Algorithmic Effects on the Diversity of Consumption on Spotify

Ashton Anderson
University of Toronto & Spotify
ashton@cs.toronto.edu

Lucas Maystre
Spotify
lucasm@spotify.com

Rishabh Mehrotra
Spotify
rishabhm@spotify.com

Ian Anderson
Spotify
iananderson@spotify.com

Mounia Lalmas
Spotify
mounia@acm.org

ABSTRACT

On many online platforms, users can engage with millions of pieces of content, which they discover either organically or through algorithmically-generated recommendations. While the short-term benefits of recommender systems are well-known, their long-term impacts are less well understood. In this work, we study the user experience on Spotify, a popular music streaming service, through the lens of *diversity*—the coherence of the set of songs a user listens to. We use a high-fidelity embedding of millions of songs based on listening behavior on Spotify to quantify how musically diverse every user is, and find that high consumption diversity is strongly associated with important long-term user metrics, such as conversion and retention. However, we also find that algorithmically-driven listening through recommendations is associated with reduced consumption diversity. Furthermore, we observe that when users become more diverse in their listening over time, they do so by shifting away from algorithmic consumption and increasing their organic consumption. Finally, we deploy a randomized experiment and show that algorithmic recommendations are more effective for users with lower diversity. Our work illuminates a central tension in online platforms: how do we recommend content that users are likely to enjoy in the short term while simultaneously ensuring they can remain diverse in their consumption in the long term?

ACM Reference Format:
Ashton Anderson, Lucas Maystre, Rishabh Mehrotra, Ian Anderson, and Mounia Lalmas. 2020. Algorithmic Effects on the Diversity of Consumption on Spotify. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3366423.3380281

1 INTRODUCTION

On many online platforms, users can engage with millions of pieces of content, which they discover either on their own or through algorithmically-generated recommendations. The user experience, in turn, is largely shaped by the content that users interact with. This dynamic gives rise to a question of central importance for online platforms: How do recommendation algorithms affect the user experience by influencing which content users consume?

A fundamental characteristic of a user's content consumption is its *diversity*—how broad is the set of pieces of content they engage with? On one extreme, a user can consume very related content, and only interact with a small part of the "space" of content available on the platform. On the other extreme, a user can consume pieces of content that are very different from each other, and therefore engage with very different parts of the content space. Whether it is decided consciously or not, every user's consumption lies somewhere on this spectrum. In this work, we investigate algorithmic effects on content consumption through the lens of diversity. In particular, we aim to answer: What is the association between algorithmic recommendation and the diversity of content users consume, and on the user experience in turn?

This question has deep ramifications for online platforms. Prior research has shown that diversity is often a desirable property of content consumption, in a variety of contexts [4, 7, 26, 27]. However, there has been widespread concern that algorithms in general, and recommender systems in particular, could encourage people to concentrate on overly narrow sets of content and get trapped in "filter bubbles" [22]. Are recommender systems on online platforms serving to increase or decrease consumption diversity? And what is the resulting effect on the user experience?

Despite the importance of the relationship between algorithmic recommendation and content diversity, it has been heretofore difficult to study. First, analyzing user-driven versus algorithmically-driven consumption in an online platform necessitates a setting where both user discovery and algorithmic recommendations are prominent, and where detailed behavioral traces are recorded. Second, quantifying the diversity of an arbitrary subset of items among the millions available on a platform is a challenging task. To do so, one needs a consistent notion of similarity that can compare any pair of items, and also a way of scoring an arbitrary subset of items on a continuous scale from "very similar" to "very diverse". Existing approaches have employed simple measures, such as entropy and the Gini coefficient, which capture the extent to which users consume different items, but fail to take into account the *similarity* between the items. Third, connecting diversity to the user experience requires reliable measures of user experience and user success, as well as the ability to measure diversity over time.

The Present Work. To investigate how algorithmic recommendations relate to consumption diversity, we conduct large-scale analyses and experiments on the music streaming platform Spotify, an ideal platform for investigating our research questions. First, Spotify users can discover music either with user-guided search and exploration, or through algorithmic recommendations — both

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380281

of which account for a substantial amount of activity. This mix of user-driven and algorithmically-driven consumption allows us to compare these two modalities. Second, we leverage large volumes of listening patterns to determine a consistent notion of similarity between any two songs, thus solving the problem of quantifying content diversity in a principled way. Third, our dataset comprises fine-grained interaction data of over 100 million users listening to music over several years. This allows us to study how long-term success metrics, such as conversion and retention, are associated with content diversity at a high resolution and large scale.

To solve the problem of quantifying consumption diversity[1], we use a high-quality *song embedding* that places songs in a vector space such that related songs are close together in the space. The sheer volume of user listening patterns on Spotify encode a tremendous amount about how songs relate to each other, and by embedding every song into the same space we can capture the similarity between any two songs while ensuring that these similarities are mutually consistent with each other. We follow recent work and adapt the generalist-specialist score (GS-score) to use vector representations of songs to measure the diversity of a user's consumption [28]. The GS-score is the expected cosine similarity between a user's center of mass vector (mean over all the songs they have consumed) and the vector representation of a randomly drawn song they have listened to. This similarity will be high if the songs are clustered together in the space (the user is a "specialist") and will be low if the songs are far apart in the space (the user is a "generalist"). In this way, we determine the content diversity of every user in our dataset using the same song embedding. We further discuss the benefits and characteristics of the GS-score, as well as its relation to other diversity metrics, in Section 3.

Overview of Results. Analyzing these musical diversity scores for each user, we uncover several important insights. First, for each user we calculate separate musical diversity scores for user-driven listening and for algorithm-driven listening. Comparing these scores, we find that user-driven listening is typically much more diverse than algorithm-driven listening. Furthermore, in a separate analysis of user diversity over time, we find that users who become more generalist over time tend to do so by drifting away from algorithmically-driven listening and gravitating towards user-driven listening. These results strongly suggest that algorithmic recommendations are associated with *reduced* consumption diversity.

Second, we investigate how key user business metrics are associated with musical diversity. Controlling for activity (the number of songs played), generalist users are up to 25 percentage points less likely to leave the platform and up to 35 percentage points more likely to become a paid subscriber. Thus, our key user metrics are very strongly associated with *diverse* listening patterns.

Finally, we conduct a randomized experiment on Spotify to understand how generalists and specialists respond to recommendations. We find that, for both groups, recommending using relevance is an effective strategy for satisfying short-term needs on Spotify, as on other platforms. However, we also find that relevance is much more important to specialists than to generalists, suggesting that there are opportunities to develop recommendation strategies that take diversity into account.

