbf{Z}_i, \sigma_i). \tag{10}$$

It is easy to see that the ATE in (7) becomes

$$\delta = g(\mathbf{1}, 1) - g(\mathbf{0}, 0). \tag{11}$$

Various response function $g(\cdot)$ can be chosen to model users' behavior. We demonstrate using two linear additive models.

Linear Additive Model I.

The first model we consider is as follows:

$$g(\mathbf{Z}_i, \sigma_i) = \alpha + \beta \mathbf{Z}_i + \gamma \sigma_i, \tag{12}$$

where β captures the treatment effects and γ the network effects. α, β, γ can be estimated from users' responses, as $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$. The additive assumption of various effects is also made by previous works, such as the "linear-in-mean" model [15]. By (11), the ATE can be estimated as

$$\hat{\delta}_{L_I} = \hat{\beta} + \hat{\gamma}.$$

Linear Additive Model II.

The linear model in (12) can be further generalized by considering different response functions for users in treatment and control groups:

$$g(\mathbf{Z}_i, \sigma_i) = \begin{cases} \alpha_0 + \gamma_0 \sigma_i, & \text{if } \mathbf{Z}_i = 0 \\ \alpha_1 + \gamma_1 \sigma_i, & \text{if } \mathbf{Z}_i = 1 \end{cases} \tag{13}$$

where α_0, γ_0 are learned from observation data of users in control group, while α_1, γ_1 are learned from observation data of users in treatment group. By (11), the ATE can be estimated as

$$\hat{\delta}_{L_{II}} = \hat{\alpha}_1 + \hat{\gamma}_1 - \hat{\alpha}_0.$$

4.2.5 Comparison of Exposure Models

While comparing the three exposure models we discussed so far, we show in this section that the SUTVA and neighborhood exposure model are special cases of the fraction neighborhood exposure model.

Lemma 1. *SUTVA is a special case of fraction neighborhood exposure model in* (12).

Proof. It can be trivially proved by setting $\gamma = 0$, *i.e.*, set the network effects to be zero. □

For the case of neighborhood exposure model, we only need to show that the assumption of the response function is a special case of the linear model in (12).

Lemma 2. *The neighborhood exposure model is a special case of the fraction neighborhood exposure model in* (12).

Proof. The response function in neighborhood exposure model can be defined as follows:

$$f_i^N(\mathbf{Z}, \xi_i) = \begin{cases} \tau_0 & \text{if } \mathbf{Z}_i = 0, \sigma_i < 1-\theta \\ \tau_1 & \text{if } \mathbf{Z}_i = 1, \sigma_i > \theta \\ \text{invalid} & \text{otherwise} \end{cases}$$

which is equivalent to

$$f_i^N(\mathbf{Z}, \xi_i) = \begin{cases} \alpha + \beta \mathbf{Z}_i & \text{if } \mathbf{Z}_i = 0, \sigma_i < 1-\theta \\ \alpha + \beta \mathbf{Z}_i & \text{if } \mathbf{Z}_i = 1, \sigma_i > \theta \\ \text{invalid} & \text{otherwise} \end{cases},$$

by re-parameterizing $\alpha = \tau_0, \beta = \tau_1 - \tau_0$. In other words, $f_i^N(\mathbf{Z}, \xi_i) = \alpha + \beta \mathbf{Z}_i$ restricted to only users who are network-exposed. □

5. EXPERIMENTAL RESULTS

In order to demonstrate the effectiveness and efficiency of fraction neighborhood model, we performed various experiments including extensive simulations and a real online experiment. These two types of experiments are complementary. We can obtain ground truth in the simulation study, so it is possible to evaluate methods under various conditions based on their bias and variance. On the other hand, real experiment offers special insights into how models compare in practice. Also note that all simulations here are based on real social network, so we only need to simulate user responses.

5.1 Simulations

Considering the heterogeneity of degrees of nodes in real social networks, we will use the Randomized Balanced Graph Partition algorithm proposed in Section 4.1.2 to determine the experiment assignment and focus on comparing the estimation models. To be more specific, we aim at answering the following questions based on the simulations:

1. How does the percentage of units under treatment (or control) influence the estimation?
2. How do different estimation models compare with respect to both bias and variance?

