Reducing Interference Bias in Online Marketplace Pricing Experiments

David Holtz

MIT Sloan School of Management, Cambridge, MA 02142, dholtz@mit.edu

Ruben Lobel, Inessa Liskovich

Airbnb Inc., San Francisco, CA 94103, ruben.lobel@airbnb.com inessa.liskovich@airbnb.com

Sinan Aral

MIT Sloan School of Management, Cambridge, MA 02142, sinan@mit.edu

Online marketplace designers frequently run A/B tests to measure the impact of proposed product changes. However, given that marketplaces are inherently connected, total average treatment effect estimates obtained through Bernoulli randomized experiments are often biased due to violations of the stable unit treatment value assumption. This can be particularly problematic for experiments that impact sellers' strategic choices, affect buyers' preferences over items in their consideration set, or change buyers' consideration sets altogether. In this work, we measure and reduce bias due to interference in online marketplace experiments by using observational data to creating clusters of similar listings, and then using those clusters to conduct cluster-randomized field experiments. We provide a lower bound on the magnitude of bias due to interference by conducting a meta-experiment that randomizes over two experiment designs: one Bernoulli randomized, one cluster randomized. In both meta-experiment arms, treatment sellers are subject to a different platform fee policy than control sellers, resulting in different prices for buyers. By conducting a joint analysis of the two meta-experiment arms, we find a large and statistically significant difference between the total average treatment effect estimates obtained with the two designs, and estimate that 32.60% of the Bernoulli-randomized treatment effect estimate is due to interference bias. We also find weak evidence that the magnitude and/or direction of interference bias depends on extent to which a marketplace is supply- or demand-constrained, and analyze a second meta-experiment to highlight the difficulty of detecting interference bias when treatment interventions require intention-to-treat analysis.

Key words: Design of experiments, Electronic markets and auctions, Interference, Cluster randomization, Airbnb

1. Introduction

As of 2020, some of the world's most highly valued technology firms (e.g., Airbnb, Uber, Etsy) are online peer-to-peer marketplaces. These platforms create markets for many different types of goods, including accommodations, transportation, artisanal goods, and dog walking. Like almost

all technology firms, online peer-to-peer marketplaces typically rely on experimentation, or A/B testing, to measure the impact of proposed changes to the platform and develop a deeper understanding of their customers. However, a randomized experiment's ability to provide an unbiased estimate of the total average treatment effect (TATE) relies on the stable unit treatment value assumption (SUTVA) (Rubin 1974), sometimes referred to as the "no interference" assumption (Cox 1958). Online marketplaces are inherently connected; sellers are likely to make strategic decisions based on the actions of their competitors, and multiple sellers may sell different items that complement or substitute for one another. As a result, SUTVA is unlikely to hold in online marketplace settings. Previous work (Blake and Coey 2014, Fradkin 2015, Holtz 2018) has shown that naive experimentation in online marketplaces can lead to TATE estimates that are overstated by up to 100%.

SUTVA violations are not unique to online marketplaces, and are a familiar problem for researchers conducting experiments in networked settings (e.g., social network experiments). In the network experimentation literature, researchers have proposed experiment designs (Eckles et al. 2017, Ugander et al. 2013) and analysis techniques (Aronow and Samii 2012, Eckles et al. 2017) that aim to reduce bias due to statistical interference (henceforth referred to as interference bias), and Saveski et al. (2017) describe a procedure for "randomizing over randomized experiments," or running meta-experiments, to detect interference bias on networks. Holtz (2018) proposes the use of bias-reduction techniques from the networks literature to reduce bias in online marketplace experiments, and investigates the viability of this approach through a simulation study using scraped Airbnb data. However, this approach has, as of yet, not been used in the field to conduct randomized experiments in online marketplaces.

In this paper, we present the results from two meta-experiments conducted on Airbnb, an online marketplace for sharing homes and experiences. Both meta-experiments make use of clusters of Airbnb listings, which are created by first using observational search behavior to create a 16-dimension "demand embedding" for each each Airbnb listing, and then segmenting the listing embedding space using a recursive partitioning tree. Each meta-experiment randomly assigns clusters of Airbnb listings to one of two randomization schemes; 25% of clusters are Bernoulli randomized (i.e., treatment assignment is randomly assigned at the listing level), whereas the remaining 75% of clusters are cluster randomized (i.e., treatment assignment is randomly assigned at the cluster level). Both of the meta-experiments we present are related to pricing on Airbnb. We focus on pricing-related treatment interventions for two reasons. First, it is crucial for both hosts and the platform intermediary to understand the price elasticity of Airbnb guests; hosts set the price of their listings, while Airbnb recommends prices to hosts and sets platform fees. Second, TATE estimates for pricing-related experiments are likely to be affected by interference bias, since hosts

observe other hosts' prices and guests usually consider many listings before choosing a listing to book.

The first meta-experiment measures the effect of a change to Airbnb's platform fee structure. In the treatment group, long-tenured hosts were subject to a platform guest fee increase, while the platform guest fee for less tenured hosts remained unchanged. In the control group, long-tenured hosts were subject to a platform guest fee *decrease*, while the platform guest fee for less tenured hosts remained unchanged. Results from the Bernoulli randomized meta-treatment arm suggest that the treatment led to a statistically significant loss of 0.207 bookings per listing over the course of the experiment.[1] However, a joint analysis of the entire meta-experimental sample finds that there is a statistically significant difference between the TATE estimates obtained in the two meta-treatment arms. We estimate that 32.60% of the Bernoulli TATE estimate on bookings is attributable to interference bias. While not statistically significant, we also report results that suggest that interference bias is more severe in markets that are demand constrained than in markets that are supply constrained.

Results from the fee meta-experiment establish the existence of interference bias in online marketplaces, and the efficacy of cluster randomization in reducing that bias. However, the guest platform fee treatment intervention is one that affects all hosts on Airbnb. Often, online marketplace designers are interested in the effect of behavioral nudges, which only cause a change in the behavior of some users. These experiments are typically analyzed with intention-to-treat (ITT) analysis. To test for interference bias in an experiment that requires ITT analysis, we conduct a second meta-experiment that measures the effect of a proposed update to the algorithm underlying Airbnb's price suggestions for hosts. On average, the treatment *increased* the prices suggested to hosts. Results from the Bernoulli randomized meta-treatment arm suggest that the treatment led to a statistically significant loss of 0.106 bookings per listing over the course of the experiment. In the cluster randomized meta-treatment arm, this treatment effect disappears; the point estimate is smaller in magnitude, and not statistically significant. However, a joint analysis of the entire meta-experimental sample fails to detect a statistically significant difference between the two sets of treatment effect estimates. Post-hoc power analysis reveals that the meta-experiment is underpowered to detect interference bias that is not extremely severe in magnitude. Although not statistically significant, our point estimates suggest that in the Bernoulli randomized pricing experiment, 54.16% of the observed treatment effect is due to interference bias. This result highlights the difficulty of detecting interference bias when a given treatment intervention only affects some users, even if the magnitude of that bias is potentially large.

---

[1] To avoid disclosing raw numbers, all raw booking, nights booked, and gross guest spend values are multiplied by a constant.

While previous research has focused on quantifying the magnitude of interference bias through simulation (Fradkin 2015, Holtz 2018) or post-hoc analysis (Blake and Coey 2014), this work is among the first empirical papers to focus on reducing interference bias in a marketplace experiment through experiment design. The experiment design techniques we employ are strongly influenced by the network experimentation literature (Eckles et al. 2017, Ugander et al. 2013, Saveski et al. 2017), and future extensions of our work might focus on adopting analysis-based approaches to reducing interference bias in network experiments (Athey et al. 2018, Aronow and Samii 2012, Eckles et al. 2017, Chin 2018) to an online marketplace setting. Future work might also focus on how to best cluster items or sellers in a marketplace. Clustering items or sellers in an online marketplace is difficult, as there is often no explicit network structure indicating which items are likely to substitute or complement for one another,[2] and measuring cross-price elasticities in markets with millions of heterogeneous goods is difficult.

The rest of this paper proceeds as follows. In Section 2, we review the related literature. In Section 3, we describe in greater detail the features of Airbnb's platform that are relevant to the two meta-experiments presented in this paper. Our meta-experiment design is described in Section 4. We present results from the fee experiment in Section 5, and results from the pricing algorithm experiment in Section 6. Finally, we discuss our findings and future extensions in Section 7.

2. Related Literature

The research in this paper connects to three bodies of academic literature: one on interference bias in online marketplace experiments, one on experimentation in networks, and one on pricing-related online marketplace interventions.

Our work is most closely related to recent research that has shown that naive marketplace experimentation can yield total average treatment effect estimates that are overstated by up to 100% (Blake and Coey 2014, Fradkin 2015, Holtz 2018). Blake and Coey (2014) arrive at this conclusion through post-hoc analysis of an experiment conducted on eBay, while Fradkin (2015) finds evidence for interference bias through a simulation of Airbnb's marketplace that has been calibrated using search and transaction data from the firm. Finally, Holtz (2018) also shows through a simple simulation of marketplace experiments on Airbnb that naive marketplace experiments are biased due to interference, and that the magnitude of this bias can be reduced through experiment design and analysis techniques.

Bias in total average treatment estimates due to statistical interference is not a problem unique to online marketplace experiments. In fact, there has been substantial research on experiment design

---

[2] When part of an online market's design, recommendation networks (Oestreicher-Singer and Sundararajan 2012a,b) do provide an explicit product network.

and analysis techniques that provide unbiased TATE estimators in settings where the stable unit treatment value assumption (Rubin 1974) is violated.[3] SUTVA assumes that the potential outcomes of a given unit of analysis are independent of the treatment assignments other units receive. However, in many settings (e.g., networks, marketplaces) SUTVA is unlikely to hold. When SUTVA is violated, the TATE estimated from a Bernoulli randomized experiment can differ substantially from the actual TATE (i.e., the average effect of the treatment under the counterfactual that every unit is treated). Network science researchers have developed experiment designs (Ugander et al. 2013, Eckles et al. 2017) and treatment effect estimators (Aronow and Samii 2012, Chin 2018) that eliminate or reduce bias due to SUTVA violations arising from network interference.

Ugander et al. (2013) propose graph cluster randomization (GCR) as an experiment design for reducing interference bias in networked experiments. In GCR, a network is first clustered, then randomized at the *cluster*-level. This can greatly reduce the probability that any ego's experimental treatment assignment is different from the treatment assignment of its alters. This will reduce the extent to which statistical interference affects experimental TATE estimates. Through simulations, Eckles et al. (2017) show that GCR can be effective in reducing interference bias in networked experiments, even when the network does not satisfy the strict requirements requirements outlined in Ugander et al. (2013). One drawback of assigning treatment at the cluster-level is that most treatment effect estimators will provide less statistical power than they would have under a Bernoulli randomized design. However, techniques such as regression adjustment and pre-stratification (Moore 2012) can be used in tandem with GCR to mitigate the loss of statistical power. Graph cluster randomization can also be used to test whether or not interference bias affects the TATE estimates obtained from a given experiment. Saveski et al. (2017) conduct a "Meta-experiment" on LinkedIn, which randomizes over two experiment designs (Bernoulli randomization and cluster randomization). By comparing the treatment effect estimates obtained in each meta-treatment arm, they are able to test for the existence of network interference for any experiment conducted on LinkedIn.

Finally, our work also connects to the literature on pricing-related online marketplace interventions. A number of recent empirical papers measure the effects of pricing-related interventions on online platforms (Dubé and Misra 2017, Filippas et al. 2019). Airbnb itself uses a customized regression model to provide pricing recommendation to hosts (Ifrach et al. 2016, Ye et al. 2018). It is crucial for both platform intermediaries and platform sellers to understand the price elasticity of their customers; sellers would like to price effectively, whereas intermediaries would like to implement effective fee structures and pricing-related market mechanisms. However, TATE estimates

---

[3] SUTVA is sometimes alternatively referred to as the 'no interference' assumption (Cox 1958).

obtained through naive experimental tests of pricing-related interventions will likely yield biased estimates of price elasticity, since marketplace sellers compete with one another, and observe each others' pricing decisions.

This paper builds on prior research by adapting experiment design techniques from the networks literature (Ugander et al. 2013, Eckles et al. 2017, Holtz 2018) and conducting meta-experiments (Saveski et al. 2017) in an online marketplace to test for the existence of interference bias. Developing methods for obtaining accurate TATE estimates in online marketplace settings is increasingly important as both researchers and practitioners continue to explore novel pricing-related interventions (Dubé and Misra 2017, Filippas et al. 2019) in online marketplace settings.

3. Setting

Airbnb is an online marketplace for accommodations and experiences. More than five million listings appear on Airbnb, and since the company's founding in 2008, over 400 million guest arrivals have occurred on the platform. On average, over two million people are staying in Airbnb listings on a given night (Airbnb 2019).

3.1. Platform Guest Fees

Airbnb earns revenue by collecting fees from guests and hosts for every transaction that occurs on the platform. In order to set fees optimally, it is crucial for the platform to understand guest price elasticity. Airbnb's fees for guests are visible in three different locations throughout the booking process. First, guest platform fees are included in the total price shown to guests when a listing appears in search. Figure 1 shows a typical Airbnb search result. Second, if a guest opens a tooltip on any search result, they are shown a price breakdown that separates the listing's nightly price and the guest platform fee. Figure 2 shows this tooltip. Finally, when viewing a listing's product detail page, a detailed pricing breakdown (including fees) is displayed next to the "Request to Book" button. Figure 3 shows this price breakdown.

3.2. Price Tips & Smart Pricing

Since the summer of 2015, Airbnb has provided tools to help hosts price more effectively. In June 2015, Airbnb launched "Price Tips," a feature that provides dynamic pricing suggestions for hosts (Airbnb 2015b). In November 2015, Airbnb launched "Smart Pricing," a tool that automatically updates hosts' prices subject to a set of constraints determined by the host (Airbnb 2015a). Both "Price Tips" and "Smart Pricing" present recommendations from the same machine learning model, which incorporates local supply and demand features to provide dynamic pricing suggestions to hosts (Ifrach et al. 2016, Ye et al. 2018). We refer the reader to Ye et al. (2018) for a more detailed description of the pricing algorithm itself. Importantly, Airbnb's pricing suggestions attempt to maximize each host's individual objectives, rather than playing the role of a central planner.

"Price tips" color codes nights on a host's calendar based on the estimated probability that a given night will be booked given the current price, and suggests an "optimal" price for each night. Importantly, "Price tips" requires hosts to manually accept prices in order to comply with the algorithm's suggestions recommended through the "Price tips" product. A screenshot of the "Price tips" UI is shown in Figure 4. "Smart pricing" was introduced to make it easier for hosts to adopt Airbnb's pricing recommendations en masse. Once "Smart pricing" is turned on, hosts automatically adopt Airbnb's recommended price, subject to constraints provided by the host. A screenshot of the "Smart Pricing" UI is shown in Figure 5.

4.   Experiment Motivation & Design

It is crucial for an online marketplace intermediary, such as Airbnb, to understand the price elasticity of its customers. This enables the firm to implement optimal pricing-related market mechanisms, such as fee structures and seller pricing suggestions. Understanding customer price elasticities can also be beneficial to sellers, who set their own prices. If the business outcomes of all Airbnb listing were independent, the firm could take an atheoretic approach to estimating price elasticity by running a randomized controlled trial, or A/B test, in which the prices of some listings were exogenously increased or decreased. However, as described in Holtz (2018), host- or listing-level experiments on Airbnb violate SUTVA due to the inherent interconnectedness of online marketplaces.

There are a number mechanisms that can lead to SUTVA violations on Airbnb. For one, if some hosts lower (raise) their prices, they may increase (decrease) demand for their listings, and, consequently, decrease (increase) demand for their competitors' listings. Furthermore, host pricing decisions may exhibit viral properties; a host may observe their competitor's pricing behavior, and copy it. Finally, Airbnb listings in a given market can also serve as complements to each other. For instance, guests may describe their positive (negative) experience with a given listing to their peers, which could increase (decrease) demand for similar listings.

