Essays on the Design of Online Marketplaces and Platforms

by

David Holtz

A.B., Princeton University (2010)
M.A., Johns Hopkins University (2013)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Management

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author . . . .
Sloan School of Management
May 7, 2021

Certified by . . . .
Sinan Aral
David Austin Professor of Management
Professor of Information Technology and Marketing
Thesis Supervisor

Accepted by . . . .
Catherine Tucker
Sloan Distinguished Professor of Management Science
Professor of Marketing
Chair, MIT Sloan PhD Program

Essays on the Design of Online Marketplaces and Platforms

by

David Holtz

Submitted to the Sloan School of Management on May 7, 2021, in partial fulfillment
of the requirements for the degree of Doctor of Philosophy in Management

Abstract

This dissertation consists of three chapters that concern the design of online marketplaces and platforms. In Chapter 1, I estimate the impact of increasing the extent to which content recommendations are personalized by analyzing the results of a randomized experiment on approximately 900,000 Spotify users across seventeen countries. I find that increasing recommendation personalization increased the number of podcasts that Spotify users streamed, but also decreased the individual-level diversity of Spotify users' podcast consumption and increased the dissimilarity between the podcast consumption patterns of different users across the population. In Chapter 2, I propose methods for obtaining unbiased estimates of the total average treatment effect (TATE) when conducting experiments in online marketplaces, and test the viability of said methods using a simulation built on top of scraped data from Airbnb. I find that blocked graph cluster randomization can reduce the bias of TATE estimates in online marketplaces by as much as 64.5%, however, this reduction in bias comes with a substantial increase in root-mean-square error (RMSE). I also find that fractional neighborhood treatment response (FNTR) exposure models and inverse probability-weighted estimators have the potential to further reduce bias, depending on the choice of FNTR threshold. In Chapter 3, I conduct two large-scale meta-experiments on Airbnb in an attempt to estimate the actual magnitude of bias in TATE estimates from marketplace interference. In both meta-experiments, some Airbnb listings are assigned to experiment conditions at the individual-level, whereas others are assigned to experiment conditions at the level of clusters of listings that are likely to substitute for one another. The two meta-experiments measure the impact of two different pricing-related interventions on Airbnb: a change to Airbnb's fee policy, and a change to the pricing algorithm that Airbnb uses to recommend prices to sellers. Results from the fee policy meta-experiment reveal that at least 32.60% of the treatment effect estimate in the Bernoulli-randomized meta-experiment arm is due to interference bias. Results from the pricing algorithm meta-experiment highlight the difficulty of detecting interference bias when treatment interventions require intention-to-treat analysis.

Thesis Supervisor: Sinan Aral
Title: David Austin Professor of Management
Professor of Information Technology and Marketing

Acknowledgments

I am incredibly fortunate and privileged to have had the opportunity to do my PhD at MIT. I cannot imagine a more vibrant intellectual community to have been a part of over the past six years, and feel I have grown immensely as a researcher and thinker. First and foremost, I thank my advisor, Sinan Aral, for his support and mentorship during this six-year-long journey. He inspired me to seek out and tackle ambitious, interesting, and important projects, and taught me how to do so in an extremely rigorous way. It has been a great pleasure to be his student. I also thank my committee members, Dean Eckles and John Horton, both of whom have regularly provided invaluable mentorship, feedback, and career advice.

Beyond my committee, I am extremely thankful to the many academic mentors and friends who have contributed to my success during graduate school. At MIT, a number of fantastic professors, teachers, and advisors, including Erik Brynjolfsson, Wanda Orlikowski, Tamara Broderick, Ben Golub, Glenn Ellison, Bengt Holmtsröm, Frank Schilbach, Victor Chernozhukov, Josh Angrist, Sara Fisher Ellison, and Rachael Meager, inspired and nurtured my interest in economics, technology, causal inference, network science, and machine learning. Outside of MIT, I've benefited tremendously from the guidance and friendship of folks like Johan Ugander, Chiara Farronato, Tianshu Sun, Jon Hersh, Arun Sundararajan, Anindya Ghose, Foster Provost, Ramesh Johari, Ravi Bapna, Ed McFowland, Zhe Zhang, Hyunjin Kim, Sam Ransbotham, Grace Gu, Marios Kokkodis, Katherine Hoffman Pham, Martin Saveski, Joel Waldfogel, Christian Catalini, Tuan Phan, and Jui Ramaprasad, all of whom have helped me navigate the crazy world of academia and make sense of my own research. I am particularly thankful to Sid Suri, who has basically been an informal second advisor to me over the past year. I can't imagine what the pandemic year would've been like without Sid's mentorship and friendship, and I look forward to meeting him in person once the world returns to normal.

During my PhD, I have been lucky to work with enough brilliant coauthors to fill an MLB roster.[1] I have learned an incredible amount from collaborating with Ruben Lobel, Inessa Liskovich, Sanaz Mobasseri, Janet Xu, Longqi Yang, Sonia Jaffe, Ben Carterette, Praveen Chandar, Henriette Cramer, Zahra Nazari, Alex Dow, Diana MacLean, Liane Scult, Seth Benzell, Jeremy Yang, Amin Rahimian, Cathy Cao, Jenny Allen, Tara Sowrirajan, Dipayan Ghosh, Jerry Zhang, Paramveer Dhillon, and Christos Nicolaides. I am particularly grateful to have had the chance to work with, learn from, and become friends with Andrey Fradkin over the past seven years. When Andrey and I began working on a paper together in 2014, I had no idea I would finish my PhD before our paper made it to publication.[2] He has continually pushed me to think about things more deeply and to work more carefully, and has played a huge role in my growth as a researcher.

My decision to pursue a PhD at a business school was in part driven by the fact that B-school academia allows me to approach research questions as an academic while still collaborating with and learning from those in industry. The things I learned in the Bay Area from 2012 to 2015 have inspired my entire research agenda, and my continued growth is in large part thanks to my interactions with friends and colleagues in the tech industry. Thank you to Terry Angelos, Vickie Peng, Chris Chen, Kevin Rice, Jason Bosinoff, Alex Rampell, and Eddie Lim for taking a chance on a physicist who knew nothing about tech. Thanks also to Elena Grewal, Bar Ifrach, Jan Overgoor, Riley Newman, Ricardo Bion, Vasyl Pihur, Hector Yee, and the rest of the A-team at Airbnb for helping me evolve into the data scientist I am today, and for sparking my interest in online marketplaces. Thanks to Aline Lerner, who believed in my skills as a data scientist before almost anyone else, and is always looking out for opportunities to do cool things with data. Thank you to Sam Way, Clark Lemke, and Briana Vecchione, with whom I had memorable late afternoon conversations on the 62nd floor of 4 WTC. Thanks to everyone at Facebook CDS, from whom I learned invaluable lessons about how to conduct top-tier computational social science research. A special thank you to Sean Taylor, who was willing to

[1] I am not suggesting my co-authors actually form a baseball team. Were they to do so, I do not think the team would be good at baseball.

[2] I've learned a lot about social science publication cycles since then.

grab a beer with someone he barely knew[3] and talk about Information Systems PhD programs. Sean has been the most supportive "academic older sibling" imaginable and without his help, I would not be where I am today.

I owe an immense amount to numerous students and post-docs at MIT, who are now lifelong friends and colleagues. Thanks to my fellow M(IT) PhD students, Emma Van Inwegen, Sebastian Steffen, Alex Moehring, Hong-Yi Tu Ye, Guillaume Saint-Jacques, and Shan Huang, who have made the past six years fly by and have been a constant source of guidance and support. I am particularly indebted to Michael Zhao, with whom I've had many engaging conversations in the cafe car of NYC-bound Amtrak trains, and Daniel Rock and Avi Collis, without whom I truly could not have navigated the final years of graduate school. Thanks to IDE post-docs Xiang Hui, Sagit Bar-Gill, Ananya Sen, and Meng Liu for your friendship and advice. Thank you to my Sloan classmates Mahreen Khan, Madhav Kumar, Heather Yang, Jenna Myers, Claire McKenna, Alex Kowalski, Simon Friis, Suzie Noh, and Georg Rickmann, who were instrumental in making it through the first two years of school and staying sane thereafter. A special thanks goes to the "Economagic" group of Joe Hazell, Layne Kirshon, Mazi Kazemi, Peter Hansen, and Tom Ernst, who have always kindly answered my dumbest questions about econometrics.

The Initiative on the Digital Economy has been a huge part of my PhD experience, and I'm incredibly appreciative of the work that David Verrill, Carrie Reynolds, Susan Young, and the rest of the IDE team have done to build such an amazing hub for impactful digitization research. I'm also thankful to Jeanne Ross, Chris Foglia, and the rest of the team at CISR. Because of CISR's generosity, I have been able to attend and present my research at conferences across the world during my PhD. I'm grateful to Matt Salganik, Chris Bail, and everyone else involved in SICSS 2018. Although my Matt Salganik back tattoo was temporary, the experiences I had in Durham were incredibly formative, and will continue to influence my research for the entirety of my career. I am lucky to have attended three straight NBER graduate student tutorials on digitization. Those tutorials led to multiple eureka moments, and I am immensely

[3] RIP to the Blind Cat.

thankful to Avi Goldfarb and Shane Greenstein for organizing them. I also owe a huge thank you to Hillary Ross, Davin Schnappauf, and Sarah Massey for all of their help and support throughout the PhD process.

I wouldn't be the person I am today were it not for the many outstanding mentors I've had throughout the years. Toby Marriage, Lyman Page, and Brian Wecht inspired and nurtured my interest in physics, and I continue to look up to them all today. Hyeseung Marriage-Song helped me navigate my 20s and escape largely unscathed. Hans Kriefall, Glen Pannell, Bob Sandberg, Carol Dunne, Joel Mercier, Jeff Snyder, Dan Trueman, James Folta, Dan Burt, Max McCal, Jim Miller, and Joe Gaudett were instrumental in my development as a musician, thespian, and comedian. Kristen Record and Linda Cline helped kickstart my interest in math and science, and encouraged me to set my sights high. Stephanie Colman, Rena Sydel, and Jimmy Martin spent countless hours with me when I was a precocious (and probably annoying) kid, and encouraged me to pursue my passions.

Finally, I am extremely fortunate to have the support of fantastic friends and family. I am immensely grateful to Steve, LAC, Zach, Halcyon, Nuss, Molly, Mike, Jason, Jon, Charlie, Billy, Jacob, Willie, Siyu, Ben, Meredith, Kevin H., Sarah, Joy, Kevin S., Roger, Rob, Lance, Albert, Chris G.D., Leila, Chris H., Anthony, and Ponchi. Thank you to my grandparents, Carole, Gilbert, Irwin, and Bernice, for pushing me to do my best and asking me to explain Bitcoin over charbroiled Outback steaks. Thank you to my parents, Tracie and Jeff, for their unconditional love and support. Above all, I am particularly indebted to my mother, who has sacrificed so much so that I could achieve my dreams.

Contents

Chapter 1

The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify[1]

1.1 Introduction

Recommender systems and algorithmic content curation play an increasingly large role in people's lives. For instance, algorithmic recommendations influence the news and entertainment that we consume, the products that we purchase, and the people with whom we develop romantic relationships. Collaborative filtering recommendation systems drive 35% of product choices on Amazon (Lamere and Green 2008) and 60% of consumption choices on Netflix (Thompson 2008). However, despite recom-

[1] With Ben Carterette (Spotify), Praveen Chandar (Spotify), Henriette Cramer (Spotify), Zahra Nazari (Spotify) and Sinan Aral (MIT Sloan School of Management). We are grateful to Samuel F. Way, Briana Vecchione, John Horton, Dean Eckles, Jui Ramaprasad, Joel Waldfogel, Daniel Rock, Michael Zhao, Katherine Hoffman Pham, Emma van Inwegen, Mazi Kazemi, Alex Moehring, Sebastian Steffen, Seth Benzell, Mahreen Khan, Hong Yi Tu Ye, Sanaz Mobasseri, and Martin Saveski for their helpful feedback. We also thank numerous other Spotify employees who have assisted with this project. We have also received helpful feedback on this work from participants at CIST 2020, SCECR 2020, ISMS 2020, ACM EC 2020, the Facebook Workshop on Responsible Recommender Systems, the LSE Information Systems & Innovation Seminar, and the MIT Sloan Marketing Seminar, as well as anonymous reviewers at *Management Science* and *Marketing Science*. This experiment was classified as exempt by the MIT Committee on the Use of Humans as Experimental Subjects under Protocol #E-1974.

mender systems' increasing ubiquity, the ways in which they impact the *types* of choices we make are still not well understood. While some scholars have speculated that recommender systems will lead to "filter bubbles" (Sunstein 2001, Pariser 2011), others hypothesize that recommender systems will homogenize user consumption, leading to the "rich getting richer" (Negroponte 1996, Van Alstyne and Brynjolfsson 2005, Salganik et al. 2006). In this paper, we analyze a large scale field experiment conducted on Spotify, one of the world's leading streaming platforms. During the experiment, both treatment and control users were recommended podcasts with the sole aim of increasing podcast consumption; while control users were recommended podcasts popular amongst those in their demographic group, treatment users were provided fully personalized recommendations based on their existing music listening history. We measure the impact of more personalized content recommendations on user engagement, as well as individual-level and aggregate podcast category diversity.

We find that the recommender system in the experiment increased the average number of podcasts streamed per user by 28.90% relative to the less-personalized, popularity-based recommendation strategy. We also test for the impact of the algorithm on user-level podcast category diversity, as measured through the Shannon entropy (Shannon 1948, Teachman 1980), and aggregate podcast category diversity, as measured through a quantity that we call "intragroup diversity" (Aral and Van Alstyne 2011).[2] We find that the more personalized algorithm *decreased* individual-level diversity, but *increased* intragroup diversity. These results indicate that recommender systems and personalization algorithms have the capacity to push individual users into homogeneous consumption patterns that are increasingly dissimilar from those of their peers. While the effects of the treatment are largest for streams originating from the section of Spotify's app where personalized recommendations are delivered, we observe evidence that the treatment also affected streams originating from other parts of the app. This suggests that exposure to personalized recommendations can also affect the diversity of content that users organically engage with. Furthermore,

[2] While we quantify diversity using these particular measures, there is a large body of academic literature discussing different approaches to measuring diversity. See, for instance, Mitchell et al. (2020).

the treatment effects that we observe quickly dissipated once the experiment had ended, indicating that on average users revert to their counterfactual baseline listening habits once personalized recommendations are no longer shown. This suggests that user exposure to personalized recommendations must be persistent for firms to realize long-term increases in user engagement, and that firms have the ability to "course correct" if they discover that personalized recommendations are having an undesirable effect on users' consumption patterns.

In aggregate, our findings highlight the potential for recommender systems to create an "engagement-diversity trade-off" for firms when recommendations are optimized solely to drive consumption; while algorithmic recommendations can increase user engagement, they can also homogenize individual users' consumption and Balkanize user content consumption. This shift in consumption diversity can negatively impact user churn rates and lifetime values (Anderson et al. 2020), and the optimal strategy for content creators, including platforms that create original content (such as Spotify or Netflix). It is possible that our findings also extend to settings where diversity is measured with respect to the ideological slant or extremity of content. If so, the "engagement-diversity trade-off" suggests that recommender systems that increase engagement/consumption can also create costs for firms, due to the high level of public scrutiny given to personalized recommendations, and impact public discourse on platforms through the creation of "filter bubbles." In light of our results, we believe it is worthwhile for academics and practitioners to continue developing personalization techniques that explicitly take into account the diversity of content recommended to users (Marler and Arora 2004, Castells et al. 2015, Lacerda 2017).[3]

1.2 Related Literature

This paper contributes to a growing body of literature that focuses on the economic and societal impacts of recommender systems, which use product metadata as well

[3] In this paper, we consider the implications of personalized recommendations from the firm's perspective. We do not consider whether personalized recommendations are welfare-increasing or -decreasing for Spotify users, although we think this is an important topic for future research.

as implicit and explicit user feedback to generate personalized product recommendations to users (Resnick and Varian 1997, Adomavicius and Tuzhilin 2005). Previous research has shown that the adoption of new digital technologies, such as online streaming platforms, can influence the types and diversity of content that people consume (Dewan and Ramaprasad 2012, Aguiar and Waldfogel 2018, Datta et al. 2018, Knox and Datta 2020), and recommender systems are likely a key mechanism in driving these consumption changes. Early research established that online recommendations impact consumer product choices (Senecal and Nantel 2004), and that recommender systems in particular often lead to increased engagement and/or purchases (Das et al. 2007, Freyne et al. 2009, De et al. 2010, Zhou et al. 2010, Oestreicher-Singer and Sundararajan 2012b, Sharma et al. 2018). However, there is no clear consensus on the impact that recommender systems have on the *diversity* of items that users consume.

Building on the work of Brynjolfsson et al. (2011), a series of papers have attempted to quantify, through models, simulations, observational analysis, and natural experiments, the effect of recommender systems on sales diversity (Fleder and Hosanagar 2009, Wu et al. 2011, Oestreicher-Singer and Sundararajan 2012a, Jannach et al. 2013, Hosanagar et al. 2013, Nguyen et al. 2014, Hervas-Drane 2015). Most, but not all, of these papers measure changes in sales diversity by looking at differences in the Lorenz curve corresponding to product consumption or sales. Many of these studies argue that recommender systems make individual consumption more diverse, while *decreasing* aggregate consumption diversity. To provide some intuition for how this might occur, imagine a platform with four users and four pieces of content: A, B, C, and D. A recommender system could shift users' consumption vectors from (A), (B), (C), (D) to (AB), (AB), (AB), (AB). While each individual users' consumption is less concentrated, aggregate consumption is more concentrated.

A separate stream of research has focused on the impact that recommender systems have on the *types* of content that people consume, and the resultant societal impacts. While some papers in this research stream argue that algorithms lead to increased ideological segregation (Flaxman et al. 2016, Tufekci 2018, Ribeiro et al.

2019), others find that users' tendency to engage with content that agrees with their ideological preferences is driven by user choice, as opposed to algorithms (Gentzkow and Shapiro 2006, Bakshy et al. 2015). Importantly, content diversity is a distinct concept from sales diversity, and the effect of recommender systems on these two types of diversity need not be the same. For instance, a recommender system could lead users to consume more long tail content that is all ideologically similar. In this paper, we focus on diversity with respect to podcast categories, rather than ideological affiliation. However, both types of diversity characterize the *type* of content that users consume, and it is possible that the application of our analytical framework to data with ideological labels would produce similar results.

Our work contributes to an emerging literature that uses randomized field experiments to measure the impact of recommender systems on the *diversity* of content that users consume (Claussen et al. 2019, Lee and Hosanagar 2019). Claussen et al. (2019) find that personalization *decreased* individual-level diversity, and that this decrease in consumption diversity spilled over to non-personalized sections of the website they studied. In contrast, Lee and Hosanagar (2019) find that the introduction of a recommender system had a neutral-to-positive effect on individual-level diversity, but decreased aggregate diversity. Our research is also closely related to Anderson et al. (2020), who use observational data from Spotify to study the relationship between personalization and listening diversity. They find that user-driven listening was more diverse than algorithmic listening, and that users who became more diverse over time did so by shifting away from algorithmic listening. Importantly, they also find that users with more diverse listening habits were less likely to leave the platform and were more likely to eventually become paid subscribers, suggesting that short-term increases in engagement due to personalization may have unintended long-term business implications.

We build on the existing literature by analyzing data from a randomized field experiment, which allows us to credibly estimate the causal effect of personalized recommendations on content consumption, and attempting to resolve the tension that exists between the two aforementioned experimental estimates. In contrast to much of the recommender systems research in economics and management, which has focused on measuring changes in sales diversity, we focus on measures of diversity that take into account the *types* of content that users consume, as measured through podcast category tags on Spotify. Like Claussen et al. (2019), we find that personalized recommendations decreased individual-level diversity, and we also find that personalized recommendations simultaneously increased aggregate diversity. These findings are at odds with Lee and Hosanagar (2019), who find that the introduction of algorithmic recommendations had a neutral-to-positive effect on individual-level diversity while decreasing aggregate diversity. We believe this contrast may be due to a number of factors. First, the impact of recommender systems can depend on a wide range of factors, including but not limited to the type of data used for training (Lin et al. 2015), the algorithm used to generate recommendations (Wu et al. 2011, Jannach et al. 2013), and the setting in which the recommender is deployed. We study the impact of a novel algorithm (which predicts podcast affinity based on a user's music listening history) in a novel setting (podcast recommendations on Spotify). Second, this paper quantifies diversity differently than Lee and Hosanagar (2019). The potential for different measures of diversity to suggest different types of effects underscores the need to develop and use a number of different measures when quantifying the impact of recommendation systems on consumption diversity.

1.3 Research Setting

1.3.1 Spotify

The setting for our study is Spotify, one of the world's leading streaming platforms. Spotify was founded in 2006, and as of December 2019 had 271 million monthly active users and 124 million paying subscribers.[4] Although Spotify launched as a music streaming platform, in 2015 the company expanded its offerings to include

[4] `https://s22.q4cdn.com/540910603/files/doc_financials/2019/q4/Shareholder-Letter-Q4-2019.pdf`

videos and, more importantly for this study, podcasts.[5] Podcasts are an increasingly popular type of content to stream online, and represent an important new vertical for Spotify. According to Edison Research, 51% of the U.S. population has listened to at least one podcast, and 32% listens to podcasts on a monthly basis. Among monthly podcast listeners, 43% have listened to a podcast on Spotify (Edison Research 2019).

Spotify users on mobile are able to access three different sections of the Spotify app via a navigation bar that runs along the bottom of the phone screen: "Your Library," "Search," and "Home." The "Your Library" section of the app allows a user to access albums, playlists and albums that they have previously saved, as well as podcasts they have previously followed and podcast episodes they have previously downloaded. The "Search" section of the app allows users to search Spotify's content library for specific pieces of content. The "Home" section of the app is most relevant to our research. It presents the user with a ranked set of "shelves," each of which contains a ranked set of "cards." Shelves correspond to different types of content, such as "content a user was recently listening to," or "music from a particular genre." Each card is essentially a link to a piece of content (e.g., a playlist or Spotify artist page). Shelves in the "Home" section of the app, and the cards within each shelf, are ordered by a combination of machine learning algorithms and human editors.[6] A screenshot of the "Home" section of the Spotify app on iOS can be seen in Figure 1.1.

In this paper, we will analyze changes to the number of podcasts users stream, as well as the *types* of podcasts users stream as measured through podcast category tags. Each podcast stream also has a "referrer" field associated with it, which indicates which part of the Spotify app the stream originated from. This field allows us to differentiate between streams that originated from the "Home" section of the app (where the experiment introduced variation in recommended podcasts) and streams that originated from other sections of the app (where the experiment did not introduce any variation).

[5] `https://www.fastcompany.com/3046504/spotify-launches-podcasts-video-and-context-based-listening`

[6] Details of Spotify's approach to ranking home content can be found in McInerney et al. (2018).

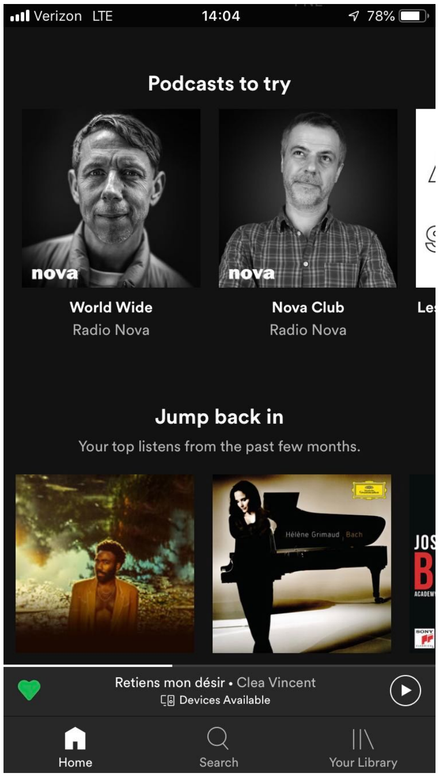
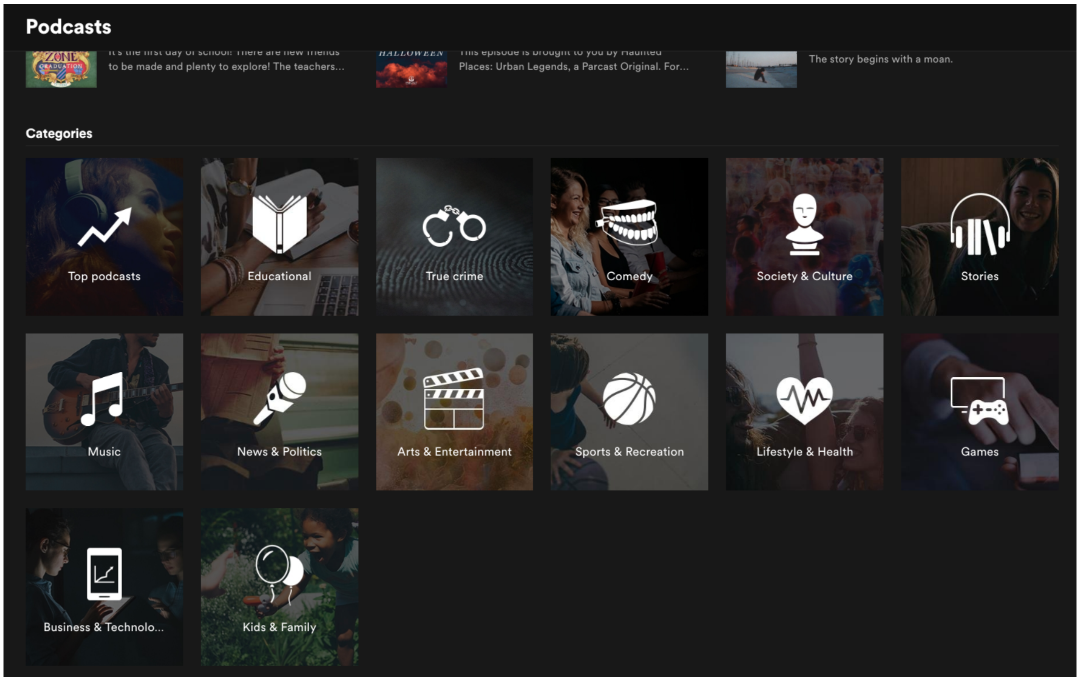


Figure 1.1: A screenshot of the "Podcasts to try" shelf on the Spotify iOS app. During the experiment, this shelf was fixed in the second slot on the "Home" section of the Spotify app.

1.3.2 Podcast categorization

At the time of the experiment, there were thirteen podcast category tags that could be associated with a podcast on Spotify: “Arts & Entertainment,” “Business & Technology,” “Comedy,” “Educational,” “Games,” “Kids & Family,” “Lifestyle & Health,” “Music,” “News & Politics,” “Society & Culture,” “Sports & Recreation,” “Stories,” and “True Crime.” Figure 1.2 shows the podcast section of the “Browse” pane on Spotify’s desktop app, which allows users to browse through podcasts by selecting one of the thirteen categories. Category tags are extracted from podcasts’ RSS feeds, and are not determined internally at Spotify. Podcast creators can specify as many category tags for a podcast as they wish, although many podcast upload tools limit podcast creators to three categories. Of the category tags podcast creators specify, no one category is identified as a “primary” category. Podcast creators are incentivized to select truthful category tags for their shows, since inaccurate tags can lead to their shows being removed from important venues, such as the iTunes store.


It's the first day of school! There are new friends to be made and plenty to explore! The teachers...

This episode is brought to you by Haunted Places: Urban Legends, a Parcast Original. For...

The story begins with a moan.

Figure 1.2: Podcast categories on Spotify, as displayed in the Podcasts section of the desktop app’s “Browse” pane.

In our dataset, there are as many as ten podcast category tags associated with any particular podcast. However, 68.23% of podcasts have only one category associated with them, and 97.50% of podcasts have three or fewer podcast categories associated with them. Figure 1.3 shows the distribution of categories per podcast for podcasts available on Spotify as of July 8th, 2019, as well as the fraction of all podcasts that have each category tag associated with them.[7] The most common podcast category is associated with 30.34% of all podcasts, whereas the least common category is associated with 0.20% of all podcasts.

We use the category tags associated with each user's podcast streams to quantify changes in the diversity of their podcast consumption. For streams of podcasts that have multiple category tags, we divide the stream evenly across each of the podcast's associated categories. For instance, the podcast "Trial By Stone: The Dark Crystal Podcast" has only one category associated with it: "Arts & Entertainment". On the other hand, the podcast "World's Best Parents" has four categories associated with it: "Comedy," "Educational", "Kids & Family," and "Stories". If a user streamed an episode of "Trial By Stone," this would count as one stream for the "Arts & Entertainment" category, whereas if a user streamed an episode of "World's Best Parents", this would count as 0.25 streams for each of the four categories with which "World's Best Parents" is associated.

1.4 Experiment Design

We analyze data from an experiment conducted on a sample of 852,937 premium Spotify users across seventeen countries[8] between April 18, 2019 and May 2, 2019 as part of a product rollout. In order to be eligible for the experiment, a user needed to have never streamed or followed a podcast on Spotify, and to have visited the "Home" section of the Spotify app during the experiment.

Users in both the treatment and control were exposed to a shelf in the "Home"

[7] Actual podcast category names are removed due to confidentiality concerns.

[8] The experiment was conducted on users located in AR, AU, BR, CA, CL, CO, DE, DK, ES, FR, GB, IT, MX, NL, NO, RS, and US.

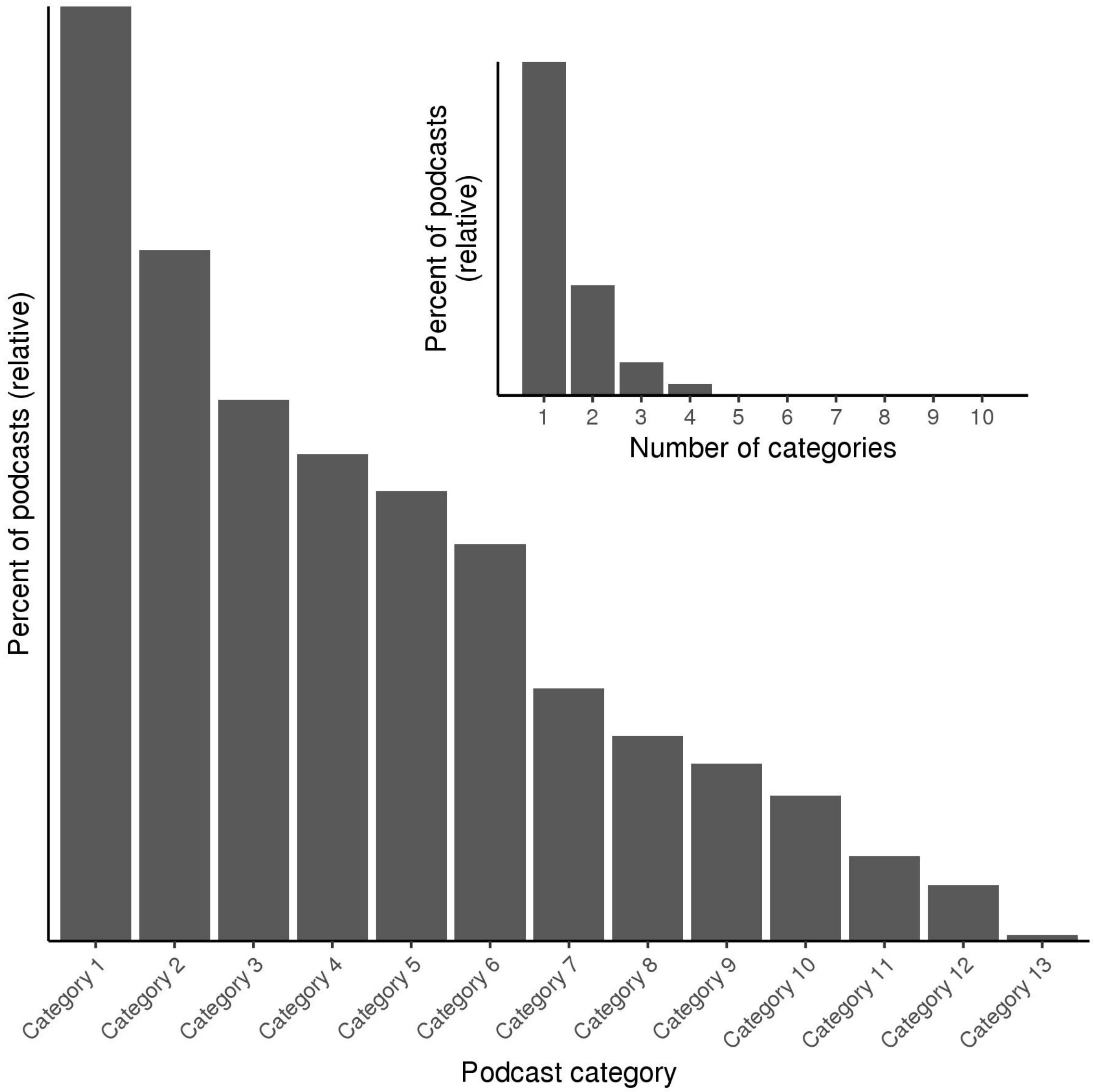


Figure 1.3: Histograms showing the frequency with which each podcast category is attached to a podcast, and the distribution of category tags per podcast. y-axis values and category names hidden due to confidentiality concerns.

section of the Spotify mobile app labeled “Podcasts to Try,” which was anchored in the second highest slot in the “Home” section. For users in the treatment, the “Podcasts to Try” shelf was populated with 10 recommendations generated by a neural network model that predicted the podcasts a user would follow based on their music listening history and demographic information.[9] For users in the control, the “Podcasts to Try” shelf was populated with the 10 most popular podcasts among users who shared the focal user’s self-reported gender, age bucket, and country.[10] Both the machine learned recommendations and the demographic-based recommendations were determined using pre-treatment data, and were not updated over the course of the experiment. For users in both treatment arms, the “Podcasts to Try” shelf was hidden once the user had streamed or followed any podcast on Spotify. Figure 1.1 shows a screenshot of the “Podcasts to try” shelf on iOS. The shelf’s UI was consistent across the control and treatment groups; the only thing exogenously varied was the set of podcasts populating the shelf.

Users were assigned to treatment arms using the following “bucket randomization” procedure. Every Spotify user was first assigned to one of ten thousand “buckets” based on a hash of their Spotify username. An equal number of these buckets were randomly assigned to the treatment and control conditions. Each user received the treatment corresponding to their bucket. A subset of buckets are also labeled as “long-term hold out” buckets, and are not included in Spotify experiments conducted on the “Home” section of the app. “Long-term hold out” buckets assigned to our treatment and control conditions were not shown the “Podcasts to Try” shelf, and are not included in our analysis. This assignment procedure resulted in 405,401 treatment users across 86 buckets, and 447,536 control users across 94 buckets. Critically, at the time of the experiment, Spotify did not create new user buckets each time an experiment was launched, which means that users within a given treatment assignment bucket share a treatment assignment history for previous experiments. To account for this, we report either cluster-robust standard errors or cluster bootstrap

[9] For a more detailed description of the neural network model, we refer the reader to Nazari et al. (2020).

[10] Age buckets are defined as follows: 18-24, 25-29, 30-34, 35-44, 45-54, 55+.

standard errors for all experiment analyses. In all cases where standard errors are bootstrapped, $n_{boot} = 1,000$. We report the balance of observable characteristics between the treatment and control groups in Appendix 1.A.

1.5 Results

In this section, we present the experiment results. We report the effects of the treatment on podcast streams, however, the effects of the treatment on podcast follows are extremely similar, and can be found in Appendix 1.B.

1.5.1 Effect on podcast consumption

We first study the impact of algorithmic podcast recommendations on the percentage of users that streamed at least one podcast, and on the average number of podcast streams per user during the experiment. Throughout the paper, we measure the effect of the treatment by estimating the following model:

$$y_i = \alpha + \beta T_i + \delta X_i + \epsilon_i, \tag{1.1}$$

where y_i is the outcome of interest for user i (in this case, either a binary outcome indicating whether the user streamed any podcasts during the experiment, or a count of the number of times the user streamed a podcast during the experiment), α is a constant, X_i is a vector of user-level covariates (age bucket, self-reported gender, and account age in days), and T_i is user i's treatment assignment. Standard errors are clustered at the user treatment assignment bucket level.

Figure 1.4 shows the distribution of podcast streams per user during the experiment in both treatment arms, both overall and conditional on the user streaming at least one podcast during the experiment. Table 1.1 reports the estimated effect of the treatment on the percentage of users streaming at least one podcast during the experiment, and Table 1.2 reports the estimated effect of the treatment on podcast streams per user during the experiment. In both cases, we report estimates obtained

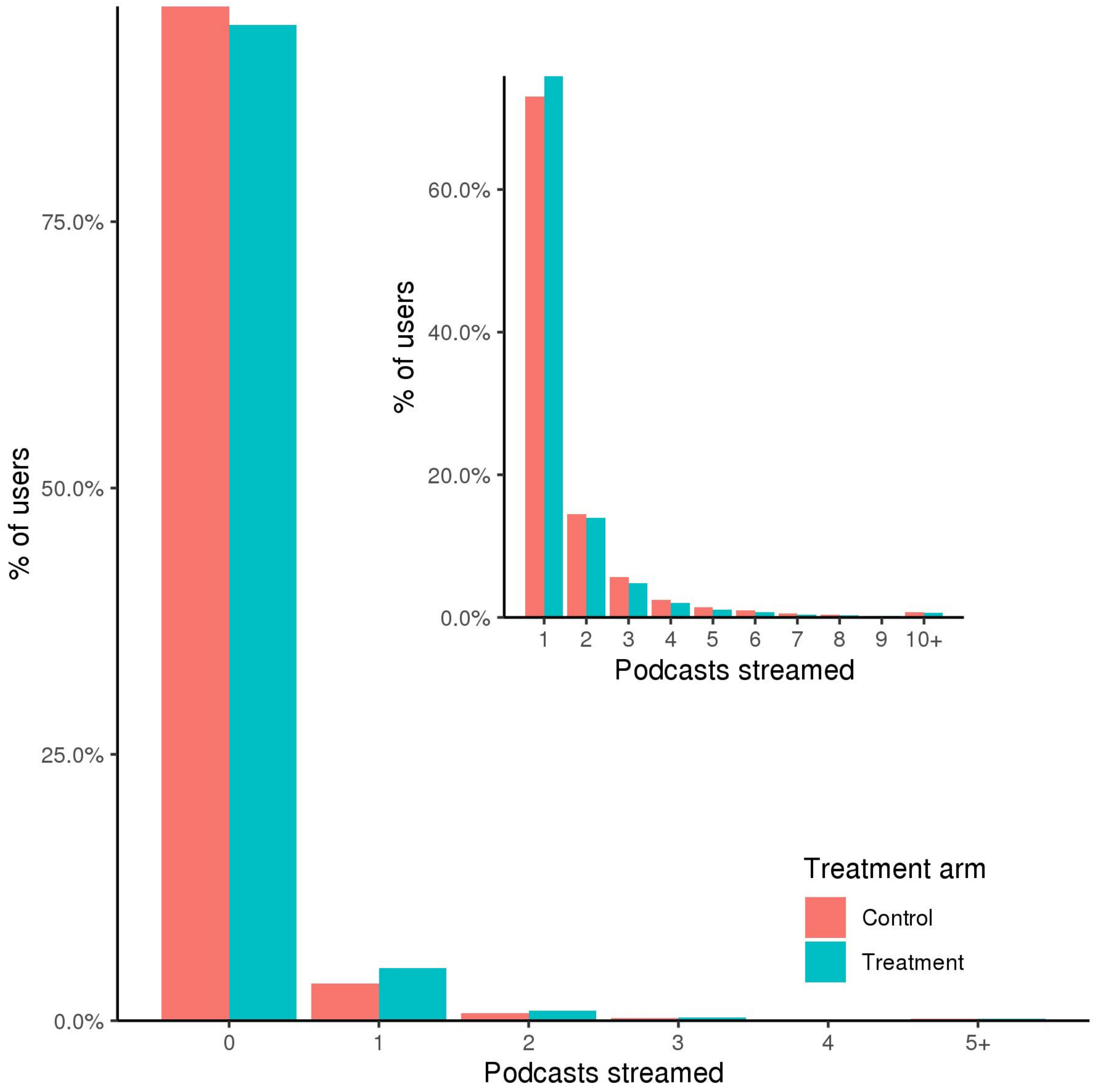


Figure 1.4: The distribution of of podcasts streamed in both treatment arms. Inset plot shows the distribution of podcasts streamed in both treatment arms conditional on streaming at least one podcast.

with and without controlling for user-level covariates. We find that the treatment increased the number of Spotify users streaming at least one podcast by 36.33% (± 3.01%), and increased the number of podcast streams per user by 28.90% (± 3.81%). This large treatment effect indicates that personalized podcast recommendations were extremely effective at increasing podcast consumption during the experiment, both in terms of unique podcast streamers and podcast streams per user.

Table 1.1: A linear probability model showing the effect of the treatment on streaming at least one podcast. Standard errors are clustered at the user bucket level.

	Dependent variable:	
	Streamed podcast	
	(1)	(2)
Treatment	0.017^{***}	0.017^{***}
	(0.001)	(0.001)
Constant	0.048^{***}	0.039^{***}
	(0.0004)	(0.001)
User Gender	No	Yes
User Age	No	Yes
User account age	No	Yes
Observations	852,937	852,937
R^2	0.001	0.004
Adjusted R^2	0.001	0.004
Residual Std. Error	0.230 (df = 852935)	0.230 (df = 852925)
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

In order to determine the extent to which the increase in podcast streams per user is driven by compositional shifts, as opposed to intensity shifts, we use the principal stratification approach detailed by Frangakis and Rubin (2002) and Ding and Lu (2017). This method allows us to estimate the causal effect of the treatment for two latent subpopulations: “always takers” (i.e., users who would have streamed a podcast whether in the control or treatment) and “compliers” (i.e., users who would have streamed a podcast if in the treatment, but not if in the control). The principal stratification methodology is detailed in Appendix 1.C. We estimate that on average,

Table 1.2: A linear model showing the effect of the treatment on number of podcasts streamed. Standard errors are clustered at the user bucket level.

	Dependent variable:	
	Podcasts streamed	
	(1)	(2)
Treatment	0.022^{***}	0.022^{***}
	(0.001)	(0.001)
Constant	0.077^{***}	0.056^{***}
	(0.001)	(0.001)
User Gender	No	Yes
User Age	No	Yes
User account age	No	Yes
Observations	852,937	852,937
R^2	0.0005	0.003
Adjusted R^2	0.0005	0.003
Residual Std. Error	0.508 (df = 852935)	0.508 (df = 852925)
Note:		*p<0.1; **p<0.05; ***p<0.01

“compliers” streamed 1.505 (95% CI: (1.488, 1.522)) more podcasts in the treatment, whereas “always takers” streamed 0.082 (95% CI: (0.055, 0.112)) *fewer* podcasts in the treatment. In other words, the observed increase in number of podcast streams per user is driven entirely by an increase in podcast streaming adoption, as opposed to an increase in the number podcasts that Spotify users stream conditional on streaming at least one podcast.

1.5.2 Effect on diversity of podcast consumption

We also measure the effect of the treatment on the diversity of content that individual users consume (henceforth referred to as “individual-level diversity”) and the diversity of content consumption across users (henceforth referred to as “intragroup diversity”).

Individual-level diversity

We quantify individual-level diversity using the Shannon entropy (Shannon 1948).[11] The Shannon entropy of user i's streams is defined as

$$th_i = -\sum_{c \in C} s_{ci} \ln(s_{ci}), \tag{1.2}$$

where C is the full set of podcast categories and s_{ci} is the share of user i's streaming coming from category c. Note that if a user did not stream any podcasts belonging to category c, that podcast category's contribution to the Shannon entropy is zero. Importantly, this also means that users who did not listen to *any* podcasts during the experiment have a Shannon entropy of zero. This, along with the fact that the treatment had a large, positive effect on the number of users streaming podcasts, could cause the observed effect of the treatment on Shannon entropy across all users to be positive, even if consumption conditional on streaming became less diverse.[12]

To account for this, we analyze the data in two ways. First, we estimate the model in Equation 1.1 for the subset of users that streamed at least one podcast during the experiment. Results of this analysis cannot be interpreted as causal, since we are conditioning on a post-treatment variable (streaming at least one podcast). Nonetheless, these results provide some insight into the extent to which increased recommendation personalization changed individual-level diversity. Figure 1.5 shows the histogram of user-level Shannon entropy for podcast streams in both treatment arms, and Table 1.3 reports the difference in the average streaming user's Shannon entropy, both with and without controlling for user-level covariates. We find that the average Shannon entropy of podcast streams among users who streamed at least one podcast was 11.51% (±1.08%) lower in the treatment. Second, we again employ the principal stratification framework described by Frangakis and Rubin (2002) and Ding and Lu (2017) to estimate the causal effect of the treatment on individual-

[11] The Shannon entropy is also sometimes referred to as the Teachman index (Teachman 1980).

[12] This is, in fact, what we observe in our data. When including all users in our analysis, we find that the treatment increased the average Shannon entropy by 21.16% (±2.89%). These results are reported in Table 1.F.1.

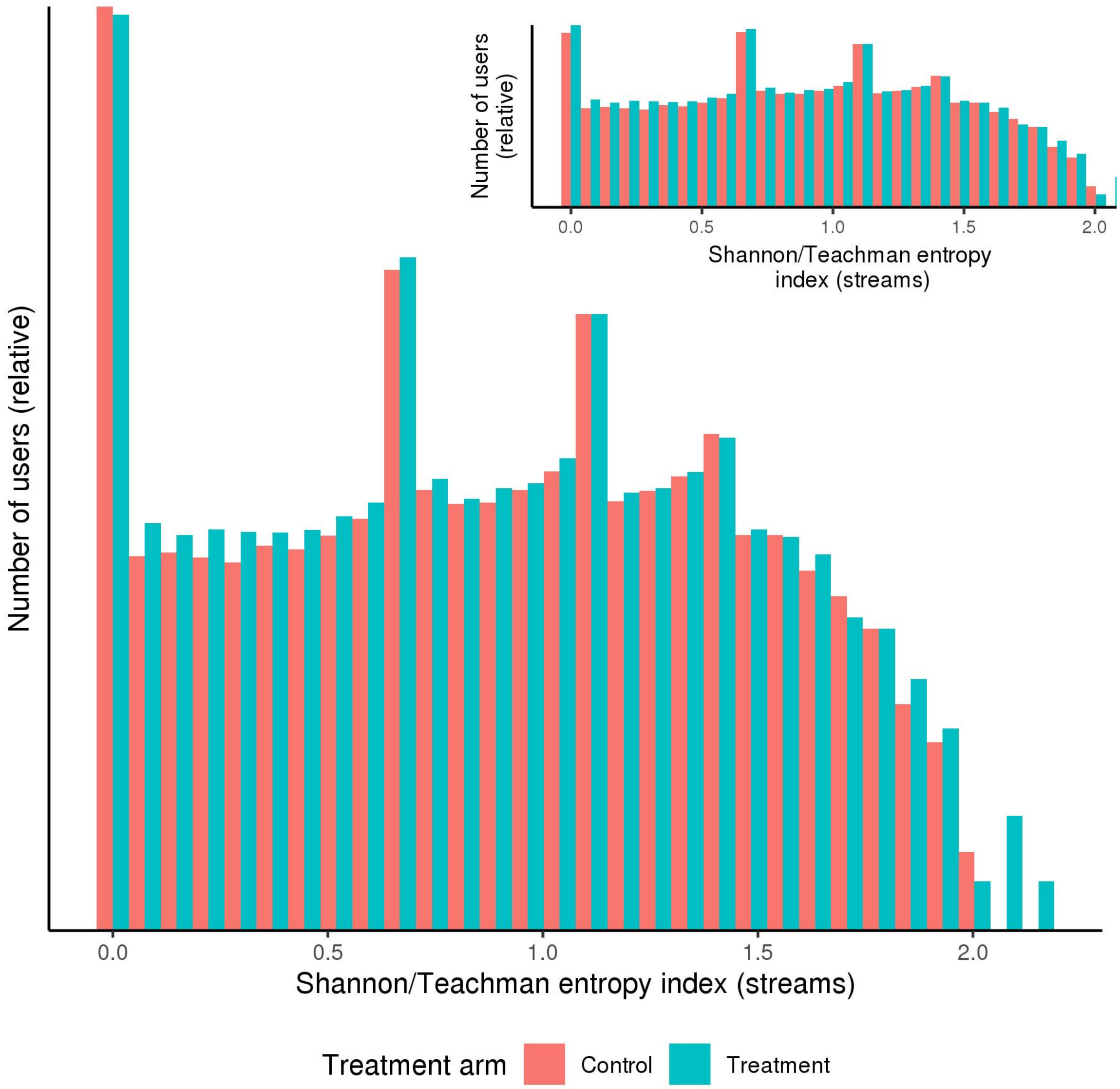


Figure 1.5: The distribution of the user-level diversity for streams in both treatment arms. Inset plot shows the distribution of user-level diversity in both treatment arms conditional on streaming at least one podcast. y-axis values are on a log scale, and are hidden due to confidentiality concerns.

