Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

David Holtz, Felipe Lobel, Ruben Lobel, Inessa Liskovich, Sinan Aral

To cite this article:
David Holtz, Felipe Lobel, Ruben Lobel, Inessa Liskovich, Sinan Aral (2025) Reducing Interference Bias in Online Marketplace Experiments Using Cluster Randomization: Evidence from a Pricing Meta-experiment on Airbnb. Management Science 71(1):390-406. https://doi.org/10.1287/mnsc.2020.01157

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 71, No. 1, January 2025, pp. 390–406
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

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

David Holtz,[a,b,*] Felipe Lobel,[c] Ruben Lobel,[d] Inessa Liskovich,[d] Sinan Aral[b,e]

[a] Management of Organizations and Entrepreneurship and Innovation, Haas School of Business, University of California, Berkeley, California 94720; [b] MIT Initiative on the Digital Economy, MIT Sloan School of Management, Cambridge, Massachusetts 02142; [c] Department of Economics, University of California, Berkeley, California 94720; [d] Airbnb, San Francisco, California 94103; [e] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

*Corresponding author

Contact: dholtz@haas.berkeley.edu, https://orcid.org/0000-0002-0896-8628 (DH); lobel@berkeley.edu, https://orcid.org/0000-0002-4603-4594 (FL); ruben.lobel@gmail.com, https://orcid.org/0000-0002-5824-8442 (RL); inessa.liskovich@airbnb.com (IL); sinan@mit.edu, https://orcid.org/0000-0002-2762-058X (SA)

Received: April 23, 2020
Revised: December 28, 2022; July 7, 2023
Accepted: July 11, 2023
Published Online in Articles in Advance: April 5, 2024

https://doi.org/10.1287/mnsc.2020.01157

Abstract. 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 because of 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 20% of the TATE estimate produced by an individual-level 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, nonstatistically 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.

History: Accepted by Chris Forman, information systems.

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2020.01157.

Keywords: 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 because of 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, Bojinov et al. 2022, Johari et al. 2022, Li et al. 2022, Bright et al. 2023) 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 nontrivial 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 it has been well studied in the research literature on unipartite social networks; in such settings, interference arises because of interactions between individuals, often referred to as peer effects (Manski 2000, Moffitt 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 the treatment assignment is then randomized at the cluster level as opposed to the individual level. Although 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 because of 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 preexisting linear model of interference to explore how online marketplace interference differs from social network interference and the implications that 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 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 product detail pages (PDPs) 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 that 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 Online 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 individual 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-level randomized meta-treatment arm (i.e., the "naive" experiment design) suggest that the treatment led to a statistically significant loss of 0.345 bookings per listing over the course of the experiment. However, when we compare this TATE estimate with 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, nonstatistically 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 that there is not a one-size-fits-all solution to interference bias in marketplace experiments and that each proposed solution (including our solution) has its strengths and weaknesses. Cluster randomization works well in marketplaces without centralized matching (in contrast to Bright et al. 2023), 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 identified 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),[5] including "two-sided randomization" (Johari et al. 2022) and "switchback" experimentation (Bojinov et al. 2022).[6]

Although each proposed solution to marketplace interference has appealing attributes, none of them offer 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. Although switchback experiments have appealing statistical properties, they can produce an inconsistent user experience for marketplace participants, and they 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, Chin 2019, Sävje et al. 2021), 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 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 with 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, Ye et al. 2018, Dubé and Misra 2023, Filippas et al. 2023). Many pricing interventions are tested and launched using randomized experiments; however, if the TATE estimates produced by these experiments are biased, marketplace designers may misestimate price elasticities and/or launch suboptimal policies. For instance, in Online 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* 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, for example, 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 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 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 20.1% of listing hosts searched at least once in their own geography, and only 12% 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 eight times. Furthermore, only 21.3% of hosts had at least one PDP view to a within-geography listing that was not their own. Among hosts who had at least one PDP view to a within-geography listing that was not their own, the median host carried out four PDP views across three distinct listings. More detailed data on search and PDP view activity in the month preceding our meta-experiment are shown in Figure 1. Given these results, in conjunction with the facts that (1) like our meta-experiment, many experiments run for much shorter periods of time than 30 days and (2) treatment interventions such as 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 because 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, an increase (decrease) in demand for their competitors' listings.[9] This is exactly the pattern we observe in the fee meta-experiment results presented in Section 5. Although 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 makes the estimated treatment effect larger in magnitude (we estimate a TATE of -0.345 bookings per listing in the individual-level randomized meta-treatment arm). We hypothesize 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 may shift their business from higher fee listings to lower 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*