Taken together, our work suggests there is a fundamental predicament that online platforms must deal with. On the one hand, to meet immediate user needs, recommendation algorithms are powerful and effective drivers of engagement. However, we also find that they tend to produce less diverse user consumption. On the other hand, we find that more diverse user consumption patterns are very strongly associated with positive long-term metrics like user conversion and retention. The predicament is this: how do we design online platforms that cater to pressing and specific user needs, but that also allow users to fully explore the world of content available to them? This is a tension between short- and long-term goals: if we need to recommend content urgently, a good strategy is to promote relevance (and thus discourage diversity), but if we want to attract and retain users, ensuring that consumption is sufficiently diverse appears important. The challenge going forward will be to develop methods and algorithms that can simultaneously deal with these conflicting incentives. Our approach of viewing online platforms and algorithmic recommendations through the lens of content diversity is the first step towards this goal.

2 RELATED WORK

Our work draws on three research areas: analyses of diversity in user behavior in recommender systems and online platforms, studies of music listening and streaming services, and work on trade-offs between short- and long-term engagement in online platforms.

Diversity in recommender systems. Ever since recommender systems exploded in popularity, researchers have been studying how they affect user consumption and behavior. Particularly relevant to our work here are investigations centered around the diversity of users' consumed content. This line of research can be traced back to Pariser, who referred to the potential of recommender systems and other personalized algorithms to expose users to an increasingly small and focused "filter bubble" of content that they are likely to agree with [22]. Various studies have tested this hypothesis, coming to mixed conclusions about whether, and to what extent, recommender systems display a filter bubble effect [2, 10, 21].

Many researchers have conducted observational analyses and randomized experiments to understand the impact of recommender systems on diversity, as we do here. Fleder and Hosanagar found that in aggregate, diversity decreases as a result of recommender system use in sales [11], whereas Zhou et al. found that the YouTube recommendation system increased the diversity of video views in 2010 [31]. Our work complements these by using a much more fine-grained definition of diversity and by connecting diversity with important user-level metrics. We discuss the relation between our definition and two common diversity measures in Section 3.3.

There is also a rich line of work proposing diversity-aware recommendation algorithms. Among others, Adamopoulous and Tuzhilin discussed over-specialization in recommendations and introduced a method that promotes diversity in a collaborative filtering framework [1], De Choudhury et al. proposed a clustering technique for recommending social media content that matches a specified level of information diversity [7], and Zhou et al. developed a recommendation algorithm that promotes diversity without sacrificing too

[1]Given our focus on music listening, we will sometimes refer to this as "musical diversity", "listener diversity", or "listening diversity".

much accuracy [32]. Particularly related to our work is Auralist, a collection of algorithms designed around "serendipitous discovery" in music [30]. Finally, there is research on the perception of diversity in recommender systems, including work by Graells-Garrido et al. showing the importance of the interface for users to be aware of diverse content recommended to them [13] and work by Hu and Pu showing that how results are presented influences users' perceptions of diversity [15].

Our work draws upon efforts to measure and study diversity more generally. The information retrieval community has been investigating diversity metrics in a number of domains for decades (e.g. see [5, 25]), and more recently, diversity measures have been used to quantify notions of fairness in online platforms [18]. In this work, we adopt Waller et al.'s generalist-specialist score (GS-score) and apply it to the study of musical diversity on Spotify [28]. This technique stems from a long history of work on generalist and specialist approaches in medicine, the arts, and beyond [3, 26]. We describe the GS-score in detail in Section 3.3.

Music listening and streaming services. As we study Spotify, our work relates to other analyses of musical diversity in online streaming services. Using a longitudinal panel dataset, Datta et al. conducted a study of music listening behavior across many streaming platforms and found that using streaming services increased the volume and diversity of music that people listened to [6]. Park et al. proposed a definition of musical diversity based on genres and correlated it with demographic and behavioral variables [23]. Our work extends this by using a more data-driven, fine-grained measure of musical diversity, studying it in a system with both user- and algorithmically-driven consumption, and running experiments.

Short-term vs long-term engagement. This work is concerned with the interplay between short-term and long-term goals. On the one hand, online platforms designers need to ensure they satisfy short-term user needs, and on the other hand, they need to ensure long-term engagement. Recommending content that users do not engage with because they are not relevant or of low quality can have severe implications long-term: users reducing their engagement with the platform and potentially even churning. Previous research has sought to understand why users churn and how to limit this churn, e.g. on streaming platforms [12] and question & answering platforms [8]. Other studies have looked at the relationships between short-term and long-term metrics, e.g. in the context of search [9] and advertising [17], as well as building models that attempt to optimize for both e.g. [16, 29]. Finally, the trade-off between short-term and long-term engagement has been studied extensively in the context of A/B experiments [14]. Our research adds to this body of work by looking at how diversity in music listening impacts short-term and long-term engagement with Spotify.

3 DATA AND METHODOLOGY

Our goal is to measure musical diversity at a global scale, and study its relationship with long-time user and platform outcomes. In this section, we describe our datasets, explain the methodology we use to construct a high-fidelity music embedding, and introduce the measure we use to quantify musical diversity.

3.1 Data

We study Spotify, an online streaming platform where users can listen to a vast selection of music from around the world. Users have access to a catalogue of over 50 million songs to play at any time through their computers, mobile devices, or other internet-connected devices. Spotify has both a free ad-supported product and a premium subscription product. On the former, users will periodically be served advertisements and have some limitations on what can be played on-demand. Additionally, ad-supported users have a limited number of skips per hour and are unable to download music for offline listening. Subscribers pay a monthly fee and are unrestricted in what they can listen to.

Our main dataset consists of the listening history of over 100 million distinct users who cumulatively listened to millions of songs around 70 billion times during the first 28 days of July 2019. Because of the aforementioned restrictions on ad-supported listeners, we restrict our study to premium users only. For our temporal analyses in Section 5, we supplement this dataset with similar datasets from the first 28 days of other months. We use 28-day periods to ensure each type of weekday appears the same number of times in all datasets.

For any music streaming service providing on-demand capabilities, the way that a user streams can be broadly broken down into two categories: user-driven ("organic") listening and algorithm-driven ("programmed") listening. On Spotify, users can search for particular songs, build collections that they can later return to, or listen to playlists made by other users. We classify any listening resulting from these interactions as *organic* listening — streams driven by user action.[2] On the other hand, a user can listen to an algorithmically personalized playlist (e.g. Discover Weekly), curated playlists, or radio stations algorithmically generated by a seed song. Though the user may have initiated the stream, songs are determined without requiring feedback from the listener. We classify these streams as *programmed* listening — which content to play was chosen algorithmically. To understand how recommendation algorithms interact with musical diversity, this is how we partition streaming behavior on Spotify.