Table 1.3: A linear model showing the effect of the treatment on the Shannon/Teachman entropy index (streams) (podcast streamers only). Standard errors are clustered at the user bucket level.

	Dependent variable:	
	Shannon/Teachman entropy index (streams)	
	(1)	(2)
Treatment	-0.071^{***}	-0.070^{***}
	(0.004)	(0.004)
Constant	0.620^{***}	0.670^{***}
	(0.002)	(0.005)
User Gender	No	Yes
User Age	No	Yes
User account age	No	Yes
R^2	0.005	0.016
Adjusted R^2	0.005	0.016
Residual Std. Error	0.479 (df = 76191)	0.477 (df = 76181)
Note:		*p<0.1; **p<0.05; ***p<0.01 Observation counts hidden due to confidentiality concerns

level diversity for the subset of users who are "always takers." Consistent with the previously reported non-causal findings, we estimate that treatment decreased the average Shannon entropy for streaming "always takers" by 11.29% (95% CI: (10.00%, 12.26%)).

The fact that higher levels of recommendation personalization decreased the average Shannon entropy for "always takers" indicates that the treatment made users' podcast consumption *more homogenous* with respect to podcast categories. Our analysis cannot identify to what extent this difference is driven by treatment users streaming podcasts that had fewer podcast categories associated with them. However, insofar as podcast categories accurately capture information about the topics covered in a particular show, it is reasonable to assume that a user who listened to podcasts with fewer category tags conditional on streaming a particular number of podcasts consumed less diverse content. Figure 1.6 shows the distributions of user-level Shannon entropy in both treatment arms conditional on streaming a particular number of podcasts during the experiment.

Intragroup diversity

We quantify intragroup diversity using a mathematical expression introduced by Aral and Van Alstyne (2011):

$$ID = \frac{1}{n_c} \sum_{j=1}^{n_c} \left[1 - \cos\left(\Gamma_j, \bar{\Gamma}\right)\right]^2, \tag{1.3}$$

where n_c is the number of users consuming at least one podcast, Γ_j is a vector describing the fraction of user j's listening belonging to each podcast category, and $\bar{\Gamma}$ is the average of Γ_j across all users streaming at least one podcast. Intuitively, ID measures the variance of all streaming users' individual-level podcast category consumption vectors. We calculate ID separately for the control and treatment groups, and test for a statistically significant difference.[13]

[13] Calculating ID requires that we restrict our analysis to the subset of control and treatment users who streamed at least one podcast during the experiment. However, since ID is a population-level outcome, as opposed to a user-level outcome, we claim that our estimates can be interpreted

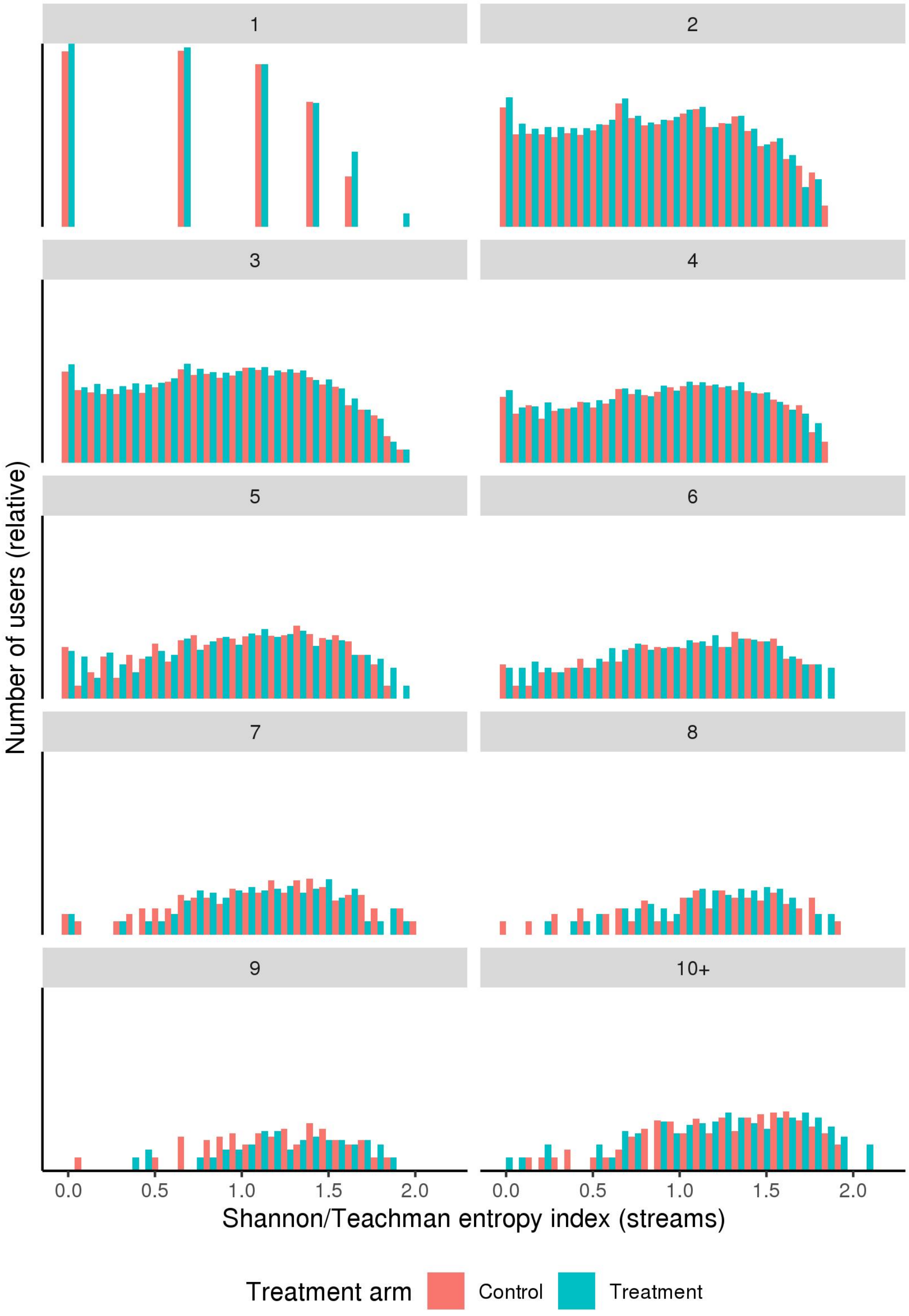


Figure 1.6: The distribution of the user-level diversity for streams in both treatment arms conditional on streaming a set number of podcasts during the experiment. y-axis values are on a log scale, and are hidden due to confidentiality concerns.

We find that the treatment increased the intragroup diversity for podcast streams by 5.96% (95% CI: 5.45%, 6.44%)), from 0.710 (95% CI: (0.708, 0.713)) in the control group to 0.753 (95% CI: (0.751, 0.754)). In other words, not only did increased recommendation personalization push podcast streamers to consume more homogenous content, it also pushed podcast streamers to listen to content that was *more* dissimilar to the content that other streamers listened to.

1.5.3 Treatment effects by stream referrer

In this subsection, we present the effects of the treatment on streams originating from different sections of the Spotify app. Because the treatment only directly affected the podcasts that were displayed on "Home," stream referrer-level treatment effects provide insight into the extent to which exposure to personalized content recommendations impacted the types of podcasts that users sought out organically. If exposure to recommendations *did* change what users sought out organically, we would expect the treatment to impact what users stream from other parts of Spotify's app, such as "Search" and "Your Library." As a result, we would observe treatment effects for streams originating from both "Home" and from non-home surfaces. On the other hand, if the treatment did not change what users sought out organically (i.e., treatment effects are entirely driven by what users consume when streaming recommended content on "Home"), we would expect to see treatment effects for "Home" streams, but no treatment effects for non-home streams.

Podcast consumption

Figure 1.7 shows the distribution of podcast streams per user on both types of referral surfaces in both treatment arms. Table 1.4 reports the estimated effect of the treatment on the number of users streaming at least one podcast from each type of referral surface during the experiment, and Table 1.5 reports the estimated effect of the treatment on podcast streams per user from each type of referral surface during

causally, and that the set of users that select into podcast streaming is one factor that contributes to the population-level potential outcomes for ID.

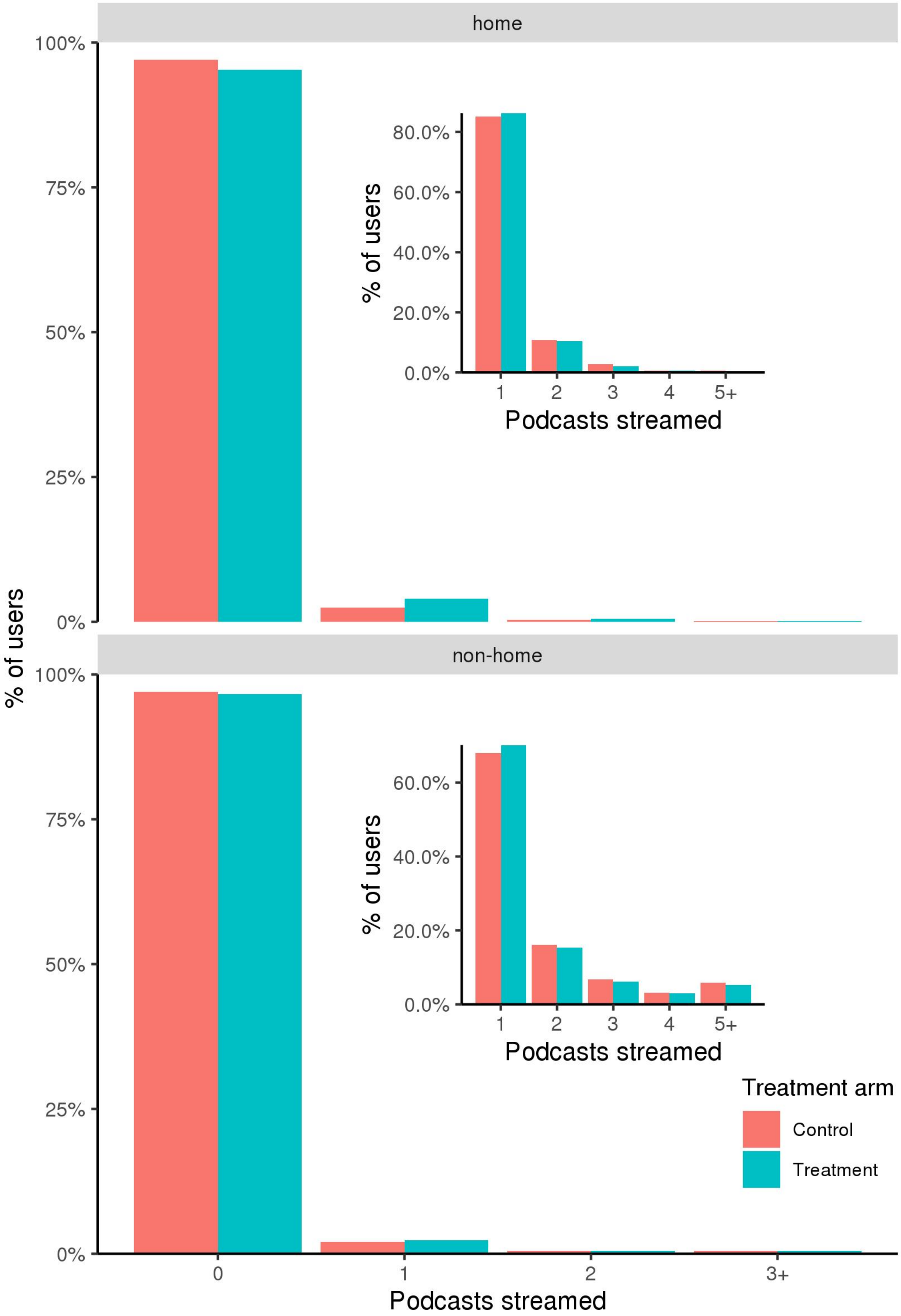


Figure 1.7: The distribution of podcasts streamed in both treatment arms by stream referrer. Inset plots shows the distribution of podcasts streamed by referrer in both treatment arms conditional on streaming at least one podcast.

Table 1.4: A linear model showing the effect of the treatment on streaming at least one podcast, both on and off of home. Standard errors are clustered at the user bucket level.

	Dependent variable:			
	Streamed podcast			
	Home		Non-home	
	(1)	(2)	(3)	(4)
Treatment	0.017^{***}	0.017^{***}	0.004^{***}	0.004^{***}
	(0.001)	(0.001)	(0.0004)	(0.0004)
Constant	0.029^{***}	0.023^{***}	0.030^{***}	0.026^{***}
	(0.0003)	(0.001)	(0.0003)	(0.001)
User Gender	No	Yes	No	Yes
User Age	No	Yes	No	Yes
User account age	No	Yes	No	Yes
Observations	852,937	852,937	852,937	852,937
R^2	0.002	0.003	0.0001	0.004
Adjusted R^2	0.002	0.003	0.0001	0.004
Residual Std. Error	0.190 (df = 852935)	0.190 (df = 852925)	0.175 (df = 852935)	0.175 (df = 852925)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 1.5: A linear model showing the effect of the treatment on number of podcasts streamed, both on and off of home. Standard errors are clustered at the user bucket level.

	Dependent variable:			
	Podcasts streamed			
	Home		Non-home	
	(1)	(2)	(3)	(4)
Treatment	0.020^{***}	0.020^{***}	0.006^{***}	0.006^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.036^{***}	0.027^{***}	0.053^{***}	0.038^{***}
	(0.0005)	(0.001)	(0.001)	(0.001)
User Gender	No	Yes	No	Yes
User Age	No	Yes	No	Yes
User account age	No	Yes	No	Yes
Observations	852,937	852,937	852,937	852,937
R^2	0.001	0.003	0.00004	0.002
Adjusted R^2	0.001	0.003	0.00004	0.002
Residual Std. Error	0.260 (df = 852935)	0.259 (df = 852925)	0.447 (df = 852935)	0.446 (df = 852925)
Note:				*p<0.1; **p<0.05; ***p<0.01

the experiment. We find that the treatment increased the number of Spotify users streaming podcasts from "Home" by 59.17% (±4.58%), and the number of podcast streams per user from "Home" by 55.75% (±5.07%). In contrast, we find that the treatment increased the number of users streaming podcasts from non-home surfaces by 12.55% (±2.94%) and the average number of podcast streams per user from non-home surfaces by 10.47% (±4.16%). In other words, the treatment not only caused an increase in podcast streaming behavior from the "Home" section of the Spotify app, but also led to a (smaller) increase in the amount of podcast streaming behavior from other sections of the app.

Individual-level diversity

Table 1.6: A linear model showing the difference in the average Shannon/Teachman entropy index (streams) by stream referral source (podcast streamers only). Standard errors are clustered at the user bucket level.

	Dependent variable:			
	Shannon/Teachman entropy index (streams)			
	Home		Non-home	
	(1)	(2)	(3)	(4)
Treatment	-0.116^{***}	-0.116^{***}	-0.018^{***}	-0.017^{***}
	(0.004)	(0.004)	(0.005)	(0.005)
Constant	0.654^{***}	0.730^{***}	0.552^{***}	0.562^{***}
	(0.003)	(0.006)	(0.003)	(0.007)
User Gender	No	Yes	No	Yes
User Age	No	Yes	No	Yes
User account age	No	Yes	No	Yes
R^2	0.016	0.029	0.0003	0.014
Adjusted R^2	0.016	0.029	0.0003	0.014
Residual Std. Error	0.456 (df = 54335)	0.453 (df = 54325)	0.500 (df = 36327)	0.497 (df = 36317)

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$
Observation counts hidden due to confidentiality concerns

Figure 1.8 shows histograms of the user-level Shannon entropy for podcast streams across both types of referral surface for users in both treatment arms, and Table 1.6 reports the differences in the average referrer-specific, user-level Shannon entropy for the subsample of users streaming at least one podcast from a given surface type during the experiment, both with and without controlling for user-level covariates. We find that for users streaming at least one podcast from "Home," the average

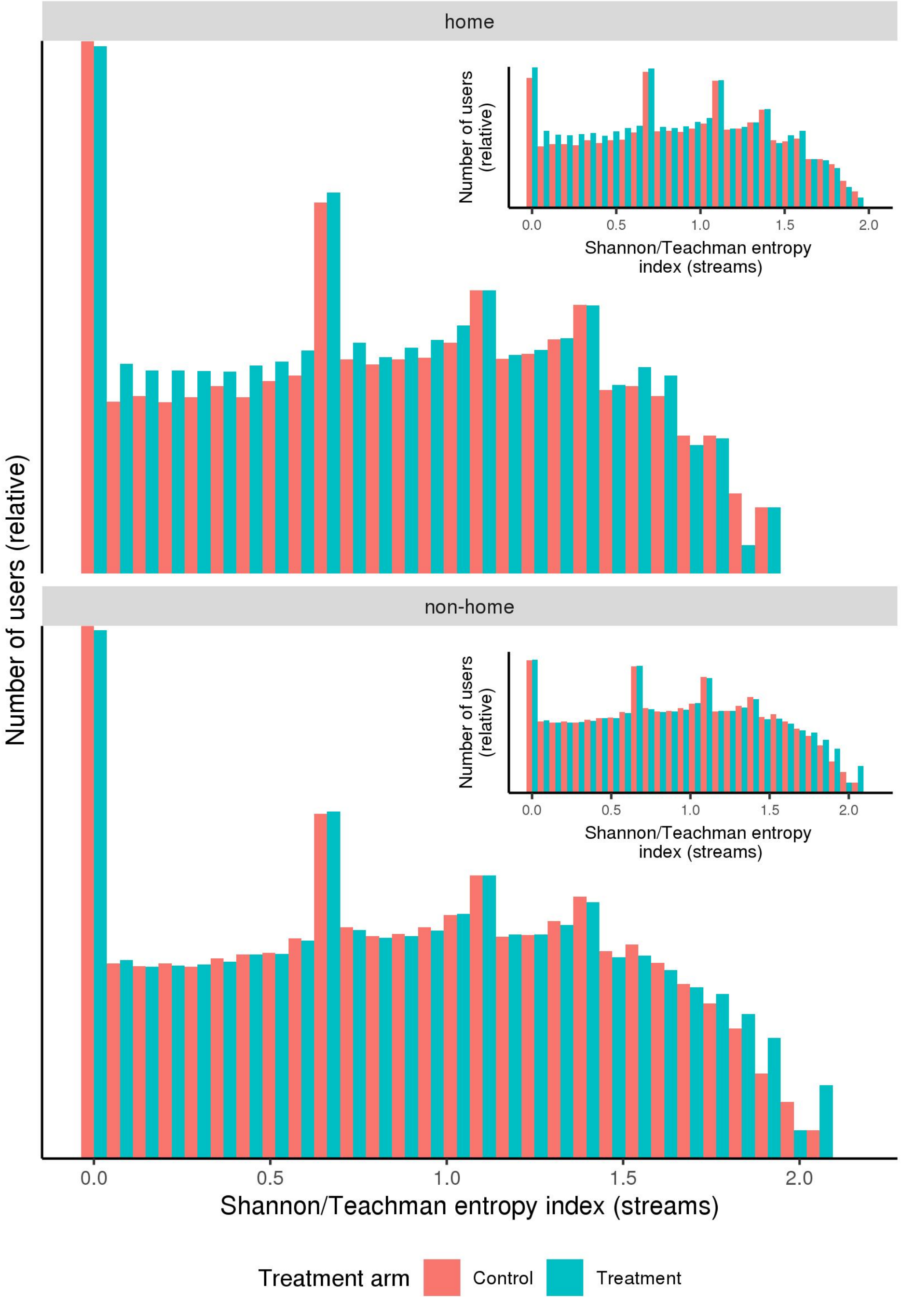


Figure 1.8: The distribution of individual-level diversity for podcast streams in both treatment arms by stream referrer. Inset plots show the distribution of individual-level diversity by referrer in both treatment arms conditional on streaming at least one podcast. y-axis values are on a log scale, and are hidden due to confidentiality concerns.

Shannon entropy of “Home” streams was 17.70% (± 1.10%) lower in the treatment group, and that for users streaming at least one podcast from a section other than “Home,” the average Shannon entropy of non-home streams was 3.31% (± 1.77%) lower in the treatment group. These results indicate that individual-level diversity in the treatment group was not only lower for streams originating from “Home,” but also for streams coming from other sections of the Spotify app.[14][15]

Intragroup diversity

We find that on “Home,” the intragroup diversity increased by 14.04% (95% CI: (13.34%, 14.66%), from 0.654 (95% CI: 0.650, 0.657) in the control group to 0.746 (95% CI: (0.744, 0.748)). We find that on non-home surfaces, the intragroup diversity increased by 0.040% (95% CI: (-0.19%, 0.96%), from 0.769 (95% CI: (0.765, 0.771)) in the control group to 0.772 (95% CI: (0.769, 0.775)). In other words, while we do find evidence of an increase in intragroup diversity for streams originating on “Home,” we do not find statistically significant evidence of an increase in intragroup diversity for non-home streams.

1.5.4 Long-term treatment effects

In this subsection, we use data collected between May 3, 2019 and July 17, 2019 to test for longer-term effects of the treatment. We repeat our main analyses on cross-sectional datasets that describe users' behavior over time intervals spanning from 3rd of the month to the 17th of the month, and from the 18th of the month to the 2nd of the month. Testing for long-term effects allows us to determine whether short-term exposure to personalized podcast recommendations has a lasting impact on the types of content that users consume, or if users revert to their counterfactual

[14] As was the case for overall effects, the effects of the treatment on referrer-specific individual-level diversity are positive when we estimate Equation 1.1 for the entire sample. We find that the treatment increased the average Shannon entropy of home streams by 29.83% (± 3.76%), and increased the average Shannon entropy of non-home streams by 7.36% (± 3.19%. These results are reported in Table 1.F.2.

[15] Our referrer-specific individual-level diversity results cannot be interpreted causally, since we are conditioning on a post-treatment variable (streaming at least one podcast from a particular referrer.

baseline podcast listening once individually personalized recommendations are no longer shown.

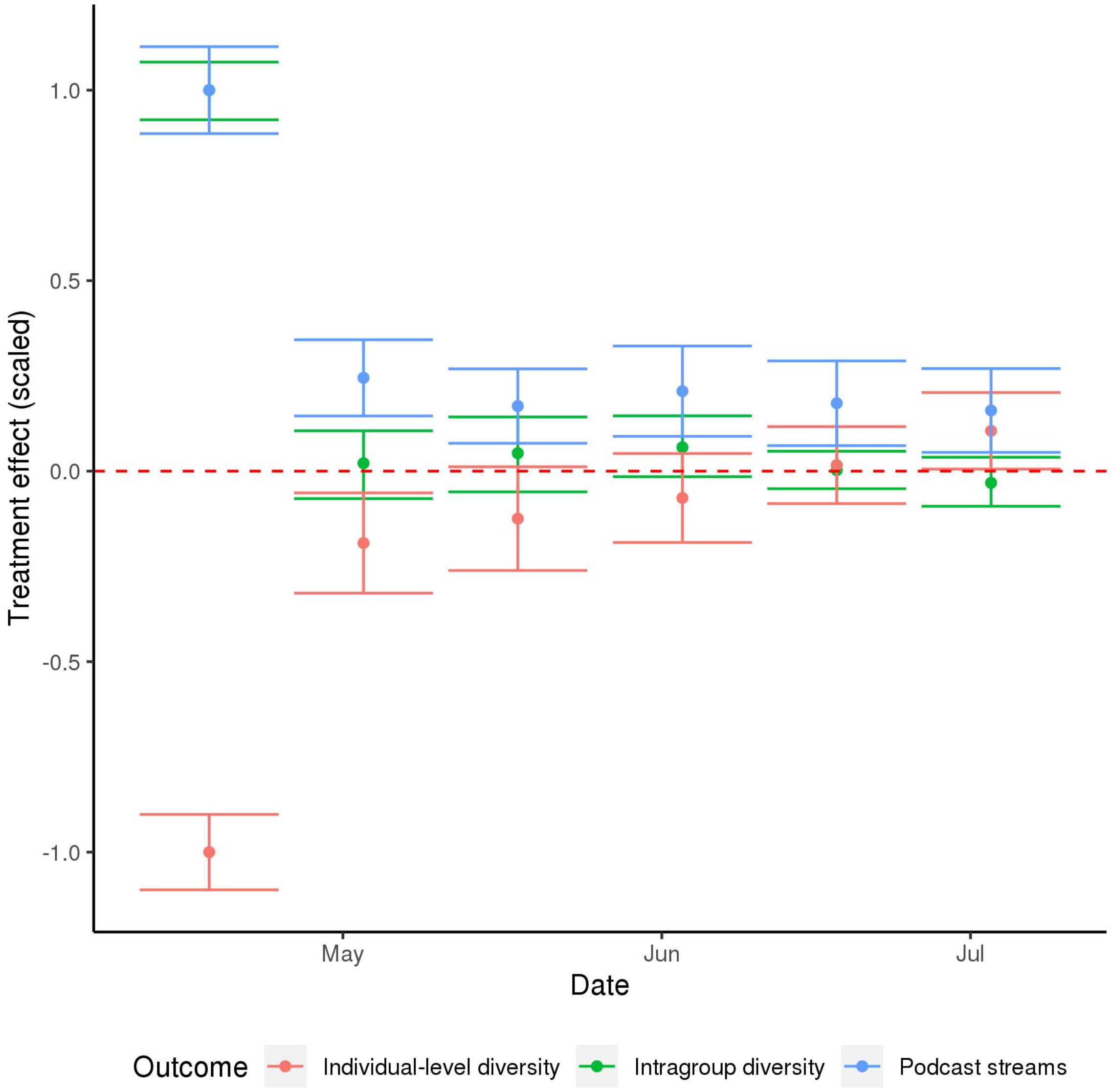


Figure 1.9: The long-term effect of the treatment on podcast streams per user, individual-level streaming diversity conditional on streaming at least one podcast, and intragroup streaming diversity. Each outcome's time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.

Figure 1.9 shows the long-term effect of the treatment on the average number of podcast streams per user, the average Shannon entropy of podcast streams conditional on streaming at least one podcast, and the intragroup diversity of podcast

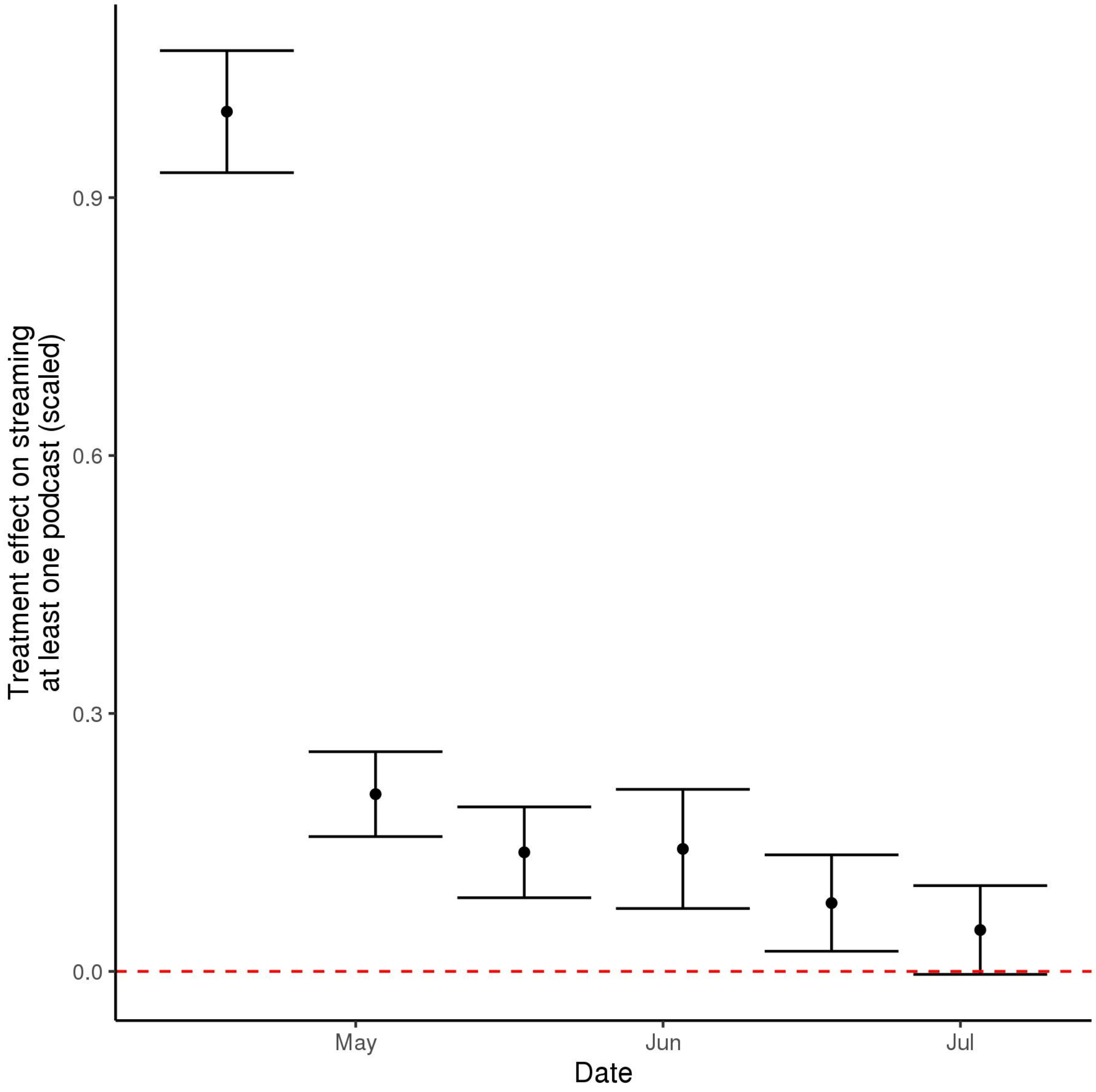


Figure 1.10: The long-term effect of the treatment on the percentage of users streaming at least one podcast. The time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.

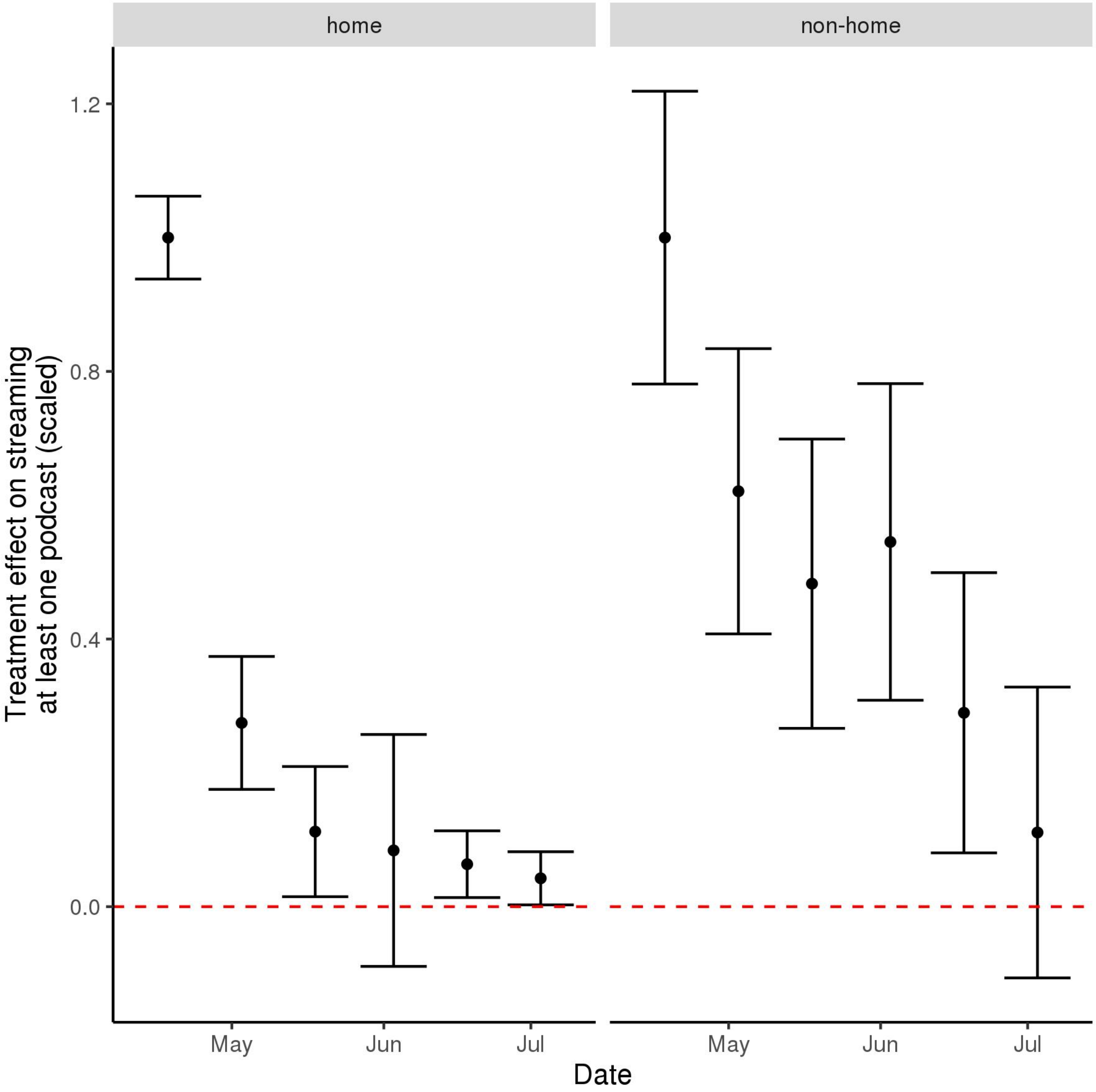


Figure 1.11: The long-term referrer-level effect of the treatment on the percentage of users streaming at least one podcast over time. Each time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.

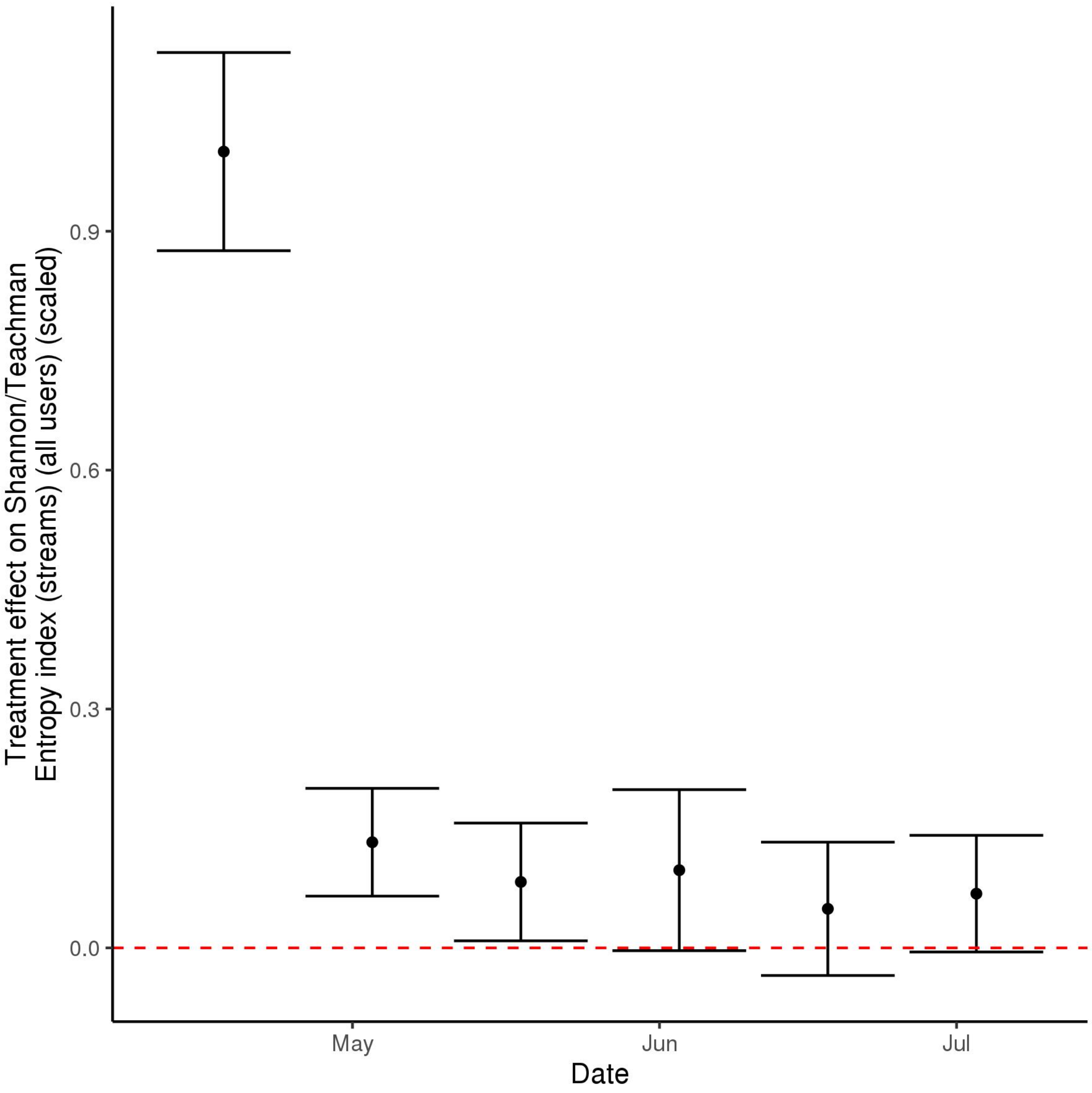


Figure 1.12: The long-term effect of the treatment on the average user-level Shannon entropy for streams. The time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.

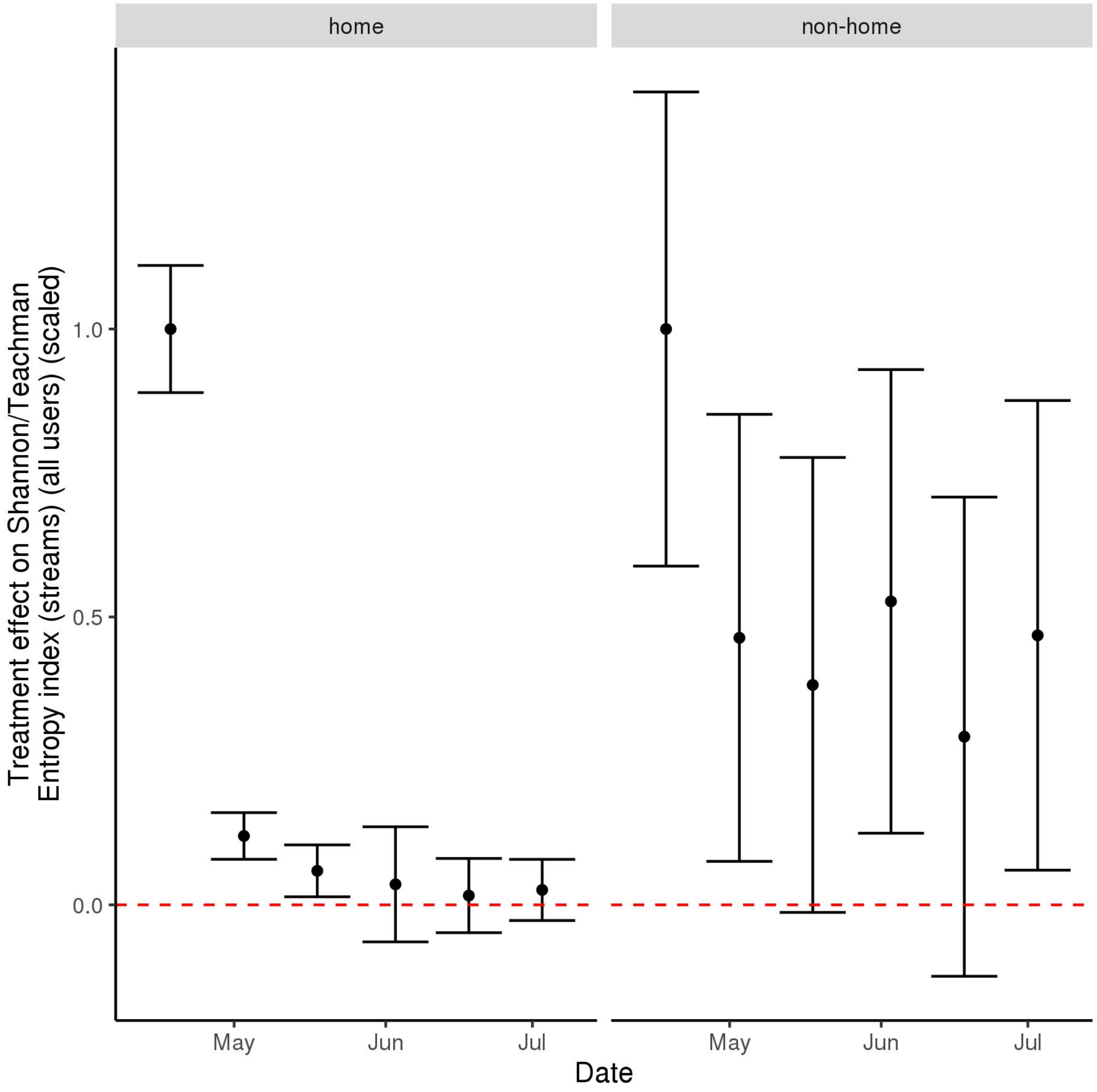


Figure 1.13: The long-term referrer-level effect of the treatment on the average user-level Shannon entropy for streams over time. Each time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.

streams. Across all of these outcomes, we observe the same trend: the large treatment effects observed during the experiment quickly shrink in magnitude, and in some cases disappear entirely, once the experiment has concluded.[16] We also measure the long-term effect of the treatment on podcast streams originating from different referral surfaces. This allows us to identify potential heterogeneity in the extent to which short-term exposure to personalized podcast recommendations has a long-term effect on consumption habits across both recommended listening and organic listening. Figure 1.14 shows the long-term effect of the treatment on average podcast streams per user, Shannon entropy for streams conditional on streaming at least one podcast, and intragroup diversity for streams originating from both home and non-home surfaces. Stream referrer-level treatment effects follow the same trend as overall effects, and this trend does not vary by stream referrer; treatment effects dissipate quickly, or disappear entirely, once the experiment has ended. The lack of long-term treatment effects suggests that short-term exposure to personalized podcast recommendations does not affect long-term listening behavior through algorithmic spillovers or through changes in what users seek out organically.[17]

1.6 Discussion

We find that personalized recommendations not only increased content consumption, but also increased the homogeneity of content consumed by individual users and increased the diversity of content consumed across users. These results suggest that an "engagement-diversity trade-off" can exist for firms that utilize personalization algorithms and recommendation systems to increase engagement and/or sales. This trade-off has multiple managerial implications. First, Anderson et al. (2020) find

[16] The treatment effects observed during the experiment also quickly dissipate for number of users streaming at least one podcast (Figures 1.10 and 1.11) and for the average individual-level diversity of podcast streams measured across all users in the experiment (Figures 1.12 and 1.13), both overall and conditional on streaming surface.

[17] The number of podcast streams per user over time was dependent on users being exposed to podcast content in the "Home" section of the Spotify app. However, the number of impressions that podcast content received on "Home" varied considerably in the months following the experiment. The reasons for this are described in Appendix 1.D.

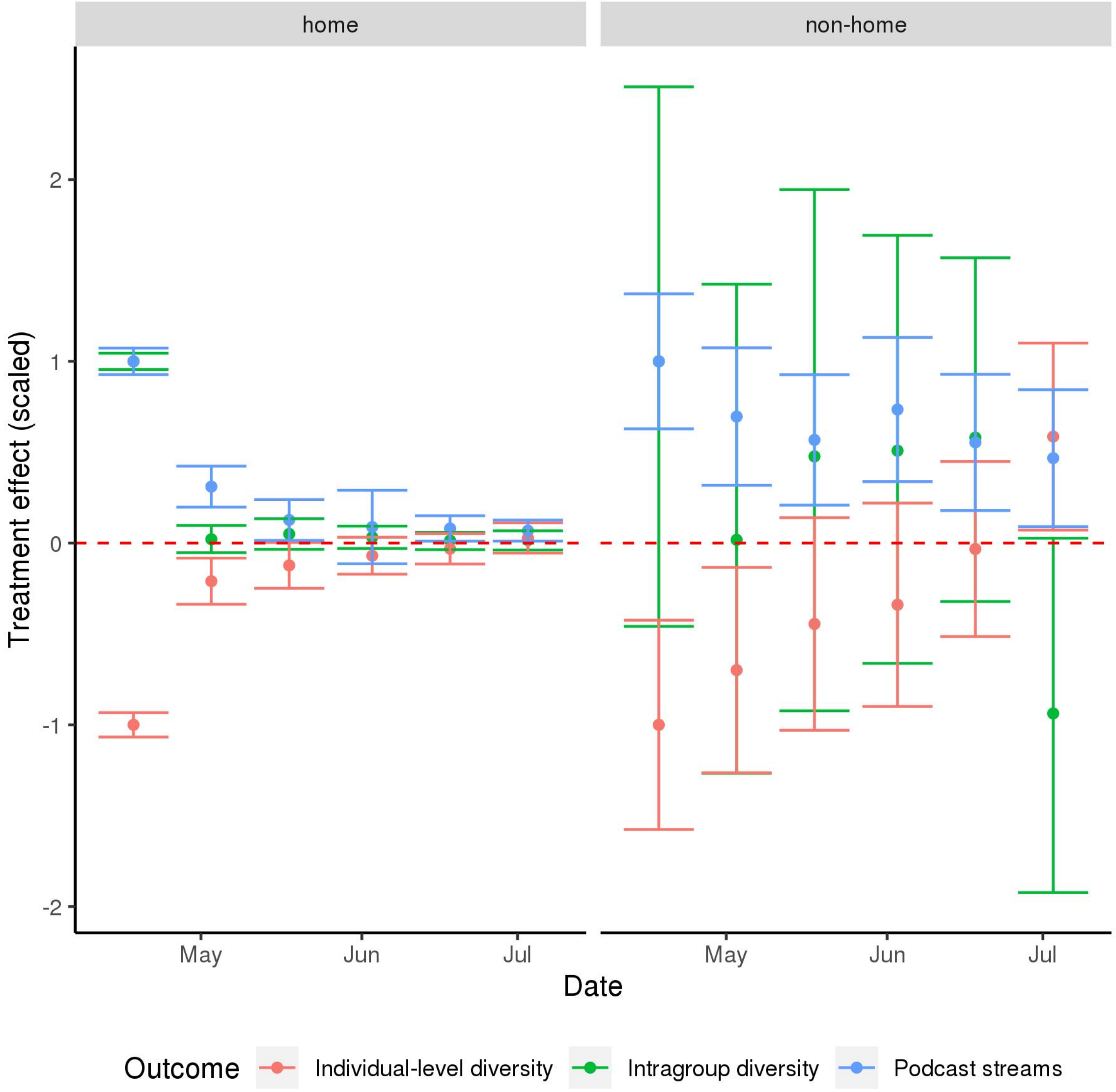


Figure 1.14: The long-term referrer-level effect of the treatment on podcast streams per user, individual-level streaming diversity conditional on streaming at least one podcast, and intragroup streaming diversity. Each outcome's time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.

that higher levels of individual-level diversity are associated with lower churn rates and higher rates of premium service subscriptions. If this relationship is causal, this would suggest that short-term increases in engagement/sales arising from the use of recommendation systems can have a neutral or even negative long-term effect on revenue. Second, the fact that recommendation systems can decrease individual-level diversity, but increase aggregate diversity may affect the optimal strategy for content creators, including platforms that produce their own original content (e.g., Spotify, Netflix). Depending on the diversity of content that users consume, content creators may find it optimal to produce large amounts of low-budget, niche content, or a small amount of high-budget content with mass appeal. Finally, in this paper, we measure the effect of increased personalization on consumption diversity measured with respect to podcast categories. However, it's possible that our analytical framework, if applied to data with ideological labels, would yield similar results. If this is the case, when the content delivered by a platform is ideological and/or extreme in nature, recommender systems that increase short term firm revenue could also create costs for firms due to the high level of public scrutiny given to personalized recommendations, and impact the nature of public discourse through the creation of "filter bubbles."

