How Intention Informed Recommendations Modulate Choices: A Field Study of Spoken Word Content

Longqi Yang
Cornell Tech, Cornell University
ylongqi@cs.cornell.edu

Michael Sobolev
Cornell Tech, Cornell University
michael.sobolev@cornell.edu

Yu Wang
Himalaya Media
yu.wang@himalaya.com

Jenny Chen
The City University of New York
jennychen0905@gmail.com

Drew Dunne
Cornell University
asd222@cornell.edu

Christina Tsangouri
The City University of New York
christinatsangouri@gmail.com

Nicola Dell
Cornell Tech, Cornell University
nixdell@cornell.edu

Mor Naaman
Cornell Tech, Cornell University
mor.naaman@cornell.edu

Deborah Estrin
Cornell Tech, Cornell University
destrin@cornell.edu

ABSTRACT

People's content choices are ideally driven by their intentions, aspirations, and plans. However, in reality, choices may be modulated by recommendation systems which are typically trained to promote popular items and to reinforce users' historical behavior. As a result, the utility and user experience of content consumption can be affected implicitly and undesirably. To study this problem, we conducted a 2×2 randomized controlled field experiment (105 urban college students) to compare the effects of intention informed recommendations with classical intention agnostic systems. The study was conducted in the context of spoken word web content (podcasts) which is often consumed through subscription sites or apps. We modified a commercial podcast app to include (1) a recommender that takes into account users' stated intentions at onboarding, and (2) a Collaborative Filtering (CF) recommender during daily use. Our study suggests that: (1) intention-aware recommendations can significantly raise users' interactions (subscriptions and listening) with channels and episodes related to intended topics by over 24%, even if such a recommender is only used during onboarding, and (2) the CF-based recommender doubles users' explorations on episodes from not-subscribed channels and improves satisfaction for users onboarded with the intention-aware recommender.

KEYWORDS

User intention; Recommendation; Field study; Podcast

ACM Reference Format:
Longqi Yang, Michael Sobolev, Yu Wang, Jenny Chen, Drew Dunne, Christina Tsangouri, Nicola Dell, Mor Naaman, and Deborah Estrin. 2019. How Intention Informed Recommendations Modulate Choices: A Field Study of Spoken Word Content. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313540

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313540

1 INTRODUCTION

People make choices about what content to consume on a day-to-day basis, such as selecting music or podcasts to listen to and identifying articles to read. Ideally, users make choices according to their intentions, plans, and aspirational preferences [31]. For example, a person may use a search engine to find an article or a piece of music with a specific title. However, choosing content on real world platforms is more complex, in part because users' choices are often sub-optimal and focus on the short-term [32], and these immediate choices then get reinforced by recommendation systems that expose users to biased sets of items. The bias of item presentations mainly comes from two sources: (1) recommenders often hold a partial and skewed view of users' preferences that are learned from observational interaction records [40, 51], and (2) recommenders are typically subject to popularity bias [51], which hinders the system from presenting relevant items. When subject to regular exposure to these biased item sets, users' original intention-related choices may be altered — on the one hand, users may explore more content, on the other hand, they may end up consuming trendy but mediocre or irrelevant content with low utility to them.

Prior recommendation systems literature was focused on *how many* [21, 41] and *what* [16, 46] items people choose but rarely addressed *why* people choose them. For example, are the choices a result of people's original intentions or their interactions with recommendation systems? In other words, how recommendations may change users' consumption from what they might have chosen, or aspire to choose? These under-explored questions are critical for recommender systems to listen to users and support users' needs, intentions, and desires [14, 26].

In this paper, we investigate the above mentioned questions, specifically, *how intention informed recommendations modulate users' choices, as compared to intention agnostic systems?* To answer this question, we designed a randomized controlled field study [27] in the domain of podcasts, where we leveraged the **topics of interest** as an indicator of user intentions. The field study is a 2×2 experiment where two factors are two stages of app usage, and two interventions within each factor are different recommendation algorithms. First, during onboarding, users expressed their topics of interest and subscribed to a set of podcast channels through a

website, where we compared a popularity-based recommender to a recommender that takes into account users' intentions (**intention-aware recommender**) in presenting channel candidates. Then, during the remainder of their participation (app usage in the field), users used a customized commercial mobile app without constraint. During this stage of the study, we compared a subscription-based recommender to a Collaborative Filtering (CF)-based recommender in populating the home feed that users interacted with everyday. Finally, participants were invited to complete a post-study survey where they gave ratings in terms of four aspects of satisfaction.