Adapting experiment design and analysis techniques from the network experimentation literature, as proposed by Holtz (2018), is one avenue for reducing interference bias in online marketplace pricing experiments. However, none of the techniques put forward by Holtz (2018) have been used yet to design or analyze an online marketplace experiment. As a first step toward empirically confirming the existence of interference bias in online marketplace experiment TATE estimates, and measuring the extent to which cluster randomization, an experiment design technique, can reduce that bias, we conduct pricing-related meta-experiments (Saveski et al. 2017) on Airbnb. Quantifying the magnitude of interference bias, as well as the extent to which cluster randomization can reduce that bias, is useful for two reasons. First, even if interference bias is a theoretical concern,

it may not be a practical one; statistical bias in TATE estimates due to interference may be small. Second, even if interference bias is large, cluster randomization may not be an effective tool to reduce that bias. If this were the case, cluster randomization would not be a worthwhile undertaking for firms; cluster randomization results in reduced statistical power relative to Bernoulli randomization, and is also more logistically complicated for firms to implement (both because of the need to identify relevant clusters, and because most corporate A/B testing tools do not support cluster randomization).

In each meta-experiment, Airbnb listings are arranged into clusters. Each of these clusters is then assigned to one of two meta-treatment arms: Bernoulli randomization, or cluster randomization. Within the Bernoulli-randomized meta-treatment arm, treatment is randomly assigned at the listing level. Within the cluster-randomized meta-treatment, treatment is randomly assigned at the cluster level. By jointly analyzing the data from both meta-treatment arms, we are able to measure whether there is a statistically significant difference between the TATEs measured separately in each meta-treatment arm.

4.1.   Treatment Assignment Mechanism

In this subsection, we describe the procedure used to arrange Airbnb listings into clusters, and then subsequently determine a given listing's meta-treatment assignment and treatment assignment.

4.1.1.   Clusters of Airbnb Listings To perform cluster randomization, it is first necessary to arrange all of Airbnb's listings into mutually exclusive clusters. Previous work (Holtz 2018) has proposed creating a network of listings that substitute for or complement one another, and then clustering that network with any of a number of graph clustering algorithms (e.g., Louvain clustering (Blondel et al. 2008)). In this subsection, we outline a different approach to clustering, which we use to generate our listing clusters. We first generate a dense, 16-dimensional demand embedding for each listing, and then cluster listings based on their location in that 16-dimensional space. Our method for generating Airbnb listing embeddings is similar to that described in Grbovic and Cheng (2018).

Our embeddings are trained on data consisting of sequences of listings that individual users view in the same search session. If, for instance, a user viewed listings L_A, L_B, and L_C in one search session, this would generate the sequence:

$$< L_A, L_B, L_C > . \tag{1}$$

We use a word2vec-like architecture (Mikolov et al. 2013b) to estimate a skip-gram model (Mikolov et al. 2013a) on this data. Given S sequences of listings, the skip-gram model attempts to maximize the objective function

$$J = \max_{W,V} \sum_{s \in S} \frac{1}{|s|} \sum_{i=1}^{|s|} \sum_{-k \leq j \leq k,\, k \neq 0} \log p\left(L_{i+j}|L_i\right), \tag{2}$$

where k is the size of a fixed moving window over the listings in a session, W and V are weight matrices in the word2vec architecture, and $p(L_{i+j}|L_i)$ is the hierarchical Softmax approximation to the regular softmax expression.

The objective function above is augmented by including listing-level attributes (e.g., a listing's market) in the search session sequences. The model is then trained using a market-level negative sampling approach. This generates a 16-dimensional vector representation for each Airbnb listing.

Once listing embeddings are estimated using the aforementioned approach, a recursive partitioning tree (Kang et al. 2016) is used to arrange the Airbnb listings into clusters. The algorithm starts from a single cluster containing all listings, and then recursively bisects clusters into two sub-clusters. The algorithm stops bisecting sub-clusters when the tree reaches a depth of 20, or when a new sub-cluster will contain less than 20 listings. Listings can then be assigned to clusters of arbitrary sizes by assigning them to the smallest sub-cluster to which they belong that has at least some threshold number of listings. For the algorithmic pricing meta-experiment, we set this threshold at 250 listings, whereas for the fee meta-experiment, we set this threshold at 1,000 listings.[4] Figure 6 depicts example clusters generated using this method in the Bay Area.

4.1.2. Pre-stratification & Treatment Assignment Once listings have been assigned to clusters, those clusters are given meta-treatment assignments and, based on those cluster-level meta-treatment assignments, listings are assigned listing-level treatment assignments.

To gain statistical power (particularly in the cluster-randomized meta-treatment arm), we group clusters into strata using a multivariate blocking procedure (Moore 2012). As a first step, we collected pre-treatment listing-level data.[5] We then aggregate data at the cluster level, and for each cluster calculate over the pre-treatment period the average number of nights booked per listing, the average number of bookings per listing, the average booking value per listing, and the number of experiment-eligible listings in the cluster.[6][7] After centering and scaling each of these metrics, we

---

[4] In choosing cluster sizes, we are attempting to balance two objectives: creating clusters that capture listings likely to interfere with one another, and designing an experiment with sufficient statistical power. Since ex ante, we expected the fee treatment intervention to have a larger effect, we chose larger clusters for that meta-experiment. For more details on the process used to determine cluster size, see Appendix A.

[5] For the fee meta-experiment, pre-treatment data was collected from January 16, 2019 to February 17, 2019. For the pricing algorithm experiment, pre-treatment data was collected from August 1, 2018 to September 25, 2018.

[6] Our experiment excludes listings in a long-term experiment holdout group, as well as listings in Airbnb's "Plus" tier.

[7] For the algorithmic pricing experiment, we also calculate the percentage of listings accepting at least one price tip during the pre-treatment period, and the percentage of listings with "Smart Pricing" enabled at the end of the pre-treatment period.

calculate the Mahalanobis distance between each pair of clusters. Finally, we use an optimal-greedy algorithm to arrange clusters into strata of size $n = 8$. At each step, the optimal-greedy algorithm finds the smallest "available" distance between two clusters[8], and assigns the two corresponding clusters to the same stratum.

Within each stratum, two clusters are assigned to the meta-control via complete random assignment. The remaining six clusters are assigned to the meta-treatment. Within the meta-control arm, Bernoulli randomization is used to assign 50% of listings to the treatment and 50% of listings to the control. Within the meta-treatment arm, three of the six clusters are assigned the treatment via complete random assignment. The remaining three clusters are assigned the control. Each listing in a meta-treatment cluster is assigned the treatment assignment corresponding to its cluster.

5. Fee Meta-experiment
5.1. Description

The fee meta-experiment ran from March 16, 2019 to March 21, 2019 on a population of 4,578,028 listings. Of those listings, 1,146,537 were assigned to the Bernoulli-randomized meta-treatment arm, and the remaining 3,431,491 were assigned to the cluster-randomized meta-treatment arm. Within the Bernoulli-randomized meta-treatment arm, 573,346 were assigned to the treatment and 573,191 listings were assigned to the control. Within the cluster-randomized meta-treatment arm, 2,982 clusters were assigned to the treatment and 2,982 clusters were assigned to the control, resulting in 1,720,147 listings assigned to the treatment and 1,711,344 listings assigned to the control. In total, across both meta-treatment arms, 2,293,493 listings were assigned to the treatment, and 2,284,535 were assigned to the control.

Figure E.1 shows the empirical CDFs for pre-treatment bookings, nights booked, and booking value across all four meta-treatment / treatment groups.[9] For each of these pre-treatment outcomes, the empirical CDFs are visually quite similar. Kolmogorov-Smirnoff tests comparing the distributions for each pre-treatment outcome in the meta-treatment and meta-control fail to reject the null of equal distributions for bookings ($p = 0.069$) and nights booked ($p = 0.647$), but do reject the null for booking value ($p = 0.021$). Kolmogorov-Smirnoff tests comparing the distributions for each pre-treatment outcome in the treatment and control groups of the meta-control arm fail to reject the null of equal distributions for bookings ($p = 1.000$), nights booked ($p = 0.627$), and booking value ($p = 0.883$). Kolmogorov-Smirnoff tests comparing the distributions for each pre-treatment outcome in the treatment and control groups of the meta-treatment arm fail to reject the null of equal distributions for bookings ($p = 0.295$), but *do* reject the null of equal distributions for nights

---

[8] A distance is "available" if that pair of clusters has not been used in a previous step.

[9] To avoid disclosing raw numbers, x-axis values are multiplied by a constant.

booked ($p = 0.012$) and booking value ($p = 0.001$). We believe that some KS tests fail because cluster randomization is a higher variance randomization procedure, due to both the smaller number of units over which randomization occurs and the correlation of pre-treatment outcomes among listings in the same cluster. Furthermore, booking value is a higher variance outcome than nights booked, which itself is a higher variance outcome than bookings.

In the fee meta-experiment, listings in the treatment had their fees increased if they were long-tenured listings (i.e., if they had been on the platform as of a certain cutoff date). Listings in the control had their fees *decreased* if they were long-tenured listings. In both treatment arms, less tenured listings (i.e., those created after the cutoff date) did not have their fees changed.[10] Conceptually, one can think of the treatment and control conditions of this meta-experiment as comparing the effect of two different fee-based incentive programs Airbnb might run. In the treatment group, new listings have lower fees (which could drive business to newer listings), whereas in the control, older listings have lower fees (which could reward long-time Airbnb hosts and reduce churn). After the conclusion of the fee meta-experiment, a "reversal experiment" was run from April 15, 2019 to April 22, 2019. In the reversal experiment, listings that had been assigned the treatment condition in the meta-experiment were assigned the control, and vice-versa. The purpose of the reversal experiment was to mitigate any negative impact of the meta-experiment on Airbnb hosts.

5.2. Results

In this section, we present results from the fee meta-experiment. We focus on a single outcome metric, bookings per listing, but the results for two alternative outcome metrics, nights booked per listing and gross guest spend per listing, are qualitatively similar and can be found in Appendix B.[11] Since, relative to the control, the treatment *increased* fees, we expect the TATE on bookings per listing to be negative.

We first present the results from separately analyzing the Bernoulli randomized arm of the meta-experiment and the cluster randomized arm of the meta-experiment. While the Bernoulli randomized arm will have ample statistical power, we expect its TATE estimate to suffer from interference bias. On the other hand, analysis of the cluster randomized arm should provide a less biased estimate of the TATE, since the amount of marketplace interference will be reduced, but will also have less statistical power. Simply comparing the point estimates obtained independently from the two meta-treatment arms is not sufficient to rigorously measure interference bias. In order to do so, we proceed to jointly analyze both the Bernoulli randomized and cluster randomized

---

[10] Due to confidentiality concerns on behalf Airbnb, we are unable to disclose the exact magnitude of the fee changes in this experiment, nor are we able to disclose the cutoff date.

[11] To avoid disclosing raw numbers, all raw booking, nights booked, and gross guest spend values are multiplied by a constant.

meta-treatment arms. Finally, we investigate the extent to which our results are contingent on how supply- or demand-constrained a given Airbnb market is.

5.2.1. Bernoulli & Cluster Randomized Results We analyze both the Bernoulli randomized and cluster randomized meta-treatment arms separately by estimating the following model,

$$Y_i = \alpha + \beta T_i + \sum_l \gamma_l \mathbb{1}(B_i = l) + \delta X_i + \epsilon_i \tag{3}$$

on listing-level data, where Y_i is the outcome of interest, T_i is the treatment assignment for listing i, B_i is a variable indicating which stratum listing i's cluster of belongs to, X_i is a vector consisting of listing i's pre-treatment bookings, nights booked, booking value, and gross guest spend, and ϵ_i is an error term.[12] For all analyses, we cluster standard errors at the Airbnb listing cluster-level.

Table 1 shows the TATE estimate for bookings per listing in both the Bernoulli randomized and cluster randomized meta-treatment arms. In the Bernoulli randomized meta-treatment arm, the TATE is -0.207 bookings per listing, whereas in the cluster randomized meta-treatment arm, the TATE is -0.142 bookings per listing. Both of these TATE estimates are statistically significant at the 95% confidence level. Figure 7 shows the estimated TATE in both meta-treatment arms, along with the corresponding 95% confidence intervals.

Although the TATE estimates obtained from the two meta-experiment arms are different, it is not clear when analyzing the two meta-experiment arms separately whether or not there is a statistically significant difference between the two estimates. By extension, it is still unclear whether or not the Bernoulli TATE estimate suffers from interference bias and/or if cluster randomization helps to mitigate this bias. In order to rigorously test for a difference, it is necessary to jointly analyze both meta-treatment arms simultaneously.

5.2.2. Joint Analysis In order to determine with statistical rigor whether the two meta-treatment arms yield different treatment effect results, we estimate the model,

$$Y_i = \alpha + (\beta + \nu M_i)T_i + \xi M_i + \sum_l \gamma_l \mathbb{1}(B_i = l) + \delta X_i + \epsilon_i, \tag{4}$$

where Y_i is the outcome of interest, M_i is a binary variable set to 1 when listing i is in the Bernoulli meta-treatment arm and 0 when i is in the cluster-randomized meta-treatment arm, T_i is a binary variable set to 1 when listing i is exposed to the treatment, B_i is a variable indicating the stratum

---

[12] Data from the cluster randomized meta-treatment arm can also be analyzed by first aggregating the data at the cluster level and then estimating a weighted version of Equation 3. We present this analysis in Appendix C. This analysis results in estimates that are nearly identical to those obtained by analyzing the experiment with listing-level data.

of clusters to which listing i belongs, X_i is a vector consisting of listing i's pre-treatment variables, and ϵ_i is the error term. Once again, we cluster standard errors at the Airbnb listing cluster-level.

In the above model, β measures the "true" effect of the treatment,[13] and ν measures the difference between the effect of the treatment in the Bernoulli arm and the effect of the treatment in the cluster randomized arm. In other words, ν should measure the extent to which cluster randomization reduces interference bias. ξ measures any baseline difference between listings in the Bernoulli-randomized arm of the meta-experiment and listings in the cluster-randomized arm of the meta-experiment. Since clusters were assigned to meta-treatment arms using the random assignment procedure described in Section 4, we expect ξ to be zero. However, it is possible that imbalances between listings in the two meta-treatment arms persist even after random assignment.

Table 2 shows the results from estimating Equation 4 for the fee meta-experiment using listing level data.[14] Figure 8 displays our point estimate for each parameter in Equation 4, along with 95% confidence intervals. We estimate that the "true" TATE is -0.139 bookings per listing, whereas -0.067 bookings per listing of the TATE measured in the Bernoulli randomized meta-treatment arm is due to interference bias. In other words, we estimate that 32.60% ($\pm$12.93%) of the TATE estimate achieved through a Bernoulli randomized experiment is due to interference bias, and is eliminated by instead running a cluster randomized experiment.

5.2.3.   The Moderating Effect of Supply and Demand Constrainedness Given that interference bias arises in part due to substitution and complementarity between Airbnb listings, one might expect that the extent to which interference causes bias in the Bernoulli randomized TATE estimate depends on the conditions in a given Airbnb market. For instance, interference bias may be *more severe* in markets that are demand constrained, and *less severe* in markets that are supply constrained.

In order to test this hypothesis, we re-estimate Equation 4 for subsets of Airbnb listings that are located in particularly supply constrained or demand constrained markets. Airbnb calculates a supply elasticity index and demand elasticity index for all markets that are above some threshold size using a Cobb-Douglas matching model a la Fradkin (2015). Of the markets for which these indices are calculated, we keep data for listings that are in markets larger than the median market (computed at the listing level). We then define a listing as being in a *supply constrained* market if

---

[13] Even when using cluster randomization, TATE estimates may be biased, since clusters do an imperfect job of capturing listings that complement and substitute for one another. Furthermore, interference may extend beyond a given listing's immediate substitutes or complements.

[14] Joint meta-experiment data can also be analyzed using a weighted combination of individual listing-level data from the Bernoulli randomized meta-treatment arm and aggregated cluster-level data from the cluster randomized meta-treatment arm. This analysis results in estimates that are nearly identical to those obtained using listing-level data from both meta-treatment arms. We present this analysis in Appendix D.

its market's supply elasticity index is above the 75th quantile of supply elasticity indices (computed at the listing level), and define a listing as being in a *demand constrained* market if its market's demand elasticity index is above the 75th quantile of demand elasticity indices (computed at the listing level).

Column 1 of Table 3 shows our results for supply constrained listings, and Column 2 of Table 3 shows our results for demand constrained listings. Neither joint analysis is able to detect interference bias with statistical significance. However, if we take our non-statistically significant point estimates as given, our results do suggest that interference bias accounts for 15.09% of the Bernoulli TATE estimate in demand constrained markets, whereas interference bias actually *reduces* the magnitude of the Bernoulli TATE estimate by 27.41% in supply constrained markets. We interpret this as weak evidence that interference bias is more likely to lead to inflated TATE estimates in demand constrained markets than supply constrained markets, although further research should be conducted to better understand this relationship.