Our results also shed light on the effect that exposure to personalized recommendations has on the types of content that users seek out organically. Although we observe stronger treatment effects on streams originating from the "Home" section of Spotify's app, the treatment did affect the volume and individual-level diversity of content that users seek out organically in other sections of app. These results suggest that personalized recommendation algorithms have the potential to affect users' preferences, and may play a role in Balkanizing online content consumption. However, we do not detect long-term changes in the diversity of Spotify users' podcast consumption choices after short-term exposure to personalized recommendations. This suggests that firms can "course correct" if they discover that personalized recommendations are impacting users' consumption patterns in undesirable ways.

While Lee and Hosanagar (2019) find that recommender systems have a neutral-to-positive effect on individual-level diversity and decrease aggregate diversity, we, like

Claussen et al. (2019), find the opposite: in our setting, personalized recommendations *decreased* individual-level diversity and *increased* aggregate diversity. We believe there are multiple reasons this may be the case. First, as argued by Lee and Hosanagar (2019), the effect of recommender systems is likely dependent on both the particular algorithm used and the setting in which it is deployed. Second, previous economics and management research (e.g., Brynjolfsson et al. (2011), Lee and Hosanagar (2019)) has typically measured changes in sales diversity, whereas we measure changes in the distribution of content categories consumed.[18] Given that recommender systems have become a common feature of content platforms, we believe it is important to measure the impact of recommender systems not just on market concentration, but also on the *types* of items that users engage with. Overall, the contrast between previous findings and ours underscores the need to study the effects of many recommendation algorithms, in many contexts, using many different measures of diversity.

Our results suggest multiple interesting extensions. First, while our experiment enables us to measure the effect of short-term exposure to personalized recommendations, we are unable to measure the impact of long-term exposure to personalized recommendations. Long-term exposure may affect content consumption and diversity differently. Second, while category tags provide a coarse sense of the type of content users are consuming, there are other important ways to quantify product diversity. For instance, it may be helpful to measure category similarity, the political skew of a piece of content, or the "extremity" of a piece of content. Third, it would be worthwhile to more explicitly consider the optimal strategy of a content producer in the presence of recommender systems that affect consumption diversity. Finally, in this paper we study personalized recommendations that are solely optimized for engagement. This single objective approach to personalization is common in practice, and our findings suggest that researchers should continue to develop personalization techniques that explicitly take into account the diversity of content recommended to users (Marler and Arora 2004, Castells et al. 2015, Lacerda 2017).

[18] In Appendix 1.E, we report the effect of the experiment on the "sales diversity" for podcast consumption.

1.7 Conclusion

In this paper, we analyze data from a randomized field experiment and measure the effect of personalized content recommendations not just on the *amount* of content that people consumed, but also on the *diversity* of content that people consumed. We find evidence that an "engagement-diversity trade-off" can exist for firms when recommendations are optimized solely to drive engagement. While more personalized recommendations increased user engagement, they also decreased the diversity of content that individual users consumed, while simultaneously *increasing* the degree of dissimilarity across users. These shifts in content consumption patterns can negatively impact the rate of churn and average lifetime value for users, and also impact the optimal strategy for content creators. We also find evidence that exposure to personalized content recommendations impacted the types of content that users sought out organically. At first glance, our results are at tension with some recent studies of recommender systems, such as Lee and Hosanagar (2019). However, we believe this contrast highlights the need for further experimental studies of recommender systems across a multitude of different business settings and algorithm specifications, as well as the need to develop new measures for quantifying the effect of recommender systems. Furthermore, we believe our results underscore the need for researchers to continue developing approaches to personalization that optimize jointly for user engagement and consumption diversity.

Appendix

1.A Experiment Balance Checks

Table 1.A.1: User bucket-level summary statistics for buckets in both the control and treatment arms of the experiment. *p*-values are computed using the Wilcoxon rank-sum test.

Metric	Mean (control)	SD (control)	Mean (treatment)	SD (treatment)	*p*-value	Sig.
Number of users	4761.021	82.808	4713.965	89.633	$< .001$	***
% of users age 18 - 24	0.342	0.007	0.339	0.008	0.077	*
% of users age 25 - 29	0.209	0.005	0.209	0.006	0.955	
% of users age 30 - 34	0.131	0.005	0.132	0.005	0.139	
% of users age 35 - 44	0.155	0.005	0.156	0.005	0.368	
% of users age 45 - 54	0.102	0.004	0.103	0.005	0.291	
% of users age 55+	0.055	0.003	0.055	0.003	0.542	
% of users of unknown age	0.006	0.001	0.006	0.001	0.4	
% of male users	0.537	0.006	0.538	0.007	0.322	
% of female users	0.454	0.006	0.453	0.007	0.291	
% of users with other gender	0.005	0.001	0.005	0.001	0.409	
% of users with gender unknown	0.004	0.001	0.004	0.001	0.967	
Average mean account age (days)	1285.698	11.569	1284.711	11.435	0.412	

Table 1.A.1 shows bucket-level summary statistics for user buckets in both the treatment and control conditions, and tests for statistically significant differences between them. With the exception of average number of exposed users per bucket, we do not find statistically significant differences between the control group and treatment group for any of the specified user-level covariates. We believe that the smaller number of exposed users per bucket in the treatment group is driven by random errors in the generation of recommendations using the neural network-based model.

1.B Podcast Follows Analysis

In this section, we repeat our analyses for podcast follows, as opposed to podcast streams. Because our results for podcast follows are extremely similar to those for podcast streams, we elect not to conduct referrer-level analysis for podcast follows.

1.B.1 Effect on podcast follows

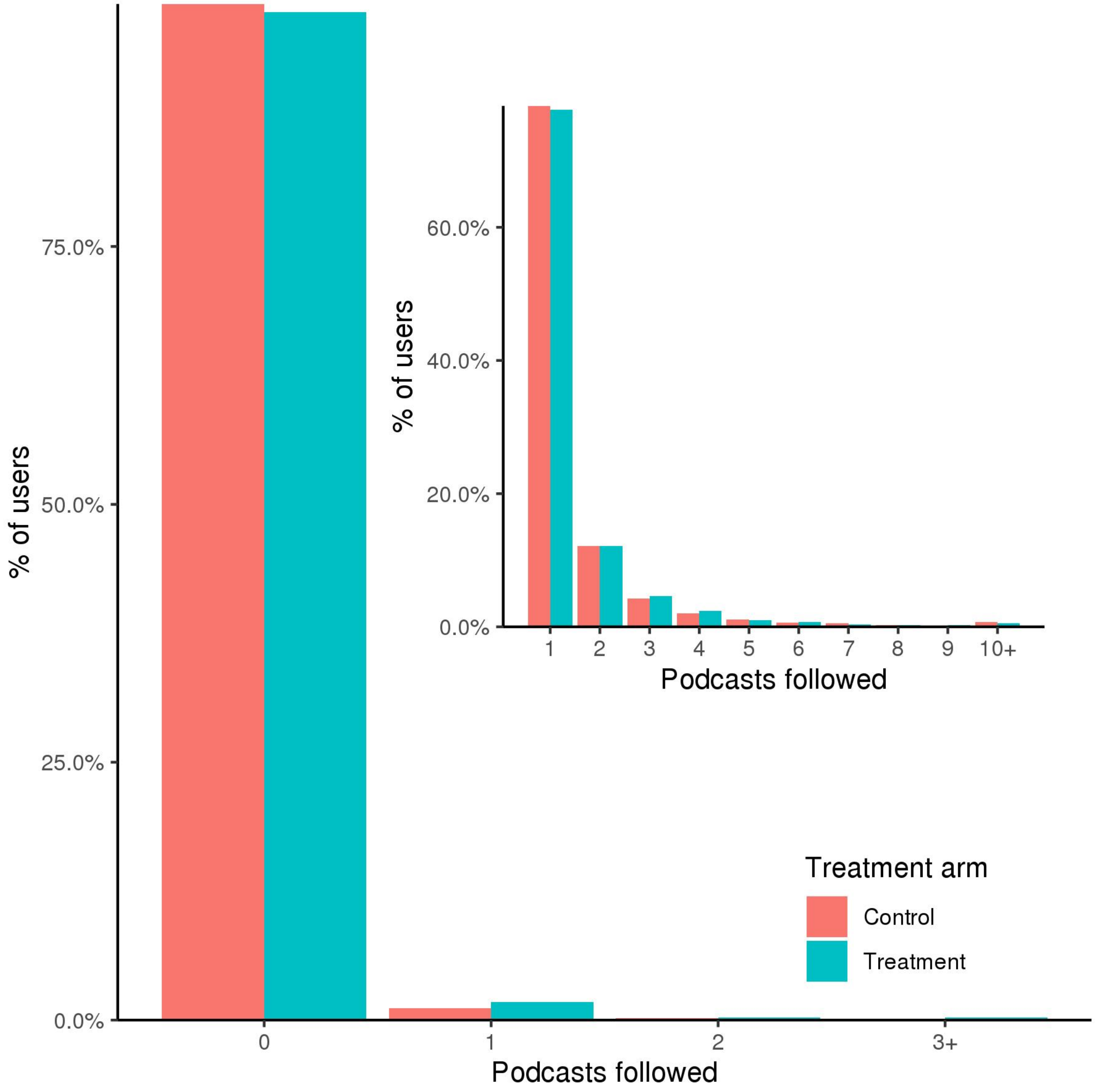


Figure 1.B.1: The distribution of podcasts followed in both treatment arms. Inset plot shows the distribution of podcasts followed in both treatment arms conditional on following at least one podcast.

Figure 1.B.1 shows the distribution of podcast follows per user during the experiment in both treatment arms. Table 1.B.1 reports the estimated effect of the treatment on podcast follows per user during the duration of the experiment, and Table 1.B.2 reports the estimated effect of the treatment on following at least one

Table 1.B.1: A linear model showing the effect of the treatment on number of podcasts followed. Standard errors are clustered at the user bucket level.

	Dependent variable:	
	Podcasts followed	
	(1)	(2)
Treatment	0.012***	0.012***
	(0.001)	(0.001)
Constant	0.023***	0.029***
	(0.0005)	(0.001)
User Gender	No	Yes
User Age	No	Yes
User account age	No	Yes
Observations	852,937	852,937
R^2	0.0004	0.001
Adjusted R^2	0.0004	0.001
Residual Std. Error	0.301 (df = 852935)	0.301 (df = 852925)
Note:		*p<0.1; **p<0.05; ***p<0.01

podcast during the experiment. We find that the treatment increased the number of Spotify users following at least one podcast by 53.45% (± 5.23%), and increased the number of podcast follows per user by 51.38% (± 7.64%).

Using the principal stratification approach (Frangakis and Rubin 2002, Ding and Lu 2017), we are able to measure the extent to which this treatment effect is driven by compositional shifts, as opposed to intensity shifts. We find that the treatment led "compliers" to follow 1.499 (95% CI: (1.472, 1.536)) more podcasts, whereas the treatment led "always takers" to follow 0.018 (95% CI: (-0.070, 0.035)) fewer podcasts. In other words, the increase in podcast following in the treatment is driven by a greater number of users following *at least one* podcast during the experiment, as opposed to an increase in the number of podcast follows from those who would have followed a podcast even if they had not been exposed to the treatment.

Table 1.B.2: A linear probability model showing the effect of the treatment on following at least one podcast. Standard errors are clustered at the user bucket level.

	Dependent variable:	
	Followed podcast	
	(1)	(2)
Treatment	0.008^{***}	0.008^{***}
	(0.0003)	(0.0003)
Constant	0.015^{***}	0.018^{***}
	(0.0002)	(0.0004)
User Gender	No	Yes
User Age	No	Yes
User account age	No	Yes
Observations	852,937	852,937
R^2	0.001	0.002
Adjusted R^2	0.001	0.002
Residual Std. Error	0.135 (df = 852935)	0.135 (df = 852925)
Note:		*p<0.1; **p<0.05; ***p<0.01

1.B.2 Effect on diversity of podcast follows

We also measure the effect of the treatment on the individual-level diversity and intragroup diversity for podcast follows.

Individual-level diversity

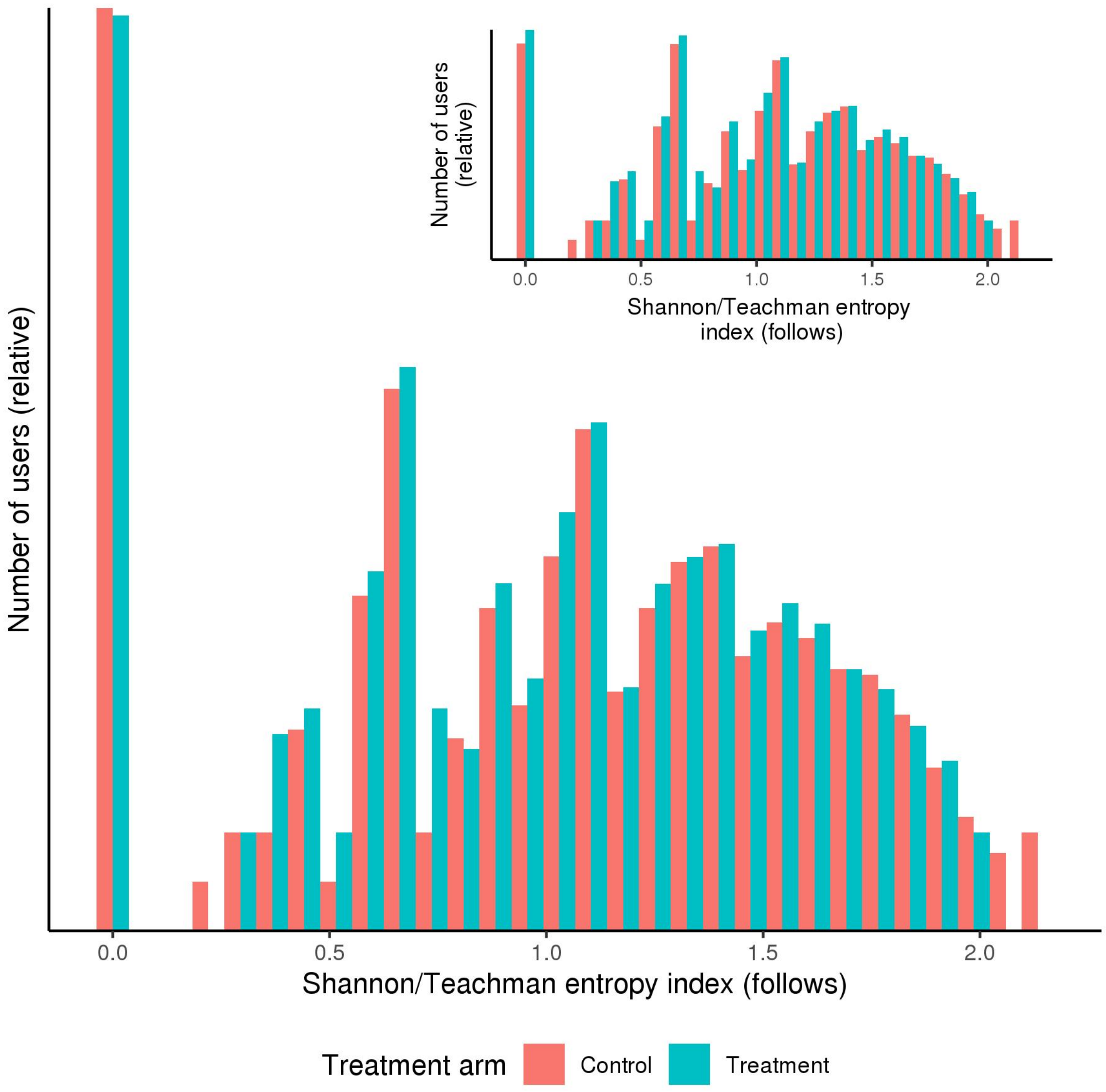


Figure 1.B.2: The distribution of the user-level diversity for follows in both treatment arms. Inset plot shows the distribution of user-level diversity in both treatment arms conditional on following at least one podcast. y-axis values are on a log scale, and are hidden due to confidentiality concerns.

Table 1.B.3: A linear model showing the difference in the average Shannon/Teachman entropy index (follows) (podcast followers only). Standard errors are clustered at the user bucket level.

	Dependent variable:	
	Shannon/Teachman entropy index (follows)	
	(1)	(2)
Treatment	-0.069^{***}	-0.068^{***}
	(0.008)	(0.008)
Constant	0.650^{***}	0.708^{***}
	(0.006)	(0.011)
User Gender	No	Yes
User Age	No	Yes
User account age	No	Yes
R^2	0.005	0.021
Adjusted R^2	0.005	0.020
Residual Std. Error	0.505 (df = 15894)	0.501 (df = 15884)
Note:		*p<0.1; **p<0.05; ***p<0.01
		Observation counts hidden due to confidentiality concerns

Figure 1.B.2 shows the histogram of the user-level Shannon entropy for podcast follows in both treatment arms, and Table 1.B.3 reports the difference in the average following user's Shannon entropy, both with and without controlling for user-level covariates. We find that the average Shannon entropy of podcast follows among users who followed at least one podcast was 10.68% (± 2.33%) lower in the treatment.

However, as was the case for our analysis of streaming behavior, this estimate is non-causal, since we condition on a post-treatment variable (the decision to follow at least one podcast). Using the principal stratification approach (Frangakis and Rubin 2002, Ding and Lu 2017), we can identify the causal effect of the treatment on the individual-level diversity of podcast follows for the subset of users who would follow a podcast, regardless of which treatment condition they were exposed to (i.e., "always takers"). We estimate that on average, the treatment decreased the individual-level diversity of always takers by 0.067 (95% CI: (0.052, 0.081)). In other words, the causal effect of the treatment on the individual-level diversity of podcast follows was negative for "always takers."

Table 1.B.4 reports the estimated effect of the treatment on the individual-level diversity of podcast follows for all users in the experiment, both with and without controlling for user-level covariates. We find that the treatment increased the Shannon entropy for podcast follows by 37.06% (± 6.17%). Figure 1.B.3 shows histograms of the user-level Shannon entropy for podcast follows in both treatment arms conditional on a user following a particular number of podcasts during the experiment.

Effect on intragroup diversity

We find that the treatment increased the intragroup diversity for follows by 7.12% (95% CI: (6.04%, 8.23%), from 0.687 (95% CI: (0.681, 0.693)) in the control group to 0.736 (95% CI: (0.732, 0.740))

Table 1.B.4: A linear model showing the effect of the treatment on the Shannon/Teachman entropy index (follows) (all users). Standard errors are clustered at the user bucket level.

	Dependent variable:	
	Shannon/Teachman entropy index (follows)	
	(1)	(2)
Treatment	0.004***	0.004***
	(0.0003)	(0.0003)
Constant	0.010***	0.013***
	(0.0002)	(0.0003)
User Gender	No	Yes
User Age	No	Yes
User account age	No	Yes
Observations	852,937	852,937
R^2	0.0003	0.001
Adjusted R^2	0.0003	0.001
Residual Std. Error	0.108 (df = 852935)	0.108 (df = 852925)
Note:		*p<0.1; **p<0.05; ***p<0.01

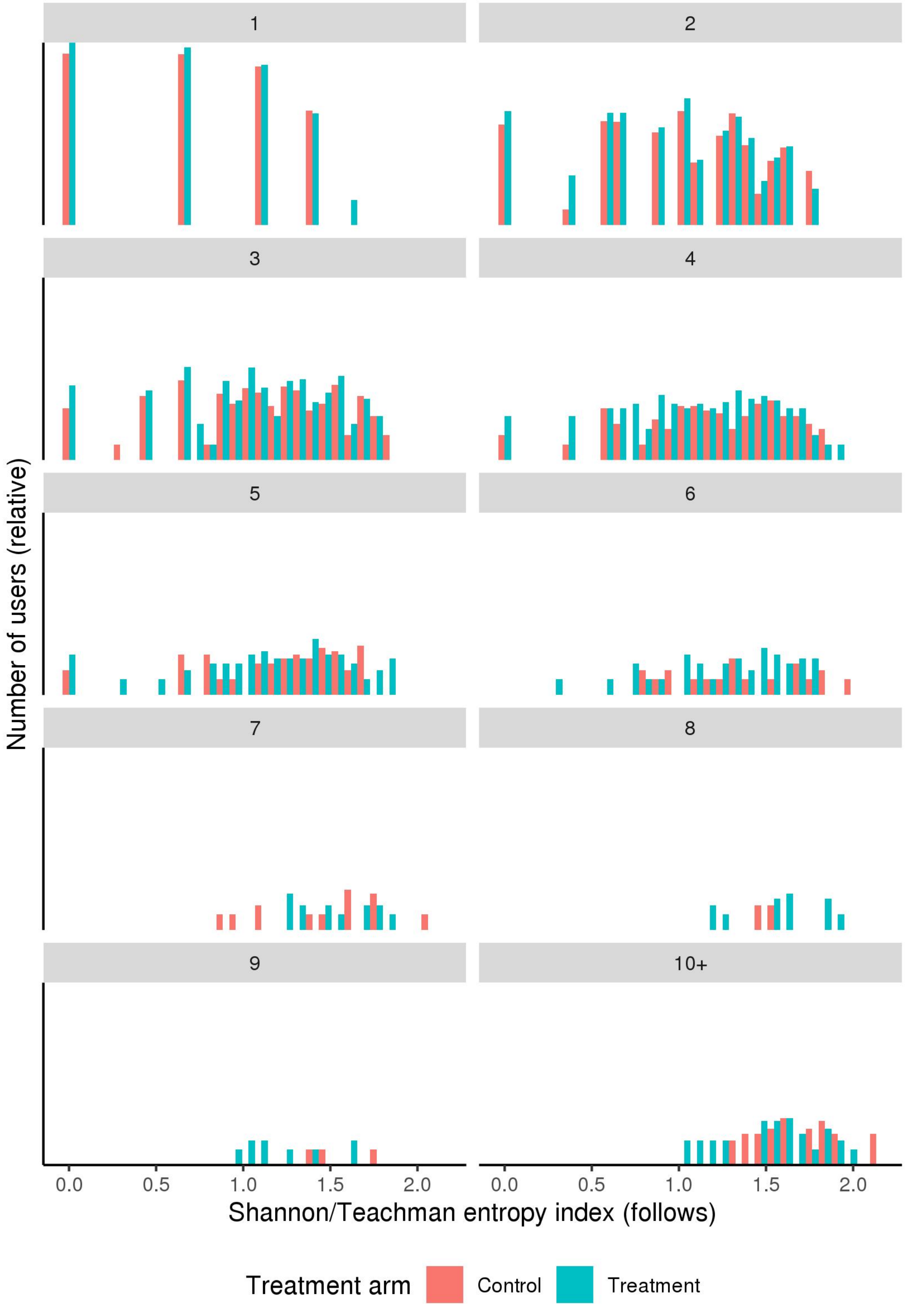


Figure 1.B.3: The distribution of the user-level diversity for follows in both treatment arms conditional on following a set number of podcasts during the experiment. y-axis values are on a log scale, and are hidden due to confidentiality concerns.

1.B.3 Long-term effects

We use data collected between May 3, 2019 and July 17, 2019 to test for longer-term effects of the treatment. We repeat our main analyses on cross-sectional datasets that describe users' behavior over time intervals spanning from the 3rd of the month to the 17th of the month, and from the 18th of the month to the 2nd of the month.

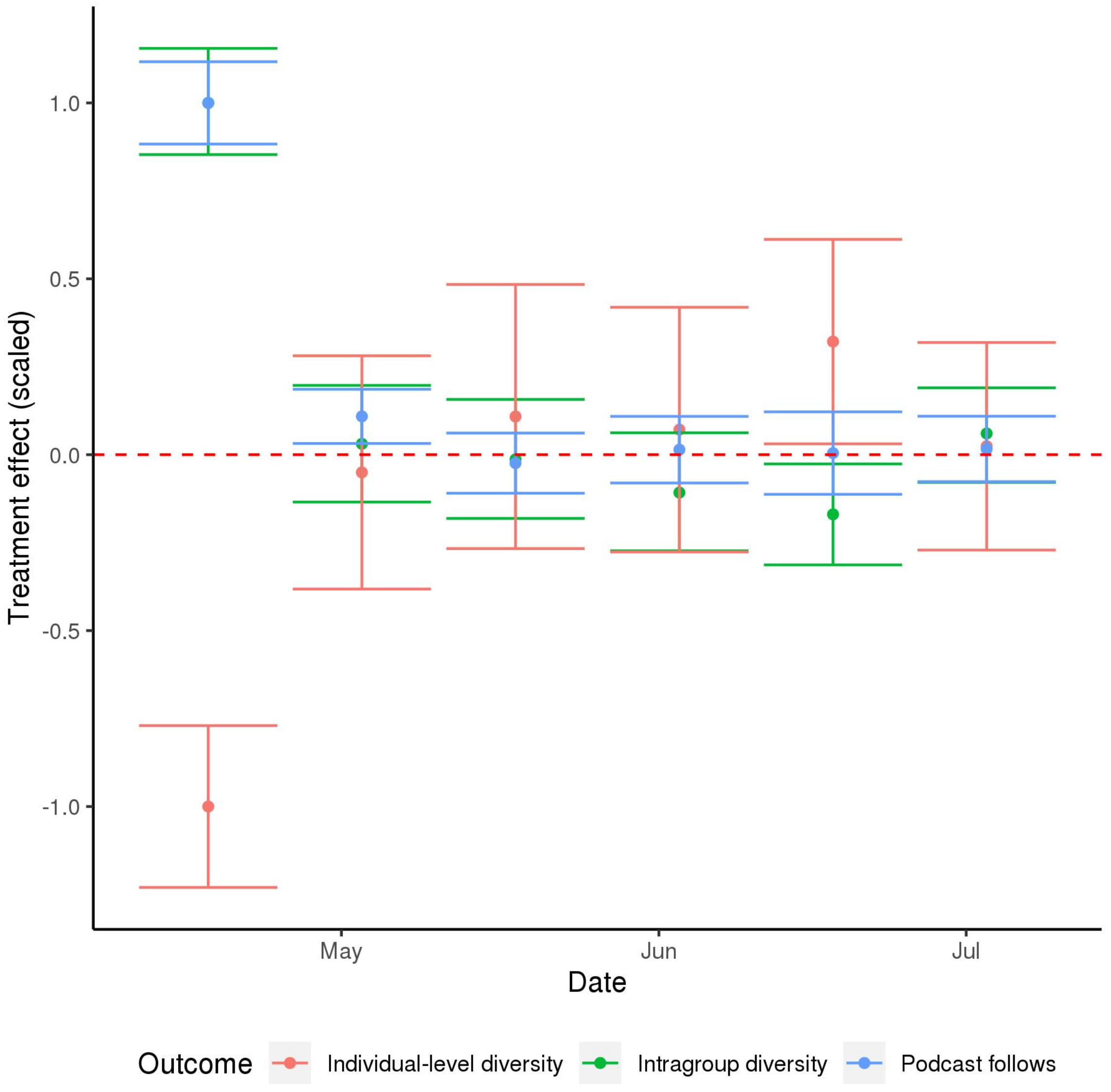


Figure 1.B.4: The long-term effect of the treatment on podcast follows per user, individual-level following diversity conditional on following at least one podcast, and intragroup following diversity. Each outcome's time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.

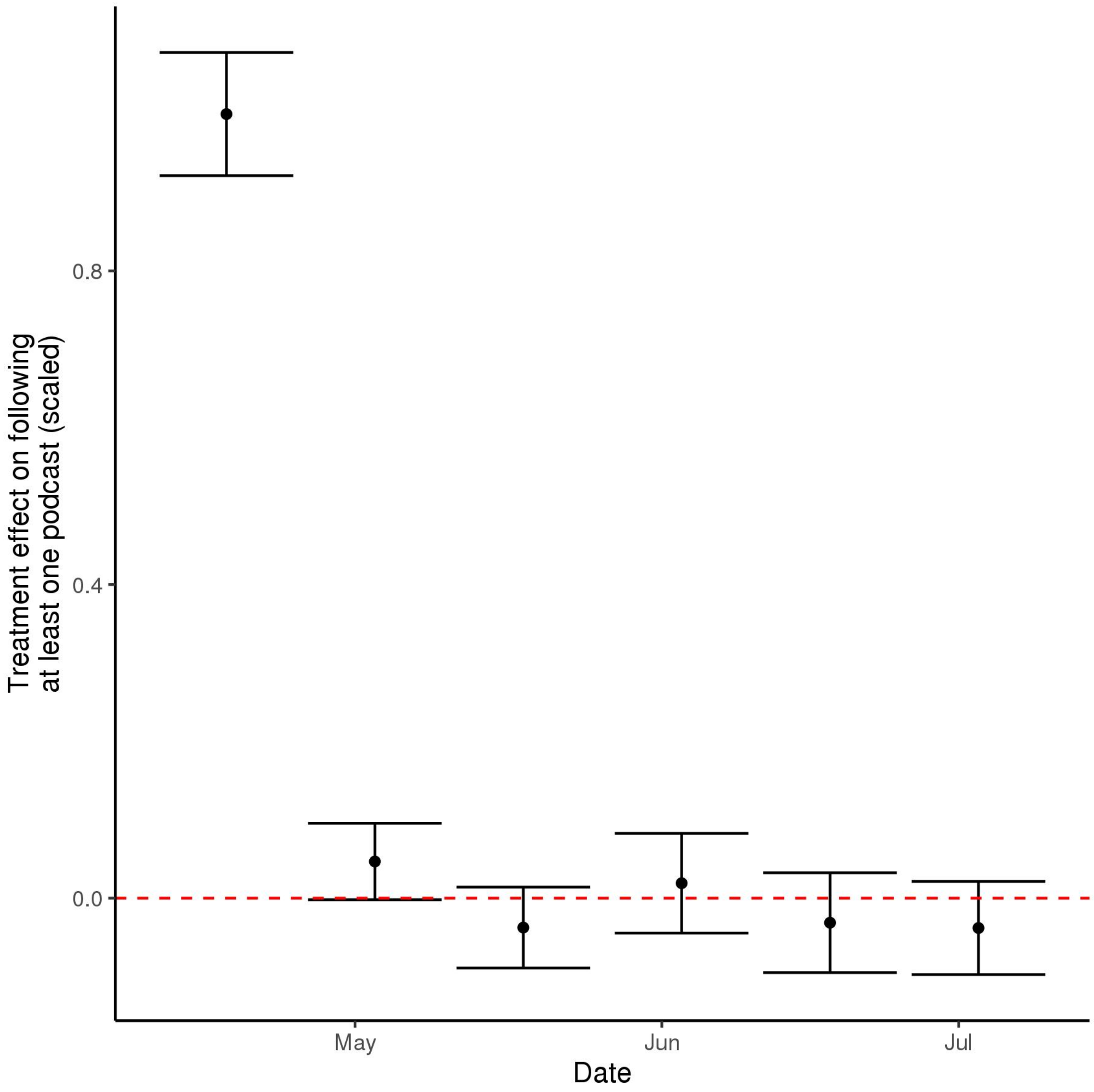


Figure 1.B.5: The long-term effect of the treatment on the percentage of users following at least one podcast. The time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.

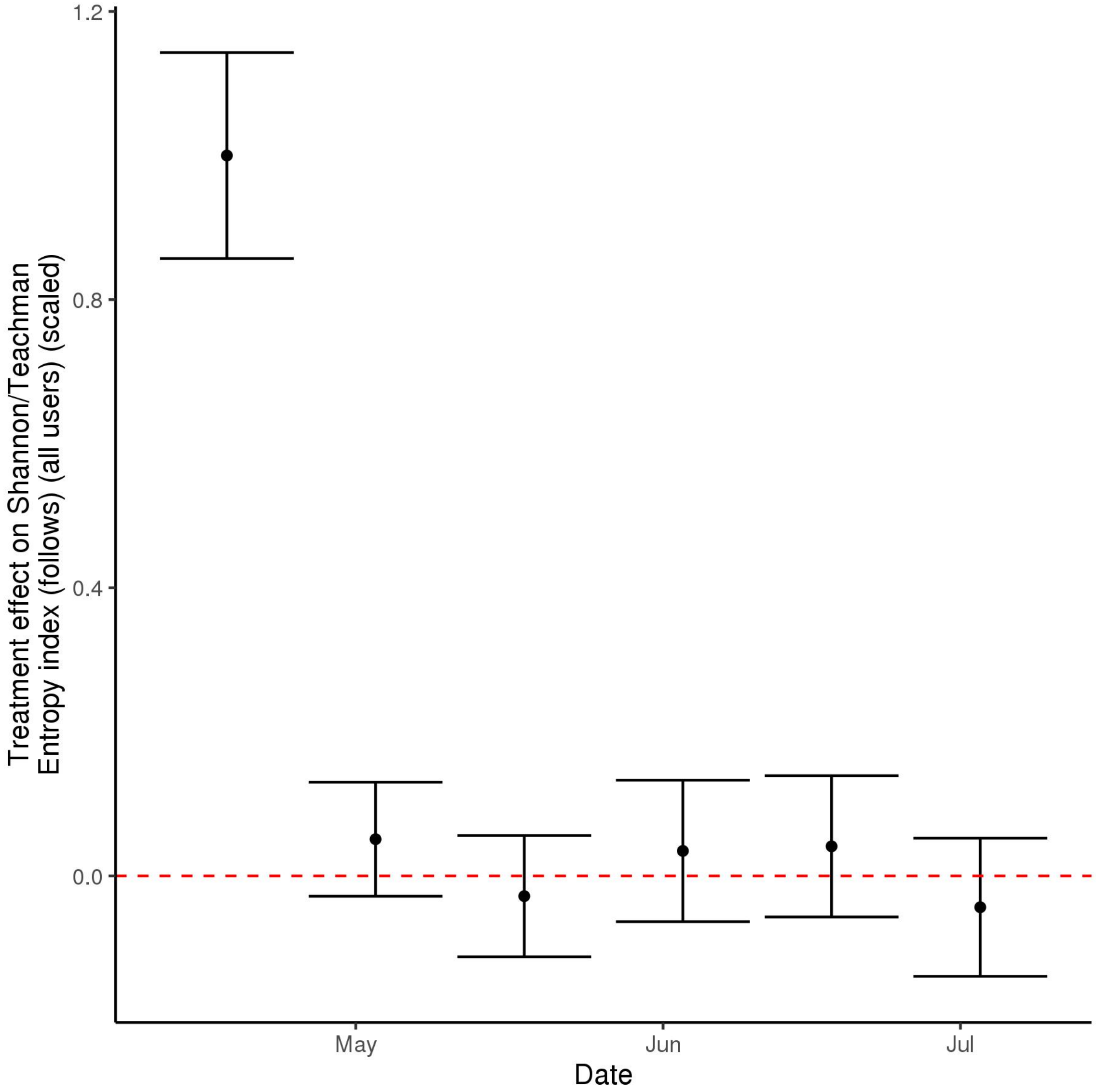


Figure 1.B.6: The long-term effect of the treatment on the average user-level Shannon entropy for follows. The time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.

Figure 1.B.4 shows the long-term effect of the treatment on the average number of podcast follows per user, the average Shannon entropy of podcast follows conditional on following at least one podcast, and the intragroup diversity of podcast follows. Across all of these outcomes, we observe the same trend: the large treatment effects observed during the experiment quickly shrink in magnitude, and in some cases disappear entirely, once the experiment has concluded. Figure 1.B.5 shows the long-term effect of the treatment on the number of users following at least one podcast, and Figure 1.B.6 shows the long-term effect of the treatment on the individual-level diversity for podcast follows across all users in our sample. For both of these time series, we also observe the rapid dissipation of treatment effects.

1.B.4 Effect on "sales diversity"

Figure 1.B.7 shows the Lorenz curve for podcast follows across the top 200 podcasts in both treatment arms of the experiment. The difference between the two curves indicates that the treatment makes podcast following *less* concentrated, and distributes a larger fraction of follows to less popular podcasts, i.e., the treatment increases the sales diversity for podcast follows. We confirm this by measuring the Gini coefficients corresponding to each treatment arm's Lorenz curve. We find that that the treatment reduces the Gini coefficient by 0.138 (95% CI: (0.116, 0.149)), from 0.588 to 0.450.

We also measure the effect of the treatment on sales diversity by estimating Equation 1.4 with follow counts and follow rank, as opposed to stream counts and stream rank. Figure 1.B.8 shows the relationship between $\ln(Follows_i + 1)$ and $\ln(Follows\,Rank_i)$ across all podcasts appearing in our dataset, and Table 1.B.5 shows the results of estimating Equation 1.4 on data from the top 200 podcasts in each treatment arm. The reported 95% confidence intervals are calculated with the cluster bootstrap (n_{boot} = 1,000). The positive estimate for β_3 also indicates that the treatment *increases* sales diversity.

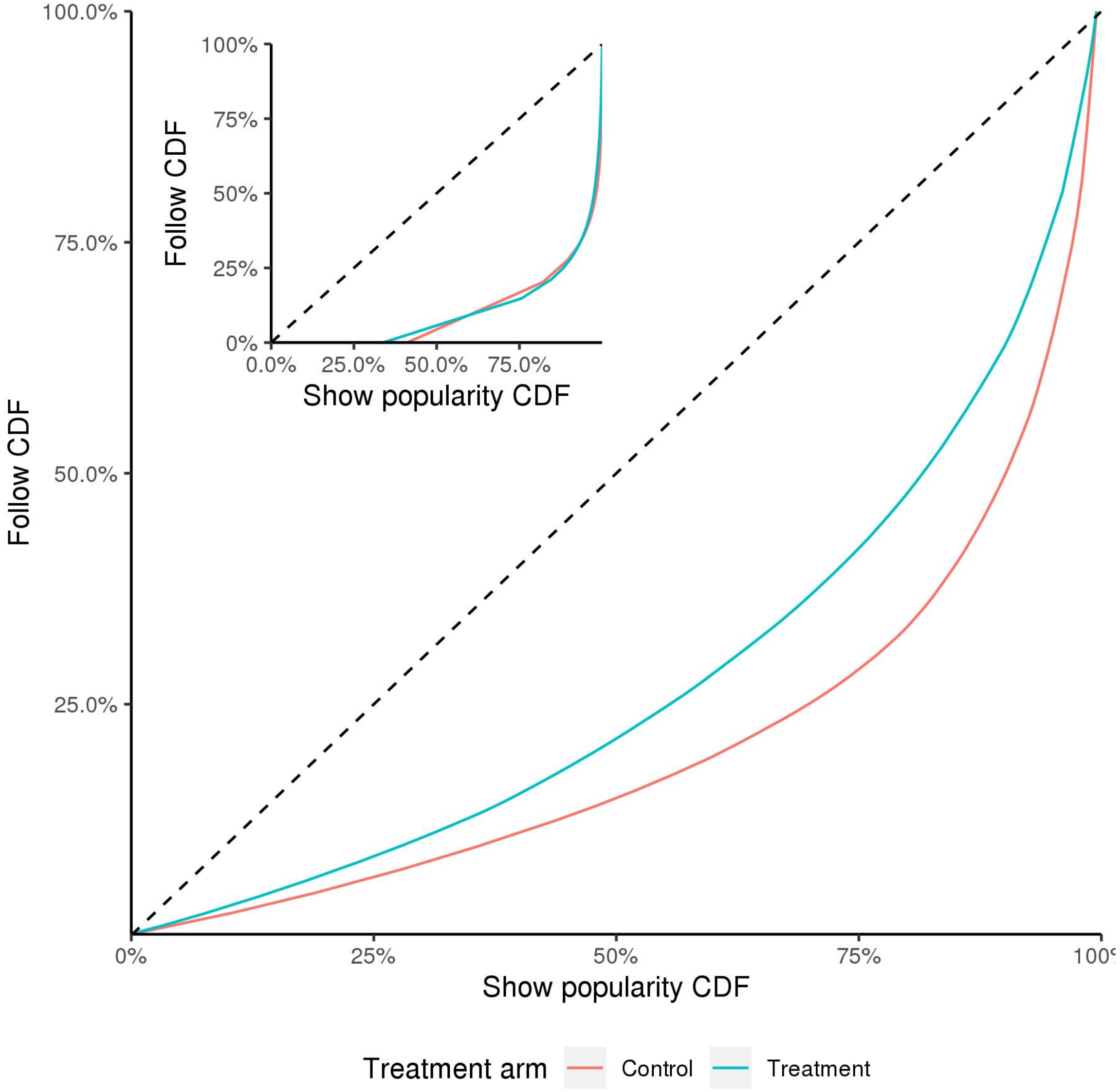


Figure 1.B.7: The Lorenz curves for podcast follows, calculated separately for users in the treatment and control. The data for each Lorenz curve is limited to the 200 most followed podcasts in the corresponding treatment arm data. The inset curve shows the Lorenz curve for follows across all podcasts.

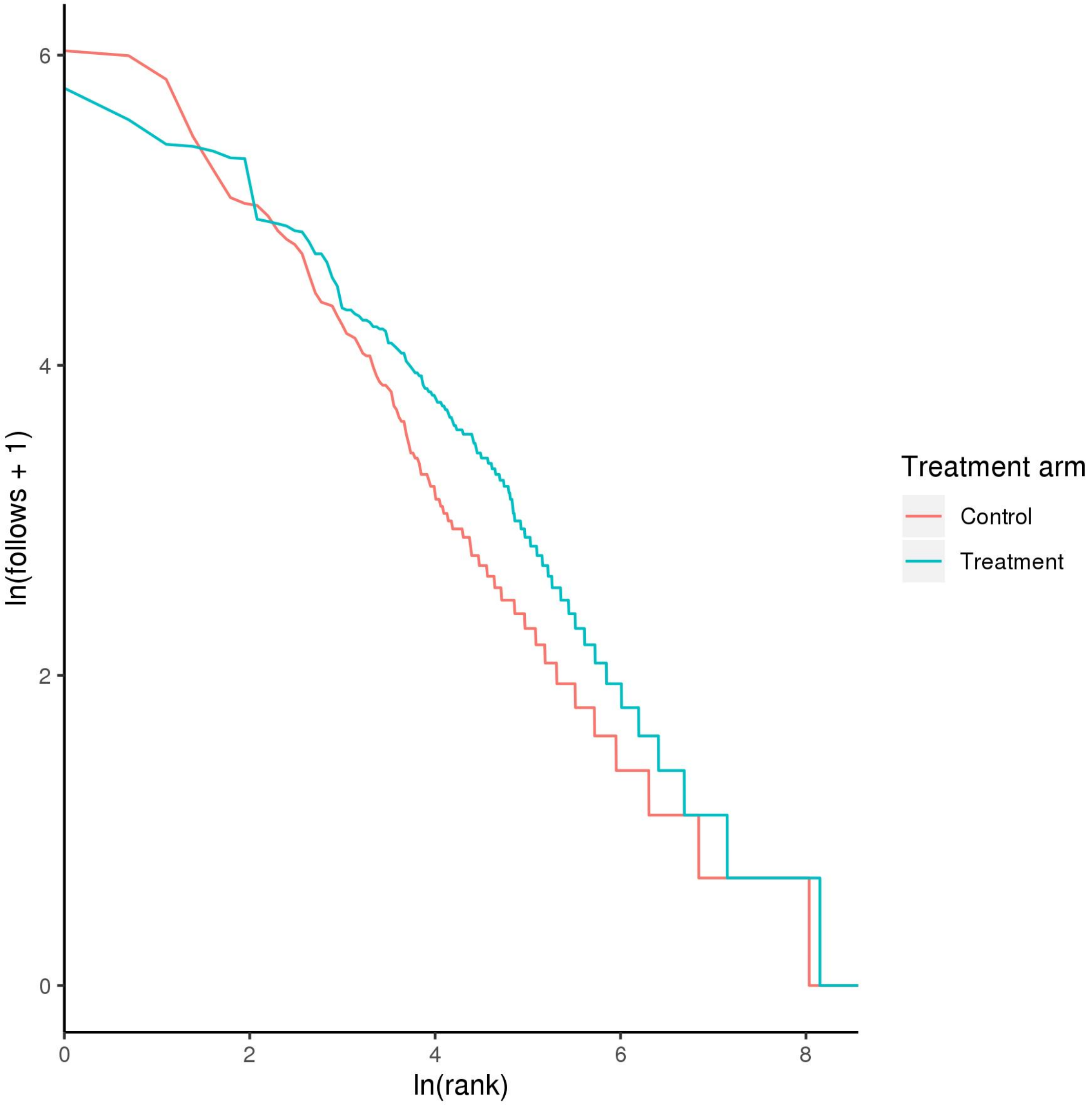


Figure 1.B.8: The relationship between ln(follows + 1) and ln(follow rank) for both the control and treatment arms of the experiment.

Table 1.B.5: Estimated coefficients for a model comparing the podcast follow Lorenz curves for control and treatment users

	Dependent variable:
	ln(follows + 1)
ln(rank)	-0.906^{***}
	$(-0.915, -0.869)$
Treatment	-0.208^{***}
	$(-0.295, -0.079)$
ln(rank) × Treatment	0.172^{***}
	(0.133, 0.194)
Constant	6.843^{***}
	(6.732, 6.892)
Observations	400
R^2	0.983
Adjusted R^2	0.983
Residual Std. Error	0.111 (df = 396)
Note:	*p<0.1; **p<0.05; ***p<0.01

1.C Principal stratification methodology

In this section, we describe our principal stratification methodology, which is based on the principal stratification approach described by Frangakis and Rubin (2002) and Ding and Lu (2017).

The principal stratification framework allows for causal inference in cases where an intermediate variable (in our case, listening to or following at least one podcast) leads to sample selection issues. Using this framework, we are able to separately measure causal effects of the treatment for "always takers," i.e., those would stream or follow a podcast, regardless of their treatment status and "compliers," i.e., those who would follow or stream a podcast only if treated. The key assumption necessary for implementing principal stratification is weak general principal ignorability (Ding and Lu 2017), which states that the expected outcome conditional on the intermediate variable (streaming or following at least one podcast) is independent of strata (complier, always taker, never taker) after controlling for covariates.

Our implementation of the principal stratification framework uses the marginal method described by Feller et al. (2017) to compute the probability that each user in our sample is a complier, always taker, or never taker. Under the principal stratification approach's monotonicity assumption, we can assume that users who do not stream or follow a podcast in the treatment are never takers, and that podcast streamers or followers in the control are always takers. For all other users, we estimate the probability that they are an always taker using a logistic regression model that is trained on control data and predicts streaming or following a podcast using user-level covariates. Similarly, we estimate the probability that a user is a never taker using a logistic regression model that is trained on treatment data and predicts streaming or following a podcast using user-level covariates. Once we have estimated $P(always\,taker)_i$ and $P(never\,taker)_i$, we can calculate $P(complier)_i$, since $P(complier)_i = 1 - P(always\,taker)_i - P(never\,taker)_i$. In cases where $P(always\,taker)_i + P(never\,taker)_i > 1$, we set $P(complier)_i = 0$ and normalize the other two probabilities so that they sum to 1. In both logistic regression models, user age bucket, user gender, and user account age (in days) are the covariates used to predict the intermediate variable.[19] Once we have computed the probability that each user belongs to each stratum, we use these probabilities as weights to construct causal stratum-level treatment effect estimators. Confidence intervals are calculated using a clustered bootstrap ($n_{boot} = 1,000$).

We test that the principal stratification model we have proposed is accurate using the balancing conditions proposed by Ding and Lu (2017). Simply put, the balancing conditions require that within each stratum, the treatment should not appear to have a causal effect on any function of the pretreatment covariates used to estimate a given unit's stratum. For both intermediate outcomes (streaming at least one podcast and following at least one podcast), we estimate the effect of the treatment on each pre-treatment user-level covariate in each stratum. The results for podcast streaming are shown in Figure 1.C.1 and the results for podcast following are shown in Figure 1.C.2. In both cases, the estimated effects are nearly zero across all strata and covariates, indicating that the balancing conditions are satisfied.