3.2 Music Embeddings

Many recommendation system approaches try to efficiently store information by using embeddings that encode latent representations between users and content. When done correctly, pieces of content that are strongly related to each other will have representations that are close to each other in the embedding space (e.g. "Yesterday" by The Beatles will be close to John Lennon's "Imagine").

To produce such an embedding space, Spotify trains the word2vec algorithm [19] on user-generated playlists. Traditionally, word2vec is used in natural language processing to embed words from a corpus of documents [20]. In the context of music, playlists are treated as "documents" that are often thematically coherent, and the songs they contain are treated as the "terms" in those documents. For our implementation, we use the continuous bag-of-words model, where the task is to predict the song in the middle of the context

[2] Although the results from a search are algorithmically ranked by relevance, any listening resulting from a search is considered organic because the user provided the seed query.

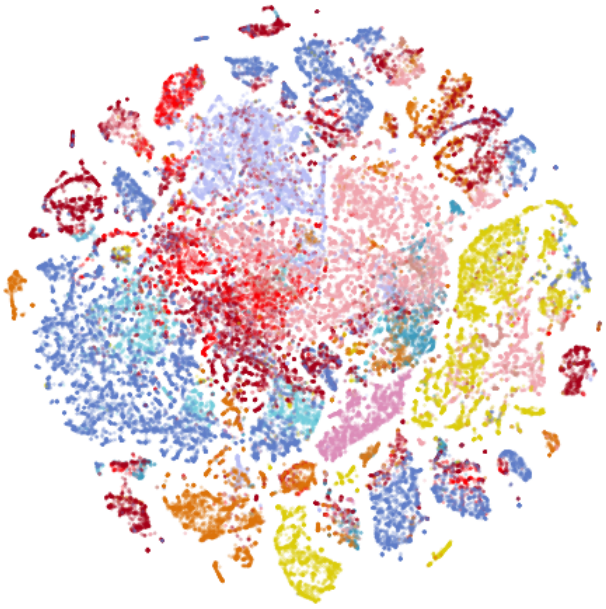

Figure 1: Two-dimensional t-SNE projection of 100,000 songs from our song embedding. Songs are represented as points, where two songs are close together if they have high usage overlap in our data. Colors represent genres of music.

window given the surrounding songs. This naturally causes songs that frequently co-occur in playlists to have nearby embeddings in the space. For example, because users frequently build genre-themed playlists, songs in the same genre will tend to appear close to each other, as seen in Figure 1.

Our embedding space was constructed using more than 850 million playlists predominantly built by users, where we have filtered out playlists that do not meet certain criteria (e.g. playlists that are too short to produce meaningful samples, playlists that are too long to have been built around a coherent theme, playlists that a user has not recently updated, etc.).

Within this embedding space, we define a user's taste profile to be the average of the song embeddings that they have listened to within the time window being considered. The vectors used in this paper are 40-dimensional and are re-trained every day to account for new content added to the platform. To calculate how similar two songs are, we directly map this to the distance between their representative points in space by calculating their *cosine similarity*. To maintain consistency we use the embedding trained on the last day in our dataset, July 28, 2019.

3.3 Generalist-Specialist Score

With our music embedding in place, we are ready to quantify the musical diversity of a given user. Recall that our goal is to measure how "spread out" a user's musical interests are in music space, while taking into account fine-grained distinctions between the similarity of songs. If a user listens to very similar songs, we want to label them a *specialist*, and if they listen to a diverse set of songs, we want to label them a *generalist*. To quantify musical diversity using our music embedding, we apply the *generalist-specialist score*, or GS-score [28]. Intuitively, a specialist's song vectors will be close together in the space, and a generalist's song vectors will be spread apart. To capture this, the GS-score measures the average cosine similarity between a song vector and the average of the user's song vectors. For specialists this similarity will be high on average, whereas for generalists it will be low on average.

Formally, say user u_i listens w_j times to song s_j, and let $\vec{s_j}$ denote s_j's vector representation in the song embedding. The first step is to define u_i's center of mass, which is simply the centroid of their songs' vectors, weighted by the number of times they listened:

$$\vec{\mu_i} = \frac{1}{\sum w_j} \cdot \sum_j w_j \vec{s_j}$$

Then u_i's GS-score is:

$$GS(u_i) = \frac{1}{\sum w_j} \sum_j w_j \frac{\vec{s_j} \cdot \vec{\mu_i}}{\|\vec{s_j}\| \cdot \|\vec{\mu_i}\|}$$

Using u_i's listening data $D_i = \{(s_j, w_j)\}$ from time period T, we refer to $GS(u_i)$ computed on D_i as the *musical diversity* of u_i at time T.

There are many definitions of consumption diversity in the literature, the most common of which are functions of the number of times each item is consumed, such as the Gini coefficient [11, 31] and entropy [7, 24]. While these capture the extent to which users consume different items, they ignore the *similarity* between these items. For example, a user who listens equally often to two songs by the Beatles and a user who listens equally often to a funk song and a Gregorian chant would be classified as equally diverse by these metrics, even though the latter listens to songs that are less similar to each other. It is thus desirable to use a diversity metric that incorporates both the number of times items are consumed as well as the similarity between the items, which the GS-score does.

In addition, it is not straightforward to incorporate item-level similarity into the study of diversity, since devising rigorous similarities between all pairs of items that are both accurate and consistent with each other is a challenging task. However, we are able to achieve this using our music embedding space—the similarity between songs is succinctly captured by the cosine similarity between their vector representations in the space. Although we believe our diversity measure is preferable to other metrics because of its granularity, natural interpretation, construction from explicit user feedback, and sensitivity to item similarities, it is correlated with other diversity metrics that we could have used. For example, we found that user GS-score and a user entropy on genre categories have a correlation coefficient of $r = -0.77$ with each other, and a user GS-score and a user Gini coefficient on genre categories have a correlation coefficient of $r = -0.60$ with each other. Thus, the results we report in this paper would likely be qualitatively similar if we instead used the Gini coefficient or entropy.

4 MUSICAL DIVERSITY ON SPOTIFY

We now apply the GS-score to quantify diversity in music listening at a global scale on Spotify.