6. Algorithmic Pricing Experiment

The fee meta-experiment results prove that interference bias can have large effects on the accuracy of TATE estimates for online marketplace experiments, and that cluster randomization can help to minimize interference bias. However, the treatment intervention in the fee meta-experiment, a uniform fee change to a well-defined set of Airbnb listings, is only one of the many types of intervention that may be of interest to practitioners. In fact, many of the interventions that online marketplace designers may want to test are behavioral nudges, which require ITT analysis. In the Airbnb context, one such intervention is a change to Airbnb's algorithmic pricing suggestions for hosts.

Previous academic research suggests that smaller firms (e.g., Airbnb hosts) often behave "behaviorally" and act sub-optimally when making managerial decisions (Kremer et al. 2019), including pricing decisions (DellaVigna and Gentzkow 2017). Airbnb uses a machine learning model (Ifrach et al. 2016, Ye et al. 2018) to suggest prices to hosts and help them achieve their business goals. Field experiments have shown that managerial training can lead to increased performance for small firms (Bloom et al. 2013, Bruhn et al. 2018), suggesting that Airbnb's algorithmic pricing suggestions can change the behavior of hosts and affect their business outcomes.

When Airbnb tests a new iteration of its pricing algorithm, not all hosts are directly affected. Some hosts do not use Airbnb's pricing suggestions at all, and hosts who access Airbnb's pricing tips through "Price Tips" often have low compliance rates due to the manual effort required to follow Airbnb's suggestions. Even those hosts who opt into "Smart Pricing" may not fully comply with Airbnb's new suggestions, since Airbnb's suggestions are often constrained by business logic

imposed by the host. Although Airbnb's pricing algorithm experiments do not directly affect all hosts, ITT analysis is required for two reasons. First, the set of hosts who *do* accept Airbnb's suggestions (and the extent to which they comply with those suggestions) is endogenous. Second, the firm is interested in the overall effect of the intervention, including the rate at which hosts accept a given set of suggestions.

In order to test the efficacy with which cluster randomization mitigates interference bias for interventions that require ITT analysis, we present the results from a second meta-experiment in which the treatment intervention is a change to Airbnb's pricing suggestions.

6.1. Description

The algorithmic pricing meta-experiment ran from September 28, 2018 to October 31, 2018 on a population of 4,557,234 listings. Of those listings, 1,139,240 were assigned to the Bernoulli-randomized meta-treatment arm, and the remaining 3,417,994 were assigned to the cluster-randomized meta-treatment arm. Within the Bernoulli-randomized meta-treatment arm, 569,821 listings were assigned to the treatment and 569,419 listings were assigned to the control. Within the Cluster-randomized meta-treatment arm, 11,631 clusters were assigned to the treatment, and 11,631 clusters were assigned to the control, resulting in 1,709,018 listings assigned to the treatment, and 1,708,976 listings assigned to the control. In total, across both meta-treatment arms, 2,278,839 listings were assigned to the treatment, and 2,278,395 listings were assigned to the control. Importantly, the sample size for the algorithmic pricing meta-experiment is approximately equal to the sample size for the fee meta-experiment.

Figure E.2 shows the empirical CDFs for pre-treatment bookings, nights booked, and booking value across all four meta-treatment / treatment groups.[15] For each of these pre-treatment outcomes, the empirical CDFs are visually quite similar. Kolmogorov-Smirnoff tests comparing the distributions for each pre-treatment outcome in the meta-treatment and meta-control fail to reject the null of equal distributions for bookings ($p = 0.387$), nights booked ($p = 0.222$), and booking value ($p = 0.180$). Kolmogorov-Smirnoff tests comparing the distributions for each pre-treatment outcome in the treatment and control groups of the meta-control arm fail to reject the null of equal distributions for bookings ($p = 1.000$), nights booked ($p = 0.888$), and booking value ($p = 0.752$). Kolmogorov-Smirnoff tests comparing the distributions for each pre-treatment outcome in the treatment and control groups of the meta-treatment arm reject the null of equal distributions for bookings ($p = 0.021$) and nights booked ($p = 0.021$), but fail to reject the null of equal distributions for booking value ($p = 0.847$). We believe that some KS tests fail for the same reasons as were outlined when describing the fee meta-experiment.

---

[15] To avoid disclosing raw numbers, x-axis values are multiplied by a constant.

For listings in the treatment group, the suggested prices surfaced through both "Price Tips" and "Smart Pricing" were generated by a new version of Airbnb's pricing algorithm. Relative to the status quo algorithm, the treatment algorithm generally increased prices. For instance, on unconstrained smart pricing nights (e.g., calendar nights in which hosts had opted into smart pricing and the suggested price was not subject to a minimum or maximum price threshold), prices increased by 4% on average.[16]

6.2. Results

In this section, we present results from the algorithmic pricing experiment. As was true for the fee meta-experiment, we report effects of the treatment on bookings per listing, but found qualitatively similar results for nights booked per listing and gross guest spend per listing, which can be found in Appendix B.[17] Since, on average, the treatment pricing algorithm increased prices, we expect the TATE on bookings per listing to be negative. We first present the results separately analyzing the Bernoulli randomized arm of the meta-experiment and the cluster randomized arm of the meta-experiment. We then proceed to jointly analyze both meta-treatment arms, in order to test for the existence of interference bias in the Bernoulli randomized experiment's TATE estimate.

6.2.1. Bernoulli & Cluster Randomized Results

We analyze both the Bernoulli randomized and cluster randomized meta-treatment arms separately by estimating equation 3 on listing-level data.[18] As was the case with the fee meta-experiment, standard errors are clustered at the Airbnb listing-cluster level.

Table 4 shows the TATE estimate for bookings per listing in both the Bernoulli randomized and cluster randomized meta-treatment arms. In the Bernoulli randomized meta-treatment arm, the TATE is -0.106 bookings per listing, and this result is statistically significant at the 95% confidence level. In the cluster randomized meta-treatment arm, our point estimate of the TATE is -0.051 bookings per listing, however, this result is not statistically significant at the 95% confidence level. Figure 9 shows the estimated TATE in both meta-treatment arms, along with the corresponding 95% confidence intervals. In order to rigorously test whether or not cluster randomization led to a reduction in interference bias, we proceed to jointly analyze both meta-treatment arms.

---

[16] Unconstrained smart pricing nights represent only a fraction of the total calendar nights on Airbnb. As a result, the average price increase across *all* calendar nights is less than 4%.

[17] To avoid disclosing raw numbers, all raw booking, nights booked, and gross guest spend values are multiplied by a constant.

[18] As was the case with the fee meta-experiment, we present aggregate-level analysis of the cluster randomized meta-treatment arm in Appendix C. The results from this analysis are nearly identical.

6.2.2. Joint Analysis In order to determine whether or not the two meta-treatment arms yield TATE estimates between which there is a statistically significant difference, we once again estimate equation 4.[19] As was the case with the fee meta-experiment, standard errors are clustered at the Airbnb listing-cluster level.

Table 5 shows our results, and Figure 10 displays our point estimate for each parameter in Equation 4, along with 95% confidence intervals. Point estimates imply that the "true" TATE is -0.050 bookings per listing, whereas -0.059 bookings per listing of the TATE measured in the Bernoulli randomized meta-treatment arm is due to interference bias. This would suggest that 54.16% ($\pm$65.05%) of the TATE achieved through a Bernoulli randomized experiment is due to interference that is eliminated by instead running a clustered experiment. However, none of these point estimates are statistically significant. A post-hoc power analysis of the algorithmic pricing experiment reveals that the meta-experiment is underpowered to detect reasonable effect sizes relative to the treatment effect estimated obtained in the Bernoulli randomized arm of the meta-experiment. Table 6 shows the calculated minimum detectable effect (MDE) for β, ν, and ξ. Each of these MDEs is also overlaid in red on Figure 10. Comparing the Bernoulli TATE estimate with the meta-experiment MDEs implies that interference bias would need to have approximately the same magnitude as our Bernoulli TATE estimate to be detectable.

This result highlights the difficulty of identifying (and reducing) interference bias using cluster randomization and meta-experimentation when the treatment intervention of interest is a behavioral nudge or some other type of intervention that will require ITT analysis. Although both the fee meta-experiment and the pricing meta-experiment have experimental samples of almost exactly the same size, one is able to detect statistically significant interference bias, while the other is not. Given that standard errors decrease with square root of the sample size, we estimate that a sample approximately 3.45 times as large would be required to detect interference bias in the algorithmic pricing meta-experiment.

7. Discussion

In this paper, we have taken the first empirical steps to understand the extent to which statistical inference can bias total average treatment effect estimates in online marketplace experiments. We have achieved this by presenting the results from two different pricing-related meta-experiments conducted on Airbnb, an online marketplace for accommodations and experiences. In each meta-experiment, some clusters of listings were assigned their experimental treatment using Bernoulli

---

[19] For the algorithmic pricing meta-experiment, X_i also includes listing i's smart pricing opt-in status at the outset of the experiment.

randomization, whereas others were assigned to their experimental treatment using cluster randomization. The motivation for our focus on pricing-related interventions was twofold; understanding customer price elasticities is crucial for both platform intermediaries and sellers, and there are strong reasons to suspect that pricing-relating experiments violate the stable unit treatment value assumption.

Analysis of our first meta-experiment, in which guest platform fees for treatment listings were increased relative to their peers in the control, provided clear evidence for interference bias in online marketplace experiments, and the potential for cluster randomization to mitigate this bias. While analysis of the Bernoulli meta-treatment arm alone suggested that the TATE was a decrease of 0.207 bookings per listing, a joint analysis of both meta-treatment arms revealed that 32.60% of the reported TATE in the Bernoulli meta-treatment arm was due to interference bias that cluster randomization was able to eliminate. This figure represents a lower bound on the magnitude of interference bias, as our clusters likely do an imperfect job of capturing Airbnb listings that interfere with one another. While many recent papers measure the impact of innovative market mechanisms through field experiments (Horton and Johari 2015, Filippas et al. 2019), very few of them explicitly account for interference bias. Based on our results, we argue that taking steps to reduce interference bias is crucial if researchers hope to estimate total average treatment effects accurately.

Analysis of the fee meta-experiment also reveals that the amount of bias in TATE estimates may depend on the extent to which a market is supply- or demand-constrained. Although our evidence is weak and comes from non-statistically significant point estimates, TATE estimates appear to be overstated due to interference bias to a greater extent in Airbnb markets that are demand constrained than in Airbnb markets that are supply constrained. Better understanding the relationships between supply elasticity, demand elasticity, and interference bias is a promising direction for future work. We also analyze a second meta-experiment, in which the treatment changes Airbnb hosts' algorithmically suggested prices, to understand how well our method can be applied to a behavioral nudge that requires ITT analysis. While point estimates suggest that the TATE estimate from the Bernoulli randomized meta-treatment arm is severely inflated due to interference bias, our results are not statistically significant, despite both meta-experiments having approximately equal sample sizes. This result highlights the difficulty of detecting interference bias for behavioral nudges and other treatment interventions that require ITT analysis. Unfortunately, these types of interventions are very common in online marketplaces. Future work might focus on developing even more sensitive tests for interference bias that will work more effectively when conducting such experiments.

In addition to cluster randomization, there are a number of analysis techniques that have been developed in the network experimentation literature, such as exposure modeling (Aronow and Samii 2012), regression adjustment (Chin 2018), and exact tests for interference (Athey et al. 2018) that, if adopted to a commerce-based setting, could help to more accurately identify and reduce interference bias in online marketplace experiments. Furthermore, there are a number of open questions regarding how to best identify the sellers most likely to interfere with one another in an online marketplace setting. The clustering method described in this paper is by no means the only (or best) way to cluster sellers before performing cluster randomization. Higher quality clusters could lead to even greater interference bias reductions. Finally, while the approach described in this work can reduce bias due to interference between sellers, it does not consider the reduction of bias due to interference between *buyers*. Given that, in general, online marketplaces have much less information about buyers, many of the approaches discussed thus far are unlikely to be effective. Developing methods that reduce interference bias on the buyer side of online marketplaces is a promising direction for future research.

Accounting for interference bias increases the logistical complexity of online marketplace experimentation. However, for many interventions, e.g., those that are designed to help platform intermediaries measure price elasticities, determining only the direction of a treatment effect is not sufficient; an accurate point estimate is required. Using pricing related meta-experiments on Airbnb as a test case, we have shown that interference bias can account for at least 32.60% of a TATE estimate in an online marketplace experiment. In light of this result, we believe that accounting for interference bias can be worth the additional effort for many marketplace designers and researchers.
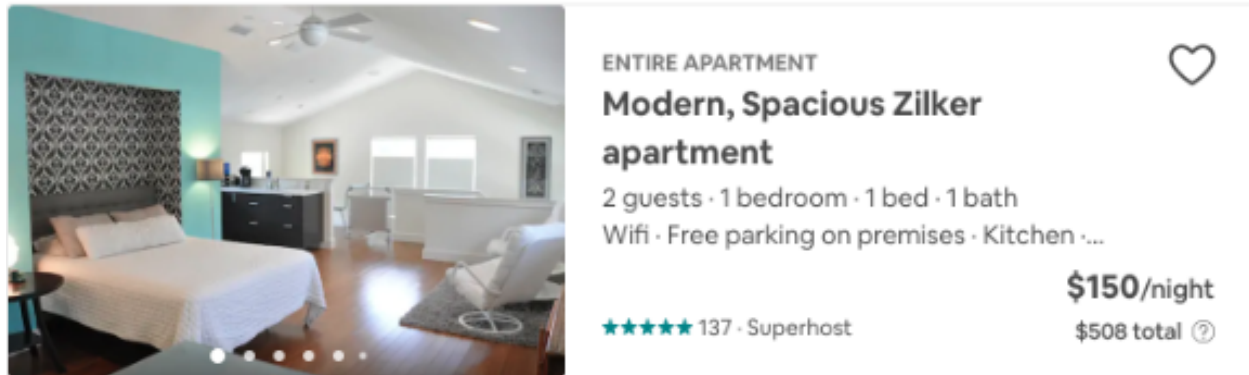
8. Figures

Figure 1   A typical search result on Airbnb. For this search result, the guest platform fee is included in the total price of $508.
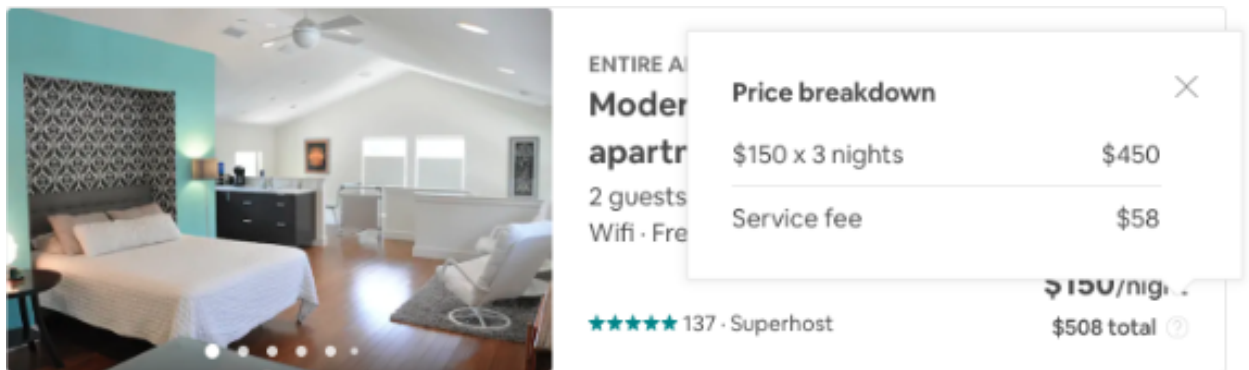

Figure 2   The price breakdown tooltip for a typical search result on Airbnb. In this tooltip, the guest platform fee (listed here as a service fee of $58) is broken out from the nightly price.