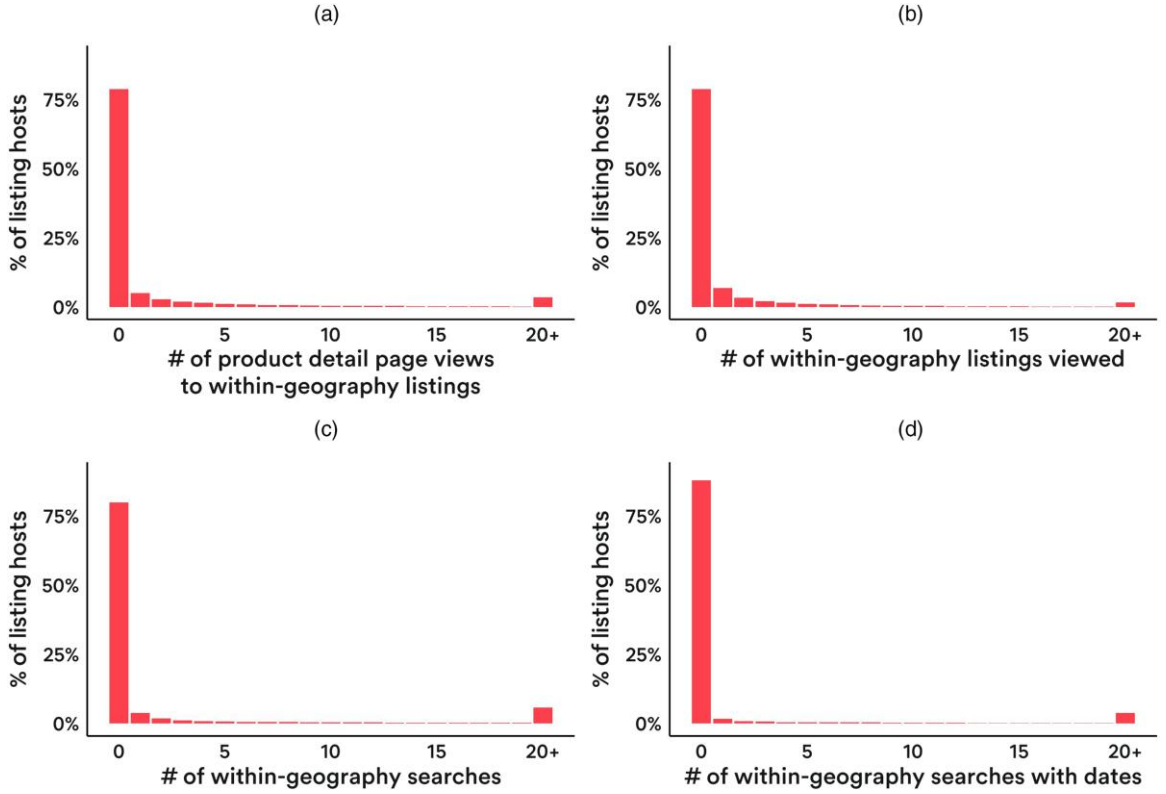
$$E[Y_i(\boldsymbol{Z})] = \alpha_i + \sum_{j \in V} B_{ij} Z_j, \tag{3}$$

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. The proof is given in Online Appendix C. □

Proposition 1 establishes that cluster randomization will never increase TATE estimation bias, but it does not provide any guidance on how to construct clusters. In any given marketplace setting, there will exist many

Figure 1. (Color online) Airbnb Host Search and PDP View Behavior in their Own Markets



Notes. Panel (a) shows the distribution of 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.

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 crossprice elasticities, to name a few possibilities. However, not all clustering 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 zero, 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 is not 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

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

Although in theory, Definition 1 provides a context-independent 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 an appropriate transformation of $\boldsymbol{B}$, it follows directly from Proposition 1 that $Q_C(\boldsymbol{P})$ can still be used to determine which one of two sets of clusters, C_1 and C_2, produces more biased difference-in-means TATE estimates.[10]

Proposition 2. *Given two matrices* $\boldsymbol{B}$ *and* $\boldsymbol{P}$ *of the same dimensions with all their elements bounded between* 0 *and* 1,

and for each element P_{ij} there exists a corresponding element B_{ij} such that $P_{ij} = f(B_{ij})$, where f is a convex and monotonically increasing function, then,

$$Q_{C_1}(\boldsymbol{P}) \leq Q_{C_2}(\boldsymbol{P}) \Rightarrow Q_{C_1}(\boldsymbol{B}) \leq Q_{C_2}(\boldsymbol{B}). \quad (5)$$

