Reducing Interference Bias in Online Marketplace Experiments using Cluster Randomization: Evidence from a Pricing Meta-Experiment on Airbnb

David Holtz
Haas School of Business, University of California, Berkeley, Berkeley, CA 94720, dholtz@haas.berkeley.edu

Felipe Lobel
University of California, Berkeley, Berkeley, CA 94720, lobel@berkeley.edu

Ruben Lobel, Inessa Liskovich
Airbnb, 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 randomized experiments to measure the impact of proposed product changes. However, given that marketplaces are inherently connected, total average treatment effect (TATE) estimates obtained through individual-level randomized experiments may be biased due to violations of the stable unit treatment value assumption, a phenomenon we refer to as "interference bias." Cluster randomization, i.e., the practice of randomizing treatment assignment at the level of "clusters" of similar individuals, is an established experiment design technique for countering interference bias in social networks, but it is unclear ex ante if it will be effective in marketplace settings. In this paper, we use a meta-experiment or "experiment over experiments" conducted on Airbnb to both provide empirical evidence of interference bias in online marketplace settings and assess the viability of cluster randomization as a tool for reducing interference bias in marketplace TATE estimates. Results from our meta-experiment indicate that at least 19.76% of the TATE estimate produced by an individual-randomized evaluation of the platform fee increase we study is attributable to interference bias and eliminated through the use of cluster randomization. We also find suggestive, non-statistically significant evidence that interference bias in seller-side experiments is more severe in demand-constrained geographies, and that the efficacy of cluster randomization at reducing interference bias increases with cluster quality.

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

1. Introduction

Many of the world's most highly valued and/or fastest growing technology firms (e.g., Airbnb, Uber, Etsy) are online peer-to-peer marketplaces. These platforms create markets for many different types of goods, including transportation, accommodations, artisanal goods, and even

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 produce an unbiased estimate of the total average treatment effect (TATE) relies on the stable unit treatment value assumption (SUTVA) (Rubin 1974), one component of which is the "no interference" assumption (Cox 1958). This assumption states that in any given experiment, each unit's outcome is a function only of their own treatment assignment, not the treatment assignments of others.

Bias in TATE estimates due to interference, which we refer to in this paper as "interference bias", is likely to occur in online marketplace settings because the buyers and sellers in marketplaces are inherently connected; different goods for sale in a marketplace are likely to complement or substitute for one another, and sellers are likely to make strategic decisions based on the actions of their competitors. Previous work (Blake and Coey 2014, Fradkin 2015) suggests that naive experimentation in online marketplace settings can lead to TATE estimates that are overstated by up to 100%, and as a result, a quickly emerging body of academic research (Liu et al. 2021, Johari et al. 2022, Bojinov et al. 2022, Bright et al. 2022, Li et al. 2022) focuses on how to properly account for interference bias specifically in the context of online marketplaces.[1] Both researchers and academics consider this an important problem to solve because decision-making based on experiment designs and analyses that fail to account for interference bias can have a non-trivial and negative financial impact for online marketplace firms.[2] However, there is still limited empirical work providing insight into the actual severity of interference bias, particularly in seller-side experiments.

Interference bias as a general phenomenon is not unique to online marketplaces, and has been well-studied in the research literature on unipartite social networks; in such settings, interference arises due to interactions between individuals, often referred to as peer effects (Manski 2000, Moffitt et al. 2001). For instance, the observed behavior of one's peers can affect voting behavior (Bond et al. 2012), exercise habits (Aral and Nicolaides 2017), and mobility levels (Holtz et al. 2020). One tool for reducing interference bias in social network experiments is graph cluster randomization (GCR) (Ugander et al. 2013, Eckles et al. 2017), an experiment design technique in which the relevant network is clustered and treatment assignment is

[1] A working version of our paper predates and is cited by much of this research.

[2] In Appendix A, we use a simple economic model to explore the potential financial ramifications of misestimating price elasticities for an online marketplace intermediary.

then randomized at the cluster-level, as opposed to the individual-level. While GCR is an established method in the network experimentation literature, it is unclear *ex ante* if cluster randomization will be an effective tool to reduce interference bias in online marketplaces. This is largely due to factors arising from the bipartite nature of online marketplaces: the mechanisms driving interference may be different than those in a social network setting,[3] and the appropriate mathematical model of interference in marketplaces may deviate from the one used to model the "self-reinforcing" spillovers seen in many unipartite network settings (i.e., positive (negative) direct effects lead to positive (negative) spillover effects).

In this paper, we use a randomized meta-experiment on Airbnb[4] to simultaneously 1) provide empirical evidence of interference bias in an online marketplace seller-side pricing experiment and 2) propose and assess the viability of utilizing cluster randomization to reduce interference bias in such settings. We test for interference bias in a pricing experiment in particular because pricing experiments are of special interest to online marketplace intermediaries; experiments related to prices help firms better understand the price elasticity of their customers, which consequently enables them to implement optimal pricing-related marketplace mechanisms such as fee structures and seller pricing suggestions. Understanding customer price elasticities can also be beneficial to sellers, who set their own prices. Results from our meta-experiment indicate that cluster randomization *is* a viable tool for reducing interference bias in seller-side marketplace experiments, and that interference bias would have accounted for at least 19.76% of the "naive" TATE estimate produced by an individual-level randomized evaluation of the treatment intervention we study.

We begin by using a pre-existing linear model of interference to explore how online marketplace interference differs from social network interference, and the implications this has for experiment design. Interference in this model is captured by a matrix $\boldsymbol{B}$, which we refer to as the "interference matrix." In order to construct an appropriate interference matrix for online marketplace settings, it is necessary to understand the mechanism(s) that drive interference. One possibility is that interference in online marketplaces operates via the same mechanism as social network interference, i.e., it is driven by sellers observing each others' actions and/or interacting. To assess whether this is plausible, we use proprietary data from Airbnb to measure the frequency with which Airbnb hosts search in their own geographies and view the

[3] As a result, if an experiment designer were to try and create a "network" of sellers and perform GCR, it is not immediately obvious how edges between sellers should be defined.

[4] Airbnb is an online marketplace for accomodations and experiences. More than six million listings appear on Airbnb, and since the company's founding in 2008, over one billion guest arrivals have occurred on the platform (Airbnb 2019).

product detail pages (PDP) of other listings. We find that over the course of a month, only 13.3% of listing hosts searched for specific dates in their own geographies and only 21.3% of hosts had at least one PDP view in their own geography. These results suggest that it is unlikely social influence is a significant contributor to interference in online marketplaces. In contrast, a simple simulation of online marketplace dynamics that does not include any seller behavior (see Appendix B) produces results consistent with the existence of interference, suggesting that competitive dynamics are likely a contributor to marketplace interference. In other words, the amount of interference between listings is at least in part determined by the extent to which they co-occur within the consideration sets of shoppers.

Another difference between social network interference and online marketplace interference is that in most social network settings, positive (negative) direct effects beget positive (negative) spillover effects, whereas we expect positive (negative) direct effects to create negative (positive) spillover effects in online marketplaces. We extend a result from Eckles et al. (2017) and show that in the presence of both same-signed and opposite-signed spillovers, cluster randomization will always reduce the bias of the difference-in-means TATE estimator. In doing so, we derive a closed form expression for the expected amount of interference bias remaining under a given clustering; this expression is a function of interference matrix $\boldsymbol{B}$, and can be used to evaluate the "quality" of a given set of clusters.

Building on these insights, we present results from an in vivo meta-experiment, or "experiment over randomized experiments" (Saveski et al. 2017) conducted on Airbnb. The treatment intervention we study in this meta-experiment is a change to Airbnb's platform fee structure; more specifically, hosts in the treatment group were charged *higher* platform fees than hosts in the control group. The meta-experiment design randomly assigned clusters of Airbnb listings to one of two randomization schemes; 25% of clusters were randomized at the individual-level (i.e., treatment is randomly assigned to listings at the indvidual level), whereas the remaining 75% of clusters were cluster randomized (i.e., treatment is randomly assigned to listings at the cluster level). Using this design, we obtain separate TATE estimates in the individual-level and cluster randomized treatment arms, and then test for a statistically significant difference between the two. Results from the individual randomized meta-treatment arm (i.e., the "naive" experiment design) suggest that the treatment led to a statistically significant loss of .345 bookings per listing over the course of the experiment. However, when we compare this TATE estimate to the estimate produced by the cluster-randomized meta-treatment arm, we find that 19.76% of the individual-level TATE estimate is eliminated by cluster randomization and attributable to interference bias. We also find suggestive, non-statistically significant

evidence that interference bias is more severe in demand-constrained geographies, and that the bias reduction from cluster randomization is larger in geographies with "higher quality" clustering.

Situating our work within the broader literature focused on interference bias in online marketplace experiments, we provide an estimate of the potential severity of interference bias in such settings, and evaluate the efficacy of cluster randomization at reducing said bias. We believe there is not a one-size-fits-all solution to interference bias in marketplace experiments, and that each proposed solution (including ours) has its strengths and weaknesses. Cluster randomization works well in marketplaces without centralized matching (in contrast to Bright et al. (2022)), for treatment interventions that must be randomized at the seller-level (in contrast to Johari et al. (2022)), and in marketplaces that are susceptible to intertemporal spillovers (in contrast to Bojinov et al. (2022)). Nonetheless, cluster randomization brings with it substantial reductions in statistical power, and many of our theoretical results apply only to treatment interventions that uniformly increase or decrease demand, but not a mixture of both. We consider both of these weaknesses promising avenues for future research.

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 interference in networks, and one on pricing-related interventions in online marketplaces.

2.1. Interference bias in online marketplace experiments

Our work is most closely related to an emerging body of research focused on the phenomenon of interference-related estimation bias in TATE estimates when conducting experiments in online marketplace settings. This issue was first identified by Blake and Coey (2014) and shortly thereafter by Fradkin (2015), who both report that naive marketplace experimentation can yield TATE estimates that are overstated by up to 100%. In the intervening years, a number of experiment design-based solutions to this problem have been proposed (Liu et al. 2021, Bojinov et al. 2022, Johari et al. 2022, Li et al. 2022) including "two-sided randomization" (Johari et al. 2022) and "switchback" experimentation (Bojinov et al. 2022).[5,6]

[5]Cluster randomization was first proposed as a solution to interference bias in online marketplaces in Holtz (2018), an unpublished master's thesis. The main results from Holtz (2018) now appear in Appendix B of this work.

[6]Analysis-based solutions to the problem have also been suggested, e.g., in Bright et al. (2022).

While each proposed solution to marketplace interference has appealing attributes, none of them offers a "silver bullet" solution. For instance, under two-sided randomization, both buyers and sellers are randomly assigned at the individual-level to treatment or control, and the treatment intervention is only delivered to buyer-seller pairs in which both the seller *and* the buyer have been assigned to the treatment. Two-sided randomization is especially well-suited to corporate experimentation settings, where existing experimentation tooling is often built specifically with individual-randomization in mind. Johari et al. (2022) show that this design reduces bias in TATE estimates due to interference without much loss of precision. However, not all treatment interventions can be delivered at the buyer-seller dyad level, e.g., a new tool for setting prices can only be delivered at the seller-level, and a new search algorithm can only be delivered at the buyer-level. In a switchback experiment design (Bojinov et al. 2022), time is discretized and the experiment designer randomizes the treatment assignment that is delivered to the entire marketplace at each time step. While switchback experiments have appealing statistical properties, they can produce an inconsistent user experience for marketplace participants, and are difficult to implement when markets do not clear quickly, creating "carryover" or temporal spillover effects. This is the case in marketplaces such as Airbnb, where guests often visit the site multiple times over the course of days or weeks before making a booking.

2.2. Interference in networks

The aforementioned papers focus on solving the problem of interference bias in online marketplace experiments, which is uniquely difficult because of the bipartite nature of marketplaces. However, the problem of estimation bias in TATE estimates arising from SUTVA violations is well-studied in settings that are not bipartite. Researchers focused on this topic have developed statistical tests for the existence of interference (Rosenbaum 2007, Aronow 2012, Bowers et al. 2013, Athey et al. 2018), techniques for conducting valid causal inference in the presence of interference (Hudgens and Halloran 2008, Tchetgen and VanderWeele 2012, Aronow and Samii 2017, Sävje et al. 2021, Chin 2018), and experiment designs that account for interference (Sinclair et al. 2012, Imai et al. 2013, Ugander et al. 2013, Liu and Hudgens 2014, Eckles et al. 2017, Saveski et al. 2017, Baird et al. 2018, Basse and Feller 2018, Ariel et al. 2019).

Our work is most closely related to that of Ugander et al. (2013), Eckles et al. (2017), and Saveski et al. (2017), which all focus on experiment designs that deliver cluster-randomized treatment to networks with the aim of obtaining less-biased TATE estimates. Ugander et al. (2013) propose graph cluster randomization (GCR), an experiment design in which, after

clustering a network, treatment assignment is randomized at the cluster-level. The authors show that under certain conditions, GCR eliminates interference bias and produces unbiased TATE estimates. Eckles et al. (2017) build on this work by showing through simulation that in instances where the conditions outlined in Ugander et al. (2013) do not hold, GCR can still greatly reduce interference bias, although it does not eliminate it entirely.[7] Saveski et al. (2017) conduct a "meta-experiment" on LinkedIn that compares the TATE estimate obtained under individual-level randomization to that obtained under GCR. This paper makes two contributions to the literature: providing a method to test for interference bias in network settings, and reporting results that highlight the efficacy of GCR at reducing said bias.

In their totality, these papers provide a thorough exploration of GCR as a method for reducing interference bias in network settings. However, because of the bipartite nature of marketplaces, differences in the mechanisms driving interference, and differences in the appropriate way to mathematically model said interference, it is unclear ex ante if cluster randomization will be as effective in the marketplace setting. Thus, in this work we propose cluster randomization as a method to reduce interference bias in marketplace experiments, and test its efficacy using a Saveski-style meta-experiment.

2.3. Pricing-related interventions in online marketplaces

Finally, our research connects to the literature on pricing-related interventions in online marketplaces. It is important 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 (Choi and Mela 2019) and pricing-related marketplace mechanisms. For instance, in recent years a growing number of online marketplaces have launched machine-learning based pricing interventions (Ifrach et al. 2016, Dubé and Misra 2017, Filippas et al. 2019, Ye et al. 2018). Many pricing interventions are tested and launched using randomized experiments, however, if the TATE estimates produced by these experiments are biased, marketplace designers may mis-estimate price elasticities and/or launch suboptimal policies. For instance, in Appendix A, we use a simple economic model to show that setting platform fees based on biased elasticity estimates reduces firm profits. These losses have the potential to wipe out the positive impacts typically associated with A/B testing (Feit and Berman 2019, Azevedo et al. 2020). Our work confirms that interference *can*

[7] One drawback of assigning treatment at the cluster-level is that most treatment effect estimators will have less statistical power than under an individual-level randomized design. However, techniques such as regression adjustment (Gerber and Green 2012) and pre- and post-stratification (Moore 2012, Miratrix et al. 2013) can be used in tandem with cluster randomization to mitigate the loss of statistical power.

bias TATE estimates when conducting pricing-related experiments in online marketplaces and establishes that cluster randomization can be an effective tool to reduce this bias.