Figure 3 **The section of the Airbnb product detail page that provides a full pricing breakdown for would-be guests. In this pricing breakdown, the guest platform fee (listed here as a service fee) is $58.**
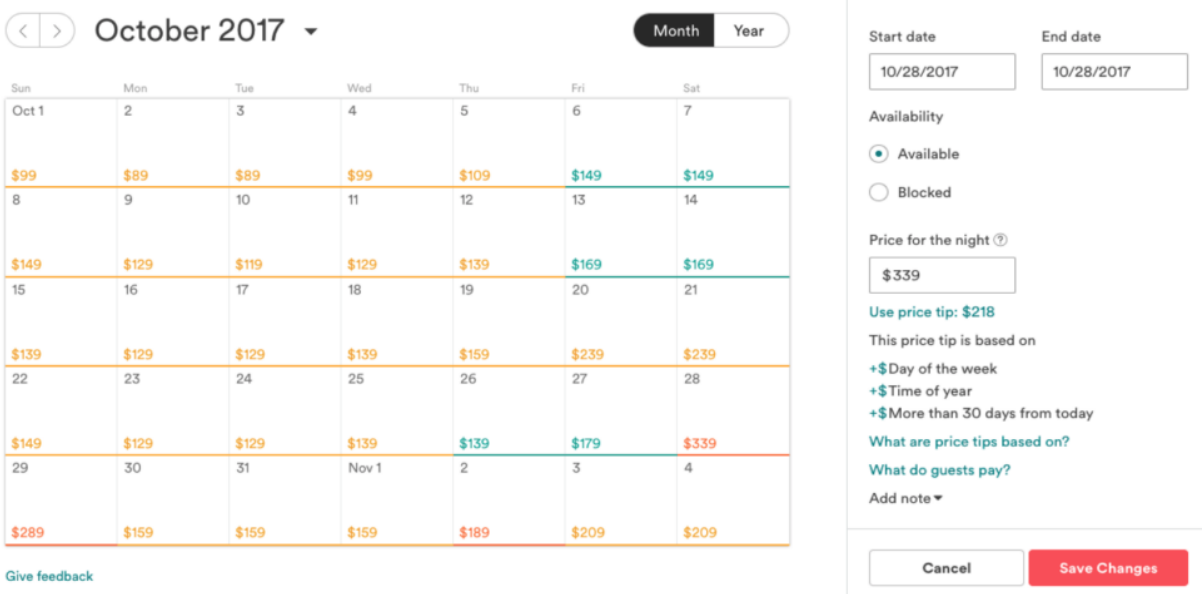
Figure 4   **A screenshot of the "Price tips" UI. "Price tips" color codes the nights on a host's calendar based on the pricing model's estimated probability that a given night will be booked. If a host selects a given calendar night, the host is shown the model's suggested price. Airbnb also presents explanations for why it is recommending a particular price (e.g., "Time of year," "More than 30 days from today"). In order for a given host to fully adopt Airbnb's recommended prices with the "Price tips" product, the host is required to visit Airbnb every day, review Airbnb's recommendations, and accept them. Image from Ye et al. (2018).**
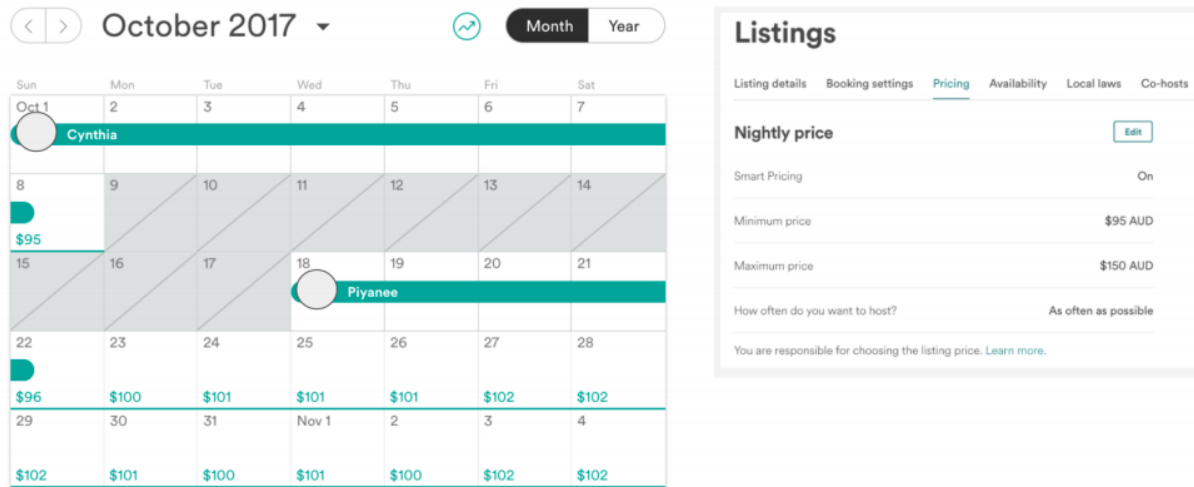
Figure 5 A screenshot of the "Smart pricing" UI. When setting up "Smart Pricing," hosts provide a minimum and maximum price. After "Smart Pricing" is turned on, hosts automatically adopt Airbnb's recommended price if it is between the host's minimum and maximum price. If Airbnb's recommendation is higher than the host's upper threshold, the price is set to the upper threshold. If Airbnb's recommendation is lower than the host's lower threshold, the price is set to the lower threshold. A screenshot of the "Smart Pricing" UI is shown in Figure 5. Image from Ye et al. (2018).

Figure 6 Example clusters generated using the hierarchical clustering scheme described in this paper. Image from Srinivasan (2018).
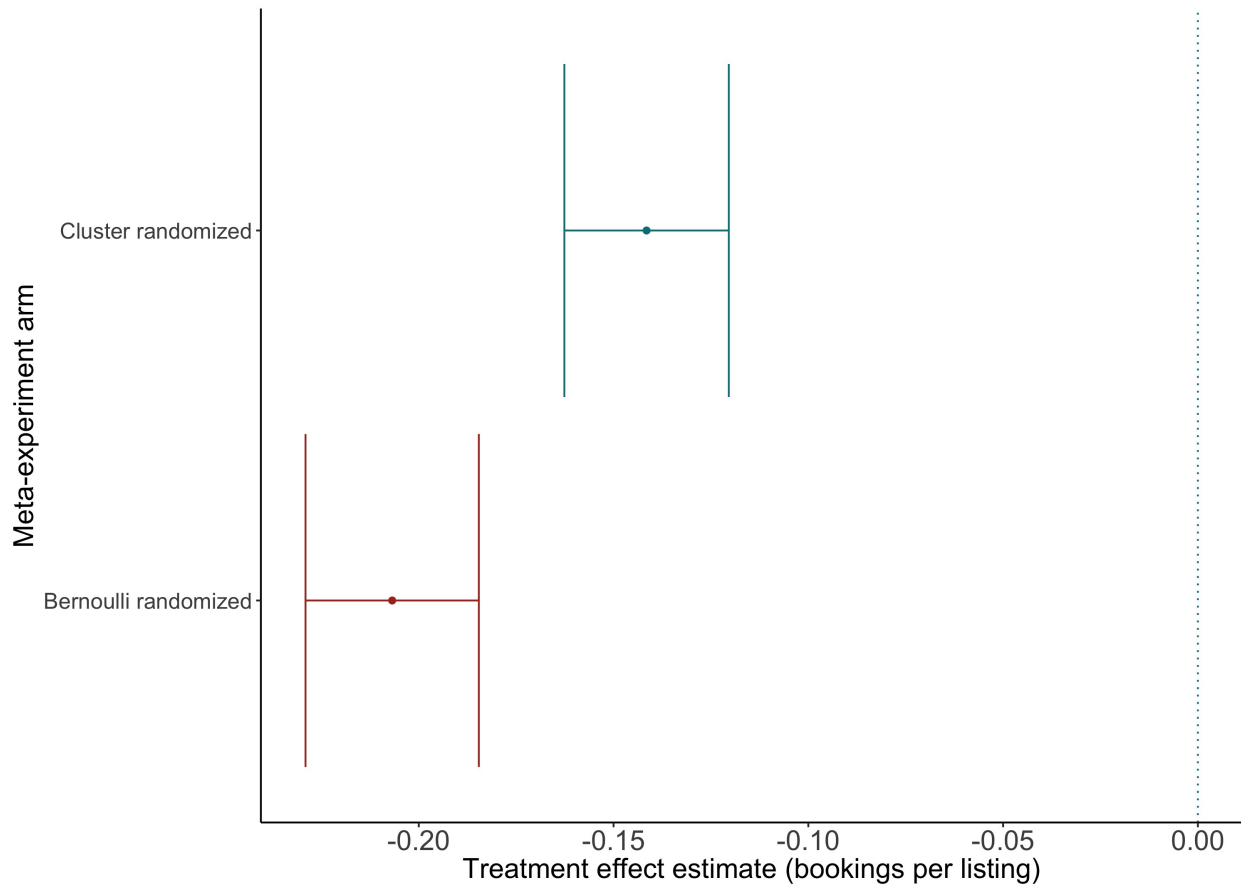
Figure 7 **Total average treatment effect estimates for the fee experiment, estimated separately in the Bernoulli randomized meta-treatment arm and the cluster randomized meta treatment arm. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0 bookings per listing.**

Figure 8 **Coefficient estimates for the joint analysis of the fee meta-experiment. Error bars represent 95% confidence intervals. The dotted blue line correponds to a treatment effect of 0 bookings per listing. The red shaded area corresponds to values that are below the MDE (80% power, 95% confidence).**

Figure 9 **Total average treatment effect estimates for the algorithmic pricing experiment, estimated separately in the Bernoulli randomized meta-treatment arm and the cluster randomized meta treatment arm. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0 bookings per listing.**
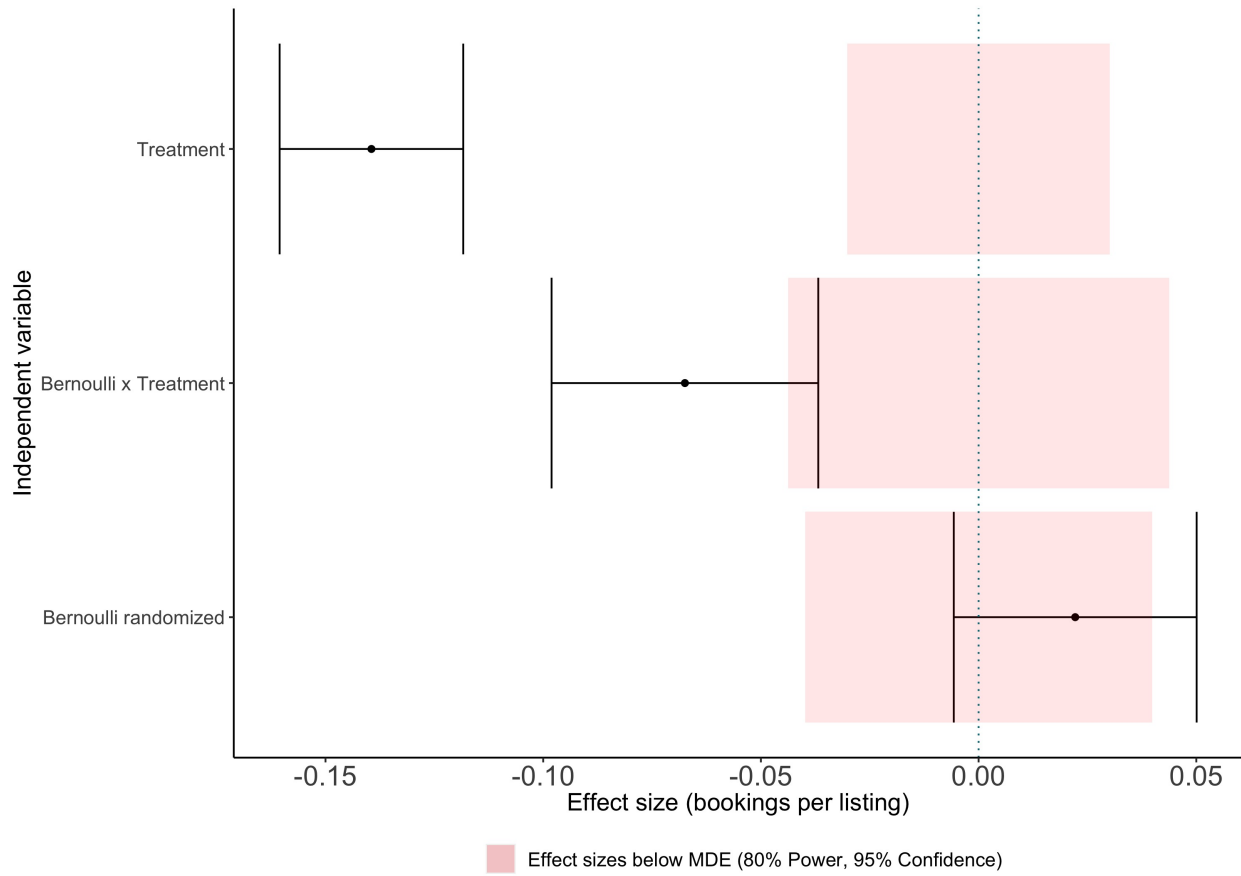
Figure 10   Coefficient estimates for the joint analysis of the algorithmic pricing meta-experiment. Error bars represent 95% confidence intervals. The dotted blue line correponds to a treatment effect of 0 bookings per listing. The red shaded area corresponds to values that are below the MDE (80% power, 95% confidence).
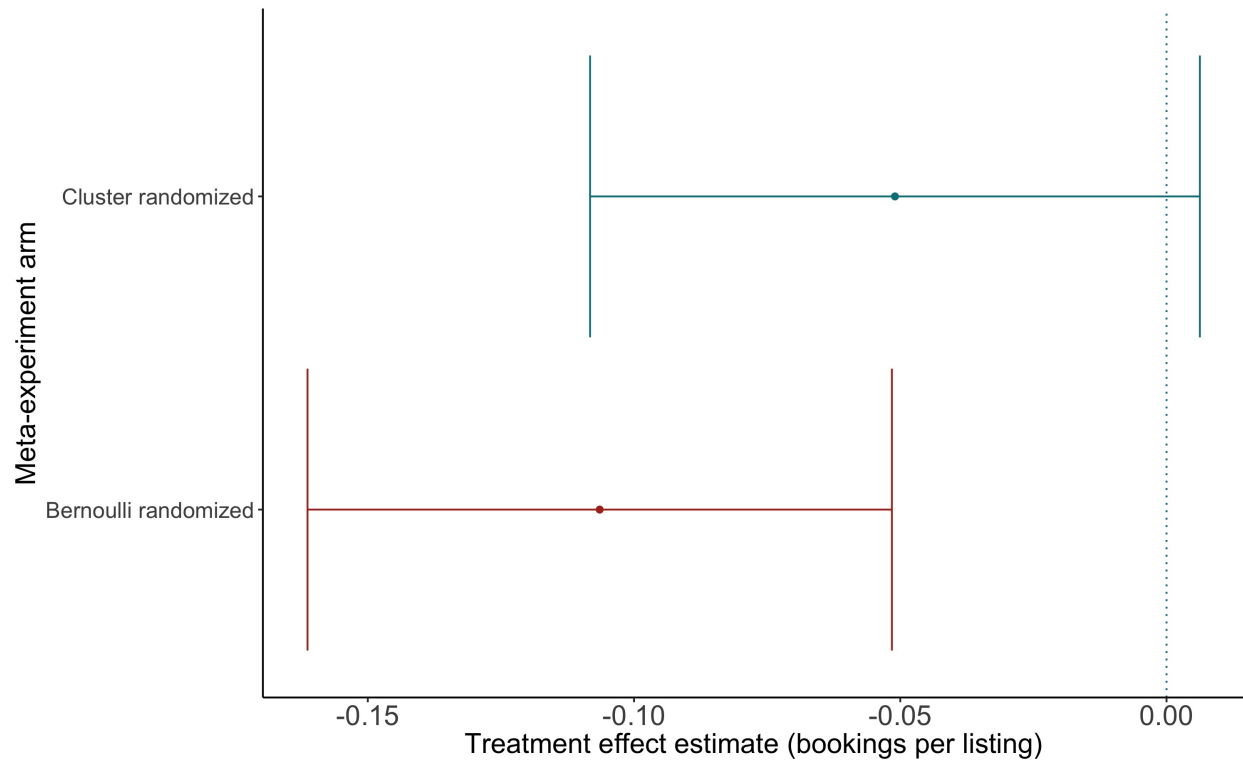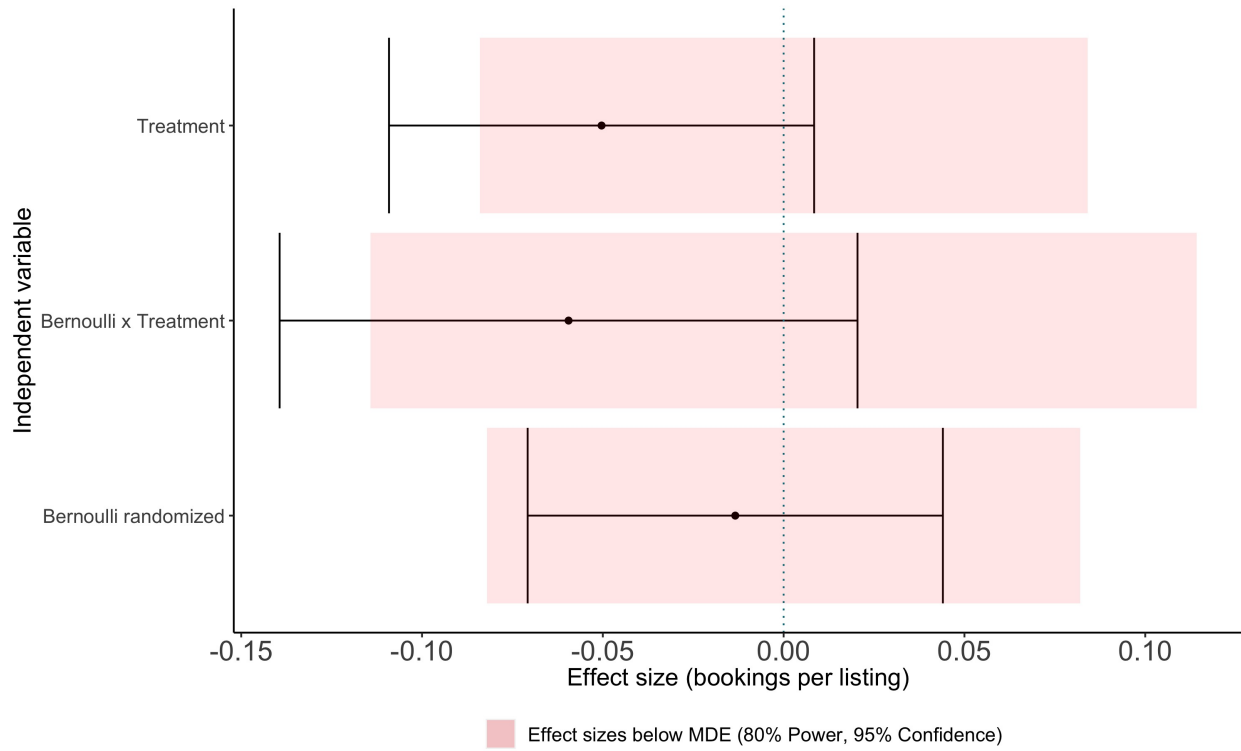
9. Tables