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 that (2) after identifying a set of clusters, $C(\cdot)$, an experiment designer can assess their quality by calculating $Q_C(\boldsymbol{P})$ given they are able to identify an appropriate proxy matrix $\boldsymbol{P}$.[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 Online 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 was a change to Airbnb's platform fees for guests. Airbnb's fees for guests were visible in three different locations throughout the booking process. First, guest platform fees were included in the total price shown to guests when a listing appeared in search (the upper panel in Figure 2). Second, if a guest opened the "price breakdown" tooltip on any search result, they were shown a price breakdown that separated out the nightly price and the guest platform fee (the lower panel in Figure 2). Finally, when viewing a listing's PDP, a detailed pricing breakdown (including fees) was 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[14] relative to the status quo.[15]

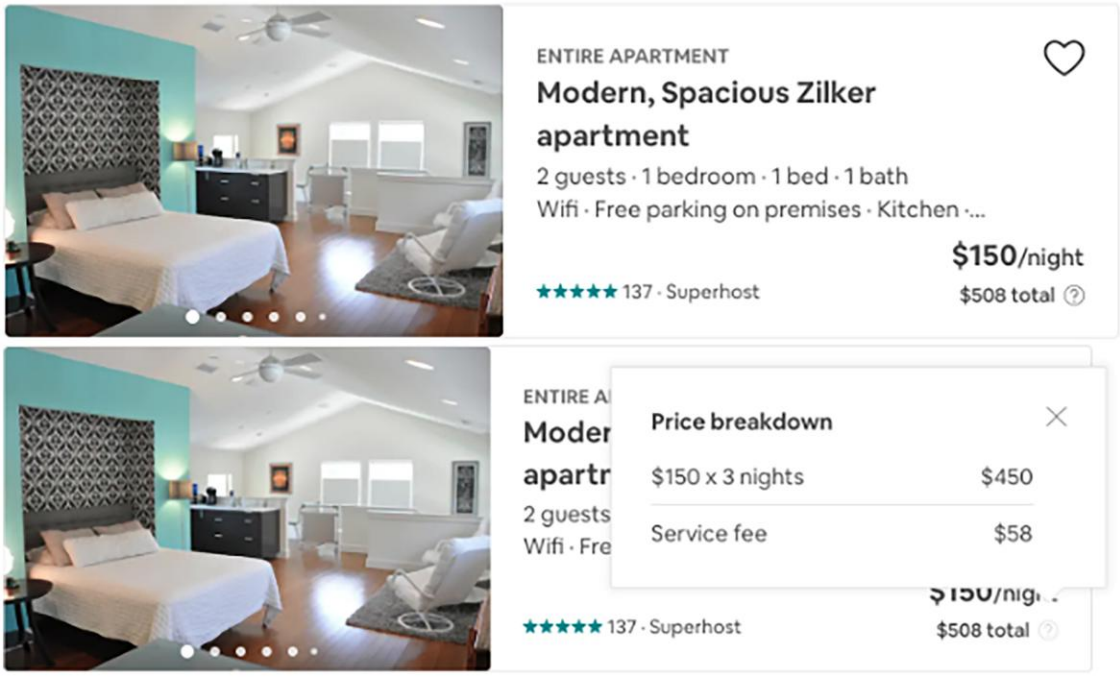
4.2. Experiment Design

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 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 used 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 listings into clusters (e.g., the simulation described in Online 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 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

Figure 2. (Color online) Airbnb Guest Platform Fees in Search



Notes. The upper panel shows a typical search result on Airbnb at the time of the experiment. In this case, the guest platform fee was included in the total price of $508. The lower panel shows what was displayed to guests after clicking the "price breakdown" tooltip; the guest platform fee (listed here as a service fee of $58) was broken out from the total nightly price.

