Limiting Bias From Test-Control Interference in Online Marketplace Experiments

David Holtz[*1]

[1]MIT Sloan School of Management

December 6, 2016

Abstract

Many internet firms use A/B tests to make product decisions. When running an A/B test, the typical objective is to measure the average treatment effect (ATE), i.e. the difference between the average outcome in the counterfactual situation where 100% of users are exposed to the treatment and the average outcome in the status quo situation (in which 100% of users are exposed to the control). However, a simple difference-in-means estimator will give a biased estimate of the ATE when outcomes of units in the control depend on the behavior or outcomes of units in the treatment - a case referred to in this work as test-control interference. Both Blake and Coey [2] and Fradkin [6] find evidence of bias due to test-control interference in online marketplace experiments. This paper considers the use of experimental designs and ATE estimators discussed by Eckles et al [5], which reduce bias in the presence of peer effects, in online marketplace experimentation. In order to do so, I model the marketplace as a network in which an edge exists between two sellers if their goods substitute for one another. Using data scraped from Airbnb [16], I create an agent-based simulation to model seller outcomes, both under the status quo and subject to "treatments" that force hosts to lower their price. Having estimated hosts' baseline outcomes and their outcomes when 100% of hosts are subject to the treatment, I use the same simulation framework to approximate ATE distributions obtained when various network experiment design and analysis techniques are used. I find that graph cluster randomization leads to bias reductions of as much as 62%. Unfortunately, the variance of ATE estimators also increases significantly. Replacing the simple difference-in-means estimator with more sophisticated ATE estimators can lead to mixed results. While some methods (i.e., exposure models) provide (small) additional reductions in bias and small reductions in variance, others (i.e., the Hajek estimator for the ATE) lead to increased bias and variance. Although further work is needed, current results suggest that experiment design and analysis techniques from the network experimentation literature are promising tools for for reducing bias due to test-control interference in marketplace experiments.

[*] dholtz@mit.edu

1 Introduction

A common way for internet firms to make product decisions is through experimentation, or 'A/B testing.' Typically, users who visit a website during an A/B test are randomly assigned to either the test or control group. These two groups experience different variants of the online product. After some amount of time, a 'winner' is chosen by evaluating the variants' relative performance. This evaluation is performed by calculating the average treatment effect (ATE) for some number of predetermined metrics (e.g., click-through rate, average revenue per user) and performing a statistical significance test [11]. Conceptually, the ATE is meant to measure the difference in average outcome across all units between the counterfactual situation in which 100% of units are exposed to the treatment, and the status quo situation in which 0% of units are exposed to the treatment.

Under the assumption that each experimental unit's response is not influenced by any other units' treatment assignment or behavior, the ATE can be identified by randomly assigning units to the treatment or control, and calculating the difference in means between the two groups. This assumption is often referred to as the 'stable unit treatment value assumption' (SUTVA) [15] or the 'no interference' assumption [4]. However, online marketplaces are by definition connected. For instance, a treatment that increases buyer demand may reduce the supply available to buyers in the control. Similarly, a treatment that induces sellers to lower prices may increase demand for treatment sellers' products and reduce the demand observed by sellers in the control. In both of these cases, the treatment affects not only the targeted units' outcomes, but also the outcomes of other units in the marketplace.

In the online experimentation literature, this phenomenon is referred to as 'test-control interference.' Blake and Coey [2] find empirical evidence for test-control interference in marketplaces by analyzing an email marketing experiment performed on eBay. They conclude that naive estimates of the ATE ignoring test-control interference exaggerate the treatment's effectiveness by about a factor of 2. Fradkin [6] simulates the counterfactual impact of implementing new search algorithms in an online marketplace (Airbnb). Comparing the true counterfactual impact of new algorithms to the counterfactual impact estimated by simulated A/B tests, Fradkin finds that A/B tests ignorant of test-control interference can overstate true treatment effects by over 100%.

Both Blake and Coey [2] and Fradkin [6] propose methods for combatting test-control interference in marketplace experiments. Rather than compare outcomes across buyers, Blake and Coey limit their analysis to auctions where a majority of bidders are in the treatment or control, and compare outcomes across auctions. While this does eliminate within-auction interactions between users that may bias results, there is still potential bias in their effect size estimates due to interference across auctions. Furthermore, it is not clear how well this methodology generalizes to marketplaces that do not offer convenient units of analysis (e.g., auctions) over which to aggregate outcomes. Fradkin proposes randomizing across well-defined markets (rather than buyers or sellers), or combining experimental results with the results of complicated structural model simulations. Market level experiments are often infeasible (due to ambiguous (or non-existent) market definitions, a limited number of markets, or market heterogeneity), and the simulations necessary to quantify equilibrium treatment effects may be too complicated and/or cumbersome for many firms or researchers to implement.

In this paper, I propose an experimentation strategy that reduces bias due to test-control interference by implementing design principles and analysis techniques from the network experimentation literature, as detailed by Eckles et al [5]. There is extensive work in the network inference literature on measuring ATEs in the presence of peer effects and interference. Aronow and Samii [1] give unbiased estimators for the ATE under certain restrictions on the extent of interference (e.g., interference is dependent on the number of treated neighbors), and derive estimators for the variance. Building on this work, Ugander et al [17] show that graph cluster randomization results in a greater number of units satisfying the conditions required for Aronow and Samii's estimators. However, Manski [13] shows that the very processes expected to produce interference make these assumptions on the extent of interference infeasible. The contribution of Eckles et al [5] is to consider methods for bias reduction (rather than elimination) in more realistic networks that violate many of the aforementioned assumptions. Eckles et al find via simulation that graph cluster randomization, along with the use of exposure models and more sophisticated ATE estimators, do in fact succeed in reducing bias.

I implement these same experimental design and analysis methods, and study how effectively they reduce bias due to test-control interference in a simulated online marketplace experiment. I begin with a scraped dataset consisting of Airbnb properties and their attributes. Using a simple decision rule that defines which listings are likely to substitute for one another, I infer an underlying network structure (an edge exists between two listings if they are likely to substitute for one another). I next simulate the process of searchers with heterogenous preferences sequentially visiting Airbnb, facing a consideration set provided by a search algorithm, and making purchase decisions. Given proposed treatments that induce listings to lower their price, I am able to quantify the booking rate and average revenue per listing both in the baseline case and in the counterfactual situation where 100% of listings are exposed to the treatment. Having quantified the true ATE, I consider different randomization schemes and estimators, and evaluate the biasedness and variance of the resultant ATE estimates. Notably, the work contained in this paper does not make any assumptions about the functional form of any potential interference between units and the marketplace simulations do not include any explicit interference.

The structure of this paper is as follows. Section 2 reviews the experimental designs and analysis methods to be evaluated via simulation. Section 3 provides a summary of the Airbnb dataset used, the network generation process, and the methodology used to define clusters on that network. Section 4 describes the simulation process used to generate outcomes. Section 5 compares the bias and variance achieved with different experimental designs and analysis methods. Section 6 summarizes these findings and discusses opportunities for future work.