1.D "Home" podcast impressions over time

The number of podcast streams per user over time was dependent on users being exposed to podcast content in the "Home" section of the Spotify app. However, the number of impressions that podcast content received on "Home" varied considerably in the months following the experiment. During the experiment, the majority of podcast content impressions on "Home" came from the "Podcasts to Try" shelf, since it was anchored in the second slot. After the experiment had ended, the "Podcasts to Try" shelf was briefly removed from the Spotify app to be productionized. The treatment version of the shelf was relaunched to 100% of Spotify users in late May, however, the shelf was no longer anchored in the second slot. As a result, there were

[19] We also calculate strata probability estimates using the EM algorithm described by Ding and Lu (2017). The point estimates obtained using this method are qualitatively similar to those obtained using the marginal method. However, we choose the marginal method for computational tractability when calculating bootstrap standard errors.

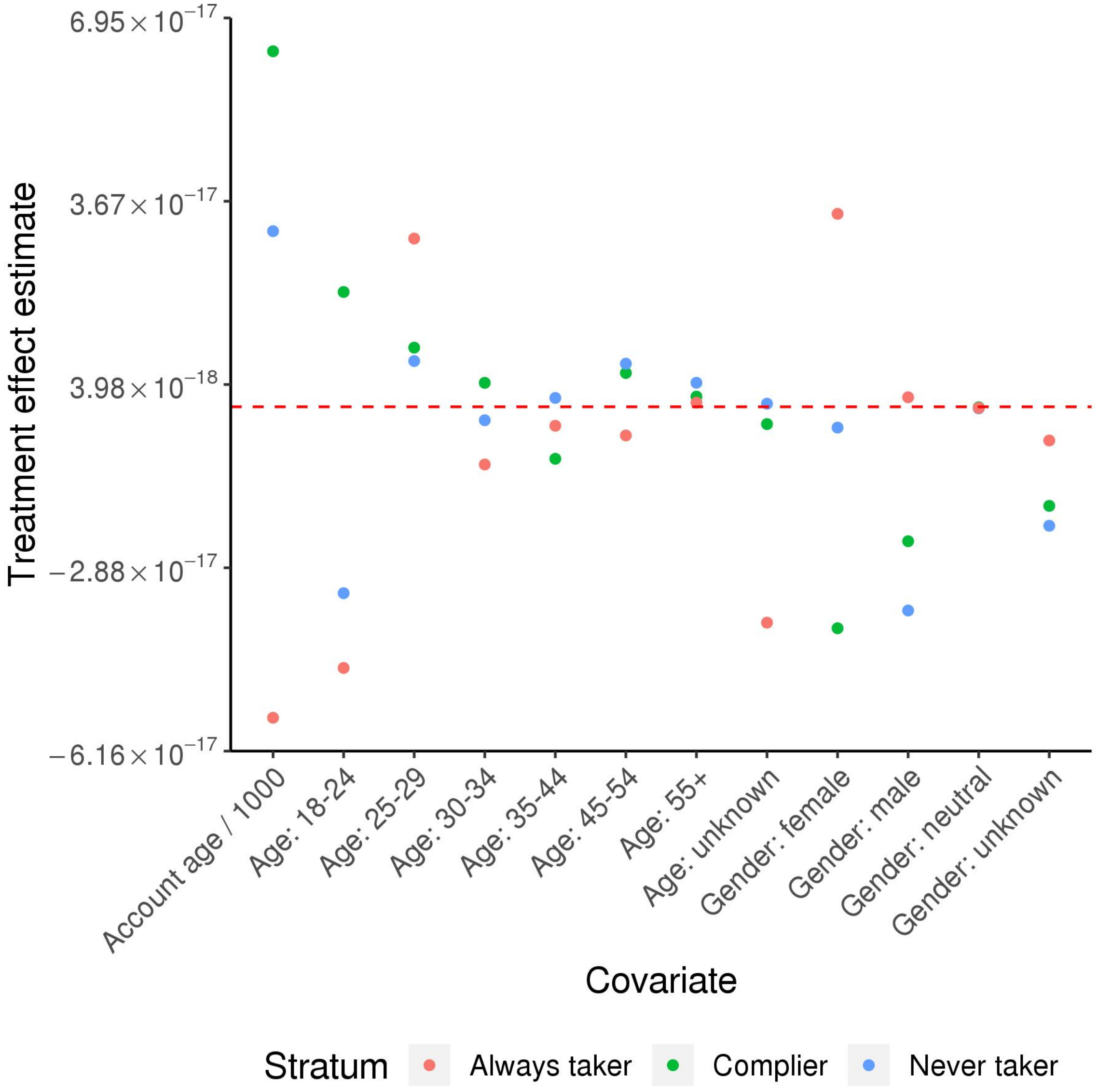


Figure 1.C.1: Results of the principal stratification balance check. The intermediate variable is whether a given user streamed at least one podcast during the experiment.

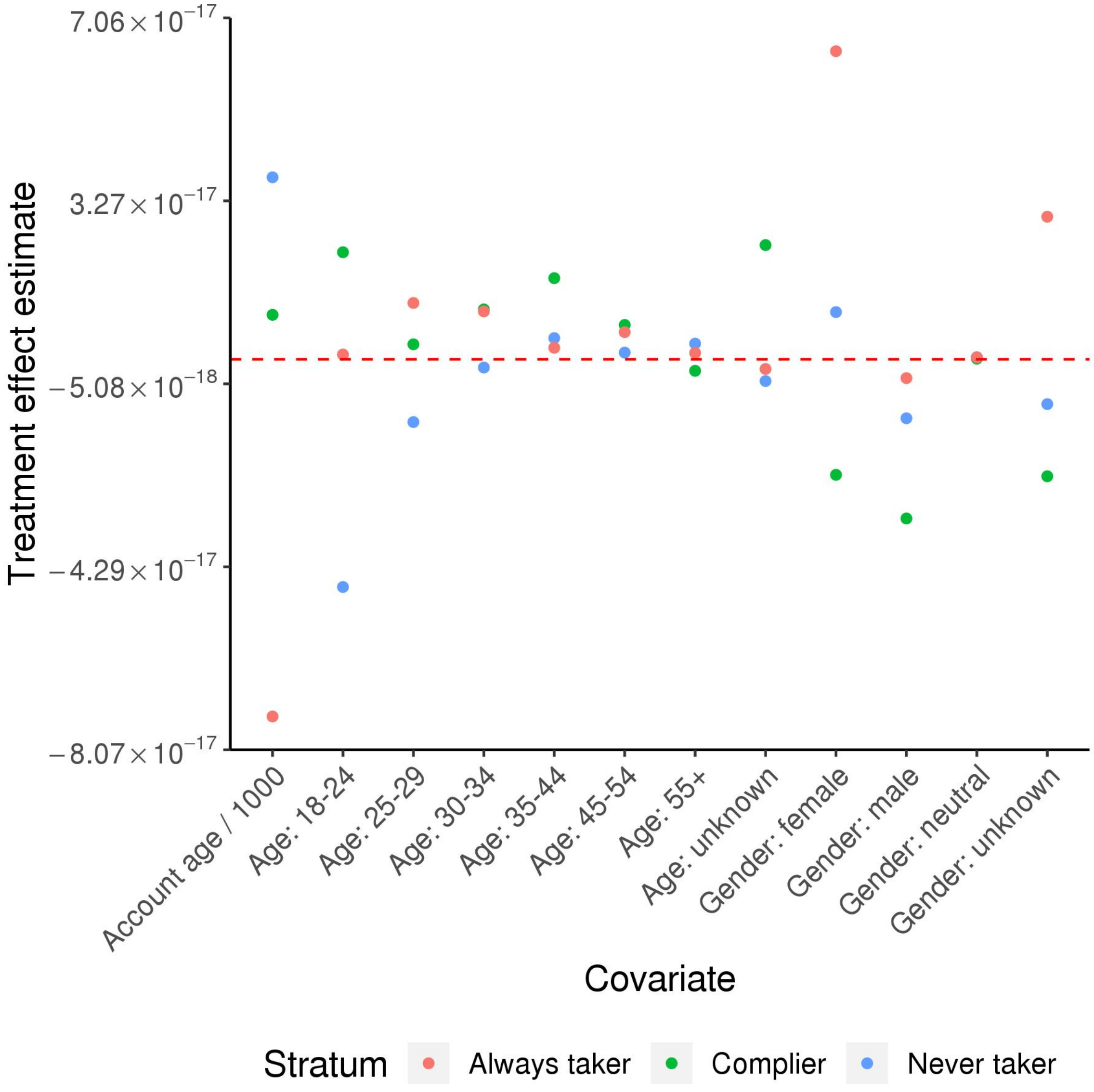


Figure 1.C.2: Results of the principal stratification balance check. The intermediate variable is whether a given user followed at least one podcast during the experiment.

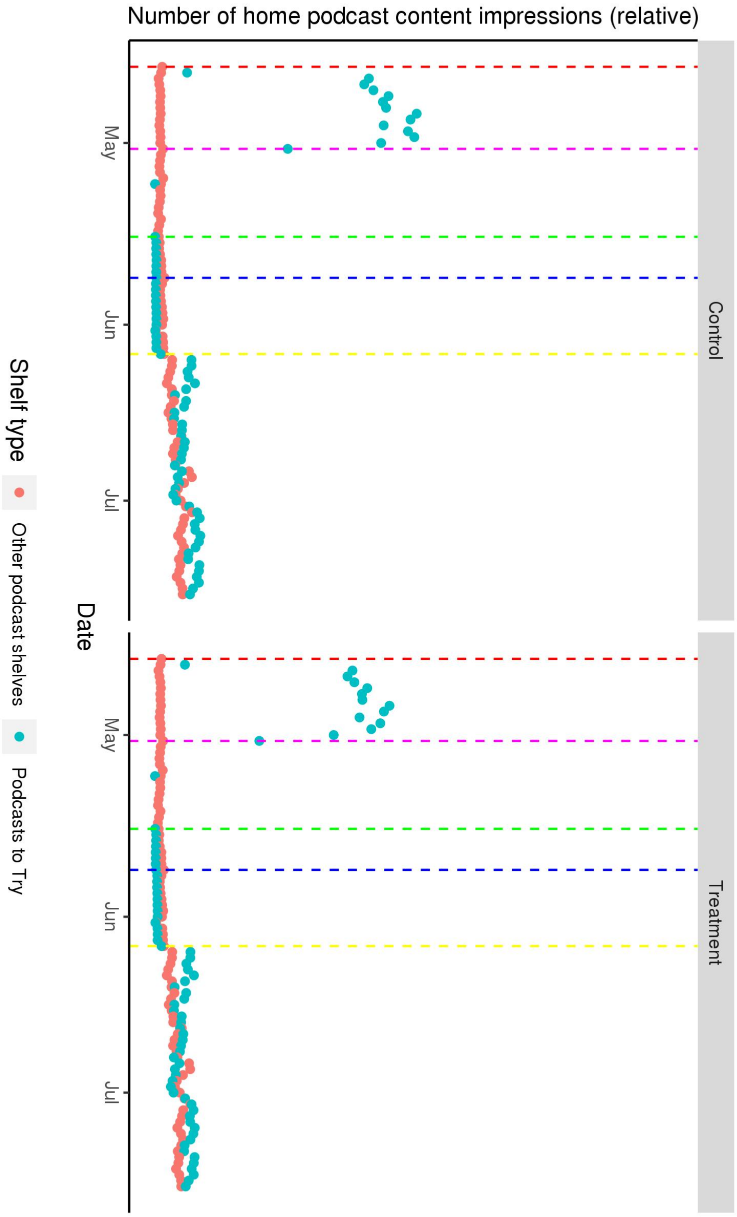


Figure 1.D.1: The number of daily podcast content impressions from both the "Podcasts to Try" shelf and other podcast-related shelves on the "Home" section of the Spotify app, shown separately for users in the two treatment arms of the experiment. The dashed red line corresponds to the experiment launch date. The dashed magenta line corresponds to the experiment end date. The dashed green line corresponds to the productization of the "Podcasts to try" shelf. The dashed blue line corresponds to the launch of the podcast shelf boosting experiment. The dashed yellow line corresponds to the end of the podcast shelf boosting experiment. y-axis values hidden due to confidentiality concerns.

far fewer impressions for all podcast related shelves, including "Podcasts to try." An experiment to determine the optimal amount of boosting for podcast shelves was launched in mid-May, and podcast shelf boosting was launched to 100% of users in early June. Figure 1.D.1 shows the number of impressions for podcast content on both "Podcasts to Try" and other podcast-related shelves for both treatment and control users over time. Note that the time series for the two experiment treatment arms are essentially identical.

1.E Effect on "sales diversity"

In this section, we measure the effect of the treatment on the "sales diversity" of podcast consumption, as measured through the Lorenz curve and Gini coefficients corresponding to podcast streaming in both treatment arms of the experiment.

Figure 1.E.1 shows the Lorenz curve for podcast streaming across the top 1,000 podcasts in both treatment arms of the experiment. The difference between the two curves indicates that the treatment makes podcast streaming *less* concentrated, and distributes a larger fraction of streams to less popular podcasts. In other words, the treatment increases sales diversity. We confirm this by measuring the Gini coefficients corresponding to each treatment arm's Lorenz curve. We find that that the treatment reduces the Gini coefficient by 0.050 (95% CI: 0.037, 0.061), from 0.692 to 0.642.

We also measure the effect of the treatment on sales diversity by estimating the following model (Brynjolfsson et al. 2011):

$$\begin{aligned}\ln(Streams_i + 1) = &\beta_0 + \beta_1 \ln(Streams\,Rank_i) + \beta_2 Treatment_i + \\ &\beta_3 Treatment_i \times \ln(Streams\,Rank_i) + \epsilon_i,\end{aligned} \tag{1.4}$$

where $Streams_i$ is how many streams podcast i received during the experiment in a particular treatment arm, $Streams\,Rank_i$ is podcast i's rank among all podcasts in that treatment arm, and $Treatment_i$ indicates the treatment arm corresponding to the observation. The coefficient of interest is β_3, which tests for a difference between

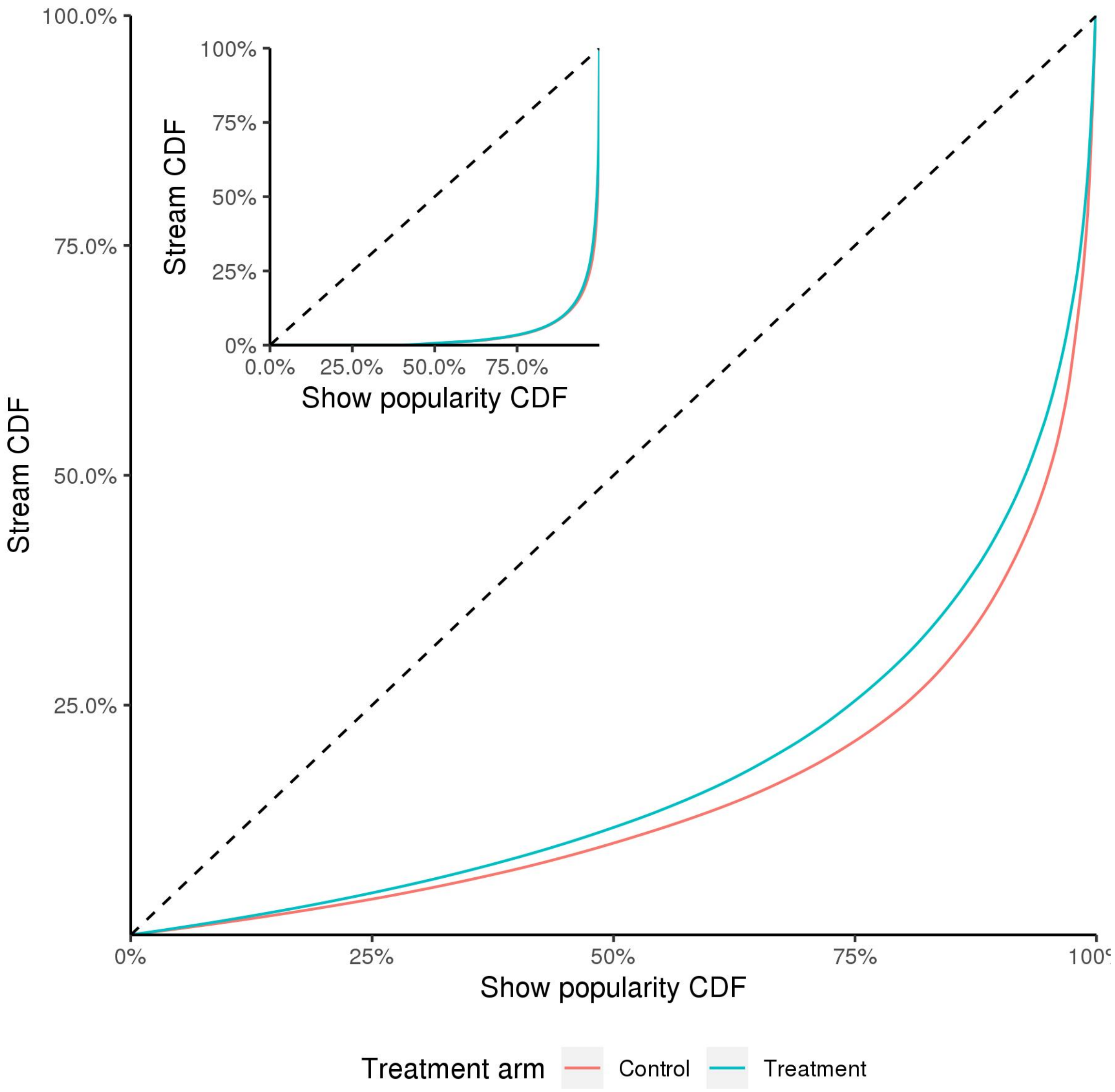


Figure 1.E.1: The Lorenz curves for podcast streams, calculated separately for users in the treatment and control. The data for each Lorenz curve is limited to the 1,000 most streamed podcasts in the corresponding treatment arm data. The inset curve shows the Lorenz curve for streams across all podcasts.

the two treatment arms in the rate at which number of streams decreases with stream rank.

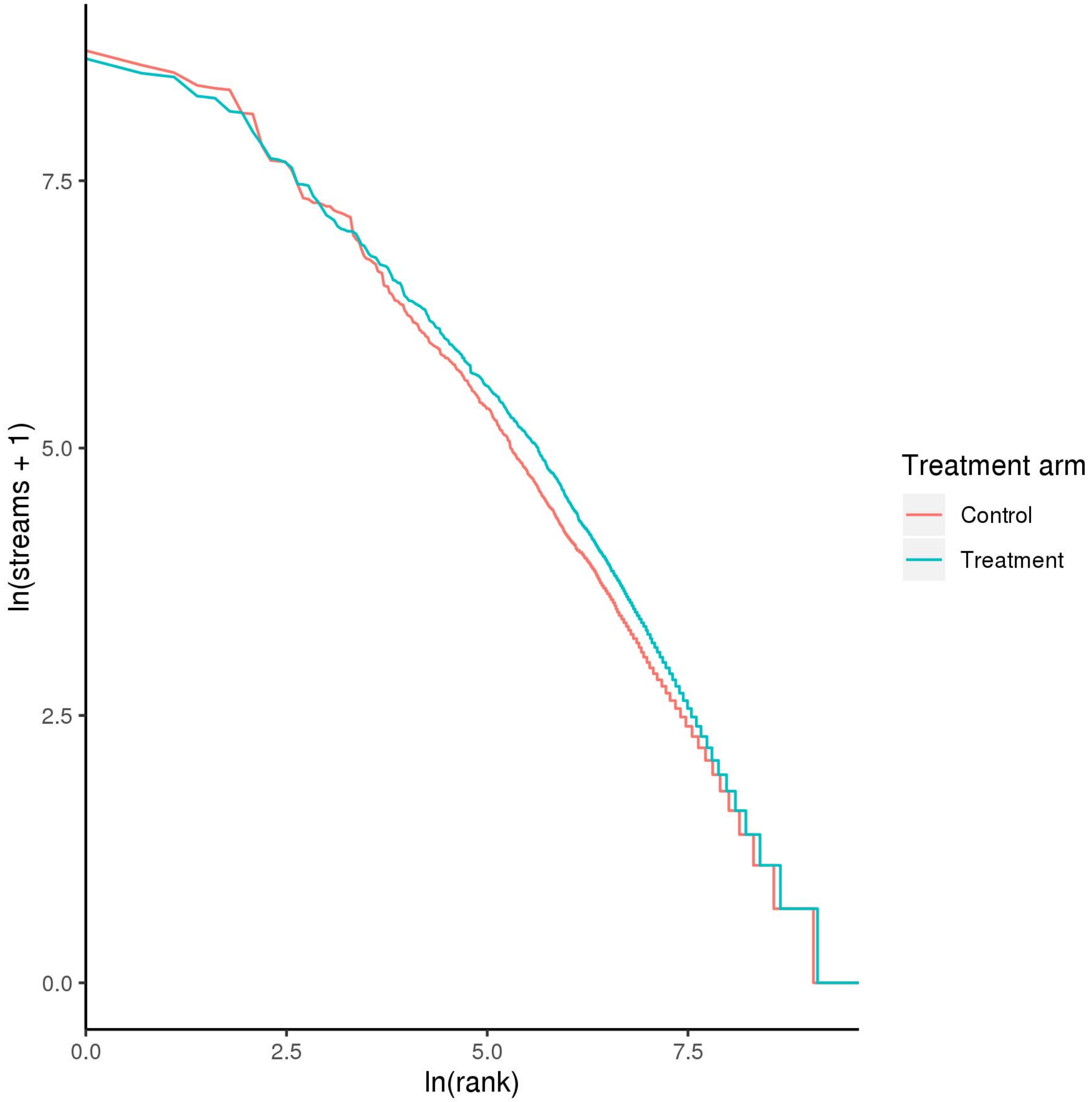


Figure 1.E.2: The relationship between ln(streams + 1) and ln(stream rank) for both the control and treatment arms of the experiment.

Figure 1.E.2 shows the relationship between $\ln(Streams_i + 1)$ and $\ln(Streams\ Rank_i)$ across all podcasts appearing in our dataset, and Table 1.E.1 shows the results of estimating Equation 1.4 on data from the top 1,000 podcasts in each treatment arm. The reported 95% confidence intervals are calculated with the cluster bootstrap (n_{boot}

Table 1.E.1: Estimated coefficients for a model comparing the podcast stream Lorenz curves for control and treatment users

	Dependent variable:
	ln(streams + 1)
ln(rank)	-1.046^{***} $(-1.064, -1.024)$
Treatment	-0.043 $(-0.178, 0.086)$
ln(rank) × Treatment	0.053^{***} $(0.027, 0.078)$
Constant	10.453^{***} $(10.380, 10.587)$
Observations	2,000
R^2	0.987
Adjusted R^2	0.987
Residual Std. Error	0.118 (df = 1996)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

= 1,000). The positive estimate for β_3 also indicates that the treatment *increases* sales diversity.

1.F Additional Tables

Table 1.F.1: A linear model showing the effect of the treatment on the Shannon/Teachman entropy index (streams) (all users). Standard errors are clustered at the user bucket level.

	Dependent variable:	
	Shannon/Teachman entropy index (streams)	
	(1)	(2)
Treatment	0.010^{***}	0.010^{***}
	(0.001)	(0.001)
Constant	0.047^{***}	0.051^{***}
	(0.0004)	(0.001)
User Gender	No	Yes
User Age	No	Yes
User account age	No	Yes
Observations	852,937	852,937
R^2	0.001	0.003
Adjusted R^2	0.001	0.003
Residual Std. Error	0.219 (df = 852935)	0.219 (df = 852925)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 1.F.2: A linear model showing the effect of the treatment on the Shannon/Teachman entropy index (streams) by stream referral source (all users). Standard errors are clustered at the user bucket level.

	Dependent variable:			
	Shannon/Teachman entropy index (streams)			
	Home		Non-home	
	(1)	(2)	(3)	(4)
Treatment	0.010^{***}	0.010^{***}	0.002^{***}	0.002^{***}
	(0.001)	(0.001)	(0.0004)	(0.0004)
Constant	0.033^{***}	0.038^{***}	0.022^{***}	0.021^{***}
	(0.0004)	(0.001)	(0.0002)	(0.0005)
User Gender	No	Yes	No	Yes
User Age	No	Yes	No	Yes
User account age	No	Yes	No	Yes
Observations	852,937	852,937	852,937	852,937
R^2	0.001	0.002	0.00003	0.002
Adjusted R^2	0.001	0.002	0.00003	0.002
Residual Std. Error	0.184 (df = 852935)	0.184 (df = 852925)	0.151 (df = 852935)	0.150 (df = 852925)
Note:				*p<0.1; **p<0.05; ***p<0.01

Chapter 2

Limiting Bias from Test-Control Interference in Online Marketplace Experiments[1]

2.1 Introduction

Some of the world's most valuable tech companies, such as Alibaba, Amazon, Airbnb, and Uber, own and operate online marketplaces. A common way for these firms to make product decisions is through experimentation, or 'A/B testing' (Kohavi et al. 2009). Typically these experiments aim to measure the total average treatment effect (TATE) of a treatment intervention, i.e., the difference between the average outcome across all units if 100% of units had been exposed to the treatment, and the average outcome across all units if no units had been exposed to the treatment, on outcomes such as click-through-rate or average revenue per user. Many firms conduct thousands of experiments a year (Clarke 2016), and often aim to measure TATEs that are small in relative terms (e.g., fractions of a percentage point), but which can correspond to large revenue gains or losses (e.g., millions of dollars). Given the high potential cost

[1] I thank Sinan Aral, Dean Eckles, Andrey Fradkin, Madhav Kumar, Ed McFowland, Rajiv Mukherjee, and Michael Zhao for their useful feedback, and are grateful to KJ Dakin for copy editing assistance. I also thank participants at CODE 2016, WISE 2016, and WEBEIS 2019, as well as anonymous reviewers at *Information Systems Research*, for their comments.

of drawing incorrect inferences about the effects of any given product change, it is crucial for online marketplace firms to develop methods for unbiased causal inference in online marketplace settings.

Under the assumption that each experimental unit's response is not influenced by any other units' treatment assignment (i.e., the "stable unit treatment value assumption" (SUTVA) (Rubin 1974) or "no interference" assumption (Cox 1958)), the TATE can be identified by the "standard" approach to experimentation, in which individual units are randomly assigned to treatment or control, and the TATE is estimated using a difference in means estimator. However, online marketplaces are by definition connected and violate SUTVA. For instance, a treatment that increases buyer demand may reduce the supply available to buyers in the control group. Similarly, a treatment that induces sellers to lower prices may increase demand for treated sellers' products, thereby reducing the demand for control sellers' products. In both of these cases, the treatment affects not only the treated units' outcomes, but also the control units' outcomes.

In this paper, I refer to violations of SUTVA as 'test-control interference.' Existing work has documented the presence of test-control interference in online marketplace settings (Blake and Coey 2014, Holtz et al. 2020), and previous research has shown both through the analysis of experimental data and through simulations that naive estimates of TATEs in online marketplaces can exaggerate the effectiveness of treatment interventions by over 100% (Blake and Coey 2014, Fradkin 2015). Both Blake and Coey (2014) and Fradkin (2015) have previously proposed methods for combating test-control interference in online marketplace experiments. However, the proposed solutions require markets that offer convenient and easily identifiable units of analysis (e.g., auctions or sub-markets) over which to aggregate outcomes or segment the marketplace, or involve the development of structural model-based simulations, which may be difficult to implement for smaller firms that do not employ a large number of economists with training in structural modeling.

Here, I contribute to the literature on test-control interference in online marketplace experiments by proposing that techniques for network experimentation be adapted to the context of online marketplaces, and using a simulation framework to assess the efficacy of these techniques at reducing the bias of TATE estimates. The techniques analyzed include graph cluster randomization (GCR) (Ugander et al. 2013), exposure modeling (Aronow et al. 2017), and inverse probability-weighted treatment effect estimators (Hájek 1971, Aronow et al. 2017, Eckles et al. 2017). The key step required to adapt these methods to online marketplaces is inferring a "product network" that connects different sellers or items within a market. After this is done, the aforementioned methods can be applied in a straightforward way when designing and/or analyzing online marketplace experiments. Prior research in the information systems literature has shown that "visible" product networks, in which an edge exists between two products if they appear on each other's checkout pages, can be an effective tool for studying competition, estimating demand spillovers, and predicting demand (Oestreicher-Singer and Sundararajan 2012b,a, Dhar et al. 2014). However, it is unclear *ex ante* if network experiment design and analysis techniques will be effective at reducing bias when applied to product networks, given that the underlying phenomena that cause spillovers (e.g., substitution and complementarity) are different. Furthermore, in many cases, online marketplace experiment designers may need to infer a product network because there is no "explicit" network that connects sellers/items in the marketplace.

The simulations use a scraped dataset that describes the full set of Airbnb properties in and around Miami, as well as a product network that is built for this market using a simple heuristic. In the simulation, which models one night of booking activity on Airbnb, a number of searchers with BLP-inspired utility functions (Berry et al. 1995, Nevo 2000) sequentially visit the market, and are served a subset of the available properties according to a "search algorithm." Based on their preferences, each searcher chooses which property, if any, they will book, after which that property does not appear in subsequent search results. Notably, my simulation framework does not make any assumptions about the functional form of interference between units. In fact, my marketplace simulations do not explicitly prescribe that different properties in the market interfere with one another at all.

Using this simulation framework, I am able to estimate the market-level booking rate and average listing revenue in the absence of a treatment intervention, and under the market-wide implementation of two different treatment interventions: one in which the prices of treated listings are reduced, and one in which the "unobservable" quality of treated listings is increased. After quantifying the true TATEs, using simulations of market-wide treatment and market-wide control, I also use my simulation framework to compare the bias, root mean square error (RMSE), and coverage probabilities of different combinations of experiment design, exposure model, and treatment effect estimator. I find that relative to the baseline of an individual-level, Bernoulli randomized experiment that is analyzed with a difference in means estimator, blocked GCR can reduce the bias of TATE estimates by as much as 64.5%. However, blocked GCR also increased the RMSE of TATE estimates by up to 204%. When analyzed with a linear regression model in which standard errors are clustered at the level of the graph cluster, blocked GCR also led to coverage probabilities that were at or above the nominal level, whereas the baseline experiment design consistently led to 95% confidence intervals with coverage probabilities as low as 6%. I also find that the combination of the FNTR exposure model and the inverse probability-weighted Hajek estimator have the potential to further reduce the bias of TATE estimates. However, the effectiveness of the FNTR exposure model is dependent on the choice of threshold, and the optimal choice may be difficult for experimenters to ascertain.

I also test the robustness of these findings to changes to the data generating process, different levels of network mis-specification, varied numbers of product network clusters, different levels of demand in the market, and an entirely different set of products (and, as a result, a different product network). In almost all cases, I find that my general results hold. I also fail to find evidence of a clear "bias-variance" trade-off as I vary the number of clusters used to segment the product network. This suggests that the extent to which the bias (RMSE) of TATE estimates decreases (increases) under GCR is not solely mediated by how finely the product network is partitioned, and may also depend on the structure of the underlying network and the particular clustering algorithm that is used.

The fact that the methods I investigate in this paper reduce bias, but at the cost of excessive variance, suggests that they may be appropriate in contexts where statistical power is high (i.e., sample size is large), the experimenter anticipates that the magnitude of test-control interference will be large, and/or there are significant concerns that bias from test-control interference may flip the sign of the TATE estimate. However, there are many cases in which the reduction in bias may not be worth the increase in RMSE. In some cases, experiment designers may prefer alternative designs, such as switchback experiments (Sneider et al. 2019, Bojinov et al. 2020) or two-sided randomization (TSR) (Johari et al. 2020). Although switchback experiment designs do not, in general, cause excessive variance, they can introduce new sources of bias in contexts where intertemporal spillovers are a significant concern. Although the TSR design can reduce bias without increasing RMSE to such a significant degree, TSR is not feasible for treatment interventions that cannot be applied at the joint shopper-product level. In other cases, in lieu of a randomized experiment, researchers and practitioners may be able to estimate the effect of a proposed treatment intervention using BLP-style demand estimation techniques (Berry et al. 1995, Nevo 2000). However, there are many cases in which this approach may be undesirable or infeasible, due to the expertise required to estimate structural models, concerns about specification error, treatment interventions that are not easily modeled in BLP-style frameworks, and/or the difficulty of using structural models to analyze markets with supply constraints and large amounts of product heterogeneity (Conlon and Mortimer 2013, Farronato and Fradkin 2018).

The structure of this paper is as follows. In Section 2.2, I review the experiment designs and analysis methods to be evaluated via simulation, and discuss how these methods can be adapted to the online marketplace setting. Section 2.3 provides a summary of the Airbnb dataset used, the network generation process, and my approach to clustering the network. I describe the simulation process in Section 2.4, and present my results in Section 2.5. In Section 2.6, I discuss my findings and in Section 2.7, I conclude.

2.2 Theoretical Motivation

In this section, I provide a description of the network interference problem, a brief overview of prior work that aims to address this issue in the context of networked experiments, and the intuition for how that prior work can be adapted to the setting of online marketplace experiments.[2]

The focus of this paper is experiments that aim to estimate the total average treatment effect (TATE), which is also sometimes referred to as the global average treatment effect (GATE). Conceptually, the TATE measures the difference between the average outcome in the population under the counterfactual in which every unit receives the treatment, and the average outcome in the population under the counterfactual in which every unit receives the control. Formally, consider an experiment with N units, and let Z be a vector of length N indicating each unit's treatment assignment (e.g., in the case of binary treatment, Z_i might be 1 if unit i is assigned to the treatment, and 0 if unit i is assigned to the control). The response of each unit i is a function of Z, and can be written as $Y_i(Z)$. In this case, the TATE estimand can be written as

$$\tau(z_1, z_0) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[Y_i(Z = z_1) - Y_i(Z = z_0)\right], \tag{2.1}$$

where z_1 is a treatment assignment vector in which all units receive the treatment, and z_0 is a treatment vector in which all units receive the control.

Non-networked experiments typically rely on the stable unit treatment value assumption (SUTVA) (Rubin 1974), which is also sometimes referred to as the no interference assumption (Cox 1958) or the individualistic treatment response assumption (Manski 2013). SUTVA requires that the response of a particular unit in an experiment relies only on the treatment delivered to that unit, and not on the treatment delivered to other units in the experiment. Put differently, SUTVA requires that $Y_i(Z) = Y_i(Z_i)$, i.e., unit i's response does not depend on any element of Z besides

[2] For a more detailed overview of the prior work on this issue in the context of networked experiments, I refer the reader to Eckles et al. (2017).

the ith element. Under SUTVA, Eq. 2.1 can be rewritten as

$$\tau(z_1, z_0) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_i(Z_i = 1) - Y_i(Z_i = 0)\right]. \tag{2.2}$$

When treatment is assigned to individual units randomly, with N_T units receiving the treatment and N_C units receiving the control, the standard difference-in-means average treatment effect (ATE) estimator

$$\hat{\mu} = \frac{1}{N_T}\sum_{i=1}^{N} Y_i \mathbb{1}(Z_i = 1) - \frac{1}{N_C}\sum_{i=1}^{N} Y_i \mathbb{1}(Z_i = 0) \tag{2.3}$$

provides an unbiased estimate of both the ATE and the TATE, as the two are equivalent. However, in cases where SUTVA is violated (i.e., there is test-control interference), the difference-in-means ATE estimator will provide a biased estimate of the TATE, since neither $Y_i(Z = z_1)$ nor $Y_i(Z = z_0)$ is observed for any unit i in the experiment. To gain some intuition for why this is the case, consider a researcher or policymaker who hopes to measure the effect of a coordinated marketing campaign delivered to *everyone* that raises awareness of a new job training program offered by the government. In a randomized experiment, some of those in the control group, who are not targeted by the campaign, may still be made aware of the job training program by contacts of theirs in the treatment group. Because of this "interference" between the control group and treatment group, the difference between program enrollment rates in the two arms of the experiment may not be equal to the difference between the enrollment rate had everyone been targeted by the marketing campaign and the enrollment rate had no one been targeted by the marketing campaign.

In order to make estimating the TATE via randomized experiment more tractable, I can introduce the assumption that unit i's treatment response, $Y_i(Z)$, depends only on a subset of the elements of Z, as opposed to the full treatment assignment vector. For instance, in a networked setting, unit i's response might only depend on Z_i and $Z_{N(i)}$, where $N(i)$ is the set of i's network neighbors and $Z_{N(i)}$ are the entries of Z corresponding to those neighbors. Such an assumption can be referred to as a constant treatment response (CTR) assumption (Manski 2013) or as an exposure model (Ugander et al. 2013, Aronow et al. 2017). More formally, an exposure model can be written as a function $g_i(\cdot)$: $Z^N \rightarrow G_i$ that maps global treatment assignment vectors to the space G_i of effective treatments for node i. Under a given exposure model $g_i(\cdot)$, $g_i(z_m) = g_i(z_n)$ implies that the treatment vectors z_m and z_n provide unit i with the same "effective treatment." For instance, under an exposure model in which unit i's effective treatment is a function of the treatment assignments for unit i and its neighbors, $N(i)$, two global treatment assignment vectors that contain different treatment assignments for unit $j \notin N(i)$ would provide unit i with the same effective treatment so long as the treatment assignments for i and $N(i)$ were all the same.

One experiment design method for reducing bias in TATE estimates is graph cluster randomization (GCR) (Ugander et al. 2013, Eckles et al. 2017), in which the network is partitioned into clusters, and then randomized at the cluster-level, as opposed to the unit-level. Even when an exposure model is not explicitly specified, GCR will still increase the percentage of units experiencing the "effective treatment" or "effective control" under a large number of reasonable exposure models, and should reduce the amount of bias due to test-control interference. More formally, in an individual-level Bernoulli randomized experiment, treatment assignment is randomized at the *unit* level by drawing each Z_i from a Bernoulli distribution

$$Z_i \sim \text{Bernoulli}(q), \tag{2.4}$$

where q is the probability of assignment to the treatment. Under GCR the network is first partitioned into N_C clusters: C_1, C_2, ..., C_{N_C}, after which each cluster is given a cluster-level treatment assignment, W_J, which is drawn from a Bernoulli distribution

$$W_j \sim \text{Bernoulli}(q). \tag{2.5}$$

After cluster-level treatment assignments are determined, each node in a given cluster is assigned its cluster's treatment assignment.

Bias can also be reduced in the analysis stage of an experiment by assuming a specific exposure model $g(i)$, and modifying the ATE estimand found in Eq. 2.1 to incorporate that exposure model, such that the revised estimand is

$$\tau(z_1, z_0)_{eff} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_i(Z)|g_i(Z) = g_i(z_1)\right] - \mathbb{E}\left[Y_i(Z)|g_i(Z) = g_i(z_0)\right]. \tag{2.6}$$

If the specified exposure model is correct, then this estimand is equivalent to the TATE. However, Eckles et al. (2017) show that even in cases where the specified exposure model is incorrect, under certain monotonicity assumptions, the more "restrictive" an exposure model is, the lower the bias of Eq. 2.6 will be relative to the TATE. In this paper, I will assess the effectiveness of fractional neighborhood treatment response (FNTR) exposure models (Ugander et al. 2013, Aronow et al. 2017, Eckles et al. 2017). In FNTR exposure models, a node is in the effective treatment (control) if it is treated (not treated) and more than some fraction λ of its neighbors are in the treatment (control). The "restrictiveness" of a given FNTR exposure model is determined by the parameter λ: the higher this threshold, the more restrictive the exposure model. In this paper, I consider $\lambda = .50$, $\lambda = .75$, and $\lambda = .95$.

A simple estimator for the ATE estimand in Eq. 2.6 would be the difference in the sample means for units in the effective treatment and effective control,

$$\hat{\tau}_{eff} = \frac{\sum_{i=1}^{N} Y_i \mathbb{1}(g_i(Z) = g_i(z_1))}{\sum_{i=1}^{N} \mathbb{1}(g_i(Z) = g_i(z_1))} - \frac{\sum_{i=1}^{N} Y_i \mathbb{1}(g_i(Z) = g_i(z_0))}{\sum_{i=1}^{N} \mathbb{1}(g_i(Z) = g_i(z_0))}. \tag{2.7}$$

However, this estimator will only be unbiased for Eq. 2.6 if the effective treatments are unconfounded, i.e., if the expected outcome of unit i under a particular effective treatment is independent of the probability that unit i receives that effective treatment. More formally, for Eq. 2.7 to be an unbiased estimator for Eq. 2.6, it is required that

$$E\left[Y_i|g_i(Z) = g_i(z)\right] \perp \mathbb{P}\left[g_i(Z) = g_i(z)\right]. \tag{2.8}$$

Unfortunately, this unconfoundedness condition is unlikely to hold in most cases.

One way for Eq. 2.8 to hold would be if either $E\left[Y_i|g_i(Z)=g_i(z)\right]$ or $\mathbb{P}\left[g_i(Z)=g_i(z)\right]$ were homogeneous across the population. However, it is extremely unlikely that $E\left[Y_i|g_i(Z)=g_i(z)\right]$ is the same for all i, and $\mathbb{P}\left[g_i(Z)=g_i(z)\right]$ will vary across the population for many exposure models of interest; for instance, in an experiment in which treatment is randomized at the individual-level, higher degree units will have a lower probability of receiving effective treatment and effective control under any FNTR exposure model. In cases where both $E\left[Y_i|g_i(Z)=g_i(z)\right]$ and $\mathbb{P}\left[g_i(Z)=g_i(z)\right]$ vary across the population, it is also easy to imagine plausible cases in which the two are correlated. For instance, one could imagine that higher degree units in a network have systematically different potential outcomes than lower degree units in a network, even conditional on being exposed to the same effective treatment. Under the aforementioned experiment in which treatment is delivered at the individual-level, this would result in a clear violation of Eq. 2.8. Even under a graph-randomized experiment design, this condition could be violated, as higher degree units may be more likely to have edges that span multiple clusters, which would also result in a lower probability of being assigned to the effective treatment or effective control conditions (Ugander et al. 2013).

One way to unconfound effective treatments is to condition analysis on the design of the experiment, which determines the probability $\pi_i(z)=\mathbb{P}\left[g_i(Z)=g_i(z)\right]$ that a given unit is assigned to the effective treatment (z_1) or the effective control (z_0). These probabilities can then be used to calculate treatment effect estimators such as the Horvitz-Thompson estimator (Horvitz and Thompson 1952),

$$\hat{\tau_{HT}}(z_1,z_0)=\frac{2}{N}\left(\sum_{i=1}^{N}\frac{Y_i\mathbb{1}\left[g_i(Z)=g_i(z_1)\right]}{\pi_i(z_1)}-\sum_{i=1}^{N}\frac{Y_i\mathbb{1}\left[g_i(Z)=g_i(z_0)\right]}{\pi_i(z_0)}\right), \tag{2.9}$$

and other inverse-probability weighted treatment effect estimators. The Horvitz-Thompson estimator often suffers from excessive variance, so in this paper, I follow Aronow et al. (2017) and Eckles et al. (2017) and use the Hajek estimator (Hájek 1971),

$$\hat{\tau_H}(z_1, z_0) = \left(\sum_{i=1}^{N} \frac{\mathbb{1}\left[g_i(Z) = g_i(z_1)\right]}{\pi_i(z_1)}\right)^{-1} \sum_{i=1}^{N} \frac{Y_i \mathbb{1}\left[g_i(Z) = g_i(z_1)\right]}{\pi_i(z_1)} - \left(\sum_{i=1}^{N} \frac{\mathbb{1}\left[g_i(Z) = g_i(z_0)\right]}{\pi_i(z_0)}\right)^{-1} \sum_{i=1}^{N} \frac{Y_i \mathbb{1}\left[g_i(Z) = g_i(z_0)\right]}{\pi_i(z_0)}, \tag{2.10}$$

to estimate treatment effects under exposure models. The Hajek estimator has non-zero bias, but the magnitude of this bias is typically small and worth the variance reduction relative to the Horvitz-Thompson estimator (Aronow et al. 2017, Eckles et al. 2017).

Although the variance of the Hajek estimator is lower than the variance of the Horvitz-Thompson estimator, all of the techniques discussed (graph cluster randomization, exposure models, inverse-probability weighted treatment effect estimators) increase the variance of TATE estimates relative to the baseline of individual-level, Bernoulli randomized experiments (Gerber and Green 2012). One way to offset the loss of precision, due to graph cluster randomization in particular, is to use a block random assignment scheme (Gerber and Green 2012, Moore 2012), such as a matched pair design. Under the matched pair design, units of randomization (in the case of GCR, clusters) are arranged into blocks of size $b = 2$ in such a way that the two units that make up each pair are as similar to one another as possible. Within each block, one unit is assigned to the treatment, and the other is assigned to the control.

The tools above (graph cluster randomization, exposure models, and inverse probability weighted treatment effect estimators) have been developed and utilized primarily in the context of network experiments. However, SUTVA violations that bias TATE estimates are not specific to networks; there is mounting evidence that 'test-control interference' frequently occurs in marketplace experiments as well, both because the items offered by different sellers substitute for and/or complement one another, and because in a market with supply constraints, shoppers "compete" to purchase inventory. For instance, Blake and Coey (2014) analyze an email marketing experiment performed on eBay and conclude that naive estimates of the TATE ignoring test-control interference exaggerate the treatment's effectiveness by about a factor of 2, and using a simulation based approach, Fradkin (2015) finds that search algorithm experiments in online marketplaces can overstate true treatment effects by over 100%.

Both Blake and Coey (2014) and Fradkin (2015) propose methods for combating bias due to test-control interference in marketplace experiments. Rather than compare outcomes across buyers, Blake and Coey (2014) limit their analysis to auctions where a majority of bidders are in the treatment or control, and compare outcomes across auctions. While this does eliminate within-auction interference between users, there is still potential bias in their effect size estimates due to interference across auctions. Furthermore, it is not clear how well this methodology generalizes to marketplaces that do not offer convenient units of analysis (e.g., auctions) over which to aggregate outcomes.

Fradkin (2015) proposes randomizing across well-defined markets (rather than at the shopper- or seller-level), or combining experimental results with results from structural model-based simulations. However, for many marketplaces, market-level experiments can be infeasible, due to ambiguous or non-existent market definitions, the existence of only a small number of markets over which to randomize, or high levels of market heterogeneity. Furthermore, developing structural model-based simulations may be beyond the scope of what is feasible for many firms conducting marketplace experiments, particularly those that are unable to hire large numbers of economists trained in structural modeling and simulation development.

In this paper, I propose that methods such as graph cluster randomization, exposure modeling, and treatment effect estimators such as the Hajek estimator can be adapted from the context of networks to the context of online marketplaces, and then test, via simulation, the efficacy of these methods at reducing bias from test-control interference in TATE estimates in this new context. The key step in adapting these tools to the online marketplace context is inferring "edges" between sellers/items that are likely to substitute for or complement one another. Many different approaches could be used to infer these edges, including the use of simple heuristics around item attributes such as price point, category, and physical location, the analysis of preexisting search, view, or purchase data, and the repurposing of existing "recommendation networks" that capture the items recommended to users that view/purchase a particular item. Prior research in the information systems literature has shown that "product networks" can be an effective tool to study competition and demand spillovers in online marketplaces. Oestreicher-Singer and Sundararajan (2012b) empirically test the hypothesis that a visible co-purchase or co-view link between two books on Amazon.com increases their demand correlation, and find that such links lead to a threefold increase in the influence that complementary products have on each others' demand levels. In a later paper, Oestreicher-Singer and Sundararajan (2012a) find that the extent to which peer-based recommendations in online marketplaces redirect demand to "niche" products in the long tail depends on the network structure of co-purchase or co-view "edges" between items in a given category. Extending this research, Dhar et al. (2014) find that when time series data is available, product networks can be used for demand prediction.

Although previous work has shown that the methods presented in this paper are effective at reducing the bias of TATE estimates in more traditional "networked" settings, it is not obvious *ex ante* that they will be effective in online marketplace settings, given that the underlying phenomena that cause spillovers (e.g., substitution and complementarity) are different, and that in many cases, the researcher will be required to construct a network based on inferred edges because there is not an explicit and/or definitive "network" that connects items or sellers in a marketplace.

2.3 Data & Network Construction

The 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.