3. Interference Bias in Online Marketplaces

Before presenting the results of our meta-experiment, we first explore the ways in which interference bias in marketplaces differs from interference bias in social networks, and the implications this has for experiment design. The basis for this exploration is the following linear parametric model of interference, which is studied in, e.g., Eckles et al. (2017) and Pouget-Abadie et al. (2018):

$$Y_i(\boldsymbol{Z}) = \alpha_i + \beta Z_i + \gamma \rho_i + \epsilon_i \tag{1}$$

where Y_i is the outcome of seller i, $\boldsymbol{Z}$ is the treatment assignment vector, β is the "direct" effect of the treatment, γ is the "indirect" effect of the treatment, ρ_i is the percentage of seller i's competitors/neighbors that are treated, and $\epsilon_i \sim N(0,1)$ is independent of ρ_i. The same linear outcome model can be represented in the following way:

$$E[Y_i(\boldsymbol{Z})] = \alpha_i + \sum_{j \in V} B_{ij} Z_j, \tag{2}$$

where Z_j indicates the treatment assignment of seller j, and $\boldsymbol{B}$ is an "interference matrix" capturing the strength of the interference between seller i and seller j.

3.1. Does "Seller Influence" Drive Interference?

The notation above makes it clear that in order to reduce interference bias through experiment design, it is helpful to have some idea how to construct an appropriate interference matrix, $\boldsymbol{B}$. In other words, it is helpful to understand the *mechanisms* that drive interference. Here, we investigate whether interference in online marketplaces operates via a similar mechanism to interference in social networks, i.e., it is driven by sellers observing the behavior of other sellers and changing their behavior in response. To do so, we reference the search and product detail page (PDP) view activity of Airbnb listing hosts in this paper's meta-experiment in the month prior to the meta-experiment's launch (February 16, 2019 to March 15, 2019). We find that the overwhelming majority of Airbnb hosts do not search in their own geographies or view the PDPs of competitors, suggesting that the "seller influence" mechanism is unlikely to play a major role in driving spillovers in our context. More specifically, in the month preceding our meta-experiment, only 22.7% of listing hosts searched at least once in their own geography,

and only 13.3% searched at least once for specific dates in their own geography. Among hosts who ran at least one search in their own geography, the median host searched only 8 times. Furthermore, only 21.3% of hosts had at least one PDP view to a within-geography listing that wasn't their own. Among hosts that had at least one PDP view to a within-geography listing that wasn't their own, the median host carried out 4 PDP views across 3 distinct listings. More detailed data on search and PDP view activity in the month preceding our meta-experiment is shown in Figure 1. Given these results, in conjunction with the fact that 1) like our meta-experiment, many experiments run for much shorter periods of time than 30 days and 2) treatment interventions like the one we study in our meta-experiment are often subtle and unlikely to be noticed by hosts after just a few search sessions or PDP views, we consider it likely that interference in online marketplaces is driven not by "seller influence," but instead by the fact that sellers co-occur in the consideration sets of potential buyers and compete with each other for transactions.[8]

3.2. Modeling Interference in Online Marketplaces

Another point of contrast between interference in online marketplaces and interference in many social network settings is the nature of the interference between units. Many network experiments study treatment interventions with "self-reinforcing" spillovers, i.e., treatment interventions in which positive (negative) treatment interventions have positive (negative) spillovers (put differently, β and γ in Equation 1 have the same sign). For instance, a vaccination encouragement intervention might increase vaccination rates not only among those that are treated, but also among their peers. Similarly, in a social media setting we would typically expect an intervention that increases the posting activity of treated users to also increase the posting activity of treated users' peers.

In contrast, many potential marketplace treatment interventions act on seller outcomes in such a way that β and γ have opposite signs, since sellers and buyers compete with one another. For instance, if an intervention caused treated Airbnb hosts to raise (lower) their prices, this could lead to an decrease (increase) in demand for their listings, and, consequently, a increase (decrease) in demand for their competitors' listings.[9] This is exactly the pattern we observe in

[8]The notion that spillovers in online marketplaces are driven by competitive dynamics is consistent with the simulation results found in Appendix B.

[9]It is also possible that Airbnb hosts in a given geography could 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. However, we consider it much more likely that accommodations on Airbnb are substitutes, and assume this to be the case throughout the rest of this work.

the fee meta-experiment results presented in Section 5. While the TATE of increasing platform fees is negative (we estimate a TATE of -0.277 bookings per listing in the cluster-randomized meta-treatment arm), the bias we observe points in the opposite direction (we estimate a TATE of -0.345 bookings per listing in the individual-randomized meta-treatment arm). We claim that this is because Airbnb customers are more likely to see a mixture of treatment and control listings under individual-level randomization, and customers who see such a mixture will shift their business from high-fee listings to low-fee listings.

Eckles et al. (2017) show that when β and γ have the same sign, i.e., when spillovers are "self-reinforcing," cluster randomization will always reduce the bias of the TATE estimator relative to individual-level randomization. However, they stop short of proving that this is true in cases where the direct and indirect treatment effects point in opposite directions, as is likely to be the case in online marketplace settings. We introduce the following proposition, which extends Theorem 2.1 from Eckles et al. (2017) and shows that cluster randomization is guaranteed to reduce the bias of TATE estimates, even in cases where the direct and indirect effects of a treatment intervention (captured by the interference matrix) have opposite signs.

Proposition 1. *Assume we have a linear outcome model for all sellers $i \in S$ that is a function of the form*

$$E[Y_i(\boldsymbol{Z})] = \alpha_i + \sum_{j \in V} B_{ij} Z_j, \tag{3}$$

where Z_j indicates the treatment assignment of seller j, and $\boldsymbol{B}$ is a matrix in which all of the diagonal entries have the same sign and all of the off-diagonal entries have the same sign. Then for any mapping of sellers to clusters $C(\cdot)$, the absolute bias of the difference-in-means TATE estimate under cluster randomization, $\hat{\tau}_{cr}$, is less than or equal to the absolute bias of the difference-in-means TATE estimate under individual-level randomization, $\hat{\tau}_{ind}$, with a fixed treatment probability p.

Proof. Given in Appendix C.

Proposition 1 establishes that cluster randomization will never increase TATE estimation bias, but does not provide any guidance on how to construct clusters. In any given marketplace setting, there will exist many different ways to cluster sellers. For instance, an experiment designer might cluster sellers based on seller-level attributes, observed rates of seller co-occurrence in search, or estimated cross-price elasticities, to name a few possibilities. However, not all clusterings will be equally effective at reducing TATE estimation bias. For instance, if a given

approach to clustering produces clusters that are essentially random, bias reduction will be very close to 0, whereas if a given clustering does a very good job of capturing the relevant marketplace dynamics, bias reduction has the potential to be much larger. Given this fact, it is natural for an experiment designer to want to identify the clustering that will lead to the greatest reduction in estimation bias.

Unfortunately, there isn't a singular optimal method for clustering; the most effective clustering strategy will vary depending on the specific research context and the treatment intervention being studied. Considering this, it is necessary to develop a concept of 'cluster quality' that is adaptable to different contexts and takes into account the relevant interference matrix, $\boldsymbol{B}$, for a specific experiment. Thankfully, our proof of Proposition 1 provides a valuable resource. The left-hand side of the final inequality in this proof helps us quantify the bias of the difference-in-means TATE estimator within a given clustering. This bias quantification can be used as an indicator of the quality for a defined set of clusters, represented as $C(\cdot)$.

Definition 1. *The quality of a given set of clusters, $Q_C(\boldsymbol{B})$, is defined as*

$$Q_C(\boldsymbol{B}) = \left| \sum_{i=1}^{N} \sum_{j=1}^{N} B_{ij} \mathbb{1}\left(C(i) \neq C(j)\right) \right|. \tag{4}$$

Although in theory, Definition 1 provides a context-dependent measure of cluster quality, in practice, the relevant interference matrix $\boldsymbol{B}$ for a given research setting and treatment intervention is almost never observable to experiment designers. However, as long as the experiment designer is able to construct some proxy matrix $\boldsymbol{P}$ that is a monotonic transformation of $\boldsymbol{B}$, it follows directly from Proposition 1 that $Q_C(\boldsymbol{P})$ can still be used to determine which of two sets of clusters, C_1 and C_2, produces more biased difference-in-means TATE estimates.[10]

Proposition 2. *Suppose that $\boldsymbol{P}$ is a monotonic transformation of $\boldsymbol{B}$. Then,*

$$Q_{C_1}(\boldsymbol{P}) \leq Q_{C_2}(\boldsymbol{P}) \implies Q_{C_1}(\boldsymbol{B}) \leq Q_{C_2}(\boldsymbol{B}). \tag{5}$$

These results suggest that 1) for seller-side marketplace interventions that uniformly increase or decrease demand for treated sellers, cluster randomization should always reduce interference bias, regardless of cluster quality (although bias reductions will increase with cluster quality) and 2) after identifying a set of clusters, $C(\cdot)$, an experiment designer can

[10]Note that because $\boldsymbol{B}$ is typically not observable, the statement that a given proxy matrix $\boldsymbol{P}$ is a monotonic transformation of $\boldsymbol{B}$ will almost always rely on a set of modeling assumptions that are not empirically testable.

assess their quality by calculating $Q_C(\boldsymbol{P})$.[11] In Section 5.4, we investigate how cluster quality moderates the extent to which cluster randomization reduces interference bias in our meta-experiment. The measure of cluster quality used in this analysis is calculated using a proxy matrix $\boldsymbol{P}$ based on listing co-occurrence in searcher-level PDP view sessions. The intuition behind this choice is that in order for two Airbnb listings to compete with one another for bookings, they need to co-occur in searchers' consideration sets. In Appendix F, we provide more detail on how we calculated this particular $Q_C(\boldsymbol{P})$ using browsing data from Airbnb.

4. Platform Fee Meta-Experiment

Although the theoretical results in the previous section suggest that cluster randomization should reduce interference bias in seller-side marketplace experiments, it is unclear if this is true in practice. Furthermore, even if interference bias in seller-side marketplace experiments is a theoretical concern, it may not be a practical one if the severity of interference bias is small. If the magnitude of interference bias is small and/or cluster randomization is not an effective bias reduction technique, cluster randomization may not be worth implementing; cluster randomization is more logistically complicated and many industry experimentation tools do not easily support cluster randomization.

In this section, we describe the design of an in-vivo meta-experiment conducted on Airbnb's platform in March 2019.[12] By analyzing this meta-experiment, we obtain an empirical lower bound on the severity of interference bias in a "naive" individual-level randomized pricing experiment on Airbnb, and also measure the extent to which cluster randomization reduces that bias.[13]

4.1. Treatment Intervention

The treatment intervention we study in our meta-experiment is a change to Airbnb's platform fees for guests. 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

[11] Alternatively, Pouget-Abadie et al. (2018) propose a meta-experiment design that can be used to empirically compare the efficacy of different sets of clusters at reducing TATE bias.

[12] We roughly follow the meta-experiment design introduced by Saveski et al. (2017). Pouget-Abadie et al. (2018) propose a similar "experiment over experiments" design. Meta-experiment designs such as these can be thought of as special cases of the randomized saturation designs discussed in, e.g., Baird et al. (2018).

[13] This meta-experiment was motivated by the simulation-based work found in Appendix B. While simulation-based work is helpful for conducting preliminary analysis, we believe that our meta-experiment provides value above and beyond simulation based work, since any simulation-based study of interference in marketplaces (including ours) will rely on assumptions about consumer behavior, the nature of the interference between units, etc.

when a listing appears in search (top panel of Figure 2). Second, if a guest opens the "price breakdown" tooltip on any search result, they are shown a price breakdown that separates out the nightly price and the guest platform fee (bottom panel of Figure 2). Finally, when viewing a listing's PDP, a detailed pricing breakdown (including fees) is displayed next to the "Request to Book" button (Figure 3).

Our meta-experiment targeted long-tenured listings (i.e., listings that had been listed on Airbnb as of a certain cutoff date). Listings in the treatment had their guest fees *increased* relative to the status quo, whereas listings in the control had their fees *decreased* relative to the status quo. Less-tenured listings (i.e., listings created after the cutoff date) did not have their fees changed relative to the status quo.[14,15]

4.2. Experiment Design

Our meta-experiment design is extremely similar to the "experiment over experiments" design described in Saveski et al. (2017). First, Airbnb listings were sorted into clusters using the process described in Section 4.2.1. Clusters were then randomly assigned to one of two meta-treatment arms: individual-level randomization (25% of clusters), or cluster randomization (75% of clusters). Within the individual-level randomized meta-treatment arm, treatment was randomly assigned to listings at the individual level. Within the cluster-randomized meta-treatment arm, treatment was randomly assigned to listings at the cluster level. The entire meta-experiment design is summarized in Figure 4.

Each meta-treatment arm can be analyzed as a standalone experiment that produces a TATE estimate, and then, by jointly analyzing the data from both meta-treatment arms, we are able to measure whether there is a statistically significant difference between these two estimates. In order to increase statistical power for this comparison, we arranged our clusters into strata and use post-stratification (Miratrix et al. 2013) when analyzing our data. The process we used to generate those strata is described in Section 4.2.3.

4.2.1. Generating Hierarchical Listing Clusters The first step in the design of our meta-experiment was arranging listings into clusters. There are many different ways to sort

[14]Due to our NDA with 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 used to determine whether listings were long-tenured. Furthermore, all of our outcome variables (bookings, nights booked, gross guest spend) are multiplied by a random constant.

[15]Because our meta-experiment only impacts fees for long-tenured listings, we restrict our analysis dataset to long-tenured listings. However, the clusters used in our experiment include all listings, regardless of tenure on the platform.

listings into clusters (e.g., the simulation described in Appendix B takes a graph clustering approach to generating clusters: edges were drawn between listings that share observable traits, and the resulting graph was clustered using Louvain clustering (Blondel et al. 2008)). For our in-vivo meta-experiment, we took an approach to clustering that made use of technical infrastructure that already existed at Airbnb. The first step in the process of generating these clusters was generating a dense, 16-dimensional demand embedding for each listing. Listings were then arranged into hierarchical clusters based on their location in that 16-dimensional space. Finally, a maximum cluster size was chosen in order to determine which subset of the hierarchical clusters to use in our meta-experiment.[16]

We generated demand embeddings for each Airbnb listing using a process similar to the one described in Grbovic and Cheng (2018). The training data used to generate our demand embeddings consisted of sequences of listings that individual users viewed 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:

$$< L_A, L_B, L_C > . \tag{6}$$

We used 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

$$J = \max_{W,V} \sum_{s \in S} \frac{1}{|s|} \sum_{i=1}^{|s|} \sum_{-k \leq j \leq k,\, k \neq 0} \log p\left(L_{i+j} | L_i\right), \tag{7}$$

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 was augmented by including listing-level attributes (e.g., a listing's geography) in the search session sequences. The model was then trained using a geography-level negative sampling approach.