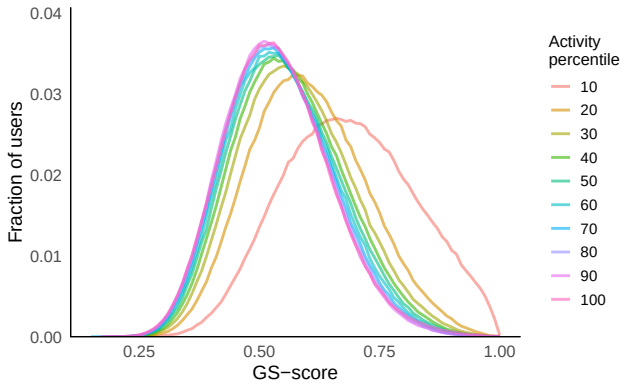

Figure 2: Distributions of diversity controlling for activity.

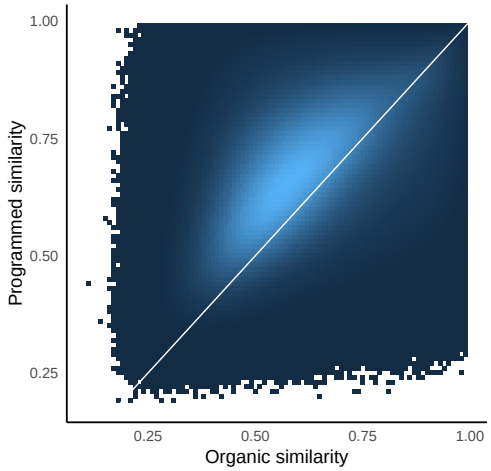

Figure 3: Distribution of organic vs. programmed diversity. The $y = x$ line is shown in white. The vast majority of users are above this line, indicating that their programmed listening is less diverse than their organic listening.

4.1 Distributions of Diversity

Diversity by activity level. First, we calculate the GS-scores of all users in our dataset. The distributions of these scores for every level of activity are shown in Figure 2. Consistent with previous work [28], we find a broad diversity of listening diversity behaviors, ranging from extreme generalists to extreme specialists, at all levels of activity. There is a slight correlation between listening diversity and activity; users who did not listen to many tracks in our timeframe are slightly more likely to be specialists, especially the lowest-activity listeners. But there is still substantial overlap in the listening diversity of these users with those who streamed a lot of tracks. We also see that as activity goes up, the distribution of listening diversity converges to a stable distribution of the GS-scores centered around 0.5.

Organic versus programmed diversity. A main focus of our work is the association between algorithmic recommendations and the breadth of user listening behavior. As discussed in Section 3, we categorize all streaming on Spotify into user-driven ("organic") listening and algorithmically-driven ("programmed") listening. Here we ask: how does the breadth of user-driven listening compare with the breadth of algorithmically-driven listening?

For this analysis, we separately compute two GS-scores for each user, one on organic listening only and another on programmed listening only. We visualize the joint distribution of these organic and programmed GS-scores in Figure 3.

Inspecting this joint distribution gives a resounding answer to our question: user-driven listening behavior is almost always more diverse than algorithmically-driven listening behavior (since most of the mass is above the $y = x$ line). This is not entirely unexpected, since recommendation algorithms supporting listening on Spotify, like analogous algorithms on many other platforms, recommend tracks that are similar to what users have already consumed. However, the magnitude of the effect is striking. Algorithmic recommendations support "focused" consumption that is more clustered together in the space, whereas organic listening is more spread out.

This could be caused by many mechanisms. It could be the case that users use recommendations to satisfy their similarity-based needs and organic streaming to satisfy their exploration-based needs. It could also be that recommendation algorithms actually push users towards less diverse listening. However, regardless of the underlying mechanism, it remains true that algorithmic recommendations are associated with less diverse consumption.

4.2 Demographic variation in musical diversity

How does listening diversity vary with demographic attributes? Here we connect the organic and programmed GS-scores with self-reported age and gender. Throughout this subsection, we will be comparing the joint distributions of organic and programmed listening (as in Figure 3) between two groups. To perform these comparisons, we use the *log odds ratio* of the two distributions. For two groups A and B, at each (x, y) pair with x organic GS-score and y programmed GS-score, we compute the log odds ratio of the two groups as follows:

$$LOR(x, y) = \log \frac{p_A(x, y)/(1 - p_A(x, y))}{p_B(x, y)/(1 - p_B(x, y))}$$

If $p_A(x, y) = p_B(x, y)$, then the log odds ratio is 0. If $p_A(x, y) > p_B(x, y)$ then the log odds ratio is positive, and it is negative if the opposite is true.

Product type. First, we comment on how which Spotify product a user chooses affects their musical diversity, in particular whether they use the free version or have a subscription ("premium"). We calculate the joint distribution of organic and programmed GS-scores for free users, and compare it to the joint distribution of organic and programmed GS-scores for premium users. For each (x, y) pair, with x organic GS-score and y programmed GS-score, we compute the log odds ratio of the free and premium probabilities. The major difference is in the diversity of organic streaming: premium users have much more diverse organic listening patterns than free users. This is due to differences in the products: Spotify's free product lets users listen to tracks without paying for a subscription, but they are

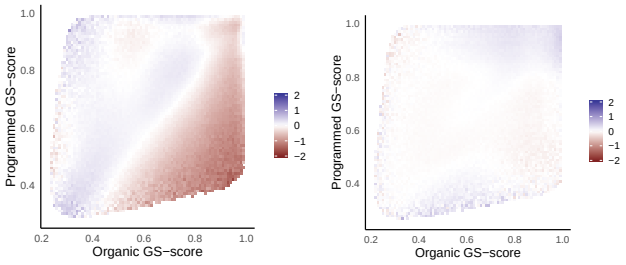

Figure 4: Log odds ratios of diversity distributions for (a) younger and older users and (b) male and female users.

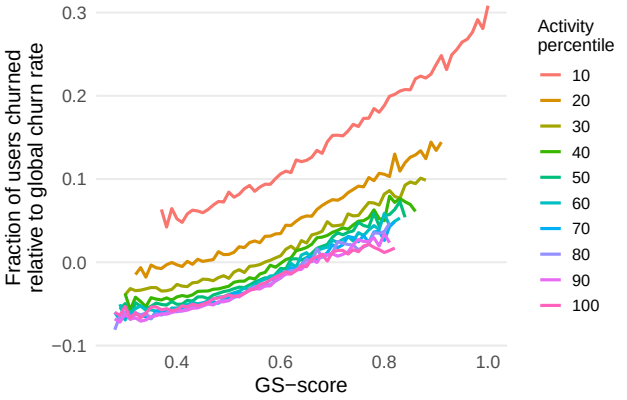

Figure 5: Probability of churning (relative to the global baseline churn rate) after one year as a function of July 2018 diversity.

correspondingly limited in their choices (e.g. number of skips per hour), which in turn affects their organic diversity. For this reason, throughout this paper we focus on premium users, including the figures already shown above.