2 Theoretical Motivation

I will first provide a brief overview of the motivation for and details of the experiment design and analysis methods considered in this paper. A more thorough discussion is provided by Eckles et al [5], from which this overview draws heavily. In all cases, the quantity of interest is the average treatment effect (ATE), which conceptually quantifies the difference between the average outcome across all units in the counterfactual situation in which all units are exposed to the treatment and the average outcome across all units in the situation in which all units are exposed to the control. Suppose we have a network with N units. Define Z as a vector of length N indicating each unit's treatment assignment. Now, define $Y_i(Z = z)$ as unit i's outcome when the global treatment vector Z is equal to z. The average treatment effect between two treatment vectors z_1 and z_0 is defined as

$$\tau(z_1, z_0) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_i(z_1) - Y_i(z_0)\right]. \tag{1}$$

Since it is impossible to observe both the counterfactual situation in which all units are assigned z_1 and the counterfactual situation in which all units are assigned to z_0, researchers often assume independence between the units (SUTVA). Under this assumption, $Y_i(Z)$ only depends on Z_i. As a result, $Y_i(Z)$ can be simplified to $Y_i(Z_i)$. In conjunction with random assignment, this assumption typically allows τ to be identified by a simple difference-in-means estimator. However, one can easily imagine situations where SUTVA is violated. One such scenario is the test-control interference in marketplace experiments that this paper considers.

One method to reduce the biasedness of the difference-in-means estimator when SUTVA is violated is to randomize units in such a way that they are "closer" to the global treatments of interest. Eckles et al [5] and Ugander et al [17] show this can be achieved via graph cluster randomization. In the case of independent assignment, each Z_i is determined by drawing from a Bernoulli distribution,

$$Z_i \sim \text{Bernoulli}(q), \tag{2}$$

where q is the probability of assignment to the treatment. In the case of graph cluster randomization, the network is partitioned into N_C clusters: $C_1, C_2, ..., C_{N_C}$. After the network is partitioned, each cluster is assigned a treatment, W_j, by drawing from a Bernoulli distribution,

$$W_j \sim \text{Bernoulli}(q), \tag{3}$$

where again, q is the probability of cluster assignment to treatment. Each vertex in a given cluster is then assigned its cluster's treatment assignment. Under graph cluster randomization, the simple difference-in-means estimator should provide a less-biased estimate of the ATE.

Another way to reduce bias in estimation of the ATE is to propose an "exposure model" that dictates the way in which a unit's neighbors' treatments or behavior impacts that unit's outcomes. Once an exposure model is specified, it is possible to define a function $g_i(\cdot)$ such that $g_i(z_m) = g_i(z_n)$ implies that treatment vectors z_m and z_n provide unit i with the same "effective treatment" [13]. Once an effective treatment (and effective control) are defined, the ATE estimator can be modified so that only units that experience the effective treatment or effective control are included:

$$\tau(z_1, z_0)_{eff} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_i(Z)|g(Z) = g(Z_1)\right] - \mathbb{E}\left[Y_i(Z)|g(Z) = g(Z_0)\right]. \tag{4}$$

This paper considers the fractional neighborhood treatment response (FNTR) exposure model discussed in [5, 17]. In the FNTR exposure model, a vertex is considered globally treated if more than fraction λ of its neighbors are treated[1].

The difference in means estimator for the ATE will only be unbiased if, for the randomization process in question [5],

$$E\left[Y_i | g_i(Z) = g_i(z)\right] \perp \mathbb{P}\left[g_i(Z) = g_i(z)\right]. \tag{5}$$

However, this assumption will often fail when units are assigned to different treatment arms using graph cluster randomization and/or exposure models. For instance, one could imagine that higher degree nodes (in this context, items that are similar to many others) might have a different expected outcome than lower degree nodes (in this context, very unique items), even conditional on being exposed to the same treatment. It may also be the case that a higher degree node is less likely to be assigned to the effective treatment or control, since that unit has a) more edges and b) potentially more edges that span multiple clusters. In this case, Equation 5 may not hold. An estimator such as the Horvitz-Thompson estimator [9] can provide unbiased estimates of the ATE under these conditions. However, the Horvitz-Thompson estimator for the ATE often exhibits high variance. In light of this, I follow the lead of Aronow & Samii and Eckles et al [1, 5] and use the Hajek estimator [8]. As Aronow & Samii point out, the Hajek estimator can often lead to efficiency gains, but this comes at the cost of finite sample bias and more complicated variance estimation [1]. The expression for the Hajek ATE estimator is

$$\begin{aligned}\hat{\tau_H}(z_1, z_0) = &\left(\sum_{i=1}^{N} \frac{\mathbb{1}\left[g_i(Z) = g_i(z_1)\right]}{\pi_i(z_1)}\right)^{-1} \sum_{i=1}^{N} \frac{Y_i \mathbb{1}\left[g_i(Z) = g_i(z_1)\right]}{\pi_i(z_1)} - \\ &\left(\sum_{i=1}^{N} \frac{\mathbb{1}\left[g_i(Z) = g_i(z_0)\right]}{\pi_i(z_0)}\right)^{-1} \sum_{i=1}^{N} \frac{Y_i \mathbb{1}\left[g_i(Z) = g_i(z_0)\right]}{\pi_i(z_0)},\end{aligned} \tag{6}$$

where $\pi_i(z) = \mathbb{P}\left[g_i(Z) = g_i(z)\right]$ is the probability that a given unit is assigned to the effective treatment (z_1) or effective control (z_0).

Finally, it is worth considering the effect that the randomization schemes and estimators heretofore discussed might have on the variance of ATE estimates. As Gerber and Green [7] point out, cluster random assignment (e.g., graph cluster randomization) can lead to large increases in ATE standard errors if the cluster-level mean outcome values exhibit high levels of variability. One way to offset the loss of precision due to graph cluster randomization is to perform block random assignment [7, 14]. Specifically, this paper considers a matched pair design, in which each block consists of only 2 units. Blocking is performed at the *cluster level*, as the cluster is the unit of assignment under graph cluster randomization. Under block random assignment, Bernoulli randomization occurs within each block (e.g., when the probability of assignment is 50%, there is exactly one treatment unit and one control unit within each block).

[1] This paper considers $\lambda = .50$, $\lambda = .75$ and $\lambda = .95$.

This paper aims to reduce bias due to test-control interference in online marketplace experiments. The baseline difference-in-means estimator with independent randomization is compared to the difference-in-means estimator with graph cluster randomization and cluster-level block random assignment, the modified difference-in-means estimator with the FNTR exposure model under graph cluster randomization and cluster-level block random assignment, and the Hajek estimator with the FNTR exposure model under graph cluster randomization and cluster-level block random assignment.