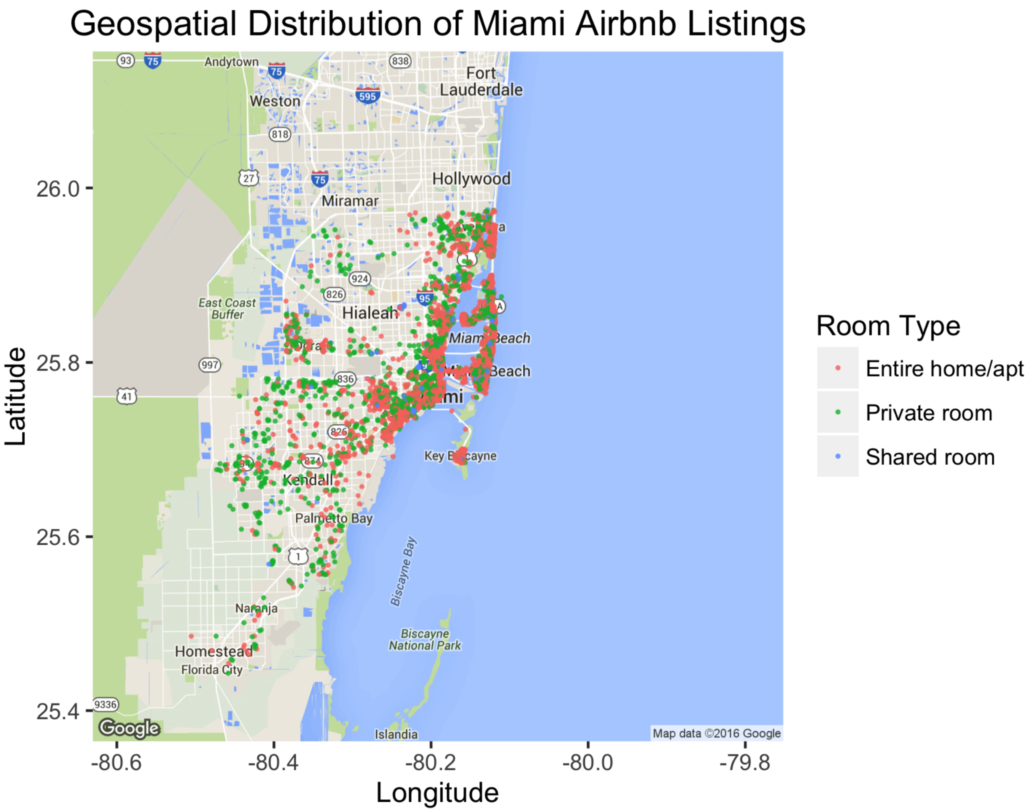


Figure 2.1: The geospatial distribution of Airbnb listings in and around Miami. Color corresponds to listing type. This figure was produced with ggmap (Kahle and Wickham 2013).

Table 2.1: Summary of Airbnb listing covariates

Statistic	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

Figure 2.1 depicts the geospatial distribution of the listings by room type, and Table 2.1 provides information about the distribution of listing-level covariates across the sample of Airbnb listings.

Before using the dataset for my analyses, I 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. I also assign each listing j in my dataset an unobservable quality component,

$$\xi_j \sim N(0,1), \tag{2.11}$$

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

I 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. I generate an edge between two listings when the following three criteria are satisfied:[3]

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

[3] 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, Srinivasan (2018) cluster items in an online marketplace based on how often they co-occur in search results.

Using the edge heuristic described above, I produce a network that has 1,538,637 edges, and a clustering coefficient of 0.74. The left pane of Figure 2.2 shows the degree distribution; the average degree of nodes in the network is 173.76.

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

$$Q = \frac{1}{2E} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2E} \right) \mathbb{1}(C_i = C_j), \tag{2.12}$$

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 my listing network, the network is partitioned into 169 clusters, which have an average size of 52.40 listings. The right pane of Figure 2.2 shows the distribution of cluster size for the 169 clusters.

As noted in Section 2.2, graph cluster randomization and the use of exposure models can increase the variance of TATE estimates. In order to counteract this increase in variance, my simulated GCR 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, I 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, I 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.

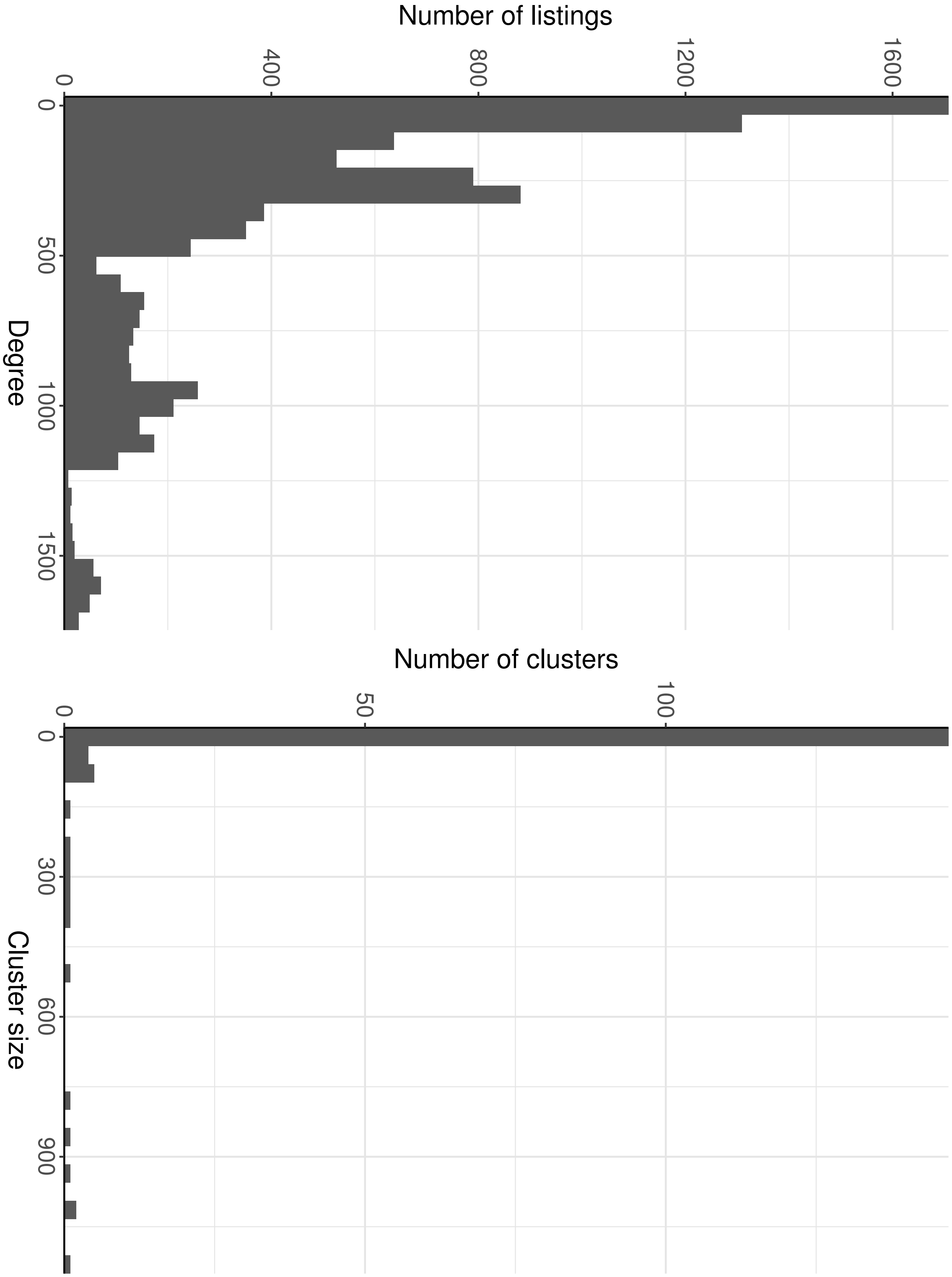


Figure 2.2: The left pane shows the degree distribution for the Airbnb listing network generated using the procedure described in Section 2.3. The right pane shows the distribution of cluster sizes across the 169 clusters generated using Louvain clustering (Blondel et al. 2008).

2.4 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, I 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.,

$$\begin{aligned} &\delta_{k0} \sim U[0,1] \text{ for } k = 1, 2, 3, ..., 9, \\ &\delta_k = \frac{\delta_{k0}}{\sum_j \delta_{k0}}. \end{aligned} \tag{2.13}$$

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.,

$$\text{Search Score}_j = \delta \cdot \mathbf{x}_j. \tag{2.14}$$

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 my market, 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 market to the 99.75th percentile of latitudes (longitudes) belonging to listings in the market.[4] 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 my simulation framework is comparable to models used in the demand estimation literature (e.g., Berry et al. (1995) and Nevo (2000)):

$$u_{ij} = \alpha_i(y_i - p_j) + \tilde{\mathbf{x}}_{\mathbf{j}}\boldsymbol{\beta}_i + \xi_j + \epsilon_{ij}, \tag{2.15}$$

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

$$\begin{aligned} y_i &\sim N(0,1) \\ \alpha_i &\sim N(0,1) \\ \beta_{ik} &\sim N(0,1)\,\forall\, k \\ \epsilon_{ij} &\sim f(x) = e^{-x}e^{e^{-x}} \text{ (the Type I extreme-value distribution).} \end{aligned} \tag{2.16}$$

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 (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

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

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, I believe it can still provide insight into the degree to which test-control interference may bias TATE estimates in online marketplace experiments, and can help determine the extent to which the proposed experiment designs and analysis techniques reduce that bias. In Appendix 2.A, I repeat my main analysis using a modified version of the simulation framework above, in which searcher preferences are drawn from different distributions. My results are qualitatively similar, suggesting that the specifics of my simulation framework are not a significant driver of my results.

2.5 Results

Using the simulation framework outlined in the previous section, I am able to conduct simulations of market activity both under market-wide policy regimes (i.e., 100% treatment and 100% control), as well as under different experiment designs. I compare the ground truth TATEs generated by contrasting outcomes under market-wide policy changes to the TATE estimates produced by different experiment designs and analysis techniques, and calculate the bias and root mean square error (RMSE) of the TATE estimates produced under different approaches to experiment design and analysis.

In each of my simulations of market activity, I am interested in two different outcomes. The first is whether or not a listing was booked, which I write as B_i. If a listing is booked, $B_i = 1$, otherwise, $B_i = 0$. The second is the amount of revenue earned by a listing. If listing i charges price p_i, then revenue will be $B_i \times p_i$. More formally, I can denote the TATE for listing bookings as

$$TATE_{\text{bookings}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[B_i(Z = z_1) - B_i(Z = z_0)\right] \tag{2.17}$$

and the TATE for listing revenue as

$$TATE_{\text{listing revenue}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[p_i(Z=z_1)\times B_i(Z=z_1) - p_i(Z=z_0)\times B_i(Z=z_0)\right]. \tag{2.18}$$

I choose these two outcomes because they are natural outcomes for both an academic researcher and a marketplace designer to be interested in when conducting online marketplace experiments. Listing revenue will also in general be a higher variance outcome than whether or not a listing is booked, so looking at both outcomes will enable us to get a sense of how my proposed techniques perform for outcome metrics with different levels of baseline variability.

I also consider two different types of treatment intervention. The first is a price reduction of .75 standard deviations for treated listings. Pricing experiments of this sort are common for online marketplaces, e.g., to estimate price elasticities (e.g., Li et al. (2015) and Holtz et al. (2020)), but in some cases could be superseded by structural models or simulations, such as the one developed in Fradkin (2015). The second is an increase of .75 standard deviations in the unobserved quality of listings. This type of treatment intervention is also common in online marketplaces, and might correspond to platform interventions that induce various difficult-to-observe changes to seller behavior (e.g., seller responsiveness or friendliness) or item quality (e.g., photo quality, review text). Treatment interventions of this sort are arguably harder to model using traditional demand estimation techniques.

2.5.1 Simulating Ground Truth

Table 2.1: True market-wide outcome distributions and TATEs

Treatment	Metric	μ_C	σ_C	μ_T	σ_T	t	p	TATE
Unobserved quality change	Listing booked	0.089	0.002	0.092	0.002	21.63	$\leq 2.2 \times 10^{-16}$	0.003
Unobserved quality change	Listing revenue	20.41	4.379	21.03	4.533	2.17	0.03	0.612
Price reduction	Listing booked	0.089	0.002	0.092	0.002	17.27	$\leq 2.2 \times 10^{-16}$	0.002
Price reduction	Listing revenue	20.41	4.379	20.87	4.490	1.63	0.1	0.458

I first use my simulation framework to simulate the distribution of market-level average outcomes in the case in which 100% of listings receive the treatment, and

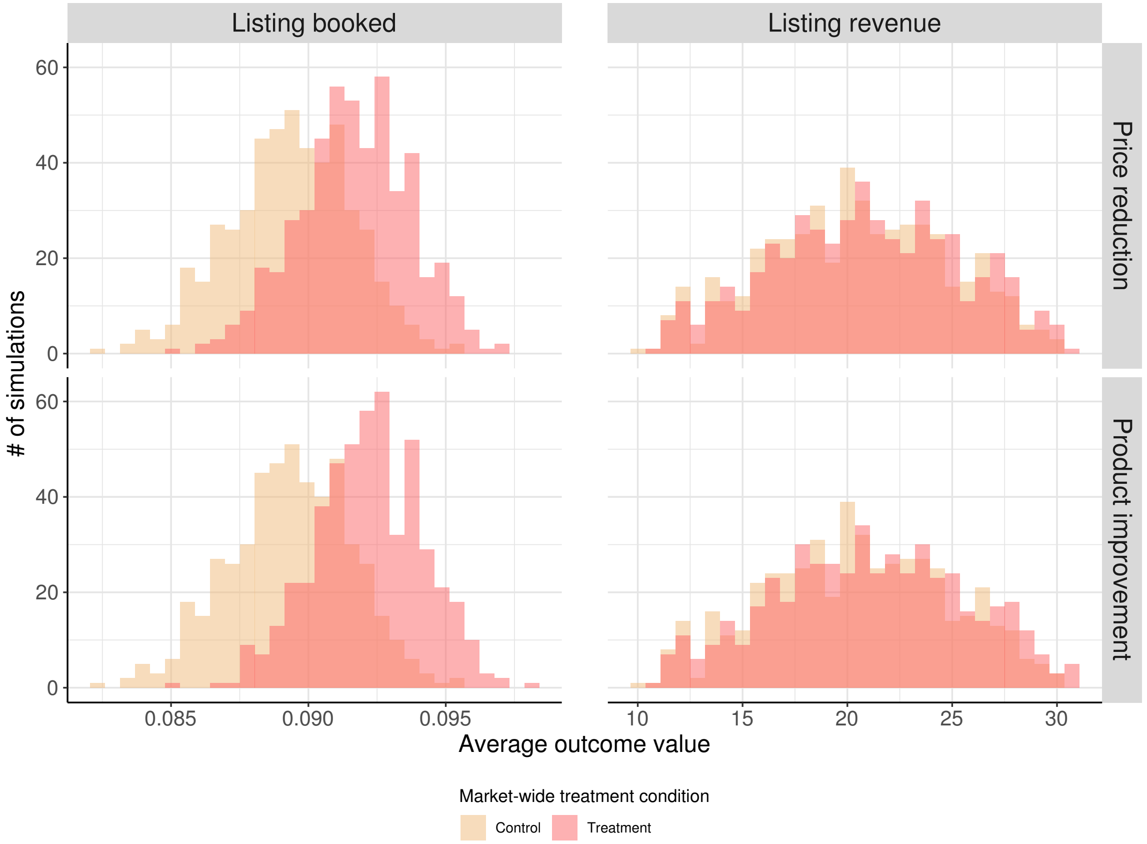


Figure 2.1: Comparison of market-wide average outcomes when either 0% or 100% of listings are assigned treatment. The top row shows distributions when the treatment is the price reduction treatment. The bottom row shows distributions when the treatment is the unobserved listing quality change treatment. The left column shows distributions for the listing booked outcome. The right column shows distributions for the listing revenue outcome.

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, I conduct 500 simulations of one night of booking activity in which 1,000 searchers visit Airbnb. Figure 2.1 compares the sampling distributions of the rate of listings being booked and the average listing revenue under all three conditions. my results are also summarized in Table 2.1.

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, I am 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.

2.5.2 Measuring bias and RMSE

Having simulated the distribution of market-level outcomes under both 100% treatment and 100% control for both my price reduction treatment and my unobservable listing quality treatment, I can now use my simulation framework to estimate the bias and RMSE of different combinations of experiment design, choice of exposure model, and treatment effect estimator for both treatments. I first use my framework to simulate 500 Bernoulli randomized individual-level 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, graph cluster randomized experiments that use the clusters described in Section 2.3. Under this design, I calculate the difference in means estimator for all listings, the difference in means estimator for units satisfying the FNTR exposure model with three different thresholds ($\lambda = .5$, $\lambda = .75$, and $\lambda = .95$), and the Hajek estimator for units satisfying the FNTR exposure model with the same three thresholds. Each of my simulated experiments emulate one night of booking activity in which 1,000 searchers visit Airbnb. Figure 2.2 summarizes the distributions of TATE estimates for each combination of experiment design, exposure model, and treatment effect estimator, as well as the *actual* distribution of treatment effects observed in my simulation framework.

Table 2.2 shows the bias and RMSE of each combination of design, exposure model, and treatment effect estimator for the booking outcome, under both the price reduction treatment and the unobserved listing quality treatments. Table 2.3 shows the same information for the listing revenue outcome under both treatments. These results are also summarized in Figure 2.3. Relative to the difference in means estimator under the Bernoulli individual-level experiment, I find that the difference in means estimator under blocked GCR reduced bias by as much as 64.5%, across both metrics and both types of treatment. However, this comes 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. Under blocked GCR, the use of

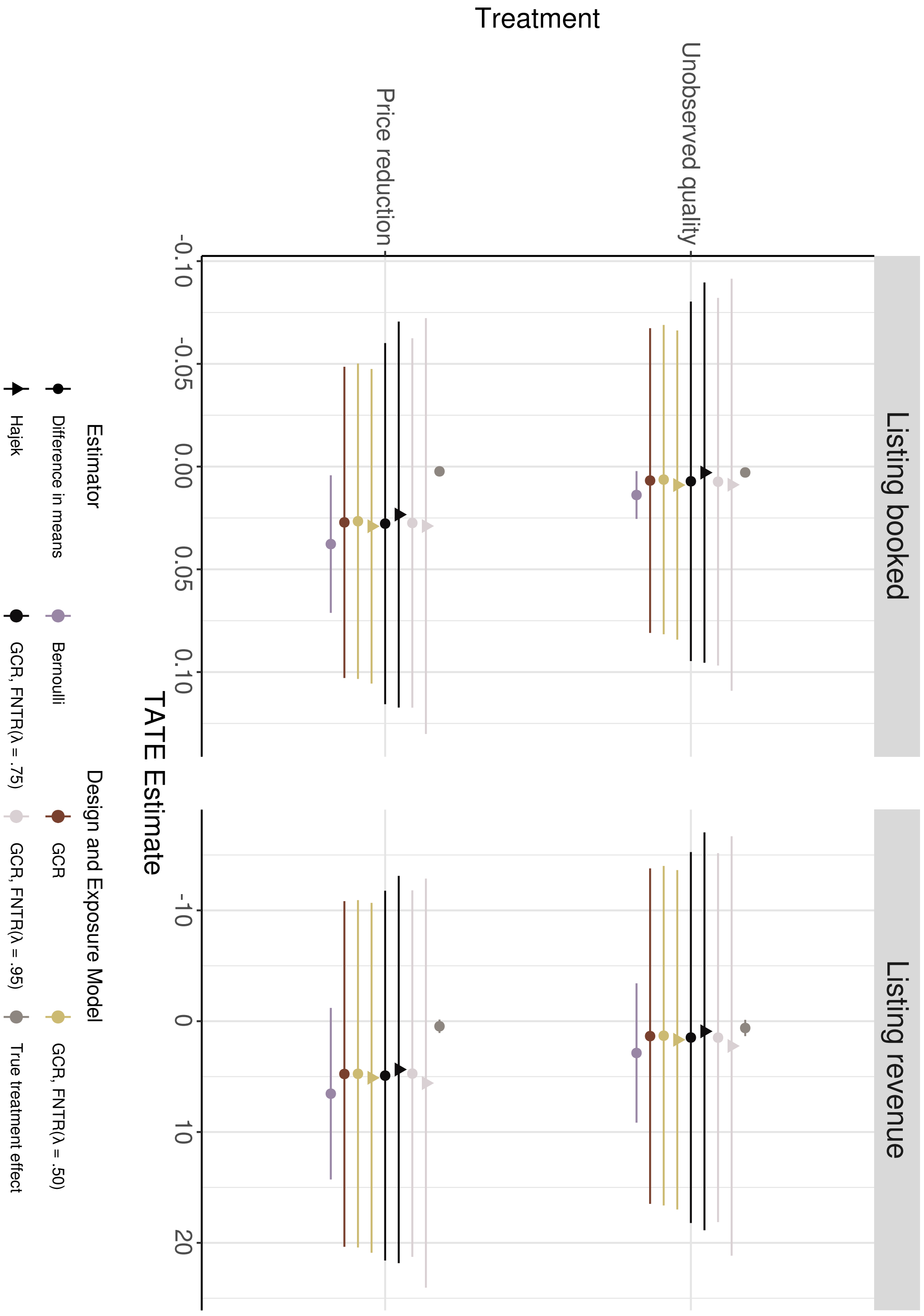


Figure 2.2: The mean TATE, as well as the 2.5th percentile and 97.5th percentile of the TATE distribution, for different combinations of experiment design, exposure model, and treatment effect estimator. All combinations reduce bias relative to the Individual-level Bernoulli experiment with difference in means, but at the cost of much higher RMSE.

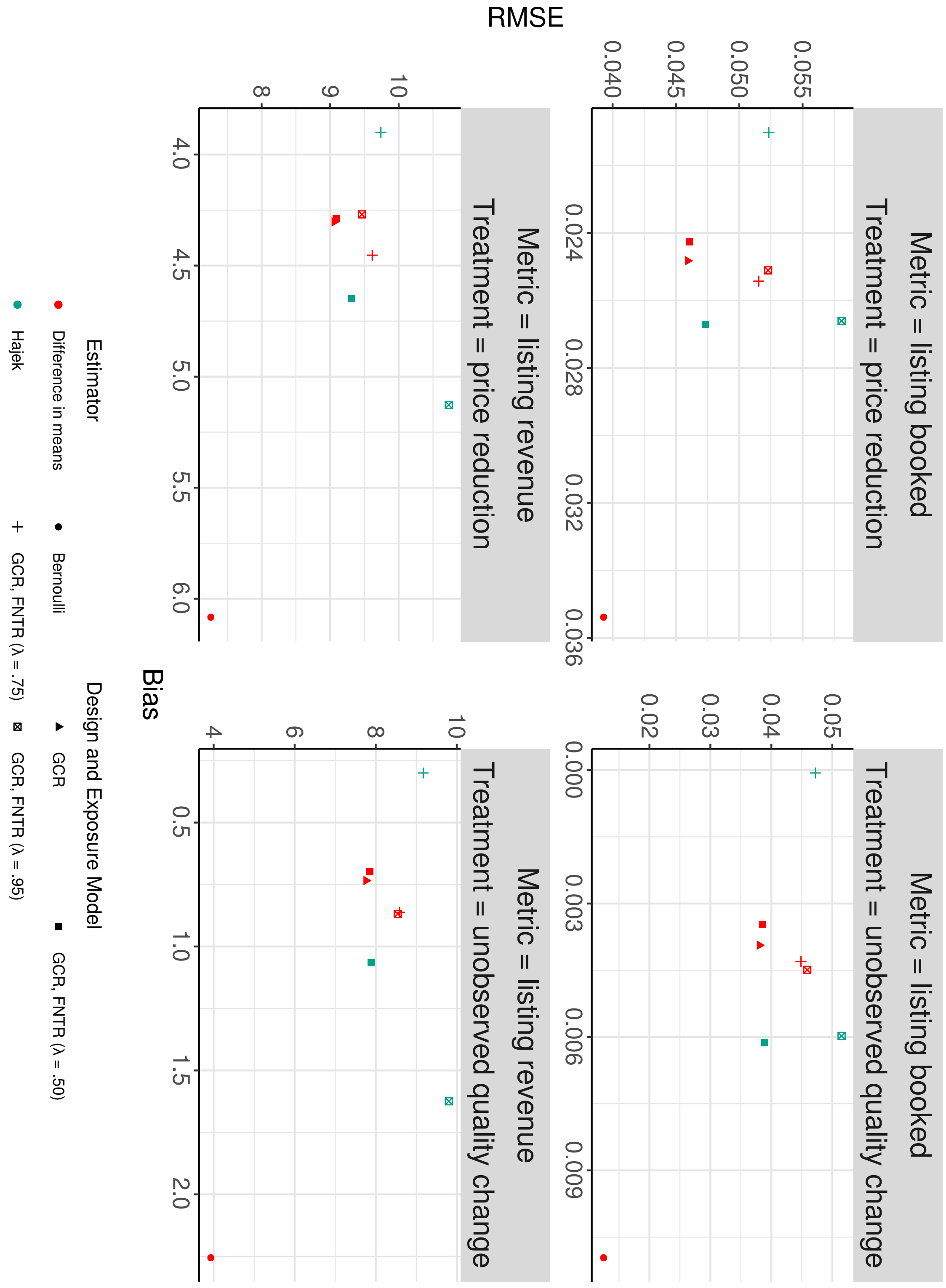


Figure 2.3: The bias and RMSE of different combinations of experiment design, exposure model, and treatment effect estimator, for both outcomes and both treatments.

)

the FNTR exposure model sometimes lowers bias further, but also sometimes leads to higher RMSE. This appears to be true, regardless of whether the difference in means estimator or the Hajek estimator is used in conjunction with the FNTR exposure model. Just how much lower the bias is under the FNTR exposure model, regardless of the choice of treatment effect estimator, appears to be sensitive to the choice of FNTR threshold, and it may be difficult for an experiment designer to determine the optimal threshold prior to their experiment.

Table 2.2: Performance comparison: outcome = bookings

Treatment	Design	Estimator	Bias	RMSE	Coverage
Price Reduction	Bernoulli	Difference in means	0.0354	0.0393	6%
Price Reduction	GCR	Difference in means	0.0248	0.0459	20%
Price Reduction	GCR	Regression + clustered S.E.	0.0248	0.0459	95%
Price Reduction	GCR, FNTR (λ = .50)	Difference in means	0.0243	0.0461	20%
Price Reduction	GCR, FNTR (λ = .50)	Hajek	0.0267	0.0473	100%
Price Reduction	GCR, FNTR (λ = .75)	Difference in means	0.0254	0.0515	19%
Price Reduction	GCR, FNTR (λ = .75)	Hajek	0.0210	0.0523	100%
Price Reduction	GCR, FNTR (λ = .95)	Difference in means	0.0251	0.0523	19%
Price Reduction	GCR, FNTR (λ = .95)	Hajek	0.0266	0.0581	100%
Unobserved quality	Bernoulli	Difference in means	0.0110	0.0125	56%
Unobserved quality	GCR	Difference in means	0.0039	0.0381	23%
Unobserved quality	GCR	Regression + clustered S.E.	0.0039	0.0381	99%
Unobserved quality	GCR, FNTR (λ = .50)	Difference in means	0.0035	0.0386	22%
Unobserved quality	GCR, FNTR (λ = .50)	Hajek	0.0061	0.0389	100%
Unobserved quality	GCR, FNTR (λ = .75)	Difference in means	0.0043	0.0449	22%
Unobserved quality	GCR, FNTR (λ = .75)	Hajek	0.0001	0.0472	100%
Unobserved quality	GCR, FNTR (λ = .95)	Difference in means	0.0045	0.0459	22%
Unobserved quality	GCR, FNTR (λ = .95)	Hajek	0.0060	0.0515	100%

Table 2.3: Performance comparison: outcome = listing revenue

Treatment	Design	Estimator	Bias	RMSE	Coverage
Price Reduction	Bernoulli	Difference in means	6.08	7.26	40%
Price Reduction	GCR	Difference in means	4.30	9.06	47%
Price Reduction	GCR	Regression + clustered S.E.	4.30	9.06	97%
Price Reduction	GCR, FNTR (λ = .50)	Difference in means	4.29	9.09	44%
Price Reduction	GCR, FNTR (λ = .50)	Hajek	4.65	9.31	100%
Price Reduction	GCR, FNTR (λ = .75)	Difference in means	4.45	9.61	45%
Price Reduction	GCR, FNTR (λ = .75)	Hajek	3.90	9.74	100%
Price Reduction	GCR, FNTR (λ = .95)	Difference in means	4.27	9.46	46%
Price Reduction	GCR, FNTR (λ = .95)	Hajek	5.13	10.73	100%
Unobserved quality	Bernoulli	Difference in means	2.26	3.93	86%
Unobserved quality	GCR	Difference in means	0.73	7.76	49%
Unobserved quality	GCR	Regression + clustered S.E.	0.73	7.76	100%
Unobserved quality	GCR, FNTR (λ = .50)	Difference in means	0.70	7.85	46%
Unobserved quality	GCR, FNTR (λ = .50)	Hajek	1.06	7.89	100%
Unobserved quality	GCR, FNTR (λ = .75)	Difference in means	0.86	8.59	46%
Unobserved quality	GCR, FNTR (λ = .75)	Hajek	0.30	9.17	100%
Unobserved quality	GCR, FNTR (λ = .95)	Difference in means	0.87	8.54	48%
Unobserved quality	GCR, FNTR (λ = .95)	Hajek	1.62	9.80	100%

In Section 2.6, I further discuss the circumstances under which a researcher and/or practitioner may find that bias reduction is important enough to accept increased RMSE, and also briefly discuss alternative solutions to the issue of test-control interference that may be preferable in cases where a dramatic increase in RMSE is not acceptable.

2.5.3 Statistical Inference

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

$$\hat{\sigma_\tau^2} = \sigma^2(Y_{iT}) + \sigma^2(Y_{iC}), \tag{2.19}$$

where $\sigma^2(Y_{iT})$ and $\sigma^2(Y_{iC})$ are the variance of outcomes in the treatment group and control group, respectively. To estimate the variance of the Hajek estimator, I use a linearized version of the conservative variance estimator for the Horvitz-Thompson estimator found in Aronow et al. (2017), in which I substitute residuals for the actual observed outcome values in my dataset. I also calculate the variance of the blocked GCR experiment design 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 the 95% confidence intervals are found in the rightmost columns of Tables 2.2 and 2.3. I find that the coverage probability of the difference in means estimator when used with the Bernoulli design is below the nominal 95% coverage in all cases, and can be as low as 6%. The blocked GCR 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; this is true both for the standard difference in means estimator and for all three FNTR exposure model variants. Both regression analysis of the blocked GCR design with clustered standard errors, and usage of the Hajek estimator in conjunction with the blocked GCR design and FNTR exposure models produce coverage probabilities that are greater than the nominal 95% coverage probability, ranging from 95% to 100%. These findings highlight the importance of statistical inference that takes into account the design of the experiment being analyzed.

2.5.4 Performance under network mis-specification

One concern about my proposed approach is that it might be sensitive to the particular approach that a given researcher uses to define the product network. The network used in my simulations is based on a listing's room type, location, and capacity, which are also three of the variables that are used by both the search algorithm and searchers. In other words, the network is "well specified." However, there might be cases in which researchers do not know the appropriate way to define an edge between two sellers or items. To test the robustness of my results to mis-specification of the product network, I estimated the performance of each experiment design, exposure model, and treatment effect estimator after rewiring random edges in the network with probabilities ranging from 1% to 15%. In addition to making the network more random and more mis-specified, this also reduces the degree to which the resulting network is clustered (Watts and Strogatz 1998). For instance, whereas the clustering coefficient of my original network is 0.74, the clustering coefficient of the network after randomly rewiring 15% of the edges is 0.37. I proceeded to cluster each of these rewired networks using the same Louvain clustering algorithm, and then repeated my main analysis using the resulting clusters.

Figure 2.4 shows how the bias of each design, exposure model, and estimator changed as a function of rewiring probability.[5] There does not appear to be a discernible relationship between rewiring probability and bias, and each combination of design, exposure model, and estimator continues to reduce bias of the TATE estimate relative to the difference in means estimator used in conjunction with individual-level Bernoulli randomization (represented in each pane by the dotted yellow line). Figure 2.5 shows how the RMSE changes as a function of rewiring probability. Again, there does not seem to be a strong relationship between rewiring probability and the

[5] No listings satisfied the requirement for the .95 FNTR threshold when the rewiring probability was .10 or .15.

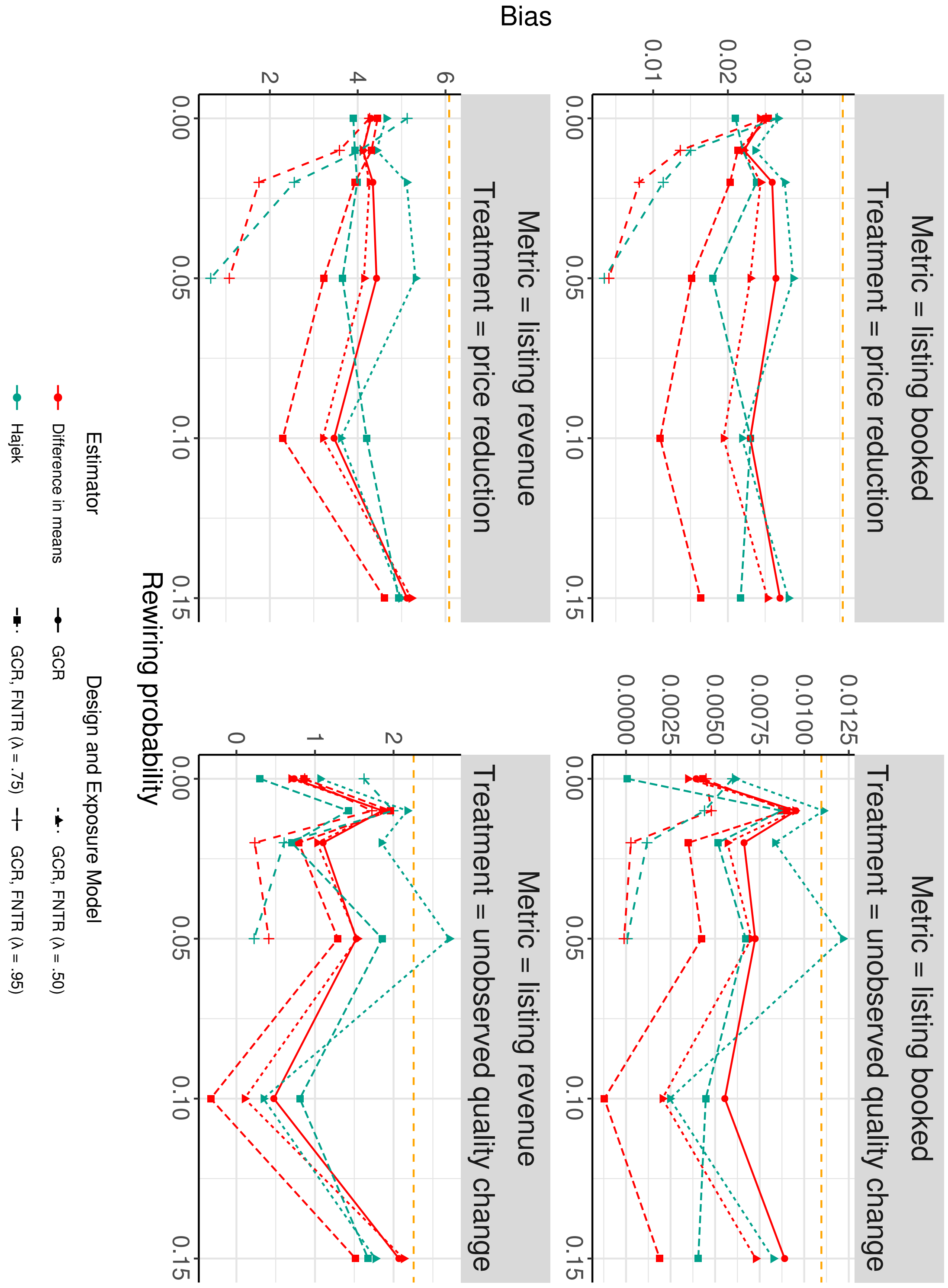


Figure 2.4: The bias of different combinations of experiment design, exposure model, and treatment effect estimator, for both outcomes and both treatments, as a function of the probability that a given edge in the product network is randomly rewired. The dashed yellow line indicates the bias of the individual-level, Bernoulli randomized experiment analyzed with the difference in means estimator.

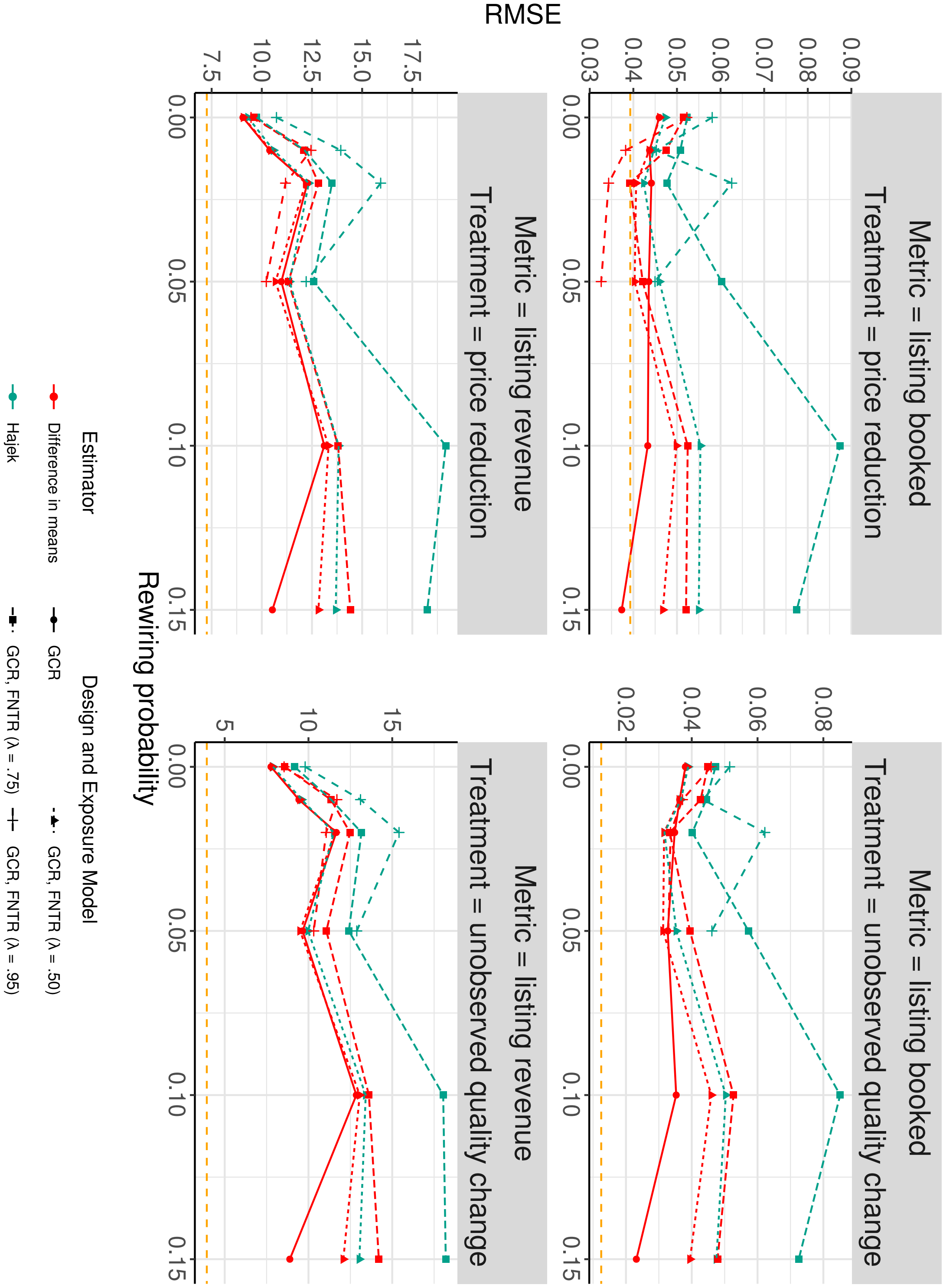


Figure 2.5: The RMSE of different combinations of experiment design, exposure model, and treatment effect estimator, for both outcomes and both treatments, as a function of the probability that a given edge in the product network is randomly rewired. The dashed yellow line indicates the RMSE of the individual-level, Bernoulli randomized experiment analyzed with the difference in means estimator.

RMSE of the TATE estimates produced under different combinations of experiment design, exposure model, and treatment effect estimator. Taken together, these results suggest that while the methods proposed in this paper continue to reduce bias in TATE estimates, even under non-trivial amounts of network mis-specification, they also continue to increase the variance of those TATE estimates.

2.5.5 Performance with varying cluster sizes

One possibility is that the "bias-variance" trade-off that exists in my results, i.e., the fact that the proposed methods do reduce bias, but increase variance and RMSE, can be controlled by changing the number of clusters produced by the algorithm used to segment the product network. In other words, with more clusters, it might be the case that bias is still reduced, albeit by a smaller amount, but with a smaller cost in terms of increased variance. To probe the extent to which this might be true, I cluster my network using a greedy algorithm that finds community structure in networks by optimizing modularity (Clauset et al. 2004). One benefit of this algorithm relative to the Louvain clustering method I deploy elsewhere in the paper is that it produces a hierarchical "dendrogram" that can be cut at various heights to produce varying numbers of clusters. I cut the dendrogram in such a way that it produced numbers of clusters varying from 100 to 1,000, and repeated my main analysis with these sets of clusters, as opposed to my primary sets of clusters.

Figure 2.6 shows how the bias of different designs, exposure models, and treatment effect estimators changed as a function of the number of clusters. I find that bias is highest, and in some cases higher than bias under the individual-level, Bernoulli randomized experiment, when the number of clusters is smallest (100). For numbers of clusters greater than 100, there is no discernible relationship between the number of clusters produced by the algorithm and the bias of the TATE estimates my simulated experiments produced. Furthermore, for all tested numbers of clusters above 100, the bias of the TATE estimates for all combinations of design, exposure model, and treatment effect estimators is generally lower than the naive experiment. Figure 2.7 shows the RMSE of my TATE estimates for each combination of design, exposure

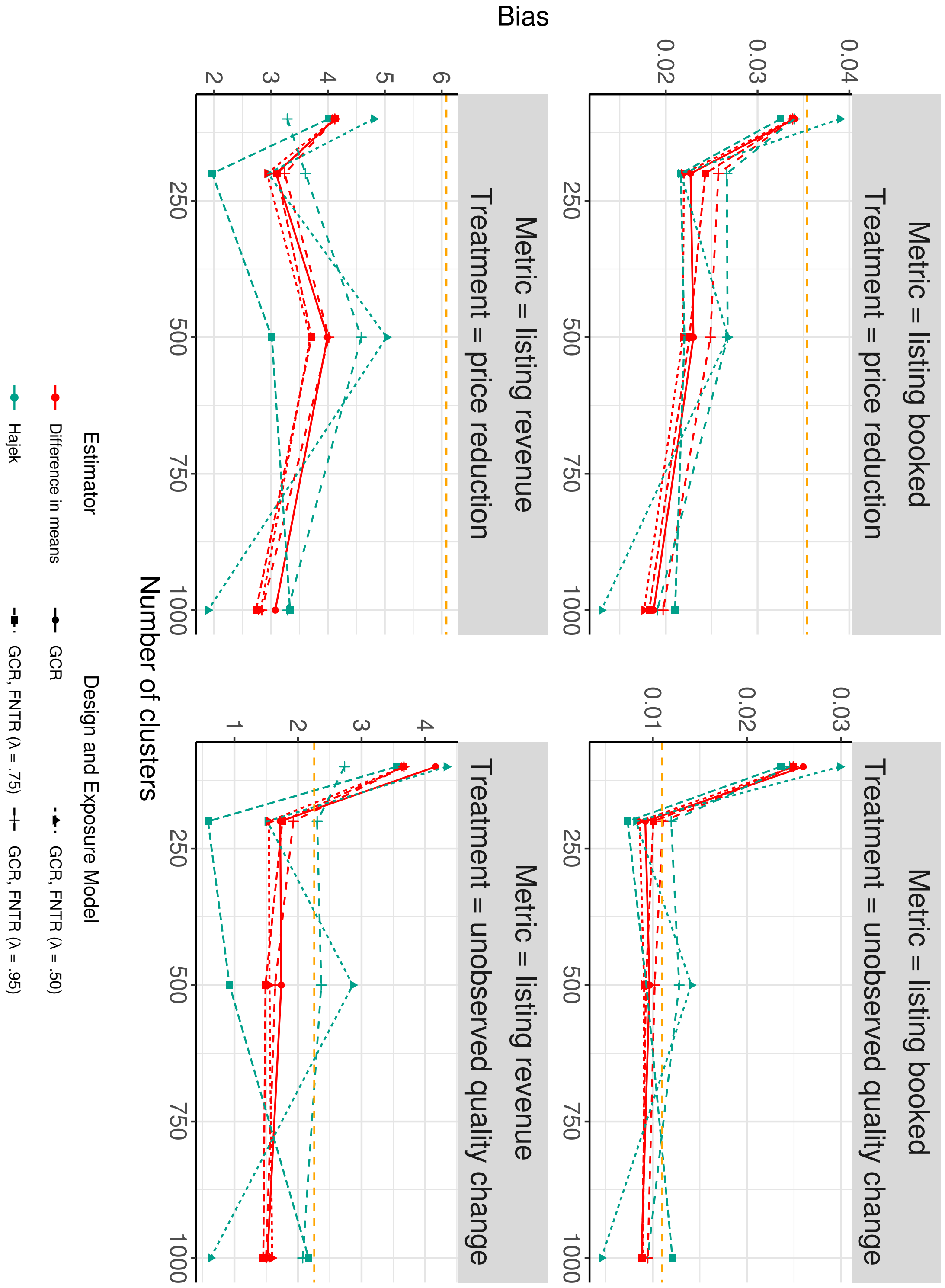


Figure 2.6: The bias of different combinations of experiment design, exposure model, and treatment effect estimator, for both outcomes and both treatments, as a function of the number of clusters specified during the hierarchical graph clustering procedure (Clauset et al. 2004). The dashed yellow line indicates the bias of the individual-level, Bernoulli randomized experiment analyzed with the difference in means estimator.