Age. We turn to measuring how diversity varies with demographic attributes. Similar to our analysis of how free and premium users differ, we compare how 18–24 and 45+ users differ in their diversity. For each (x, y) pair of organic GS-score x and programmed GS-score y, we compute the probabilities that younger users and older users have these exact GS-scores, then take the log odds ratio of these probabilities to compare them. This comparison is shown in Figure 4a, where areas in which younger users are disproportionately likely to occur are shown in blue and vice-versa in red. The pattern is very apparent: as age increases, organic diversity goes down and programmed diversity goes up. This effect is very practically significant, with log odds ratios ranging up to 2 (corresponding to one probability being 7 times higher than the other).

Gender. We also analyze how musical diversity varies with self-reported gender. Similarly to the above analyses, we compute the log odds ratio of the probabilities that male users and female users have organic GS-score x and programmed GS-score y. We see in Figure 4b that there are no practically significant differences between the diversity of male and female users. This is reassuring, since we have no reason to expect males and females to diverge in their organic or programmed diversity.

4.3 User Retention and Conversion

Retention. One important metric for virtually any online platform is user retention: are users remaining on the platform? Which users tend to stay, and which tend to leave? Here we study this question on Spotify through the lens of musical diversity — do more diverse or less diverse listeners tend to stay on the platform longer?

To answer this question, we consider the tens of millions of premium users who were active in July 2018, measure their generalist-specialist scores for their activity during this month, and compute the empirical probability that they are active one year later as a function of their diversity. As before, we control for activity by binning users according to their number of tracks streamed in July 2018. We calculate the global baseline average churn rate and report all figures relative to this average (i.e. if the global average churn rate is 20%, we report a 15% churn rate as $15\% - 20\% = -5\%$).

The results of this analysis are shown in Figure 5. The trend is strikingly clear: at every activity level, generalist users are much more likely to remain on Spotify than specialist users. Furthermore, the effect is very large. For example, among the least active users, specialists churn at a rate 30 percentage points (p.p.) higher than the global average, whereas the same figure is only 5 p.p. above average for generalists at the same activity level. Among highly active users, the specialists' churn rate is 3 p.p above average while generalists churn 7 p.p *below* average. Additionally, at the higher activity levels, changes in the GS-score are much more associated with changes in user retention than changes in activity are (the high-activity curves in Figure 5 start to overlap each other). This is especially remarkable given that user engagement (i.e. activity) is traditionally the strongest predictor of user retention.

While it is possible that this is a causal effect (i.e. that diverse listening causes one to remain on the platform), there are many other explanations as well. For example, it could be that both listening diversely and remaining active on music platforms are signs of an engaged music lover. However, this strong association is important regardless of the underlying mechanism, as increases or decreases in diversity could be indicative of fundamental long-term outcomes. Also, knowledge of which users are more likely to churn can be very valuable to online platforms.

Conversion. Another crucial metric for many online platforms is user conversion: are users who try the free product likely to convert into paid subscribers? Which users are more likely to become premium members? We consider this question through the lens of diversity: are generalists or specialists more likely to become premium members? Again, we calculate the global average conversion rate and report all figures relative (additively) to this baseline.

For this analysis, we consider the tens of millions of users who were using the free product during July 2018, calculate their GS-scores on all activity during this month, then measure how often

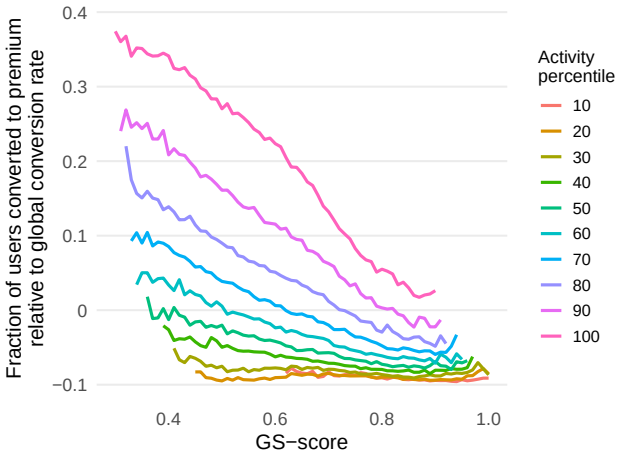

Figure 6: Probability of converting from free to premium after one year as a function of July 2018 diversity.

free users convert to premium as a function of their musical diversity, controlling for activity.

The effect of diversity on conversion is perhaps even more striking. As can be seen in Figure 6, Spotify users on the free product are far more likely to convert to premium if they are generalists than if they are specialists. This effect is particularly acute for high-activity users; for the highest activity level, extreme generalists convert 38 p.p above average, whereas extreme specialists only convert 3 p.p. above average — a difference of 35 percentage points. The effect attenuates as activity drops, since low-activity users are on the whole less likely to convert into paid subscribers.

Discussion. Taken together, these results strongly suggest that positive user outcomes such as conversion and retention are associated with *greater* content diversity. We also find, however, that algorithmic consumption is *less* diverse than organic consumption. Thus, it appears that recommendation algorithms, although they drive a significant amount of engagement, are associated with lower content diversity. When combined with the association between diversity and key user metrics, this raises the possibility that recommendation algorithms can be effective in the short-term while potentially being partially counterproductive in the long-term.

We note that there are many possible explanations underlying these results, and uncovering exactly why generalist users are so much more likely to convert to premium and remain on Spotify and other platforms is an interesting direction for future work. Furthermore, uncovering whether this is a causal effect is also important. Regardless of the underlying mechanisms, however, the associations we have discovered are valuable insights. For example, knowing which users are likely to churn is a central problem for online platforms, and content diversity is a very strong correlate in our data. Similarly, knowing which users are likely to convert to subscriptions and which are not helps platforms direct their efforts efficiently. At a higher level, knowing that diverse consumption patterns are highly associated with important outcomes — whether it is driven by fundamental user differences, indicative of greater engagement, or a causal effect — should drive platforms to better understand this relationship.

5 MUSICAL DIVERSITY OVER TIME

In the previous section, we established that the GS-score, measured in a static snapshot of one month of user data, is associated with important outcomes. Now, we build on these results by better understanding how diversity changes over time. How stable is a user's musical diversity? When users do become more or less musically diverse, what are the *mechanisms* by which this happens? Here, we investigate the evolution of musical diversity and how it relates with algorithmic recommendations.