Table 1   Independent results of the fee meta-experiment

	Dependent variable:	
	Bookings	
	Bernoulli randomized	Cluster randomized
	(1)	(2)
Treatment	−0.207***	−0.142***
	(0.011)	(0.011)
Pre-treatment bookings	0.173***	0.174***
	(0.001)	(0.001)
Pre-treatment nights booked	−0.003***	−0.003***
	(0.000)	(0.000)
Pre-treatment booking value	0.000	0.000***
	(0.000)	(0.000)
Pre-treatment gross guest spend	−0.000**	−0.000***
	(0.000)	(0.000)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	No	Yes
R^2	0.407	0.405
Adjusted R^2	0.406	0.405

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2   Results of the fees Meta-experiment

	Dependent variable:
	Bookings
Treatment	−0.139***
	(0.011)
Bernoulli Randomized	0.022
	(0.014)
Bernoulli × Treatment	−0.067***
	(0.016)
Pre-treatment bookings	0.174***
	(0.001)
Pre-treatment nights booked	−0.003***
	(0.000)
Pre-treatment booking value	0.000***
	(0.000)
Pre-treatment gross guest spend	−0.000***
	(0.000)
Stratum F.E.	Yes
Robust s.e.	Yes
Clustered s.e.	Yes
R^2	0.405
Adjusted R^2	0.405
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3   **Results of the fee meta-experiment for supply- and demand-constrained listings**

	Dependent variable:	
	Bookings	
	Supply constrained	Demand constrained
	(1)	(2)
Treatment	−0.241***	−0.200***
	(0.051)	(0.038)
Bernoulli Randomized	−0.029	−0.031
	(0.060)	(0.059)
Bernoulli × Treatment	0.052	−0.036
	(0.059)	(0.052)
Pre-treatment bookings	0.170***	0.174***
	(0.002)	(0.002)
Pre-treatment nights booked	−0.003***	−0.003***
	(0.000)	(0.000)
Pre-treatment booking value	0.000	0.000***
	(0.000)	(0.000)
Pre-treatment gross guest spend	−0.000**	−0.000***
	(0.000)	(0.000)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	Yes
R^2	0.421	0.389
Adjusted R^2	0.420	0.388
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 4   **Independent results of the algorithmic pricing meta-experiment**

	Dependent variable:	
	Bookings	
	Bernoulli randomized	Cluster randomized
	(1)	(2)
Treatment	−0.106***	−0.051*
	(0.028)	(0.029)
Pre-treatment bookings	0.822***	0.828***
	(0.004)	(0.002)
Pre-treatment nights booked	−0.018***	−0.017***
	(0.001)	(0.000)
Pre-treatment booking value	0.000*	0.000***
	(0.000)	(0.000)
Pre-treatment gross guest spend	−0.000**	−0.000***
	(0.000)	(0.000)
Smart pricing pre-treatment	0.587***	0.586***
	(0.033)	(0.020)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	No	Yes
R^2	0.580	0.578
Adjusted R^2	0.578	0.578

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5   Results of the algorithmic pricing meta-experiment

	Dependent variable:
	Bookings
Treatment	-0.050^{*}
	(0.030)
Bernoulli Randomized	-0.013
	(0.037)
Bernoulli $\times$ Treatment	-0.059
	(0.041)
Pre-treatment bookings	0.827^{***}
	(0.002)
Pre-treatment nights booked	-0.017^{***}
	(0.000)
Pre-treatment booking value	0.000^{***}
	(0.000)
Pre-treatment gross guest spend	-0.000^{***}
	(0.000)
Smart pricing pre-treatment	0.577^{***}
	(0.017)
Stratum F.E.	Yes
Robust s.e.	Yes
Clustered s.e.	Yes
R^2	0.577
Adjusted R^2	0.577
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6   Minimum detectable effects for algorithmic pricing meta-experiment analysis (power = 80%, confidence level = 95%)

Regressor	Bookings
Treatment	0.084
Bernoulli x Treatment	0.114
Bernoulli randomized	0.082

References

Airbnb (2015a) Airbnb unveils expansive suite of personalized tools to empower hosts. URL `https://www.airbnb.com/press/news/airbnb-unveils-expansive-suite-of-personalized-tools-to-empower-hosts`.

Airbnb (2015b) Using data to help set your price. URL `https://blog.atairbnb.com/using-data-to-help-set-your-price/`.

Airbnb (2019) Airbnb press room: Fast facts. URL `https://press.airbnb.com/fast-facts/`.

Aronow PM, Samii C (2012) Estimating average causal effects under general interference. *Summer Meeting of the Society for Political Methodology, University of North Carolina, Chapel Hill, July*, 19–21 (Citeseer).

Athey S, Eckles D, Imbens GW (2018) Exact p-values for network interference. *Journal of the American Statistical Association* 113(521):230–240.

Blake T, Coey D (2014) Why marketplace experimentation is harder than it seems: The role of test-control interference. *Proceedings of the fifteenth ACM conference on Economics and computation*, 567–582 (ACM).

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008(10):P10008.

Bloom N, Eifert B, Mahajan A, McKenzie D, Roberts J (2013) Does management matter? evidence from india. *The Quarterly Journal of Economics* 128(1):1–51.

Bruhn M, Karlan D, Schoar A (2018) The impact of consulting services on small and medium enterprises: Evidence from a randomized trial in mexico. *Journal of Political Economy* 126(2):635–687.

Chin A (2018) Central limit theorems via stein's method for randomized experiments under interference. *arXiv preprint arXiv:1804.03105* .

Cox DR (1958) Planning of experiments. .

DellaVigna S, Gentzkow M (2017) Uniform pricing in us retail chains. Technical report, National Bureau of Economic Research.

Dubé JP, Misra S (2017) Scalable price targeting. Technical report, National Bureau of Economic Research.

Eckles D, Karrer B, Ugander J (2017) Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference* 5(1).

Filippas A, Jagabathula S, Sundararajan A (2019) Managing market mechanism transitions: A randomized trial of decentralized pricing versus platform control. *Proceedings of the 2019 ACM Conference on Economics and Computation* (ACM).

Fradkin A (2015) Search frictions and the design of online marketplaces. *Work. Pap., Mass. Inst. Technol* .

Grbovic M, Cheng H (2018) Real-time personalization using embeddings for search ranking at airbnb. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 311–320 (ACM).

Holtz DM (2018) *Limiting bias from test-control interference in online marketplace experiments.* Master's thesis, Massachusetts Institute of Technology.

Horton JJ, Johari R (2015) At what quality and what price?: Eliciting buyer preferences as a market design problem. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 507–507 (ACM).

Ifrach B, Holtz DM, Yee YH, Zhang L (2016) Demand prediction for time-expiring inventory. US Patent App. 14/952,576.

Kang JH, Park CH, Kim SB (2016) Recursive partitioning clustering tree algorithm. *Pattern Analysis and Applications* 19(2):355–367.

Kremer M, Rao G, Schilbach F (2019) Behavioral development economics. *Handbook of Behavioral Economics* 2.

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Moore RT (2012) Multivariate continuous blocking to improve political science experiments. *Political Analysis* 20(4):460–479.

Oestreicher-Singer G, Sundararajan A (2012a) Recommendation networks and the long tail of electronic commerce. *MIS Quarterly* 36(1):65–83.

Oestreicher-Singer G, Sundararajan A (2012b) The visible hand? demand effects of recommendation networks in electronic markets. *Management science* 58(11):1963–1981.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5):688.

Saveski M, Pouget-Abadie J, Saint-Jacques G, Duan W, Ghosh S, Xu Y, Airoldi EM (2017) Detecting network effects: Randomizing over randomized experiments. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1027–1035 (ACM).

Srinivasan S (2018) Learning market dynamics for optimal pricing. URL `https://medium.com/airbnb-engineering/learning-market-dynamics-for-optimal-pricing-97cffbcc53e3`.

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

Ye P, Qian J, Chen J, Wu Ch, Zhou Y, De Mars S, Yang F, Zhang L (2018) Customized regression model for airbnb dynamic pricing. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 932–940 (ACM).

Appendix A: Method for cluster size selection

In this section, we detail the methodology that was used in deciding to conduct the fee meta-experiment with clusters with a listing threshold of 1,000, as opposed to 250. Although this analysis was originally conducted using clusters and data from February 2019, we present analyses using clusters generated on January 5, 2020, listing views occurring between January 5, 2020 and January 12, 2020, and bookings occurring between January 5, 2020 and January 26, 2020. However, the results we report and the corresponding conclusions are qualitatively similar to those obtained using 2019 data.

In choosing a cluster size threshold, the fundamental trade-off is between statistical power and capturing Airbnb demand. While smaller clusters will yield more statistical power (since there will be more of them), they will also do a poorer job of capturing demand, since a given user search session is more likely to contain listings from many different clusters. On the other hand, larger clusters will provide less statistical power, but will do a better job of capturing demand. Power analysis suggested that a week-long experiment shifting fees in the same manner as our fee experiment would have an MDE of 0.9% for interference bias if clusters with a threshold size of 250 were used, whereas the same experiment would have an MDE of 1.05% for interference bias if clusters with a threshold size of 1,000 were used. In order to determine whether this reduction in "ideal" MDE is worthwhile, we needed to measure differences in the extent to which the two sets of clusters capture demand.

We began our investigation by defining two different measures related to demand capture:

$$\% \text{ in single cluster } = \frac{1}{n_{users}} \sum_{\text{all users}} \mathbb{1}\left(n_{clusters} = 1\right) \tag{5}$$

$$\text{Demand capture } = \frac{1}{n_{users}} \sum_{\text{all users}} \left(1 - \frac{n_{clusters}}{n_{listings}}\right) \tag{6}$$

The first measures the percentage of users for whom all listings viewed fall within a single cluster. The second is a less strict measure that captures the extent to which all viewed listings are contained within a small number of clusters. Importantly, both measures will be close or equal to 1 if users never compare listings across different clusters and $n_{listings}$ is sufficiently large, and will be equal to 0 if the number of listings a user compares is equal to the number of clusters needed to cover them. Figure E.3 shows both of these measures for listing views occurring between January 5, 2020 and January 12, 2020, for cluster size thresholds ranging from 100 to entire markets. As expected, as the size of clusters increases, both of these demand capture metrics move closer to 1. Importantly, even when markets are defined as "clusters," they are unable to capture 100% of demand, regardless of which measure we use.

Based on statistical power considerations, we decided that a cluster size threshold of 1,000 was the maximum threshold worth considering. Once this decision was made, we began to more directly compare the status quo threshold of 250 listings (which had been used for the algorithmic pricing meta-experiment) to the maximum threshold of 1,000 listings.[20] In doing so, we created an alternative demand capture measure

---

[20] The 250 listing threshold was chosen for the algorithmic pricing meta-experiment in an ad-hoc manner.

that asked the following question: given a set of clusters, what percentage of listing viewers have at least $x\%$ of their listings captured by one cluster? Figure E.4 plots this measure for both the 250 listing threshold clusters and the 1,000 listing threshold clusters, with demand capture thresholds of 67%, 75%, and 90%. As expected, the clusters with the 1,000 listing threshold do a better job of capturing demand than the 250 listing threshold clusters.

In order to make a principled decision, we assumed that the "ideal" MDEs mentioned earlier in this appendix were reduced by poor demand capture according to the relationship below:

$$MDE_{actual} = \frac{MDE_{ideal}}{\text{Demand capture}}. \tag{7}$$

In other words, as a given set of clusters' demand capture moved closer to 1, the MDE would approach the ideal MDE. Given this assumed relationship between actual MDE, ideal MDE, and demand capture, we determined that the 1,000 listing threshold clusters would be preferable to the 250 listing threshold clusters when

$$\frac{\text{Demand capture}_{1,000}}{\text{Demand capture}_{250}} > \frac{MDE_{ideal_{250}}}{MDE_{ideal_{1,000}}} \rightarrow \frac{\text{Demand capture}_{1,000}}{\text{Demand capture}_{250}} > \frac{1.05\%}{0.9\%} \tag{8}$$

Table F.1 shows the ratio of demand capture for clusters with a threshold of 1,000 listings to the demand capture for clusters with a threshold of 250 clusters according to five different demand capture measures: the average share of search listings belonging to a cluster, the average user-level Herfindahl-Hirschman index across clusters, and the percentage of users for which one cluster accounts for at least 67%, 75%, and 90% of listings viewed. Across all five of these demand capture metrics, and across different user subpopulations, the demand capture ratio is consistently above $\frac{1.05\%}{0.9\%} = 1.17$. Based on this calculation, we determined that clusters with a threshold of 1,000 listings were preferable.

Appendix B: Interference bias for nights booked and gross guest spend

In addition to bookings per listing, we also conducted the main analyses in our paper for both nights booked per listing and gross guest spend per listing. In this appendix, we present the results of our analyses for these additional outcomes. Qualitatively, our results for nights booked per listing and gross guest spend per listing are extremely similar to our results for bookings per listing.

B.1.   Fee meta-experiment

Table F.2 shows the estimated effect of the fee treatment in both the Bernoulli randomized meta-treatment arm and the cluster randomized meta-treatment arm on both nights booked per listing and gross guest spend per listing. Our TATE estimates for each outcome are also depicted, along with 95% confidence intervals, in Figure E.5. We estimate in the Bernoulli randomized meta-treatment arm that the treatment led to a statistically significant loss of 0.768 nights booked per listing and $79.68 in gross guest spend per listing, whereas we estimate in the cluster randomized meta-treatment arm that the treatment led to a statistically significant loss of 0.579 nights booked per listing and $63.39 in booking value per listing.