All simulations are based on the employee network summarized in Table 2, which has a heterogeneous degree distribution with a high variance. We have also run similar simulation analysis on different networks extracted from LinkedIn's social network graph, but the results are similar.

5.1.1 Simulation Model

The observed user response is generated based on the following probit model [8][3]

$$\begin{aligned} \widetilde{Y}_{i,t} &= \lambda_0 + \lambda_1 \cdot \mathbf{Z}_i + \lambda_2 \cdot \frac{\boldsymbol{A}_{\cdot i}^\top \cdot \boldsymbol{Y}_{t-1}}{D_{ii}} + U_{i,t} \\ Y_{i,t} &= \mathbf{I}(\widetilde{Y}_{i,t} > 0), \end{aligned} \tag{14}$$

where $Y_{i,t}$ is the response of i at time t and $U_{i,t} \sim \mathcal{N}(0,1)$ is a stochastic component capturing user specific traits. λ_0 is a constant that $\lambda_0 = -1.5$. λ_1 captures the strength of treatment effects, and λ_2 is for the network effects. We initialize $Y_{i,0} = 0$ for all users, and then run the iterative process to generate users' responses for T steps with different combinations of λ_1 and λ_2 where $\lambda_1 \in \{0.25, 0.5, 0.75, 1.0\}$, and $\lambda_2 \in \{0, 0.1, 0.5, 1.0\}$. The step length is set to $T = 3$.

It is worth noting that in this data generation model, one's response depends directly on the behavior of his/her neighbors at the previous time stamp, not simply the neighbors' experiment assignments, which is quite realistic.

5.1.2 Estimators

We have discussed different estimators in Section 4.2.1 according to different exposure models. In addition to these discussed, we will also include in the comparison the *Hajek estimator*[2]. Specifically, we will study the following five estimators:

1. Under SUTVA, the sample mean estimator as in (2).
2. Under the Neighborhood Exposure Model, the sample mean estimator as in (9).
3. Under the Neighborhood Exposure Model, the *Hajek estimator*[2] defined as follows

$$\begin{aligned} \widehat{\tau}_{0,H} &= \frac{\sum_{i=1}^N \mathbf{I}(\mathbf{z}_i = 0, \sigma_i < 1-\theta) \cdot Y_i / \pi(\mathcal{C}_i^\theta)}{\sum_{i=1}^N \mathbf{I}(\mathbf{z}_i = 0, \sigma_i < 1-\theta) \cdot 1/\pi(\mathcal{C}_i^\theta)} \\ \widehat{\tau}_{1,H} &= \frac{\sum_{i=1}^N \mathbf{I}(\mathbf{z}_i = 1, \sigma_i > \theta) \cdot Y_i / \pi(\mathcal{T}_i^\theta)}{\sum_{i=1}^N \mathbf{I}(\mathbf{z}_i = 1, \sigma_i > \theta) \cdot 1/\pi(\mathcal{T}_i^\theta)} \\ \hat{\delta}_H^N &= \widehat{\tau}_{1,H} - \widehat{\tau}_{0,H} \end{aligned} \tag{15}$$

   where $\pi(\mathcal{C}_i^\theta)$ and $\pi(\mathcal{T}_i^\theta)$ are the probabilities for user i to be neighborhood exposed to control and treatment given the cluster level randomization and threshold θ. We set $\theta = 0.75$.
4. Under the Fraction Neighborhood Exposure Model, the linear model I estimator (12).
5. Under the Fraction Neighborhood Exposure Model, the linear model II estimator (13).

[3]For fair comparison, the simulation process is exactly the same as [8] except that we use real networks while [8] uses simulated random networks.

We did 1000 simulations to compute both bias and variance for each of these estimators. For each simulation, the true ATE is estimated by putting all users in treatment or all in control.

5.1.3 Percentage of Users in Treatment

When running an A/B test, how to split the traffic between treatment and control is one of the key questions one has to decide. This is even more crucial if there are network effects, as the overall percentage of users in treatment (ρ) will substantially influence the fraction of one's neighbors in treatment, which leads to different level of treatment exposure. This is demonstrated by how the empirical cumulative distribution function of σ_i changes with different ρ's in Figure 3.