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 these data. Given S sequences of listings, the skip-gram model attempts to maximize the objective function

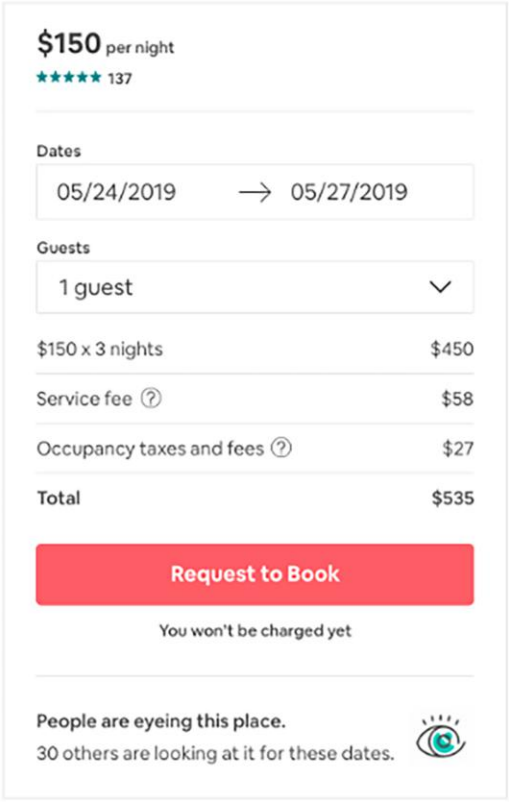
$$J = \max_{W,V} \sum_{s \in S} \frac{1}{|s|} \sum_{i=1}^{|s|} \sum_{-k \le j \le k,\ k \ne 0} \log\, p(L_{i+j} | L_i), \tag{7}$$

where k is the size of a fixed moving window over the listings in a session, $\boldsymbol{W}$ and $\boldsymbol{V}$ 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 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.

After 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 clusters. The algorithm starts from a single cluster containing all listings and then recursively bisects clusters into two subclusters. The algorithm stops bisecting subclusters when the tree reaches a depth of 20 or when a new subcluster will contain fewer 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 Online Appendix D.

4.2.2. Treatment Assignment Randomization. After 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%

Figure 3. (Color online) Airbnb Guest Platform Fees on the Product Detail Page



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

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 Poststratification. 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 pretreatment 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 nonexperimental 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 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 pretreatment 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 that our randomization procedure was sound.

5. Results

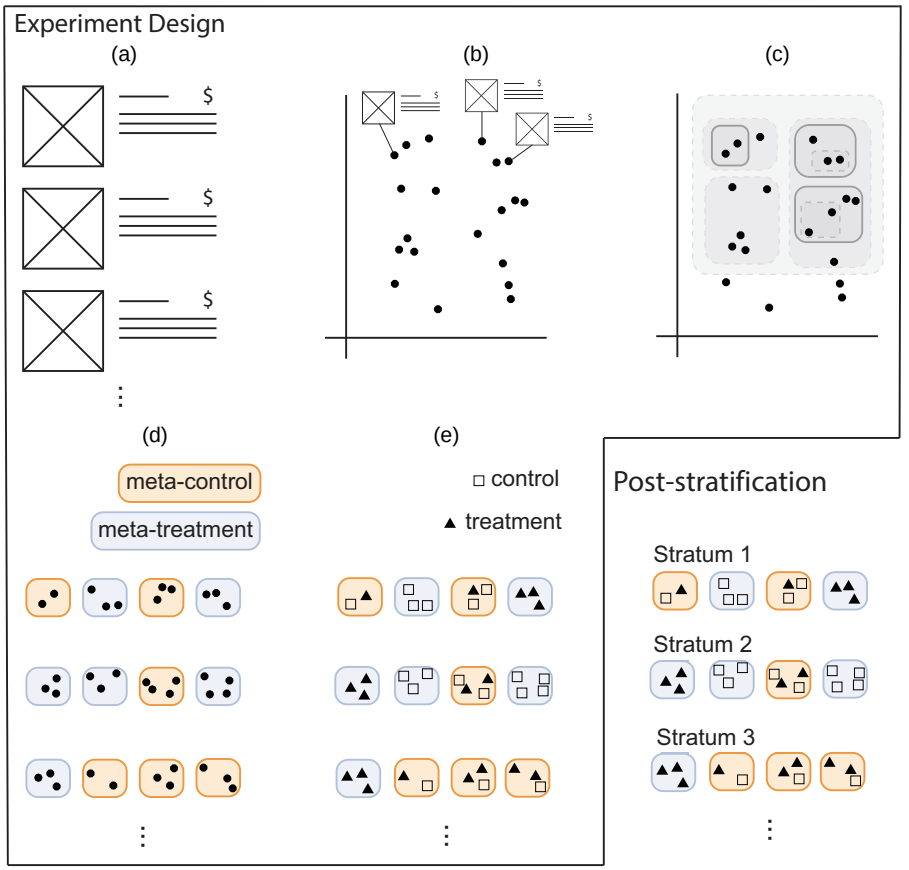
In this section, we present results from the fee meta-experiment. We focus on a single outcome metric, bookings, but the results for two alternative outcome metrics, nights booked and gross guest spend, are qualitatively similar and can be found in Online Appendix E. Because relative to the control, 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. Although 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 because the amount of marketplace interference will be reduced, but it will also have less precision. 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 and Cluster-randomized Results