Once listing embeddings were generated using the aforementioned approach, a recursive partitioning tree (Kang et al. 2016) was used to arrange the Airbnb listings into hierarchical

[16] We believe that providing guidance on cluster construction is beyond the scope of this paper, given that the "optimal" set of clusters for cluster randomization will vary depending on the research setting and the treatment intervention of interest. However, the cluster quality metric provided in Definition 1 can be a useful tool for adjudicating between two candidate sets of clusters. We also believe that the analyses and theoretical results in this paper provide a roadmap of sorts that other researchers can draw on when designing clusters for the purpose of a cluster-randomized marketplace experiment. We discuss this point further in Section 6.

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 maximum size by applying a cut to the hierarchy of clusters generated by the recursive partitioning tree. Figure 5 depicts example clusters generated using this method in the San Francisco Bay Area. Using an ad-hoc approach, we chose a cluster size threshold of 1,000 for the fee meta-experiment. This ad-hoc approach is described in Appendix D.

4.2.2. Treatment assignment randomization Once each Airbnb listing was assigned to a cluster, 75% of clusters were randomly assigned to the "meta-treatment" (cluster randomization) and 25% of clusters were randomly assigned to the "meta-control" (individual-level randomization). Within the meta-control arm, Bernoulli individual-level randomization was used to assign 50% of listings to the treatment and 50% of listings to the control. Within the meta-treatment arm, Bernoulli cluster randomization was used to assign 50% of clusters to the treatment and 50% of clusters to the control. Each listing in a meta-treatment cluster was assigned the treatment assignment corresponding to its cluster.

4.2.3. Strata for post-stratification In our meta-experiment analysis, we use post-stratification (Miratrix et al. 2013) to increase statistical power. The strata we use for this purpose were generated using a multivariate blocking procedure (Moore 2012). As a first step, we collected pre-treatment listing-level data for the period running from January 16, 2019 to February 17, 2019. Across this period, we calculated cluster-level summary statistics: the average number of nights booked per listing, the average number of bookings per listing, the average gross guest spend per listing, and the number of non-experimental holdout listings in the cluster.[17] After centering and scaling each of these metrics, we calculated the Mahalanobis distance (Mahalanobis 1936) between each pair of clusters. Finally, we used an optimal-greedy algorithm to arrange clusters into strata of maximum size $n = 8$.

4.3. Experiment Preliminaries

The meta-experiment was run from March 16, 2019 to March 21, 2019 on a sample of 2,602,782 listings.[18] Of those listings, 647,377 were assigned to the listing-randomized meta-control

[17] At the time of our meta-experiment, experiments on Airbnb excluded listings in a long-term experiment holdout group, as well as listing in Airbnb's "Plus" tier.

[18] Shortly after the meta-experiment's conclusion, 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 potential negative impact of the meta-experiment on Airbnb hosts.

arm, and the remaining 1,955,405 were assigned to the cluster-randomized meta-treatment arm. Within the listing-randomized meta-treatment arm, 323,734 listings were assigned to the control and 323,643 listings were assigned to the treatment. Within the cluster-randomized meta-treatment arm, 2,981 clusters were assigned to the treatment and 2,979 clusters were assigned to the control, resulting in 979,015 listings assigned to the treatment and 976,390 listings assigned to the control. In total, across both meta-treatment arms, 1,300,124 listings were assigned to the control and 1,302,568 listings were assigned to the treatment. We check for balance on pre-treatment outcome variables between the meta-treatment and meta-control clusters, and between the control and treatment groups in both meta-treatment arms (see Table 1); we do not detect any statistically significant differences, indicating our randomization procedure was sound.

5. Results

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

We first present the results from separately analyzing the individual-level randomized and cluster randomized arms of the meta-experiment. While the individual-level 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 individual-level randomized and cluster randomized meta-treatment arms. Finally, we investigate the extent to which our results vary as a function of 1) the level of supply- or demand-constrainedness in an Airbnb marketplace and 2) the geography-level quality of our clusters.

5.1. Individual-level & Cluster Randomized Results

We analyze both the individual-level randomized and cluster randomized meta-treatment arms separately by estimating the following model on listing-level data,

$$Y_i = \alpha + \beta T_i + \sum_l \gamma_l \mathbb{1}(B_i = l) + \delta X_i + \epsilon_i \tag{8}$$

where Y_i is the number of bookings, 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, gross guest spend, calendar nights available, and geography-level number of searches per available night in the month prior to the meta-experiment, and ϵ_i is an error term. For the cluster-randomized meta-treatment arm, we cluster standard errors at the Airbnb listing cluster-level.[19]

Table 2 shows the TATE estimate for bookings in both the individual-level randomized (column 1) and cluster randomized (column 2) meta-treatment arms. In the individual-level randomized meta-treatment arm, the TATE is -0.345 bookings per listing, whereas in the cluster randomized meta-treatment arm, the TATE is -0.277 bookings per listing. Both of these TATE estimates are statistically significant at the 95% confidence level.

5.2. Joint Analysis

In order to determine whether the difference between the TATE estimates generated by the two meta-treatment arms is statistically significant, we estimate the model

$$Y_i = \alpha + (\beta + \nu M_i)T_i + \xi M_i + \sum_l \gamma_l \mathbb{1}(B_i = l) + \delta X_i + \epsilon_i, \tag{9}$$

where Y_i is the outcome of interest, M_i is a binary variable set to 1 when listing i is in the individual-level 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 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. Standard errors are clustered at the individual-level for listings in the individual-level randomized meta-treatment arm, and at the Airbnb listing cluster-level for listings in the cluster-randomized meta-treatment arm.[20]

In the above model, β measures the "true" effect of the treatment, and ν measures the difference between the estimated effect of the treatment in the individual-level randomized arm and the estimated effect of the treatment in the cluster randomized arm. In other words, ν should measure the extent to which cluster randomization reduces interference bias, and also

[19] In order to increase statistical power, our preferred model specification is Equation 8, which utilizes post-stratification (Miratrix et al. 2013) through the inclusion of stratum-level indicators. Results obtained from estimating a more straightforward model that regresses bookings only on treatment assignment can be found in Table H.7.

[20] In order to increase statistical power, our preferred model specification is Equation 9, which utilizes post-stratification (Miratrix et al. 2013) through the inclusion of stratum-level indicators. Results obtained from estimating a more straightforward model that regresses bookings only on meta-treatment assignment, treatment assignment, and their interaction can be found in Table H.8.

provide a lower bound on the amount of interference bias in the individual-level randomized meta-treatment arm.[21] ξ measures any baseline difference between listings in the individual-level randomized arm of the meta-experiment and listings in the cluster-randomized arm of the meta-experiment; since clusters were randomly assigned to meta-treatment arms, we expect ξ to be zero. Once we have estimated Equation 9, our estimate of the interference bias is

$$\Omega = \frac{\hat{\nu}}{\hat{\nu} + \hat{\beta}}, \tag{10}$$

i.e., the percentage of the listing-randomized meta-treatment arm TATE estimate that does *not* appear in the cluster-randomized meta-treatment arm TATE estimate. We calculate standard errors on this quantity using the delta method (we use the `deltamethod` function in the R library `msm`).

Column 1 of Table 3 and Figure 6 show the results from estimating Equation 9 on our entire sample. We estimate that the "true" TATE is -0.277 bookings per listing, whereas -0.068 bookings per listing of the TATE measured in the listing-randomized meta-treatment arm is due to interference bias. Plugging these point estimates into Equation 10, we estimate that 19.76% ($\pm$9.06%) of the TATE estimate achieved through the individual-level randomized experiment is due to interference bias, and was eliminated through cluster randomization.

5.3. The Moderating Effect of Supply and Demand Constrainedness

We hypothesize that the extent to which the TATE estimate under listing-level randomization suffers from interference bias will depend on marketplace conditions. More specifically, we expect that interference bias will be *larger* in geographies that are demand constrained, and *smaller* in geographies that are supply constrained. The intuition for this is as follows: in an extremely supply-constrained geography, all listings will eventually get booked, which will push the interference bias to zero, whereas in an extremely demand-constrained geography, only "more appealing" listings (i.e., only those in the treatment or control, depending on the treatment intervention) will be booked, maximizing interference bias. Simulation-based evidence motivating this hypothesis can also be found in Johari et al. (2022).

To test this hypothesis, we re-estimate Equation 9 separately for listings that are above/below the median listing in terms of the supply-constrainedness of their geography. Our measure of "supply constrainedness" is relatively crude, but effective: we divide the number

[21] Recall that even when using cluster randomization, TATE estimates will likely remain biased to some extent, since any given clustering will do an imperfect job of capturing every pair of listings that interfere with one another.

of searches occurring in a given geography in the month prior to our meta-experiment by the number of calendar nights available in the geography at the outset of the month prior to our experiment. Columns 2 and 3 of Table 3 display our results for supply-constrained and demand-constrained geographies, respectively; these results are also visualized in Figure 7. We estimate that 12.05% (±11.55%) of the listing-level randomized TATE estimate in supply-constrained geographies can be attributed to interference bias, whereas 28.65% (±14.91%) of the listing-level randomized TATE estimate in demand-constrained geographies can be attributed to interference bias. While these results are consistent with both our hypothesis and the results reported in Johari et al. (2022), the difference between these two point estimates is not statistically significant (see Column 1 of Table H.9), and hence these results should only be considered suggestive.

5.4. The Moderating Effect of Cluster Quality

We also hypothesize that geographies with higher quality clusters (as defined in Definition 1) should see a greater reduction in interference bias. Using a process described in Appendix F, we construct a geography-level measure of cluster quality. Under this measure, which uses a proxy for the "true" interference matrix $\boldsymbol{B}$ based on user-level PDP view sessions, a given clustering is considered "higher quality" if listings tend to co-occur with listings from the same cluster in user-level PDP view sessions. We proceed to split listings into those that are above or below the median listing in terms of geography-level clustering quality, and separately estimate Equation 9 on these two samples. Columns 4 and 5 of Table 3 display our results for low-quality and high-quality clustering, respectively; these results are also visualized in Figure 7. We find that clustering reduces the TATE estimate by 25.92% (±15.14%) in geographies with high-quality clusters, and reduces the TATE estimate by 14.98% (±11.69%) in geographies with low-quality clusters. As was the case for our heterogeneity analysis with respect to supply-constrainedness, although these results are consistent with our hypothesis, we consider them suggestive since the difference between these two estimates of interference bias reduction is not statistically significant (see Column 2 of Table H.9).[22]

[22]We conduct the same analysis with an alternate definition of cluster quality that is based on observable listing attributes, as opposed to consumer search data. To construct this alternative measure, we classify two listings as "substitutable" if they are in the same geography-level decile for the following three variables: share of 5 stars trips, person capacity, and price. At the geography-level, we then calculate the average percentage of a listing's "substitutable" listings (including itself) that are in the same cluster. Table H.10 shows our results using this alternative cluster quality measure; they are qualitatively similar to those found in Table 3.

6. Discussion

In this paper, we have highlighted the ways in which interference bias in online marketplaces differs from interference bias in social networks, and presented results from an in vivo meta-experiment conducted on Airbnb. Results from this meta-experiment provide empirical evidence that interference has the potential to cause substantial statistical bias in online marketplace seller-side experiment TATE estimates, and establish that cluster randomization is a promising tool for reducing said bias. More specifically, we find that at least 19.76% of the TATE estimate obtained from our individual-randomized meta-treatment arm was due to interference bias. We also find suggestive, non-statistically significant evidence that interference bias is more severe in demand-constrained geographies, and that higher-quality clusters lead to greater bias reduction in TATE estimates.

While our results show that there *can* be a sizable amount of interference bias in online marketplace experiments, it is possible that different treatment interventions in different marketplaces would be less (or more) prone to estimation bias. Although we are unable to make evidence-based claims on this topic, we believe that the analyses described in this paper provide something of a roadmap for researchers and firms hoping to assess the potential severity of interference bias in their setting and/or use cluster randomization to mitigate it. For instance, researchers might begin by estimating the potential financial impact of interference bias in their setting (Appendix A), conducting observational analysis to better understand the potential mechanisms driving interference in their setting (Section 3.1) and/or running simulated experiments (Appendix B).

When interference bias seems worth accounting for, an appropriate next step would be to weigh the pros and cons of cluster randomization relative to other proposed solutions such as two-sided randomization (Johari et al. 2022) and switchback experimentation (Bojinov et al. 2022). In general, both two-sided randomization and switchback experimentation will reduce TATE estimation bias relative to the individual-level randomized baseline. The extent to which this bias reduction comes at the price of reduced statistical power depends on the amount of supply-demand imbalance (in the case of two-sided randomization) or the strength of temporal "carryover" effects (in the case of switchback experimentation). There are also some treatment interventions for which switchback experimentation and/or two-sided randomization may not be viable (for instance, data-driven decision-making aids cannot be assigned at the buyer-seller dyad level, as is required for two-sided randomization). Beyond relying on domain knowledge and intuition, managers and researchers may find it informative to run simulated

experiments that make reasonable assumptions about, e.g., the strength of carryover effects or the types of sellers that might interfere with one another, and compare the bias and statistical power of different experiment designs and treatment effect estimators in these simulations. As previously mentioned, relative to alternatives, our belief is that cluster randomization is well-suited to seller-side interventions that are susceptible to intertemporal spillovers.

In cases that are best suited to cluster randomization, researchers can consider many different sets of clusters and either calculate and compare the "quality" of said clusters (Appendix F) or conduct a meta-experiment using the design described in Pouget-Abadie et al. (2018) to identify which clustering will provide the greatest bias reduction. Having chosen a set of clusters, one can imagine either running a straightforward cluster randomized experiment to obtain a TATE estimate, or conducting a meta-experiment similar to ours (Section 4.2) to obtain a lower-bound on the actual amount of interference bias present.

We believe our work leaves open multiple promising avenues for future research, the most pressing of them being the development of methods to increase the statistical power of cluster-randomized experiments in online marketplaces. Even in cases where cluster randomization is well-suited to the treatment intervention under evaluation, one major barrier to the adoption of cluster randomization in online marketplaces is the fact that clustering greatly reduces the precision of TATE estimates. Loss of statistical power due to clustering can also make it difficult to estimate the severity of interference bias. This is evidenced by the fact that the confidence interval around our interference bias estimate is still quite wide, despite our meta-experiment including over 2 million Airbnb listings.[23] Future work might focus on, e.g., using meta-experiments to estimate underlying structural parameters of marketplaces (such as price elasticities), and subsequently using those structural parameter estimates to optimize the design of future experiments and/or predict the amount of interference bias associated with other potential treatment interventions.