3 Data & Network Construction

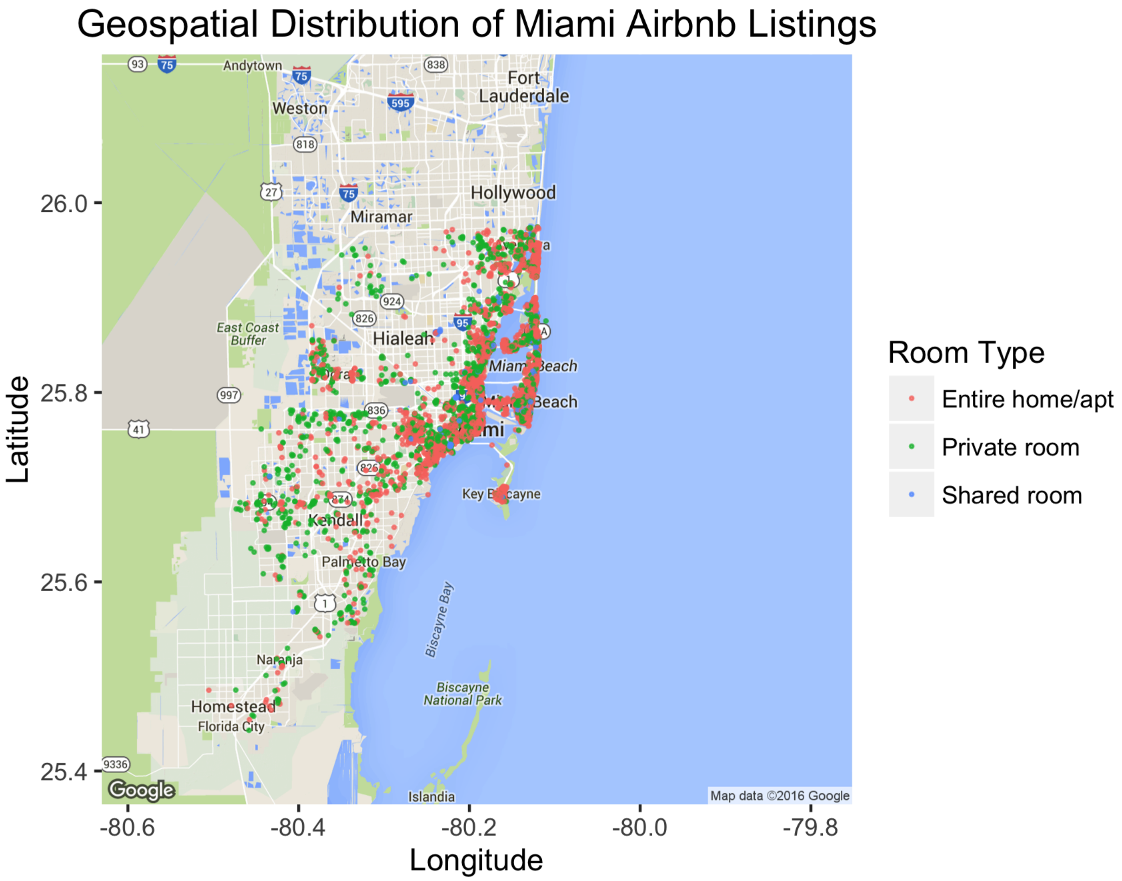


Figure 1: The geospatial distribution of Airbnb listings in and around Miami. Color corresponds to listing type. This figure was produced with ggmap [10].

Simulations in this paper use data scraped from www.Airbnb.com by Slee [16] in January 2016. The scraped dataset provides information on 8,855 Airbnb listings located in and around Miami. The Miami market is chosen because it is large enough to produce a relatively dense substitution network and is representative of other major metropolitan areas on Airbnb. It should therefore provide a good sense of how effective the proposed designs and methods may be in reducing bias. On the other hand, the dataset is not so large as to make simulations exceedingly computationally expensive. The dataset contains information on each listing's room type, number of reviews, average 'overall satisfaction' rating, guest capacity, number of bedrooms, number of bathrooms, price per night (USD), minimum length of stay, latitude, and longitude. Figure 1 depicts the geospatial distribution of the listings by room type. Table 1 provides the distribution of listings by room type. Table 2 summarizes the distributions of other covariates across the sample.

Table 1: Distribution of Airbnb Listing Room Types

	Entire home/apt	Private room	Shared room
Count	6,566	2,062	227

Table 2: Distribution of Airbnb Listing Covariates

	Reviews	Satisfaction	Capacity	Beds	Baths	Price	Min Stay
Min	0	1.00	1	0	0	$15	1
1st Qu.	0	4.50	2	1	1	$89	1
Median	3	4.50	3	1	1	$140	2
Mean	11.40	4.59	3.06	1.40	1.37	$226	3.29
3rd Qu.	12	5.00	4	2	2	$249	3
Max	304	5.00	8	10	8	$10,000	365
N/A	0	2,422	2,226	12	933	0	437

Some small amount of pre-processing is done on the dataset before the substitution network is generated and simulations are run. Missing guest capacity, bedroom, and bathroom values are imputed by the modal value for each variable. Minimum length of stay values are capped at 30, and missing minimum length of stay values are imputed by the modal value for minimum length of stay. Missing overall satisfaction values are imputed using the mean value of non-empty entries.

Each row in this dataset will then constitute a vertex in a network that approximates which listings are likely to substitute for one another when searchers are making purchasing decisions. The next step is to generate edges between vertices. Recall that an edge between two listings is meant to suggest that listing i and listing j might reasonably substitute for one another when a potential guest is making a purchase decision. In order for an edge to be generated between two listings, the following three criteria must be satisfied[2]:

1. The listings are within 1 mile of each other
2. The listings have the same room type
3. The difference between the guest capacity of the two listings is not greater than 1 in absolute magnitude

[2] One could imagine using a subset of these criteria (e.g., all listings within 1 mile of each other may be substitutes), or a totally unrelated criteria (e.g., listings must have co-occurred in search more than x times).

After imposing the three criteria above on the full set of listings, an undirected 'substitution network' is initialized. The resulting network has 1,538,637 edges, a clustering coefficient of 0.74, and vertices in the network have an average degree of 173.76. In order to perform graph cluster randomization on the network, it must be divided into clusters. This is done using Louvain clustering [3]. The Louvain clustering algorithm attempts to maximize modularity, a value that falls between -1 and 1 and is defined as

$$Q = \frac{1}{2E} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2E} \right) \mathbb{1}(C_i = C_j), \tag{7}$$

where E is the total number of edges in the graph, A_{ij} is a $\{0, 1\}$ variable that indicates whether or not an edge exists between nodes i and j, d_i and d_j are the degrees of nodes i and j, respectively, and $\mathbb{1}(C_i = C_j)$ is an indicator function that is equal to 1 only when i and j belong to the same cluster. Conceptually, Louvain clustering attempts to maximize the density of links inside communities relative to links between communities. This results in the network being partitioned into 169 clusters, which have an average size of 52.40 listings.

The resultant clusters are then divided into matched pairs to accommodate block random assignment. Within each block, the average number of reviews, average overall satisfaction score, average number of beds, average number of bathrooms, average minimum stay, average latitude, average longitude, percentage of private room listings, and percentage of shared room listings are calculated. Distance between clusters is then calculated using the Mahalanobis distance [12] and blocks are chosen to minimize the sum of the distances between clusters in the same block.