Controlling for Activity. As discussed in Section 4, diversity is slightly correlated with activity — users who stream more are on the whole more generalist than users who stream less. Since the activity of a given user can change significantly over time, it is no longer possible to control for user activity by assigning users to a single activity bucket. To address this issue, we introduce the *activity-adjusted* GS-score. For a given time period T, a user activity-adjusted GS-score is the percentile rank of their GS-score relative to all users in the same activity bucket at time T. Now when we compare users across time, the GS-score is always relative to other users with the same activity. This enables comparing GS-scores across time, controlling for any activity biases.

5.1 The Stability of Musical Diversity

How stable is musical diversity over time? If it is a measure of an inherent quality of how a user consumes music, and if we assume many people have relatively stable music consumption patterns over time on Spotify, then we should observe that the GS-score is stable over time. In the following analysis, we validate our score of musical diversity and measure the stability of GS-scores.

To examine how GS-scores vary over time, we consider multiple snapshots of user activity and compare them. We complement the July 2019 dataset described in Section 3 with twelve additional datasets consisting of user activity during the first 28 days of each month between July 2018 and June 2019. Since free and premium users exhibit different patterns of diversity (see Section 4), we restrict our attention to premium users only. For every month m leading up to July 2019, we consider only users who streamed at least one track during both July 2018 and month m. For each user present in both time periods, we compute their GS-score on their July 2018 activity and their GS-score on their activity during month m. These scores are then adjusted for activity, as described above, to control any activity effect on our measure of musical diversity.

Figure 7 displays how the activity-adjusted GS-score changes over the one-year period we consider. More specifically, it depicts the probability distribution of an activity-adjusted July 2019 GS-score, conditioned on a given activity-adjusted July 2018 GS-score. First and foremost, we observe that most of the probability mass is concentrated on the diagonal, meaning that the GS-score remains stable after one year. Second, we observe that extreme generalists and extreme specialists are disproportionately likely to display the same musical diversity behavior from one year to the next, as indicated by the concentration of mass in the corners of Figure 7.

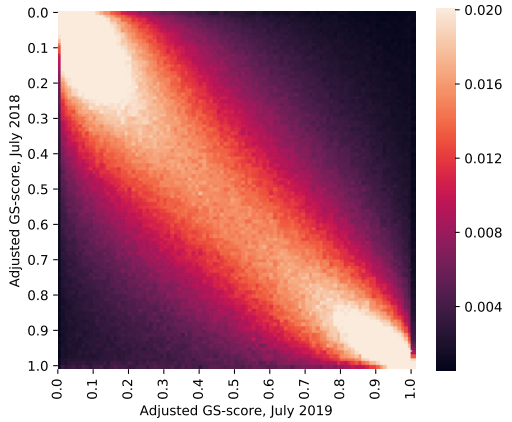

Figure 7: Distribution of GS-score in 2019 given a GS-score in 2018, for a sample of premium users.

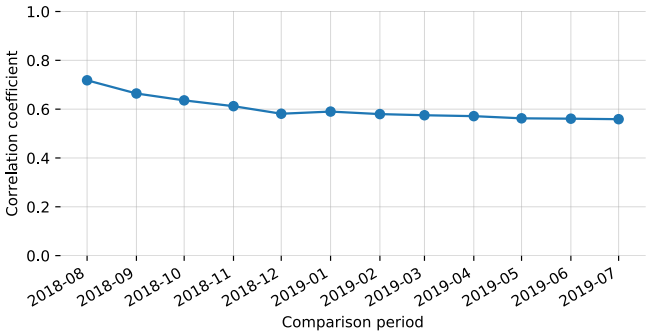

Figure 8: Pearson correlation coefficient between July 2018 GS-score and the GS-score of each of the 12 subsequent months.

This analysis compares two distant time periods. What about how the GS-score varies over a continuous period of time? Figure 8 displays the Pearson correlation coefficient between the GS-score of July 2018 and each of the 12 subsequent months. The coefficient between the two first contiguous months is 0.72, suggesting that a month-long window is still subject to observation noise. Remarkably, the correlation coefficient remains stable over time, ending at 0.56 after one year. This analysis provides more evidence that the GS-score captures a fundamental, latent characteristic of users. Musical diversity is very stable over long periods of time, indicating that it is an inherent behavioral trait on Spotify.

This also suggests that users' first experiences on Spotify are particularly important. Since musical diversity is stable over time, ensuring users start with a diverse experience could have long-term consequences. We leave a full exploration on how diversity relates to the cold-start problem to future work.

5.2 Mechanisms of Change

Despite the relative stability of the GS-score, some users *do* move on the diversity spectrum over time. As we have seen that higher musical diversity is associated with positive outcomes, we would like to know how this happens. When a user goes from the specialist side to the generalist side of the musical diversity spectrum, what is the mechanism by which they do this? Of particular interest to us here is understanding the relationship between user-driven vs. algorithmically-driven listening and a user's musical diversity. To answer this question, we conduct a fine-grained analysis of where each stream originates from, and contrast the patterns of users who increase their activity-adjusted content diversity to those who decreased their diversity.

We partition users into two sets: those whose GS-score has increased from July 2018 to July 2019, and those whose GS-score has decreased.[3] For each user, we compute how their streams are distributed over different play contexts, such as the user's library, the radio, etc. Recall that each play context is classfied as either being user-driven (e.g. searching, playing from user-made collections) or algorithmically-driven (e.g. algorithmically personalized playlists, radio). Finally, we aggregate these distributions across our two partitions, then compute log odds ratios to compare how users increase and decrease their content diversity.

Figure 9 displays the log odds ratios of streams from users whose GS-score has increased and decreased, respectively, across different play contexts. The results of this analysis are strikingly consistent. We observe that users who become more diverse do so by curating and listening to their own library playlists, while decreasing their relative consumption of algorithmically-driven content. Programmed streaming, shown in orange, universally goes down, while organic streaming, shown in blue, almost universally goes up. Thus, while programmed content (such as editorial and personalized playlists) has the benefit of satisfying user needs to find similar content, it seems to discourage diversity overall. Again, it appears that algorithmic recommendations are associated with lower diversity, and organic consumption is associated with higher diversity. As before, it is not clear if the effect of algorithmic recommendations on diversity is causal, but even the association is significant. Users who increase their intake of algorithmic recommendations become less diverse in their consumption. This highlights opportunities for recommendation algorithms and interfaces: there is potential to better serve the needs of diversity-seeking users.

Discussion. We analyzed how diversity changes over time and found two main results. First, the GS-score is quite stable over time, clarifying that it captures an inherent behavioral characteristic of users. Second, we explicitly analyzed users who increased their diversity over time, and found that these changes are accompanied by increases in their organic streaming and decreases in their algorithmically-influenced consumption. These lend further support to our earlier findings that using recommendation algorithms is associated with reduced consumption diversity.