We believe our work leaves open multiple promising avenues for future research, the most pressing of them being the development of methods to increase the statistical power of cluster-randomized experiments in online marketplaces. Even in cases where cluster randomization is well-suited to the treatment intervention under evaluation, one major barrier to the adoption of cluster randomization in online marketplaces is the fact that clustering greatly reduces

[23]To further emphasize this point, let us provide an explanatory anecdote: prior to the meta-experiment reported in this paper, we conducted a different pricing-related meta-experiment on Airbnb with a less intense treatment intervention. Because the treatment intervention was less extreme, this meta-experiment was underpowered to detect interference bias, despite having a sample size in the millions.

the precision of TATE estimates. Loss of statistical power due to clustering can also make it difficult to estimate the severity of interference bias. This is evidenced by the fact that the confidence interval around our interference bias estimate is still quite wide, despite our meta-experiment including over 2 million Airbnb listings.[24] Future work might focus on, e.g., using meta-experiments to estimate underlying structural parameters of marketplaces (such as price elasticities), and subsequently using those structural parameter estimates to optimize the design of future experiments and/or predict the amount of interference bias associated with other potential treatment interventions.

Furthermore, the results we present in Section 3.2 are somewhat specific to treatment interventions that lead to uniform increases/decreases in demand. However, many treatment interventions of interest, including algorithmic pricing interventions (Ifrach et al. 2016, Dubé and Misra 2017, Filippas et al. 2019, Ye et al. 2018) increase demand for some sellers while decreasing demand for others. Future research might explore theoretical guarantees around cluster randomization in marketplaces when treatment interventions are more complicated than those considered in this paper and/or conduct meta-experiments similar to ours to assess the efficacy of cluster randomization when the treatment intervention under evaluation is more complex.

[24] To further emphasize this point, let us provide an explanatory anecdote: prior to the meta-experiment reported in this paper, we conducted a different pricing-related meta-experiment on Airbnb with a less intense treatment intervention. Because the treatment intervention was less extreme, this meta-experiment was underpowered to detect interference bias, despite having a sample size in the millions.

7. Figures

Figure 1 Panel A shows the distribution of product detail page (PDP) views to within-geography listings, whereas Panel B shows the distribution of unique within-geography listings with at least one PDP view. Panel C shows the distribution of within-geography searches, whereas Panel D shows the distribution of within-geography searches with dates. Searches with dates are generally considered to be higher intent to book.

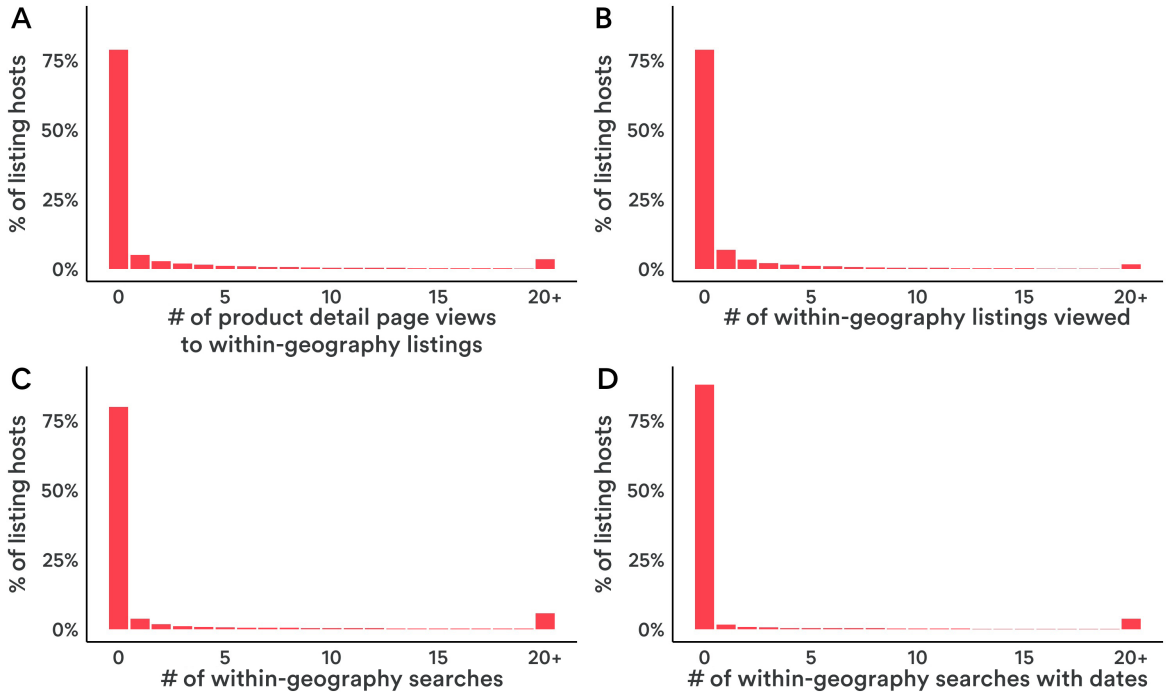


Figure 2 **The top panel shows a typical search result on Airbnb. In this case, the guest platform fee is included in the total price of $508. The bottom panel shows what is displayed to guests after clicking the "price breakdown" tooltip: the guest platform fee (listed here as a service fee of $58) is broken out from the total nightly price.**

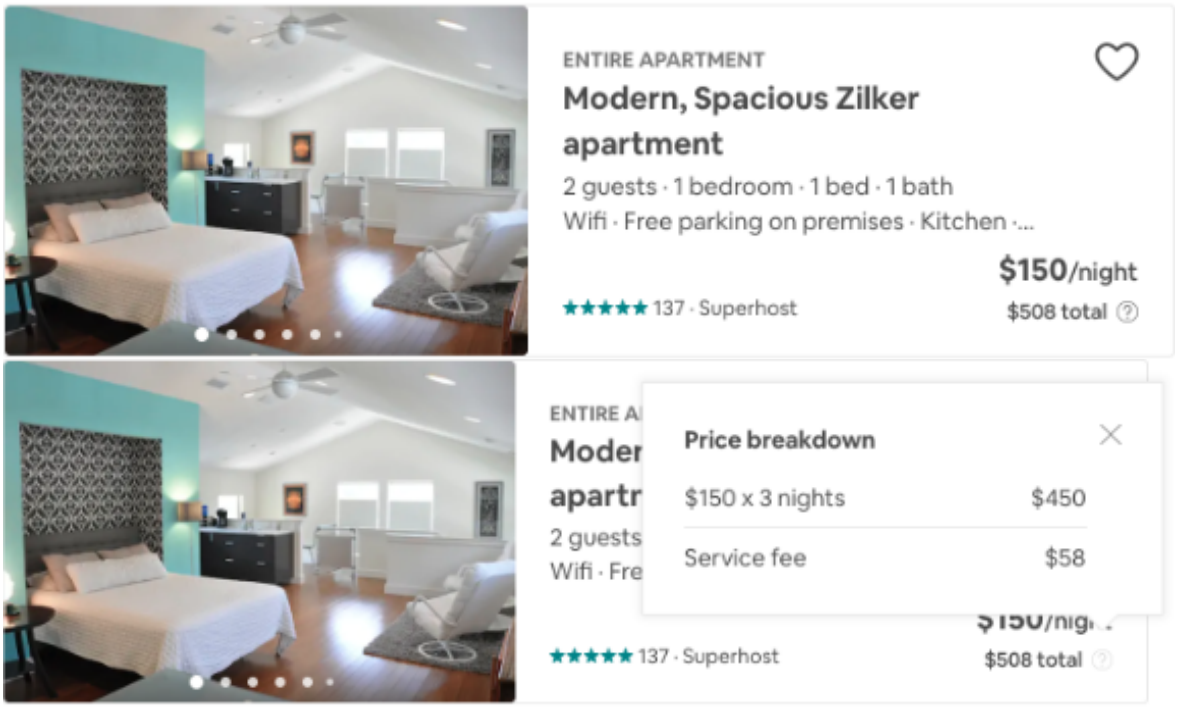


Figure 3 **The section of the Airbnb product detail page that provides a full pricing breakdown for would-be guests. In this pricing breakdown, the guest platform fee (listed here as a service fee) is $58.**

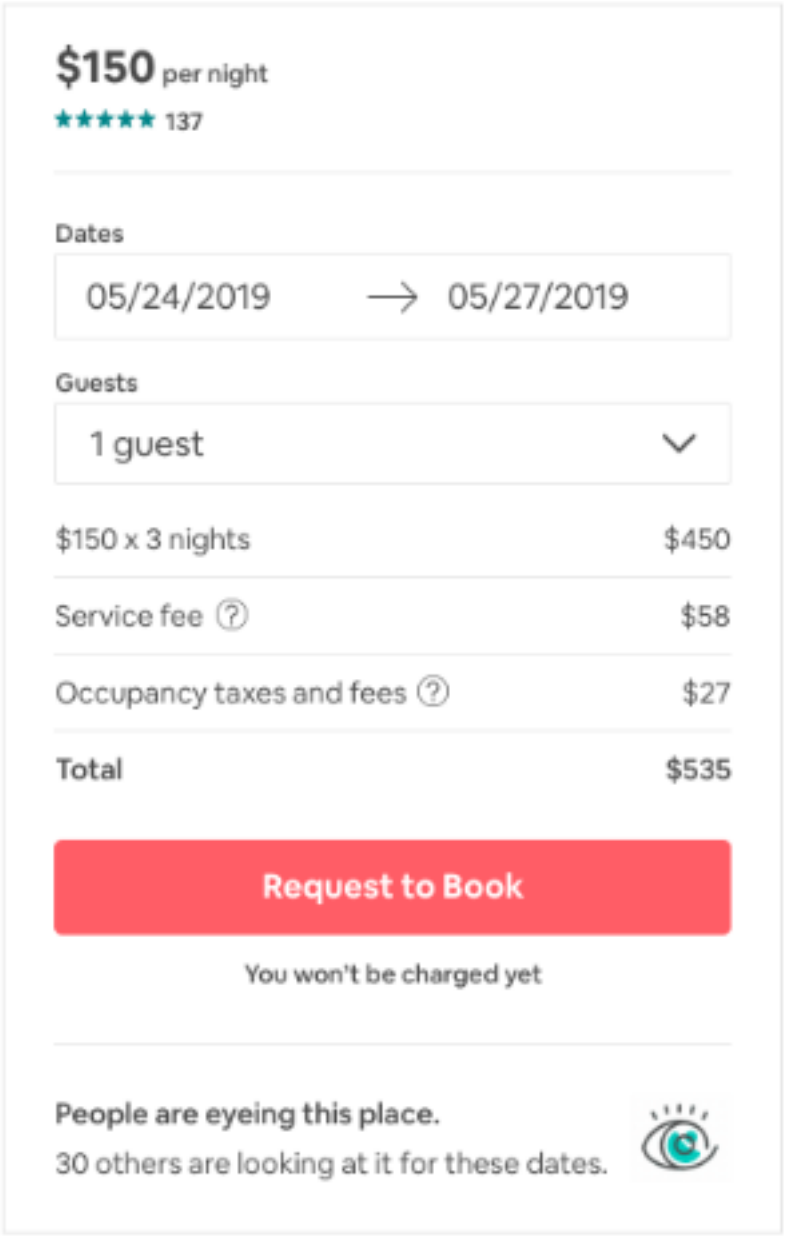


Figure 4 This figure depicts the experiment design process. We use listing-level co-occurrence in search (a) in order to learn "demand embeddings" (b). A hierarchical clustering algorithm is then applied to those embeddings in order to generate clusters (c). Clusters are randomly assigned to meta-treatment or meta-control (d); within meta-control, treatment is assigned at the individual-listing level, whereas in meta-treatment, treatment is assigned at the cluster-level (e). We arrange clusters into strata after treatment assignment to facilitate post-stratification (Miratrix et al. 2013).

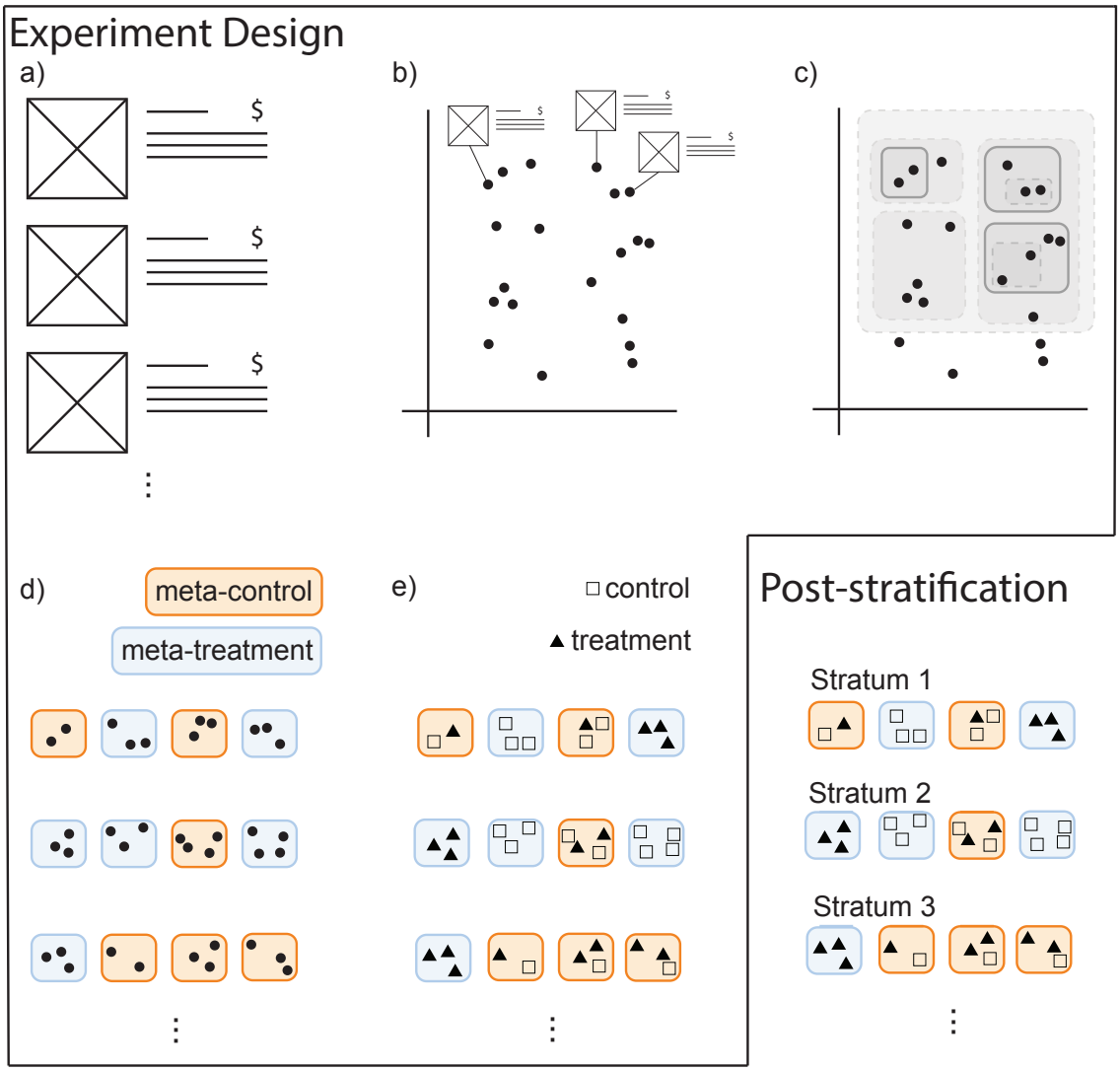


Figure 5 **These maps illustrate clusters generated using the hierarchical clustering scheme described in this paper. Image from Srinivasan (2018).**