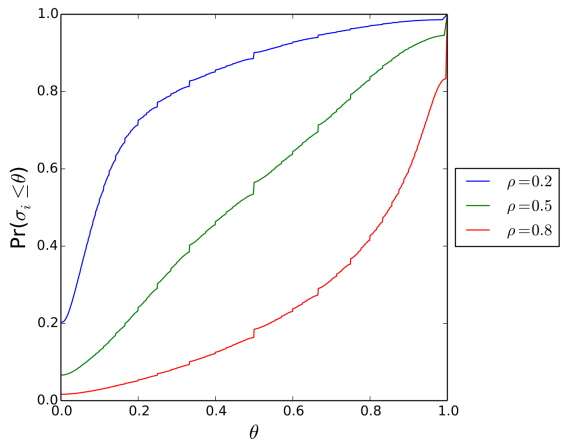
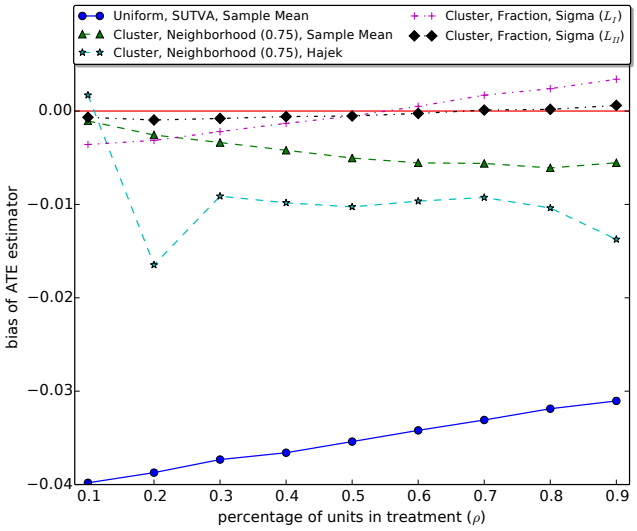


Figure 3: By changing the percentage of units in treatment ρ, the distribution of σ_i changes significantly.

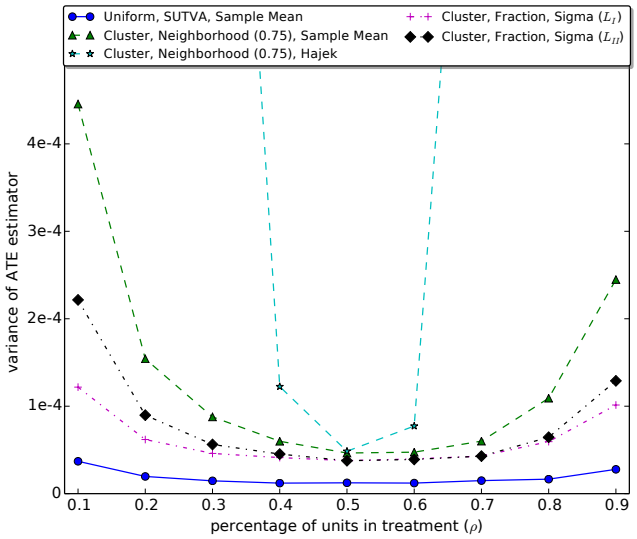
A more interesting question is how the performance of different estimators change with ρ. As far as we know, we are the first one to study this problem and the results are shown in Figure 4. We have the following observations. (i) The bias of each estimator does not change much as ρ changes, but the variance changes significantly and reaches the smallest when $\rho = 0.5$ (as expected since Var $\sim 1/\sqrt{1/N_0 + 1/N_1}$). (ii) For $\rho < 0.3$ and $\rho > 0.7$, the variance of the Hajek estimator is large because the scale factors π_i's (probabilities of being neighborhood exposed) are small. (iii) The Hajek estimator performs worse than sample mean estimator under the same neighborhood exposure model. It is interesting to note that, in random networks such difference is minimal [8], showing the importance of the network structure when evaluating estimators. (iv) Our linear estimators under the Fraction Neighborhood Exposure model achieve the smallest bias and variance. The L_I estimator achieves smaller variance, while L_{II} estimator achieves smaller bias. The different between L_I estimator and L_{II} is minimal when $\rho = 0.5$.