6 IMPACT OF RECOMMENDATIONS FOR GENERALISTS & SPECIALISTS

We have observed strong associations between algorithmic recommendations and long-term reductions in content diversity. The

[3] We conducted a series of robustness checks and concluded that our results remain unchanged if we vary exactly how to partition users, e.g. restricting to large changes in diversity, etc.

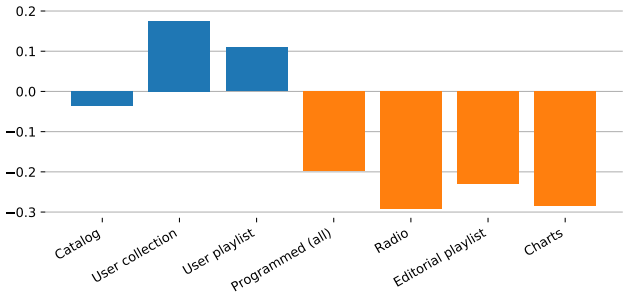

Figure 9: Log odds ratios of July 2019 streams from diversity-seekers vs. diversity-avoiders as a function of play context. Organic stream contexts are displayed in blue, programmed contexts in orange.

natural follow-up question is: What are the causal effects between algorithmic recommendations, consumption diversity, and user outcomes? This question, however, is a very difficult one to answer. For example, "diverse" consumption can range between being well-curated to being completely random, and thus straightforward experiments might fail to test the right notion of diversity. We hope and expect that a combined research effort will be devoted to uncovering these causal relationships in the future.

Here, we instead consider how recommendation systems might be tailored to cater to users' diversity preferences. Intuitively, specialists should be easier to recommend content to, since they are interested in a smaller portion of the overall content space, whereas generalists may have preferences that are not well-captured by classical recommendation approaches. We thus hypothesize that users respond differently to recommendations based on their consumption diversity. To test this, we design and deploy an online randomized experiment.

Experimental setup. We ran our experiment on a set of seven popular algorithmic playlists that are focused around particular decades of music, starting with the 1950s and ending with the 2010s ("All Out 50s" through "All Out 10s"). This set provides an ideal experimental context on Spotify, because whenever a user chooses to listen to one of these playlists, the songs are first ranked by some algorithm and then displayed to the user. In this experiment, when users navigate to one of these playlists, we randomly vary the ranking algorithm used to order the songs within the playlist and subsequently measure the effectiveness of the ranking. By comparing how ranking algorithms differentially affect generalists and specialists, we can shed light on our question.

For this experiment, we only consider users using the free version of the product, since the experience is uniform and simple for these users: when they start streaming from the playlist, songs are chosen randomly from the top k songs in the ranking. Thus, our ranking algorithms are essentially being evaluated on their ability to place relevant songs in the top k songs in their ranking. This removes potential confounds stemming from user choice and focuses our attention on the ranking itself.

To train our ranking algorithms, we use a dataset consisting of interaction information from a random sample of 4 million users and 6.6 million user sessions with the playlists considered. For each user-playlist interaction session, we store song listens and skips as outcome metrics, and we extract user-level (e.g. user embedding vector, region), song-level (e.g. song embedding vector, genres), and interaction-level (e.g. the similarity between user and song vectors) features to train our ranking algorithms.

Song Ranking Approaches. Given a playlist p, our goal is to rank a set of songs from the overall pool of songs T_p for the playlist p and display them to the user according to this ranking. To investigate whether different users can be satisfied by different recommendation strategies, we employed three approaches for ranking songs for users:

(1) **Popularity ranker:** The first ranking algorithm is the popularity ranking, where we simply rank songs by popularity in descending order. Given a large pool of songs for a given playlist ($t \in T_p$), we select the k most popular songs to show to the user. Based on our analysis presented above, we hypothesize that ranking based on popularity would be more satisfying to generalists than to specialists, since generalists are less tied to a particular point in the music space.
(2) **Relevance ranker:** The second ranking algorithm is the relevance ranking, where we simply rank songs by their relevance to the user. Given a large pool of songs and a given user u, we select the k most relevant songs, where "relevance" is measured as the cosine similarity score between the user and song vectors in the music embedding. We hypothesize that ranking based on relevance is very effective, and disproportionately so for specialist users, since they have indicated stronger preferences for similar music.
(3) **Learned ranker:** The third ranking algorithm is a model learned based on user preferences. We train a neural regression model that scores each song for a given user based on user-level, song-level, and interaction-level features. The user-level features are the user's country, the user's affinity for various genres, and their vector representation in the song embedding; the song-level features are its popularity, its genres, and its vector representation; and the interaction-level features are the cosine similarity between the user and song vectors, and the user's affinity for the song's main genre. We train a five-layer, fully-connected neural network on a random sample of 6 million user-song interactions with song completion as the objective. We use cross entropy loss, which penalizes the distance between the label (y_i) and the model's output (p_i), i.e. $L = \Sigma_i y_i \log(p_i)$. Compared to the popularity and relevance rankers, we hypothesize that this model will better reflect the relationship between user characteristics and music content, and thereby be able to learn user preferences based on historic song streaming information.

The three rankers approach song ranking differently, and capture varied aspects of user preferences. This enables us to investigate how effective these approaches are with specialists and generalists.

Online A/B Test Results. We conducted a one-week test on a random sample of 540,000 users subscribed to the free version of the service who interacted with the *All Out* playlists. When a free

Comparison	User Type	Song Streams	Song Skips
Relevance over Popularity	Generalists	+10.03%	+4.71%
	Specialists	+25.66%	+2.89%
Learned over Relevance	Generalists	+1.82%	+0.90%
	Specialists	+1.30%	−9.76%

Table 1: Relative performance of the different ranking algorithms in our online experiment.

user decides to listen to an All Out playlist, an algorithmic ranker first selects a set of 70-100 songs out of the bigger pool of over 400 songs to create the corresponding playlist for the user. The users were randomly split into three test buckets corresponding to which ranker was used to order the songs. Table 1 presents results from the experiment comparing the three ranking approaches on two metrics: (i) song streams, and (ii) song skips, where a skip denotes that the user decided to skip the song. The *song streams* metric serves as a proxy for user satisfaction, and *song skips* serves as a proxy for user dissatisfaction. Higher values of song streams and lower values of song skips are better.

First, we observe that ranking by relevance is far better than ranking by popularity. The relevance ranker outperforms the popularity ranker for both user groups, specialists and generalists, which highlights the effectiveness of recommending songs based on similarity. We also observe that this increase in streaming comes at the cost of more skipping; however, the amount of additional streams is much higher than the amount of additional skips for the relevance ranker.