4 Simulation Process

A simple simulation of the process by which different Airbnb listings might come to be booked for a given calendar night is used to quantify average treatment effects, as well as the bias and variance of various ATE estimators given different experimental designs. Each set of simulated outcomes has the following steps:

1. A 'search algorithm' is randomly drawn
   - The search algorithm takes into account number of reviews, overall satisfaction score, number of beds, number of baths, minimum stay length, and price
   - The search algorithm randomly draws a vector of search algorithm weights for these six attributes. The weights always sum to 1
   - A search algorithm score is assigned to each listing. The search algorithm score is the weighted average (using the search algorithm weights) of centered and scaled versions of the six aforementioned variables
2. 1,000 'searchers' sequentially look for an Airbnb listing in and around Miami. For each searcher:

- The searcher randomly draws a region of interest in latitude/longitude space. The locations of the box edges are drawn with uniform probability from the range of listing latitudes and longitudes. The search algorithm will only return listings in this box
- The searcher draws a room type preference from a uniform distribution over {entire home/apt, private room, shared room}. The search algorithm will only return listings of this type
- The searcher draws a minimum guest capacity from a uniform distribution over {1,2,3,4}. The search algorithm will only return listings with a guest capacity greater than the searcher's minimum
- The searcher randomly draws a vector of searcher weights for the same six attributes considered by the search algorithm. Again, these weights always sum to 1
- The searcher randomly draws a reserve score from the uniform distribution over {0, 2}.
- A searcher score is assigned to each listing. The searcher score is the weighted average (using the searcher weights) of centered and scaled versions of the six aforementioned variables
- The search algorithm provides the searcher with a consideration set. The consideration set consists of 10 listings satisfying the searcher's location, room type, and guest capacity preferences with the highest search algorithm scores
- The searcher chooses the listing in the consideration set with the highest searcher score. As long as the searcher score of that listing is above the searcher's reserve score, that listing is 'booked' and can no longer appear in future searchers' consideration sets
- If the search algorithm returns an empty consideration set or the searcher score of the highest ranked listing in the consideration set is below the searcher's reserve score, the searcher does not book any listings and exits the platform

The process above simplifies Airbnb marketplace dynamics in countless ways. To name a few: no analysis has been done to ensure that the search algorithm or searcher behavior described above matches the empirically observed search algorithm or searcher behavior on Airbnb. Many of the random variables drawn from uniform distributions (e.g., searchers' location and covariate preferences) are unlikely to be uniformly distributed in reality. For instance, searchers are much more likely to look for listings in busy downtown areas than on the outskirts of the city. Finally, the proposed model does not do a good job of taking into account various general equilibrium effects, such as those described in [2]. For example, a reduction in prices would likely cause shifts in both the amount of demand and amount of supply on the site. Although this is captured to some extent by searchers' reserve scores, the uniform distribution from which they are drawn is completely arbitrary, and the current simulation does not take into account supply-side effects (i.e., listings may exit the marketplace if they cannot earn some minimum amount per night, and listings may enter the marketplace if the potential earnings are sufficiently lucrative).

Nonetheless, the aforementioned simulation provides an effective toy model to begin studying to what degree test-control interference may bias estimates of ATEs in online marketplace experimentation. It can also help determine to what extent the proposed experimental design and analysis methodology can reduce that bias.

5 Results

This paper considers two different outcomes across Airbnb listings via simulation. The first is whether or not a listing gets booked, referred to as Y_i. The average of Y_i over the sample is referred to as the booking rate, BR:

$$BR = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[Y_i\right]. \tag{8}$$

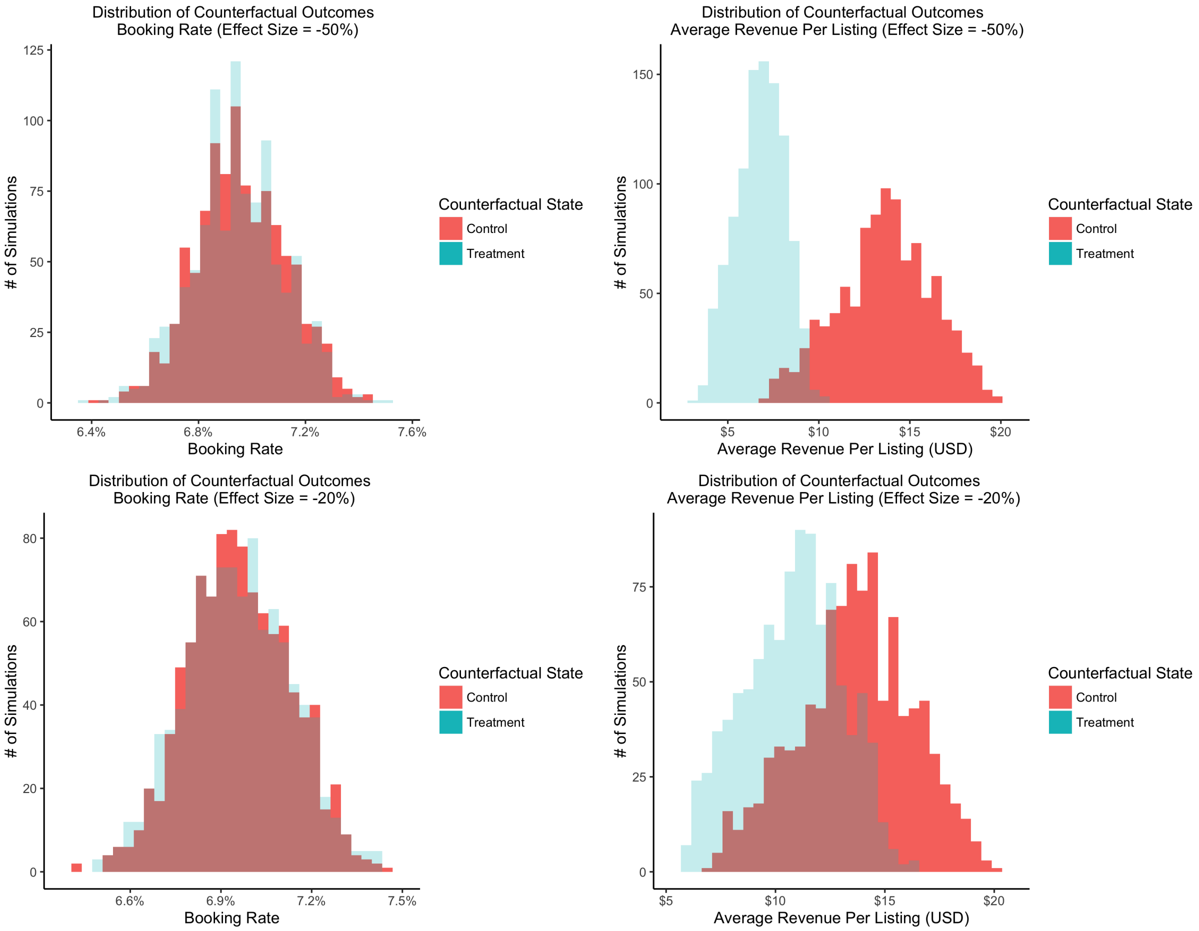
The other outcome discussed in this paper is a listing's revenue, which is defined as $R_i = Y_i p_i$. The average of R_i over the sample is referred to as the average revenue per listing, $ARPL$:

$$ARPL = \frac{1}{N} \sum_{i=1}^{N} p_i \times \mathbb{E}\left[Y_i\right]. \tag{9}$$

This paper considers a treatment that induces all listings to lower their price. Two different effect sizes are considered: one that induces listings to lower their price by 20% and one that induces listings to lower their price by 50%. The first effect size is more realistic, whereas the second effect size should yield more pronounced test-control interference. I aim to answer a number of questions. First, what is the true ATE? In other words, what is the difference between the booking rate and average revenue per listing when 100% of listings experience the control, and when 100% of listings lower their price by 20% or 50%? Second, what is the magnitude of the bias in estimation of the ATE on these outcomes due to test-control interference, given listing treatments were determined with independent random assignment and the ATE is estimated using a simple difference-in-means estimator? Third, how to what extent do various experiment design and analysis methods reduce this bias, and what effect do they have on the RMSE of the the estimate?