We analyze both the individual-level randomized and cluster-randomized meta-treatment arms separately by

Figure 4. (Color online) The Experiment Design Process



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

Figure 5. (Color online) These Maps Illustrate Clusters Generated Using the Hierarchical Clustering Scheme Described in This Paper

Source. Srinivasan (2018).

Table 1. Confirming Balance Between Conditions

	Individual-level randomized			Cluster-randomized			Meta-experiment		
	Control	Treatment	*p*-value	Control	Treatment	*p*-value	Meta-control	Meta-treatment	*p*-value
Pretreatment 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	

Notes. This table tests for statistically significant differences in pretreatment 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.

estimating the following model on listing-level data,

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

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 belongs to; $\boldsymbol{X_i}$ is a vector consisting of listing i's pretreatment 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

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

where Y_i is the outcome of interest, M_i is a binary variable set to one when listing i is in the individual-level meta-treatment arm and zero when i is in the cluster-randomized meta-treatment arm, T_i is a binary variable set to one when listing i is exposed to the treatment, B_i is a variable indicating the stratum of clusters to which listing i belongs, $\boldsymbol{X_i}$ is a vector consisting of listing i's pretreatment 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]

Table 2. The TATE Results Obtained by Analyzing the Two Meta-treatment Arms Separately

	Dependent variable: *Bookings*	
	Individual-level randomized (1)	Cluster-randomized (2)
Treatment	−0.345*** (0.013)	−0.277*** (0.012)
Pretreatment bookings	0.174*** (0.001)	0.175*** (0.001)
Pretreatment nights booked	−0.003*** (0.000)	−0.003*** (0.000)
Pretreatment gross guest spend	−0.000*** (0.000)	−0.000*** (0.000)
Pretreatment nights available	0.002*** (0.000)	0.001*** (0.000)
Pretreatment searches/night	0.267*** (0.027)	0.033** (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

Notes. Individual-level randomized results are found in column (1), and cluster-randomized results are found in column (2). F.E., fixed effect; s.e., standard error.

**$p < 0.05$; ***$p < 0.01$.

In the 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 provide a lower bound on the amount of interference bias in the individual-level randomized meta-treatment arm.[21] Once we have estimated Equation (9), our estimate of the interference bias is

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

which is 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 because of interference bias. Plugging these point estimates into Equation (10), we estimate that 19.76% (±9.06%) of the TATE estimate achieved through the individual-level randomized experiment is because of interference bias and was eliminated through cluster randomization.

5.3. The Moderating Effect of Supply and Demand Constrainedness

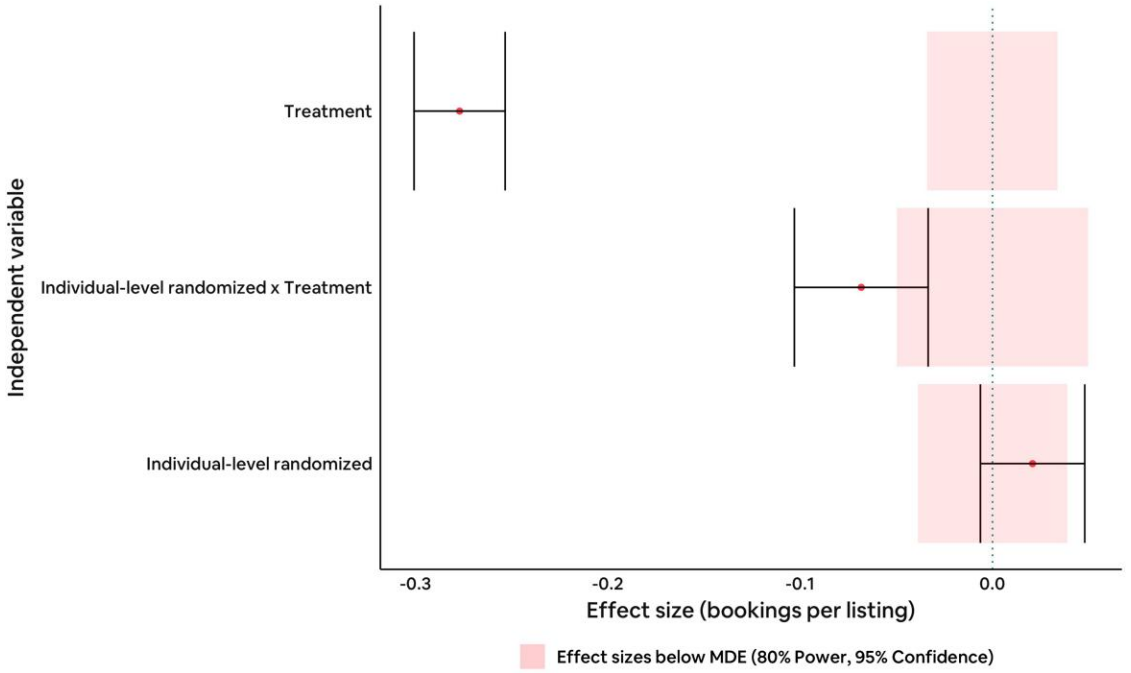
We hypothesize that the extent to which the TATE estimate under listing-level randomization suffers from interference bias will depend on marketplace conditions. More specifically, we expect that interference bias will be *larger* in geographies that are demand constrained and *smaller* in geographies that are supply constrained. The intuition for this is as follows; in an extremely