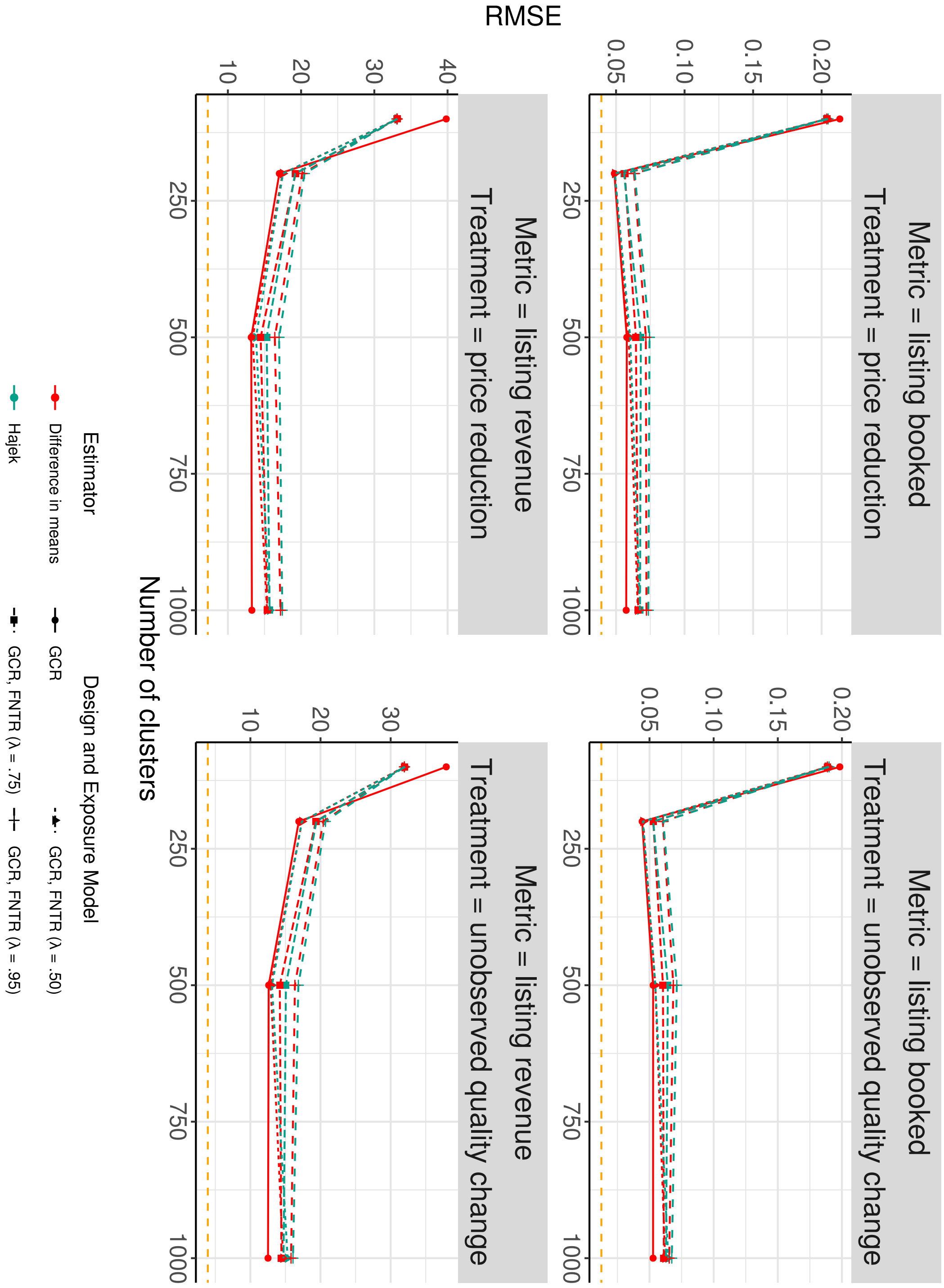


Figure 2.7: The RMSE of different combinations of experiment design, exposure model, and treatment effect estimator, for both outcomes and both treatments, as a function of the number of clusters specified during the hierarchical graph clustering procedure (Clauset et al. 2004). The dashed yellow line indicates the RMSE of the individual-level, Bernoulli randomized experiment analyzed with the difference in means estimator.

model, and treatment effect estimator as I vary the number of clusters. The results are similar to the results for the bias of the TATE estimates. RMSE is high when the number of clusters is 100; after that, RMSE is much lower, and there does not seem to be a relationship between the number of clusters and TATE estimate RMSE.

These results suggest that the relationship between the bias and RMSE of graph clustered experiment designs is mediated by factors beyond just the number of clusters produced by the clustering algorithm. This relationship is likely a function of other factors as well, such as the network structure of the product network, the particular algorithm used to cluster the item/seller network, and the distribution of cluster sizes conditional on a given number of clusters.

2.5.6 Performance with different levels of demand

It is also possible that the extent to which the proposed methods reduce bias (and/or increase RMSE) depends on the amount of demand observed in the market. Holtz et al. (2020) find weak evidence in a large-scale meta-experiment that test-control interference is more severe in markets that are demand-constrained, as opposed to supply constrained. This is consistent with findings by Johari et al. (2020), who develop an analytical framework to study bias in two-sided marketplace experiments and find that test-control interference is larger in demand-constrained markets. In order to test how my findings vary across low- and high-demand regimes, I change the $n_{searchers}$ parameter in my analysis from 1,000 to 500, and then from 500 to 250, and repeat my main analysis.

Figure 2.8 shows how the bias of TATE estimates produced for each combination of experiment design, exposure model, and treatment effect estimator changes as a function of the number of searchers in the market. Generally speaking, as the number of searchers in the market increases, the relative magnitude of the TATE estimate bias decreases across all experiment designs, exposure models, and treatment effect estimators. However, blocked GCR, the FNTR exposure model, and the Hajek estimator continue to reduce bias relative to the individual-level, Bernoulli randomized experiment analyzed with a difference in means treatment effect estimator. Figure 2.9

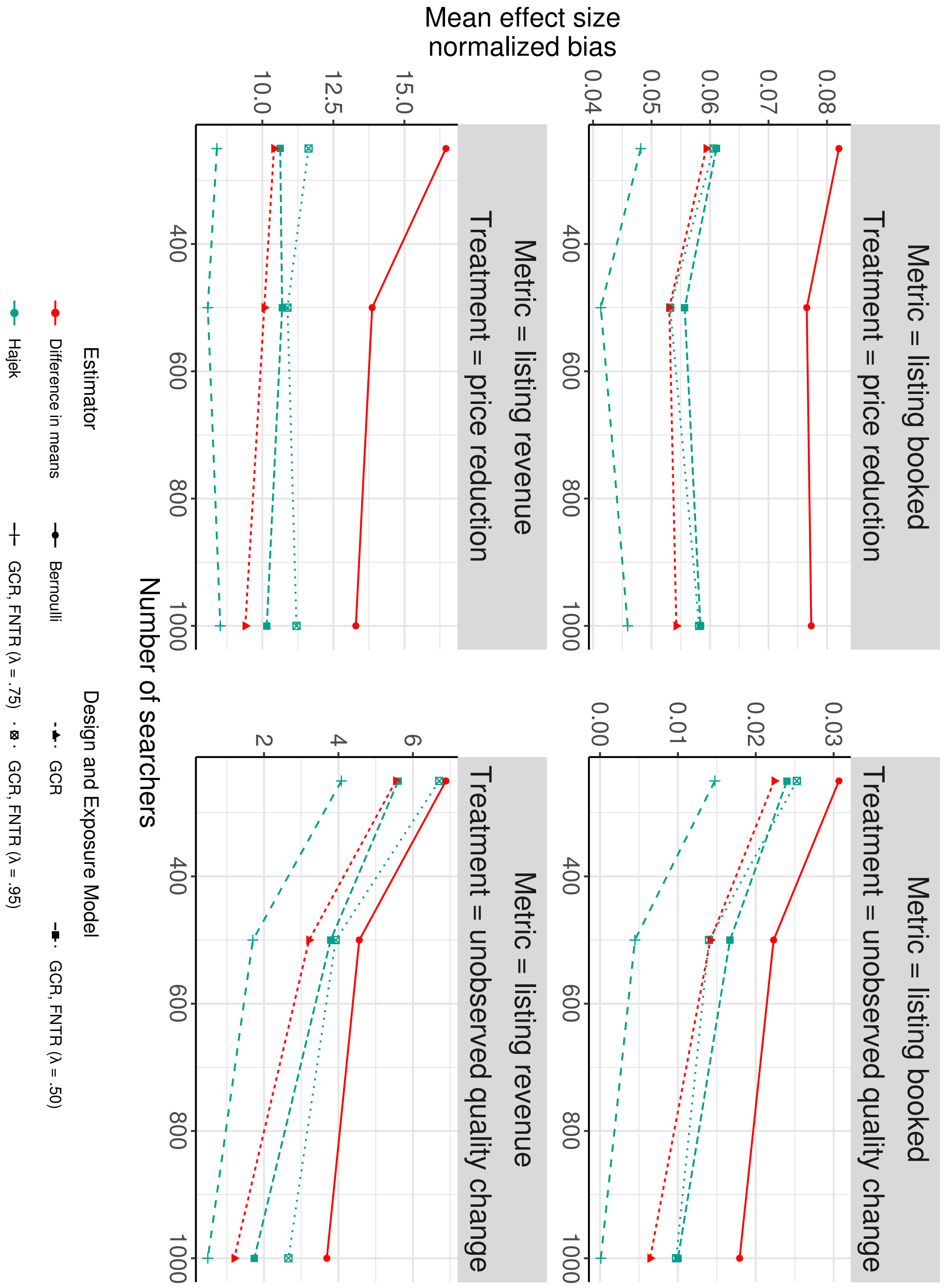


Figure 2.8: The bias of different combinations of experiment design, exposure model, and treatment effect estimator, for both outcomes and both treatments, as a function of the amount of demand, i.e., the number of searchers visiting the marketplace.

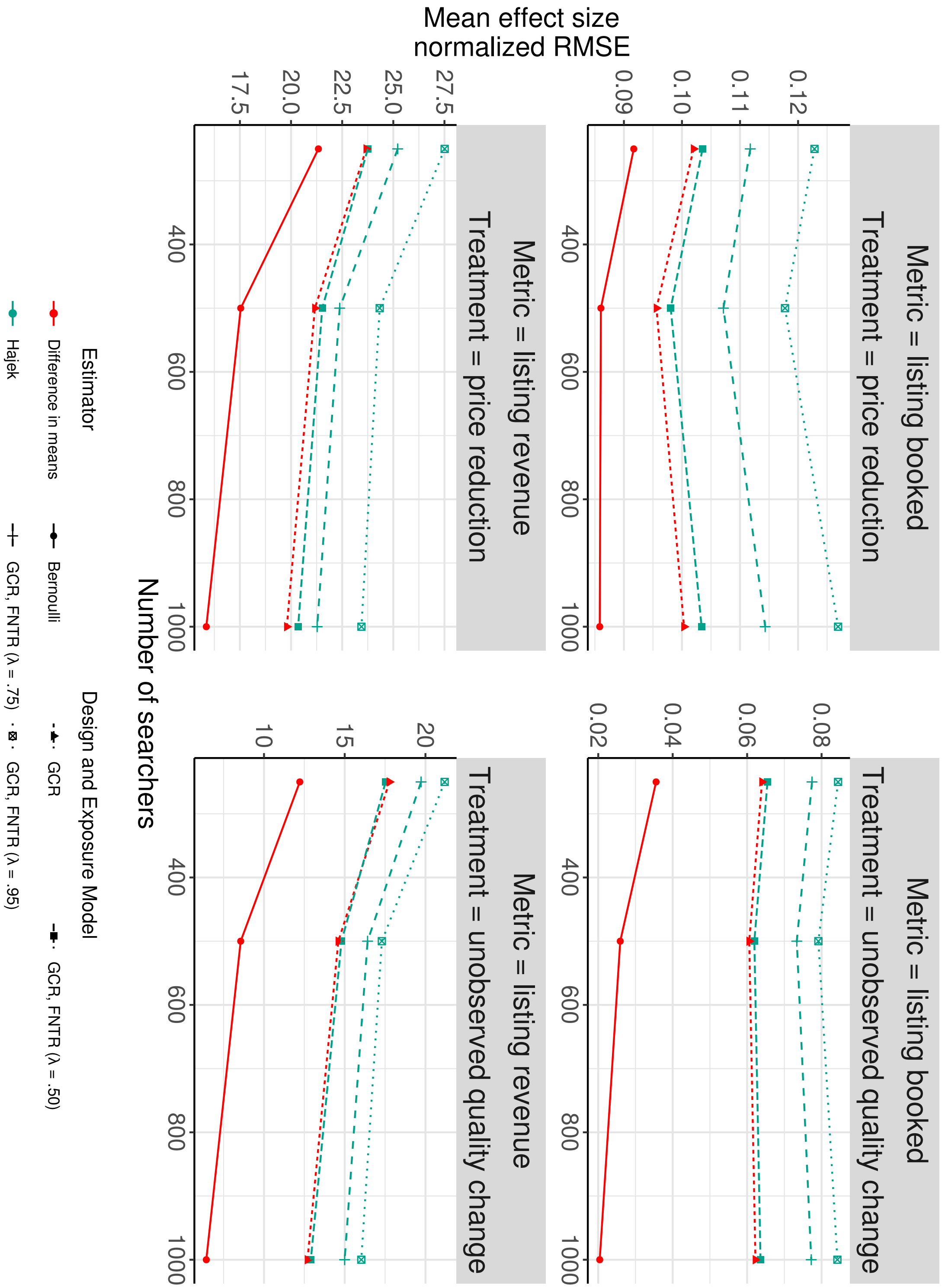


Figure 2.9: The RMSE of different combinations of experiment design, exposure model, and treatment effect estimator, for both outcomes and both treatments, as a function of the amount of demand, i.e., the number of searchers visiting the marketplace.

shows how the RMSE under different designs, exposure models, and treatment effect estimators varies as a function of the amount of demand. I find that RMSE also tends to decrease as the number of searchers increase, however, each combination of design, exposure model, and treatment effect estimator I discuss in this paper consistently exhibits higher RMSE than the baseline of a individual-level, Bernoulli randomized experiment analyzed with a difference in means estimator.

2.5.7 Performance in a different simulated market

Table 2.4: Summary of Airbnb listing covariates (Washington, D.C.)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Private room	4,304	0.302	0.459	0	0	1	1
Shared room	4,304	0.029	0.167	0	0	0	1
Entire home/apt	4,304	0.670	0.470	0	0	1	1
Reviews	4,304	15.985	31.416	0	1	17	385
Overall satisfaction	3,213	4.718	0.428	1.000	4.500	5.000	5.000
Capacity	3,748	2.662	1.104	1.000	2.000	4.000	6.000
Beds	4,285	1.231	0.850	0.000	1.000	1.000	10.000
Price (USD)	4,304	139.668	121.384	10	80	150	2,000
Min Stay	4,150	2.968	8.744	1.000	1.000	3.000	180.000
Lat.	4,304	38.914	0.020	38.825	38.902	38.926	38.993
Lon.	4,304	−77.025	0.025	−77.111	−77.041	−77.007	−76.950

Finally, in order to test that my results hold in a different simulated market, with different network structure and a different distribution of product attributes, I repeated my analysis using another Airbnb dataset scraped by Slee (2015), which describes the set of Airbnb listings in Washington, D.C. as of February 21, 2016.[6] Table 2.4 provides information about the distribution of listing-level covariates across this alternate sample of Airbnb listings. After preprocessing the data, creating a network, and clustering that network using the same procedure described in Section 2.3, I used my simulation framework to measure the bias and RMSE of different combinations of experiment design, exposure model, and TATE estimator for this alternate market under both the price reduction treatment and the unobserved listing quality treatment. Figure 2.10 summarizes my results, which are found in full in Table 2.5 and Table 2.6. These results are qualitatively similar to those found in Figure 2.3. In other words, I do not find evidence that my results are sensitive to the particular

[6] One difference between the Miami dataset and the Washington, D.C. dataset is the D.C. dataset does not contain information on the number of bathrooms in each listing. When calculating search algorithm scores and searcher utilities, I set this term to 0.

network structure or distribution of product-level attributes found in my primary Airbnb dataset.

Table 2.5: Performance comparison: outcome = booked listings (Washington, D.C.)

Treatment	Design	Estimator	Bias	RMSE	Coverage
Price Reduction	Bernoulli	Difference in means	0.1027	0.1165	6%
Price Reduction	GCR	Difference in means	0.0807	0.1035	12%
Price Reduction	GCR	Regression + clustered S.E.	0.0807	0.1035	50%
Price Reduction	GCR, FNTR ($\lambda = .50$)	Difference in means	0.0806	0.1039	12%
Price Reduction	GCR, FNTR ($\lambda = .50$)	Hajek	0.0788	0.1023	100%
Price Reduction	GCR, FNTR ($\lambda = .75$)	Difference in means	0.0806	0.1106	13%
Price Reduction	GCR, FNTR ($\lambda = .75$)	Hajek	0.0752	0.1040	100%
Price Reduction	GCR, FNTR ($\lambda = .95$)	Difference in means	0.0874	0.1369	17%
Price Reduction	GCR, FNTR ($\lambda = .95$)	Hajek	0.0946	0.1375	100%
Unobserved quality	Bernoulli	Difference in means	0.0204	0.0242	63%
Unobserved quality	GCR	Difference in means	0.0097	0.0514	34%
Unobserved quality	GCR	Regression + clustered S.E.	0.0097	0.0514	94%
Unobserved quality	GCR, FNTR ($\lambda = .50$)	Difference in means	0.0101	0.0523	35%
Unobserved quality	GCR, FNTR ($\lambda = .50$)	Hajek	0.0093	0.0520	100%
Unobserved quality	GCR, FNTR ($\lambda = .75$)	Difference in means	0.0114	0.0645	34%
Unobserved quality	GCR, FNTR ($\lambda = .75$)	Hajek	0.0085	0.0600	100%
Unobserved quality	GCR, FNTR ($\lambda = .95$)	Difference in means	0.0123	0.0929	30%
Unobserved quality	GCR, FNTR ($\lambda = .95$)	Hajek	0.0217	0.0909	100%

Table 2.6: Performance comparison: outcome = listing revenue (Washington, D.C.)

Treatment	Design	Estimator	Bias	RMSE	Coverage
Price Reduction	Bernoulli	Difference in means	12.21	13.52	11%
Price Reduction	GCR	Difference in means	10.65	16.49	16%
Price Reduction	GCR	Regression + clustered S.E.	10.65	16.49	52%
Price Reduction	GCR, FNTR ($\lambda = .50$)	Difference in means	10.62	16.67	17%
Price Reduction	GCR, FNTR ($\lambda = .50$)	Hajek	10.38	16.34	100%
Price Reduction	GCR, FNTR ($\lambda = .75$)	Difference in means	10.48	18.72	18%
Price Reduction	GCR, FNTR ($\lambda = .75$)	Hajek	9.75	16.96	100%
Price Reduction	GCR, FNTR ($\lambda = .95$)	Difference in means	11.04	23.74	17%
Price Reduction	GCR, FNTR ($\lambda = .95$)	Hajek	12.52	22.97	100%
Unobserved quality	Bernoulli	Difference in means	2.47	3.34	74%
Unobserved quality	GCR	Difference in means	1.67	11.36	29%
Unobserved quality	GCR	Regression + clustered S.E.	1.67	11.36	82%
Unobserved quality	GCR, FNTR ($\lambda = .50$)	Difference in means	1.74	11.65	29%
Unobserved quality	GCR, FNTR ($\lambda = .50$)	Hajek	1.65	11.48	100%
Unobserved quality	GCR, FNTR ($\lambda = .75$)	Difference in means	1.94	14.40	26%
Unobserved quality	GCR, FNTR ($\lambda = .75$)	Hajek	1.62	12.78	100%
Unobserved quality	GCR, FNTR ($\lambda = .95$)	Difference in means	2.16	20.08	18%
Unobserved quality	GCR, FNTR ($\lambda = .95$)	Hajek	3.98	18.66	100%

2.6 Discussion

The fact that the experiment designs, exposure models, and treatment effect estimators I study decrease bias, but also cause significant increases in TATE estimates suggests an interesting trade-off for online marketplace researchers and practitioners. In some cases, firms may prefer a biased estimate with much lower RMSE. Although unbiased experiments are desirable, the corresponding loss of statistical power can lead to minimum detectable effect sizes that are much higher than acceptable, given that large online marketplace firms often conduct A/B tests with the hopes of detecting treatment effects that are on the order of fractions of a percent. I believe the

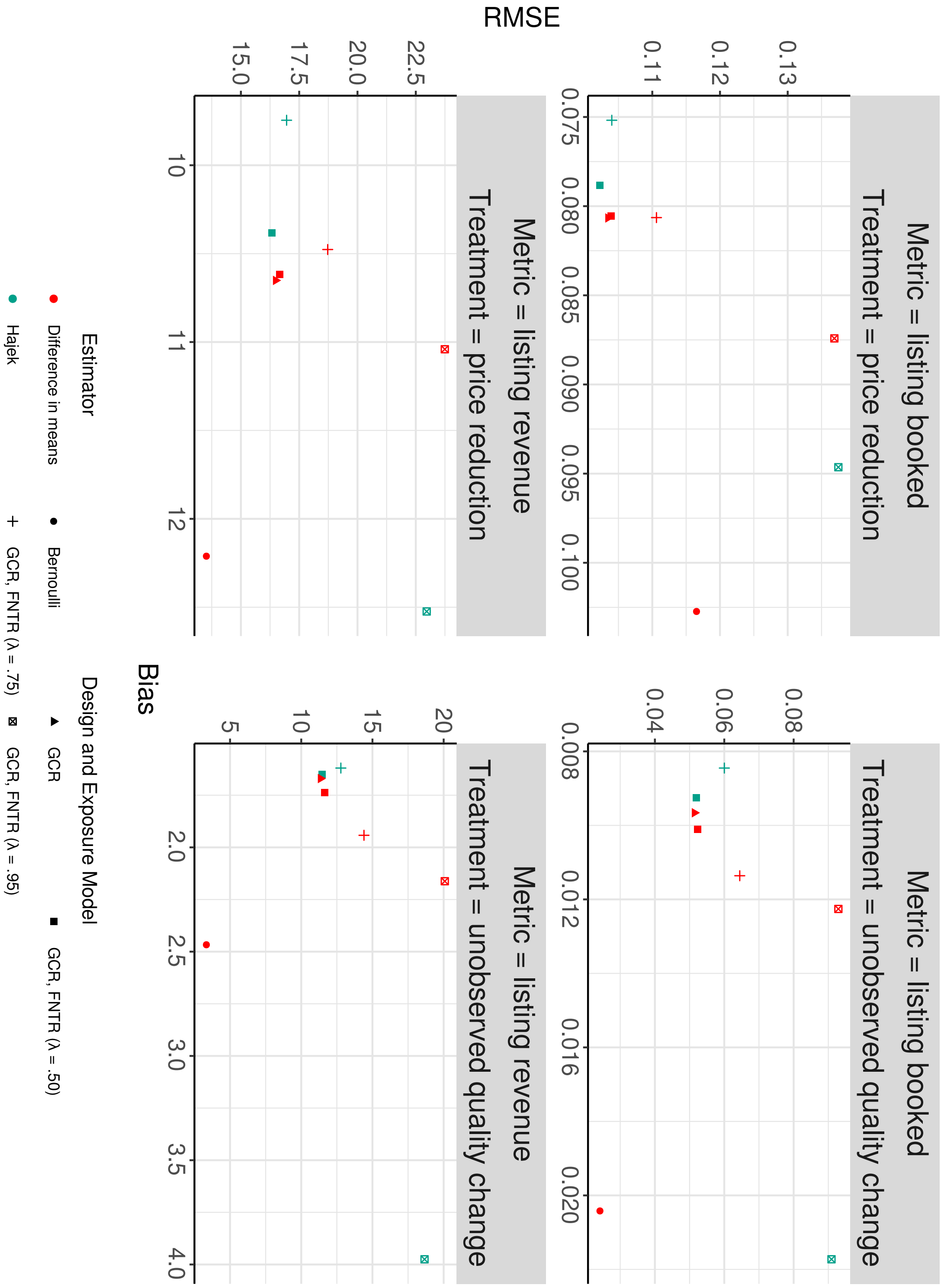


Figure 2.10: The bias and RMSE of different combinations of experiment design, exposure model, and treatment effect estimator, for both outcomes and both treatments using data from Washington, D.C.

methods I explore are best suited to contexts in which there is ample statistical power (i.e., large sample sizes and/or the ability to increase precision through methods such as regression adjustment), there is a strong belief that test-control interference will be severe for the treatment intervention of interest, or there is significant concern that test-control interference could lead to a sign flip in the TATE estimate.

In other cases, alternative experiment designs for online marketplaces may be preferable. Recent research has explored the viability of “switchback” designs (Sneider et al. 2019, Bojinov et al. 2020), in which the units of randomization are chunks of time, as opposed to shoppers or sellers. Switchback experiments do not suffer from excessive variance, however they may introduce *new* sources of bias if a given treatment intervention has the potential to create intertemporal spillovers. For instance, in the context of a real-time two-sided market such as Uber, riders may choose to wait some duration of time if they open the app and find that wait times or prices are too high. Another design that has recently been proposed is “two-sided randomization” (TSR) (Johari et al. 2020), in which randomization is delivered at the level of the shopper-seller or shopper-item pair. Johari et al. (2020) report that TSR is able to reduce bias in TATE estimates with only modest increases in RMSE. However, there are many treatment interventions for which TSR is not feasible, due to the requirement that randomization be conducted at the shopper-seller/shopper-item level. For instance, in the Airbnb context, a new calendar management or pricing tool must be introduced at the host level, not at the level of the searcher-host pair.

There are also markets and treatment interventions for which structural modeling techniques, such as the BLP framework (Berry et al. 1995, Nevo 2000), can be used to estimate the TATE. In this paper, I focus on the assessment of experiment design and analysis techniques because experimentation is the dominant tool in industry to assess potential product changes, and represents the “gold standard” for causal inference in academic research. Many online marketplace firms may not have the resources required to hire PhD economists who can build structural models, and randomized experiments may still be preferable in cases where specification error is a large concern. There are also many treatment interventions that may induce changes to behaviors or attributes that are difficult to incorporate into the BLP framework, such as changes to seller responsiveness or friendliness, changes to the text of user reviews, or new ways of presenting information to shoppers. Finally, the standard BLP framework assumes there are no constraints on product availability (Conlon and Mortimer 2013), which is not true in many large online marketplaces, including Airbnb, where each listing-night is a totally unique item that can only be purchased once. Although recent research (Farronato and Fradkin 2018) has applied demand estimation techniques in the context of Airbnb, in order to do so it was necessary to aggregate heterogeneous listings into groups that are assumed to be perfectly substitutable (e.g., "entire apartments in NYC"). In other words, this approach assumes away much of the heterogeneity that is present in online markets such as Airbnb, eBay, and Etsy.

Although I believe my simulations provide useful insight into the extent to which blocked graph cluster randomization, exposure modeling, and inverse probability-weighted estimators can affect the bias and RMSE of TATE estimates when adapted to the context of online marketplace experiments, my framework has a number of limitations. For instance, demand parameters estimated on actual Airbnb data could be used to tune my simulation. The simulation could also be extended to simulate more than one calendar night, to support social contagion among sellers who are able to observe each others' behavior, and to model treatment interventions that have seller non-compliance. Furthermore, while I believe my framework is helpful for developing intuition, it does not provide insight into what approach should be used in a given research context to build the product network that is used to design and analyze a given experiment. It is worth noting that my findings in Section 2.5.4 suggest that network mis-specification does not significantly impact the performance of the methods I evaluate. That being said, meta-experiments such as those presented by Pouget-Abadie et al. (2017), Saveski et al. (2017), and Holtz et al. (2020) may provide a solution to this problem. Under such meta-experiments, two field experiments are conducted simultaneously on different segments of the market with different designs, but the same treatment intervention. Subsequently, the treatment effect estimates obtained under the two experiment designs are compared in a statistically rigorous way. One particularly intriguing approach to constructing product networks is to use historical browsing and purchase data. Evaluation of this approach is beyond the scope of this paper, because I do not have access to actual user data. However, Holtz et al. (2020) find that using this type of data to conduct cluster randomized experiments can be extremely effective at reducing bias from test-control interference in TATE estimates.

2.7 Conclusion

Given the ubiquity of large online marketplace firms, it is crucial for both academics and practitioners that methods are developed to obtain accurate causal estimates through experimentation in online marketplaces. In this work, I have proposed adapting experiment designs and analysis techniques from the networks literature to the context of online marketplaces. I then assessed the performance of those methods through a simulation framework. my methods show that block randomized graph cluster randomization in particular can be effective at reducing bias in total average treatment effect estimates, but this comes at the cost of excessive variance and much higher RMSE. While there are some contexts in which experiment designers may find this trade-off worthwhile, there are many other contexts in which other experiment designs or structural modeling may be preferred. The fact that graph cluster randomization increases the variance of treatment effect estimators is not specific to the context of online marketplaces; this is also true in the context of network experiments, and has likely impeded broader adoption of graph cluster randomized network experiments. However, recent work in the network experimentation literature (Ugander and Yin 2020) suggests that it may be possible to obtain more precise treatment effect estimates under graph cluster randomized designs, providing a potential path forward for graph cluster randomization in both the network and online marketplace contexts.

Appendix

2.A Modified Simulation Framework

In order to test the robustness of my results to different underlying data generating processes, I repeat my analysis using a modified version of the simulation framework described in Section 2.4. Under the modified simulation framework, each searcher draws a preference vector, $\boldsymbol{\gamma}_i$, with each element of $\boldsymbol{\gamma}_i$ being generated by first drawing from a uniform distribution over the interval [0,1] and then normalizing so that the sum of the elements of $\boldsymbol{\gamma}_i$ is one, i.e.,

$$\begin{aligned} &\boldsymbol{\gamma}_{i0} \sim U[0,1] \text{ for } k = 1, 2, 3, ..., 9, \\ &\boldsymbol{\gamma}_i = \frac{\gamma_{i0}}{\sum_j \gamma_{i0}}. \end{aligned} \tag{2.1}$$

Each searcher also draws a "search quality cutoff," η_i, from the uniform distribution over the interval [-2, 2], i.e.,

$$\eta_i \sim U[-2, 2], \tag{2.2}$$

and calculates a "searcher score" for all listings presented by the search algorithm by taking the inner product of $\boldsymbol{\gamma}_i$ with the product attribute vector $\boldsymbol{\beta}_j$,

$$\text{Searcher Score}_{ij} = \gamma_i \cdot \mathbf{x}_j. \tag{2.3}$$

The searcher books the listing presented by the search algorithm that has the highest searcher score, so long as that searcher score is greater than or equal to η_i. If no listings meeting this requirement are presented, or if the search algorithm presents an empty set, the searcher does not book. In the modified simulation framework, there are no changes to the search algorithm itself.

Table 2.A.1: Performance comparison: outcome = booked listings (modified simulation)

Treatment	Design	Estimator	Bias	RMSE	Coverage
Price Reduction	Bernoulli	Difference in means	0.0256	0.0275	4%
Price Reduction	GCR	Difference in means	0.0189	0.0314	24%
Price Reduction	GCR	Regression + clustered S.E.	0.0189	0.0314	96%
Price Reduction	GCR, FNTR ($\lambda = .50$)	Difference in means	0.0183	0.0312	23%
Price Reduction	GCR, FNTR ($\lambda = .50$)	Hajek	0.0200	0.0323	100%
Price Reduction	GCR, FNTR ($\lambda = .75$)	Difference in means	0.0192	0.0344	23%
Price Reduction	GCR, FNTR ($\lambda = .75$)	Hajek	0.0164	0.0347	100%
Price Reduction	GCR, FNTR ($\lambda = .95$)	Difference in means	0.0189	0.0345	24%
Price Reduction	GCR, FNTR ($\lambda = .95$)	Hajek	0.0205	0.0388	100%

Table 2.A.2: Performance comparison: outcome = listing revenue (modified simulation)

Treatment	Design	Estimator	Bias	RMSE	Coverage
Price Reduction	Bernoulli	Difference in means	5.00	6.06	49%
Price Reduction	GCR	Difference in means	3.40	7.84	48%
Price Reduction	GCR	Regression + clustered S.E.	3.40	7.84	98%
Price Reduction	GCR, FNTR ($\lambda = .50$)	Difference in means	3.28	7.84	45%
Price Reduction	GCR, FNTR ($\lambda = .50$)	Hajek	3.40	7.99	100%
Price Reduction	GCR, FNTR ($\lambda = .75$)	Difference in means	3.41	8.00	48%
Price Reduction	GCR, FNTR ($\lambda = .75$)	Hajek	3.27	8.04	100%
Price Reduction	GCR, FNTR ($\lambda = .95$)	Difference in means	3.21	7.63	50%
Price Reduction	GCR, FNTR ($\lambda = .95$)	Hajek	4.16	8.71	100%

I repeated my main analysis using this modified simulation framework, as opposed to the one described in Section 2.4. Since there is no unobserved quality component in this simulation framework, I am only able to measure the bias and RMSE of TATE estimates obtained for the price reduction treatment intervention. Figure 2.A.1 depicts the bias and RMSE of different combinations of experiment design, exposure model, and treatment effect estimator under this modified simulation framework. my results under this modified framework are also shown in Tables 2.A.1 and 2.A.2. my results are qualitatively similar to my main results, suggesting that my findings are robust to different underlying data generating processes.

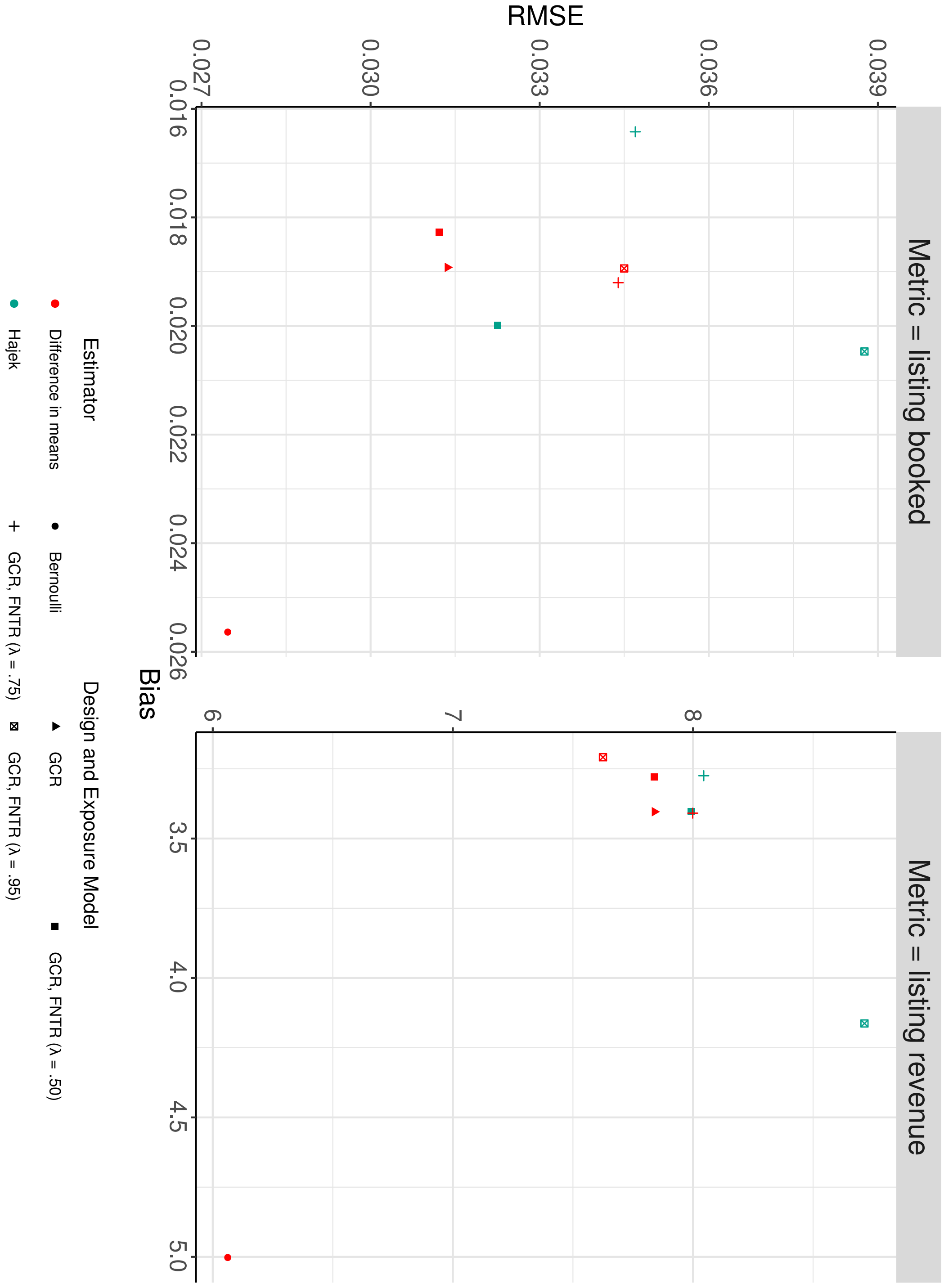


Figure 2.A.1: The bias and RMSE of different combinations of experiment design, exposure model, and treatment effect estimator, for both outcomes and both treatments under the modified simulation framework.

Chapter 3

Reducing Interference Bias in Online Marketplace Pricing Experiments[1]

3.1 Introduction

As of 2020, some of the world's most highly valued technology firms (e.g., Airbnb, Uber, Etsy) are online peer-to-peer marketplaces. These platforms create markets for many different types of goods, including accommodations, transportation, artisanal goods, and dog walking. Like almost all technology firms, online peer-to-peer marketplaces typically rely on experimentation, or A/B testing, to measure the impact of proposed changes to the platform and develop a deeper understanding of their customers. However, a randomized experiment's ability to provide an unbiased estimate of the total average treatment effect (TATE) relies on the stable unit treatment value assumption (SUTVA) (Rubin 1974), sometimes referred to as the "no interference" assumption (Cox 1958). Online marketplaces are inherently connected; sellers are likely to make strategic decisions based on the actions of their competitors, and mul-

[1] With Ruben Lobel (Airbnb) and Inessa Liskovich (Airbnb). We 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 also appreciate the helpful feedback we have received from Sinan Aral, Dean Eckles, Andrey Fradkin, Alex Moehring, Hong Yi Tu Ye, attendees of the WCBA 2019, CIST 2020, and the HBS Digital Doctoral Workshop, and anonymous reviewers at *Management Science*. This experiment was classified as exempt by the MIT Committee on the Use of Humans as Experimental Subjects under Protocol #1807452488.

tiple sellers may sell different items that complement or substitute for one another. As a result, SUTVA is unlikely to hold in online marketplace settings. Previous work (Blake and Coey 2014, Fradkin 2015, Holtz 2018) has shown that naive experimentation in online marketplaces can lead to TATE estimates that are overstated by up to 100%.

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

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

The first meta-experiment measures the effect of a change to Airbnb's platform fee structure. In the treatment group, long-tenured hosts were subject to a platform guest fee increase, while the platform guest fee for less tenured hosts remained unchanged. In the control group, long-tenured hosts were subject to a platform guest fee *decrease*, while the platform guest fee for less tenured hosts remained unchanged. Results from the Bernoulli randomized meta-treatment arm suggest that the treatment led to a statistically significant loss of 0.207 bookings per listing over the course of the experiment.[2] However, a joint analysis of the entire meta-experimental sample finds that there is a statistically significant difference between the TATE estimates obtained in the two meta-treatment arms. We estimate that 32.60% of the Bernoulli TATE estimate on bookings is attributable to interference bias. While not statistically significant, we also report results that suggest that interference bias is more severe in markets that are demand constrained than in markets that are supply constrained.

Results from the fee meta-experiment establish the existence of interference bias in online marketplaces, and the efficacy of cluster randomization in reducing that bias. However, the guest platform fee treatment intervention is one that affects all hosts on Airbnb. Often, online marketplace designers are interested in the effect of behavioral nudges, which only cause a change in the behavior of some users. These experiments are typically analyzed with intention-to-treat (ITT) analysis. To test for interference bias in an experiment that requires ITT analysis, we conduct a second meta-experiment that measures the effect of a proposed update to the algorithm underlying Airbnb's price suggestions for hosts. On average, the treatment *increased* the prices suggested to hosts. Results from the Bernoulli randomized meta-treatment arm suggest that the treatment led to a statistically significant loss of 0.106 bookings per

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

listing over the course of the experiment. In the cluster randomized meta-treatment arm, this treatment effect disappears; the point estimate is smaller in magnitude, and not statistically significant. However, a joint analysis of the entire meta-experimental sample fails to detect a statistically significant difference between the two sets of treatment effect estimates. Post-hoc power analysis reveals that the meta-experiment is underpowered to detect interference bias that is not extremely severe in magnitude. Although not statistically significant, our point estimates suggest that in the Bernoulli randomized pricing experiment, 54.16% of the observed treatment effect is due to interference bias. This result highlights the difficulty of detecting interference bias when a given treatment intervention only affects some users, even if the magnitude of that bias is potentially large.

While previous research has focused on quantifying the magnitude of interference bias through simulation (Fradkin 2015, Holtz 2018) or post-hoc analysis (Blake and Coey 2014), this work is among the first empirical papers to focus on reducing interference bias in a marketplace experiment through experiment design. The experiment design techniques we employ are strongly influenced by the network experimentation literature (Eckles et al. 2017, Ugander et al. 2013, Saveski et al. 2017), and future extensions of our work might focus on adopting analysis-based approaches to reducing interference bias in network experiments (Athey et al. 2018, Aronow et al. 2017, Eckles et al. 2017, Chin 2019) to an online marketplace setting. Future work might also focus on how to best cluster items or sellers in a marketplace. Clustering items or sellers in an online marketplace is difficult, as there is often no explicit network structure indicating which items are likely to substitute or complement for one another,[3] and measuring cross-price elasticities in markets with millions of heterogeneous goods is difficult.

The rest of this paper proceeds as follows. In Section 3.2, we review the related literature. In Section 3.3, we describe in greater detail the features of Airbnb's platform that are relevant to the two meta-experiments presented in this paper. Our

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

meta-experiment design is described in Section 3.4. We present results from the fee experiment in Section 3.5, and results from the pricing algorithm experiment in Section 3.6. Finally, we discuss our findings and future extensions in Section 3.7.

3.2 Related Literature

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

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

Bias in total average treatment estimates due to statistical interference is not a problem unique to online marketplace experiments. In fact, there has been substantial research on experiment design and analysis techniques that provide unbiased TATE estimators in settings where the stable unit treatment value assumption (Rubin 1974) is violated.[4] SUTVA assumes that the potential outcomes of a given unit of analysis are independent of the treatment assignments other units receive. However, in many settings (e.g., networks, marketplaces) SUTVA is unlikely to hold. When SUTVA is violated, the TATE estimated from a Bernoulli randomized experiment can differ substantially from the actual TATE (i.e., the average effect of the treatment under the counterfactual that every unit is treated). Network science researchers have developed

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

experiment designs (Ugander et al. 2013, Eckles et al. 2017) and treatment effect estimators (Aronow et al. 2017, Chin 2019) that eliminate or reduce bias due to SUTVA violations arising from network interference.

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

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

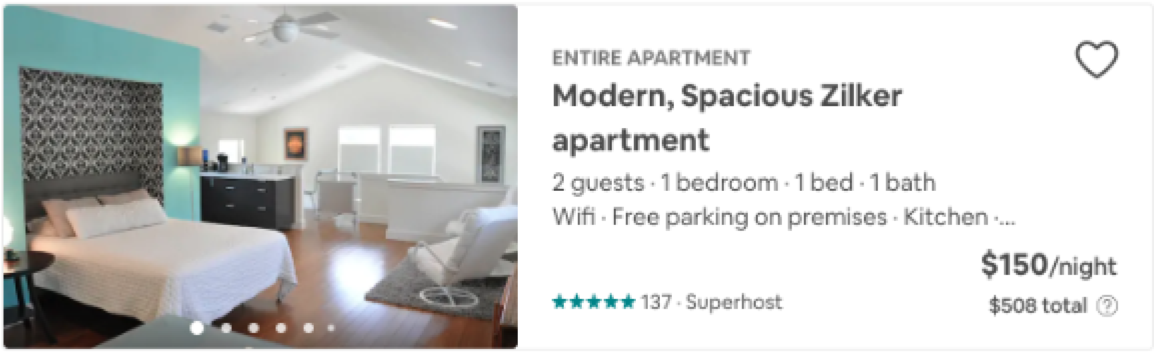
This paper builds on prior research by adapting experiment design techniques from the networks literature (Ugander et al. 2013, Eckles et al. 2017, Holtz 2018) and conducting meta-experiments (Saveski et al. 2017) in an online marketplace to test for the existence of interference bias. Developing methods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

3.3 Setting

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

3.3.1 Platform Guest Fees

Airbnb earns revenue by collecting fees from guests and hosts for every transaction that occurs on the platform. In order to set fees optimally, it is crucial for the platform to understand guest price elasticity. Airbnb's fees for guests are visible in three different locations throughout the booking process. First, guest platform fees are included in the total price shown to guests when a listing appears in search. Figure 3.1 shows a typical Airbnb search result. Second, if a guest opens a tooltip on any search result, they are shown a price breakdown that separates the listing's nightly price and the guest platform fee. Figure 3.2 shows this tooltip. Finally, when viewing a listing's product detail page, a detailed pricing breakdown (including fees) is displayed next to the "Request to Book" button. Figure 3.3 shows this price breakdown.



ENTIRE APARTMENT

Modern, Spacious Zilker apartment

2 guests · 1 bedroom · 1 bed · 1 bath

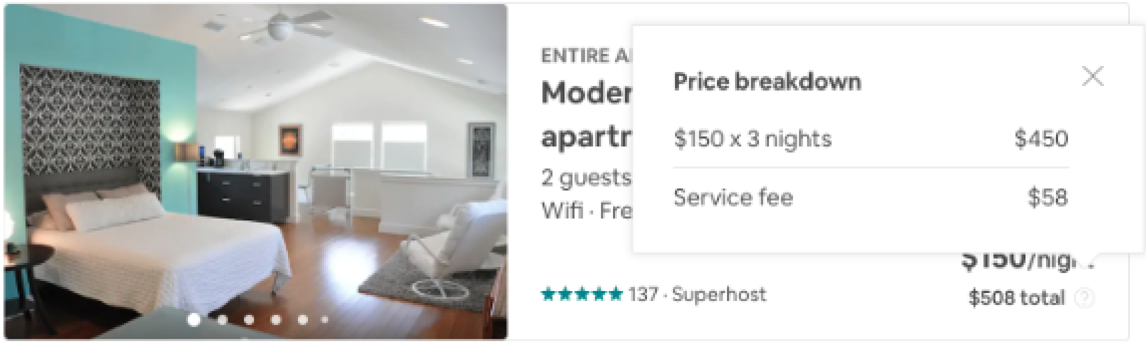
Wifi · Free parking on premises · Kitchen ·...

★★★★★ 137 · Superhost

$150/night

$508 total

Figure 3.1: A typical search result on Airbnb. For this search result, the guest platform fee is included in the total price of $508.



ENTIRE A

Moder

apartr

2 guests

Wifi · Fre

Price breakdown

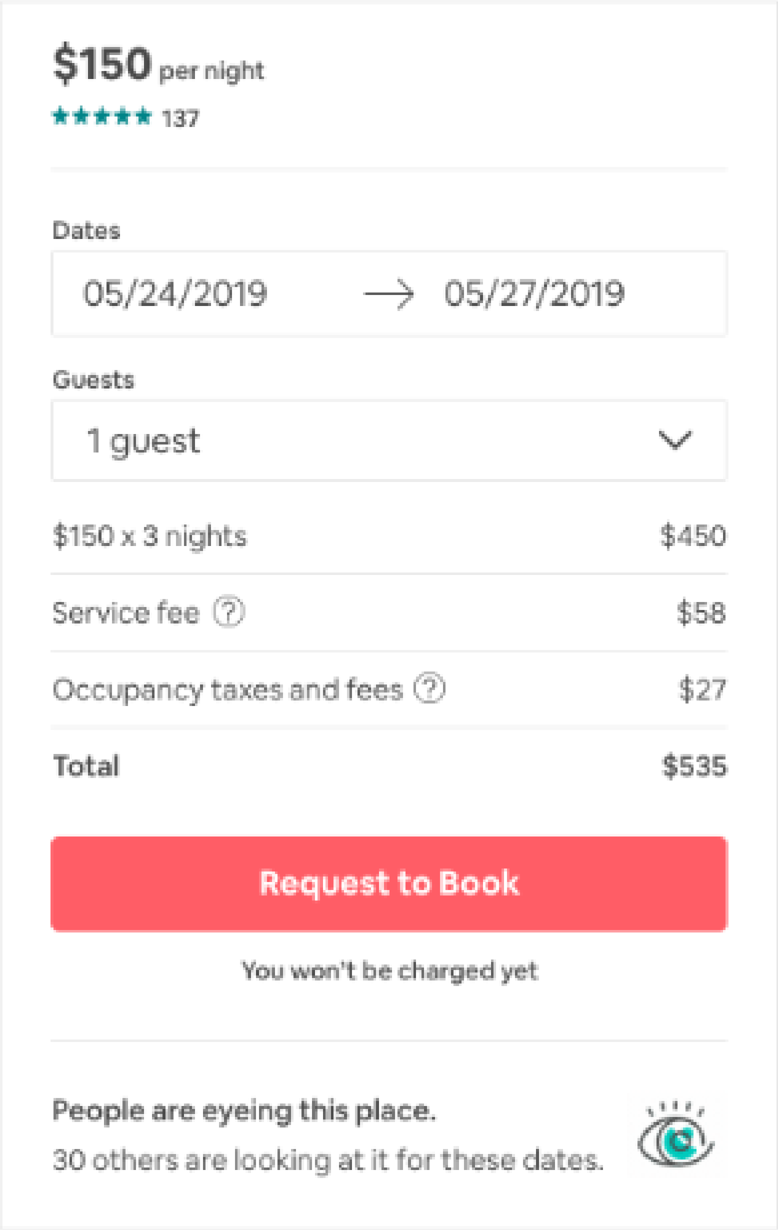
$150 x 3 nights	$450
Service fee	$58

$150/nig

★★★★★ 137 · Superhost

$508 total

Figure 3.2: The price breakdown tooltip for a typical search result on Airbnb. In this tooltip, the guest platform fee (listed here as a service fee of $58) is broken out from the nightly price.