It is first necessary to establish the true values of the estimands in question (the ATE for two outcomes: receiving a booking and listing revenue) for both the 20% price reduction and 50% price reduction treatments. In order to approximate these difference, 1,000 simulations each are run for the baseline case (0% treatment), the first counterfactual case (100% -20% price treatment), and the second counterfactual case (100% -50% price treatment). Each simulation follows the steps outlined in Section 4.

Figure 2 compares the distribution of average booking rates and average revenue per listing measured in the baseline case and the two treatment cases. A two-sided t-test between the baseline distribution of booking rates and the counterfactual distribution of booking rates under the -20% price reduction treatment yields a t-statistic of $t = 2.17$ ($p = 0.03$) Although this test gives a statistically significant effect, the effect size is extremely small ($ATE = 1.05 \times 10^{-5}$). This statistically significant difference may be caused by buyers



Distribution of Counterfactual Outcomes Booking Rate (Effect Size = -50%)

Distribution of Counterfactual Outcomes Average Revenue Per Listing (Effect Size = -50%)

Distribution of Counterfactual Outcomes Booking Rate (Effect Size = -20%)

Distribution of Counterfactual Outcomes Average Revenue Per Listing (Effect Size = -20%)

Figure 2: Comparison of the simulated average counterfactual outcome distributions when either 0% or 100% of listings are assigned the treatment. Top left: -50% price reduction treatment, booking rate. Top right: -50% price reduction treatment, average revenue per listing. Bottom left: -20% price reduction treatment, booking rate. Bottom right: -20% price reduction treatment, average revenue per listing.

exiting the marketplace due to a lack of sufficiently "high quality" listings more often in the baseline case. A two-sided t-test between the baseline distribution of booking rates and the counterfactual distribution of booking rates under the -50% price reduction treatment yields a t-statistic of $t = 1.12$ ($p = 0.26$). Hence, I am unable to reject the null hypothesis that $BR_c = BR_t$ at any conventionally accepted level of statistical significance for the -50% price treatment. For the -20% price case, while I fail to reject the null, the measured effect size is extremely small. In combination, these results suggest that the true effect on the booking rate under both treatments is equal to or close to zero[3]. Intuitively, this makes sense; regardless of the heterogeneous searcher preferences and the search algorithm that is drawn, under the simulation's model uniformly lowering price should result in approximately the same number of bookings in either the 0% or 100% treatment case. Any changes in booking

[3] If anything, we would expect the magnitude of the effect (if there were one) to be larger for the -50% price treatment than for the -20% price treatment.

rate should come purely as a result of more or fewer searchers exiting the marketplace due to a failure to find listings above their reserve quality.

Two-sided t-tests between the baseline distribution of average revenue per listing and the average revenue per listing under the two treatment cases, on the other hand, are easily able to reject the null hypotheses that $ARPL_c = ARPL_t$. A two-sided t-test between the baseline distribution of average revenue per listing and the counterfactual distribution of average revenue per listing under the -50% price reduction treatment yields a t-statistic of $t = 45.86$ ($p < 2.2 \times 10^{-16}$). A two-sided t-test between the baseline distribution of average revenue per listing and the counterfactual distribution of average revenue per listing under the -20% price reduction treatment yields a t-statistic of $t = 14.54$ ($p < 2.2 \times 10^{-16}$). Again, intuitively, this makes sense; assuming that the booking rate does in fact remain constant between the two treatments, the listings that do get booked after all listings have lowered their prices will make strictly less money. I take as point estimates of the ATE in both cases the differences in distribution means. A summary of these findings can be found in Table 3.

Table 3: True counterfactual outcome distributions and ATEs

	μ_C	σ_C	μ_T	σ_T	t	p	ATE
-50% (revenue)	\$7.25	\$2.26	\$3.65	\$1.13	45.86	$< 2.2 \times 10^{-16}$	-\$3.60
-50% (bookings)	0.024	0.002	0.024	0.002	2.17	0.20	0
-20% (revenue)	\$7.25	\$2.26	\$5.93	\$1.86	14.54	$< 2.2 \times 10^{-16}$	-\$1.33
-20% (bookings)	0.024	0.002	0.024	0.002	1.12	0.03	1.05×10^{-5}

Having quantified the true value of the estimands we want to measure (the ATE on booking rate and average revenue per listing), the next step is to determine the bias and RMSE of estimates achieved under different combinations of treatment assignment methods and ATE estimators. The first estimation method considered is one that is totally ignorant to test-control interference; an experiment with independent randomization that calculates a difference-in-means estimator. 1,000 simulations of such an experiment are run. In each simulation, 50% of units are assigned to the treatment and 50% are assigned to the control. The left column of Figure 3 show the ATEs for booking rate and average revenue per listing under both the -50% price treatment and the -20% price treatment. It is apparent from looking at these distributions and comparing them to the ATEs in Table 3 that these estimators are biased. As expected, ATEs for bookings overstate effect size (treatment listings cannibalize bookings from control), whereas ATEs for revenue per listing understate the negative effect (a higher relative proportion of listings receive bookings, leading to a higher average revenue). Table 4 provides the bias, RMSE, and distributions for these ATE estimators.

This baseline case can be compared to the case where each unit is assigned to a treatment using graph cluster randomization and block random assignment, and the difference-in-means estimator is used. Clusters and blocks are generated according to the methodology described in Section 3. One cluster in each block is assigned to the treatment, and each vertex receives its cluster's assignment. The right column of Figure 3 show the ATE distributions for booking rate and average revenue per listing under this design and analysis methodology for both the -50% price treatment and the -20% price treatment. The bias of the difference-in-means

Table 4: ATE mean, standard deviation, bias, and RMSE under random assignment and difference-in-means

	μ_{ATE}	σ_{ATE}	Bias	RMSE
-50% (revenue)	-\$0.86	\$1.95	\$2.74	\$3.37
-50% (bookings)	0.005	.003	0.005	0.007
-20% (revenue)	-\$0.31	\$2.18	\$1.02	\$2.41
-20% (bookings)	0.002	.003	0.002	0.004

estimator is uniformly decreased. However, there are large increases in variance. Thus, the net effect is increased RMSE. Table 5 provides the mean, standard error, bias and RMSE for these ATE estimators.