Table 3. Summary of the Meta-experiment Results for the Number of Bookings

	Dependent variable: *Bookings*				
	Overall (1)	Supply constrained (2)	Demand constrained (3)	Low-quality clusters (4)	High-quality clusters (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)
Pretreatment bookings	0.175***	0.174***	0.175***	0.172***	0.178***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Pretreatment nights booked	−0.003***	−0.003***	−0.003***	−0.003***	−0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Pretreatment gross guest spend	−0.000***	−0.000***	−0.000***	−0.000***	−0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Pretreatment nights available	0.001***	0.003***	0.000***	0.002***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Pretreatment 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
Semiclustered 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

Notes. 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 cluster quality. F.E., fixed effect; s.e., standard error.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Figure 6. (Color online) Coefficient Estimates for the Joint Analysis of the Fee Meta-experiment



Notes. Error bars represent 95% confidence intervals. The dotted line corresponds to a treatment effect of zero bookings per listing. The shaded area corresponds to values that are below the minimum detectable effect (MDE) (80% power, 95% confidence).

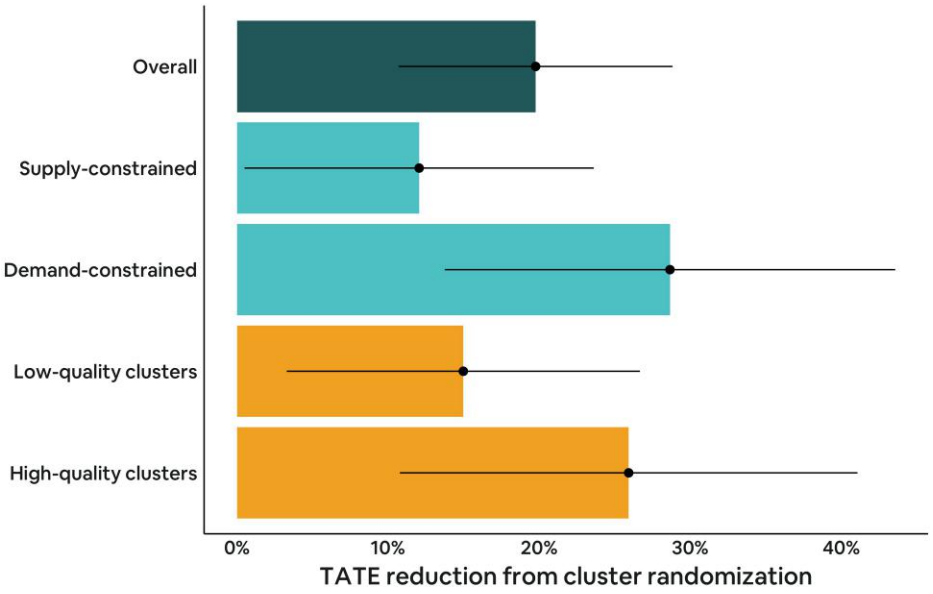
supply-constrained geography, all listings will eventually get booked, which will push the interference bias to zero, whereas in an extremely demand-constrained geography, only "more appealing" listings (i.e., only those in the treatment or control depending on the treatment intervention) will be booked, 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 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 meta-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. Although 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 in the online appendix), 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 Online 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 because the difference between these two estimates of interference bias

Figure 7. (Color online) Reduction in Bias from Cluster Randomization



Note. 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.

reduction is not statistically significant (see column (2) of Table H.9 in the online appendix).[22]

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-level randomized meta-treatment arm was because of interference bias. We also find suggestive, nonstatistically 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.