In order to test whether or not there is a statistically significant difference between the TATE estimates in the two meta-treatment arms, we conduct a joint analysis of both meta-treatment arms simultaneously. Table

F.3 shows our results. Our results are also depicted in Figure E.6, along with 95% confidence intervals. We find statistically significant evidence of interference bias in the Bernoulli TATE estimate for nights booked per listing at the 95% confidence level, but do not find statistically significant evidence of interference bias in the Bernoulli TATE estimate for gross guest spend per listing. Our point estimates suggest that interference accounts for 24.79% of the Bernoulli TATE estimate for nights booked per listing (stat sig.) and 21.04% of the Bernoulli TATE estimate for gross guest spend per listing (not stat. sig).

B.2.   Algorithmic pricing meta-experiment

Table F.4 shows the estimated effect of the algorithmic pricing treatment in both the Bernoulli randomized meta-treatment arm and the cluster randomized meta-treatment arm on both nights booked per listing and gross guest spend per listing. Our TATE estimates for each outcome are depicted, along with 95% confidence intervals, in Figure E.7, We estimate in the Bernoulli randomized meta-treatment arm that the treatment let do a statistically significant loss of 0.288 nights booked per listing and $37.38 in gross guest spend per listing, whereas we do not detect a statistically significant treatment effect for either outcome in the cluster randomized meta-treatment arm.

In order to test whether or not there is a statistically significant difference between the TATE estimates in the two meta-treatment arms, we conduct a joint analysis of both meta-treatment arms simultaneously. Table F.5 shows our results. Our results are also depicted in Figure E.8, along with 95% confidence intervals. We do not find statistically significant evidence for interference bias for either outcome. While not statistically significant, our point estimates suggest that interference accounts for 36.86% of the Bernoulli TATE estimate for nights booked per listing and 104.73% of the Bernoulli TATE estimate for gross guest spend per listing.

Appendix C: Cluster-level analysis of cluster-randomized meta-treatment arm

Rather than analyzing data from the cluster randomized meta-treatment arm of our experiments at the individual level with clustered standard errors, it is also possible to aggregate data at the *cluster* level and instead estimated a weighted version of Equation 3, where each cluster is weighted according to the number of experiment-eligible listings in that cluster. In this appendix, we compare the cluster randomized TATE estimates obtained using these two different approaches.

C.1.   Fee meta-experiment

Table F.6 compares the TATE estimates obtained from the cluster randomized meta-treatment arm of the fee meta-experiment when analyzing the data at both the individual listing level and at the cluster level. Our results are also depicted in Figure E.9. We find that both approaches yield almost identical TATE point estimates and standard errors.

C.2.   Algorithmic pricing meta-experiment

Table F.7 compares the TATE estimates obtained from the cluster randomized meta-treatment arm of the algorithmic pricing meta-experiment when analyzing the data at both the individual listing level and at the cluster level. Our results are also depicted in Figure E.10. We find that both approaches yield almost identical TATE point estimates and standard errors.

Appendix D: Results with mixed units of analysis

In addition to performing joint analysis of our meta-experiments with listing-level data, it is possible to analyze the meta-experiments with a mixture of listing-level data and data aggregated at the cluster level. For both meta-experiments, we estimate Equation 4 on listing-level data from the Bernoulli randomized meta-treatment arm and cluster-level data from the cluster randomized meta-treatment arm. Observations are weighted by the number of listings making up that observation (i.e., listings receive a weight of 1, whereas clusters receive a weight equal to the number of experiment eligible listings in that cluster). In this appendix, we compare results obtained using this approach with those obtained analyzing the meta-experiment entirely with listing level data.

D.1.   Fee meta-experiment

Table F.8 compares results obtained by analyzing the fee meta-experiment at the listing level and with mixed units of analysis. Our results are also depicted in Figure E.11. We find that both approaches yield almost identical results.

D.2.   Algorithmic pricing meta-experiment

Table F.9 compares results obtained by analyzing the fee meta-experiment at the listing level and with mixed units of analysis. Our results are also depicted in Figure E.12. We find that both approaches yield almost identical results.

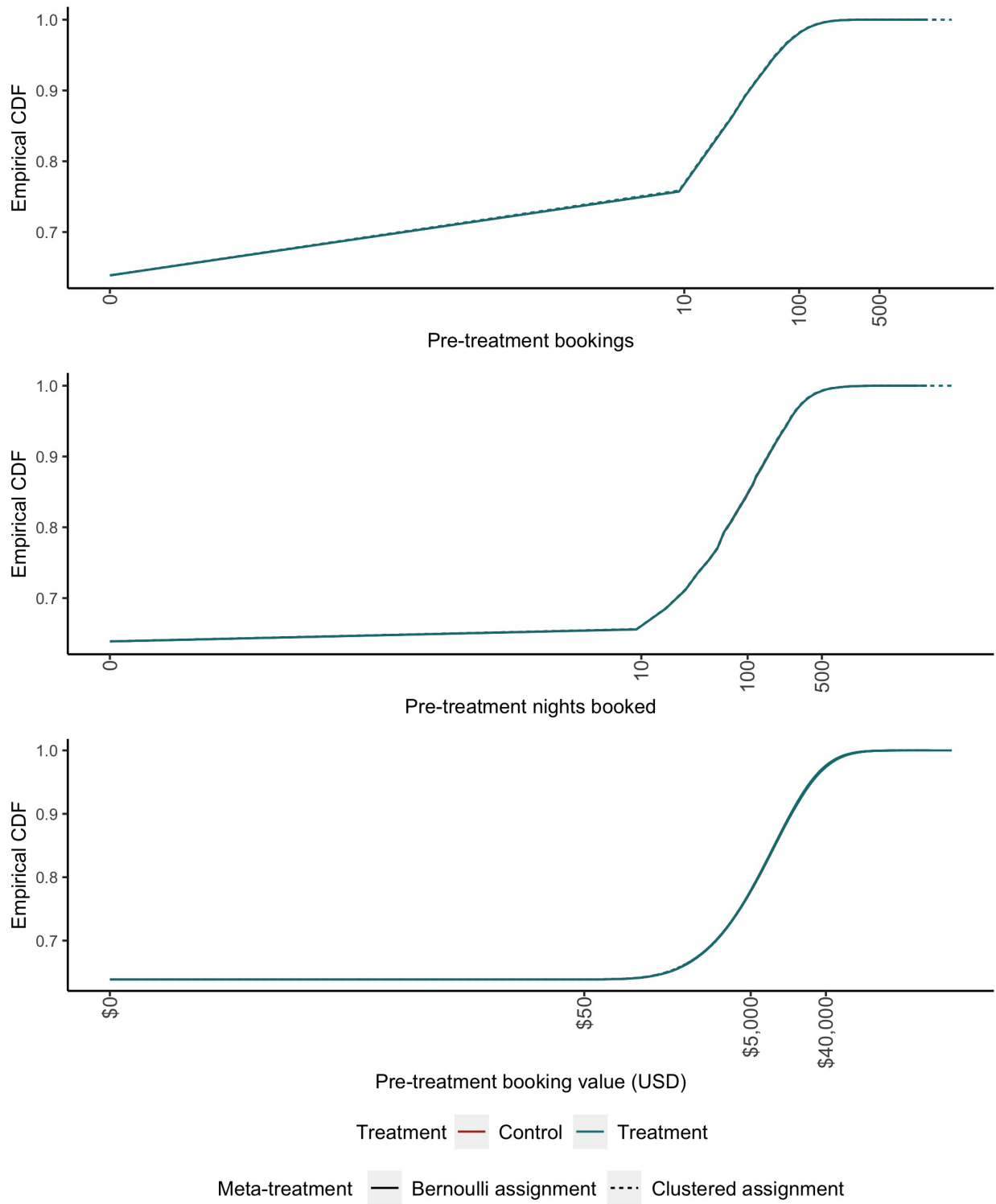## Appendix E: Additional Figures

Figure E.1   The empirical CDFs for pre-treatment bookings, nights booked, and booking value in each of the four treatment/meta-treatment groups for the fee meta-experiment.
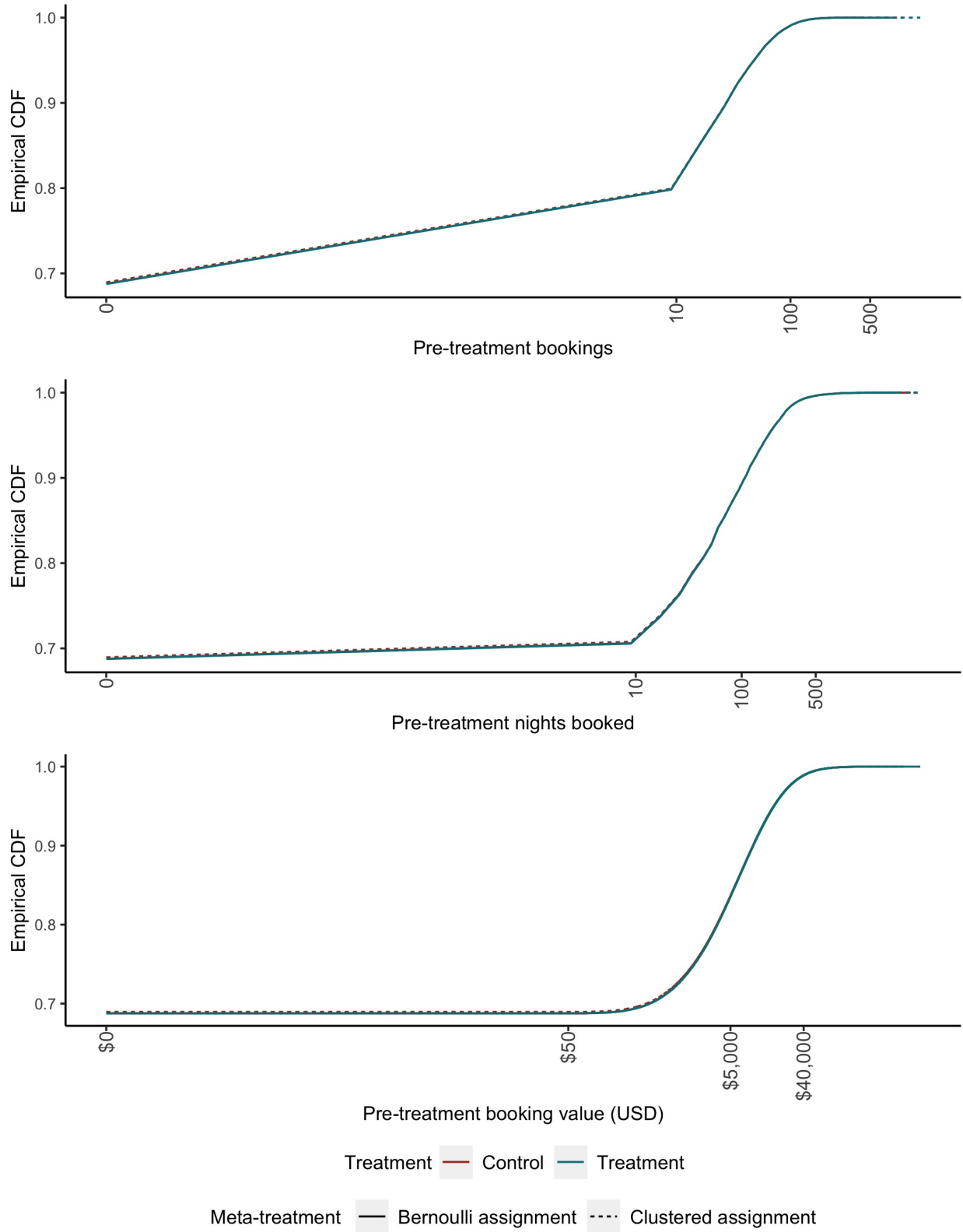
Figure E.2   The empirical CDFs for pre-treatment bookings, nights booked, and booking value in each of the four treatment/meta-treatment groups for the algorithmic pricing meta-experiment.
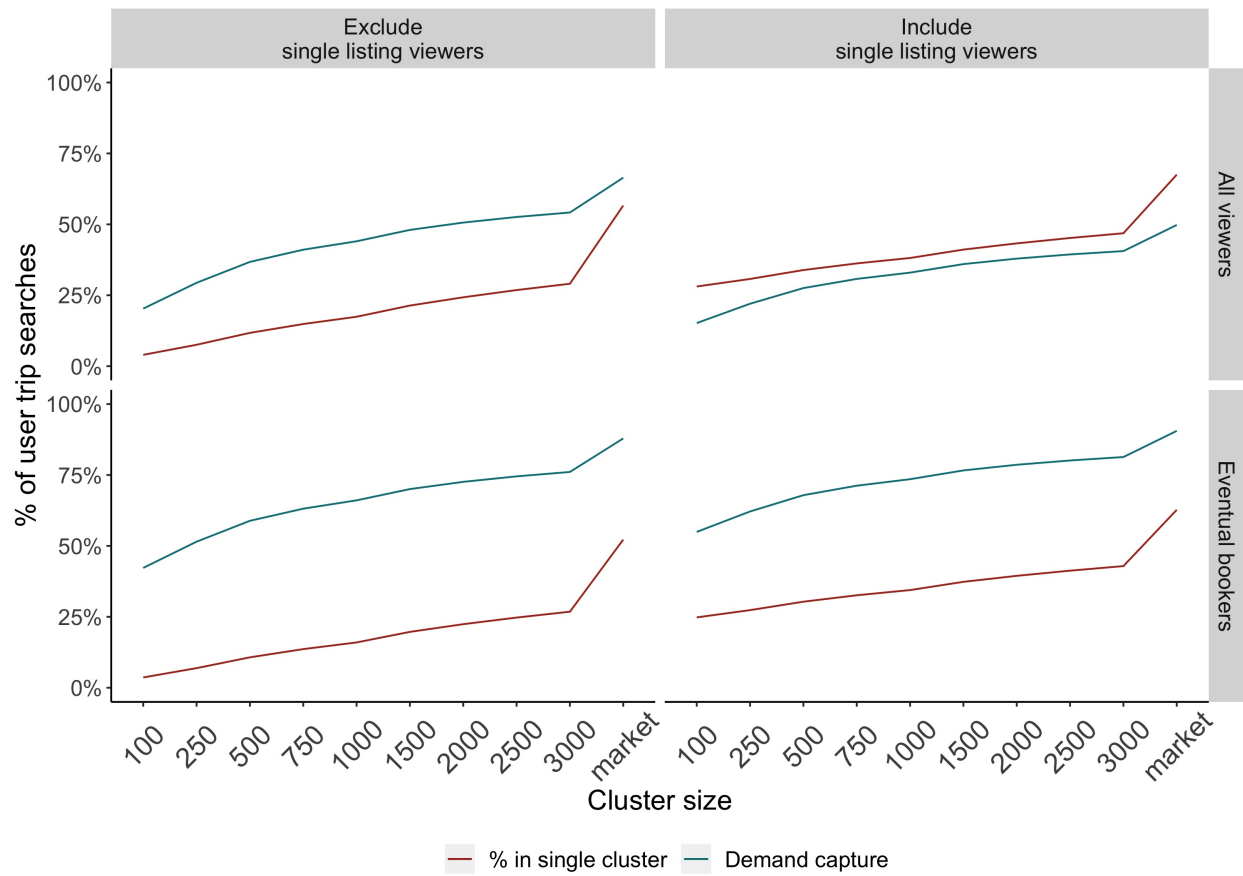
Figure E.3 **The relationship between cluster size and demand capture for two different metrics. The left column excludes users who only view a single Airbnb listing, whereas the right column includes them. The top row includes all listing viewers, whereas the bottom row only includes Airbnb users who go on to eventually book a listing.**
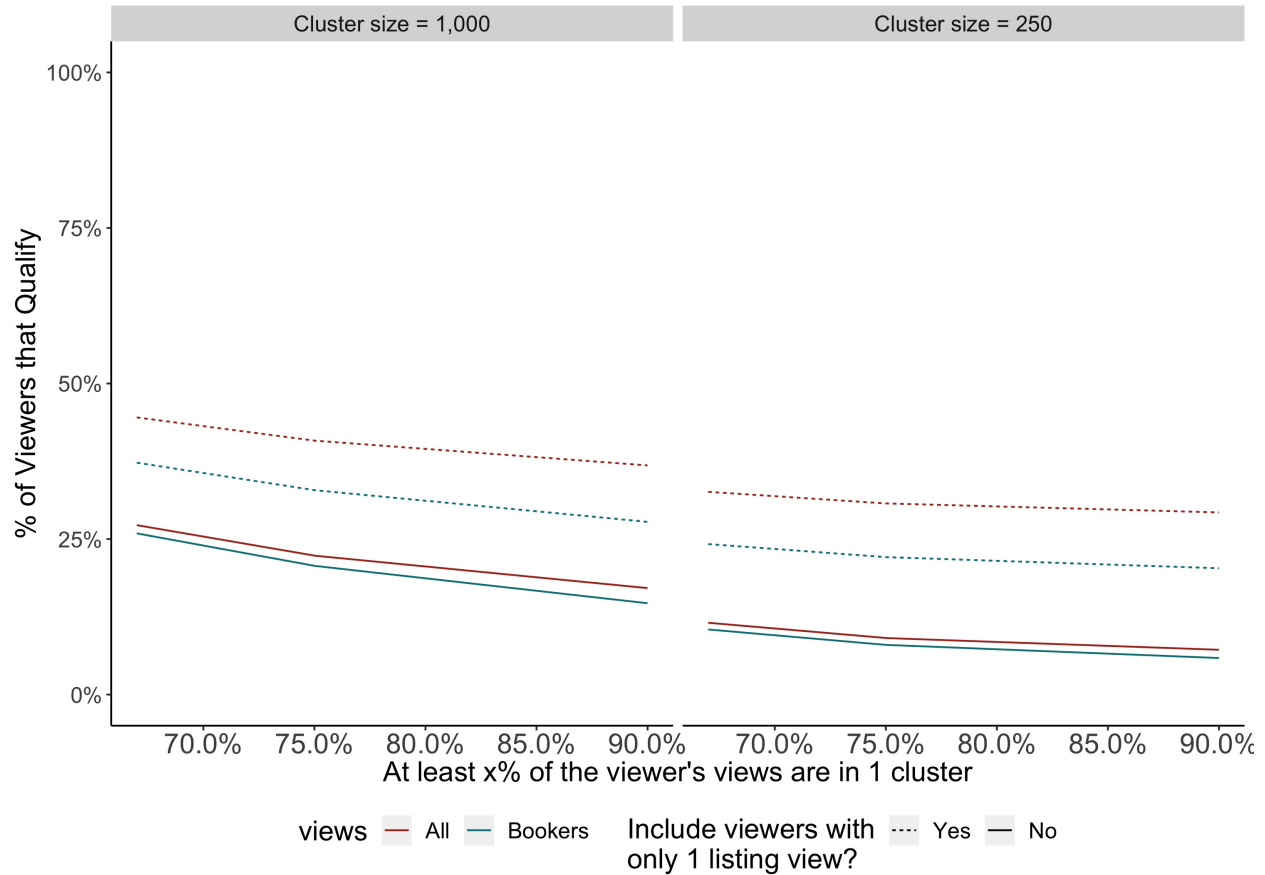
Figure E.4 **A direct comparison of the demand capture of clusters with a 1,000 listing threshold, and clusters with a 250 listing threshold. Curves show the percentage of viewers for whom at least $x\%$ of their views are contained by one cluster. Red curves include all listing viewers, whereas blue curves only include Airbnb users who go on to eventually book a listing. Dashed lines include users who only view a single Airbnb listing, whereas solid lines do not.**
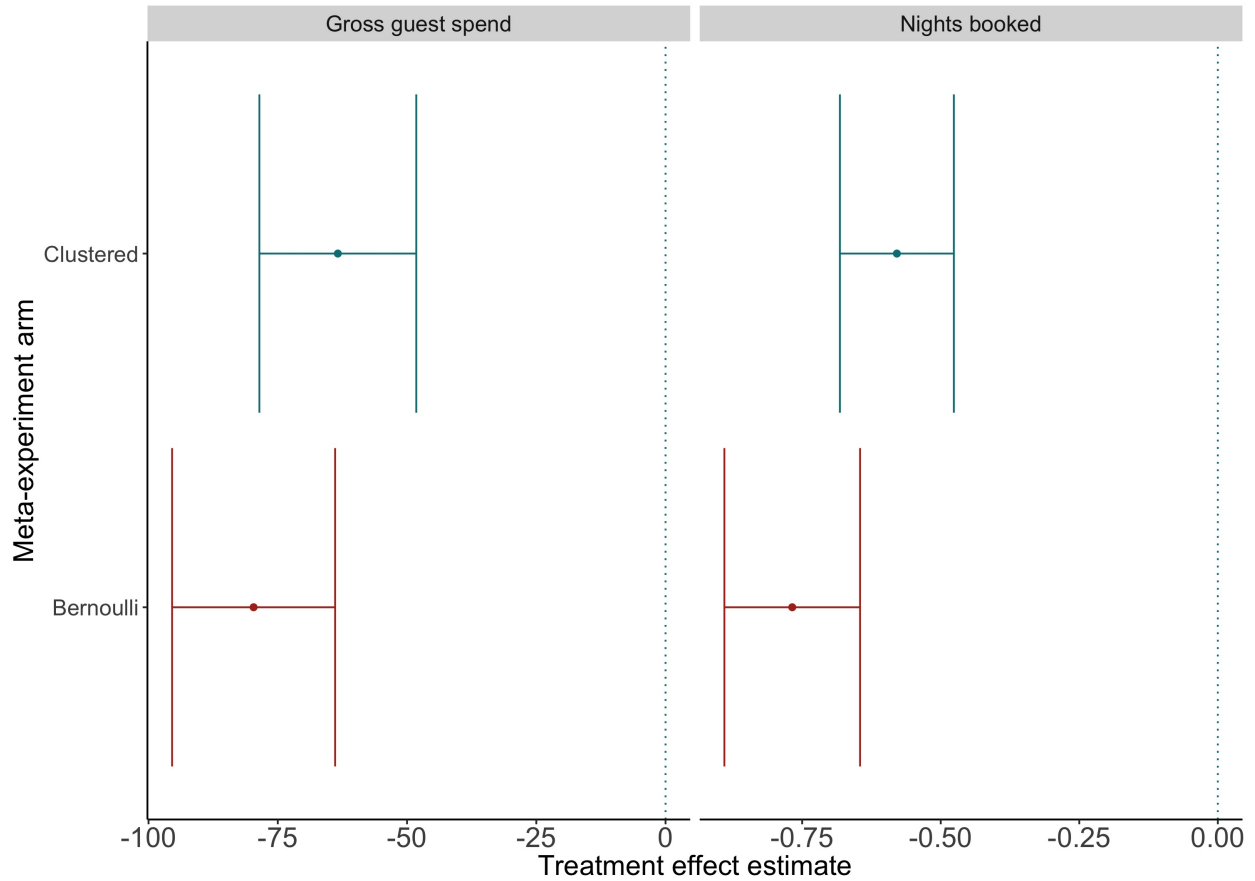
Figure E.5   Total average treatment effect estimates (nights booked per listing and gross guest spend per listing) for the fee experiment, estimated separately in the Bernoulli randomized meta-treatment arm and the cluster randomized meta treatment arm. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0.