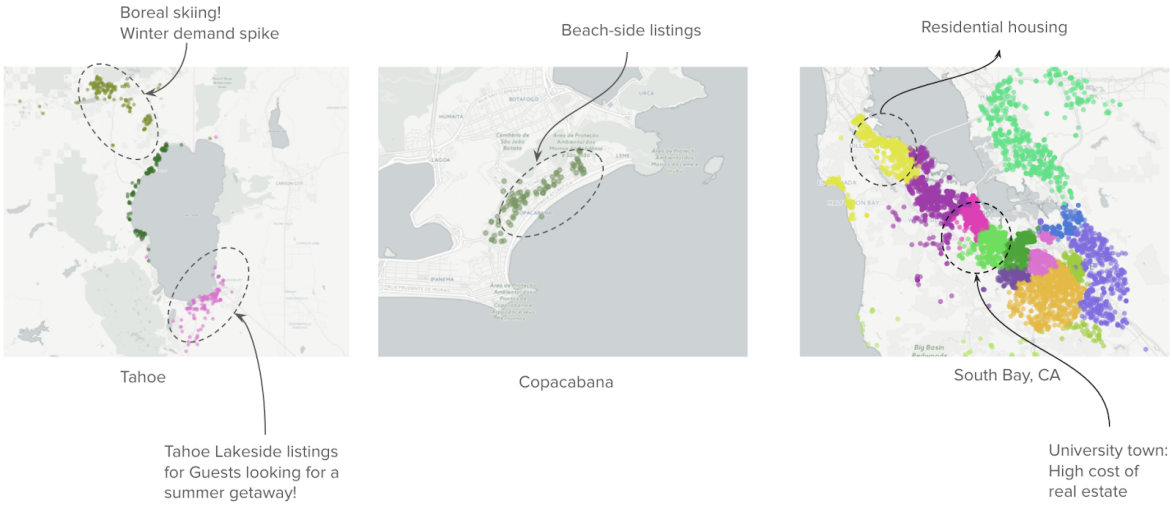


Figure 6 **Coefficient estimates for the joint analysis of the fee meta-experiment. Error bars represent 95% confidence intervals. The dotted blue line correponds to a treatment effect of 0 bookings per listing. The red shaded area corresponds to values that are below the MDE (80% power, 95% confidence).**

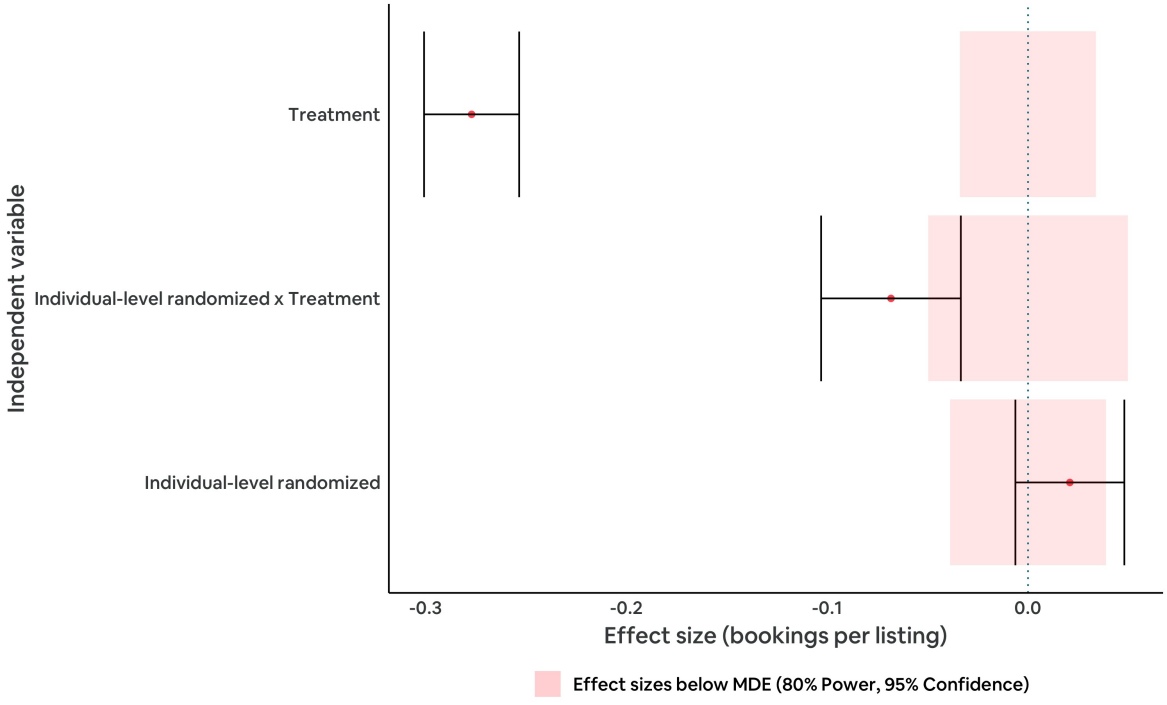
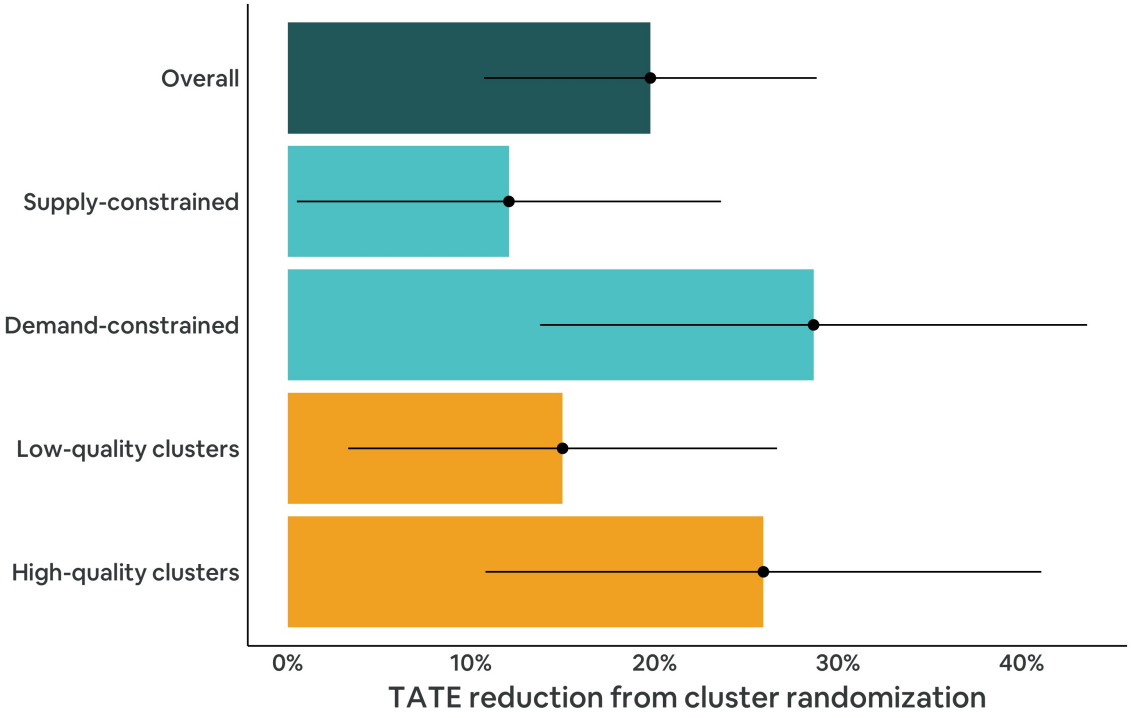


Figure 7 **This graph visualizes the reduction in interference bias from cluster randomization that we estimate across different samples: overall, listings in supply-constrained geographies, listings in demand-constrained geographies, listings in geographies with low-quality clusters, and listings in high-quality clusters.**



8. Tables

Table 1 **This table tests for statistically significant differences in pre-treatment outcomes between treatment and control in the individual-level randomized meta-treatment arm, treatment and control in the cluster-randomized meta-treatment arm, and meta-treatment and meta-control. Each comparison uses a two-sided t-test. Analysis is conducted at the individual-level within the meta-control arm, and at the cluster-level within the meta-treatment arm and when comparing the two meta-treatment arms.**

	Individual-randomized			Cluster-randomized			Meta-experiment		
	Control	Treatment	p-value	Control	Treatment	p-value	Meta-control	Meta-treatment	p-value
Pre-treatment statistics									
Bookings	11.864 (26.275)	11.882 (26.174)	0.78	11.760 (10.559)	11.572 (10.256)	0.49	11.790 (10.664)	11.666 (10.408)	0.65
Nights Booked	44.984 (101.570)	44.953 (102.677)	0.90	43.288 (34.339)	42.497 (33.646)	0.37	43.195 (34.517)	42.893 (33.994)	0.73
Gross Guest Spend	5,920.370 (15,751.420)	5,934.694 (15,824.250)	0.72	5,554.392 (6,764.090)	5,399.833 (6,412.172)	0.37	5,587.642 (6,953.921)	5,477.087 (6,590.321)	0.53
$N_{\text{individuals}}$	323,734	323,643							
N_{clusters}				2,979	2,981		1,987	5,960	

Table 2 This table reports the TATE results obtained by analyzing the two meta-treatment arms separately. Individual-level randomized results are found in Column (1), and cluster randomized results are found in Column (2).

	Dependent variable: Bookings	
	Individual-level randomized	Cluster randomized
	(1)	(2)
Treatment	-0.345^{***}	-0.277^{***}
	(0.013)	(0.012)
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 gross guest spend	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)
Pre-treatment nights available	0.002^{***}	0.001^{***}
	(0.000)	(0.000)
Pre-treatment searches/night	0.267^{***}	0.033^{**}
	(0.027)	(0.015)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	No	Yes
R^2	0.408	0.405
Adjusted R^2	0.407	0.405

*p<0.1; **p<0.05; ***p<0.01

Table 3 This table summarizes the meta-experiment results for number of bookings. Column (1) presents the overall results. Columns (2) and (3) explore heterogeneity with respect to supply/demand-constrainedness. Columns (4) and (5) explore heterogeneity with respect to to cluster quality.

	Dependent variable:				
	Bookings				
	Overall	Supply constrained	Demand constrained	Low-quality clusters	High-quality clusters
	(1)	(2)	(3)	(4)	(5)
Treatment	-0.277^{***}	-0.433^{***}	-0.140^{***}	-0.360^{***}	-0.196^{***}
	(0.012)	(0.022)	(0.011)	(0.019)	(0.016)
Individual-level Randomized	0.021	0.019	0.013	0.021	0.015
	(0.014)	(0.025)	(0.014)	(0.022)	(0.018)
Individual-level Randomized × Treatment	-0.068^{***}	-0.059^{*}	-0.056^{***}	-0.063^{**}	-0.069^{***}
	(0.018)	(0.031)	(0.018)	(0.027)	(0.023)
Pre-treatment bookings	0.175^{***}	0.174^{***}	0.175^{***}	0.172^{***}	0.178^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Pre-treatment nights booked	-0.003^{***}	-0.003^{***}	-0.003^{***}	-0.003^{***}	-0.003^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Pre-treatment gross guest spend	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Pre-treatment nights available	0.001^{***}	0.003^{***}	0.000^{***}	0.002^{***}	0.001^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Pre-treatment searches/night	0.050^{**}	0.021^{**}	0.775^{***}	0.203^{***}	0.028^{**}
	(0.020)	(0.010)	(0.062)	(0.024)	(0.013)
Interference bias estimate	19.76%	12.05%	28.65%	14.98%	25.92%
	(± 9.06%)	(± 11.55%)	(± 14.91%)	(± 11.69%)	(± 15.14%)
Stratum F.E.	Yes	Yes	Yes	Yes	Yes
Robust s.e.	Yes	Yes	Yes	Yes	Yes
Semi-clustered s.e.	Yes	Yes	Yes	Yes	Yes
R^2	0.405	0.404	0.365	0.408	0.402
Adjusted R^2	0.405	0.404	0.364	0.407	0.402

*p<0.1; **p<0.05; ***p<0.01

References

Airbnb (2019) Airbnb news: About us. URL `https://news.airbnb.com/about-us/`.

Aral S, Nicolaides C (2017) Exercise contagion in a global social network. *Nature communications* 8(1):1–8.

Ariel B, Sutherland A, Sherman LW (2019) Preventing treatment spillover contamination in criminological field experiments: the case of body-worn police cameras. *Journal of Experimental Criminology* 15(4):569–591.

Aronow PM (2012) A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research* 41(1):3–16.

Aronow PM, Samii C (2017) Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* 11(4):1912–1947.

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

Azevedo EM, Deng A, Montiel Olea JL, Rao J, Weyl EG (2020) A/b testing with fat tails. *Journal of Political Economy* 128(12):4614–000.

Baird S, Bohren JA, McIntosh C, Özler B (2018) Optimal design of experiments in the presence of interference. *Review of Economics and Statistics* 100(5):844–860.

Basse G, Feller A (2018) Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association* 113(521):41–55.

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society* 841–890.

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.

Bojinov I, Simchi-Levi D, Zhao J (2022) Design and analysis of switchback experiments. *Management Science* .

Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–298.

Bowers J, Fredrickson MM, Panagopoulos C (2013) Reasoning about interference between units: A general framework. *Political Analysis* 21(1):97–124.

Bright I, Delarue A, Lobel I (2022) Reducing marketplace interference bias via shadow prices. *arXiv preprint arXiv:2205.02274* .

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

Choi H, Mela CF (2019) Monetizing online marketplaces. *Marketing Science* 38(6):948–972.

Cox DR (1958) Planning of experiments. .

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).

Feit EM, Berman R (2019) Test & roll: Profit-maximizing a/b tests. *Marketing Science* 38(6):1038–1058.

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* .

Gerber AS, Green DP (2012) *Field experiments: Design, analysis, and interpretation* (WW Norton).

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 D, Zhao M, Benzell SG, Cao CY, Rahimian MA, Yang J, Allen J, Collis A, Moehring A, Sowrirajan T, et al. (2020) Interdependence and the cost of uncoordinated responses to covid-19. *Proceedings of the National Academy of Sciences* 117(33):19837–19843.

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

Hudgens MG, Halloran ME (2008) Toward causal inference with interference. *Journal of the American Statistical Association* 103(482):832–842.

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

Imai K, Tingley D, Yamamoto T (2013) Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176(1):5–51.

Johari R, Li H, Liskovich I, Weintraub GY (2022) Experimental design in two-sided platforms: An analysis of bias. *Management Science* .

