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

David Holtz
MIT Sloan School of Management, Cambridge, MA 02139, dholtz@mit.edu

Benjamin Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer
Spotify, Inc., New York, NY 10007, benjaminc@spotify.com, praveenrchandar@spotify.com, zahran@spotify.com, henriette@spotify.com

Sinan Aral
MIT Sloan School of Management, Cambridge, MA 02139, sinan@mit.edu

It remains unknown whether personalized recommendations increase or decrease the diversity of content people consume. We present results from a randomized field experiment on Spotify testing the effect of personalized recommendations on consumption diversity. In the experiment, both control and treatment users were given podcast recommendations, with the sole aim of increasing podcast consumption. Treatment users' recommendations were personalized based on their music listening history, whereas control users were recommended popular podcasts among users in their demographic group. We find that, on average, the treatment increased podcast streams by 28.90%. However, the treatment also decreased the average individual-level diversity of podcast streams by 11.51%, and increased the aggregate diversity of podcast streams by 5.96%, indicating that personalized recommendations have the potential to create patterns of consumption that are homogenous within and diverse across users, a pattern reflecting Balkanization. Our results provide evidence of an "engagement-diversity trade-off" when recommendations are optimized solely to drive consumption: while personalized recommendations increase user engagement, they also affect the diversity of consumed content. This shift in consumption diversity can affect user retention and lifetime value, and impact the optimal strategy for content producers. We also observe evidence that our treatment affected streams from sections of Spotify's app not directly affected by the experiment, suggesting that exposure to personalized recommendations can affect the content that users consume organically. We believe these findings highlight the need for academics and practitioners to continue investing in personalization methods that explicitly take into account the diversity of content recommended.

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 recommender 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 Dhillon 2016).[1] 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 engage with organically.

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 can impact 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

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

cases where diversity is measured with respect to the ideological slant or extremity of content. If so, the "engagement-diversity tradeoff" 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).

Our work contributes to an emerging research 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). Notably, our results indicate that recommender systems can decrease individual-level diversity while increasing aggregate diversity, while Lee and Hosanagar (2019) 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 is due to differences in how the two studies quantify diversity, as well as differences in the recommendation algorithms being used and the research settings in which the two experiments are conducted. These conflicting results highlight the importance of measuring the impact of many different recommendation algorithms in a variety of settings, and the need to develop a number of different methodological approaches for measuring the impact of recommendation systems on consumption diversity.

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 as implicit and explicit user feedback to generate personalized product recommendations to users (Resnick and Varian 1997, Adomavicius and Tuzhilin 2005). 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). 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.

Three recent papers closely related to our research are Claussen et al. (2019), Lee and Hosanagar (2019) and Anderson et al. (2020). Both Claussen et al. (2019) and Lee and Hosanagar (2019) conduct online field experiments that measure the effect of personalized recommendations on consumption diversity. Claussen et al. (2019) find that personalization *decreases* individual-level diversity, and that this decrease in consumption diversity spills over to non-personalized sections of the website they study. On the other hand, 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. Anderson et al. (2020) use observational data from Spotify to study the relationship between personalization and listening diversity. They find that user-driven listening is more diverse than algorithmic listening, and that users who become more diverse over time do so by shifting away from algorithmic listening. Importantly, they also find that users with more diverse listening habits are less likely to leave the platform and are more likely to eventually become paid subscribers.

Our research builds on the existing literature in multiple ways. First, in contrast to many observational studies of recommender systems, we analyze data from a randomized field experiment, which enables us to credibly estimate the causal effect of personalized recommendations on content consumption. Second, whereas most recommender systems research in economics and management 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. Finally, 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). Given that 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, it is important for researchers to continue studying the impact of many different recommendation algorithms in a number of different settings.

3. Research Setting

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, has 271 million monthly active users and 124 million paying subscribers.[2] Although Spotify launched as a music streaming platform, in 2015 the company expanded its offerings to include videos and, more importantly for this study, podcasts.[3] 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 "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 on home, and the cards within each shelf, are ordered by a combination of machine learning algorithms and human editors.[4] A screenshot of the "Home" section of the Spotify app on iOS can be seen in Figure 1.

In this paper, we will analyze changes to the number of podcasts users stream, as well as the *types* of podcasts users stream. Each podcast stream 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 on 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).

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

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

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

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." 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 podcasts associated with them. 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. A more detailed description of podcast categorization on Spotify can be found in Appendix A.

4. Experiment Design

We analyze data from an experiment conducted on a sample of 852,937 premium Spotify users across seventeen countries[5] 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 were assigned to treatment arms using a "bucket randomization" procedure. That randomization procedure, along with the balance of observable characteristics between the treatment and control groups, is described in Appendix B.

Users in both the treatment and control were exposed to a shelf in the "Home" 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.[6] 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.[7] 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

[5] 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.

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

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

podcast on Spotify. Figure 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.

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 C.

5.1. Effect on podcast consumption

We first study the impact of algorithmic podcast recommendations on the average number of podcast streams per user during the experiment.[8][9] We estimate the effect of the treatment by estimating the following model:

$$y_i = \alpha + \beta T_i + \delta X_i + \epsilon_i, \tag{1}$$

where y_i is the outcome of interest for user i (in this case, podcast streams), 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 2 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 reports the estimated effect of the treatment on podcast streams per user during the experiment, both with and without controlling for user-level covariates. We find that the treatment increased the number of podcast streams per user by 28.90% ($\pm$ 3.81%). This large treatment effect indicates that personalized podcast recommendations were extremely effective at increasing podcast consumption during the experiment.

Using the principal stratification approach detailed by Frangakis and Rubin (2002) and Ding and Lu (2017), we are also able to measure the extent to which this treatment effect is driven by compositional shifts, as opposed to intensity shifts.[10] We estimate that on average, "compliers" (i.e., those who would only stream at least one podcast if exposed to the treatment) streamed 1.505 (95% CI: (1.488, 1.522)) more podcasts in the treatment, whereas "always takers" (i.e., those who would stream at least one podcast whether in the control or treatment) streamed 0.082 (95% CI:

[8] The long-term effects of the experiment are reported in Appendix D.

[9] The effect of the treatment on the number of users streaming at least one podcast is reported in Appendix E.

[10] The principal stratification methodology is detailed in Appendix F.

(0.055, 0.112)) *fewer* podcasts in the treatment. In other words, the increase in podcast streaming is driven by a greater number of users streaming *at least one* podcast during the experiment, as opposed to an increase in the amount of podcast streaming from those who would have streamed at least one podcast even if they had not been exposed to the treatment.

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").

5.2.1. 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

$$th_i = -\sum_{c \in C} s_{ci} \ln(s_{ci}), \tag{2}$$

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.

To account for this, we conduct our analysis on the subset of users that streamed at least one podcast during the experiment.[12] Results of this analysis cannot be interpreted as causal user-level effects, 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 3 shows the histogram of user-level Shannon entropy for podcast streams in both treatment arms, and Table 2 reports the difference in the average streaming user's Shannon entropy, both with and without controlling for user-level covariates.[13] 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. Although this result is non-causal,

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

[12] Analysis on the full sample can be found in Appendix G. As expected, the measured effect of the treatment is positive.

[13] Figure I.1 shows the distributions of user-level Shannon entropy in both treatment arms conditional on streaming a particular number of podcasts during the experiment.

we can use the principal stratification approach to estimate the causal effect of the treatment on individual-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 0.070 (95% CI: (0.062, 0.076)).

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.

5.2.2. Intragroup diversity We quantify intragroup diversity using a mathematical expression introduced by Aral and Dhillon (2016):

$$ID = \frac{1}{n_c} \sum_{j=1}^{n_c} \left[1 - \cos\left(\Gamma_j, \bar{\Gamma}\right)\right]^2, \tag{3}$$

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.

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)).[14] These results indicate that the treatment had a causal effect on the variance of podcast streamers' individual-level category consumption vectors. 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.

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 affects the podcasts that are displayed on "Home," stream referrer-level treatment effects provide insight into the extent to

[14] 95% confidence intervals calculated with the cluster bootstrap ($n_{boot} = 1,000$).

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.

5.3.1. Podcast consumption Figure 4 shows the distribution of podcast streams per user on both types of referral surfaces in both treatment arms, and Table 3 reports the estimated effect of the treatment on podcast streams per user from either type of referral surface during the experiment, both with and without controlling for user-level covariates. We find that the treatment increased the average number of podcast streams per user from home by 55.75% (±5.07%), and the average number of podcast streams per user from non-home surfaces by 10.47% (±4.16%).

5.3.2. Individual-level diversity Figure 5 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 4 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 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.

5.3.3. 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.

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 tradeoff" 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 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-run 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 does 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.

While Lee and Hosanagar (2019) find that recommender systems have a neutral-to-positive effect on individual-level diversity and decrease aggregate diversity, we 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.[15] Given that recommender systems have become a common feature of content platforms, we believe

[15] In Appendix H, we report the effect of the experiment on the "sales diversity" for podcast consumption.

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).

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 tradeoff" 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 consume, 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 methods for measuring 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.

8. Figures

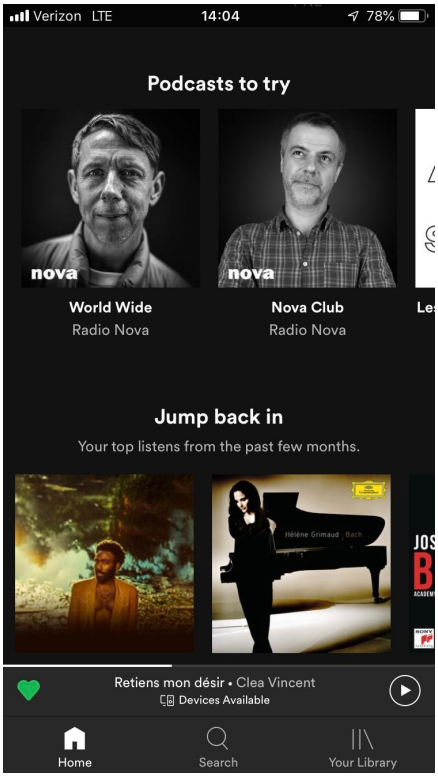

Figure 1 **A screenshot of the "Podcasts to try" shelf on the Spotify iOS app. During the experiment, this shelf was fixed in the second slot on the "Home" section of the Spotify app.**

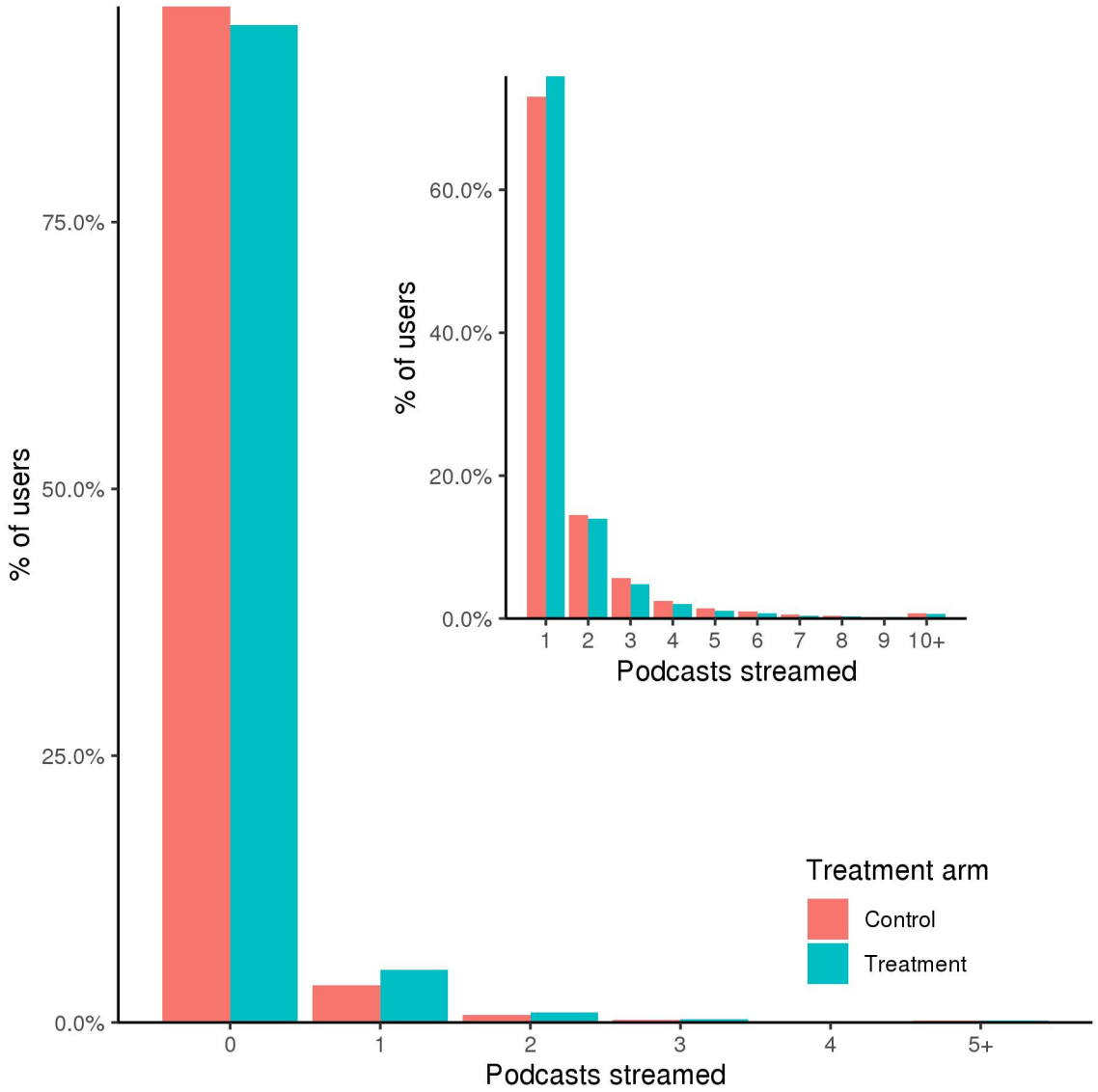

Figure 2 **The distribution of of podcasts streamed in both treatment arms. Inset plot shows the distribution of podcasts streamed in both treatment arms conditional on streaming at least one podcast.**