$150 per night

★★★★★ 137

Dates

05/24/2019 → 05/27/2019

Guests

1 guest

$150 x 3 nights	$450
Service fee ?	$58
Occupancy taxes and fees ?	$27
Total	**$535**

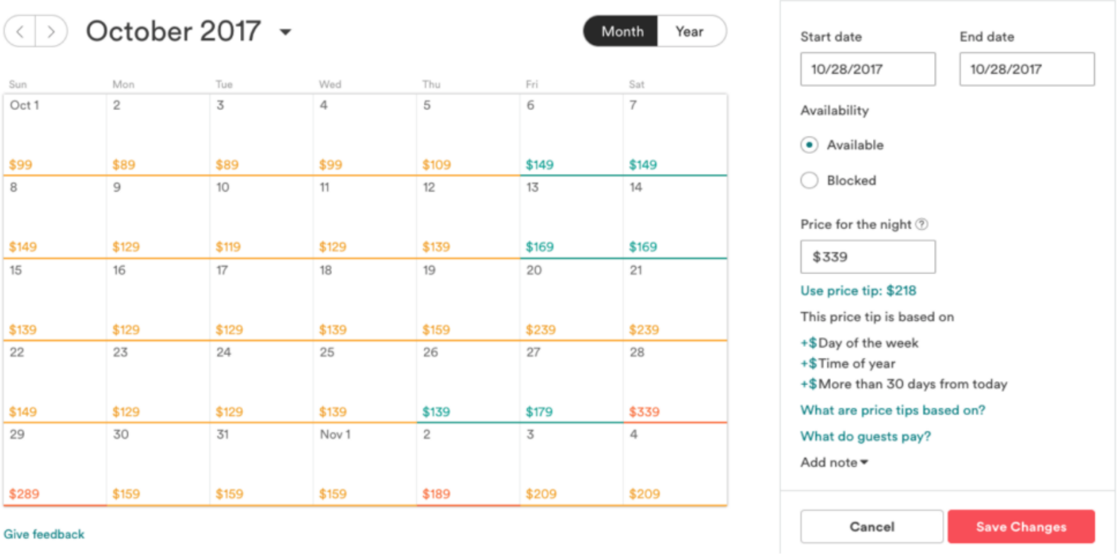
You won't be charged yet

People are eyeing this place.
30 others are looking at it for these dates.

Figure 3.3: 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.

3.3.2 Price Tips & Smart Pricing

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



Sun	Mon	Tue	Wed	Thu	Fri	Sat
Oct 1 $99	2 $89	3 $89	4 $99	5 $109	6 $149	7 $149
8 $149	9 $129	10 $119	11 $129	12 $139	13 $169	14 $169
15 $139	16 $129	17 $129	18 $139	19 $159	20 $239	21 $239
22 $149	23 $129	24 $129	25 $139	26 $139	27 $179	28 $339
29 $289	30 $159	31 $159	Nov 1 $159	2 $189	3 $209	4 $209

Start date: 10/28/2017
End date: 10/28/2017

Availability

● Available

○ Blocked

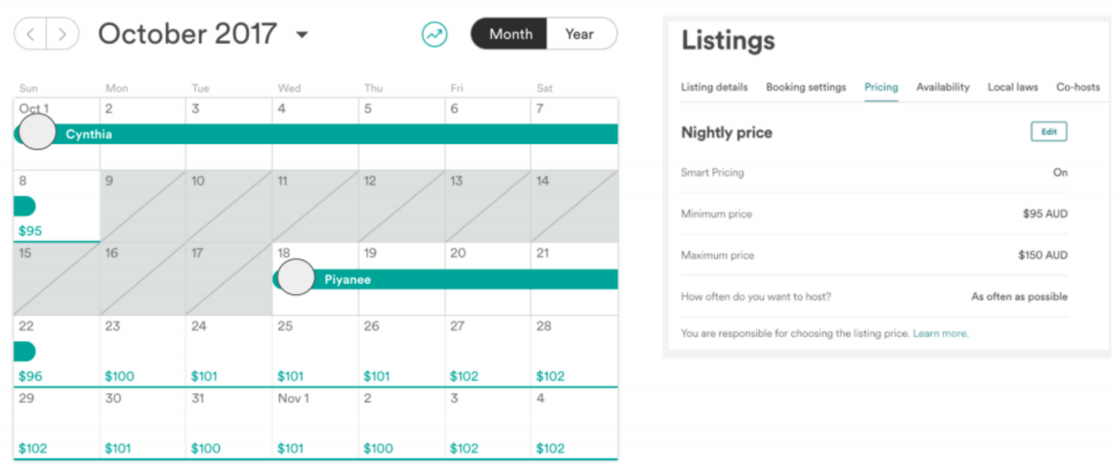
Price for the night: $339

Use price tip: $218

This price tip is based on

+$Day of the week
+$Time of year
+$More than 30 days from today

Figure 3.4: A screenshot of the "Price tips" UI. "Price tips" color codes the nights on a host's calendar based on the pricing model's estimated probability that a given night will be booked. If a host selects a given calendar night, the host is shown the model's suggested price. Airbnb also presents explanations for why it is recommending a particular price (e.g., "Time of year," "More than 30 days from today"). In order for a given host to fully adopt Airbnb's recommended prices with the "Price tips" product, the host is required to visit Airbnb every day, review Airbnb's recommendations, and accept them. Image from Ye et al. (2018).



Sun	Mon	Tue	Wed	Thu	Fri	Sat
Oct 1 Cynthia	2	3	4	5	6	7
8 $95	9	10	11	12	13	14
15	16	17	18 Piyanee	19	20	21
22 $96	23 $100	24 $101	25 $101	26 $101	27 $102	28 $102
29 $102	30 $101	31 $100	Nov 1 $101	2 $100	3 $102	4 $102

Listings

Nightly price

Smart Pricing	On
Minimum price	$95 AUD
Maximum price	$150 AUD
How often do you want to host?	As often as possible

You are responsible for choosing the listing price. Learn more.

Figure 3.5: A screenshot of the “Smart pricing” UI. When setting up“Smart Pricing,” hosts provide a minimum and maximum price. After “Smart Pricing” is turned on, hosts automatically adopt Airbnb’s recommended price if it is between the host’s minimum and maximum price. If Airbnb’s recommendation is higher than the host’s upper threshold, the price is set to the upper threshold. If Airbnb’s recommendation is lower than the host’s lower threshold, the price is set to the lower threshold. A screenshot of the “Smart Pricing” UI is shown in Figure 3.5. Image from Ye et al. (2018).

"Price tips" color codes nights on a host's calendar based on the estimated probability that a given night will be booked given the current price, and suggests an "optimal" price for each night. Importantly, "Price tips" requires hosts to manually accept prices in order to comply with the algorithm's suggestions recommended through the "Price tips" product. A screenshot of the "Price tips" UI is shown in Figure 3.4. "Smart pricing" was introduced to make it easier for hosts to adopt Airbnb's pricing recommendations en masse. Once "Smart pricing" is turned on, hosts automatically adopt Airbnb's recommended price, subject to constraints provided by the host. A screenshot of the "Smart Pricing" UI is shown in Figure 3.5.

3.4 Experiment Motivation & Design

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

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

Adapting experiment design and analysis techniques from the network experimentation literature, as proposed by Holtz (2018), is one avenue for reducing interference bias in online marketplace pricing experiments. However, none of the techniques put forward by Holtz (2018) have been used yet to design or analyze an online marketplace experiment. As a first step toward empirically confirming the existence of interference bias in online marketplace experiment TATE estimates, and measuring the extent to which cluster randomization, an experiment design technique, can reduce that bias, we conduct pricing-related meta-experiments (Saveski et al. 2017) on Airbnb. Quantifying the magnitude of interference bias, as well as the extent to which cluster randomization can reduce that bias, is useful for two reasons. First, even if interference bias is a theoretical concern, it may not be a practical one; statistical bias in TATE estimates due to interference may be small. Second, even if interference bias is large, cluster randomization may not be an effective tool to reduce that bias. If this were the case, cluster randomization would not be a worthwhile undertaking for firms; cluster randomization results in reduced statistical power relative to Bernoulli randomization, and is also more logistically complicated for firms to implement (both because of the need to identify relevant clusters, and because most corporate A/B testing tools do not support cluster randomization).

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

3.4.1 Treatment Assignment Mechanism

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

Clusters of Airbnb Listings

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

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

$$< L_A, L_B, L_C > . \tag{3.1}$$

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

$$J = \max_{W,V} \sum_{s \in S} \frac{1}{|s|} \sum_{i=1}^{|s|} \sum_{-k \leq j \leq k,\, k \neq 0} \log p\left(L_{i+j} | L_i\right), \tag{3.2}$$

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

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

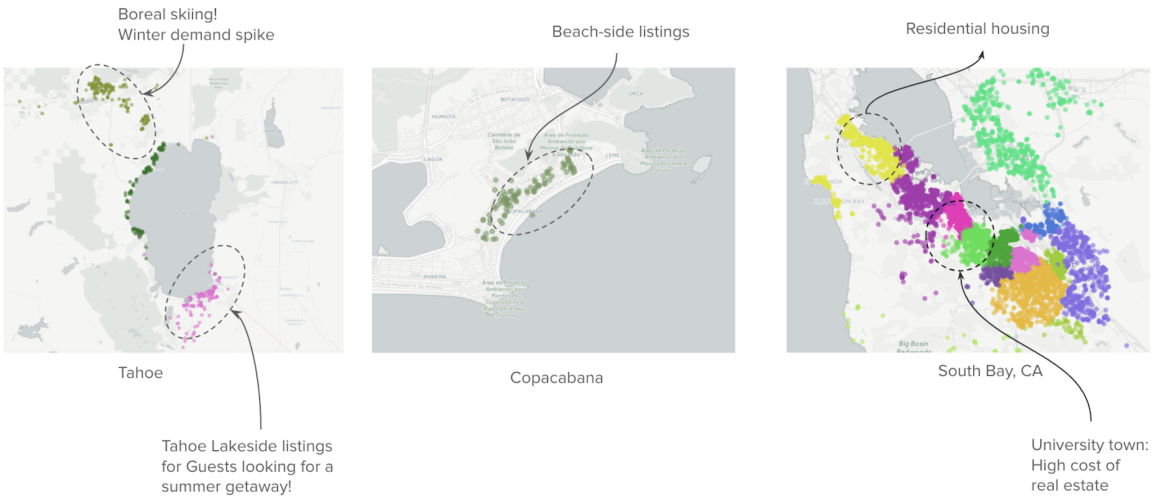


Figure 3.1: Example clusters generated using the hierarchical clustering scheme described in this paper. Image from Srinivasan (2018).

Once listing embeddings are estimated using the aforementioned approach, a recursive partitioning tree (Kang et al. 2016) is used to arrange the Airbnb listings into clusters. The algorithm starts from a single cluster containing all listings, and then recursively bisects clusters into two sub-clusters. The algorithm stops bisecting sub-clusters when the tree reaches a depth of 20, or when a new sub-cluster will contain less than 20 listings. Listings can then be assigned to clusters of arbitrary sizes by assigning them to the smallest sub-cluster to which they belong that has at least some threshold number of listings. For the algorithmic pricing meta-experiment, we set this threshold at 250 listings, whereas for the fee meta-experiment, we set this threshold at 1,000 listings.[5] Figure 3.1 depicts example clusters generated using this method in the Bay Area.

Pre-stratification & Treatment Assignment

Once listings have been assigned to clusters, those clusters are given meta-treatment assignments and, based on those cluster-level meta-treatment assignments, listings

[5] In choosing cluster sizes, we are attempting to balance two objectives: creating clusters that capture listings likely to interfere with one another, and designing an experiment with sufficient statistical power. Since ex ante, we expected the fee treatment intervention to have a larger effect, we chose larger clusters for that meta-experiment. For more details on the process used to determine cluster size, see Appendix 3.A.

are assigned listing-level treatment assignments.

To gain statistical power (particularly in the cluster-randomized meta-treatment arm), we group clusters into strata using a multivariate blocking procedure (Moore 2012). As a first step, we collected pre-treatment listing-level data.[6] We then aggregate data at the cluster level, and for each cluster calculate over the pre-treatment period the average number of nights booked per listing, the average number of bookings per listing, the average booking value per listing, and the number of experiment-eligible listings in the cluster.[7][8] After centering and scaling each of these metrics, we calculate the Mahalanobis distance between each pair of clusters. Finally, the smallest "available" distance between two clusters[9], and assigns the two corresponding clusters to the same stratum.

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

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

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

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

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

3.5 Fee Meta-experiment

3.5.1 Description

The fee meta-experiment ran from March 16, 2019 to March 21, 2019 on a population of 4,578,028 listings. Of those listings, 1,146,537 were assigned to the Bernoulli-randomized meta-treatment arm, and the remaining 3,431,491 were assigned to the cluster-randomized meta-treatment arm. Within the Bernoulli-randomized meta-treatment arm, 573,346 were assigned to the treatment and 573,191 listings were assigned to the control. Within the cluster-randomized meta-treatment arm, 2,982 clusters were assigned to the treatment and 2,982 clusters were assigned to the control, resulting in 1,720,147 listings assigned to the treatment and 1,711,344 listings assigned to the control. In total, across both meta-treatment arms, 2,293,493 listings were assigned to the treatment, and 2,284,535 were assigned to the control. Figure 3.1 shows the empirical CDFs for pre-treatment bookings, nights booked, and booking value across all four meta-treatment / treatment groups.[10] For each of these pre-treatment outcomes, the empirical CDFs are quite similar.

In the fee meta-experiment, listings in the treatment had their fees increased if they were long-tenured listings (i.e., if they had been on the platform as of a certain cutoff date). Listings in the control had their fees *decreased* if they were long-tenured listings. In both treatment arms, less tenured listings (i.e., those created after the cutoff date) did not have their fees changed.[11] Conceptually, one can think of the treatment and control conditions of this meta-experiment as comparing the effect of two different fee-based incentive programs Airbnb might run. In the treatment group, new listings have lower fees (which could drive business to newer listings), whereas in the control, older listings have lower fees (which could reward long-time Airbnb hosts and reduce churn). After the conclusion of the fee meta-experiment, a "reversal experiment" was run from April 15, 2019 to April 22, 2019. In the reversal experiment, listings that had been assigned the treatment condition in the meta-experiment were

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

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

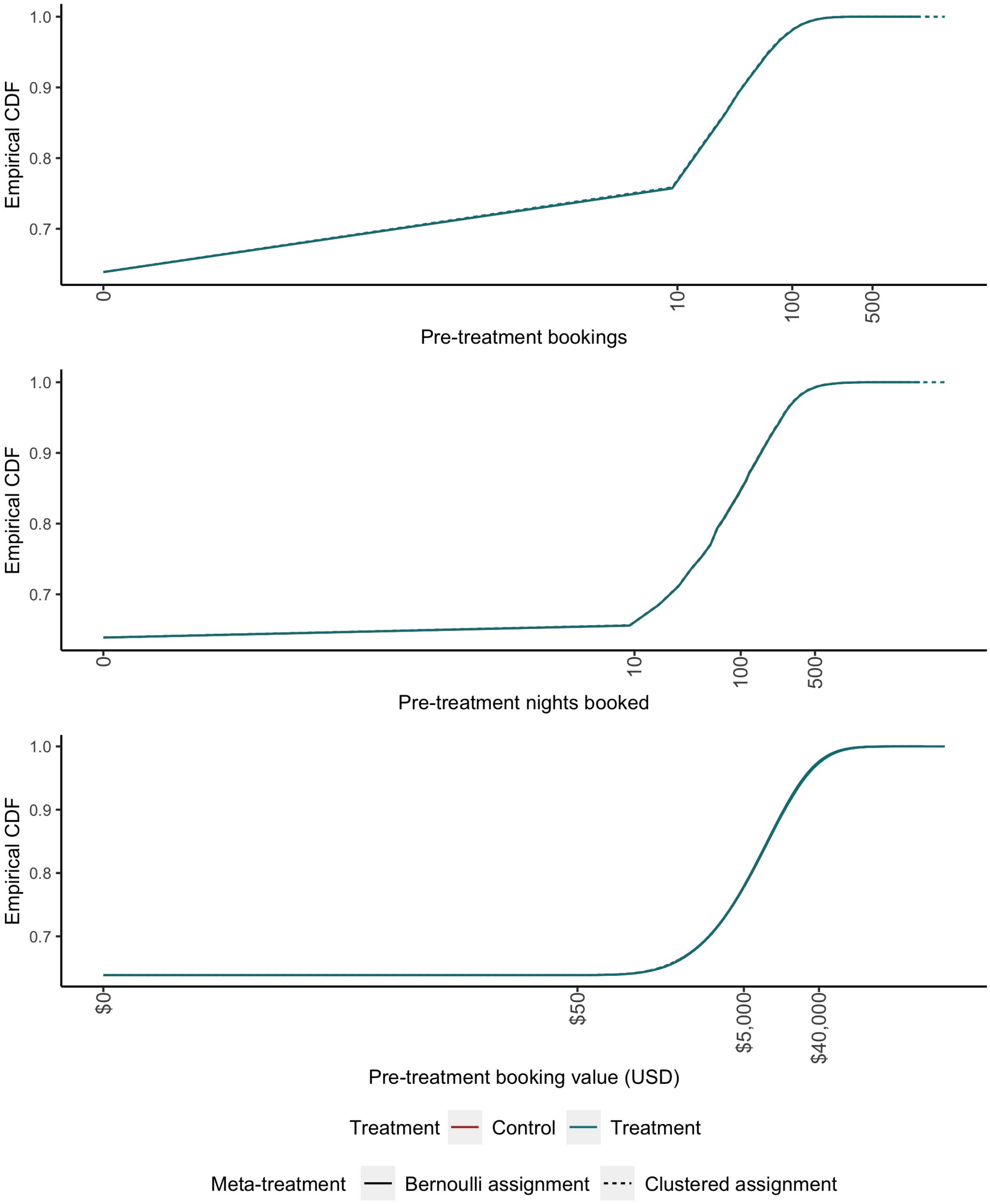


Figure 3.1: The empirical CDFs for pre-treatment bookings, nights booked, and booking value in each of the four treatment/meta-treatment groups for the fee meta-experiment.

assigned the control, and vice-versa. The purpose of the reversal experiment was to mitigate any negative impact of the meta-experiment on Airbnb hosts.

3.5.2 Results

In this section, we present results from the fee meta-experiment. We focus on a single outcome metric, bookings per listing, but the results for two alternative outcome metrics, nights booked per listing and gross guest spend per listing, are qualitatively similar and can be found in Appendix 3.B.[12] Since, relative to the control, the treatment *increased* fees, we expect the TATE on bookings per listing to be negative.

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

Bernoulli & Cluster Randomized Results

We analyze both the Bernoulli randomized and cluster randomized meta-treatment arms separately by estimating the following model,

$$Y_i = \alpha + \beta T_i + \sum_l \gamma_l \mathbb{1}(B_i = l) + \delta X_i + \epsilon_i \tag{3.3}$$

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

on listing-level data, where Y_i is the outcome of interest, T_i is the treatment assignment for listing i, B_i is a variable indicating which stratum listing i's cluster of belongs to, X_i is a vector consisting of listing i's pre-treatment bookings, nights booked, booking value, and gross guest spend, and ϵ_i is an error term.[13] For all analyses, we cluster standard errors at the Airbnb listing cluster-level.

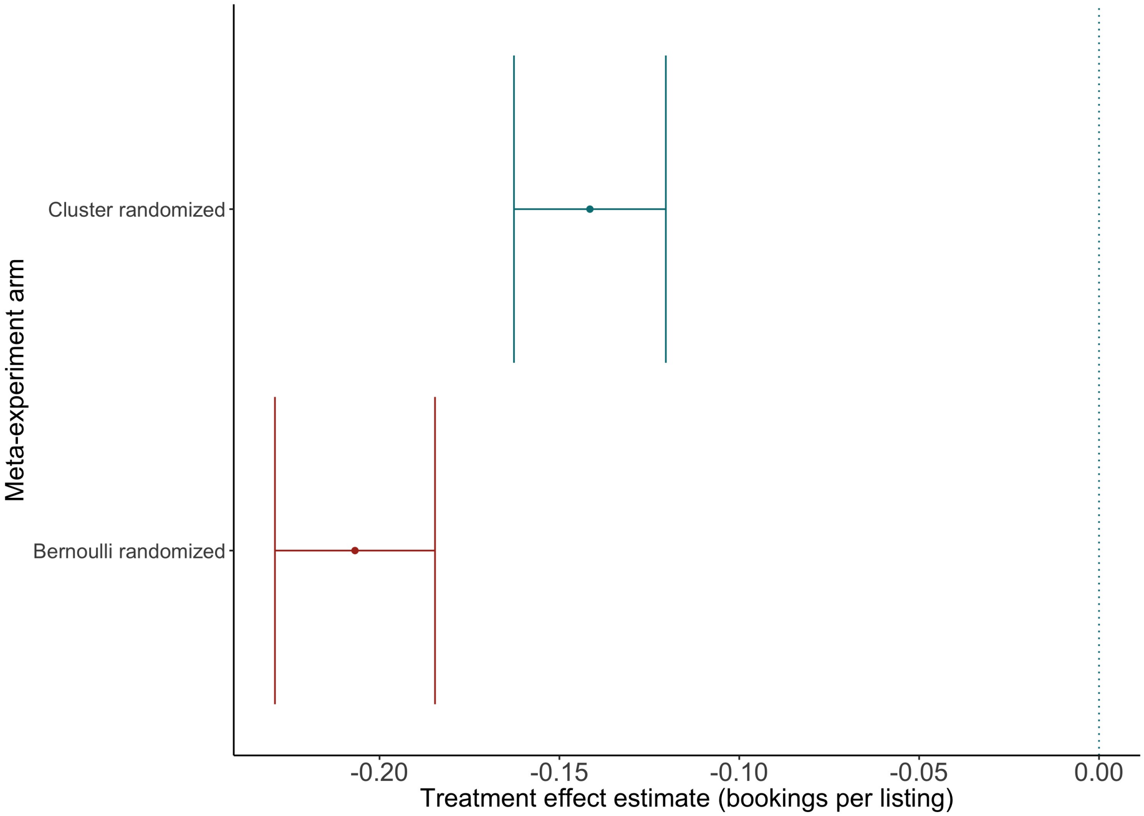


Figure 3.2: Total average treatment effect estimates for the fee experiment, estimated separately in the Bernoulli randomized meta-treatment arm and the cluster randomized meta treatment arm. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0 bookings per listing.

Table 3.1 shows the TATE estimate for bookings per listing in both the Bernoulli randomized and cluster randomized meta-treatment arms. In the Bernoulli randomized meta-treatment arm, the TATE is -0.207 bookings per listing, whereas in the cluster randomized meta-treatment arm, the TATE is -0.142 bookings per listing.

[13] Data from the cluster randomized meta-treatment arm can also be analyzed by first aggregating the data at the cluster level and then estimating a weighted version of Equation 3.3. We present this analysis in Appendix 3.C. This analysis results in estimates that are nearly identical to those obtained by analyzing the experiment with listing-level data.

Table 3.1: Independent results of the fee meta-experiment

	Dependent variable:	
	Bookings	
	Bernoulli randomized	Cluster randomized
	(1)	(2)
Treatment	-0.207^{***}	-0.142^{***}
	(0.011)	(0.011)
Pre-treatment bookings	0.173^{***}	0.174^{***}
	(0.001)	(0.001)
Pre-treatment nights booked	-0.003^{***}	-0.003^{***}
	(0.000)	(0.000)
Pre-treatment booking value	0.000	0.000^{***}
	(0.000)	(0.000)
Pre-treatment gross guest spend	-0.000^{**}	-0.000^{***}
	(0.000)	(0.000)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	No	Yes
R^2	0.407	0.405
Adjusted R^2	0.406	0.405
Note:		*p<0.1; **p<0.05; ***p<0.01

Both of these TATE estimates are statistically significant at the 95% confidence level. Figure 3.2 shows the estimated TATE in both meta-treatment arms, along with the corresponding 95% confidence intervals.

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

Joint Analysis

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

$$Y_i = \alpha + (\beta + \nu M_i)T_i + \xi M_i + \sum_l \gamma_l \mathbb{1}(B_i = l) + \delta X_i + \epsilon_i, \tag{3.4}$$

where Y_i is the outcome of interest, M_i is a binary variable set to 1 when listing i is in the Bernoulli meta-treatment arm and 0 when i is in the cluster-randomized meta-treatment arm, T_i is a binary variable set to 1 when listing i is exposed to the treatment, B_i is a variable indicating the stratum of clusters to which listing i belongs, X_i is a vector consisting of listing i's pre-treatment variables, and ϵ_i is the error term. Once again, we cluster standard errors at the Airbnb listing cluster-level.

In the above model, β measures the "true" effect of the treatment,[14] and ν measures the difference between the effect of the treatment in the Bernoulli arm and the effect of the treatment in the cluster randomized arm. In other words, ν should measure the extent to which cluster randomization reduces interference bias. ξ measures any baseline difference between listings in the Bernoulli-randomized arm of the

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

meta-experiment and listings in the cluster-randomized arm of the meta-experiment. Since clusters were assigned to meta-treatment arms using the random assignment procedure described in Section 3.4, we expect ξ to be zero. However, it is possible that imbalances between listings in the two meta-treatment arms persist even after random assignment.

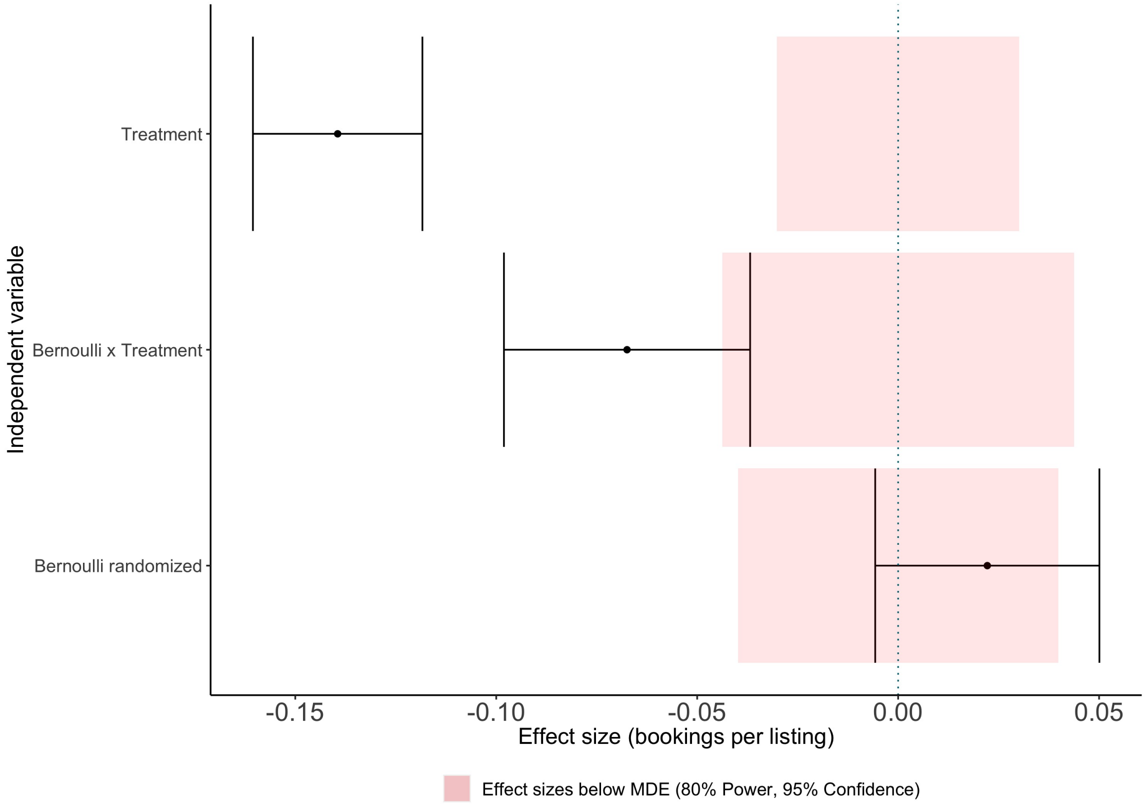


Figure 3.3: Coefficient estimates for the joint analysis of the fee meta-experiment. Error bars represent 95% confidence intervals. The dotted blue line correponds to a treatment effect of 0 bookings per listing. The red shaded area corresponds to values that are below the MDE (80% power, 95% confidence).

Table 3.2 shows the results obtained for the fee meta-experiment by estimating Equation 3.4 using listing level data.[15] Figure 3.3 displays our point estimate for each parameter in Equation 3.4, along with 95% confidence intervals. We estimate that

[15] Joint meta-experiment data can also be analyzed using a weighted combination of individual listing-level data from the Bernoulli randomized meta-treatment arm and aggregated cluster-level data from the cluster randomized meta-treatment arm. This analysis results in estimates that are nearly identical to those obtained using listing-level data from both meta-treatment arms. We present this analysis in Appendix 3.D.

Table 3.2: Results of the fees Meta-experiment

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

the "true" TATE is -0.139 bookings per listing, whereas -0.067 bookings per listing of the TATE measured in the Bernoulli randomized meta-treatment arm is due to interference bias. In other words, we estimate that 32.60% (±12.93%) of the TATE estimate achieved through a Bernoulli randomized experiment is due to interference bias, and is eliminated by instead running a cluster randomized experiment.

The Moderating Effect of Supply and Demand Constrainedness

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

In order to test this hypothesis, we re-estimate Equation 3.4 for subsets of Airbnb listings that are located in particularly supply constrained or demand constrained markets. Airbnb calculates a supply elasticity index and demand elasticity index for all markets that are above some threshold size using a Cobb-Douglas matching model a la Fradkin (2015). Of the markets for which these indices are calculated, we keep data for listings that are in markets larger than the median market (computed at the listing level). We then define a listing as being in a *supply constrained* market if its market's supply elasticity index is above the 75th quantile of supply elasticity indices (computed at the listing level), and define a listing as being in a *demand constrained* market if its market's demand elasticity index is above the 75th quantile of demand elasticity indices (computed at the listing level).

Column 1 of Table 3.3 shows our results for supply constrained listings, and Column 2 of Table 3.3 shows our results for demand constrained listings. Neither joint analysis is able to detect interference bias with statistical significance. However, if we take our non-statistically significant point estimates as given, our results do suggest that interference bias accounts for 15.09% of the Bernoulli TATE estimate in demand constrained markets, whereas interference bias actually *reduces* the magnitude of the Bernoulli TATE estimate by 27.41% in supply constrained markets. We interpret

Table 3.3: Results of the fee meta-experiment for supply- and demand-constrained listings

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

this as weak evidence that interference bias is more likely to lead to inflated TATE estimates in demand constrained markets than supply constrained markets, although further research should be conducted to better understand this relationship.

3.6 Algorithmic Pricing Experiment

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

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

When Airbnb tests a new iteration of its pricing algorithm, not all hosts are directly affected. Some hosts do not use Airbnb's pricing suggestions at all, and hosts who access Airbnb's pricing tips through "Price Tips" often have low compliance rates due to the manual effort required to follow Airbnb's suggestions. Even those hosts who opt into "Smart Pricing" may not fully comply with Airbnb's new suggestions, since Airbnb's suggestions are often constrained by business logic imposed by the host. Although Airbnb's pricing algorithm experiments do not directly affect all hosts, ITT analysis is required for two reasons. First, the set of hosts who *do* accept Airbnb's suggestions (and the extent to which they comply with those suggestions) is endogenous. Second, the firm is interested in the overall effect of the intervention, including the rate at which hosts accept a given set of suggestions.

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

3.6.1 Description

The algorithmic pricing meta-experiment ran from September 28, 2018 to October 31, 2018 on a population of 4,557,234 listings. Of those listings, 1,139,240 were assigned to the Bernoulli-randomized meta-treatment arm, and the remaining 3,417,994 were assigned to the cluster-randomized meta-treatment arm. Within the Bernoulli-randomized meta-treatment arm, 569,821 listings were assigned to the treatment and 569,419 listings were assigned to the control. Within the Cluster-randomized meta-treatment arm, 11,631 clusters were assigned to the treatment, and 11,631 clusters were assigned to the control, resulting in 1,709,018 listings assigned to the treatment, and 1,708,976 listings assigned to the control. In total, across both meta-treatment arms, 2,278,839 listings were assigned to the treatment, and 2,278,395 listings were assigned to the control. Importantly, the sample size for the algorithmic pricing meta-experiment is approximately equal to the sample size for the fee meta-experiment. Figure 3.1 shows the empirical CDFs for pre-treatment bookings, nights booked, and booking value across all four meta-treatment / treatment groups.[16] For each of these pre-treatment outcomes, the empirical CDFs are quite similar.

For listings in the treatment group, the suggested prices surfaced through both "Price Tips" and "Smart Pricing" were generated by a new version of Airbnb's pricing algorithm. Relative to the status quo algorithm, the treatment algorithm generally increased prices. For instance, on unconstrained smart pricing nights (e.g., calen-

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

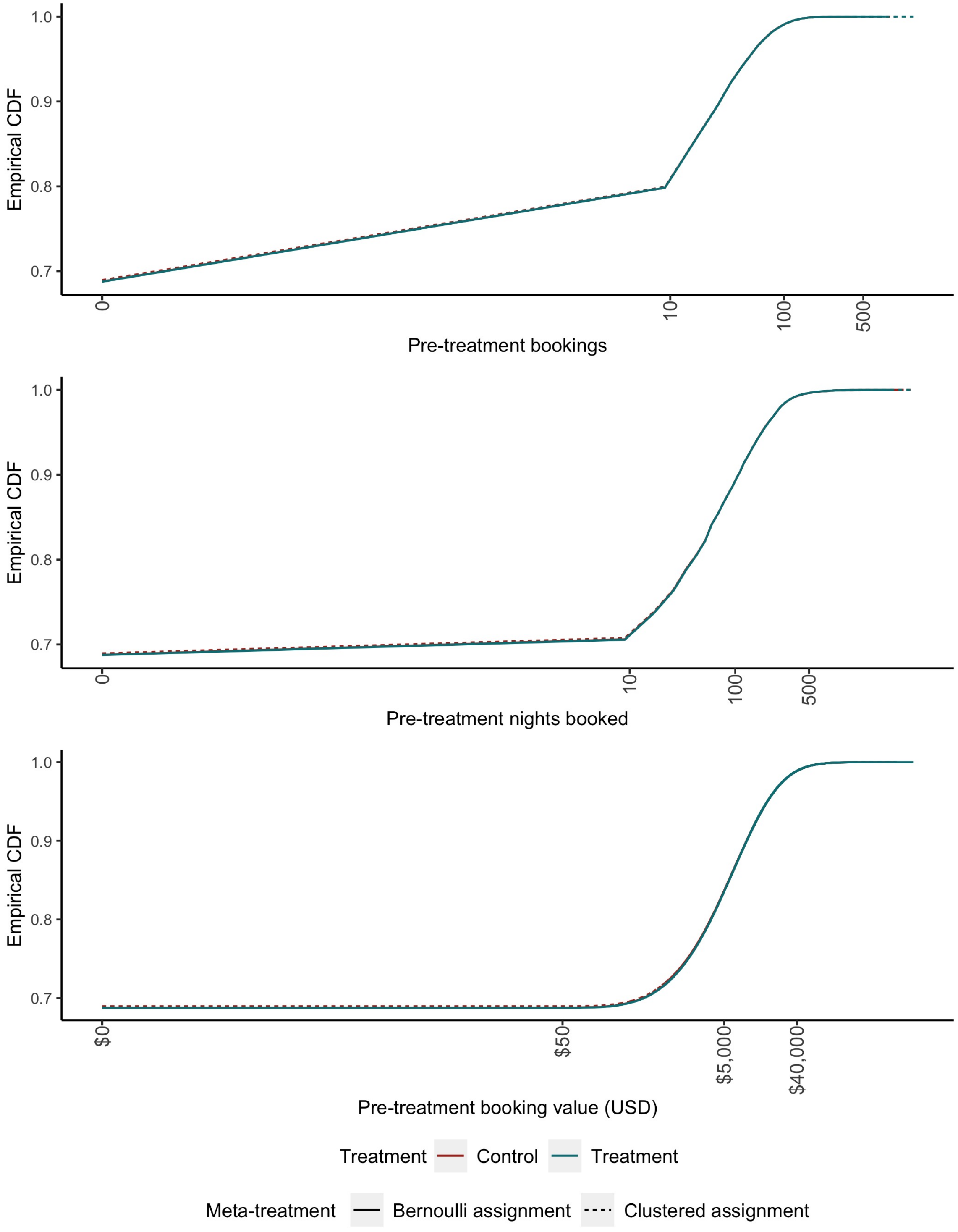


Figure 3.1: The empirical CDFs for pre-treatment bookings, nights booked, and booking value in each of the four treatment/meta-treatment groups for the algorithmic pricing meta-experiment.

dar nights in which hosts had opted into smart pricing and the suggested price was not subject to a minimum or maximum price threshold), prices increased by 4% on average.[17]

3.6.2 Results

In this section, we present results from the algorithmic pricing experiment. As was true for the fee meta-experiment, we report effects of the treatment on bookings per listing, but found qualitatively similar results for nights booked per listing and gross guest spend per listing, which can be found in Appendix 3.B.[18] Since, on average, the treatment pricing algorithm increased prices, we expect the TATE on bookings per listing to be negative. We first present the results separately analyzing the Bernoulli randomized arm of the meta-experiment and the cluster randomized arm of the meta-experiment. We then proceed to jointly analyze both meta-treatment arms, in order to test for the existence of interference bias in the Bernoulli randomized experiment's TATE estimate.

Bernoulli & Cluster Randomized Results

We analyze both the Bernoulli randomized and cluster randomized meta-treatment arms separately by estimating equation 3.3 on listing-level data.[19] As was the case with the fee meta-experiment, standard errors are clustered at the Airbnb listing-cluster level.

Table 3.1 shows the TATE estimate for bookings per listing in both the Bernoulli randomized and cluster randomized meta-treatment arms. In the Bernoulli randomized meta-treatment arm, the TATE is -0.106 bookings per listing, and this result is statistically significant at the 95% confidence level. In the cluster randomized meta-treatment arm, our point estimate of the TATE is -0.051 bookings per listing,

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

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

[19] As was the case with the fee meta-experiment, we present aggregate-level analysis of the cluster randomized meta-treatment arm in Appendix 3.C. The results from this analysis are nearly identical.

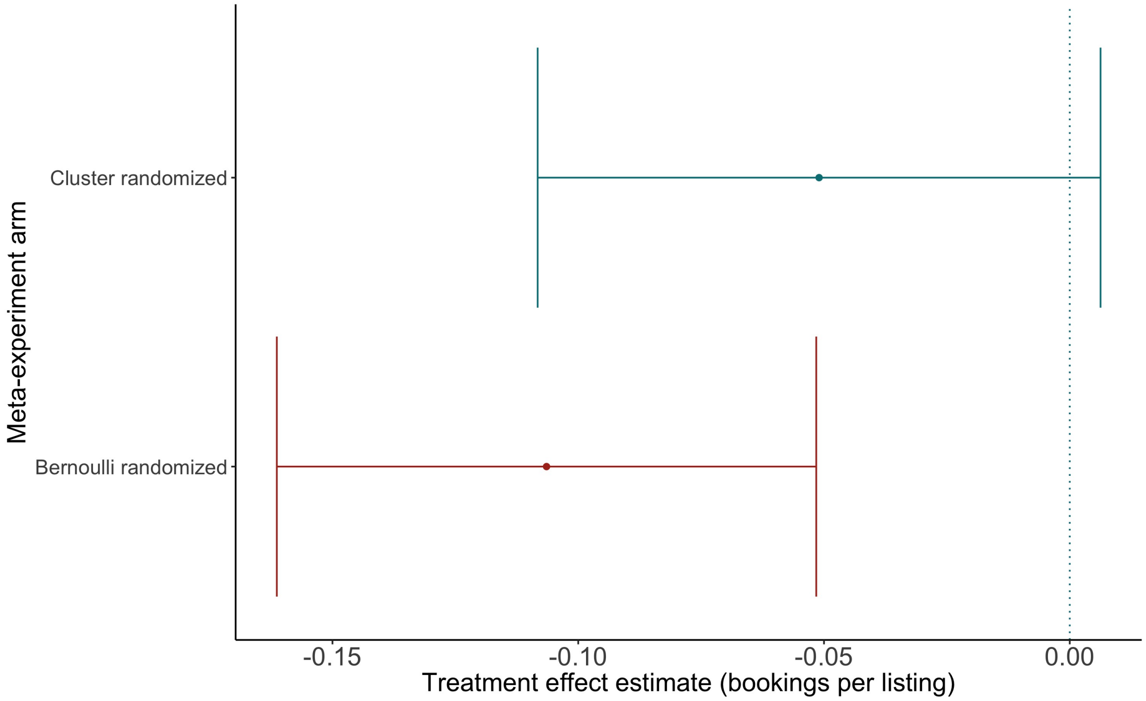


Figure 3.2: Total average treatment effect estimates for the algorithmic pricing experiment, estimated separately in the Bernoulli randomized meta-treatment arm and the cluster randomized meta treatment arm. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0 bookings per listing.

Table 3.1: Independent results of the algorithmic pricing meta-experiment

	Dependent variable:	
	Bookings	
	Bernoulli randomized	Cluster randomized
	(1)	(2)
Treatment	-0.106^{***}	-0.051^{*}
	(0.028)	(0.029)
Pre-treatment bookings	0.822^{***}	0.828^{***}
	(0.004)	(0.002)
Pre-treatment nights booked	-0.018^{***}	-0.017^{***}
	(0.001)	(0.000)
Pre-treatment booking value	0.000^{*}	0.000^{***}
	(0.000)	(0.000)
Pre-treatment gross guest spend	-0.000^{**}	-0.000^{***}
	(0.000)	(0.000)
Smart pricing pre-treatment	0.587^{***}	0.586^{***}
	(0.033)	(0.020)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	No	Yes
R^2	0.580	0.578
Adjusted R^2	0.578	0.578
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

however, this result is not statistically significant at the 95% confidence level. Figure 3.2 shows the estimated TATE in both meta-treatment arms, along with the corresponding 95% confidence intervals. In order to rigorously test whether or not cluster randomization led to a reduction in interference bias, we proceed to jointly analyze both meta-treatment arms.

Joint Analysis

In order to determine whether or not the two meta-treatment arms yield TATE estimates between which there is a statistically significant difference, we once again estimate equation 3.4.[20] As was the case with the fee meta-experiment, standard errors are clustered at the Airbnb listing-cluster level.

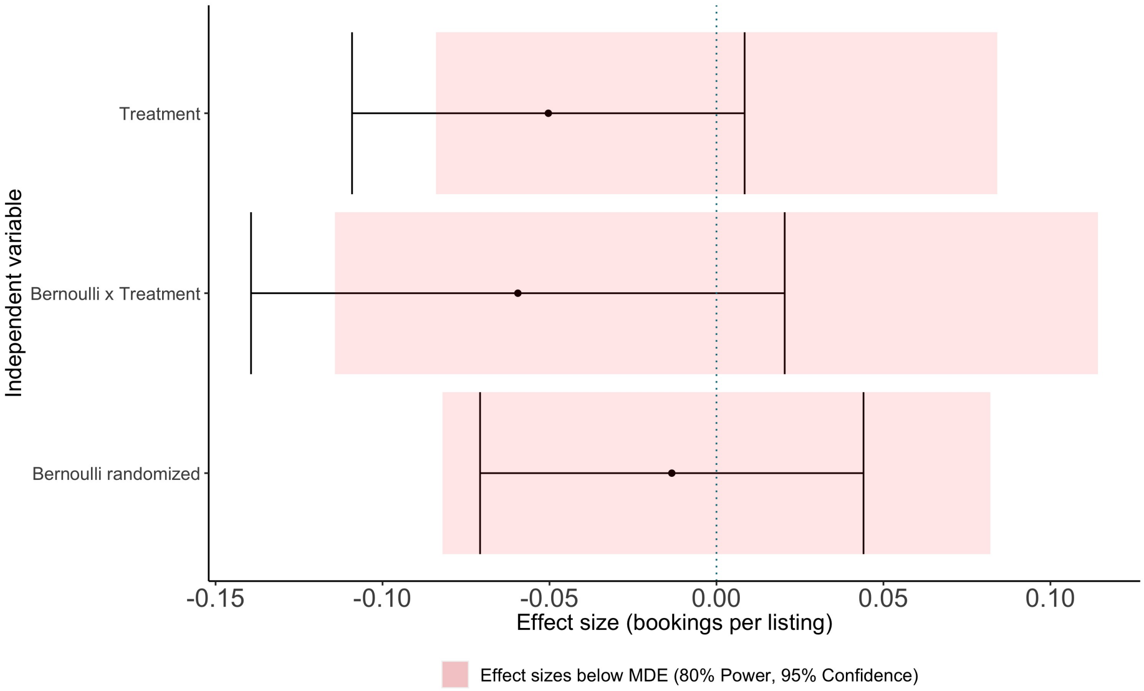


Figure 3.3: Coefficient estimates for the joint analysis of the algorithmic pricing meta-experiment. Error bars represent 95% confidence intervals. The dotted blue line correponds to a treatment effect of 0 bookings per listing. The red shaded area corresponds to values that are below the MDE (80% power, 95% confidence).

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

Table 3.2: Results of the algorithmic pricing meta-experiment

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

Table 3.2 shows our results, and Figure 3.3 displays our point estimate for each parameter in Equation 3.4, along with 95% confidence intervals. Point estimates imply that the "true" TATE is -0.050 bookings per listing, whereas -0.059 bookings per listing of the TATE measured in the Bernoulli randomized meta-treatment arm is due to interference bias. This would suggest that 54.16% (±65.05%) of the TATE achieved through a Bernoulli randomized experiment is due to interference that is eliminated by instead running a clustered experiment. However, none of these point estimates are statistically significant. A post-hoc power analysis of the algorithmic pricing experiment reveals that the meta-experiment is underpowered to detect reasonable effect sizes relative to the treatment effect estimated obtained in the Bernoulli randomized arm of the meta-experiment. Table 3.3 shows the calculated minimum detectable effect (MDE) for β, ν, and ξ. Each of these MDEs is also overlaid in red on Figure 3.3. Comparing the Bernoulli TATE estimate with the meta-experiment MDEs implies that interference bias would need to have approximately the same magnitude as our Bernoulli TATE estimate to be detectable.

Table 3.3: Minimum detectable effects for algorithmic pricing meta-experiment analysis (power = 80%, confidence level = 95%)

Regressor	Bookings
Treatment	0.084
Bernoulli x Treatment	0.114
Bernoulli randomized	0.082

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

3.7 Discussion

In this paper, we have taken the first empirical steps to understand the extent to which statistical inference can bias total average treatment effect estimates in online marketplace experiments. We have achieved this by presenting the results from two different pricing-related meta-experiments conducted on Airbnb, an online marketplace for accommodations. In each meta-experiment, some clusters of listings were assigned their experimental treatment using Bernoulli randomization, whereas others were assigned to their experimental treatment using cluster randomization. The motivation for our focus on pricing-related interventions was twofold; understanding customer price elasticities is crucial for both platform intermediaries and sellers, and there are strong reasons to suspect that pricing-relating experiments violate the stable unit treatment value assumption.

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

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

In addition to cluster randomization, there are a number of analysis techniques that have been developed in the network experimentation literature, such as exposure modeling (Aronow et al. 2017), regression adjustment (Chin 2019), and exact tests for interference (Athey et al. 2018) that, if adopted to a commerce-based setting, could help to more accurately identify and reduce interference bias in online marketplace experiments. Furthermore, there are a number of open questions regarding how to best identify the sellers most likely to interfere with one another in an online marketplace setting. The clustering method described in this paper is by no means the only (or best) way to cluster sellers before performing cluster randomization. Higher quality clusters could lead to even greater interference bias reductions. Finally, while the approach described in this work can reduce bias due to interference between sellers, it does not consider the reduction of bias due to interference between *buyers.* Given that, in general, online marketplaces have much less information about buyers, many of the approaches discussed thus far are unlikely to be effective. Developing methods that reduce interference bias on the buyer side of online marketplaces is a promising direction for future research.