Second, comparing the performance differences across specialists and generalists, we observe that sorting by relevance is disproportionately effective for specialists — the gain in using relevance over popularity is over twice as high for specialists than it is for generalists. This is in line with our hypothesis, since specialists are more concentrated in music space. Generalists, on the other hand, still prefer relevance over popularity, but by a smaller margin than specialists. Furthermore, the increase in song skips is higher for generalists (4.7%) than for specialists (2.8%), which highlights that the trade-off between increased streams at the cost of increased skips is more prevalent for generalists. These findings suggest that while specialists prefer relevance-based over popularity-based recommendations, generalists do not exhibit such a strong preference for relevance over popularity.

Finally, comparing the relevance ranker with the learned ranker, we observe a slight increase in song streams for both user types. Additionally, specialists experience a 9.8% reduction in skips for the learned ranker. This highlights that the trained model is better at learning user preferences for specialists. However, these results also demonstrate that simply sorting by relevance is competitive with a model learned from many useful features and trained on an appropriate loss function. This further underscores the fact that algorithmically recommending content based on relevance is an effective strategy for meeting short-term user needs.

In summary, these results show that consumption diversity is a useful signal in understanding how users will respond to recommendations. In particular, classical recommendation models perform much better for specialists than for generalists. This hints at the need to re-think recommendation strategies for generalists, and motivates the development of more sophisticated *diversity-aware ranking methods* that take a user's content diversity into account.

7 DISCUSSION

Understanding how algorithmic recommendations are associated with consumption diversity is a central question for online platforms. In this work, we studied this effect on content diversity on Spotify, and on key user metrics in turn. To do this, we employed a diversity measure that captures fine-grained distinctions between item similarities using a high-fidelity song embedding. We found this diversity measure to be stable over long periods of time, indicating that it captures inherent characteristics of user behavior.

In a randomized experiment, we found that algorithmically ranking songs by relevance to the user is very effective for satisfying short-term user needs (more songs are streamed). In our analyses, however, we observed that recommendation algorithms are associated with reduced diversity in listening. Furthermore, when users become more diverse over time, they do so by reducing their algorithmically-driven consumption and increasing their user-driven consumption. We also discovered that key user metrics, conversion to subscriptions and retention on the platform, are very strongly associated with greater content diversity. These effect sizes were very large, ranging up to 35 percentage point differences between generalists and specialists at the same activity level.

These results have deep implications for online platforms beyond the one studied here. A main contribution of our work is to illuminate a balancing act that online platforms in general must perform: simultaneously recommending content to users that they are likely to enjoy in the short term while ensuring that users can explore the content space and remain satisfied in the long term. Our work suggests that there are risks to algorithmic over-specialization in online platforms, and to measuring the effectiveness of recommender systems too narrowly. It also motivates the need to develop diversity-aware recommendation methods that can reap the short-term benefits of using relevance for recommendations while simultaneously optimizing for the long-term rewards of serving users with diverse content.

REFERENCES

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in collaborative filtering systems. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 153–160.
[2] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
[3] Isaiah Berlin. 1953. *The hedgehog and the fox.* Weidenfeld & Nicolson.
[4] Elisabeth Bublitz and Florian Noseleit. 2014. The skill balancing act: when does broad expertise pay off? *Small Business Economics* 42, 1 (2014).
[5] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446
[6] Hannes Datta, George Knox, and Bart J Bronnenberg. 2017. Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery. *Marketing Science* 37, 1 (2017), 5–21.
[7] Munmun De Choudhury, Scott Counts, and Mary Czerwinski. 2011. Identifying relevant social media content: leveraging information diversity and user cognition. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia.*

ACM, 161–170.

[8] Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. 2012. Churn Prediction in New Users of Yahoo! Answers. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. ACM, New York, NY, USA, 829–834. https://doi.org/10.1145/2187980.2188207

[9] Georges Dupret and Mounia Lalmas. 2013. Absence Time and User Engagement: Evaluating Ranking Functions. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*. ACM, New York, NY, USA, 173–182. https://doi.org/10.1145/2433396.2433418

[10] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80, S1 (2016), 298–320.

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

[12] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.* 6, 4, Article 13 (Dec. 2015), 19 pages. https://doi.org/10.1145/2843948

[13] Eduardo Graells-Garrido, Mounia Lalmas, and Ricardo Baeza-Yates. 2016. Encouraging Diversity- and Representation-Awareness in Geographically Centralized Content. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, New York, NY, USA, 7–18. https://doi.org/10.1145/2856767.2856775

[14] Henning Hohnhold, Deirdre O'Brien, and Diane Tang. 2015. Focus on the Long-Term: It's better for Users and Business. In *Proceedings 21st Conference on Knowledge Discovery and Data Mining*. Sydney, Australia. http://dl.acm.org/citation.cfm?doid=2783258.2788583

[15] Rong Hu and Pearl Pu. 2011. Helping Users Perceive Recommendation Diversity.. In *DiveRS@ RecSys*. 43–50.

[16] David Hubbard, Benoit Rostykus, Yves Raimond, and Tony Jebara. 2019. Beta Survival Models. *CoRR* abs/1905.03818 (2019).

[17] Mounia Lalmas, Janette Lehmann, Guy Shaked, Fabrizio Silvestri, and Gabriele Tolomei. 2015. Promoting Positive Post-Click Experience for In-Stream Yahoo Gemini Users. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1929–1938. https://doi.org/10.1145/2783258.2788581

[18] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off Between Relevance, Fairness, and Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 2243–2251.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*.

[21] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*. ACM, 677–686.

[22] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[23] Minsu Park, Ingmar Weber, Mor Naaman, and Sarah Vieweg. 2015. Understanding musical diversity via online social media. In *Ninth International AAAI Conference on Web and Social Media*.

[24] Lijing Qin and Xiaoyan Zhu. 2013. Promoting diversity in recommendation by entropy regularizer. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

[25] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are "Good"? 595–604. https://doi.org/10.1145/3331184.3331215

[26] Kurt C Stange. 2009. The generalist approach. *The Annals of Family Medicine* 7, 3 (2009).

[27] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. 2012. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences* 109, 16 (2012), 5962–5966.

[28] Isaac Waller and Ashton Anderson. 2019. Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms. In *The World Wide Web Conference*. ACM, 1954–1964.

[29] Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. 2017. Returning is Believing: Optimizing Long-term User Engagement in Recommender Systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 1927–1936. https://doi.org/10.1145/3132847.3133025

[30] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 13–22.

[31] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. 2010. The impact of YouTube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 404–410.

[32] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515.