5.1.4 Network Effect

We now compare the performance of different ATE estimators under different levels of treatment effects and network effects, by changing the values of λ_1 and λ_2 in the simulation model (14). We use $\rho = 0.5$ and the results are



(a) The bias of different estimators

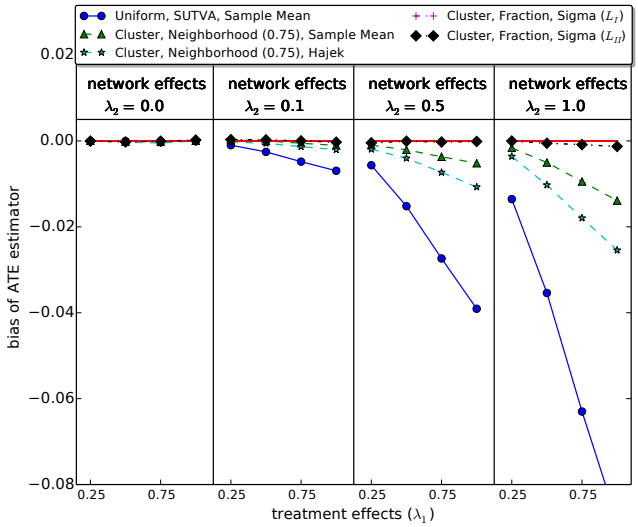


(b) The variance of different estimators

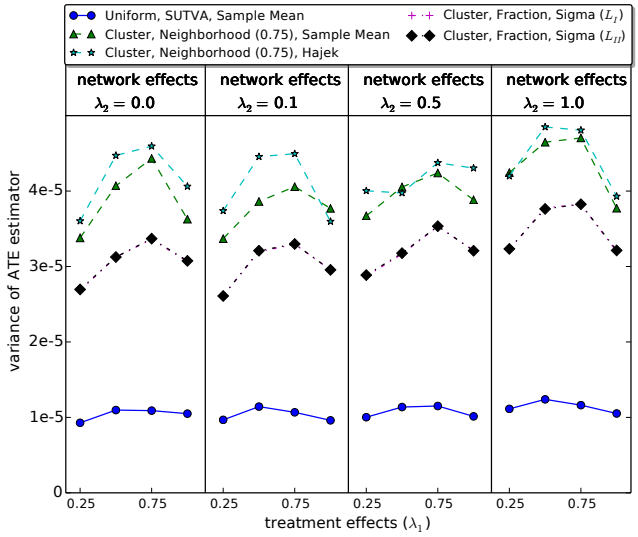
Figure 4: Behavior of different estimators with different percentage of neighbors in treatment.

shown in Figure 5. The results with $\rho = 0.1$ are similar and hence omitted.

We have the following observations. (i) The bias of both sample mean and Hajek estimators under the neighborhood exposure model becomes larger as the network effects and treatment effects become larger; (ii) The variance of the estimators is fairly consistent across different levels of treatment effects and network effects. This is because the number of observations used in estimation does not change as λ_1 and λ_2 change. In addition, the linear estimators under our fraction neighborhood exposure model achieve a smaller variance than the estimators under the neighborhood exposure model because all the observations are included in the estimation. (iii) Our linear estimators achieve the best bias and variance across different levels of network effects and treatment effects.



(a) The bias of different estimators



(b) The variance of different estimators

Figure 5: Behavior of different estimators with different parameters, which controls the strength of treatment effects (λ_1) and network effects (λ_2), in the simulation model. The overall percentage of nodes in treatment $\rho = 0.5$.

5.2 Real Online Experiment

In addition to the extensive simulations, we have also conducted a real online experiment at LinkedIn using the network A/B testing framework we have proposed. Specifically, we have done the following:

1. Select a country as the sub-network to experiment on.
2. Apply our randomized balanced graph partition algorithm to assign users from this country into treatment and control groups.
3. Apply different Feed algorithms to the treatment and control groups. Estimate the ATE after running the experiment for two weeks.

Country	$N_{\mathcal{S}}$ ($\times 10^6$)	$R_{\mathcal{S}}$	$\overline{d_{\mathcal{S}}}$
Brazil	19.9	0.932	41.6
United States	119.3	0.910	54.3
Netherlands	**6.1**	**0.868**	**93.0**
Chile	2.8	0.866	38.4
New Zealand	1.3	0.654	29.4

Table 4: Basic statistics about several countries. $N_{\mathcal{S}}$ is the size of the subnetwork, selected by the corresponding country; $R_{\mathcal{S}}$ is the self-containment measure defined in (16); $\overline{d_{\mathcal{S}}}$ is the average degree of the selected subnetwork defined in (17).