Although 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 road map 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 (Online Appendix A), conducting observational analysis to better understand the potential mechanisms driving interference in their setting (Section 3.1), and/or running simulated experiments (Online 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, for example, 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 (Online 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 our experiment (Section 4.2) to obtain a lower bound on the actual amount of interference bias present.

We believe that 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 because of 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, for example, 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, Ye et al. 2018, Dubé and Misra 2023, Filippas et al. 2023), 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 our experiment to assess the efficacy of cluster randomization when the treatment intervention under evaluation is more complex.

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. The authors thank numerous other Airbnb employees who have assisted this project and are grateful to Iris Fung, Jonathan Niles-Weed, and Peter Coles for their assistance. The authors also appreciate the helpful feedback they received from Dean Eckles, Andrey Fradkin, Hannah Li, Alex Moehring, Hong Yi Tu Ye, the Berkeley Haas Management of Organizations (MORS) Macro Research Lunch, the Massachusetts Institute of Technology (MIT) Sloan Social Analytics Laboratory, and attendees of the 2019 Winter Conference on Business Analytics and the Harvard Business School Digital Doctoral Workshop. This experiment was classified as exempt by the MIT Committee on the Use of Humans as Experimental Subjects under Protocol 1807452488. David Holtz was previously an employee of Airbnb between 2014 and 2016 and was an unpaid contractor while working on this paper. He currently holds a material financial interest in Airbnb, Inc. Ruben Lobel and Inessa Liskovich are currently full-time employees of Airbnb. Felipe Lobel was previously a paid intern at Airbnb and was an unpaid contractor while working on this paper. Airbnb had the right to review this paper subject to the following stipulations: (1) delaying disclosure or publication until any intellectual property rights have been adequately protected; (2) removing Airbnb confidential information from the work product; and (3) adding Airbnb as an author, as appropriate, of the work product and/or adding an appropriate disclosure concerning Airbnb's support of the project.

Endnotes

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

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

[3] As a result, if an experiment designer was 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 accommodations and experiences. More than 6 million listings appear on Airbnb, and since the company's founding in 2008, over 1 billion guest arrivals have occurred on the platform (Airbnb 2019).

[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 Online Appendix B.

[6] Analysis-based solutions to the problem have also been suggested (e.g., in Bright et al. 2023).

[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.

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

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

[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, for example, Baird et al. (2018).

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

[14] Because of our non-disclosure agreement 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 data set to long-tenured listings. However, the clusters used in our experiment include all listings, regardless of tenure on the platform.

[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 road map 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.

[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.

[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 in the online appendix.

[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 in the online appendix.

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

[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 five 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 in the online appendix shows our results using this alternative cluster quality measure; they are qualitatively similar to those found in Table 3.

[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 milder treatment intervention. Because the treatment intervention was milder, this meta-experiment was underpowered to detect interference bias, despite having a sample size in the millions.

References

Airbnb (2019) Airbnb news: About us. Accessed September 1, 2019, https://news.airbnb.com/about-us/.

Aral S, Nicolaides C (2017) Exercise contagion in a global social network. *Nat. Comm.* 8(1):14753.

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

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

Aronow PM, Samii C (2017) Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Statist.* 11(4):1912–1947.

Athey S, Eckles D, Imbens GW (2018) Exact p-values for network interference. *J. Amer. Statist. Assoc.* 113(521):230–240.

Azevedo EM, Deng A, Montiel Olea JL, Rao J, Weyl EG (2020) A/B testing with fat tails. *J. Political Econom.* 128(12):4614.

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

Basse G, Feller A (2018) Analyzing two-stage experiments in the presence of interference. *J. Amer. Statist. Assoc.* 113(521):41–55.

Blake T, Coey D (2014) Why marketplace experimentation is harder than it seems: The role of test-control interference. *Proc. Fifteenth ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 567–582.

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J. Statist. Mech. Theory Experiment.* 2008(10):P10008.

Bojinov I, Simchi-Levi D, Zhao J (2022) Design and analysis of switchback experiments. *Management Sci.* 69(7):3759–3777.

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 Anal.* 21(1):97–124.