Table 5: ATE mean, standard deviation, bias, and RMSE under graph cluster assignment, black random assignment, and difference-in-means

	μ_{ATE}	σ_{ATE}	Bias	RMSE
-50% (revenue)	-\$1.47	\$2.97	\$2.13	\$3.65
-50% (bookings)	0.003	0.007	0.003	0.008
-20% (revenue)	-\$0.68	\$3.45	\$0.65	\$3.51
-20% (bookings)	0.001	0.007	0.001	0.007

Another augmentation that can be considered is to pair graph cluster randomization and block random assignment with an analysis procedure that still uses the difference-in-means estimator, but only considers vertices that are subjected to the effective treatment or effective control (as defined by some exposure model). I estimate the ATE under this scheme using the FNTR exposure model (as described in Section 2), with three different values of λ: $\lambda = .5, \lambda = .75$, and $\lambda = .95$. Again, within each block one cluster is assigned to the treatment, and each vertex receives its cluster's assignment. Figure 4 shows the ATE distributions for these estimators. The impact of this approach is unclear. For all values of λ, the effect on bias is weakly positive (bias is either unchanged or decreases). For small values of λ ($\lambda \in \{.5, .75\}$), the effect on RMSE is also weakly positive (RMSE is unchanged or smaller). However, for $\lambda = .95$ the effect on RMSE is weakly negative. This may indicate that large values of λ cause the exposure model to throw out too many listings and/or select out a non-representative sample (since only listings with $\pi(z_0) > 0$ and $\pi(z_1) > 0$ according to the exposure model are compared). Even when the introduction of the exposure model is weakly positive, the reduction in bias is modest compared to the initial gains from graph cluster randomization and block random assignment, and the RMSEs obtained are still large relative to those for the difference-in-means estimator under independent assignment. Table 6 provides the mean, standard error, bias and RMSE for these ATE estimators.

The final combination of experimental design and analysis techniques considered is the combination of graph cluster randomization and block random assignment, the FNTR exposure model, and the Hajek estimator detailed in Section 2. The same FNTR thresholds

Table 6: ATE mean, standard deviation, bias, and RMSE under graph cluster assignment, block random assignment, and difference-in-means for units in the effective treatment or effective control (FNTR)

	λ	μ_{ATE}	σ_{ATE}	Bias	RMSE
-50% (revenue)	0.5	-$1.49	$2.92	$2.11	$3.61
-50% (bookings)	0.5	0.003	0.007	0.003	0.008
-20% (revenue)	0.5	-$0.69	$3.41	$0.64	$3.47
-20% (bookings)	0.5	0.001	0.007	0.001	0.007
-50% (revenue)	0.75	-$1.56	$2.93	$2.04	$3.57
-50% (bookings)	0.75	0.003	0.008	0.003	0.008
-20% (revenue)	0.75	-$0.72	$3.41	$0.61	$3.47
-20% (bookings)	0.75	0.001	0.008	0.001	0.008
-50% (revenue)	0.95	-$1.60	$3.05	$2.00	$3.65
-50% (bookings)	0.95	0.003	0.008	0.003	0.008
-20% (revenue)	0.95	-$0.74	$3.54	$0.59	$3.59
-20% (bookings)	0.95	0.001	0.008	0.001	0.008

of $\lambda = .5, \lambda = .75$, and $\lambda = .95$ are considered. One cluster in each block is assigned to the treatment, and each vertex receives its cluster's assignment. Figure 5 shows the ATE distributions for these estimators. The introduction of the Hajek estimator seems neutral to negative. In no case are the bias or RMSE of the Hajek estimator better than what is obtained using simple difference in means. This may be a combination of the same sample issues discussed with regards to the the exposure model difference-in-means estimator, coupled with the finite sample bias of the Hajek estimator. To study this further, future work might focus on the use of the Horvitz-Thompson estimator rather than the Hajek estimator. Table 7 provides the bias, RMSE, and distribution of these ATE estimators.

Overall, the combination of graph cluster randomization and block random assignment seems to lead to large reductions in bias from test-control interference (at the cost of increases in estimator variance and RMSE). Further reductions in bias and slight reductions in RMSE can be achieved through the use of an exposure model (i.e., the FNTR model). However, RMSE under the exposure model is still greater than RMSE given a simple difference-in-means estimator and random assignment. The introduction of the Hajek estimator seems to have a negative effect on both bias and RMSE.

6 Conclusions

This paper considers the efficacy of tools proposed in Eckles et al [5] in reducing bias in online marketplace experimentation due to test-control interference. In order to test how well these designs and analysis techniques reduce bias, I simulate a toy marketplace experiment inducing sellers on Airbnb to lower prices. Two different effect sizes are considered - a price reduction of 50% and a price reduction of 20%. While a price reduction of 20% is more realistic, test-control interference effects are larger in the 50% case. This paper focuses on the

Table 7: ATE mean, standard deviation, bias, and RMSE under graph cluster assignment and block random assignment using the Hajek estimator for units in the effective treatment or effective control (FNTR)

	λ	μ_{ATE}	σ_{ATE}	Bias	RMSE
-50% (revenue)	0.5	-\$1.19	\$2.94	\$2.41	\$3.81
-50% (bookings)	0.5	0.004	5.4×10^{-5}	0.004	0.008
-20% (revenue)	0.5	-\$0.933	\$3.45	\$1.00	\$3.60
-20% (bookings)	0.5	0.001	5.4×10^{-5}	0.001	0.007
-50% (revenue)	0.75	-\$1.47	\$2.90	\$2.13	\$3.60
-50% (bookings)	0.75	0.003	0.009	0.003	0.01
-20% (revenue)	0.75	-\$0.63	\$3.45	\$0.70	\$3.52
-20% (bookings)	0.75	0.001	8.7×10^{-5}	0.001	0.009
-50% (revenue)	0.95	-\$1.41	\$2.97	\$2.19	\$3.69
-50% (bookings)	0.95	0.003	0.009	0.003	0.01
-20% (revenue)	0.95	-\$0.52	\$3.51	\$0.81	\$3.60
-20% (bookings)	0.95	0.002	0.009	0.002	0.009

effect this experiment has on seller bookings and seller revenue. A network of sellers whose goods are likely to substitute for each other is initialized using a simple heuristic. Simulations show that the combination of graph cluster randomization and block random assignment significantly reduce bias from test-control interference. However, this comes at the cost of increased RMSE. While the use of an exposure model (i.e., the FNTR model) does lead to additional (modest) reductions in bias and small reductions in RMSE, reductions in bias are small compared to those obtained through graph cluster randomization/block random assignment, and RMSE values remain far above RMSE for the difference-in-means ATE estimator under independent randomization. FNTR exposure models with high λ values lead to increases in bias and RMSE. Further, the use of the Hajek estimator for the ATE also leads to increases in bias and RMSE. This may be caused by a small or unrepresentative sample that is obtained when the FNTR exposure model threshold is too high. Furthermore, the Hajek estimator may lead to increased bias due to the estimator's finite sample bias.

The results in this paper represent a first step in evaluating how well methods developed for the estimation of average treatment effects in the presence of peer effects might reduce bias due to test-control interference in online marketplace experiments. There are a number of follow-ups that can and should be performed. First, the current simulation relies heavily on a number of researcher degrees of freedom. The distributions from which both searcher and search engine parameters are drawn are admittedly arbitrary. Miami may be a unique Airbnb market. Results may change when a different clustering algorithm is used. It would be worthwhile to redo this analysis with different choices and determine how robust and generalizable these findings are. Specifically in the case of searcher and search algorithm parameters, it would be optimal to redo this analysis with values and/or mechanisms that more closely resemble the actual booking process on Airbnb.