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

Appendix

3.A Method for cluster size selection

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

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

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

$$\% \text{ in single cluster } = \frac{1}{n_{users}} \sum_{\text{all users}} \mathbb{1}\left(n_{clusters} = 1\right) \tag{3.5}$$

$$\text{Demand capture } = \frac{1}{n_{users}} \sum_{\text{all users}} \left(1 - \frac{n_{clusters}}{n_{listings}}\right) \tag{3.6}$$

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

Based on statistical power considerations, we decided that a cluster size threshold of 1,000 was the maximum threshold worth considering. Once this decision was made, we began to more directly compare the status quo threshold of 250 listings (which had been used for the algorithmic pricing meta-experiment) to the maximum threshold of 1,000 listings.[21] In doing so, we created an alternative demand capture measure that asked the following question: given a set of clusters, what percentage of listing viewers have at least $x\%$ of their listings captured by one cluster? Figure 3.A.2 plots this measure for both the 250 listing threshold clusters and the 1,000 listing threshold clusters, with demand capture thresholds of 67%, 75%, and 90%. As expected, the clusters with the 1,000 listing threshold do a better job of capturing demand than the 250 listing threshold clusters.

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

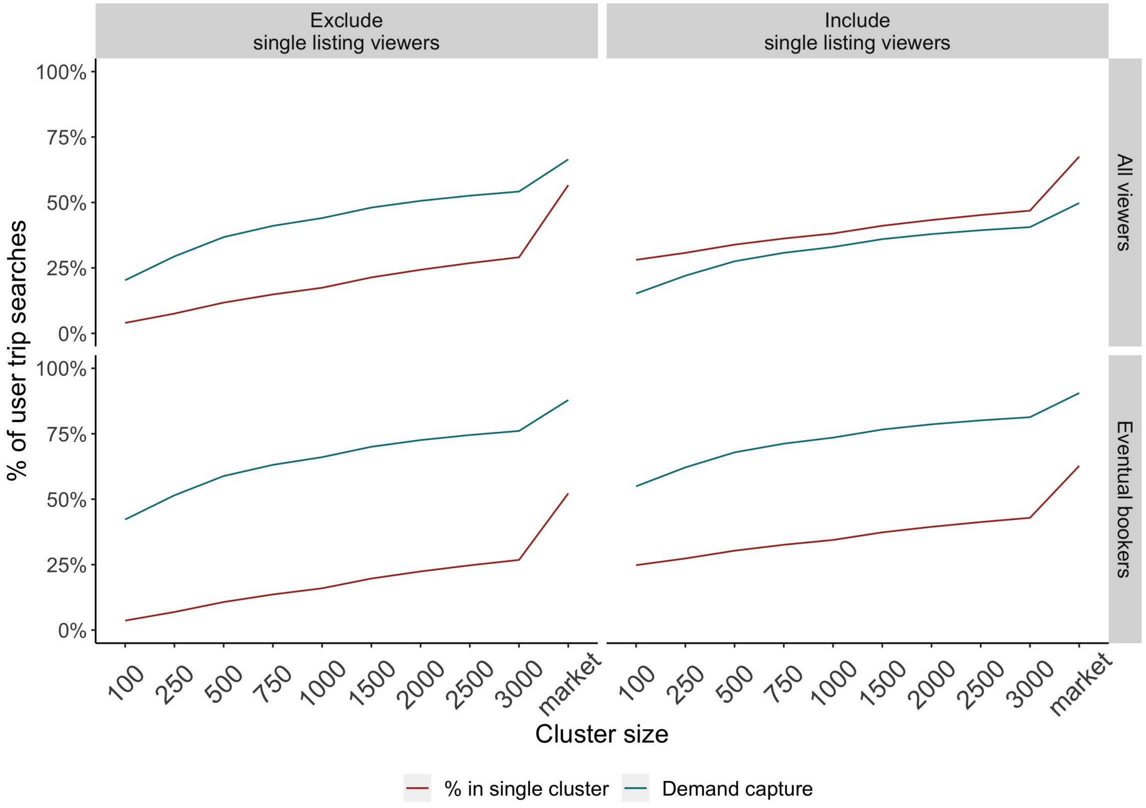


Figure 3.A.1: The relationship between cluster size and demand capture for two different metrics. The left column excludes users who only view a single Airbnb listing, whereas the right column includes them. The top row includes all listing viewers, whereas the bottom row only includes Airbnb users who go on to eventually book a listing.

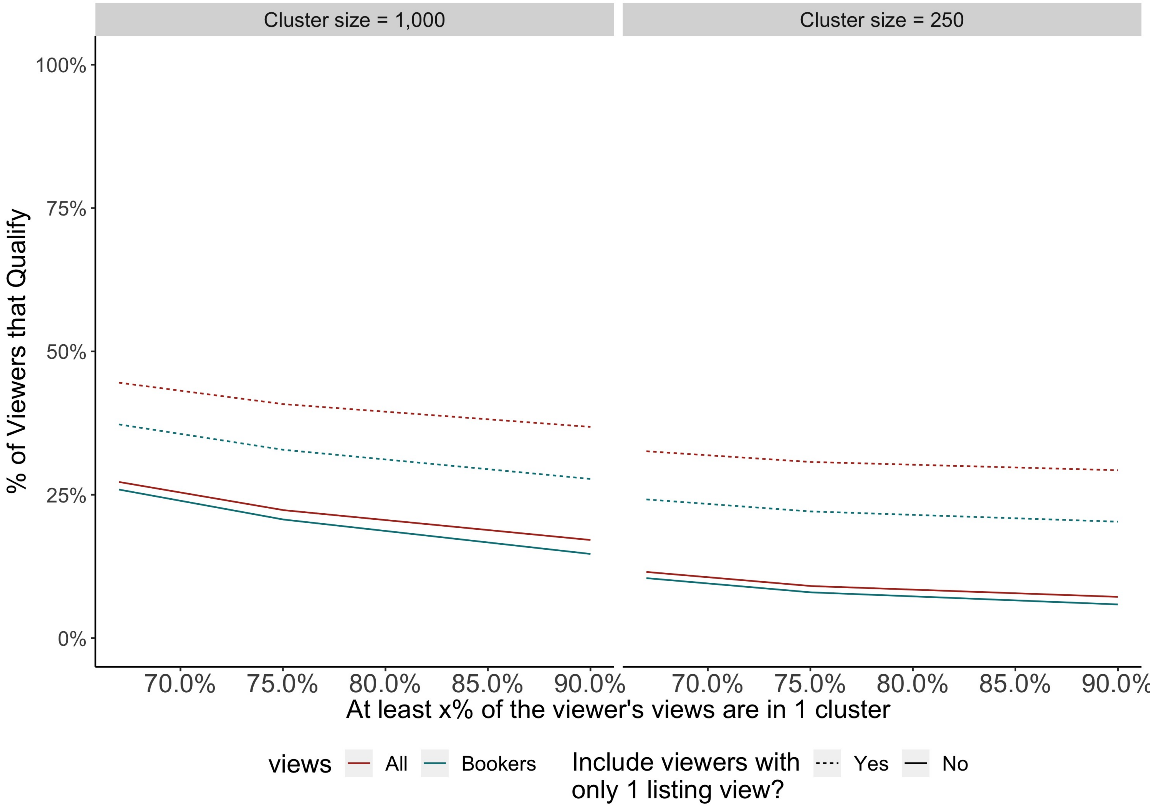


Figure 3.A.2: A direct comparison of the demand capture of clusters with a 1,000 listing threshold, and clusters with a 250 listing threshold. Curves show the percentage of viewers for whom at least $x\%$ of their views are contained by one cluster. Red curves include all listing viewers, whereas blue curves only include Airbnb users who go on to eventually book a listing. Dashed lines include users who only view a single Airbnb listing, whereas solid lines do not.

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

$$MDE_{actual} = \frac{MDE_{ideal}}{\text{Demand capture}}. \tag{3.7}$$

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

$$\frac{\text{Demand capture}_{1,000}}{\text{Demand capture}_{250}} > \frac{MDE_{ideal_{250}}}{MDE_{ideal_{1,000}}} \rightarrow \frac{\text{Demand capture}_{1,000}}{\text{Demand capture}_{250}} > \frac{1.05\%}{0.9\%} \tag{3.8}$$

Table 3.A.1 shows the ratio of demand capture for clusters with a threshold of 1,000 listings to the demand capture for clusters with a threshold of 250 clusters according to five different demand capture measures: the average share of search listings belonging to a cluster, the average user-level Herfindahl-Hirschman index across clusters, and the percentage of users for which one cluster accounts for at least 67%, 75%, and 90% of listings viewed. Across all five of these demand capture metrics, and across different user subpopulations, the demand capture ratio is consistently above $\frac{1.05\%}{0.9\%} = 1.17$. Based on this calculation, we determined that clusters with a threshold of 1,000 listings were preferable.

Table 3.A.1: The ratio of demand capture for 1,000 listing threshold clusters and 250 listing threshold clusters, using different demand capture metrics and user subpopulations.

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

3.B Interference bias for nights booked and gross guest spend

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

3.B.1 Fee meta-experiment

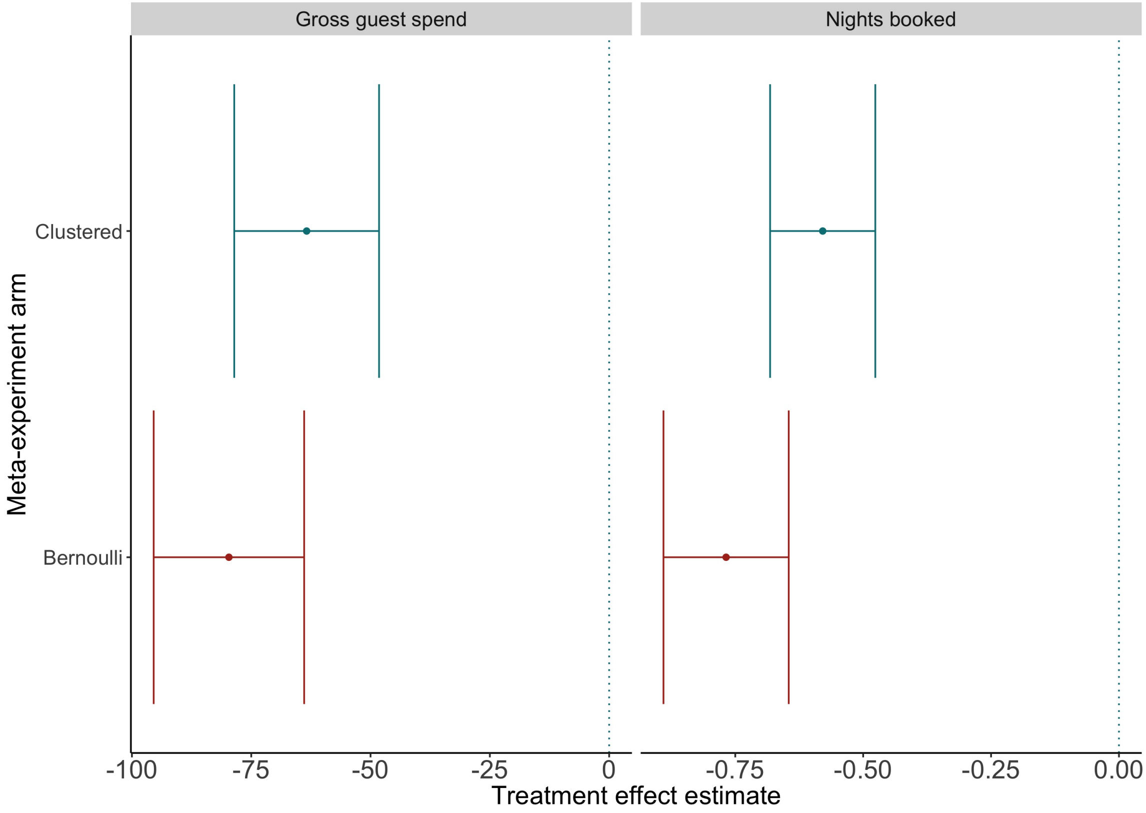


Figure 3.B.1: Total average treatment effect estimates (nights booked per listing and gross guest spend per listing) for the fee experiment, estimated separately in the Bernoulli randomized meta-treatment arm and the cluster randomized meta treatment arm. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0.

Table 3.B.1: Independent results of the fee meta-experiment (nights booked and gross guest spend)

	Dependent variable:			
	Nights booked		Gross guest spend	
	Bernoulli	Cluster	Bernoulli	Cluster
	(1)	(2)	(3)	(4)
Treatment	−0.768***	−0.579***	−79.677***	−63.388***
	(0.062)	(0.052)	(8.044)	(7.741)
Pre-treatment bookings	0.281***	0.288***	23.220***	22.626***
	(0.005)	(0.003)	(0.750)	(0.372)
Pre-treatment nights booked	0.038***	0.037***	−4.289***	−3.698***
	(0.002)	(0.001)	(0.433)	(0.129)
Pre-treatment booking value	−0.000***	−0.000***	−0.060	−0.148***
	(0.000)	(0.000)	(0.085)	(0.021)
Pre-treatment gross guest spend	0.000***	0.000***	0.153**	0.226***
	(0.000)	(0.000)	(0.070)	(0.017)
Stratum F.E.	Yes	Yes	Yes	Yes
Robust s.e.	Yes	Yes	Yes	Yes
Clustered s.e.	No	Yes	No	Yes
R^2	0.115	0.118	0.166	0.176
Adjusted R^2	0.114	0.118	0.165	0.176
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3.B.1 shows the estimated effect of the fee treatment in both the Bernoulli randomized meta-treatment arm and the cluster randomized meta-treatment arm on both nights booked per listing and gross guest spend per listing. Our TATE estimates for each outcome are also depicted, along with 95% confidence intervals, in Figure 3.B.1. We estimate in the Bernoulli randomized meta-treatment arm that the treatment led to a statistically significant loss of 0.768 nights booked per listing and $79.68 in gross guest spend per listing, whereas we estimate in the cluster randomized meta-treatment arm that the treatment led to a statistically significant loss of 0.579 nights booked per listing and $63.39 in booking value per listing.

In order to test whether or not there is a statistically significant difference between the TATE estimates in the two meta-treatment arms, we conduct a joint analysis of both meta-treatment arms simultaneously. Table 3.B.2 shows our results. Our results are also depicted in Figure 3.B.2, along with 95% confidence intervals. We find statistically significant evidence of interference bias in the Bernoulli TATE estimate for nights booked per listing at the 95% confidence level, but do not find statistically

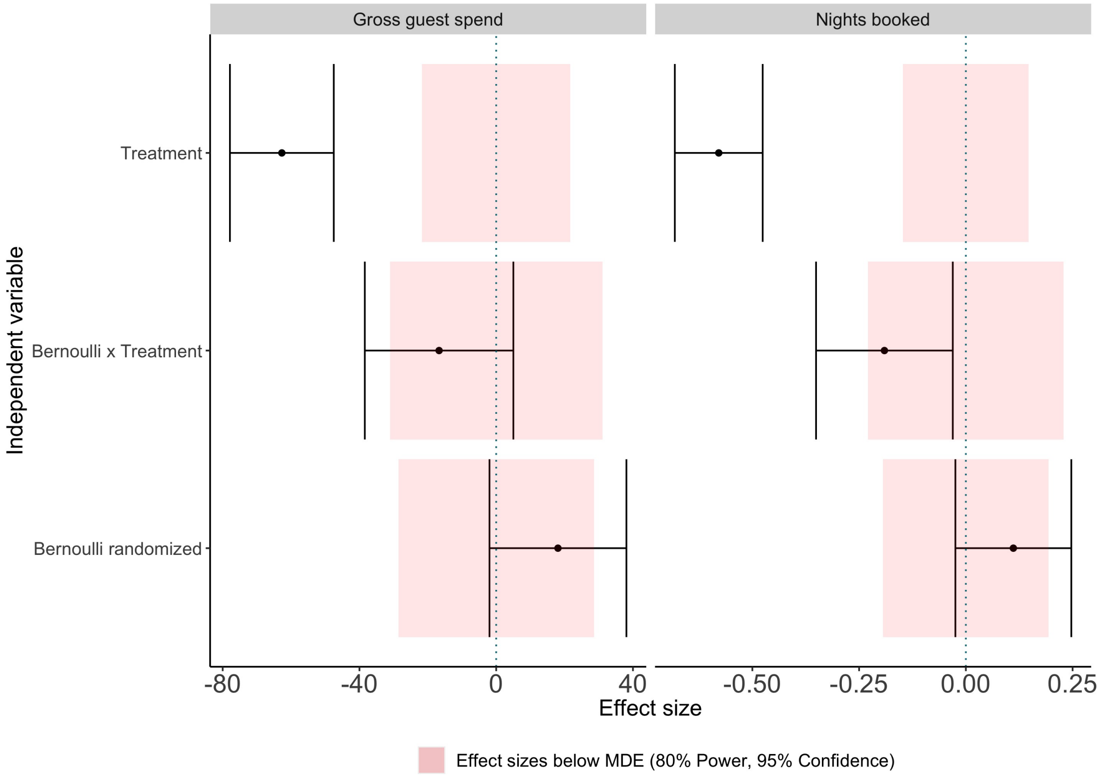


Figure 3.B.2: Coefficient estimates for the joint analysis of the fee meta-experiment (nights booked per listing and gross guest spend per listing). Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0. The red shaded area corresponds to values that are below the MDE (80% power, 95% confidence).

Table 3.B.2: Results of the fees Meta-experiment (nights booked and gross guest spend)

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

significant evidence of interference bias in the Bernoulli TATE estimate for gross guest spend per listing. Our point estimates suggest that interference accounts for 24.79% of the Bernoulli TATE estimate for nights booked per listing (stat sig.) and 21.04% of the Bernoulli TATE estimate for gross guest spend per listing (not stat. sig).

3.B.2 Algorithmic pricing meta-experiment

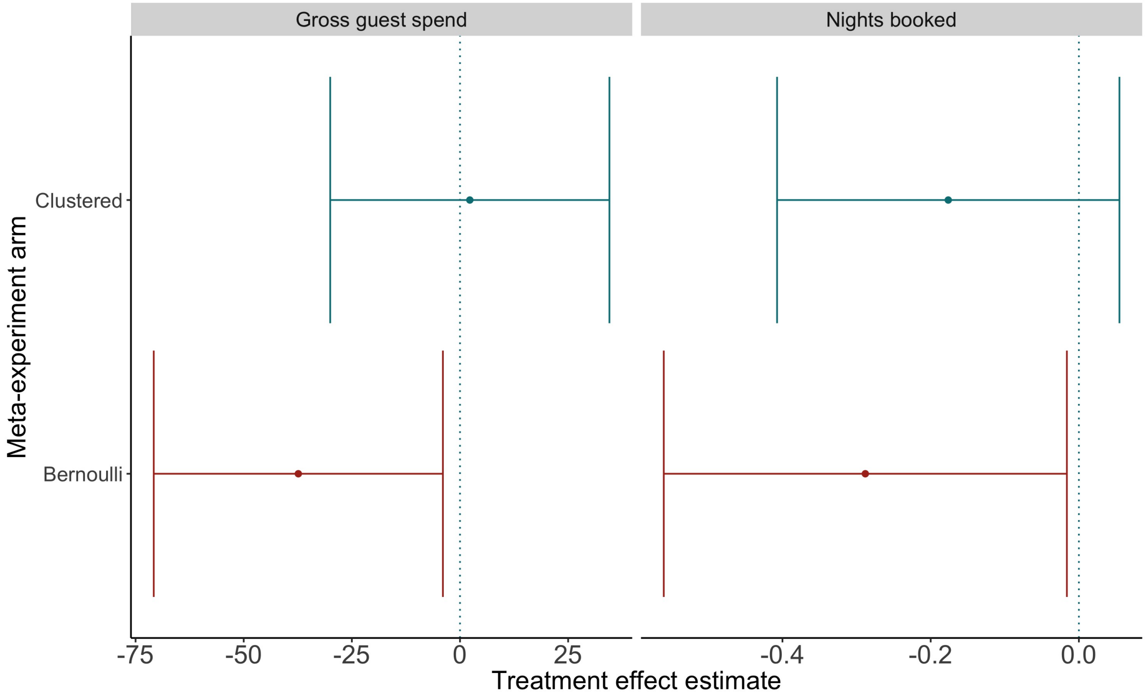


Figure 3.B.3: Total average treatment effect estimates (nights booked per listing and gross guest spend per listing) for the algorithmic pricing experiment, estimated separately in the Bernoulli randomized meta-treatment arm and the cluster randomized meta treatment arm. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0.

Table 3.B.3 shows the estimated effect of the algorithmic pricing treatment in both the Bernoulli randomized meta-treatment arm and the cluster randomized meta-treatment arm on both nights booked per listing and gross guest spend per listing. Our TATE estimates for each outcome are depicted, along with 95% confidence intervals, in Figure 3.B.3, We estimate in the Bernoulli randomized meta-treatment arm that the treatment let do a statistically significant loss of 0.288 nights booked

Table 3.B.3: Independent results of the algorithmic pricing meta-experiment (nights booked and gross guest spend)

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

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

per listing and $37.38 in gross guest spend per listing, whereas we do not detect a statistically significant treatment effect for either outcome in the cluster randomized meta-treatment arm.

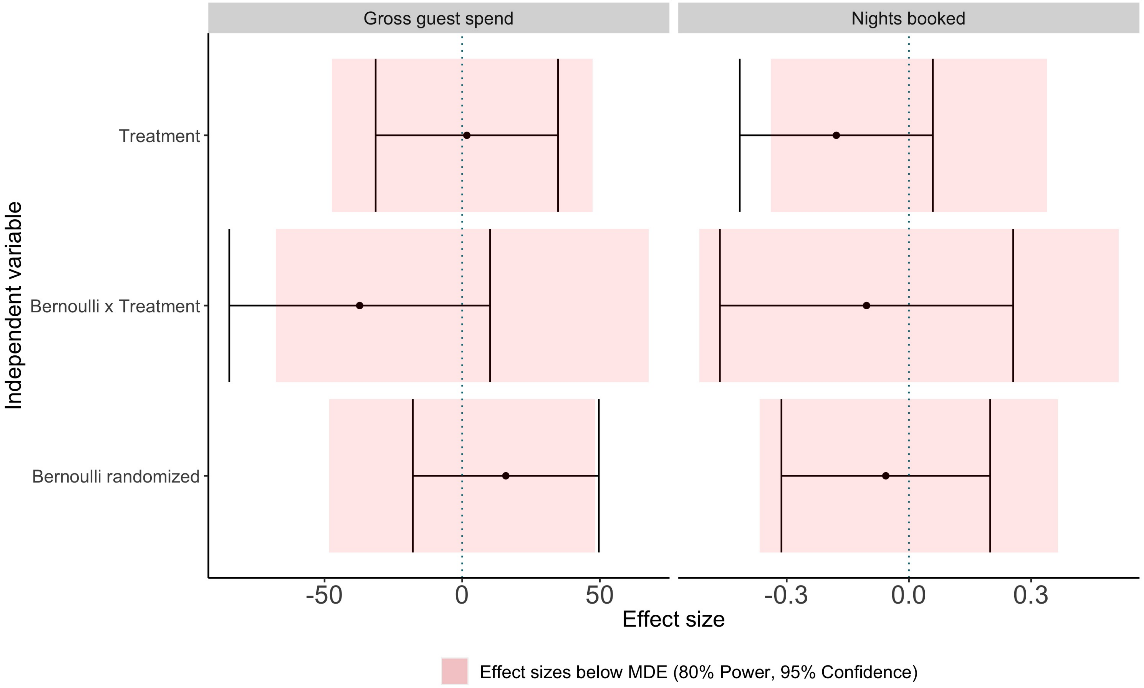


Figure 3.B.4: Coefficient estimates for the joint analysis of the algorithmic pricing meta-experiment (nights booked per listing and gross guest spend per listing). Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0. The red shaded area corresponds to values that are below the MDE (80% power, 95% confidence).

In order to test whether or not there is a statistically significant difference between the TATE estimates in the two meta-treatment arms, we conduct a joint analysis of both meta-treatment arms simultaneously. Table 3.B.4 shows our results. Our results are also depicted in Figure 3.B.4, along with 95% confidence intervals. We do not find statistically significant evidence for interference bias for either outcome. While not statistically significant, our point estimates suggest that interference accounts for 36.86% of the Bernoulli TATE estimate for nights booked per listing and 104.73% of the Bernoulli TATE estimate for gross guest spend per listing.

Table 3.B.4: Results of the algorithmic pricing meta-experiment (nights booked and gross guest spend)

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

3.C Cluster-level analysis of cluster-randomized meta-treatment arm

Rather than analyzing data from the cluster randomized meta-treatment arm of our experiments at the individual level with clustered standard errors, it is also possible to aggregate data at the *cluster* level and instead estimated a weighted version of Equation 3.3, where each cluster is weighted according to the number of experiment-eligible listings in that cluster. In this appendix, we compare the cluster randomized TATE estimates obtained using these two different approaches.

3.C.1 Fee meta-experiment

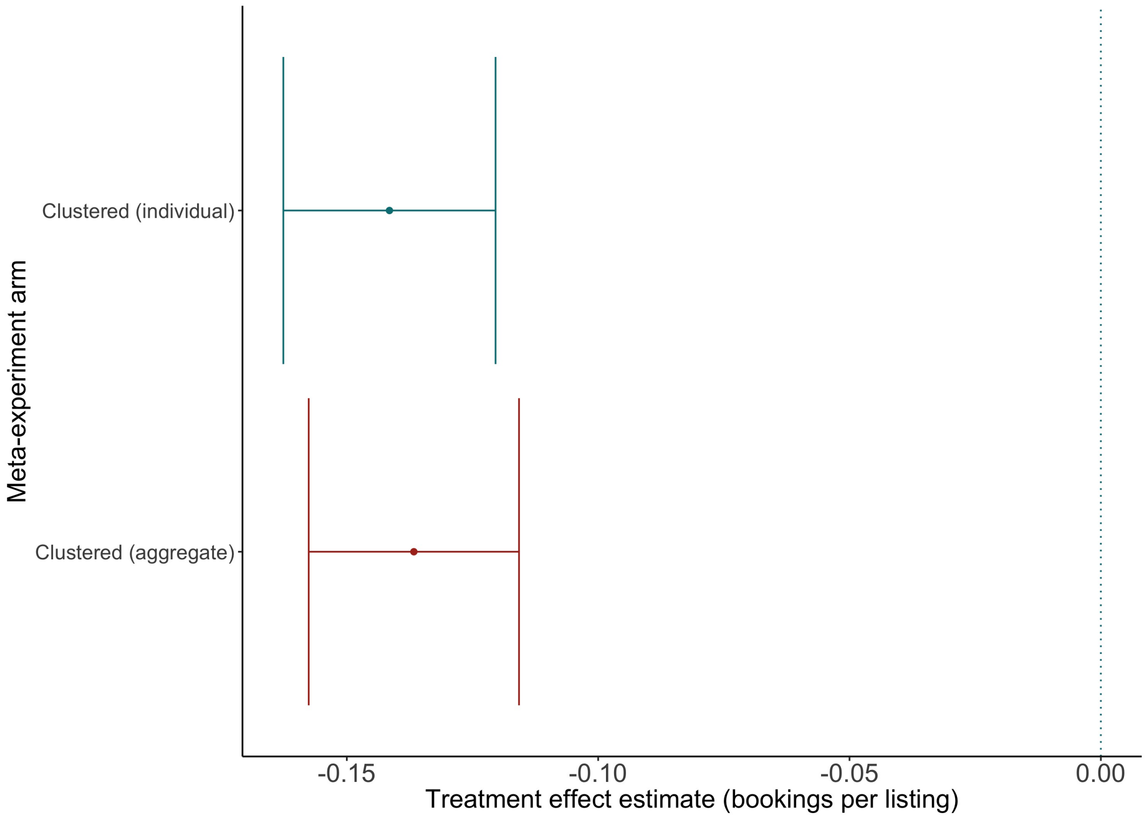


Figure 3.C.1: Comparison of the TATE estimates from the cluster randomized meta-treatment arm of the fees experiment, obtained analyzing data at either the individual listing level or at the cluster level. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0 bookings per listing.

Table 3.C.1: Cluster randomized fees experiment (individual- and cluster-level analysis)

	Dependent variable:	
	Individual-level	Cluster-level
	(1)	(2)
Treatment	-0.142^{***}	-0.137^{***}
	(0.011)	(0.011)
Pre-treatment bookings	0.174^{***}	0.206^{***}
	(0.001)	(0.006)
Pre-treatment nights booked	-0.003^{***}	0.003^{*}
	(0.000)	(0.002)
Pre-treatment booking value	0.000^{***}	-0.000
	(0.000)	(0.000)
Pre-treatment gross guest spend	-0.000^{***}	0.000
	(0.000)	(0.000)
Stratum F.E.	Yes	Yes
Robust s.e.	Yes	Yes
Clustered s.e.	Yes	No
R^2	0.405	0.973
Adjusted R^2	0.405	0.968
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3.C.1 compares the TATE estimates obtained from the cluster randomized meta-treatment arm of the fee meta-experiment when analyzing the data at both the individual listing level and at the cluster level. Our results are also depicted in Figure 3.C.1. We find that both approaches yield almost identical TATE point estimates and standard errors.

3.C.2 Algorithmic pricing meta-experiment

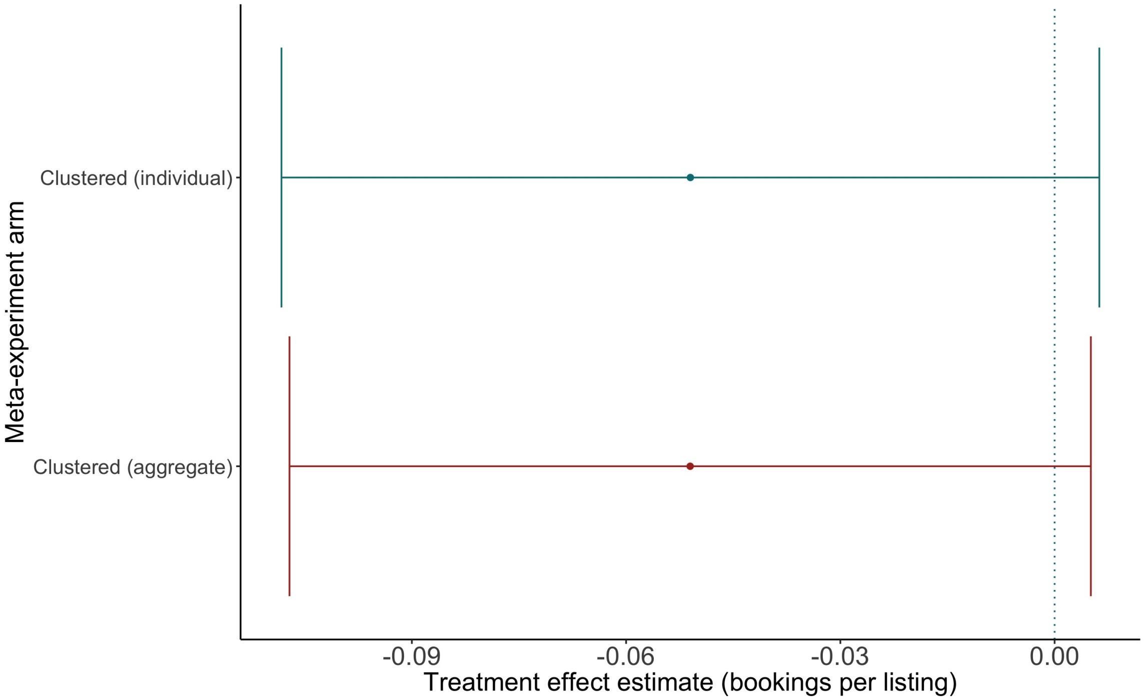


Figure 3.C.2: Comparison of the TATE estimates from the cluster randomized meta-treatment arm of the algorithmic pricing experiment, obtained analyzing data at either the individual listing level or at the cluster level. Error bars represent 95% confidence intervals. The dotted blue line corresponds to a treatment effect of 0 bookings per listing.

Table 3.C.2 compares the TATE estimates obtained from the cluster randomized meta-treatment arm of the algorithmic pricing meta-experiment when analyzing the data at both the individual listing level and at the cluster level. Our results are also depicted in Figure 3.C.2. We find that both approaches yield almost identical TATE point estimates and standard errors.

Table 3.C.2: Cluster randomized algorithmic pricing experiment (individual- and cluster-level analysis)

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

3.D Results with mixed units of analysis

In addition to performing joint analysis of our meta-experiments with listing-level data, it is possible to analyze the meta-experiments with a mixture of listing-level data and data aggregated at the cluster level. For both meta-experiments, we estimate Equation 3.4 on listing-level data from the Bernoulli randomized meta-treatment arm and cluster-level data from the cluster randomized meta-treatment arm. Observations are weighted by the number of listings making up that observation (i.e., listings receive a weight of 1, whereas clusters receive a weight equal to the number of experiment eligible listings in that cluster). In this appendix, we compare results obtained using this approach with those obtained analyzing the meta-experiment entirely with listing level data.

3.D.1 Fee meta-experiment

Table 3.D.1 compares results obtained by analyzing the fee meta-experiment at the listing level and with mixed units of analysis. Our results are also depicted in Figure 3.D.1. We find that both approaches yield almost identical results.

3.D.2 Algorithmic pricing meta-experiment

Table 3.D.2 compares results obtained by analyzing the fee meta-experiment at the listing level and with mixed units of analysis. Our results are also depicted in Figure 3.D.2. We find that both approaches yield almost identical results.

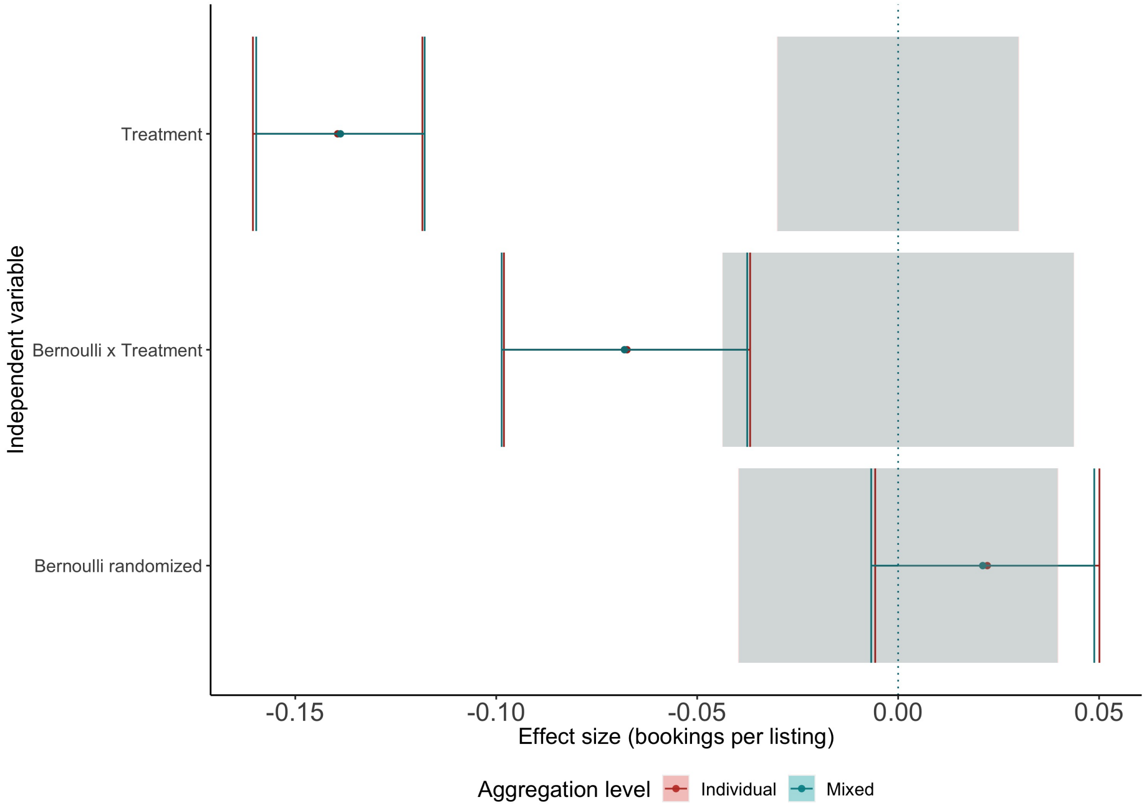


Figure 3.D.1: Comparison of fee experiment meta-analysis estimates obtained analyzing data at the individual level of analysis, and the mixed level of analysis. In the mixed analysis, Data from Bernoulli randomized listings is included at the listing level, whereas data from cluster randomized listings is aggregated at the cluster level. Error bars correspond to 95% confidence intervals. Shaded areas represent effect sizes below the MDE threshold (80% power, 95% confidence).

Table 3.D.1: Results of the fees Meta-experiment (individual and mixed analysis)

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

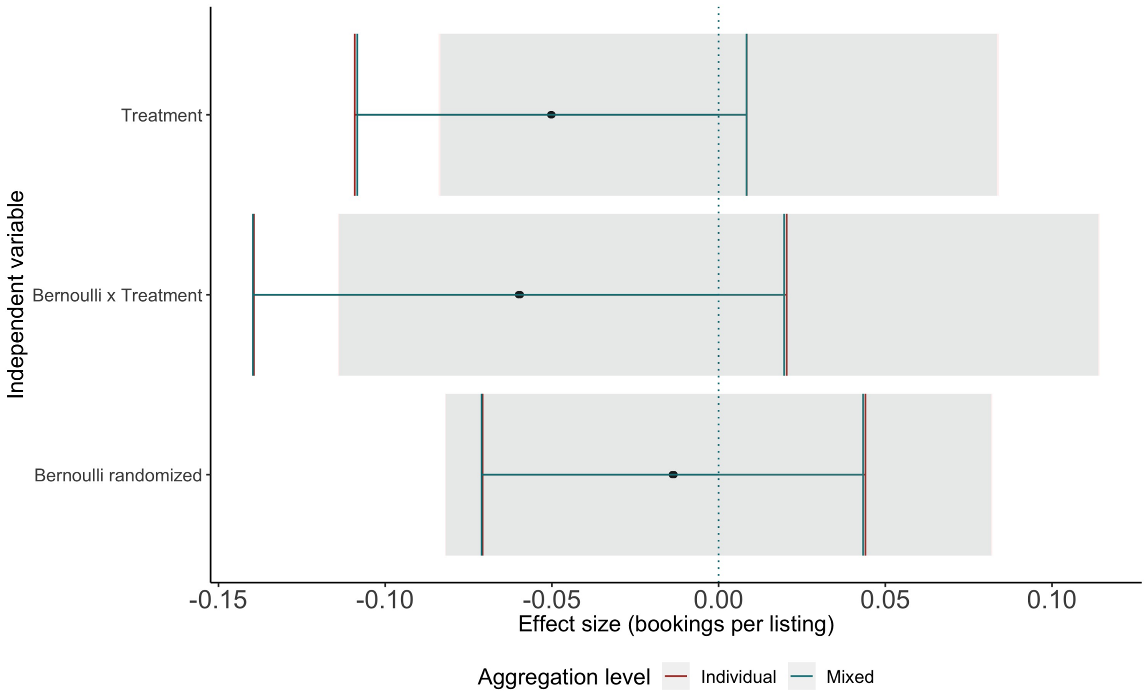


Figure 3.D.2: Comparison of algorithmic pricing experiment meta-analysis estimates obtained analyzing data at the individual level of analysis, and the mixed level of analysis. In the mixed analysis, Data from Bernoulli randomized listings is included at the listing level, whereas data from cluster randomized listings is aggregated at the cluster level. Error bars correspond to 95% confidence intervals. Shaded areas represent effect sizes below the MDE threshold (80% power, 95% confidence).

Table 3.D.2: Results of the algorithmic pricing meta-experiment (individual and mixed analysis)

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

Bibliography

Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering* (6):734–749.

Aguiar L, Waldfogel J (2018) Platforms, promotion, and product discovery: Evidence from spotify playlists. Technical report, National Bureau of Economic Research.

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

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

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

Anderson A, Maystre L, Anderson I, Mehrotra R, Lalmas M (2020) Algorithmic effects on the diversity of consumption on spotify. *Proceedings of The Web Conference 2020*, 2155–2165.

Aral S, Van Alstyne M (2011) The diversity-bandwidth trade-off. *American journal of sociology* 117(1):90–171.

Aronow PM, Samii C, et al. (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.

Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

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.

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

Bojinov I, Simchi-Levi D, Zhao J (2020) Design and analysis of switchback experiments. *Available at SSRN 3684168* .

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

Brynjolfsson E, Hu Y, Simester D (2011) Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science* 57(8):1373–1386.

Castells P, Hurley NJ, Vargas S (2015) Novelty and diversity in recommender systems. *Recommender systems handbook*, 881–918 (Springer).

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

Clarke B (2016) Why these tech companies keep running thousands of failed experiments. URL `https://www.fastcompany.com/3063846/why-these-tech-companies-keep-running-thousands-of-failed`.

Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. *Physical review E* 70(6):066111.

Claussen J, Peukert C, Sen A (2019) The editor vs. the algorithm: Targeting, data and externalities in online news. *Data and Externalities in Online News (June 5, 2019)* .

Conlon CT, Mortimer JH (2013) Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics* 5(4):1–30.

Cox DR (1958) Planning of experiments. .

Das AS, Datar M, Garg A, Rajaram S (2007) Google news personalization: scalable online collaborative filtering. *Proceedings of the 16th international conference on World Wide Web*, 271–280 (ACM).

Datta H, Knox G, Bronnenberg BJ (2018) Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery. *Marketing Science* 37(1):5–21.

De P, Hu Y, Rahman MS (2010) Technology usage and online sales: An empirical study. *Management Science* 56(11):1930–1945.

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

Dewan S, Ramaprasad J (2012) Research note—music blogging, online sampling, and the long tail. *Information Systems Research* 23(3-part-2):1056–1067.

Dhar V, Geva T, Oestreicher-Singer G, Sundararajan A (2014) Prediction in economic networks. *Information Systems Research* 25(2):264–284, ISSN 15265536, URL `http://dx.doi.org/10.1287/isre.2013.0510`.

Ding P, Lu J (2017) Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79(3):757–777.

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).

Edison Research (2019) The podcast consumer 2019. URL `https://www.edisonresearch.com/the-podcast-consumer-2019`.

Farronato C, Fradkin A (2018) The welfare effects of peer entry in the accommodation market: The case of airbnb. Technical report, National Bureau of Economic Research.

Feller A, Mealli F, Miratrix L (2017) Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics* 42(6):726–758.

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).

Flaxman S, Goel S, Rao JM (2016) Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80(S1):298–320.

Fleder D, Hosanagar K (2009) Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science* 55(5):697–712.

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

Frangakis CE, Rubin DB (2002) Principal Stratification in Causal Inference. *Biometrics* 58(1):21–29, ISSN 1541-0420, URL `http://dx.doi.org/10.1111/j.0006-341X.2002.00021.x`.

Freyne J, Jacovi M, Guy I, Geyer W (2009) Increasing engagement through early recommender intervention. *Proceedings of the third ACM conference on Recommender systems*, 85–92 (ACM).

Gentzkow M, Shapiro JM (2006) Media bias and reputation. *Journal of political Economy* 114(2):280–316.

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).

Hájek J (1971) Comment on "an essay on the logical foundations of survey sampling, part one,". *The Foundations of Survey Sampling* 236.

Hervas-Drane A (2015) Recommended for you: The effect of word of mouth on sales concentration. *International Journal of Research in Marketing* 32(2):207–218.

Holtz D, Lobel R, Liskovich I, Aral S (2020) Reducing interference bias in online marketplace pricing experiments. *arXiv preprint arXiv:2004.12489* .

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

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

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47(260):663–685.

Hosanagar K, Fleder D, Lee D, Buja A (2013) Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. *Management Science* 60(4):805–823.

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

Jannach D, Lerche L, Gedikli F, Bonnin G (2013) What recommenders recommend–an analysis of accuracy, popularity, and sales diversity effects. *International Conference on User Modeling, Adaptation, and Personalization*, 25–37 (Springer).

Johari R, Li H, Liskovich I, Weintraub G (2020) Experimental design in two-sided platforms: An analysis of bias. *arXiv preprint arXiv:2002.05670* .

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.

Knox G, Datta H (2020) Streaming services and the homogenization of music consumption .

Kohavi R, Longbotham R, Sommerfield D, Henne RM (2009) Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18(1):140–181.

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

Lacerda A (2017) Multi-objective ranked bandits for recommender systems. *Neurocomputing* 246:12–24.

Lamere P, Green S (2008) Project aura: recommendation for the rest of us. *Presentation at Sun JavaOne Conference.*

Lee D, Hosanagar K (2019) How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment. *Information Systems Research* 30(1):239–259.

Li JQ, Rusmevichientong P, Simester D, Tsitsiklis JN, Zoumpoulis SI (2015) The value of field experiments. *Management Science* 61(7):1722–1740.

Lin Z, Goh KY, Heng CS (2015) The demand effects of product recommendation networks: An empirical analysis of network diversity and stability. *Forthcoming in MIS Quarterly* .

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

Manski CF (2013) Identification of treatment response with social interactions. *The Econometrics Journal* 16(1):S1–S23.

Marler RT, Arora JS (2004) Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization* 26(6):369–395.

McInerney J, Lacker B, Hansen S, Higley K, Bouchard H, Gruson A, Mehrotra R (2018) Explore, exploit, and explain: personalizing explainable recommendations with bandits. *Proceedings of the 12th ACM Conference on Recommender Systems*, 31–39 (ACM).

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.

Mitchell M, Baker D, Moorosi N, Denton E, Hutchinson B, Hanna A, Gebru T, Morgenstern J (2020) Diversity and inclusion metrics in subset selection. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 117–123.

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

Nazari Z, Charbuillet C, Pages J, Laurent M, Charrier D, Vecchione B, Carterette B (2020) Recommending podcasts for cold-start users based on music listening and taste. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1041–1050.

Negroponte N (1996) *Being digital* (Vintage).

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.

Nguyen TT, Hui PM, Harper FM, Terveen L, Konstan JA (2014) Exploring the filter bubble: the effect of using recommender systems on content diversity. *Proceedings of the 23rd international conference on World wide web*, 677–686 (ACM).

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

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

Pariser E (2011) *The filter bubble: How the new personalized web is changing what we read and how we think* (Penguin).

Pouget-Abadie J, Saveski M, Saint-Jacques G, Duan W, Xu Y, Ghosh S, Airoldi E (2017) Testing for arbitrary interference on experimentation platforms. *preprint* .

Resnick P, Varian HR (1997) Recommender systems. *Communications of the ACM* 40(3):56–59.

Ribeiro MH, Ottoni R, West R, Almeida VA, Meira W (2019) Auditing radicalization pathways on youtube. *arXiv preprint arXiv:1908.08313* .

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

Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311(5762):854–856.

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).

Senecal S, Nantel J (2004) The influence of online product recommendations on consumers' online choices. *Journal of retailing* 80(2):159–169.

Shannon CE (1948) A mathematical theory of communication. *Bell system technical journal* 27(3):379–423.

Sharma A, Hofman JM, Watts DJ, et al. (2018) Split-door criterion: Identification of causal effects through auxiliary outcomes. *The Annals of Applied Statistics* 12(4):2699–2733.

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

Sneider C, Tang Y, Tang Y (2019) Experiment rigor for switchback experiment analysis. URL `https://doordash.engineering/2019/02/20/experiment-rigor-for-switchback-experiment-analysis/`.

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

Sunstein CR (2001) *Republic.com* (Princeton university press).

Teachman JD (1980) Analysis of population diversity: Measures of qualitative variation. *Sociological Methods & Research* 8(3):341–362.

Thompson C (2008) If you liked this, you're sure to love that. *The New York Times* 21.

Tufekci Z (2018) Youtube, the great radicalizer. *The New York Times* 10.

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).

Ugander J, Yin H (2020) Randomized graph cluster randomization. *arXiv preprint arXiv:2009.02297* .

Van Alstyne M, Brynjolfsson E (2005) Global village or cyber-balkans? modeling and measuring the integration of electronic communities. *Management Science* 51(6):851–868.

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world'networks. *nature* 393(6684):440–442.

Wu LL, Joung YJ, Chiang TE (2011) Recommendation systems and sales concentration: The moderating effects of consumers' product awareness and acceptance to recommendations. *2011 44th Hawaii International Conference on System Sciences*, 1–10 (IEEE).

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).

Zhou R, Khemmarat S, Gao L (2010) The impact of youtube recommendation system on video views. *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 404–410 (ACM).