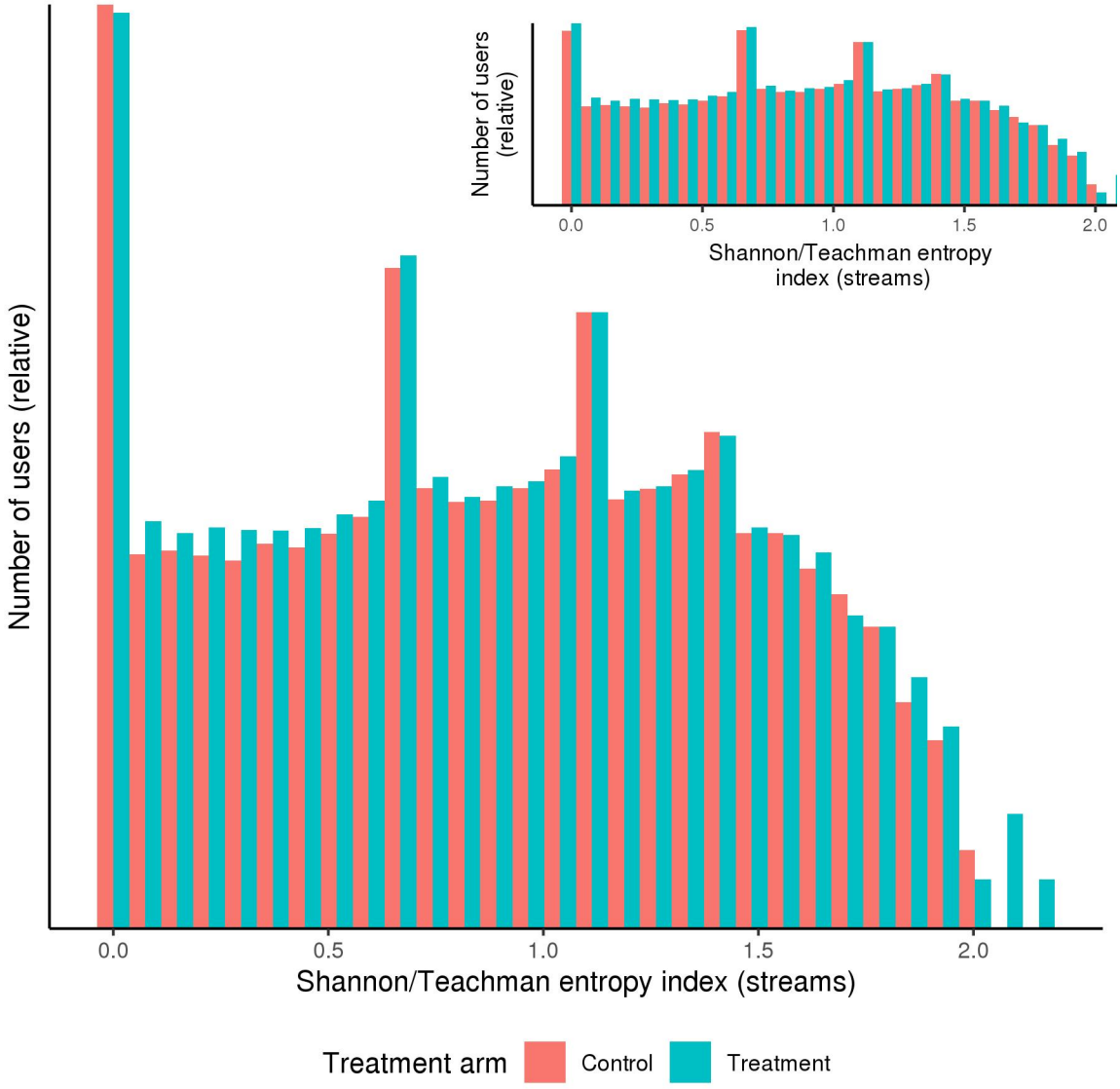

Figure 3 **The distribution of the user-level diversity for streams in both treatment arms. Inset plot shows the distribution of user-level diversity in both treatment arms conditional on streaming at least one podcast. y-axis values are on a log scale, and are hidden due to confidentiality concerns.**

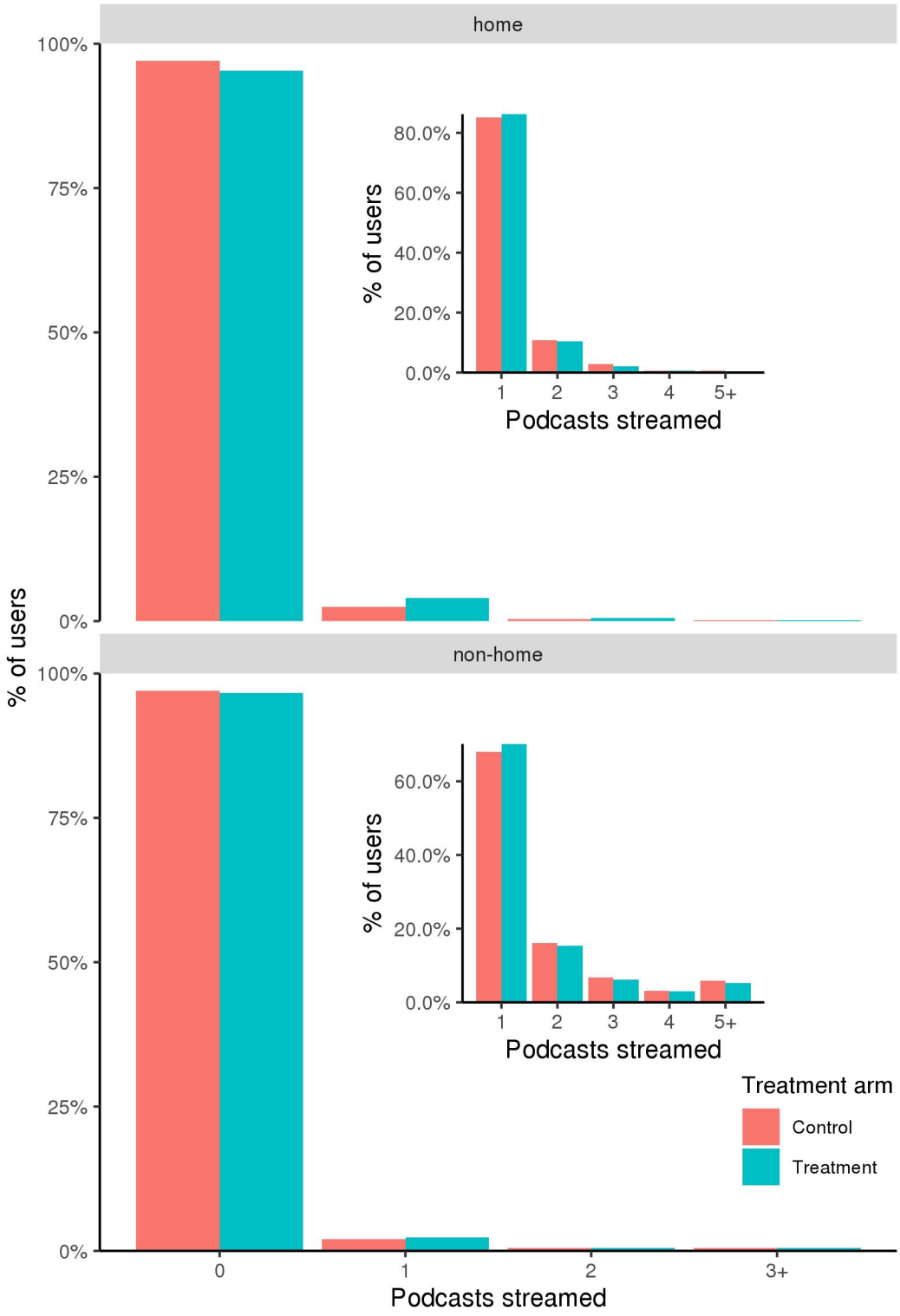

Figure 4 **The distribution of podcasts streamed in both treatment arms by stream referrer. Inset plots shows the distribution of podcasts streamed by referrer in both treatment arms conditional on streaming at least one podcast.**

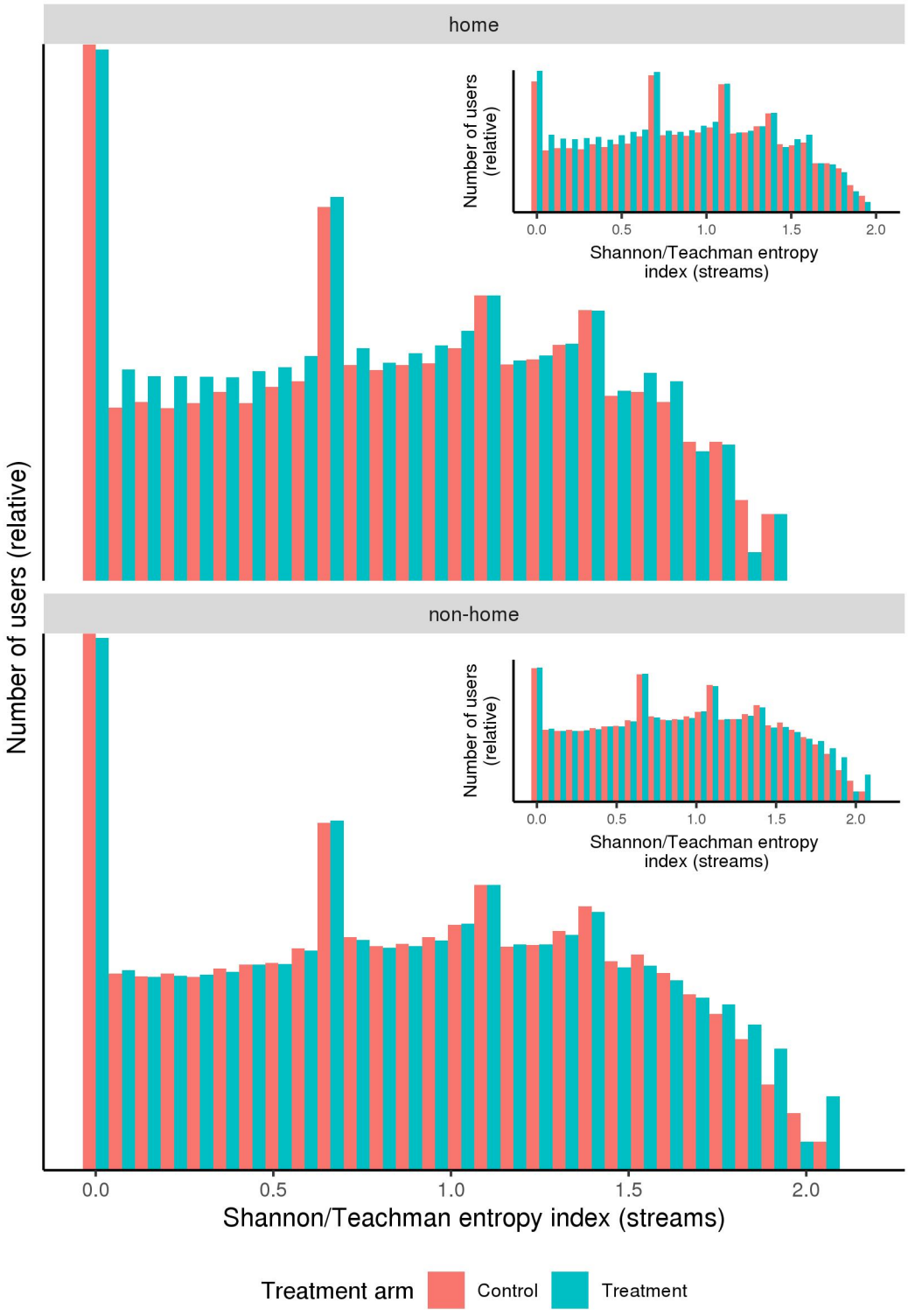

Figure 5 **The distribution of individual-level diversity for podcast streams in both treatment arms by stream referrer. Inset plots show the distribution of individual-level diversity by referrer in both treatment arms conditional on streaming at least one podcast. y-axis values are on a log scale, and are hidden due to confidentiality concerns.**

9. Tables

Table 1 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

Table 2 A linear model showing the difference in the average 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	0.549***	0.600***
	(0.003)	(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

Table 3 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

Table 4 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

References

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.

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

Aral S, Dhillon P (2016) Unpacking novelty: The anatomy of vision advantages. *Available at SSRN 2388254* .

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

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).

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)* .

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).

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

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.

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

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.

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.

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.

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.

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.

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).

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.

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* .

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).

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.

Nazari Z, Charbuillet C, Pages J, Laurent M, Charrier D, Vecchione B, Carterette B (Forthcoming) Recommending podcasts for cold-start users based on music listening and taste.

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

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* 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).

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* .

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

Senecal S, Nantel J (2004) The influence of online product recommendations on consumersâ 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.

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, youâre sure to love that. *The New York Times* 21.

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

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.

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).

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).

Appendix A: Podcast Categorization

In this section, we provided a more detailed description of how podcasts are categorized on Spotify. In our dataset, there are as many as ten podcast category tags associated with any particular podcast. 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". 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.

Figure I.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. Figure I.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.[16] 68.23% of podcasts have only one category associated with them, and 97.50% of podcasts have three or fewer categories associated with them. 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.

As mentioned in the main text, when a podcast has multiple category tags associated with it, we divide that podcast's streams evenly across all of the categories with which it is associated. For example, 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.

Appendix B: Bucket Randomization Procedure & Balance Checks

In this section, we describe the randomization procedure used to assign users to the treatment and control arms of the experiment we analyze, and report the balance of different user-level covariates across the two treatment conditions.

Spotify users were assigned to treatment arms using the following procedure: every Spotify user is 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 receives 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 results 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

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

this, we report either cluster-robust standard errors or cluster bootstrap standard errors for all experiment analyses.

Table J.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.

Appendix C: 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.

C.1. Effect on podcast follows

Figure I.4 shows the distribution of podcast follows per user during the experiment in both treatment arms, and Table J.2 reports the estimated effect of the treatment on podcast follows per user during the duration of the experiment, both with and without controlling for user-level covariates. We find that the treatment increased the number of podcast follows per user by 51.38% ($\pm$ 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 J.3 reports the estimated effect of the treatment on following at least one podcast during the experiment, both with and without controlling for user-level covariates. We find that the treatment increased the number of Spotify users following at least one podcast by 53.45% ($\pm$ 5.23%).

C.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.

C.2.1. Individual-level diversity Figure I.5 shows the histogram of the user-level Shannon entropy for podcast follows in both treatment arms, and Table J.4 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% ($\pm$ 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 J.5 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 I.6 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.

C.2.2. 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))

C.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 I.7 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 I.8 shows the long-term effect of the treatment on the number of users following at least one podcast. Figure I.9 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.

C.4. Effect on "sales diversity"

Figure I.10 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 4 with follow counts and follow rank, as opposed to stream counts and stream rank. Figure I.11 shows the relationship between $\ln(Follows_i + 1)$ and $\ln(Follows\,Rank_i)$ across all podcasts appearing in our dataset, and Table J.6 shows

the results of estimating Equation 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.

Appendix D: 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 their counterfactual baseline podcast listening once individually personalized recommendations are no longer shown.

Figure I.12 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 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.

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 I.13 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 changes in what users seek out organically.

It is worth noting that the number of podcast streams per user over time is 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 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 I.14 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.

Appendix E: Effect of the treatment on distinct podcast streamers

In this section, we report the effect of the treatment on the number of users streaming at least one podcast during the experiment.