Additionally, the current work focuses on the case where (by design) edges are constructed in a way that is well-aligned with the parameters upon which algorithms and searchers make their decisions. It is worth understanding how robust these results are to cases where researchers may ‘mis-specify‘ the network. Furthermore, one might wonder how well this methodology works when one of the variables that determines item substitutability is exogenously varied in the experimental treatment (e.g., price). Finally, it may be interesting to consider how well the proposed techniques work in cases where items can be complements (in addition to substitutes). For instance, in a marketplace that sells both record players and records, there may be *positive* spillovers between units in the treatment and control arms of an experiment.

Despite the large amount of future work required, the current results suggest that relatively straightforward experiment design and analysis improvements (i.e., graph cluster randomization, block random assignment, and the use of exposure models) can provide researchers and practitioners with a relatively lightweight way to reduce bias in average treatment effect estimates in online marketplace experiments. Unfortunately, these methods also lead to non-trivial increases in the RMSE of the ATE estimator. In short, future work should validate these results, and focus on ways to maintain the achieved reductions in bias while minimizing the resulting increases in RMSE.

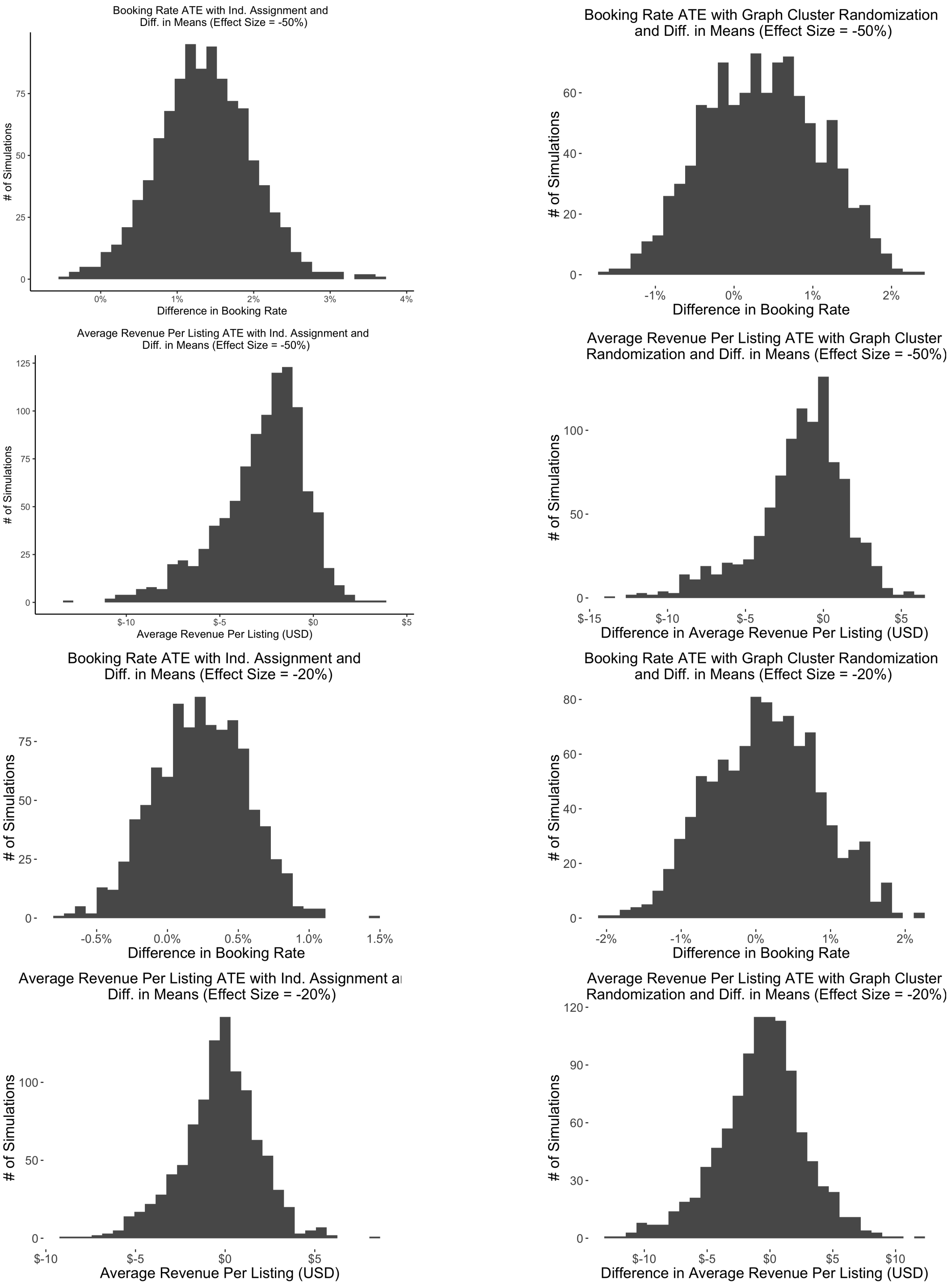


Figure 3: Distribution of the ATE on booking rate and average revenue per listing over 1,000 simulations using a simple difference-in-means estimator. Left panels correspond to ATE using independent randomization. Right panels correspond to ATE using graph cluster randomization and block random assignment. First two rows show ATE distributions for effect size of -50%. Third and fourth rows show ATE for effect size of -20%. Graph cluster randomization and block random assignment reduce bias, but increases variance.

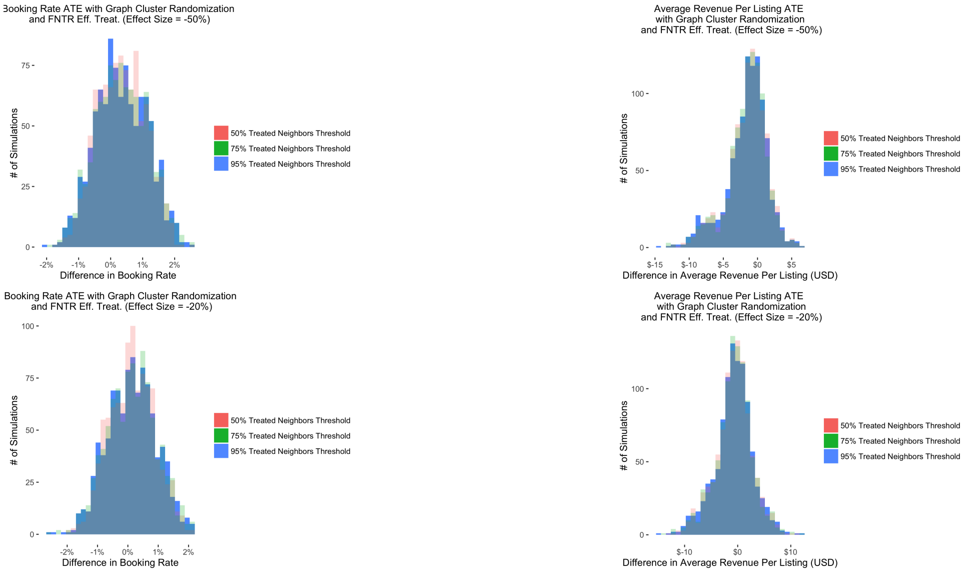


Figure 4: The distribution of ATE measurements over 1,000 simulations when graph cluster randomization and block random assignment are used along with a difference in means estimator for only the units in the 'effective treatment.' Effective treatments correspond to an FNTR exposure model with $\lambda = .5$, $\lambda = .75$ and $\lambda = .95$. Top left: -50% price reduction treatment, booking rate. Top right: -50% price reduction treatment, average revenue per listing. Bottom left: -20% price reduction treatment, booking rate. Bottom right: -20% price reduction treatment, average revenue per listing.