Figure E.6   **Coefficient estimates for the joint analysis of the fee meta-experiment (nights booked per listing and gross guest spend per listing). Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0. The red shaded area corresponds to values that are below the MDE (80% power, 95% confidence).**

Figure E.7 **Total average treatment effect estimates (nights booked per listing and gross guest spend per listing) for the algorithmic pricing experiment, estimated separately in the Bernoulli randomized meta-treatment arm and the cluster randomized meta treatment arm. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0.**
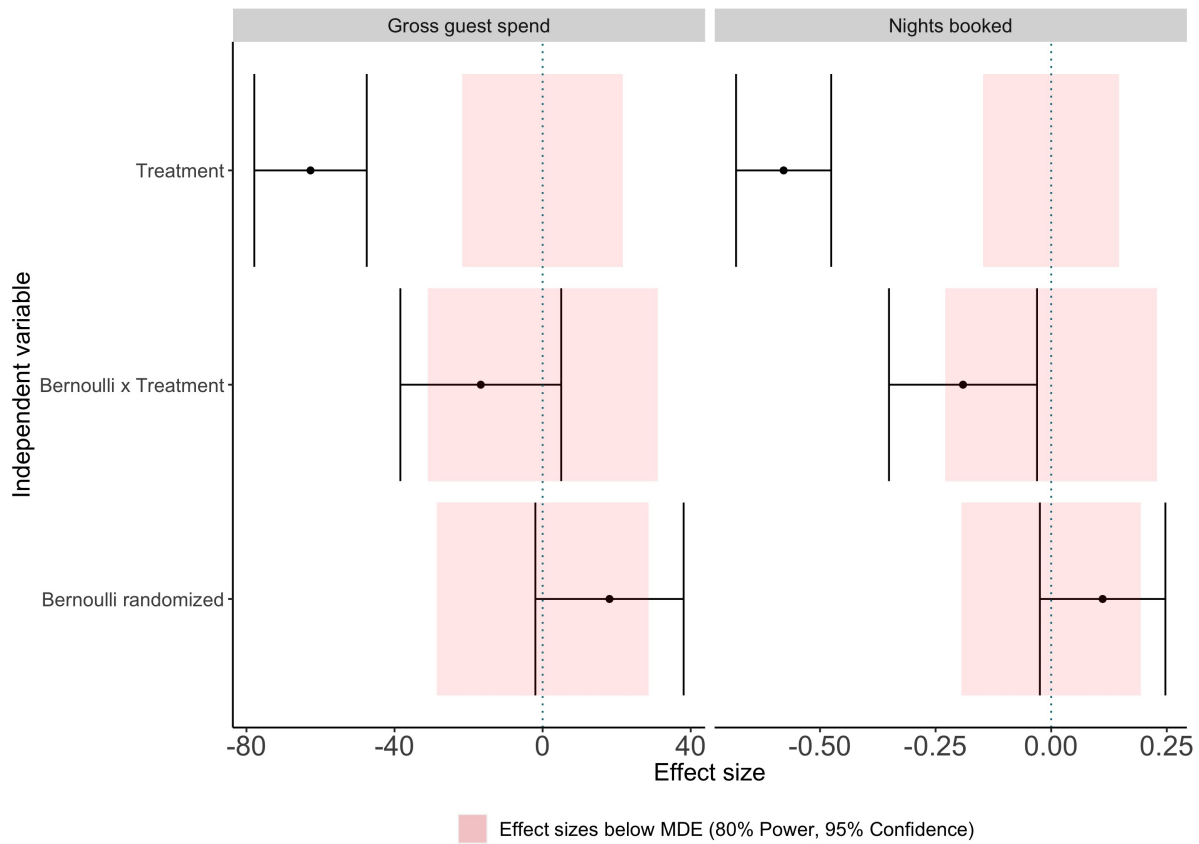
Figure E.8   Coefficient estimates for the joint analysis of the algorithmic pricing meta-experiment (nights booked per listing and gross guest spend per listing). Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0. The red shaded area corresponds to values that are below the MDE (80% power, 95% confidence).
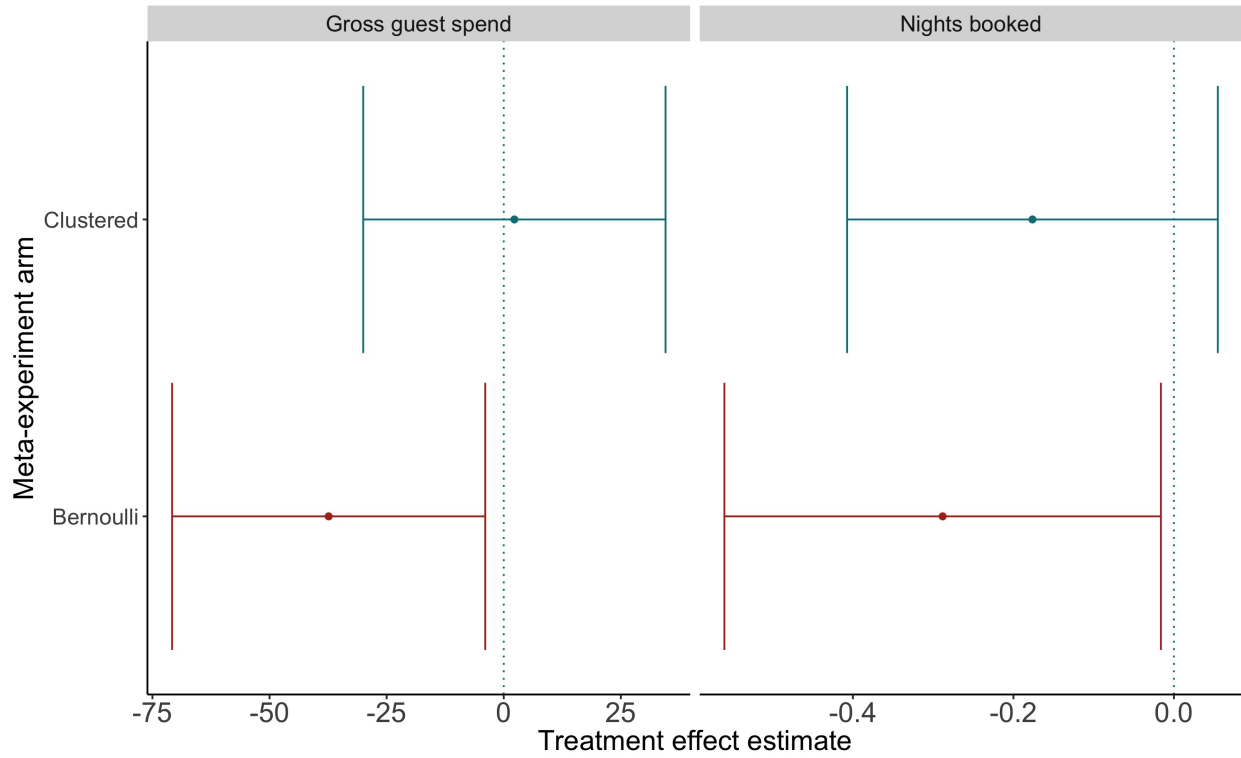
Figure E.9 **Comparison of the TATE estimates from the cluster randomized meta-treatment arm of the fees experiment, obtained analyzing data at either the individual listing level or at the cluster level. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0 bookings per listing.**