Table J.7 reports the estimated effect of the treatment, both with and without controlling for user-level covariates. We find that the treatment increased the number of Spotify users streaming podcasts by 36.33% (±3.01%). Table J.8 reports the estimated effect of the treatment on streaming at least one podcast during the experiment from either type of referral surface, both with and without controlling for user-level covariates. We find that the treatment increased the number of Spotify users streaming podcasts from "Home" by 59.17% (±4.58%), and the number of users streaming podcasts from non-home surfaces by 12.55% (±2.94%).

Figures I.15 and I.16 show the long-term effect of the treatment on the number of users streaming at least one podcast, both overall and conditional on streaming surface. We find that the large treatment effects we observe during the experiment rapidly dissipate once the experiment has concluded.

Appendix F: 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 measuring 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.[17] 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 I.17 and the results for podcast following are shown in Figure I.18. In both cases, the estimated effects are nearly zero across all strata and covariates, indicating that the balancing conditions are satisfied.

Appendix G: Effect of the treatment on individual-level diversity (all users)

In this section, we report the effect of the treatment on the individual-level diversity when including all users in our analysis, regardless of whether or not they streamed any podcasts during the experiment. Table J.9 reports the estimated effect of the treatment on the Shannon entropy, both with and without controlling for user-level covariates. We find that the treatment increased the average Shannon entropy for podcast streams by 21.16% (± 2.89%). Table J.10 reports the estimated effect of the treatment on the referrer-specific user-level Shannon entropy, both with and without controlling for user-level covariates. 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%).

Figures I.19 and I.20 show the long-term effect of the treatment on the individual-level diversity of podcast streams across all users in the experiment, both overall and conditional on streaming surface. We find that the large treatment effects we observe during the experiment rapidly dissipate once the experiment has concluded.

Appendix H: 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 I.21 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.

[17] 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.

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

$$\begin{aligned}\ln(Streams_i + 1) =& \beta_0 + \beta_1 \ln(Streams\,Rank_i) + \beta_2 Treatment_i + \\ & \beta_3 Treatment_i \times \ln(Streams\,Rank_i) + \epsilon_i,\end{aligned} \tag{4}$$

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

Figure I.22 shows the relationship between $\ln(Streams_i + 1)$ and $\ln(Streams\,Rank_i)$ across all podcasts appearing in our dataset, and Table J.11 shows the results of estimating Equation 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} = 1,000). The positive estimate for β_3 also indicates that the treatment *increases* sales diversity.

Appendix I: Additional Figures

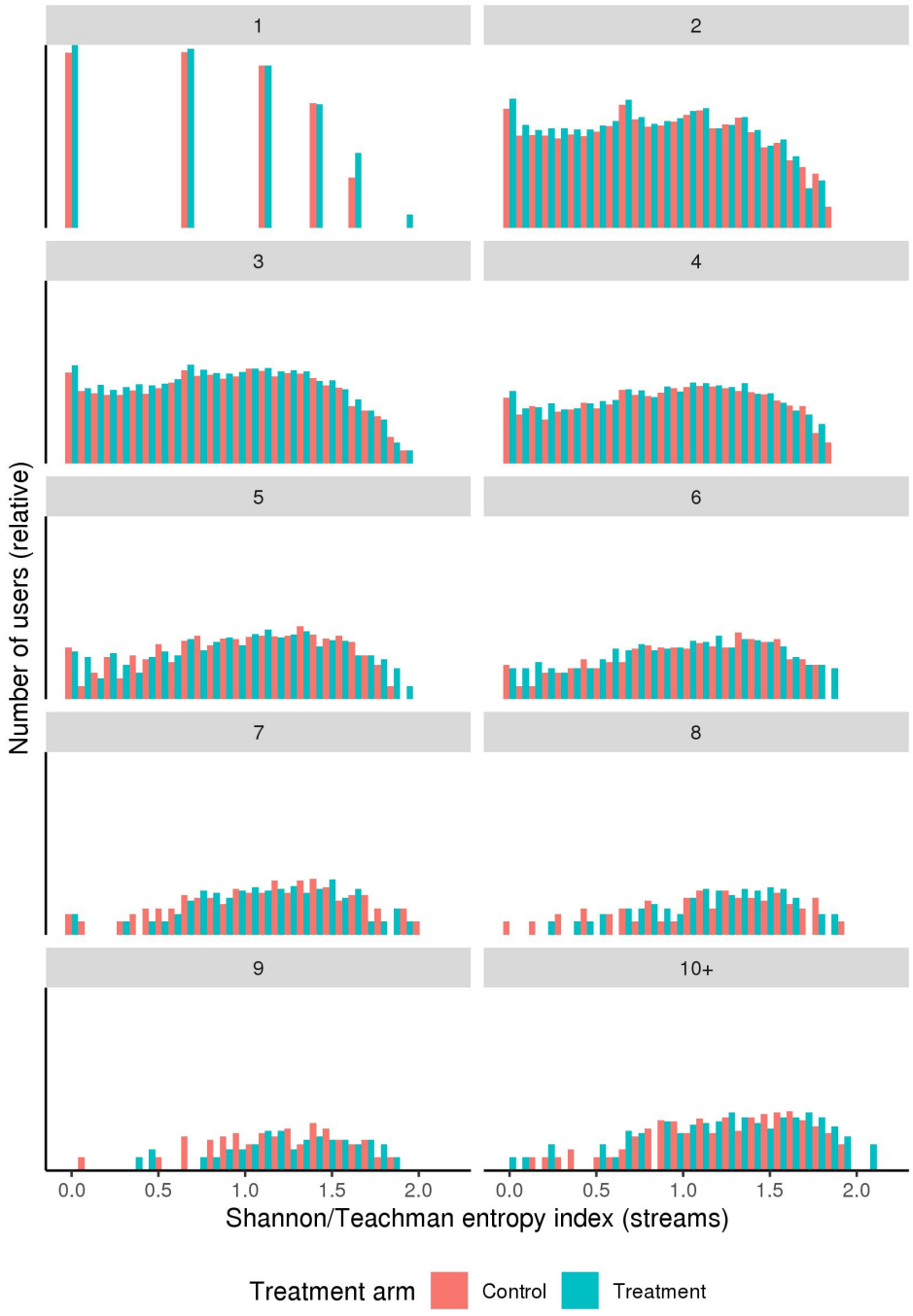

Figure I.1 The distribution of the user-level diversity for streams in both treatment arms conditional on streaming a set number of podcasts during the experiment. y-axis values are on a log scale, and are hidden due to confidentiality concerns.

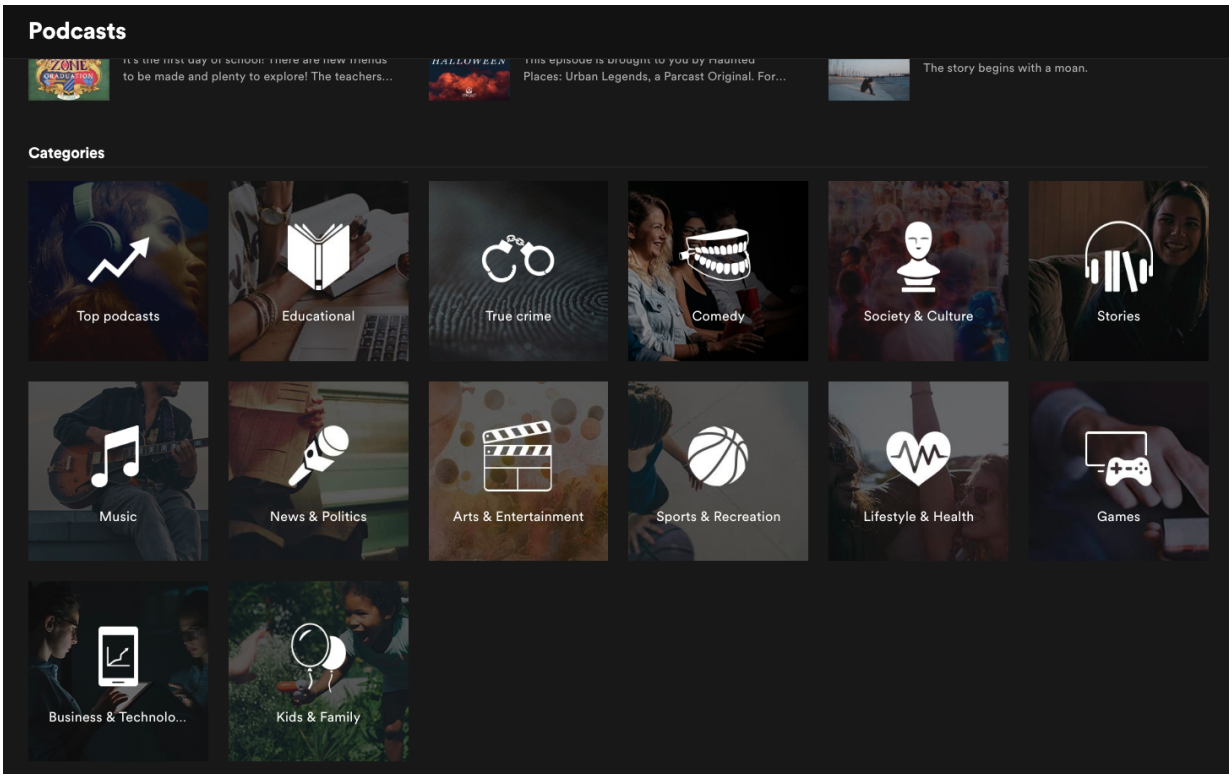

Figure I.2 Podcast categories on Spotify, as displayed in the Podcasts section of the desktop app's "Browse" pane.

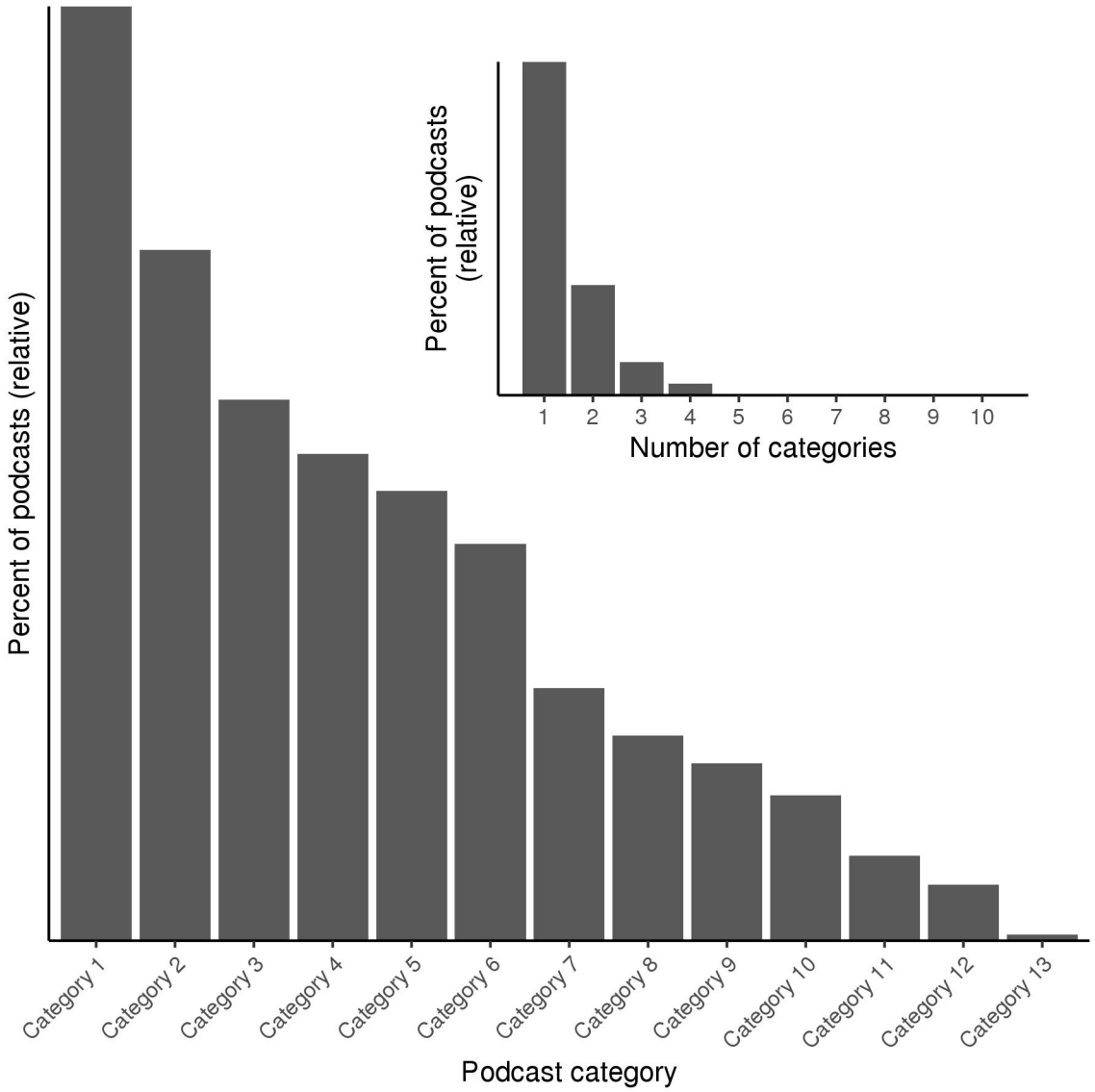

Figure I.3 Histograms showing the frequency with which each podcast category is attached to a podcast, and the distribution of category tags per podcast. y-axis values and category names hidden due to confidentiality concerns.