Kahle D, Wickham H (2013) ggmap: Spatial visualization with ggplot2. *The R Journal* 5(1):144–161, URL `http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf`.

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

Li H, Zhao G, Johari R, Weintraub GY (2022) Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. *Proceedings of the ACM Web Conference 2022*, 182–192.

Liu L, Hudgens MG (2014) Large sample randomization inference of causal effects in the presence of interference. *Journal of the american statistical association* 109(505):288–301.

Liu M, Mao J, Kang K (2021) Trustworthy and powerful online marketplace experimentation with budget-split design. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 3319–3329.

Mahalanobis PC (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)* 2:49–55.

Manski CF (2000) Economic analysis of social interactions. *Journal of economic perspectives* 14(3):115–136.

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.

Miratrix LW, Sekhon JS, Yu B (2013) Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(2):369–396.

Moffitt RA, et al. (2001) Policy interventions, low-level equilibria, and social interactions. *Social dynamics* 4(45-82):6–17.

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

Nevo A (2000) A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of economics & management strategy* 9(4):513–548.

Pouget-Abadie J, Mirrokni V, Parkes DC, Airoldi EM (2018) Optimizing cluster-based randomized experiments under monotonicity. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2090–2099.

Rolnick D, Aydin K, Pouget-Abadie J, Kamali S, Mirrokni V, Najmi A (2019) Randomized experimental design via geographic clustering. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2745–2753.

Rosenbaum PR (2007) Interference between units in randomized experiments. *Journal of the american statistical association* 102(477):191–200.

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).

Sävje F, Aronow P, Hudgens M (2021) Average treatment effects in the presence of unknown interference. *Annals of statistics* 49(2):673.

Sinclair B, McConnell M, Green DP (2012) Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science* 56(4):1055–1069.

Slee T (2015) Airbnb data collection: Methodology and accuracy. URL `http://tomslee.net/airbnb-data-collection-methodology-and-accuracy`.

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

Tchetgen EJT, VanderWeele TJ (2012) On causal inference in the presence of interference. *Statistical methods in medical research* 21(1):55–75.

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: The impact of interference bias on platform profit

In this appendix, we use a toy model to quantify the potential profit loss associated with interference bias.[25] In the model, a firm chooses a price to maximize profits given a fixed demand curve. The demand function depends on the output elasticity with respect to price, which is ex-ante unknown by the firm. Pricing experiments like the one presented in this paper are one tool available to firms in order to pin down the demand elasticity of consumers. If the firm estimates the wrong elasticity due to interference bias, then the firm's optimization procedure will lead to suboptimal profit.

To be more concrete: Define P as price and Q as quantity. In this context it is important to differentiate the actual demand elasticity from the estimated one, which can suffer from interference bias. Denote the latter as the observed elasticity (η'). Assume the demand function is iso-elastic, $Q = P^{-\eta}$, and the cost function is linear with a slope 1 for simplicity: Q. The profit is defined as a function of P: $\pi(P) = PQ - Q$. The firm equates marginal cost and marginal benefit based on their assessment of demand elasticity η', which does not align with η in the presence of interference. Therefore the price chosen is such that:

$$(1-\eta')P^{\eta'} = \eta P^{-\eta'-1} \Rightarrow P = \left(\frac{\eta'}{\eta'-1}\right).$$

Based on this choice of price, the quantity is defined based on the real (unobserved) elasticity η that drives demand,

$$Q = \left(\frac{\eta'}{\eta'-1}\right)^{-\eta}.$$

Therefore, the firm's profit is given by,

$$\pi(\eta'|\eta) = \left(\frac{\eta'}{\eta'-1}\right)^{1-\eta} - \left(\frac{\eta'}{\eta'-1}\right)^{-\eta}. \tag{11}$$

We define b as the elasticity bias ($b = \frac{\eta'-\eta}{\eta}$). We can restate the profit as a function of the true elasticity and the bias:

$$\pi(b,\eta) = \left(\frac{\eta(b+1)}{\eta(b+1)-1}\right)^{1-\eta} - \left(\frac{\eta(b+1)}{\eta(b+1)-1}\right)^{-\eta}. \tag{12}$$

In our setting, the parameter η could be estimated using a cluster-randomized experiment, the parameter η' could be estimated using an individual-level randomized experiment, and the bias b could be estimated by taking the difference of the two. Finally, to assess the loss due to interference bias we could calculate:

[25] The inclusion of this toy model in our paper does not imply or suggest that Airbnb sets guest fees or make any other platform design choices to maximize profits.

$$\Delta(b|\eta) = \frac{\pi(b|\eta) - \pi(0|\eta)}{\pi(0|\eta)}. \tag{13}$$

Due to our NDA, we cannot reveal Airbnb's estimated demand elasticity. However, Figure G.1 shows the profit loss of a hypothetical firm for a particular estimated demand elasticity given different levels of bias.[26] We can see that the profit loss is increasing in the size of the bias, and occurs both for under- and over-estimates of the actual demand elasticity (η). As expected, profit is maximized when bias is zero.

[26]Recall that we estimate that interference bias accounts for at least 19.76% of naive TATE estimates for pricing applications on Airbnb.

Appendix B: Interference Simulation

In this section, we design a simulation of booking behavior for one calendar night in a single Airbnb geography (Miami). We use this simulation framework to determine whether individual-level randomization yields biased TATE estimates, and perform a preliminary investigation into the viability of cluster randomization at reducing that bias.

B.1. Data & Network Construction

Our simulation framework is built on top of a dataset scraped by Slee (2015), which describes all of the Airbnb listings in and around Miami as of February 13, 2016. This dataset details the room type, number of reviews, average "overall satisfaction" rating, guest capacity, number of bedrooms, number of bathrooms, price per night (USD), minimum length of stay, latitude, and longitude of 8,855 Airbnb listings. Figure G.2 depicts the geospatial distribution of the listings by room type, and Table H.1 provides information about the distribution of listing-level covariates across the sample of Airbnb listings.

Before using the dataset for our analyses, we impute missing values in a number of fields: missing guest capacity, bedroom, and bathroom values are imputed using the modal value for each variable. Minimum length of stay values are capped at 30, and missing minimum length of stay values are imputed using the modal value for minimum length of stay. Missing overall satisfaction values are imputed using the mean value of non-empty entries. We also assign each listing j in our dataset an unobservable quality component,

$$\xi_j \sim N(0,1), \tag{14}$$

which is kept constant across all simulations. This unobserved quality component is observable to searchers, but not observable to the search algorithm or the platform. Depending on the quality of a given platform's data, factors that contribute to a listing's unobservable quality might include the quality of its photos, the responsiveness of the seller, and/or the text content of the listing's reviews

We proceed to build a "product network" for listings in this dataset. Each listing in the dataset constitutes a node in the network, and an edge between two listings implies that the listings are likely to substitute for one another when searchers are making purchase decisions. We generate an edge between two listings when the following three criteria are satisfied:[27]

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

[27] One could imagine using a subset of these criteria (e.g., all listings within 1 mile of each other are substitutes), or a totally unrelated criteria (e.g., listings must have co-occurred in search more than x times). For instance, in the main body of this paper and in Srinivasan (2018), items in an online marketplace are clustered based on how often they co-occur in search results.

Using the edge heuristic described above, we produce a network that has 1,538,637 edges, and a clustering coefficient of 0.74. The average degree of nodes in the network is 173.76.

In order to simulate cluster randomized experiments, we need to divide this network into clusters. We do so using the Louvain clustering algorithm (Blondel et al. 2008). Louvain clustering attempts to maximize modularity, which is defined as

$$Q = \frac{1}{2E} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2E} \right) \mathbb{1}(C_i = C_j), \tag{15}$$

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

As noted in the main body of this paper, cluster randomization can increase the variance of TATE estimates. In order to counteract this increase in variance, our simulated cluster randomization experiments use block random assignment, with blocks of size $b = 2$, to assign cluster-level treatment. To arrange clusters into pairs that will be used in that block random assignment procedure, we first calculate the average number of reviews, the average overall satisfaction score, the average number of beds, the average number of bathrooms, the average minimum stay, the average latitude, the average longitude, the percentage of private room listings, and the percentage of shared rooms for each cluster. After concatenating these metrics into a vector representing each cluster, we calculate the Mahalanobis distance (Mahalanobis 1936) between every possible pair of clusters, and select pairs of clusters using a greedy algorithm that attempts to minimize the sum of the Mahalanobis distances between each chosen pair.

B.2. Simulation Process

In order to estimate the true TATE under different treatment interventions, as well as the bias and sampling variance of the TATE estimator under different experiment designs and analysis approaches, we create a framework for simulating the Airbnb booking process for one calendar night. Each set of simulated outcomes is generated using the following procedure.

First, a "search algorithm," $\boldsymbol{\delta}$, is drawn, with each element of $\boldsymbol{\delta}$ being generated by first drawing from the uniform distribution over the interval $[0,1]$ and then normalizing so that the sum of the elements of $\boldsymbol{\delta}$ is one, i.e.,

$$\begin{aligned} &\delta_{k0} \sim U[0,1] \text{ for } k = 1,2,3,...,9, \\ &\delta_k = \frac{\delta_{k0}}{\sum_j \delta_{k0}}. \end{aligned} \tag{16}$$

The nine elements of δ correspond to the weight that the algorithm puts on normalized versions of the following listing-level attributes: number of reviews, average satisfaction score, number of bedrooms, number of bathrooms, minimum stay, price, whether the listing is for an entire home/apt, whether the listing is for a private room, and whether a listing is a shared room. The "search algorithm" can then determine a "score" for each listing by taking the inner product of $\boldsymbol{\delta}$ and $\mathbf{x_j}$, the full vector of the listing i's centered and scaled attributes, i.e.,

$$\text{Search Score}_j = \delta \cdot \mathbf{x}_j. \tag{17}$$

Conditional on being issued a query by a searcher with certain geographic or attribute constraints, the algorithm will return to the searcher the ten unbooked listings with the highest search score. In cases where ten listings meeting the searcher's criteria are not available, the algorithm will return all of the listings satisfying the searcher's criteria. This allows for the possibility that the algorithm returns no listings if there are none that satisfy the searcher's requirements.

Then, $n_{searchers}$ "searchers" sequentially arrive at Airbnb and look for an available listing in our marketplace, i.e., Miami. Each searcher randomly draws a region of interest in latitude/longitude space. The locations of the box edges are drawn with uniform probability from the interval spanning from the .25th percentile of the latitudes (longitudes) belonging to listings in the geography to the 99.75th percentile of latitudes (longitudes) belonging to listings in the geography.[28] The searcher also draws a minimum guest capacity from a uniform distribution over {1,2,3,4}. The geographic boundaries and minimum guest capacity constitute the searcher's "query," and only listings that satisfy the searcher's geographic and capacity requirements will be returned by the search algorithm.

Searcher i's utility from booking listing j is given by the following equation, which is chosen so that our simulation framework is comparable to models used in the demand estimation literature (e.g., Berry et al. (1995) and Nevo (2000)):

$$u_{ij} = \alpha_i(y_i - p_j) + \tilde{\mathbf{x}}_\mathbf{j}\boldsymbol{\beta}_i + \xi_j + \epsilon_{ij}, \tag{18}$$

where $\tilde{\mathbf{x}}_\mathbf{j}$ is the vector of listing j's attributes *besides* price, and

$$\begin{aligned} y_i &\sim N(0,1) \\ \alpha_i &\sim N(0,1) \\ \beta_{ik} &\sim N(0,1)\,\forall\, k \\ \epsilon_{ij} &\sim f(x) = e^{-x}e^{e^{-x}} \text{ (the Type I extreme-value distribution).} \end{aligned} \tag{19}$$

Searcher i uses the above utility function to determine which of the up to 10 listings provided by the search algorithm they would like to book. If none of the listings have a utility greater than 0

[28] This is done to account for the potential that there are listings in our dataset that are geographic outliers.

(representing the outside option), or if the search algorithm does not return any listings meeting the searcher's query parameters, the searcher chooses not to book and exits the marketplace. Otherwise, the searcher "books" the listing that provides the highest utility to them. After this point, that listing cannot appear in future searchers' consideration sets.

Although this simulation framework simplifies the marketplace dynamics of a platform like Airbnb, we believe it can still provide insight into the degree to which interference may bias TATE estimates in online marketplace experiments, and can help determine the extent to which cluster randomization reduces that bias. We conduct simulations of marketplace activity both under marketplace-wide policy regimes (i.e., 100% treatment and 100% control), as well as under different experiment designs. We then compare the ground truth TATEs generated by contrasting outcomes under marketplace-wide policy changes to the TATE estimates produced by different experiment designs, and calculate the bias and root mean square error (RMSE) of the TATE estimates produced under different approaches to experiment design. In each of our simulations of marketplace activity, we are interested in two different outcomes. The first is whether or not a listing was booked. The second is the amount of revenue earned by a listing. We also consider two different types of treatment intervention. The first is a price reduction of .75 standard deviations for treated listings. The second is an increase of .75 standard deviations in the unobserved quality of listings.

B.3. Simulating Ground Truth

We first use our simulation framework to simulate the distribution of marketplace-level average outcomes in the case in which 100% of listings receive the treatment, and the case in which 100% of listings receive the control. For the control, as well as both the price reduction treatment and the unobserved listing quality treatment, we conduct 500 simulations of one night of booking activity in which 1,000 searchers visit Airbnb. Figure G.3 compares the sampling distributions of the rate of listings being booked and the average listing revenue under all three conditions.

A two-sided t-test between the distribution of booking rates under the control and the distribution of booking rates under the price reduction treatment yields a t-statistic of $t = 17.27$ ($p \leq 2.2 \times 10^{-16}$), with an average TATE of 0.002, whereas a two-sided t-test between the distribution of average listing revenue under the control and the distribution of average listing revenue under the price reduction treatment yields a t-statistic of $t = 1.63$ ($p = 0.10$), i.e., at the 95% level, we are unable to reject the null hypothesis that the average TATE is equal to zero. This pair of results is somewhat intuitive: when sellers lower prices, the rate at which listings are booked increases, because a greater share of listings dominate the outside option. However, that increase in booking rate does not translate into an increase in revenue, since those listings are being booked at a lower price.

A two-sided t-test between the distribution of booking rates under the control and the distribution of booking rates under the unobserved listing quality treatment yields a t-statistics of $t = 21.63$ ($p \leq 2.2 \times 10^{-16}$), with an average TATE of 0.003, whereas a two-sided t test between the distribution of

average listing revenue under the control and the distribution of average listing revenue under the unobserved listing quality change treatment yields a t-statistic of 2.17 ($p = 0.03$), with an average TATE of 0.612. This pair of results is also intuitive: when the unobservable quality of listings increases, the rate at which listings are booked increases, again because a greater share of listings dominate the outside option. Because this increase in booking rate does not come hand in hand with a reduction in price, this increase in booking rate translates into an increase in revenue.