Figure E.10   Comparison of the TATE estimates from the cluster randomized meta-treatment arm of the algorithmic pricing experiment, obtained analyzing data at either the individual listing level or at the cluster level. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0 bookings per listing.
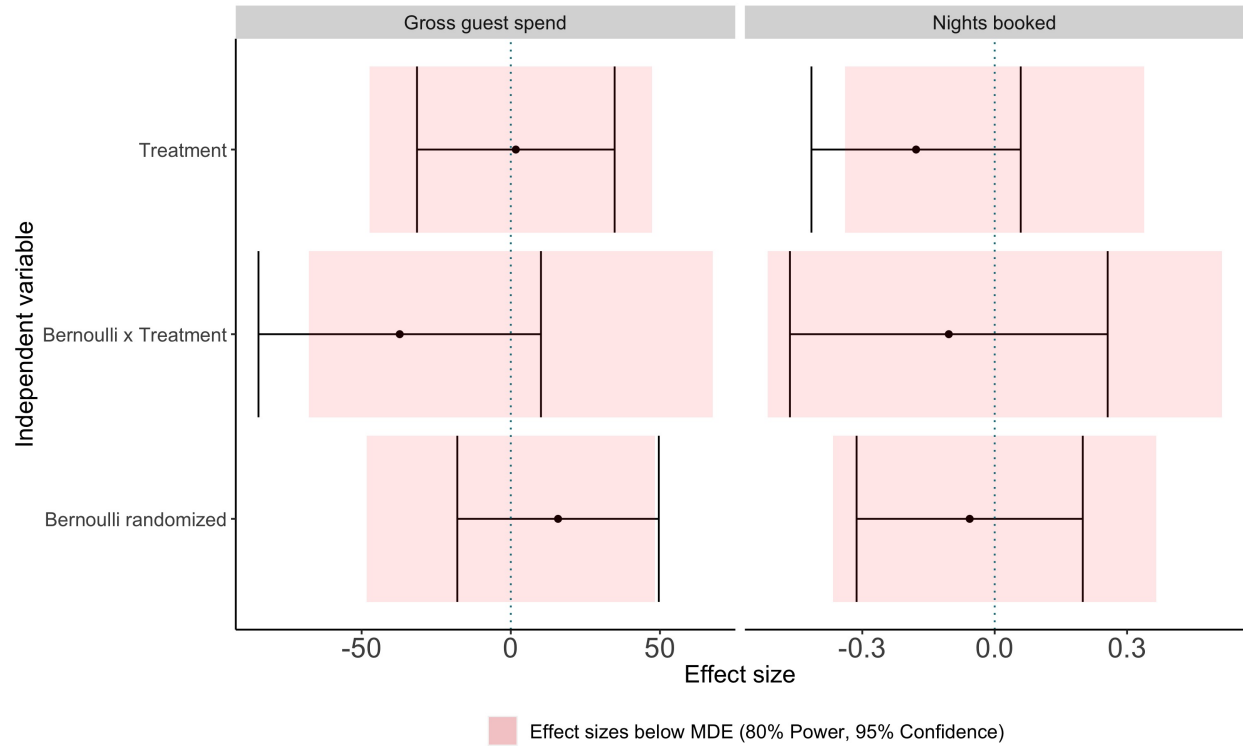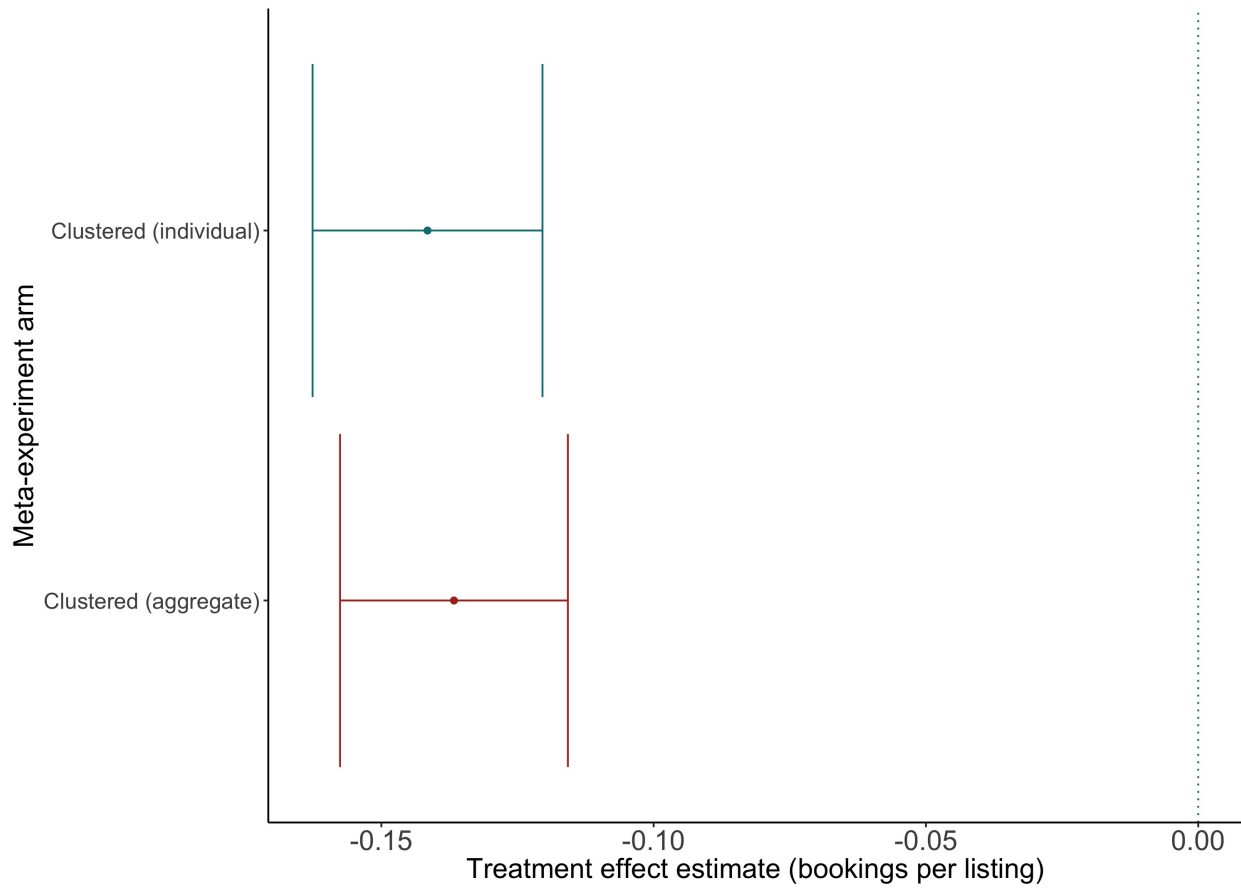
Figure E.11   Comparison of fee experiment meta-analysis estimates obtained analyzing data at the individual level of analysis, and the mixed level of analysis. In the mixed analysis, Data from Bernoulli randomized listings is included at the listing level, whereas data from cluster randomized listings is aggregated at the cluster level. Error bars correspond to 95% confidence intervals. Shaded areas represent effect sizes below the MDE threshold (80% power, 95% confidence).

Figure E.12   **Comparison of algorithmic pricing experiment meta-analysis estimates obtained analyzing data at the individual level of analysis, and the mixed level of analysis. In the mixed analysis, Data from Bernoulli randomized listings is included at the listing level, whereas data from cluster randomized listings is aggregated at the cluster level. Error bars correspond to 95% confidence intervals. Shaded areas represent effect sizes below the MDE threshold (80% power, 95% confidence).**
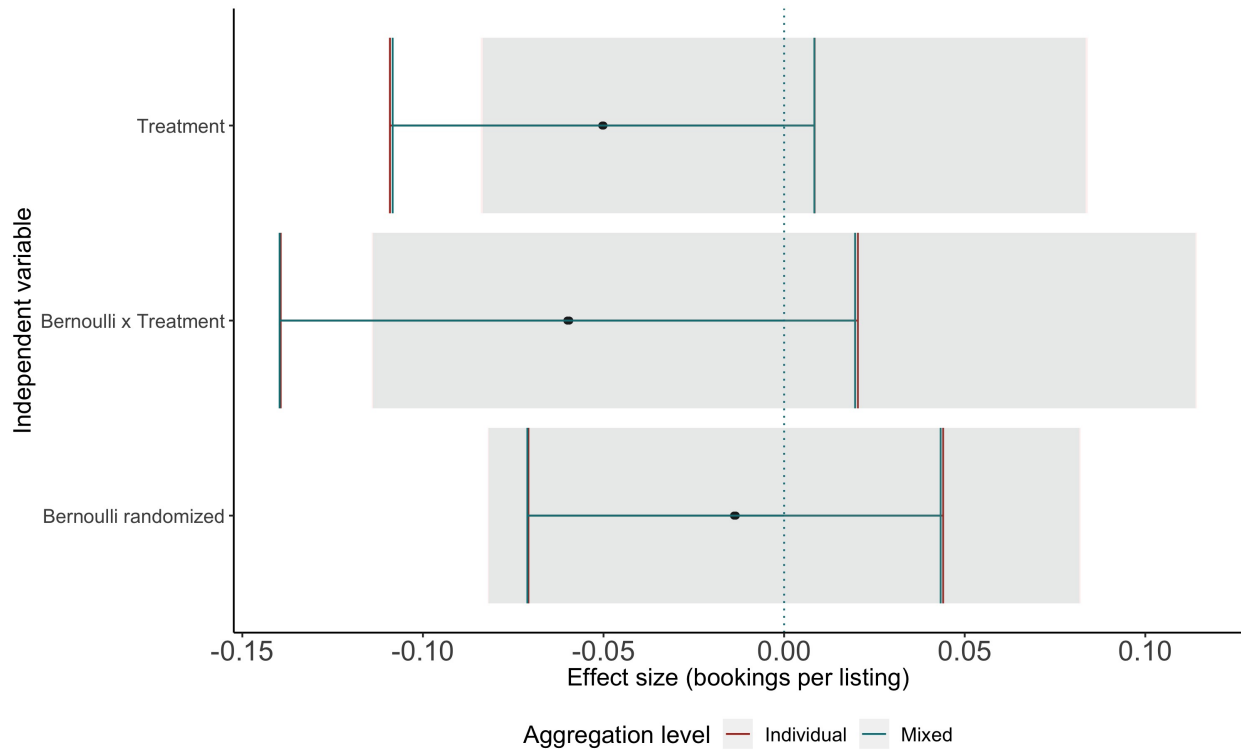
Appendix F: Additional Tables

Table F.1   **The ratio of demand capture for 1,000 listing threshold clusters and 250 listing threshold clusters, using different demand capture metrics and user subpopulations.**

Single views?	Type of viewers	avg. cluster share	avg. HHI	% over 67%	% over 75%	% over 90%
No	All	1.32	1.36	2.36	2.46	2.38
No	Bookers	1.38	1.43	2.48	2.59	2.50
Yes	All	1.16	1.19	1.37	1.33	1.26
Yes	Bookers	1.23	1.27	1.54	1.49	1.37

Table F.2   **Independent results of the fee meta-experiment (nights booked and gross guest spend)**

	Dependent variable:			
	Nights booked		Gross guest spend	
	Bernoulli randomized	Cluster randomized	Bernoulli randomized	Cluster randomized
	(1)	(2)	(3)	(4)
Treatment	−0.768***	−0.579***	−79.677***	−63.388***
	(0.062)	(0.052)	(8.044)	(7.741)
Pre-treatment bookings	0.281***	0.288***	23.220***	22.626***
	(0.005)	(0.003)	(0.750)	(0.372)
Pre-treatment nights booked	0.038***	0.037***	−4.289***	−3.698***
	(0.002)	(0.001)	(0.433)	(0.129)
Pre-treatment booking value	−0.000***	−0.000***	−0.060	−0.148***
	(0.000)	(0.000)	(0.085)	(0.021)
Pre-treatment gross guest spend	0.000***	0.000***	0.153**	0.226***
	(0.000)	(0.000)	(0.070)	(0.017)
Stratum F.E.	Yes	Yes	Yes	Yes
Robust s.e.	Yes	Yes	Yes	Yes
Clustered s.e.	No	Yes	No	Yes
R^2	0.115	0.118	0.166	0.176
Adjusted R^2	0.114	0.118	0.165	0.176

Note: *p<0.1; **p<0.05; ***p<0.01

Table F.3   Results of the fees Meta-experiment (nights booked and gross guest spend)

	Dependent variable:	
	Nights booked	Gross guest spend
	(1)	(2)
Treatment	−0.579***	−62.696***
	(0.052)	(7.749)
Bernoulli Randomized	0.111	18.063*
	(0.069)	(10.217)
Bernoulli × Treatment	−0.191**	−16.704
	(0.082)	(11.085)
Pre-treatment bookings	0.287***	22.787***
	(0.002)	(0.342)
Pre-treatment nights booked	0.038***	−3.849***
	(0.001)	(0.147)
Pre-treatment booking value	−0.000***	−0.123***
	(0.000)	(0.028)
Pre-treatment gross guest spend	0.000***	0.206***
	(0.000)	(0.023)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	Yes
R^2	0.117	0.173
Adjusted R^2	0.117	0.173
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table F.4   Independent results of the algorithmic pricing meta-experiment (nights booked and gross guest spend)

	Dependent variable:			
	Nights booked		Gross guest spend	
	Bernoulli randomized	Cluster randomized	Bernoulli randomized	Cluster randomized
	(1)	(2)	(3)	(4)
Treatment	−0.288**	−0.176	−37.377**	2.268
	(0.139)	(0.118)	(17.052)	(16.466)
Pre-treatment bookings	1.342***	1.370***	87.218***	85.714***
	(0.013)	(0.008)	(1.842)	(1.095)
Pre-treatment nights booked	0.152***	0.147***	−19.907***	−19.948***
	(0.004)	(0.003)	(0.963)	(0.471)
Pre-treatment booking value	−0.006***	−0.006***	−1.782***	−1.722***
	(0.000)	(0.000)	(0.168)	(0.091)
Pre-treatment gross guest spend	0.005***	0.005***	2.083***	2.038***
	(0.000)	(0.000)	(0.141)	(0.078)
Smart pricing pre-treatment	3.376***	3.437***	362.779***	348.078***
	(0.164)	(0.096)	(23.840)	(13.857)
Stratum F.E.	Yes	Yes	Yes	Yes
Robust s.e.	Yes	Yes	Yes	Yes
Clustered s.e.	No	Yes	No	Yes
R^2	0.282	0.283	0.381	0.373
Adjusted R^2	0.280	0.282	0.379	0.373

Note: *p<0.1; **p<0.05; ***p<0.01

Table F.5   **Results of the algorithmic pricing meta-experiment (nights booked and gross guest spend)**

	Dependent variable:	
	Nights booked	Booking value
	(1)	(2)
Treatment	−0.178	1.682
	(0.121)	(16.904)
Bernoulli Randomized	−0.057	15.840
	(0.154)	(20.941)
Bernoulli × Treatment	−0.104	−37.238
	(0.184)	(23.988)
Pre-treatment bookings	1.366***	86.295***
	(0.007)	(0.941)
Pre-treatment nights booked	0.149***	−20.025***
	(0.002)	(0.429)
Pre-treatment booking value	−0.005***	−1.717***
	(0.000)	(0.080)
Pre-treatment gross guest spend	0.005***	2.033***
	(0.000)	(0.068)
Smart pricing pre-treatment	3.382***	344.350***
	(0.084)	(12.096)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	Yes
R^2	0.281	0.374
Adjusted R^2	0.280	0.373
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table F.6   Cluster randomized fees experiment (individual- and cluster-level analysis)

	Dependent variable:	
	Individual-level	Cluster-level
	(1)	(2)
Treatment	−0.142***	−0.137***
	(0.011)	(0.011)
Pre-treatment bookings	0.174***	0.206***
	(0.001)	(0.006)
Pre-treatment nights booked	−0.003***	0.003*
	(0.000)	(0.002)
Pre-treatment booking value	0.000***	−0.000
	(0.000)	(0.000)
Pre-treatment gross guest spend	−0.000***	0.000
	(0.000)	(0.000)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	No
R^2	0.405	0.973
Adjusted R^2	0.405	0.968

Note: *p<0.1; **p<0.05; ***p<0.01

Table F.7   **Cluster randomized algorithmic pricing experiment (individual- and cluster-level analysis)**

	Dependent variable:	
	Individual-level	Cluster-level
	(1)	(2)
Treatment	−0.051*	−0.051*
	(0.029)	(0.029)
Pre-treatment bookings	0.828***	1.114***
	(0.002)	(0.017)
Pre-treatment nights booked	−0.017***	−0.006
	(0.000)	(0.005)
Pre-treatment booking value	0.000***	0.000**
	(0.000)	(0.000)
Pre-treatment gross guest spend	−0.000***	−0.000*
	(0.000)	(0.000)
Smart pricing pre-treatment	0.586***	−0.777***
	(0.020)	(0.172)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	No
R^2	0.578	0.951
Adjusted R^2	0.578	0.941
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table F.8   Results of the fees Meta-experiment (individual and mixed analysis)

	Dependent variable:	
	Bookings	
	(1)	(2)
Treatment	−0.139***	−0.139***
	(0.011)	(0.011)
Bernoulli Randomized	0.022	0.021
	(0.014)	(0.014)
Bernoulli × Treatment	−0.067***	−0.068***
	(0.016)	(0.016)
Pre-treatment bookings	0.174***	0.175***
	(0.001)	(0.001)
Pre-treatment nights booked	−0.003***	−0.003***
	(0.000)	(0.000)
Pre-treatment booking value	0.000***	0.000
	(0.000)	(0.000)
Pre-treatment gross guest spend	−0.000***	−0.000
	(0.000)	(0.000)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	No
R^2	0.405	0.515
Adjusted R^2	0.405	0.515
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table F.9   **Results of the algorithmic pricing meta-experiment (individual and mixed analysis)**

	Dependent variable:	
	Bookings	
	(1)	(2)
Treatment	−0.050*	−0.050*
	(0.030)	(0.030)
Bernoulli Randomized	−0.013	−0.014
	(0.037)	(0.037)
Bernoulli × Treatment	−0.059	−0.060
	(0.041)	(0.041)
Pre-treatment bookings	0.827***	0.838***
	(0.002)	(0.004)
Pre-treatment nights booked	−0.017***	−0.018***
	(0.000)	(0.001)
Pre-treatment booking value	0.000***	0.000
	(0.000)	(0.000)
Pre-treatment gross guest spend	−0.000***	−0.000*
	(0.000)	(0.000)
Smart pricing pre-treatment	0.577***	0.358***
	(0.017)	(0.037)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	No
R^2	0.577	0.692
Adjusted R^2	0.577	0.691
Note:	*p<0.1; **p<0.05; ***p<0.01	

Acknowledgments