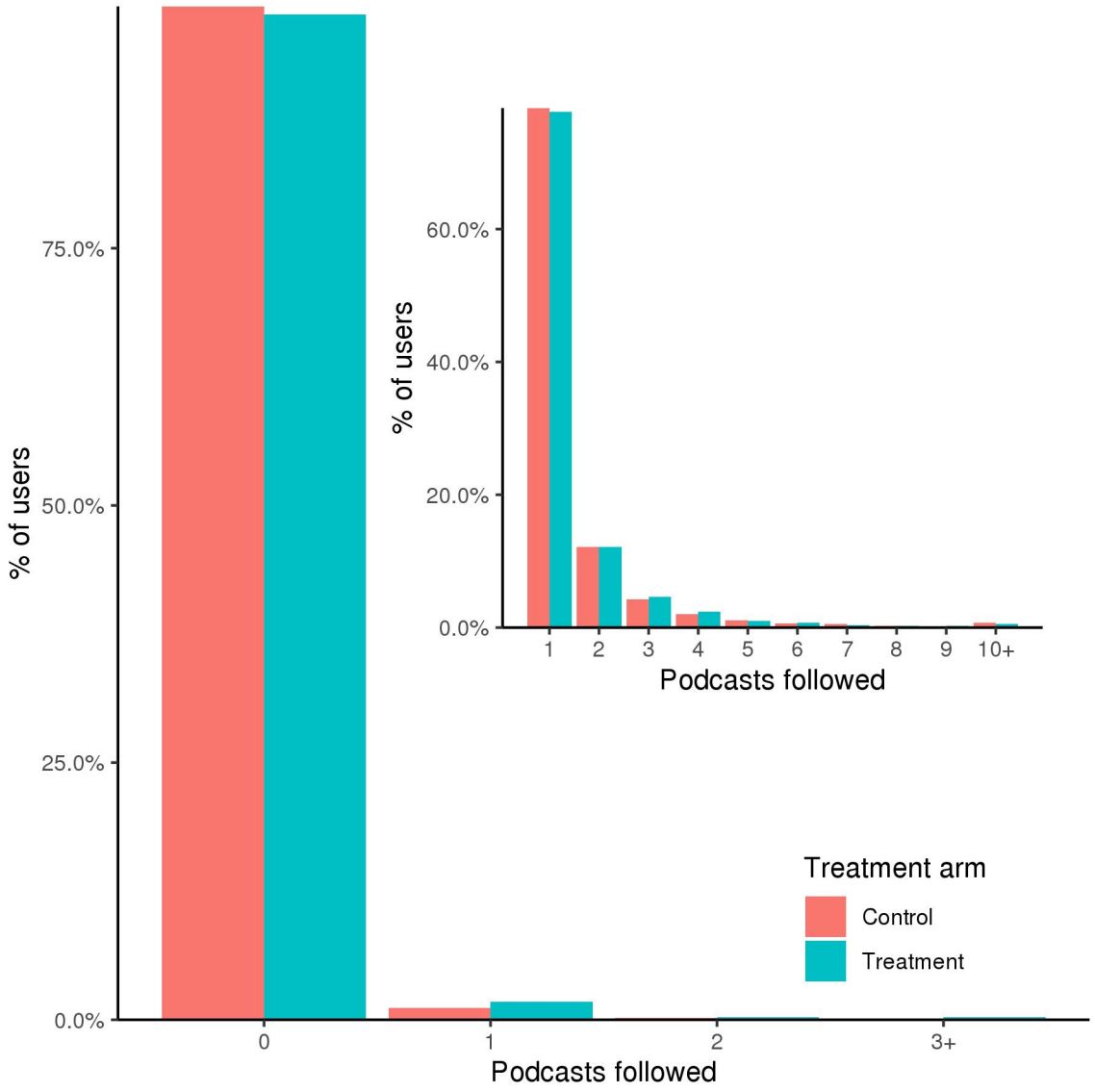

Figure I.4 **The distribution of podcasts followed in both treatment arms. Inset plot shows the distribution of podcasts followed in both treatment arms conditional on following at least one podcast.**

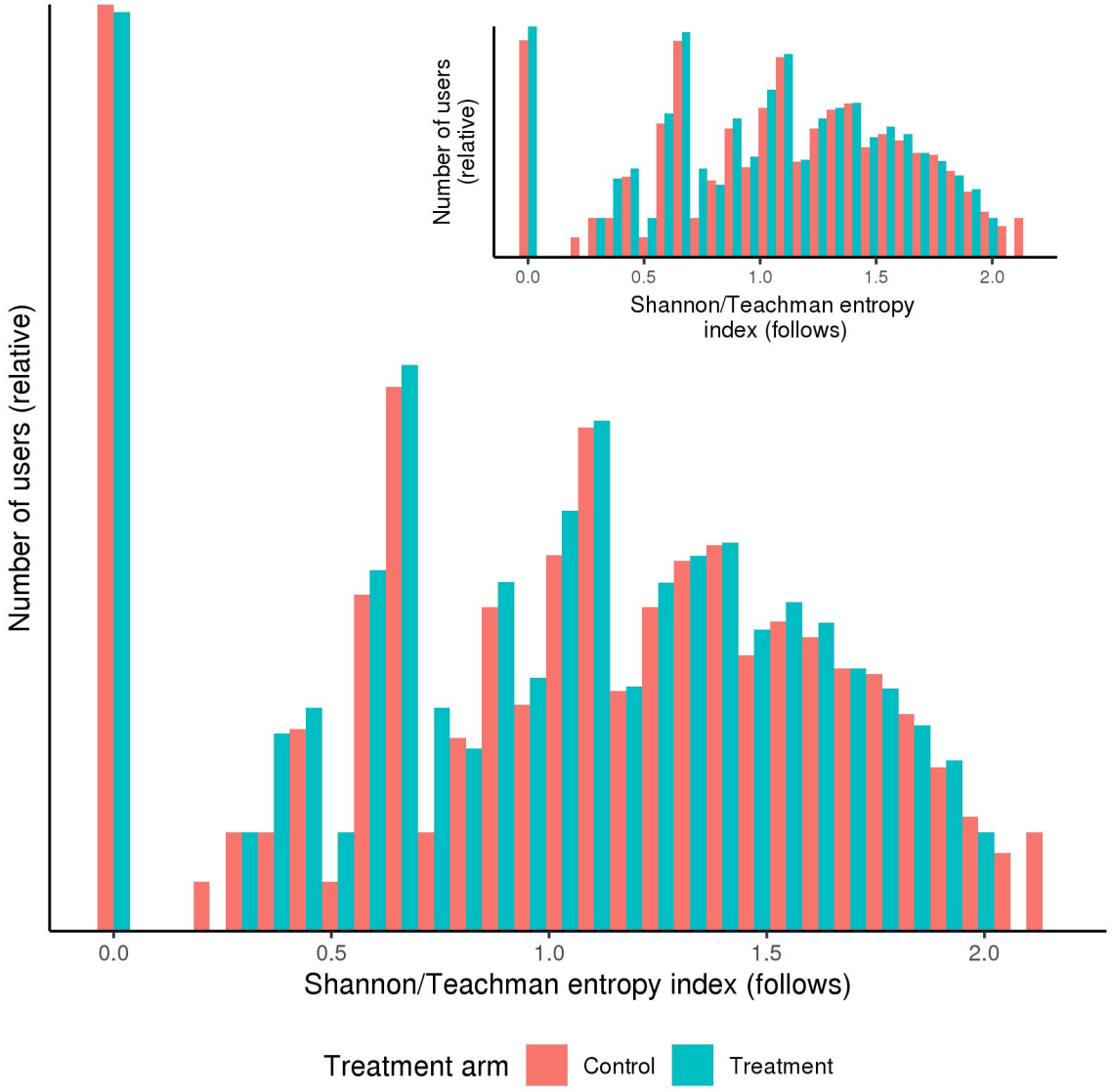

Figure I.5 **The distribution of the user-level diversity for follows in both treatment arms. Inset plot shows the distribution of user-level diversity in both treatment arms conditional on following at least one podcast. y-axis values are on a log scale, and are hidden due to confidentiality concerns.**

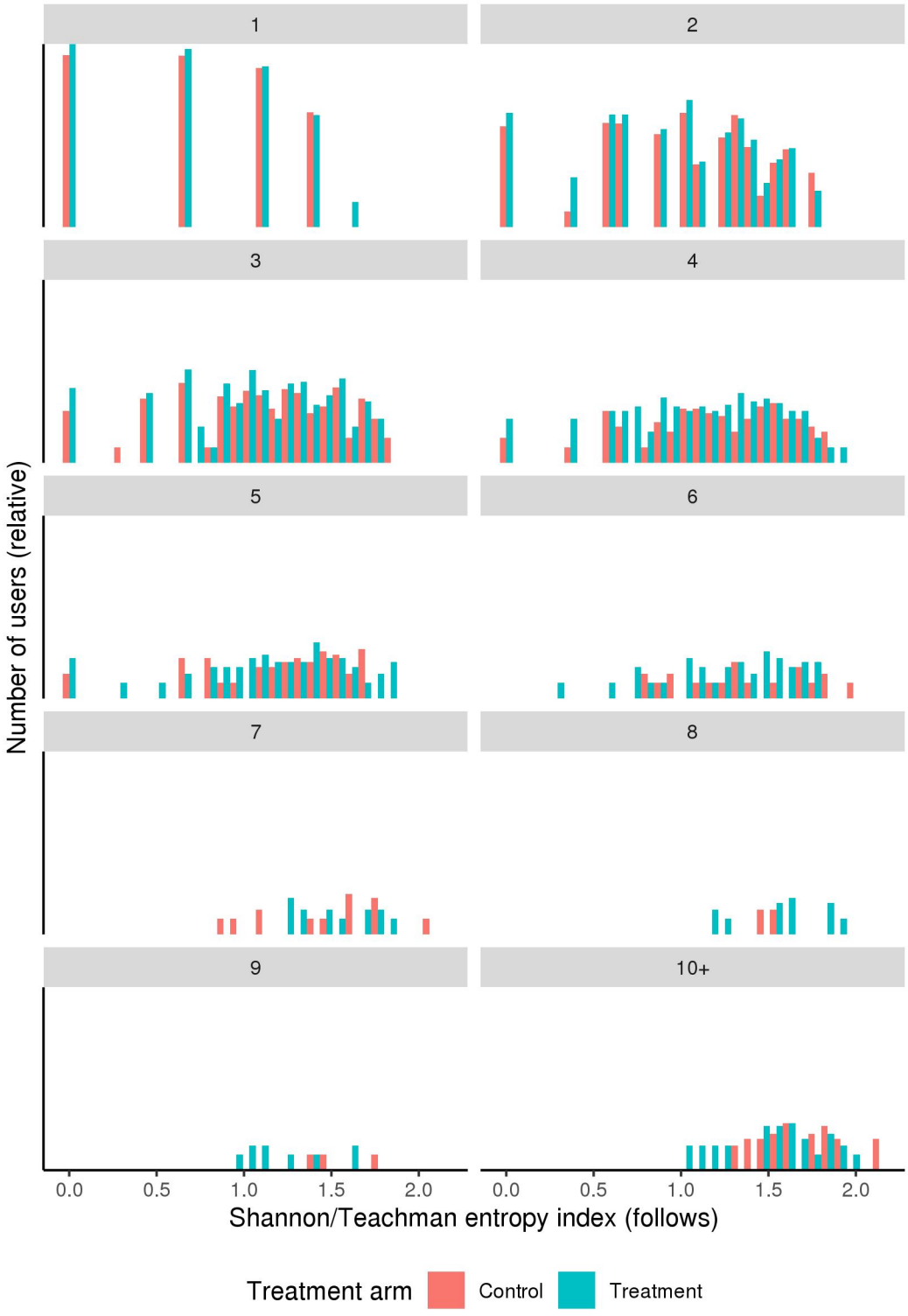

Figure I.6 **The distribution of the user-level diversity for follows in both treatment arms conditional on following a set number of podcasts during the experiment. y-axis values are on a log scale, and are hidden due to confidentiality concerns.**

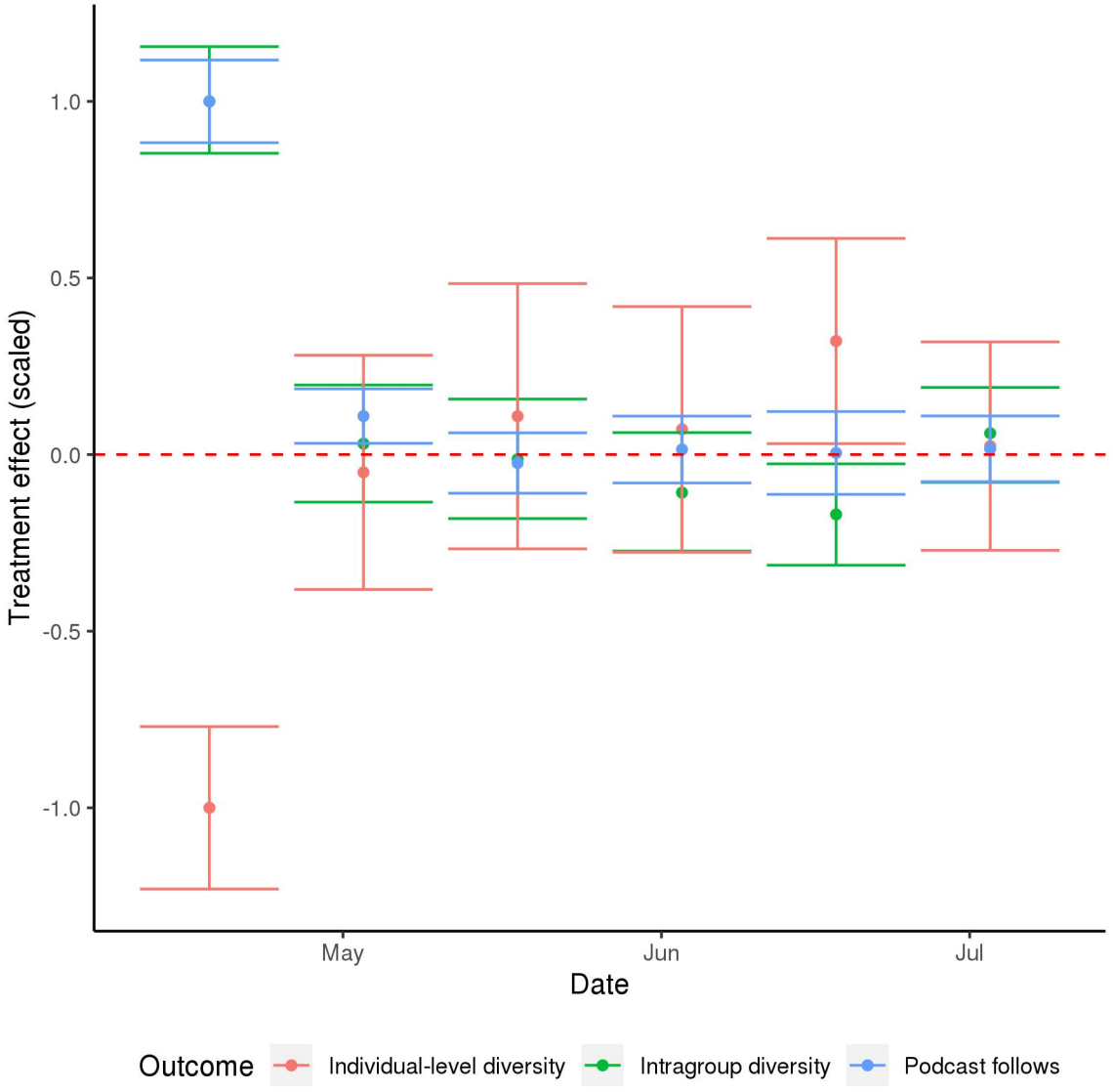

Figure I.7 **The long-term effect of the treatment on podcast follows per user, individual-level following diversity conditional on following at least one podcast, and intragroup following diversity. Each outcome's time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.**