We note that unlike simulations, there is no ground truth for this real world experiment. The Feed team has, however, compared these two Feed algorithms globally in a uniformly randomized A/B test, and the treatment Feed algorithm was significantly better than control.

Our goal for the real world experiment is two-fold. First, we would like to compare results from different estimators in a real application setting to complement the observations from simulations. In particular, we want to compare results with and without taking into consideration of the network effects, and further, how our fraction neighborhood exposure model compares with the neighborhood exposure model. Second, as far as we know, we are the first to run a real network A/B test. We would like to establish a process for running network A/B test in practice. As we have seen how the conclusions can differ drastically in real networks compared to simulated networks, we hope this can bridge the gap and encourage more research focusing on real applications in the area of network A/B testing.

5.2.1 Country Selection

We would like to select a country that has a well self-contained LinkedIn social network. Ideally, It should be an isolated sub-network that has as few connections to the outside of the country as possible to prevent network influence to and from users outside. We use the following ratio to quantify such "self-containment" for a set $\mathcal{S}$ of users:

$$R_{\mathcal{S}} = \frac{\sum_{i,j\in\mathcal{S}} A_{i,j}}{\sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}\cup\mathcal{S}^c} A_{i,j}}, \tag{16}$$

where $\mathcal{S}^c$ is the complement of $\mathcal{S}$ in the population network. Remark that $R_{\mathcal{S}}$ is the ratio between the edge count within the set $\mathcal{S}$ and all the edges with one end in $\mathcal{S}$.

In addition, we consider the average internal degree such that the selected subnetwork is well connected. Considering selected sub-network $\mathcal{S}$, the average internal degree is defined as

$$\overline{d_{\mathcal{S}}} = \frac{\sum_{i,j\in\mathcal{S}} A_{i,j}}{|\mathcal{S}|}. \tag{17}$$

We calculate $R_{\mathcal{S}}$ and $\overline{d_{\mathcal{S}}}$ for all countries on LinkedIn, and among the ones with top $R_{\mathcal{S}}$ ratios (shown in Table 4), we decide to pick Netherlands as it has the highest average internal degrees and a reasonably large sized network (around 6 million users).

After selecting the Netherlands as the sub-network, we applied the randomized balanced graph partition algorithm

to divide it into 600 shards and randomly picked 300 of them to receive treatment while the rest 300 to receive control. Before performing the A/B test, we have also conducted an A/A test, which is a controlled experiment where treatment is identical to control. This was to confirm that no bias was introduced during the experiment assignment process.

5.2.2 *Online Results*

We let the experiment run for two weeks before we collected data for analysis. The metric we use for evaluation is the average number of social gestures on Feed, such as "like", "comment" or "share".

We compute the ATE based on the various estimators described in Section 5.1.2. The results are shown in Table 5. We have the following observations. (i) The ATE estimators with consideration of network effects are all larger than the estimate under SUTVA. This is yet another good confirmation that there are indeed network effects presented in the A/B experiment. (ii) The choice of θ in the neighborhood exposure model matters. The sample mean estimator almost doubles when θ changes from 0.75 to 0.9. On the other hand, the Hajek estimator gives a smaller estimate when θ changes from 0.75 to 0.9, this is due to small values of π_i's when θ is large. (iv) The fraction neighborhood exposure model gives larger estimates than existing methods.

Method	ATE for social gestures
SUTVA	0.168
Neighbor. Exposure $\theta = 0.75$	0.264
Neighbor. Exposure $\theta = 0.9$	0.520
Hajek. Exposure $\theta = 0.75$	0.625
Hajek. Exposure $\theta = 0.9$	0.133
Fraction Exposure (I)	0.687
Fraction Exposure (II)	0.714

Table 5: ATE estimates from different models for the online Feed experiment.