B.4. Measuring bias and RMSE

Having simulated the distribution of marketplace-level outcomes under both 100% treatment and 100% control for both our price reduction treatment and our unobservable listing quality treatment, we can now use our simulation framework to estimate the bias and RMSE of different experiment designs for both treatments. We first use our framework to simulate 500 individual-level randomized experiments, in which treatment effects are estimated using a difference in means treatment effect estimator, and then use the simulation framework to simulate 500 blocked cluster randomized experiments. Under this design, we calculate the difference in means estimator and also estimate the treatment effect using a linear regression with clustered standard errors.

Table H.2 shows the bias and RMSE of each experiment design for the booking outcome, under both the price reduction treatment and the unobserved listing quality treatments. Table H.3 shows the same information for the listing revenue outcome under both treatments. Relative to the difference in means estimator under the individual-level randomized experiment, we find that the difference in means estimator under blocked cluster randomization reduced bias by as much as 64.5%, across both metrics and both types of treatment. However, this came at the cost of increasing RMSE by as much as 204%. In other words, although the TATE estimates are on average closer to the ground truth TATE, the variance of the distribution of those estimates is much higher, i.e., statistical power is much lower.

B.5. Statistical Inference

In addition to measuring the true bias and RMSE of different experiment designs, we also assess the coverage probability associated with the 95% confidence interval that each of these approaches yields. For our difference in means estimators, we calculate the variance of the treatment effect estimate using the following expression,

$$\hat{\sigma_\tau^2} = \sigma^2(Y_{iT}) + \sigma^2(Y_{iC}), \tag{20}$$

where $\sigma^2(Y_{iT})$ and $\sigma^2(Y_{iC})$ are the variance of outcomes in the treatment group and control group, respectively. We also calculate the variance of the blocked cluster randomized TATE estimate when analyzed with a linear model that clusters standard errors at the level of the cluster. This approach to analyzing the data better takes into account the design of the experiment, and should lead to 95% confidence intervals with a coverage probability closer to the nominal level.

The coverage probabilities corresponding to our 95% confidence intervals are found in the rightmost columns of Tables H.2 and H.3. We find that the coverage probability of the difference in means estimator when used with the individual-level randomized design is below the nominal 95% coverage in all cases, and can be as low as 6%. The blocked cluster randomized design, when used in conjunction with the difference in means estimator, tends to move the coverage probability closer to the nominal coverage probability for the price reduction treatment, but negatively impacts the coverage probability for the unobserved quality change treatment. Regression analysis of the blocked GCR design with clustered standard errors produces coverage probabilities that are greater than the nominal 95% coverage probability, ranging from 95% to 100%.

Appendix C: Proof of Proposition 1

Proof. Under independent randomization,

$$\begin{aligned}\hat{\tau}_{ind} &= \frac{1}{N}\sum_{i=1}^{N}(E[Y_i|z_i=1]-E[Y_i|z_i=0]) \\ &= \frac{1}{N}\sum_{i=1}^{N}([B_{ii}+p\sum_{j\neq i}B_{ij}]-[p\sum_{j\neq 1}B_{ij}]) \\ &= \sum_{i=1}^{N}B_{ii}.\end{aligned} \tag{21}$$

Under cluster randomization,

$$\begin{aligned}\hat{\tau}_{cr} &= \frac{1}{N}\sum_{i=1}^{N}\left(E\left[Y_i|z_i=1\right]-E[Y_i|z_i=0]\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}([B_{ii}+\sum_{\substack{i\neq j\\ C(j)=C(i)}}B_{ij}+p\sum_{\substack{j\neq i\\ C(j)\neq C(i)}}B_{ij}]-p\sum_{\substack{j\neq 1\\ C(j)\neq C(i)}}B_{ij}) \\ &= \frac{1}{N}\sum_{i=1}^{N}([B_{ii}+\sum_{\substack{i\neq j\\ C(j)=C(i)}}B_{ij}]) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}B_{ij}\mathbb{1}[C(i)=C(j)].\end{aligned} \tag{22}$$

If the bias of the treatment effect under graph cluster randomization is less than the bias under independent randomization, then $|\tau(1,0)-\hat{\tau}_{cr}| \leq |\tau(1,0)-\hat{\tau}_{ind}|$, which implies that

$$\frac{1}{N}\left|\sum_{i=1}^{N}\sum_{j=1}^{N}B_{ij}-\sum_{i=1}^{N}\sum_{j=1}^{N}B_{ij}\mathbb{1}\left(C(i)=C(j)\right)\right| \leq \frac{1}{N}\left|\sum_{i=1}^{N}\sum_{j=1}^{N}B_{ij}-\sum_{i=1}^{N}B_{ii}\right|. \tag{23}$$

This expression can be simplified to

$$\left|\sum_{i=1}^{N}\sum_{j=1}^{N}B_{ij}\mathbb{1}\left(C(i)\neq C(j)\right)\right| \leq \left|\sum_{i=1}^{N}\sum_{j=1}^{N}B_{ij}\mathbb{1}(i\neq j)\right|. \tag{24}$$

Since the set of sellers not equal to i is a superset of the sellers not in the same cluster as i, and since all of the off-diagonal elements of $\boldsymbol{B}$ have the same sign, this will always hold true. □

Appendix D: Determining Cluster Size

Previously attempted pricing meta-experiments at Airbnb had used clusters of minimum size 250, so this was considered the "status quo" cluster size. We also decided based on statistical power considerations that a cluster size threshold of 1,000 was the maximum feasible threshold. Given these facts, the choice of cluster size threshold became a direct comparison between a minimum size of 250 and a minimum size of 1,000. In choosing a cluster size threshold, the fundamental trade-off is between statistical power and capturing Airbnb demand. While smaller clusters yield more statistical power (since there are 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. As a consequence, cluster quality and bias reduction will both be lower. On the other hand, larger clusters will provide less statistical power, but will do a better job of capturing demand and reducing bias. Power analysis suggested that without taking into account differences in cluster quality, our fee meta-experiment would have a minimum detectable effect (MDE) for interference bias that was 1.17 times as large if clusters of minimum size 1,000 were used as opposed to clusters of size 250. In order to determine whether this degradation in "ideal" MDE was worthwhile, we needed to measure differences in the extent to which the two sets of clusters captured demand on the platform.[29,30]

In order to make a principled decision between the two different minimum cluster sizes, we assumed that the "ideal" MDEs obtained via our power calculations would be reduced due to poor demand capture according to the relationship below:

$$MDE_{actual} = \frac{MDE_{ideal}}{\text{Demand capture}}. \tag{25}$$

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 if

$$\frac{\text{Demand capture}_{1,000}}{\text{Demand capture}_{250}} > \frac{MDE_{ideal_{250}}}{MDE_{ideal_{1,000}}} \rightarrow \frac{\text{Demand capture}_{1,000}}{\text{Demand capture}_{250}} > 1.17 \tag{26}$$

Table H.4 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

[29]The analysis we describe below was originally conducted using data and clusters from February 2019, however, we present analyses using clusters generated on January 5, 2020, PDP views occurring between January 5, 2020 and January 12, 2020, and bookings occurring between January 5, 2020 and January 26, 2020. The results we report and the corresponding conclusions are qualitatively similar to those obtained using 2019 data.

[30]The meta-experiment design process occurred prior to the creation of the cluster quality metric introduced in Definition 1. Moving forward, we would recommend others use the cluster quality metric found elsewhere in this paper, as opposed to any of the demand capture metrics described below.

measures calculated across one week of PDP views: the average share of PDPs belonging to a given cluster, the average user-level PDP 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 1.17. Based on this calculation, we determined that clusters with a size threshold of 1,000 listings were preferable to those with a size threshold of 250.

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

In this appendix, we present the results of our analyses for two additional outcomes: nights booked per listing and gross guest spend per listing. Qualitatively, our results for nights booked per listing and gross guest spend per listing are extremely similar to our results for bookings per listing.

Table H.5 shows the estimated effect of the fee treatment in both the individual-level randomized meta-treatment arm and the cluster randomized meta-treatment arm on both nights booked per listing and gross guest spend per listing. We estimate in the individual-level randomized meta-treatment arm that the treatment led to a statistically significant loss of 0.308 nights booked per listing and \$29.92 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 0.257 nights booked per listing and \$26.56 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. Our results are displayed in Table H.6 and Figure G.4. We find statistically significant evidence of interference bias in the individual-level randomized TATE estimate for nights booked per listing, but do not find statistically significant evidence of interference bias in the individual-level randomized TATE estimate for gross guest spend per listing. Our point estimates suggest that interference accounts for 15.26% of the Bernoulli TATE estimate for nights booked per listing (stat sig.) and 9.98% of the Bernoulli TATE estimate for gross guest spend per listing (not stat. sig).

Appendix F: Estimating cluster quality using browsing data

In this appendix, we describe the process used to estimate a geography-level version of the "cluster quality" metric found in Definition 1 for our clusters. In Section 5.4, we use this metric to estimate heterogeneity in the amount of interference bias with respect to geography-level cluster quality.

Because this paper focuses on pricing-related interventions, the "true" interference matrix $\boldsymbol{B}$ that we wish to construct in order to assess cluster quality is likely the matrix of listing cross-price elasticities. Unfortunately, the full set of cross-price elasticities on Airbnb is extremely difficult, if not impossible to estimate. However, given that cross-price elasticities are at least partially driven by co-occurrence in searchers' consideration sets, we argue that search or PDP view data can be used to construct an appropriate proxy matrix for a pricing experiment on Airbnb.

In place of constructing a full proxy matrix, we use a procedure similar to the one described by Rolnick et al. (2019) to calculate the quality score $Q_C(\boldsymbol{P})$ for a given set of clusters by "folding" the underlying bipartite graph between searchers and/or PDP viewers and clusters. More specifically, let s_{ik} be the number of search impressions or PDP views by searcher i to listings in cluster C_k. We calculate the normalized folded edge between clusters k and k':

$$F_{kk'} = \sum_i \frac{s_{ik} s_{ik'}}{\sqrt{\sum_i s_{ik} \sum_i s_{ik'}} \sqrt{\sum_k s_{ik} \sum_k s_{ik}}}. \tag{27}$$

It follows that $Q_C(\boldsymbol{P})$, i.e., the total edge weight not captured by the clustering C consisting of M clusters is

$$Q_C(\boldsymbol{P}) = \frac{1}{M} \sum_k (1 - F_{kk}), \tag{28}$$

where the normalization factor of M ensures that the maximum value of $Q_C(\boldsymbol{P})$ is 1. This expression for cluster quality is higher when listings from cluster k tend to co-occur in search or PDP view sessions with other listings in cluster k, and will be maximized when listings in cluster k *only* co-occur in search or PDP view sessions with other listings in cluster k.

For computational tractability, we choose to construct our proxy matrix using PDP views, as opposed to search impressions. However, we expect that the proxy matrices constructed using these two datasets would be extremely similar.

Appendix G: Additional Figures

Figure G.1 **This figure plots the impact of interference bias on firm profits. The figure shows that profit is maximized when the bias does not exist. Losses increase with interference bias, and this is true for positive and negative bias.**

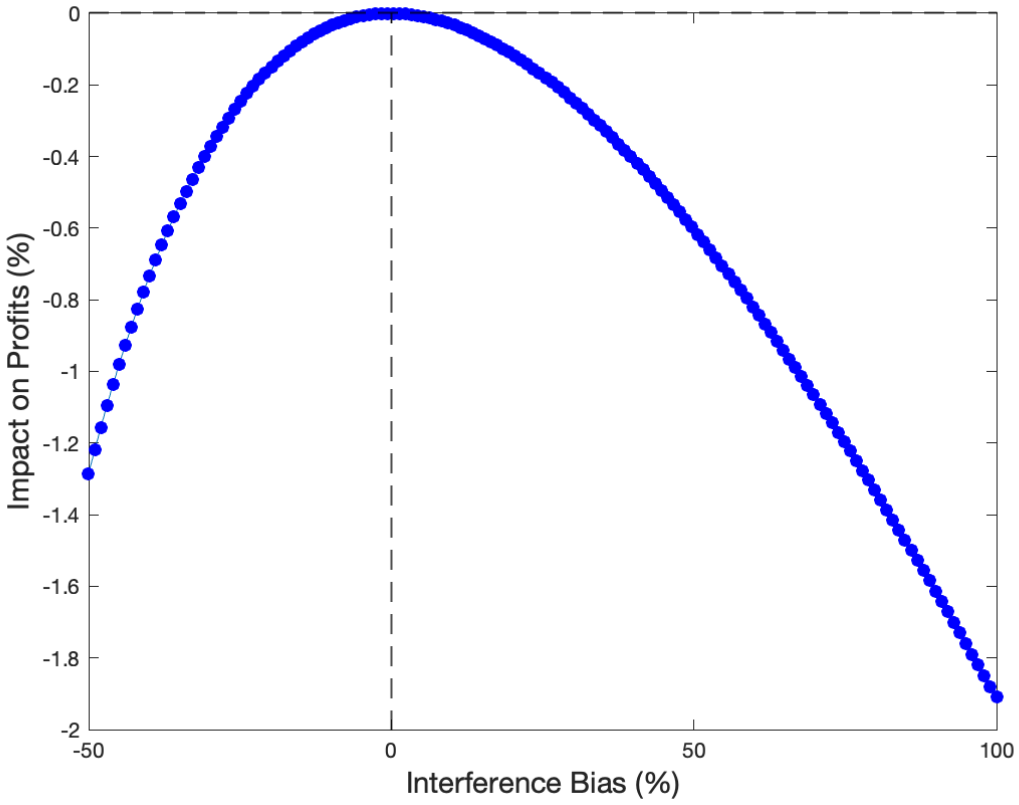


Figure G.2 **The geospatial distribution of Airbnb listings in and around Miami. Color corresponds to listing type. This figure was produced with ggmap (Kahle and Wickham 2013).**