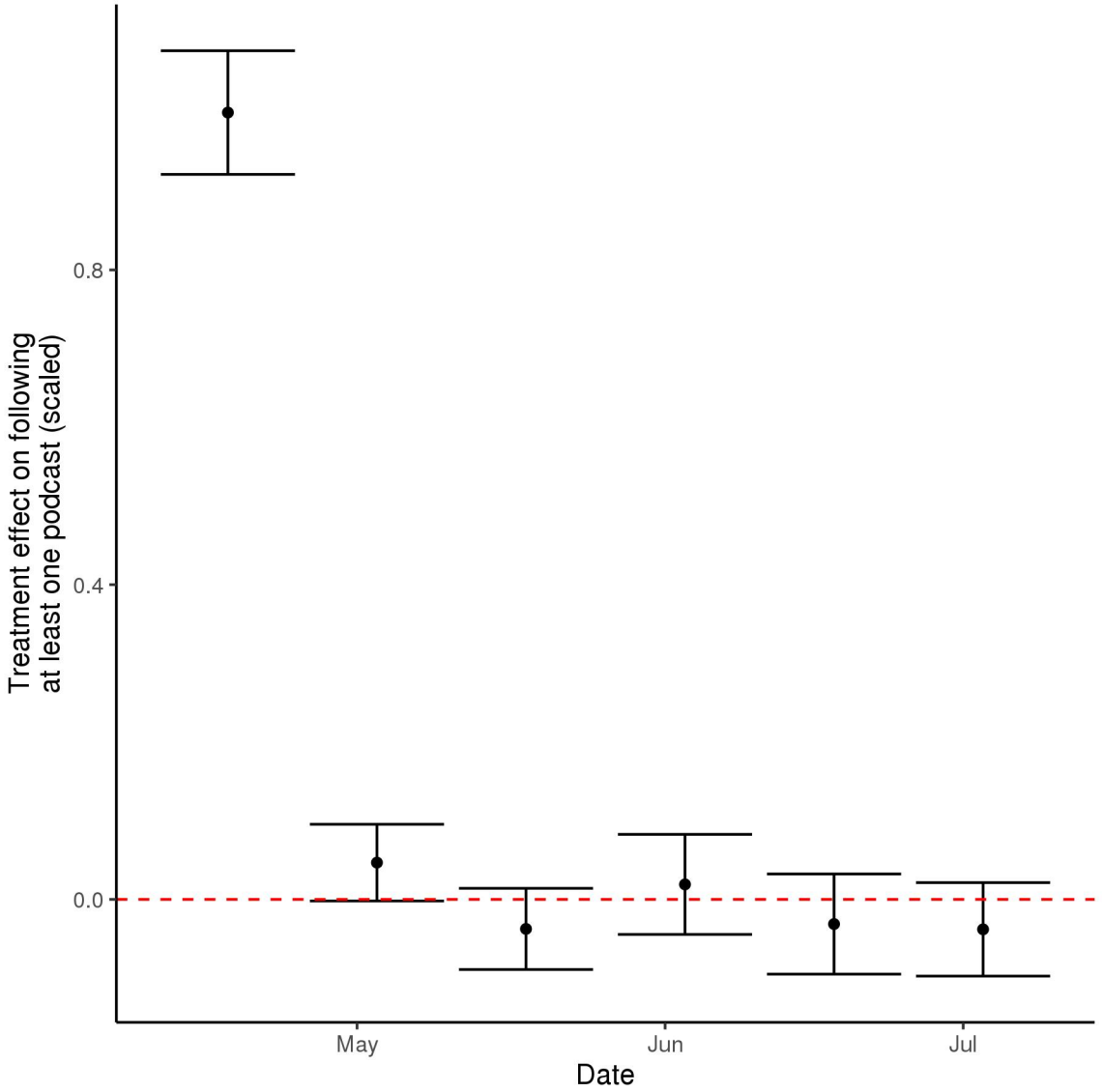

Figure I.8 **The long-term effect of the treatment on the percentage of users following at least one podcast. The time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.**

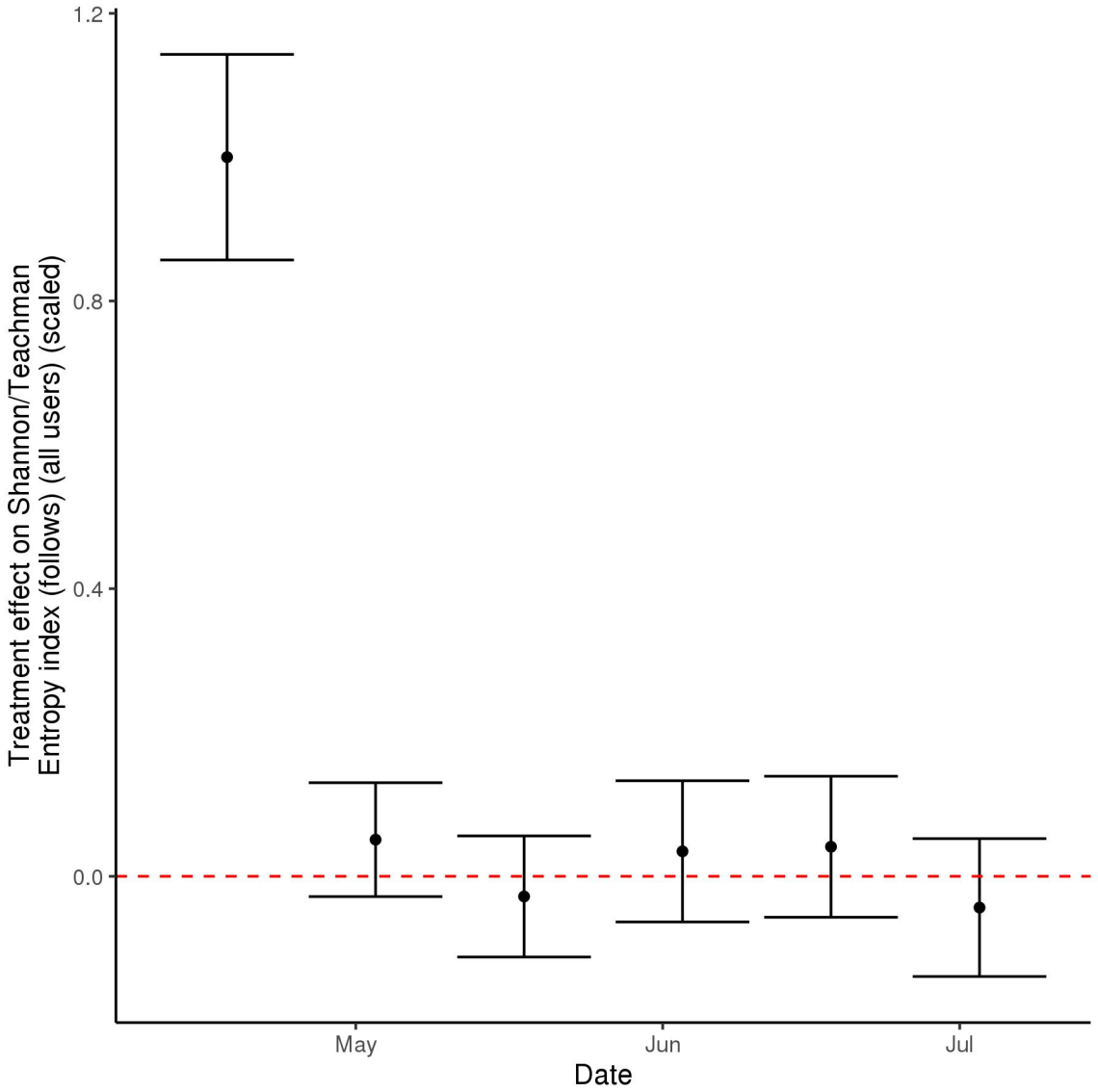

Figure I.9 **The long-term effect of the treatment on the average user-level Shannon entropy for follows. The time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.**

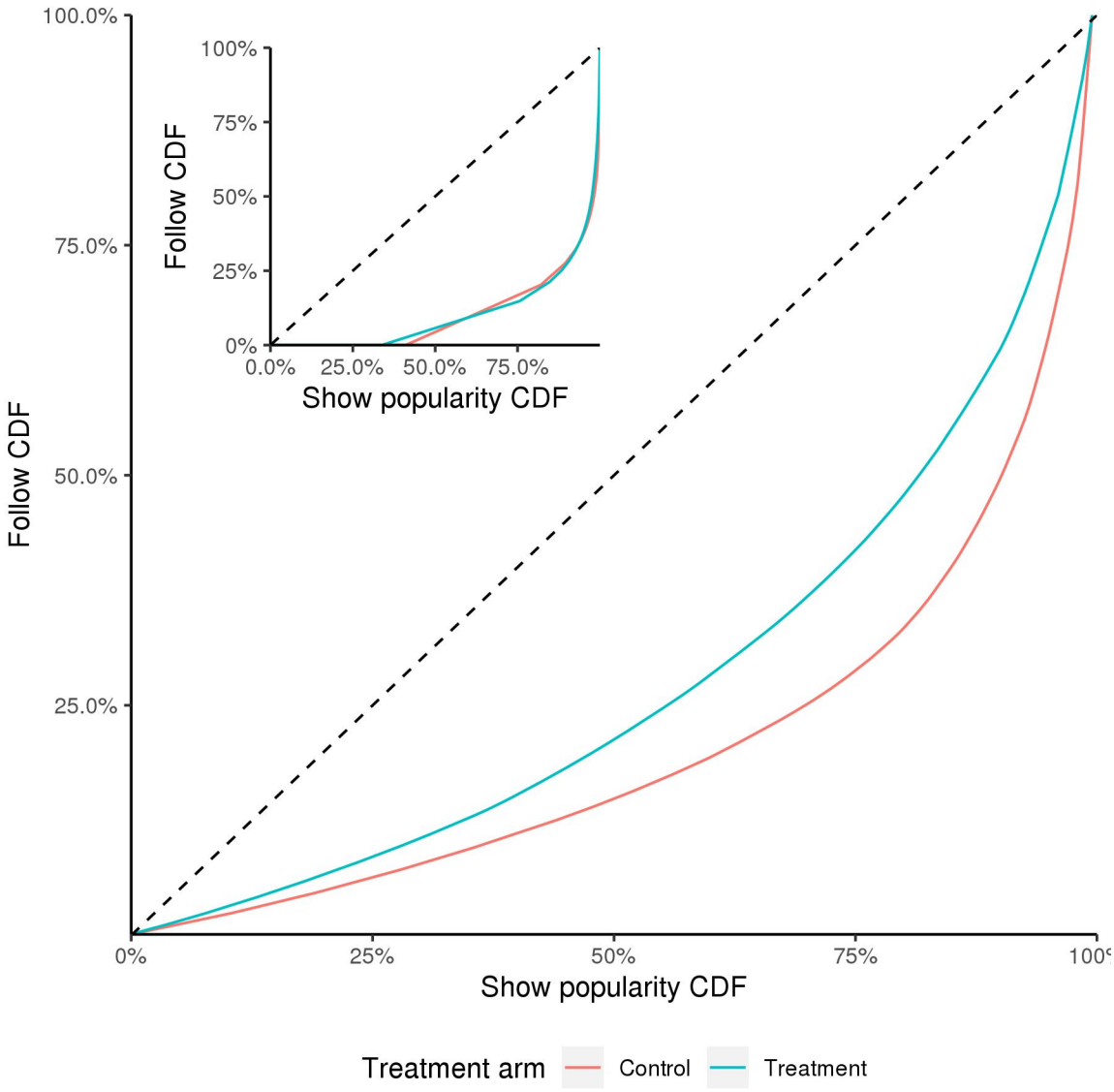

Figure I.10 **The Lorenz curves for podcast follows, calculated separately for users in the treatment and control. The data for each Lorenz curve is limited to the 200 most followed podcasts in the corresponding treatment arm data. The inset curve shows the Lorenz curve for follows across all podcasts.**

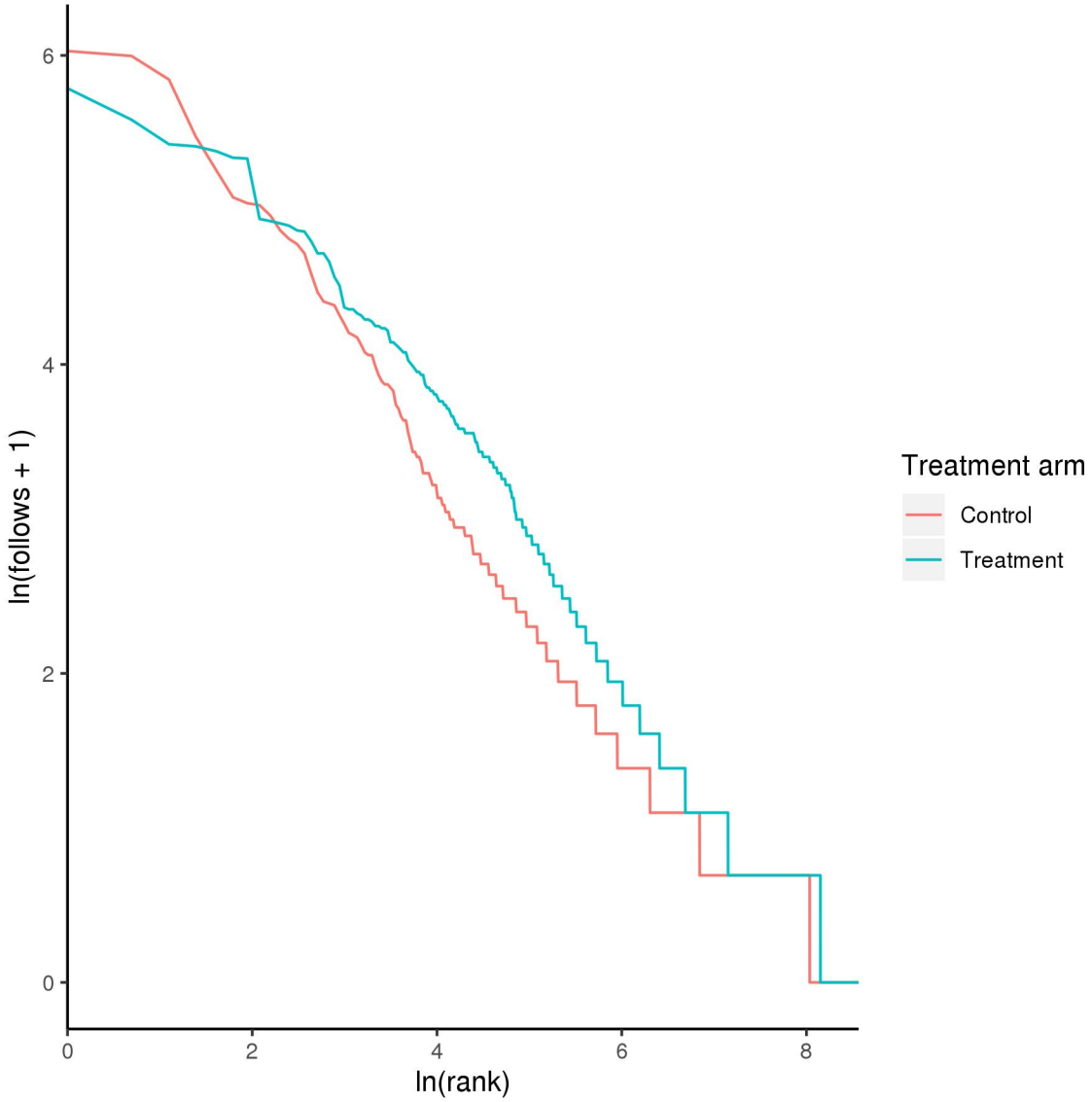

Figure I.11 **The relationship between ln(follows + 1) and ln(follow rank) for both the control and treatment arms of the experiment.**