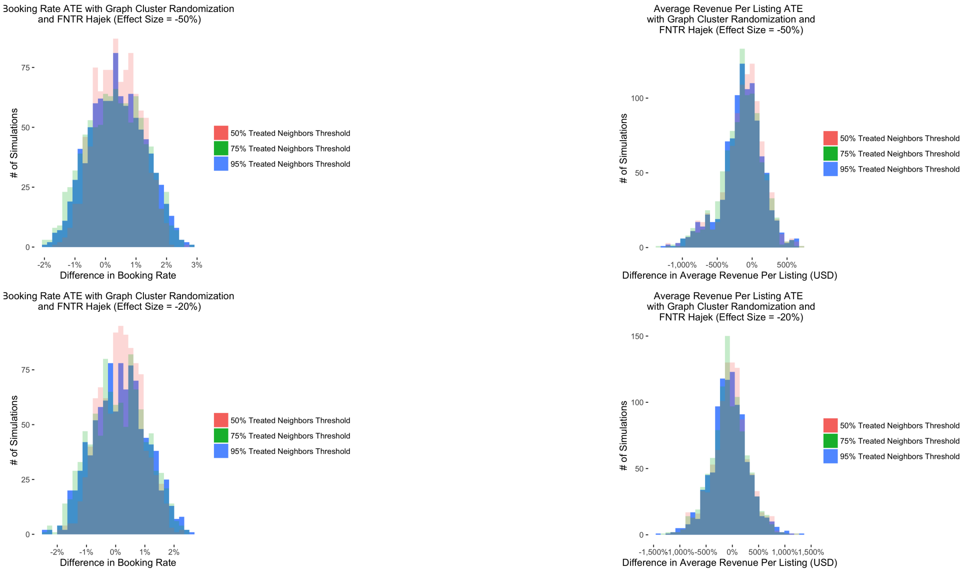


Figure 5: The distribution of ATE measurements over 1,000 simulations when graph cluster randomization and block random assignment are used along with a Hajek estimator for only the units in the 'effective treatment.' Effective treatments correspond to an FNTR exposure model with $\lambda = .5$, $\lambda = .75$ and $\lambda = .95$. Top left: -50% price reduction treatment, booking rate. Top right: -50% price reduction treatment, average revenue per listing. Bottom left: -20% price reduction treatment, booking rate. Bottom right: -20% price reduction treatment, average revenue per listing.

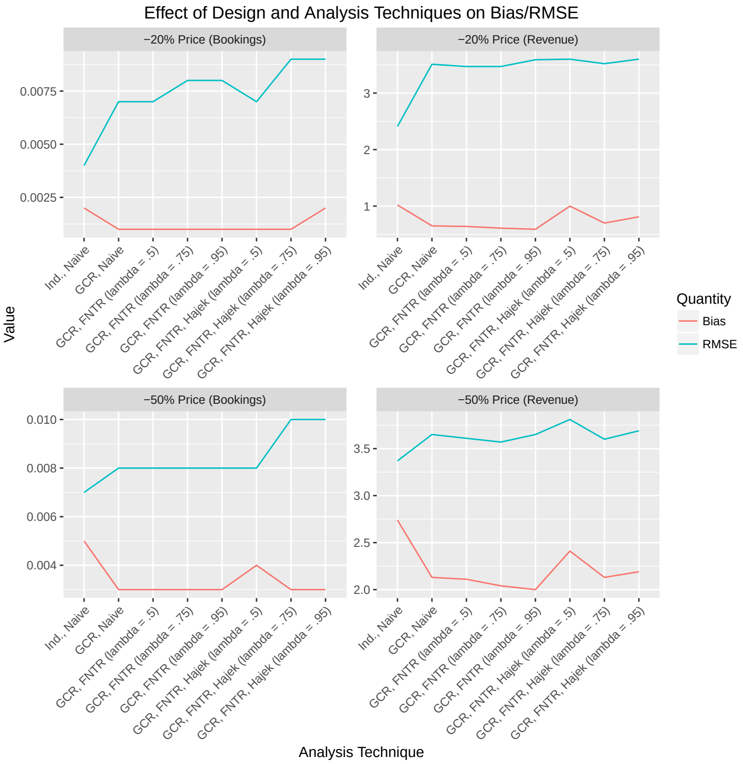


Figure 6: Changes in bias and RMSE of ATE estimates using different experiment designs and ATE estimators. Top panels correspond to -20% price change treatment. Bottom panels correspond to -50% price change treatment. Left panels correspond to booking rate ATE. Right panels correspond to average revenue per listing ATE. In general, graph cluster randomization reduces bias, but increases variance. FNTR exposure models with small λ can provide further bias reduction with neutral to positive (albeit small) effect on RMSE. The introduction of the Hajek estimator appears to increase both bias and RMSE

Acknowledgements

The author would like to thank Dean Eckles, Sinan Aral and Andrey Fradkin for their useful comments. I would also like to thank the participants at the 2016 MIT Conference on Digital Experimentation and the MIT Simple Person's Applied Math Seminar for their useful questions and feedback. I'd also like to extend my gratitude to Tom Slee for publicly sharing Airbnb data that he has periodically scraped from the site.

References

[1] Peter M Aronow and Cyrus Samii. Estimating average causal effects under general interference. In *Summer Meeting of the Society for Political Methodology, University of North Carolina, Chapel Hill, July*, pages 19–21. Citeseer, 2012.

[2] Thomas Blake and Dominic Coey. Why marketplace experimentation is harder than it seems: The role of test-control interference. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 567–582. ACM, 2014.

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[4] David Roxbee Cox. Planning of experiments. 1958.

[5] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530*, 2014.

[6] Andrey Fradkin. Search frictions and the design of online marketplaces. Technical report, Working Paper, 2014.

[7] Alan S Gerber and Donald P Green. *Field experiments: Design, analysis, and interpretation*. WW Norton, 2012.

[8] Jaroslav Hájek. Comment on an essay on the logical foundations of survey sampling, part one,. *The Foundations of Survey Sampling*, 236, 1971.

[9] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[10] David Kahle and Hadley Wickham. ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013.

[11] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

[12] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

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

[14] Ryan T Moore. Multivariate continuous blocking to improve political science experiments. *Political Analysis*, 20(4):460–479, 2012.

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

[16] Tom Slee. Airbnb data collection: Methodology and accuracy, 2015.

[17] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon 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.