6. CONCLUSION AND FUTURE WORK

In this paper, we study the problem of network A/B testing in real networks. We start by examining a recent A/B experiment conducted on LinkedIn without considering network structures, which motivates us to set up a framework to study both the sampling and the estimation aspects of the network A/B testing problem. To address the challenge of degree heterogeneity in real social networks, we come up with a new randomization scheme based on balanced graph partitioning, for which an efficient and distributed algorithm is proposed. Based on new sampling scheme, we propose a new method to estimate the average treatment effect (ATE) that is able to take into consideration of the level of network exposure. Extensive simulations are conducted to evaluate these methods and the results show that our new proposals can achieve both a smaller bias and a smaller variance. We have also conducted a real online experiment under the framework we have proposed and the results further validate many observations from simulations.

On the other hand, there are still many open problems in the field of network A/B testing that remain to be addressed, especially with respect to real world applications. First of all, we did not consider the influence strength between pairs of nodes, which may have significant impact on determining users exposure status; Secondly, real social networks are growing all the time, leading to rapid change of network structures, which makes network A/B testing even more challenging considering the effects of newly added edges and nodes. To further complicate the problem, many real experiments on social networks are aiming at increasing network density, making the temporal variability a real, noticeable issue. Thirdly, there are different forms of network interference to be considered. For instance, in discussion *groups*, information propagates from one user to all other users of the same group, so every group acts as a fully connected sub-network. However each user can belong to multiple groups. In this case, the graph clustering randomization can no longer split users into treatment and control under the new information propagation structure. Lastly, our focus here has been on ATE estimation, and we have not touched upon how virality works and how to preserve it in a network A/B testing setting. Given the complex structure of real social networks and the way viral information propagates, the framework proposed here may not be sufficient.

A/B testing in general is widely used and also well studied in the industry as it offers the best scientific approach to understand the causal impact of product changes on end user behavior. However, the problem of A/B testing in a social network setting is no where near solved. A lot of work still remains to be done to make it a well-understood problem in real world applications. We hope our work here can bridge some of the gaps and encourage more research in this area.

Acknowledgement

Firstly, we wish to thank our colleagues and friends who have held many enlightening discussions with us: Deepak Agarwal, Mathieu Bastian, Evion Kim, Wayne Tai Lee, Haishan Liu, Mitul Tiwari, Xiaolong Wang, and many members of the A/B testing team at LinkedIn. Secondly, we wish to thank Bee-Chung Chen, Liang Zhang, Pannagadatta Shivaswamy, Cory Hicks and Caroline Gaffney for working with us to run the network A/B test on LinkedIn Feeds. Lastly, we wish to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

[1] K. Andreev and H. Racke. Balanced graph partitioning. *Theory of Computing Systems*, 39(6):929–939, 2006.

[2] P. M. Aronow and C. Samii. Estimating average causal effects under general interference. Citeseer, 2012.

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

[4] D. A. Bader, H. Meyerhenke, P. Sanders, and D. Wagner. *Graph partitioning and graph clustering*, volume 588. American Mathematical Soc., 2013.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[6] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. International World Wide Web Conferences Steering Committee, 2014.

[7] N. Cressie and H. Whitford. How to use the two sample t-test. *Biometrical Journal*, 28(2):131–148, 1986.

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

[9] H. Gui, Y. Sun, J. Han, and G. Brova. Modeling topic diffusion in multi-relational bibliographic information networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 649–658. ACM, 2014.

[10] M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482), 2008.

[11] K. L. Judd. The law of large numbers with a continuum of iid random variables. *Journal of Economic theory*, 35(1):19–25, 1985.

[12] 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.

[13] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. 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, 2013.

[14] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. Seven rules of thumb for web site experimenters. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[15] C. F. Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

[16] C. F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

[17] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[18] J. A. Middleton and P. M. Aronow. Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Available at SSRN 1803849*, 2011.

[19] M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[20] S. W. Raudenbush. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2):173, 1997.

[21] M. G. Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf. Uncovering the structure and temporal dynamics of information propagation. *Network Science*, 2(01):26–65, 2014.

[22] P. R. Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477), 2007.

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

[24] M. E. Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

[25] L. Tang, R. Rosales, A. Singh, and D. Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM, 2013.

[26] 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.

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

[28] J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. 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. ACM, 2013.

[29] F. Yates. Sir ronald fisher and the design of experiments. *Biometrics*, 20(2):307–321, 1964.