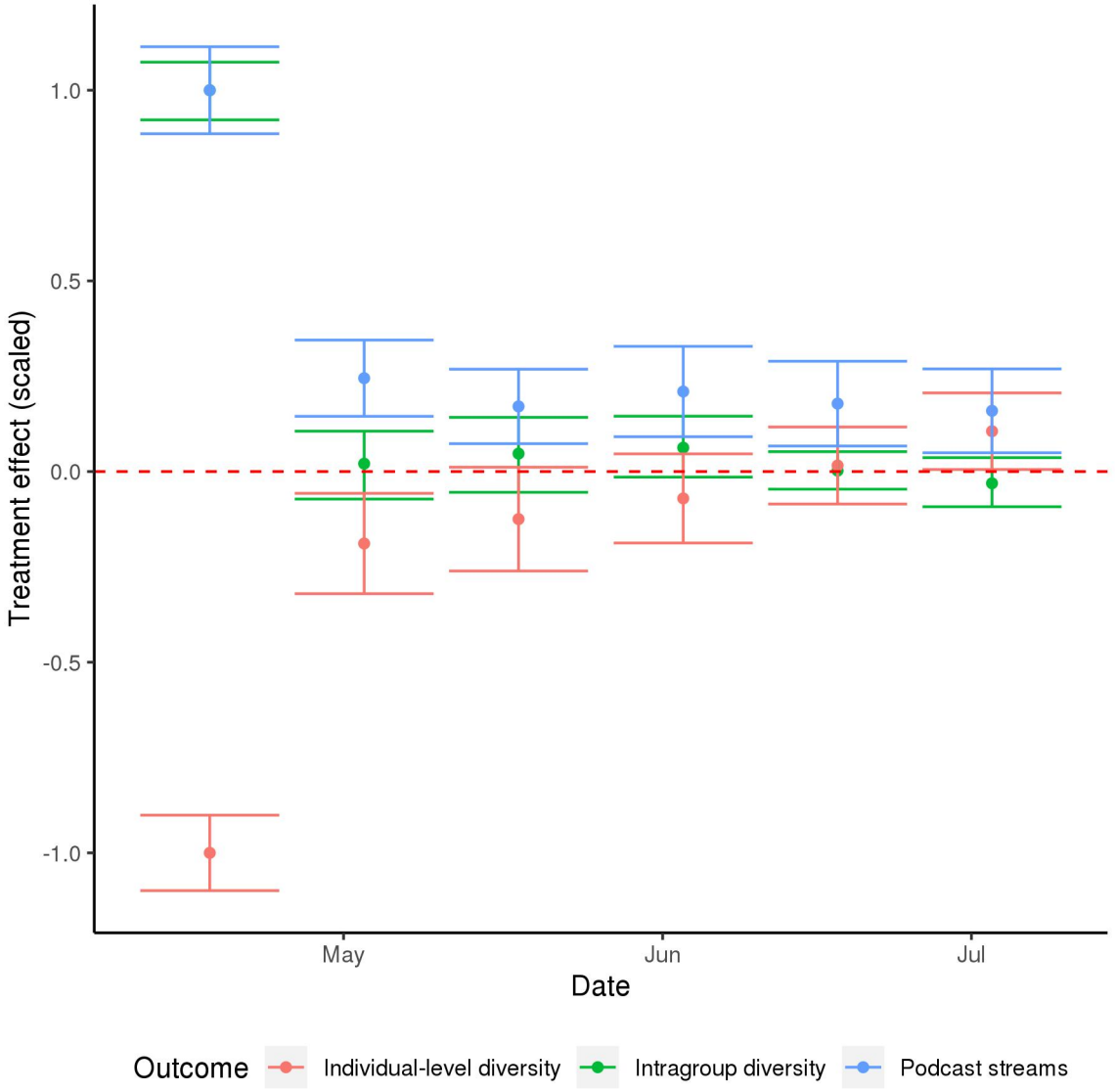

Figure I.12 **The long-term effect of the treatment on podcast streams per user, individual-level streaming diversity conditional on streaming at least one podcast, and intragroup streaming diversity. Each outcome's time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.**

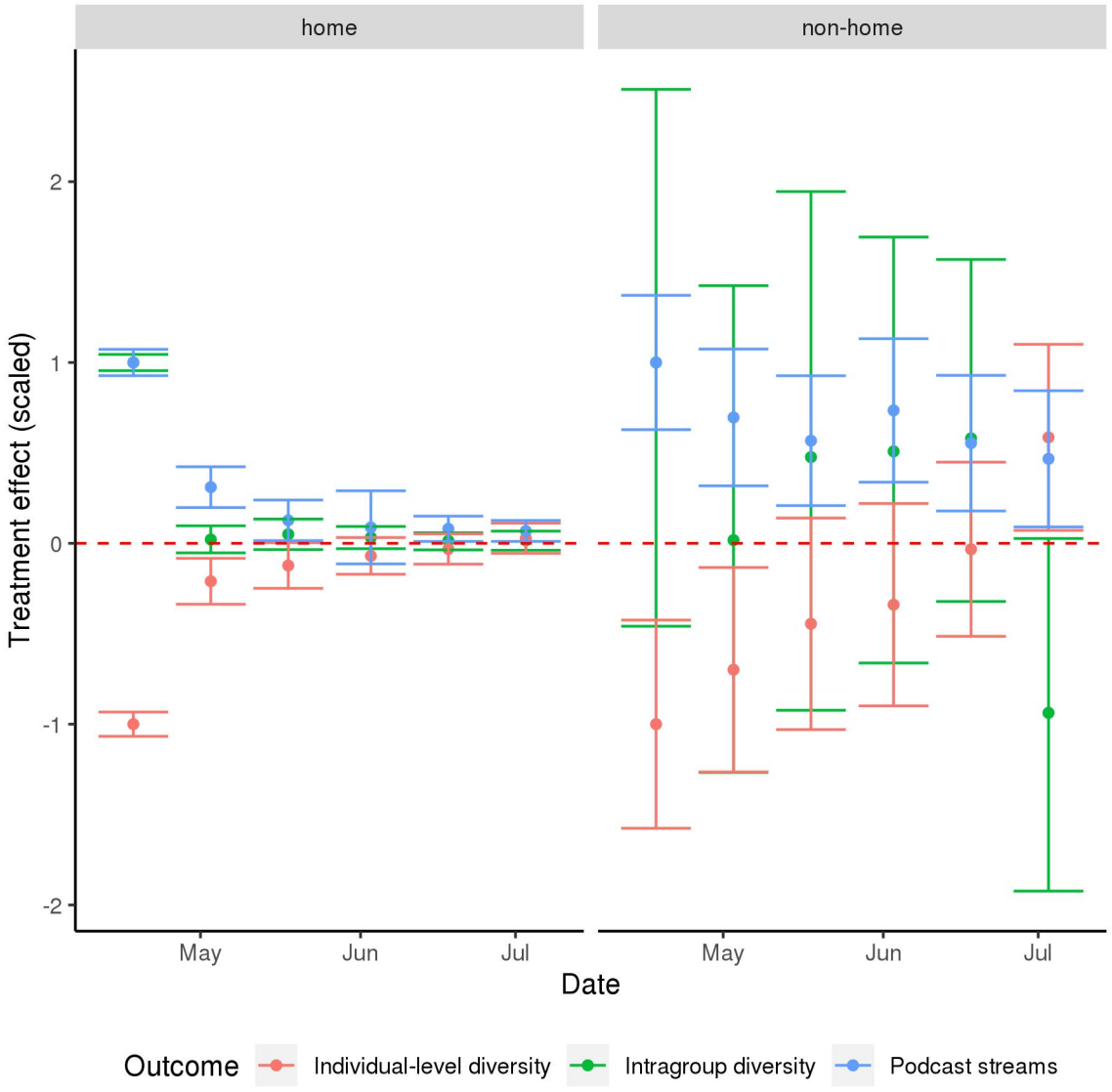

Figure I.13 **The long-term referrer-level effect of the treatment on podcast streams per user, individual-level streaming diversity conditional on streaming at least one podcast, and intragroup streaming diversity. Each outcome's time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.**

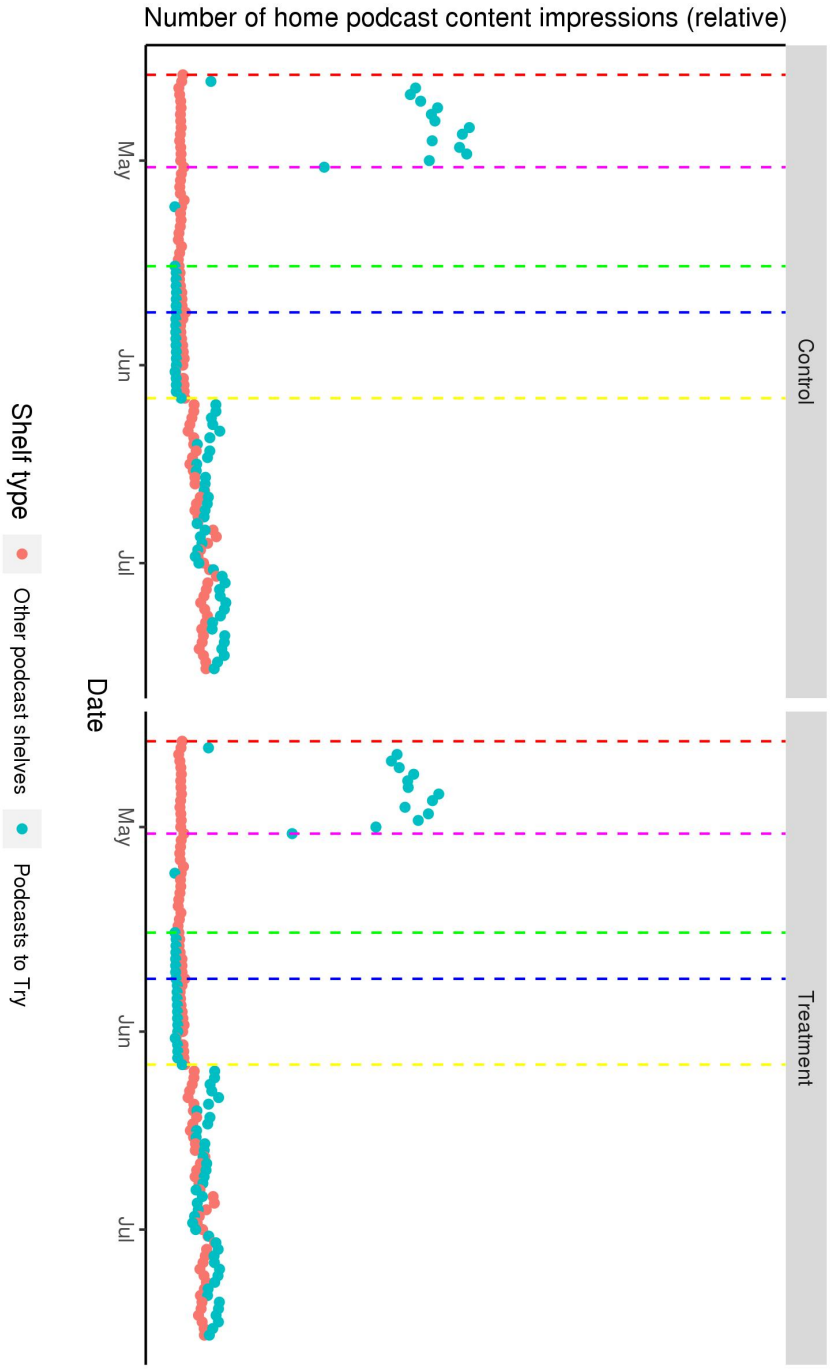

Figure I.14 **The number of daily podcast content impressions from both the "Podcasts to Try" shelf and other podcast-related shelves on the "Home" section of the Spotify app, shown separately for users in the two treatment arms of the experiment. The dashed red line corresponds to the experiment launch 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.**

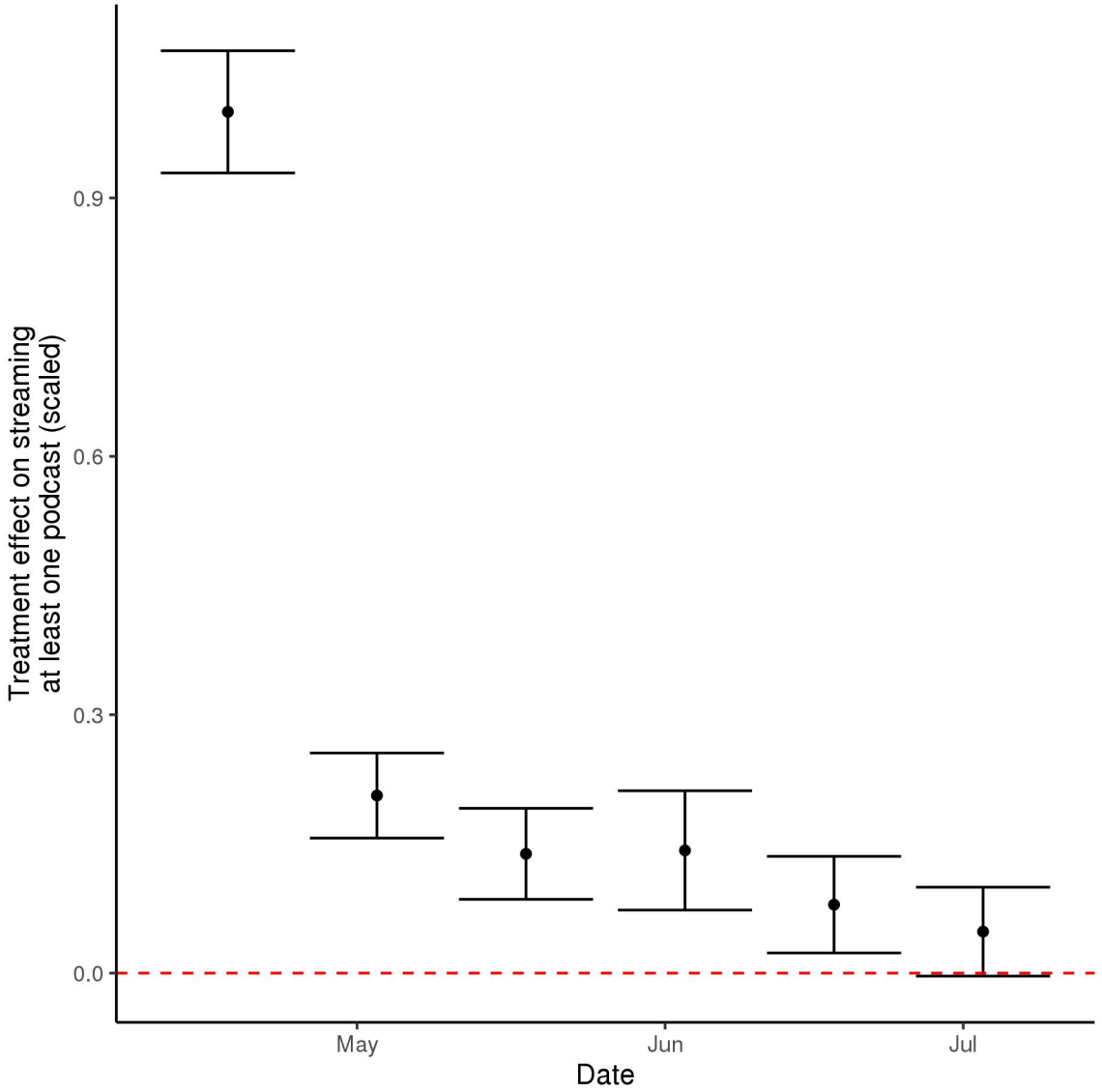

Figure I.15 **The long-term effect of the treatment on the percentage of users streaming at least one podcast. The time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.**