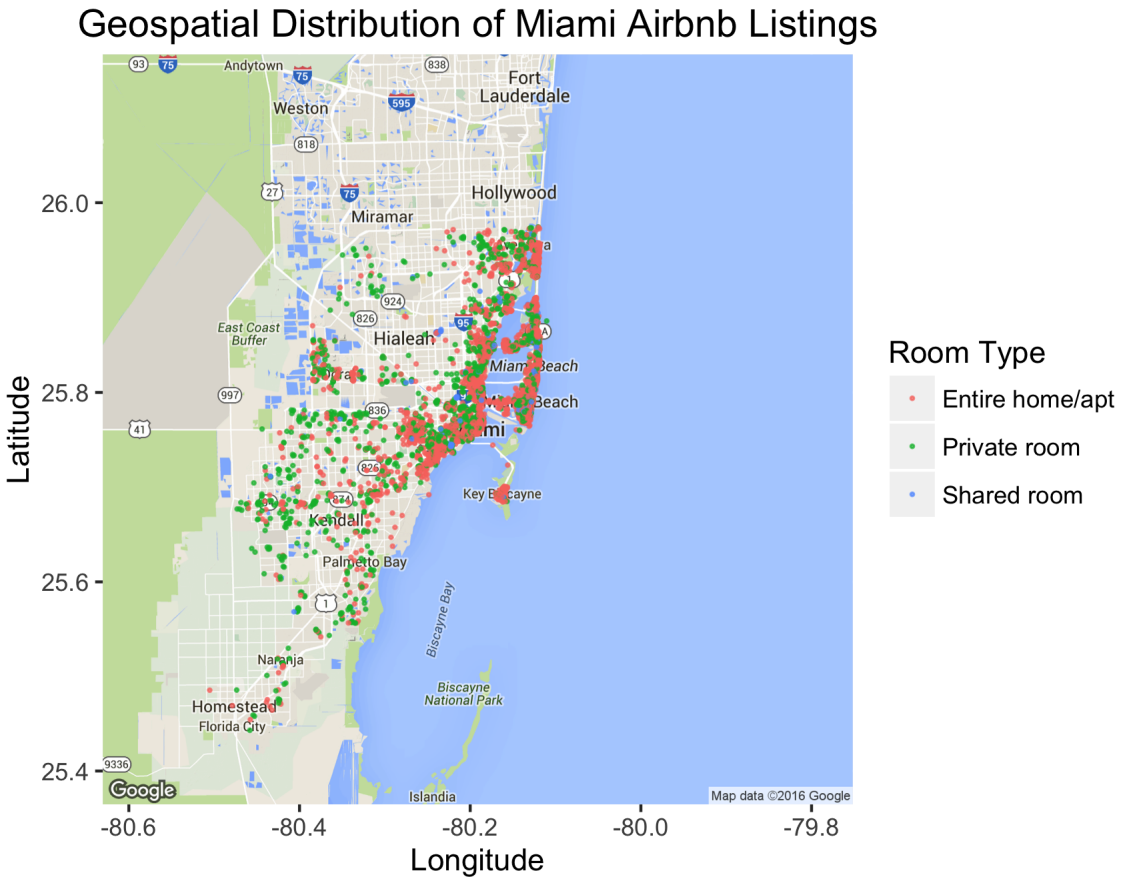
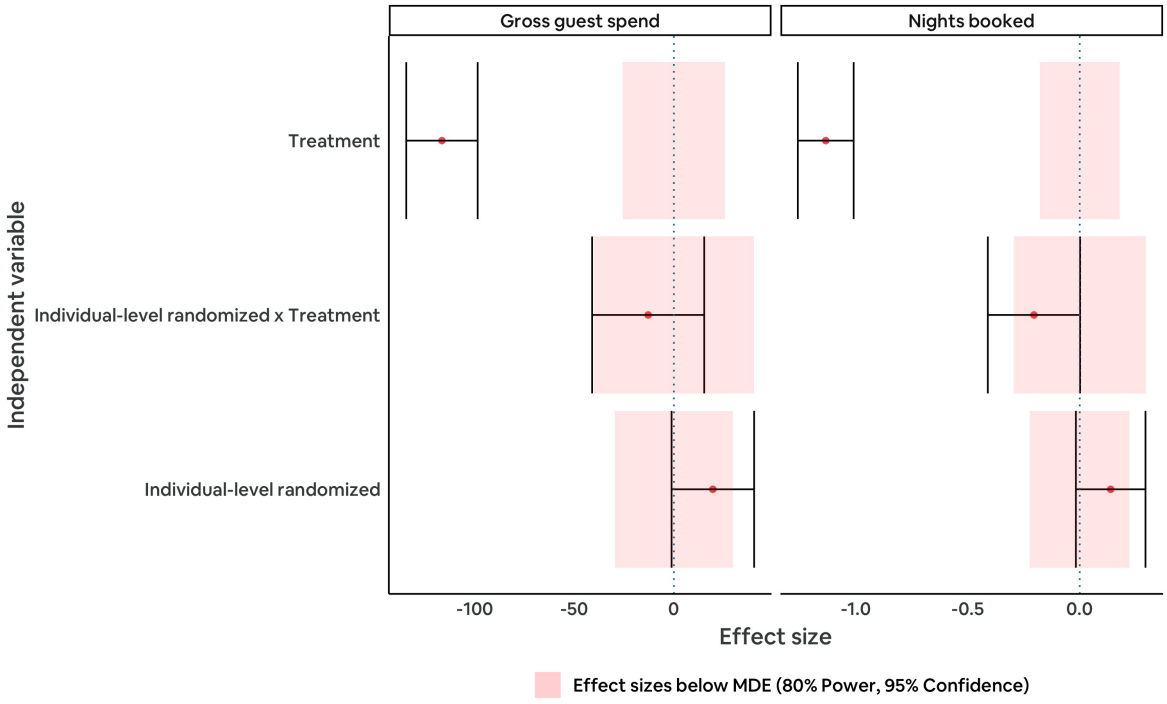


Figure G.3 **Comparison of simulated marketplace-wide average outcomes when either 0% or 100% of listings are assigned treatment. The top row shows distributions when the treatment is the price reduction treatment. The bottom row shows distributions when the treatment is the unobserved listing quality change treatment. The left column shows distributions for the listing booked outcome. The right column shows distributions for the listing revenue outcome.**

Figure G.4 **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).**



Appendix H: Additional Tables

Table H.1 Summary of Airbnb listing covariates for interference simulation

	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Private room	8,855	0.233	0.423	0	0	0	1
Shared room	8,855	0.026	0.158	0	0	0	1
Entire home/apt	8,855	0.742	0.438	0	0	1	1
Reviews	8,855	11.397	22.366	0	0	12	304
Overall satisfaction	6,433	4.588	0.539	1.000	4.500	5.000	5.000
Capacity	6,629	3.060	1.152	1.000	2.000	4.000	8.000
Beds	8,843	1.399	1.028	0.000	1.000	2.000	10.000
Baths	7,922	1.370	0.695	0.000	1.000	2.000	8.000
Price (USD)	8,855	226.016	406.892	15	89	249	10,000
Min Stay	8,418	3.293	9.309	1.000	1.000	3.000	365.000
Lat.	8,855	25.808	0.072	25.443	25.773	25.844	25.974
Lon.	8,855	−80.176	0.070	−80.505	−80.193	−80.129	−80.110

Table H.2 Simulated performance comparison: outcome = bookings

Treatment	Design	Estimator	Bias	RMSE	Coverage
Price Reduction	Individual-level randomization	Difference in means	0.0354	0.0393	6%
Price Reduction	Cluster randomization	Difference in means	0.0248	0.0459	20%
Price Reduction	Cluster randomization	Regression + clustered S.E.	0.0248	0.0459	95%
Unobserved quality	Individual-level randomization	Difference in means	0.0110	0.0125	56%
Unobserved quality	Cluster randomization	Difference in means	0.0039	0.0381	23%
Unobserved quality	Cluster randomization	Regression + clustered S.E.	0.0039	0.0381	99%

Table H.3 Simulated performance comparison: outcome = listing revenue

Treatment	Design	Estimator	Bias	RMSE	Coverage
Price Reduction	Individual-level randomization	Difference in means	6.08	7.26	40%
Price Reduction	Cluster randomization	Difference in means	4.30	9.06	47%
Price Reduction	Cluster randomization	Regression + clustered S.E.	4.30	9.06	97%
Unobserved quality	Individual-level randomization	Difference in means	2.26	3.93	86%
Unobserved quality	Cluster randomization	Difference in means	0.73	7.76	49%
Unobserved quality	Cluster randomization	Regression + clustered S.E.	0.73	7.76	100%

Table H.4 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 H.5 Independent results of the fee meta-experiment (nights booked and gross guest spend)

	Dependent variable:			
	Nights booked		Gross guest spend	
	Individual-level randomized	Cluster randomized	Individual-level randomized	Cluster randomized
	(1)	(2)	(3)	(4)
Treatment	-1.340^{***}	-1.117^{***}	-130.021^{***}	-115.442^{***}
	(0.084)	(0.064)	(11.164)	(9.230)
Pre-treatment bookings	0.293^{***}	0.298^{***}	24.754^{***}	24.742^{***}
	(0.005)	(0.003)	(0.558)	(0.466)
Pre-treatment nights booked	0.035^{***}	0.034^{***}	-4.788^{***}	-4.265^{***}
	(0.002)	(0.001)	(0.230)	(0.154)
Pre-treatment gross guest spend	-0.000^{**}	-0.000^{*}	0.102^{***}	0.098^{***}
	(0.000)	(0.000)	(0.003)	(0.002)
Pre-treatment nights available	0.013^{***}	0.010^{***}	1.563^{***}	1.116^{***}
	(0.002)	(0.001)	(0.210)	(0.090)
Pre-treatment searches/night	1.334^{***}	0.152^{**}	231.887^{***}	29.231^{**}
	(0.160)	(0.068)	(29.159)	(12.818)
Stratum F.E.	Yes	Yes	Yes	Yes
Robust s.e.	Yes	Yes	Yes	Yes
Clustered s.e.	No	Yes	No	Yes
R^2	0.110	0.111	0.167	0.167
Adjusted R^2	0.109	0.110	0.165	0.166

*p<0.1; **p<0.05; ***p<0.01

Table H.6 Results of the fee meta-experiment (nights booked and gross guest spend)

	Dependent variable:	
	Nights booked	Gross guest spend
	(1)	(2)
Treatment	-1.136^{***}	-116.667^{***}
	(0.064)	(9.163)
Individual-level Randomized	0.138^{*}	19.602^{*}
	(0.079)	(10.593)
Individual-level Randomized × Treatment	-0.205^{*}	-12.931
	(0.105)	(14.384)
Pre-treatment bookings	0.297^{***}	24.769^{***}
	(0.003)	(0.376)
Pre-treatment nights booked	0.035^{***}	-4.407^{***}
	(0.001)	(0.129)
Pre-treatment gross guest spend	-0.000^{**}	0.100^{***}
	(0.000)	(0.001)
Pre-treatment nights available	0.011^{***}	1.174^{***}
	(0.001)	(0.082)
Pre-treatment searches/night	0.230^{**}	42.669^{**}
	(0.093)	(17.007)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Semi-clustered s.e.	Yes	Yes
R^2	0.110	0.166
Adjusted R^2	0.110	0.166

*p<0.1; **p<0.05; ***p<0.01

Table H.7 Independent results of the fee meta-experiment (simple specification)

	Dependent variable: Bookings	
	Individual-level randomized	Cluster randomized
	(1)	(2)
Treatment	-0.343^{***}	-0.291^{***}
	(0.017)	(0.058)
Constant	2.578^{***}	2.520^{***}
	(0.013)	(0.043)
Clustered s.e.	No	Yes
R^2	0.001	0.000
Adjusted R^2	0.001	0.000

*p<0.1; **p<0.05; ***p<0.01

Table H.8 Results of the fee meta-experiment (simple specification)

	Dependent variable:		
	Bookings	Nights booked	Gross guest spend
	(1)	(2)	(3)
Treatment	-0.291^{***}	-1.261^{***}	-123.942^{***}
	(0.058)	(0.195)	(36.306)
Individual-level Randomized	0.058	0.182	28.039
	(0.045)	(0.162)	(28.211)
Individual-level Randomized × Treatment	−0.052	−0.080	−3.427
	(0.060)	(0.214)	(38.309)
Constant	2.520^{***}	9.517^{***}	$1{,}215.845^{***}$
	(0.043)	(0.148)	(26.830)
Semi-clustered s.e.	Yes	Yes	Yes
R^2	0.001	0.000	0.000
Adjusted R^2	0.001	0.000	0.000

*p<0.1; **p<0.05; ***p<0.01

Table H.9 Treatment effect heterogeneity for the fee meta-experiment (interacted)

	Dependent variable:	
	Bookings	
	Supply/demand-constrained	Cluster quality
	(1)	(2)
Treatment	-0.092^{***}	-0.349^{***}
	(0.005)	(0.018)
Individual-level Randomized	0.009	0.032
	(0.005)	(0.021)
Individual-level Randomized × Treatment	-0.021^{***}	-0.076^{***}
	(0.007)	(0.027)
Demand-constrained	-0.073^{***}	
	(0.004)	
High-quality cluster		-0.040^{**}
		(0.018)
Pre-treatment bookings	0.175^{***}	0.175^{***}
	(0.001)	(0.001)
Pre-treatment nights booked	-0.003^{***}	-0.003^{***}
	(0.000)	(0.000)
Pre-treatment gross guest spend	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)
Pre-treatment nights available	0.000^{***}	0.001^{***}
	(0.000)	(0.000)
Pre-treatment searches/night	0.005^{**}	0.051^{**}
	(0.002)	(0.020)
Individual-level randomized × Demand-constrained	-0.010^{*}	
	(0.006)	
Treatment × Demand-constrained	0.055^{***}	
	(0.005)	
Individual-level Randomized × Treatment × Demand-constrained	0.013	
	(0.008)	
Individual-level randomized × High-quality cluster		−0.024
		(0.028)
Treatment × High-quality cluster		0.142^{***}
		(0.024)
Individual-level Randomized × Treatment × High-quality cluster		0.018
		(0.035)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	Yes
R^2	0.405	0.405
Adjusted R^2	0.405	0.405

*p<0.1; **p<0.05; ***p<0.01

Table H.10 Treatment effect heterogeneity for the fee meta-experiment w.r.t. cluster quality (attribute-based definition)

	Dependent variable:	
	Bookings	
	Low-quality clusters (attributes)	High-quality clusters (attributes)
	(1)	(2)
Treatment	-0.312^{***}	-0.231^{***}
	(0.016)	(0.017)
Individual-level Randomized	0.010	0.033
	(0.019)	(0.020)
Individual-level Randomized × Treatment	-0.052^{**}	-0.092^{***}
	(0.024)	(0.026)
Pre-treatment bookings	0.173^{***}	0.176^{***}
	(0.001)	(0.001)
Pre-treatment nights booked	-0.002^{***}	-0.003^{***}
	(0.000)	(0.000)
Pre-treatment gross guest spend	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)
Pre-treatment nights available	0.002^{***}	0.002^{***}
	(0.000)	(0.000)
Pre-treatment searches/night	0.199^{***}	0.035^{*}
	(0.022)	(0.018)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	Yes
R^2	0.405	0.406
Adjusted R^2	0.405	0.406

*p<0.1; **p<0.05; ***p<0.01

Acknowledgments

The authors are grateful to Lanbo Zhang, Minyong Lee, and Sharan Srinivasan for their assistance with the design and analysis of the experiments in this paper. We also thank numerous other Airbnb employees who have assisted with this project. We are also grateful to Jonathan Niles-Weed for his assistance. We also appreciate the helpful feedback we have received from Dean Eckles, Andrey Fradkin, Hannah Li, Iris Fung, Alex Moehring, Hong Yi Tu Ye, the Berkeley Haas MORS Macro Research Lunch, the MIT Sloan Social Analytics Lab, attendees of the 2019 Winter Conference on Business Analytics and the HBS Digital Doctoral Workshop. This experiment was classified as exempt by the MIT Committee on the Use of Humans as Experimental Subjects under Protocol #1807452488.