Bright I, Delarue A, Lobel I (2023) Reducing marketplace interference bias via shadow prices. *Proc. 24th ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 300.

Chin A (2019) Regression adjustments for estimating the global treatment effect in experiments with interference. *J. Causal Inference* 7(2):20180026

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

Cox DR (1958) *Planning of Experiments* (Wiley, New York).

Dubé JP, Misra S (2023) Personalized pricing and consumer welfare. *J. Political Economy* 131(1):131–189.

Eckles D, Karrer B, Ugander J (2017) Design and analysis of experiments in networks: Reducing bias from interference. *J. Causal Inference* 5(1):20150021.

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

Filippas A, Jagabathula S, Sundararajan A (2023) The limits of centralized pricing in online marketplaces and the value of user control. *Management Sci.* 69(12):7202–7216.

Fradkin A (2015) Search frictions and the design of online marketplaces. Working paper, Massachusetts Institute of Technology, Cambridge.

Gerber AS, Green DP (2012) *Field Experiments: Design, Analysis, and Interpretation* (W. W. Norton, New York).

Grbovic M, Cheng H (2018) Real-time personalization using embeddings for search ranking at Airbnb. *Proc. 24th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 311–320.

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

Holtz D, Zhao M, Benzell SG, Cao CY, Rahimian MA, Yang J, Allen J, et al. (2020) Interdependence and the cost of uncoordinated responses to COVID-19. *Proc. Natl. Acad. Sci. USA* 117(33):19837–19843.

Hudgens MG, Halloran ME (2008) Toward causal inference with interference. *J. Amer. Statist. Assoc.* 103(482):832–842.

Ifrach B, Holtz DM, Yee YH, Zhang L (2016) Demand prediction for time-expiring inventory. U.S. Patent Application No. 14/952,576.

Imai K, Tingley D, Yamamoto T (2013) Experimental designs for identifying causal mechanisms. *J. Roy. Statist. Soc. Ser. A* 176(1):5–51.

Johari R, Li H, Liskovich I, Weintraub GY (2022) Experimental design in two-sided platforms: An analysis of bias. *Management Sci.* 68(10):7069–7089.

Kang JH, Park CH, Kim SB (2016) Recursive partitioning clustering tree algorithm. *PAA Pattern Anal. Appl.* 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. *Proc. ACM Web Conf. 2022* (Association for Computing Machinery, New York), 182–192.

Liu L, Hudgens MG (2014) Large sample randomization inference of causal effects in the presence of interference. *J. Amer. Statist. Assoc.* 109(505):288–301.

Liu M, Mao J, Kang K (2021) Trustworthy and powerful online marketplace experimentation with budget-split design. *Proc. 27th ACM SIGKDD Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 3319–3329.

Mahalanobis PC (1936) On the generalized distance in statistics. *Proc. Natl. Inst. Sci. (Calcutta)* 2:49–55.

Manski CF (2000) Economic analysis of social interactions. *J. Econom. Perspect.* 14(3):115–136.

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. Preprint, submitted January 16, https://arxiv.org/abs/1301.3781.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 26 (Curran Associates Inc. Red Hook, NY), 3111–3119.

Miratrix LW, Sekhon JS, Yu B (2013) Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. Roy. Statist. Soc. Ser. B Statist. Methodology* 75(2):369–396.

Moffitt RA (2001) Policy interventions, low-level equilibria, and social interactions. Durlauf SN, Young HP, eds. *Social Dynamics* (MIT Press, Cambridge, MA), 6–17.

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

Pouget-Abadie J, Mirrokni V, Parkes DC, Airoldi EM (2018) Optimizing cluster-based randomized experiments under monotonicity. *Proc. 24th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 2090–2099.

Rosenbaum PR (2007) Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* 102(477):191–200.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psych.* 66(5):688–701.

Saveski M, Pouget-Abadie J, Saint-Jacques G, Duan W, Ghosh S, Xu Y, Airoldi EM (2017) Detecting network effects: Randomizing over randomized experiments. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 1027–1035.

Sävje F, Aronow P, Hudgens M (2021) Average treatment effects in the presence of unknown interference. *Ann. Statist.* 49(2): 673–701.

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

Srinivasan S (2018) Learning market dynamics for optimal pricing. *Medium* (August 10), 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. *Statist. Methods Medical Res.* 21(1):55–75.

Ugander J, Karrer B, Backstrom L, Kleinberg J (2013) Graph cluster randomization: Network exposure to multiple universes. *Proc. 19th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 329–337.

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. *Proc. 24th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 932–940.