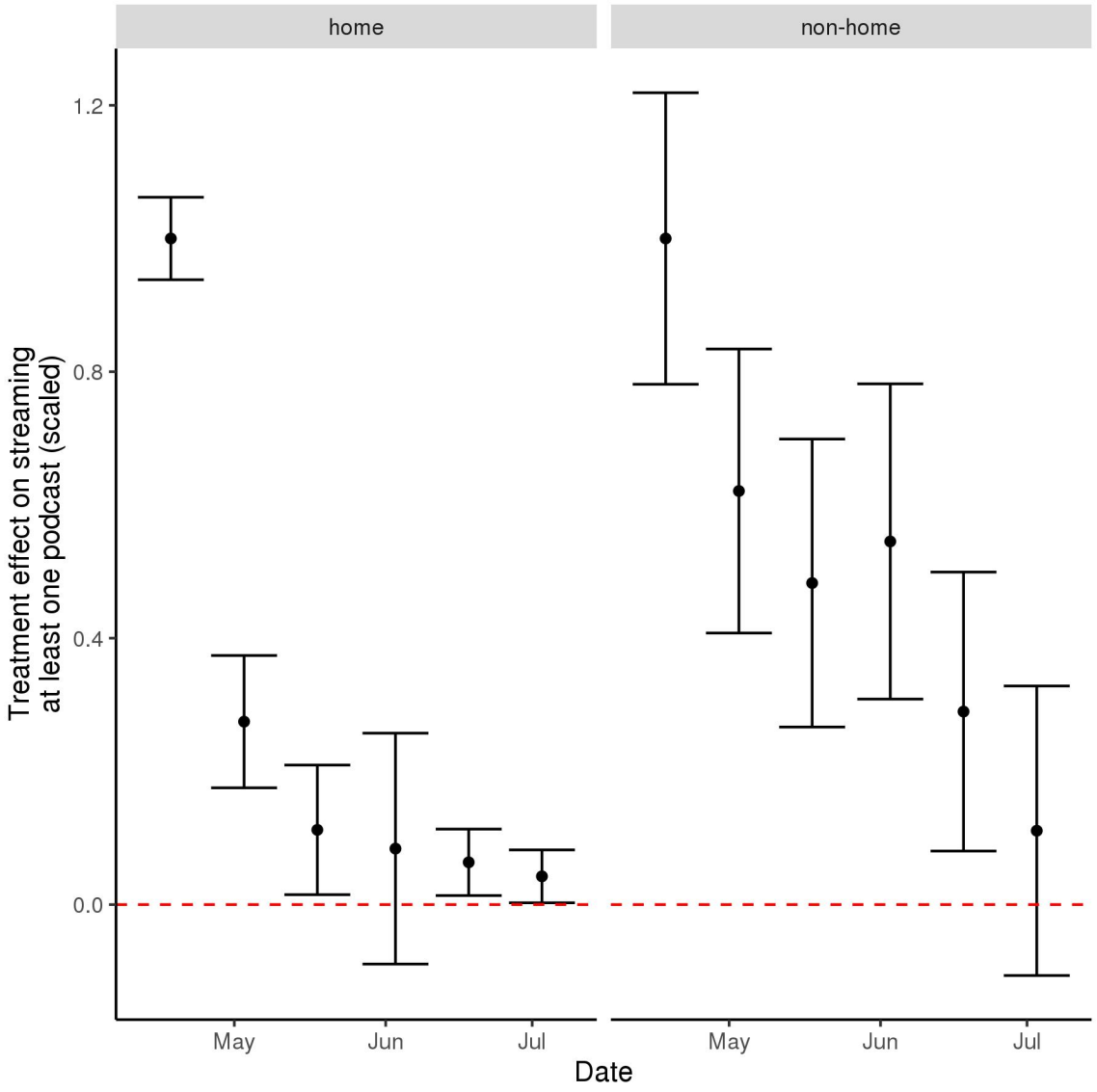

Figure I.16 **The long-term referrer-level effect of the treatment on the percentage of users streaming at least one podcast over time. Each time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.**

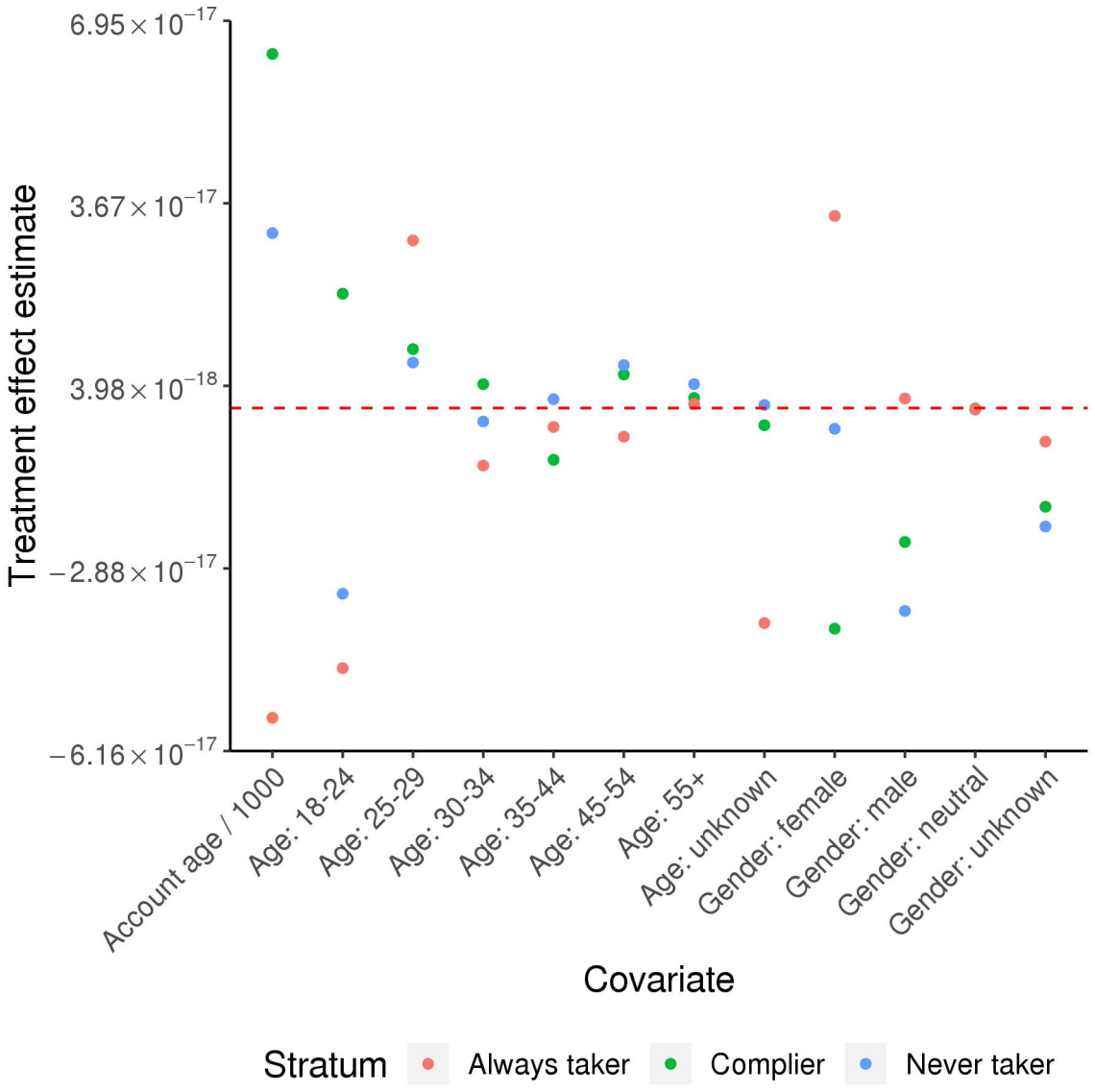

Figure I.17 **Results of the principal stratification balance check. The intermediate variable is whether a given user streamed at least one podcast during the experiment.**

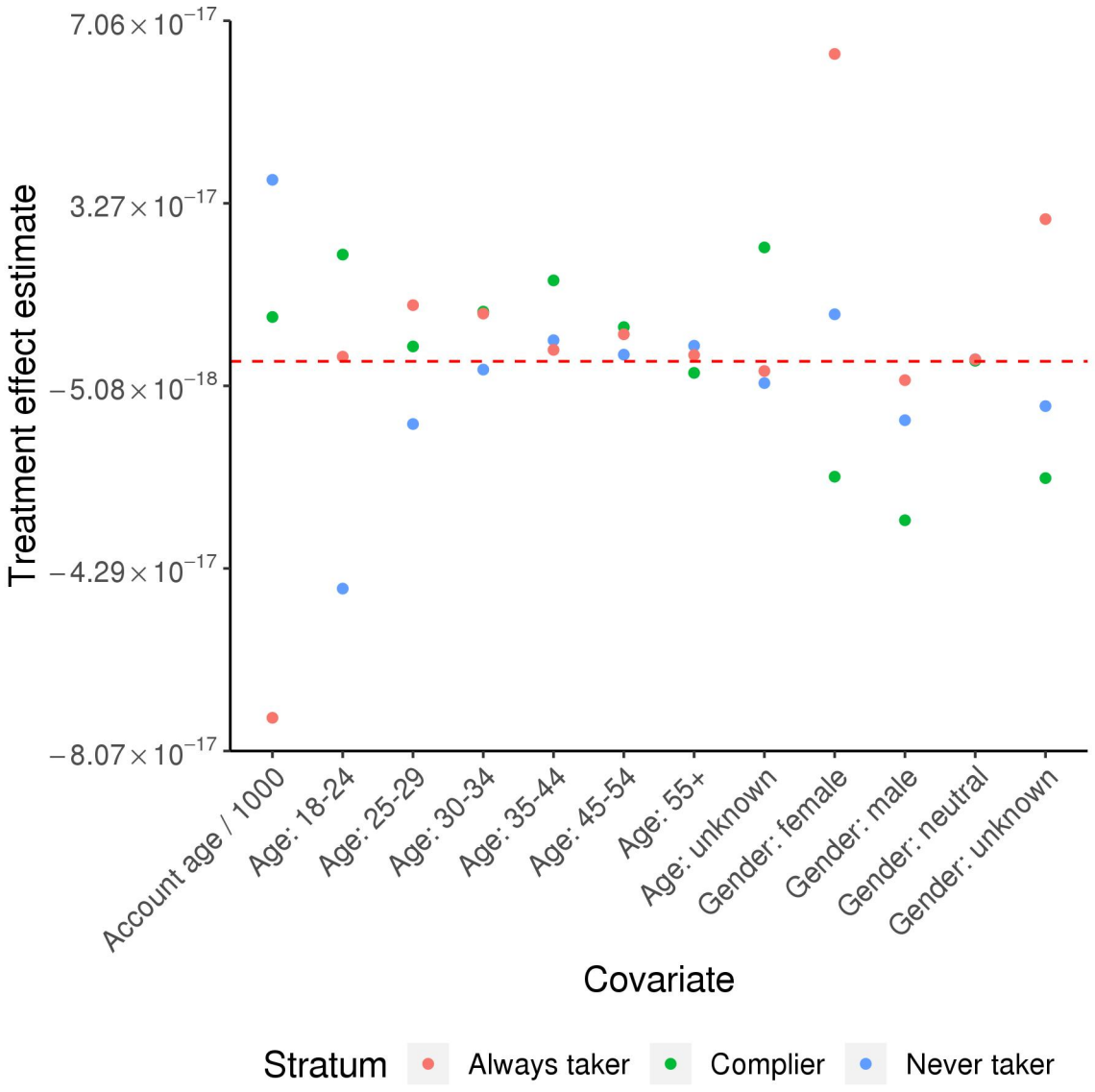

Figure I.18 **Results of the principal stratification balance check. The intermediate variable is whether a given user followed at least one podcast during the experiment.**

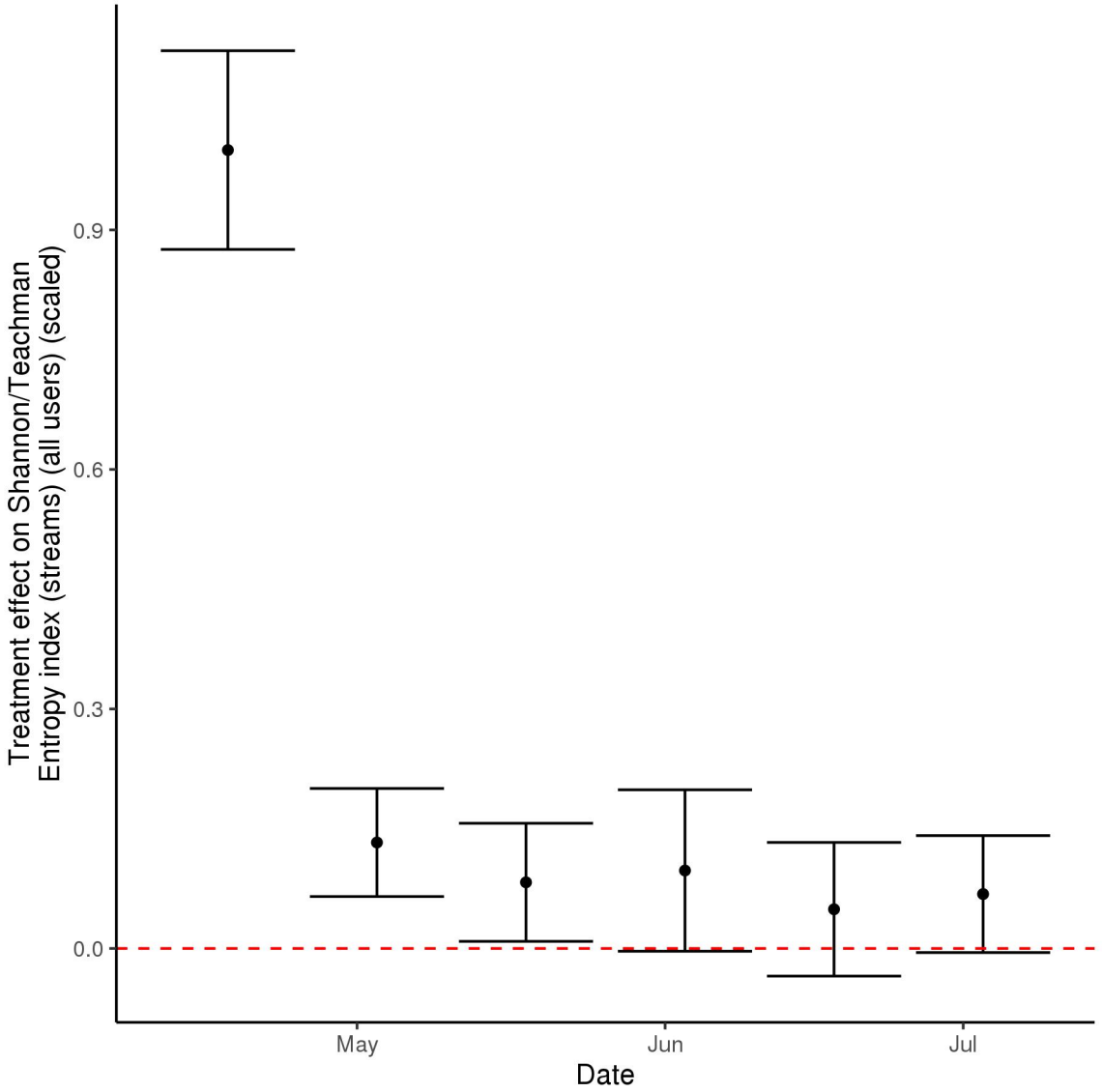

Figure I.19 **The long-term effect of the treatment on the average user-level Shannon entropy for streams. The time series is scaled by the absolute value of the magnitude of the treatment effect during the experiment.**

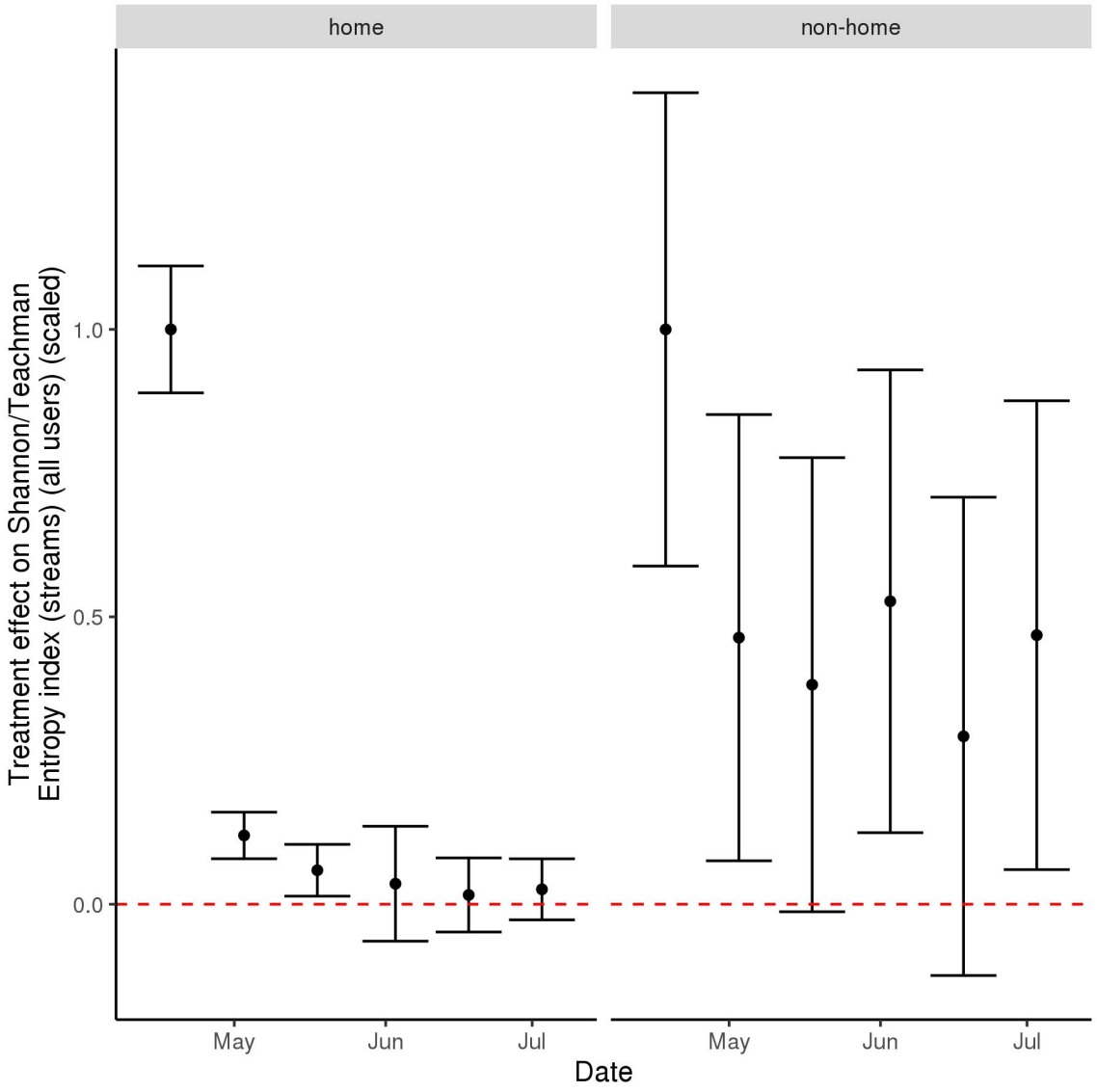

Figure I.20 The long-term referrer-level effect of the treatment on the average user-level Shannon entropy for streams over time. Each time series is scaled by the absolute value of the magnitude of the treatment effectduring the experiment.

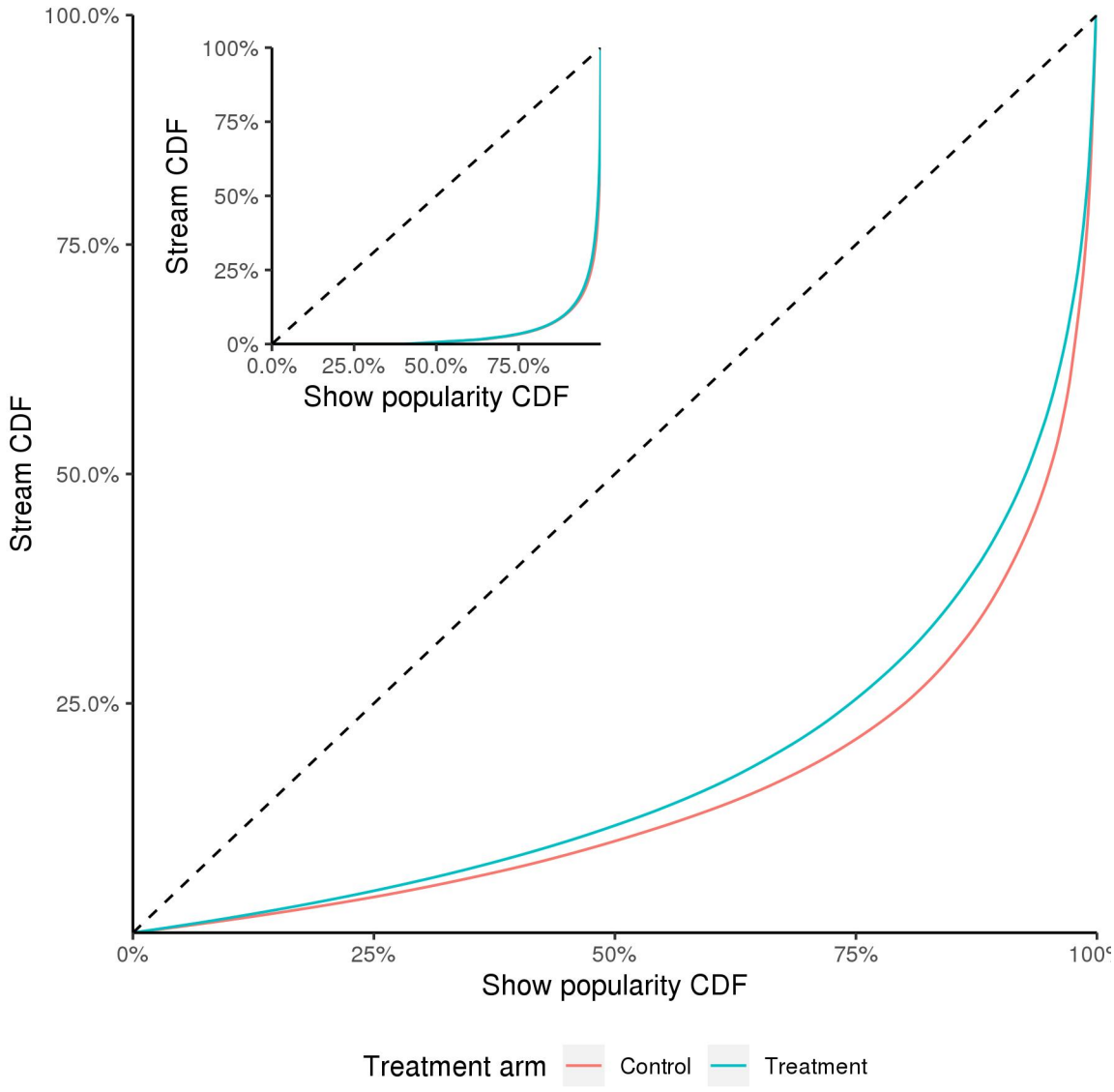

Figure I.21 **The Lorenz curves for podcast streams, calculated separately for users in the treatment and control. The data for each Lorenz curve is limited to the 1,000 most streamed podcasts in the corresponding treatment arm data. The inset curve shows the Lorenz curve for streams across all podcasts.**

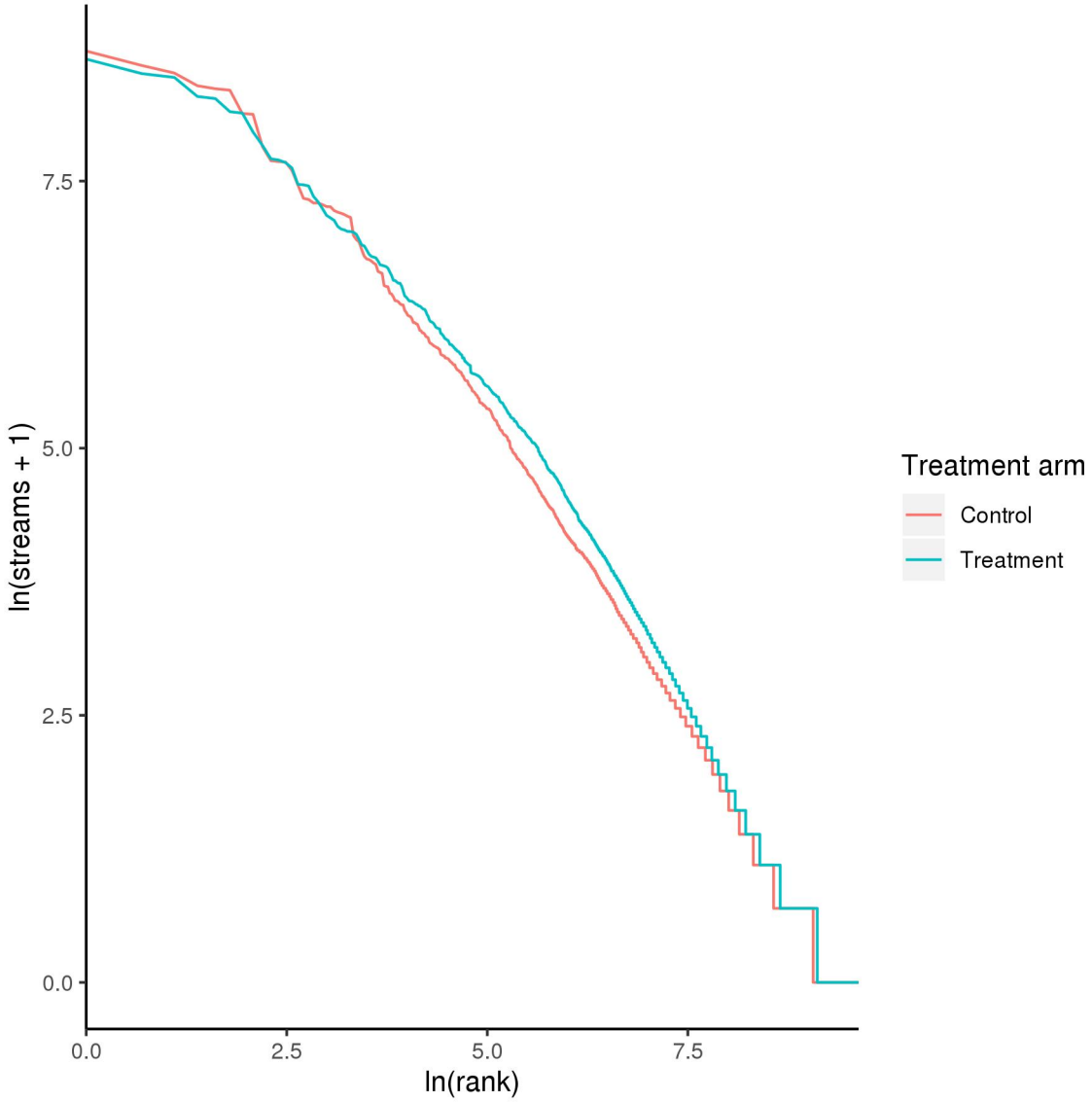

Figure I.22 **The relationship between ln(streams + 1) and ln(stream rank) for both the control and treatment arms of the experiment.**

Appendix J: Additional Tables

Table J.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	Stat. 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 J.2 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	

Table J.3 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

Table J.4 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	

Table J.5 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

Table J.6 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

Table J.7 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

Table J.8 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:			
	Stremed 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 J.9 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 J.10 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

Table J.11 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

Acknowledgments

Analysis for this paper was conducted while David Holtz was an intern at Spotify during the summer of 2019. The authors 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.