We choose podcasts as the study domain for two main reasons. First, traditional podcast content consumption is typically based on subscriptions and therefore clearly relates to user intentions — users subscribe to RSS feeds of the channels they plan to listen to and then regularly consume released episodes from those channels. Second, recommendation systems for podcasts is of growing importance but currently under-explored (Section 2.4).

We conducted the study with 105 urban college students, which consists of 52.5 hours of one-by-one onboarding, four weeks of field experiments with daily communications and weekly reminders, and a follow-up survey with each participant. Our key findings include:

- **Effects of onboarding recommendations:** Compared to commonly used popularity-based ranking of channels, intention-aware recommendations for user onboarding significantly raised the ratio of channel subscription and episode listening that were aligned with users' topic-wise aspirations (improvements: 72.1% and 36.5% in terms of subscriptions at onboarding and in the field, and 24.9% in terms of listening time).
- **Effects of field recommendations:** Home feeds that were populated by the CF-based recommendations significantly increased the ratio of episode listening to not-subscribed channels by 127.5%, as compared to the traditional home feeds that were filled purely with episodes from subscribed channels.
- **Interaction effects:** User satisfaction was jointly affected by the recommendation algorithms used in the two stages — the CF-based recommender improved satisfaction for users onboarded with the intention-aware recommender, whereas for others, the CF-based recommender was shown to have negative effects.

These findings suggest that recommendations can implicitly but significantly modulate users' intention-related choices — they can encourage or discourage users to pursue their aspirations and intentions. The positive modulation effects can be leveraged to support healthy behavior and benefit an individual's aspired long-term growth, as discussed in Section 5. Also, our study suggests a hybrid form of recommender for podcasts and subscription-based media, consisting of an intention-aware recommender for onboarding and a CF-based recommender for home feed generation. Together, these recommenders support user aspirations, encourage content exploration, and provide satisfying user experiences.

Through our study, we also find that signals regarding the utility of user engagement is not reflected in intention-agnostic statistics (e.g., total listening time and total number of subscriptions) that are commonly employed to understand user experiences (Section 4.1). This highlights the importance of using metrics conditioned on individual intentions to complement the understanding of recommendation effects (Section 5.5).

2 RELATED WORK

Our work builds on and contributes to four lines of research: (1) studying the effects of recommendations, (2) investigating user intentions in using intelligent systems, (3) building recommendation systems beyond optimizing for accuracy, and (4) analyzing and leveraging spoken word content on the web.

2.1 Effects of recommendations

Recommendation systems were shown to increase traffic and user engagement [41], but it was recently recognized in the research community that they can also significantly affect end users' behavior and the structure of a society. Prior work in studying the effects of recommendations mainly focused on the social network structures [11, 42, 43] and the filtering bubble problem [3, 7, 16, 20, 34, 36]. For example, the former line of research demonstrated that introducing friend-based recommendations into social network platforms exacerbates popularity bias (i.e., rich gets richer) [43] and establishes an algorithmic ceiling for minority groups of users [42]. The latter line of research illustrated how recommendations affect users' information exposure by either limiting users' information exposure to a biased scope [16, 36] or enabling users to explore ideologically diverse opinions [3, 16]. As a result, consumers and users may be fragmented [20]. Most recently, Chaney et al. [8] used a simulation to show that recommendations may lead to a homogenization of users' choices. For contextual-aware recommendations, prior work has raised the concern about their potential alternation of users' content consumption context [1].

Although prior research has revealed significant effects of recommendations in the global and individual levels, these effects are user intention-agnostic and are measured and interpreted from system designers' and experts' perspectives. It is unknown whether recommendations' effects are aligned with or deviated from users' own intentions. Our study measures effects from users' angle and contributes findings that are critical to the future user-centric recommendation systems [14, 21, 26].

2.2 User intentions

Understanding and leveraging user intentions is an important theme in designing intelligent systems. For example, in the context of web search, previous research [12, 29, 39, 44, 48] discovered diverse user intents in using search engines [44], e.g., for the same query, users may look for different information. The understanding and prediction of users' intents is an essential component for personalized search experience [12, 44, 48]. In other domains, such as arts and fashion [10], and psychology [15, 38], user intentions were also investigated and were shown to be predictable from behavior logs [10]. In the context of recommendation systems, prior work leveraged interactive systems to elicit signals about user intentions, such as conversation-based [25], survey-based [52], and critique-based [9] systems. Recently, Tomkins et al. [46] presented a system that recommends appropriate products for users who intend to maintain a sustainable behavior.

However, when incorporating users' intentions, the intelligent systems were often evaluated against intention-agnostic metrics, such as click-through rate, dwelling time, and etc., which do not answer the questions of how these systems alter users' choices from

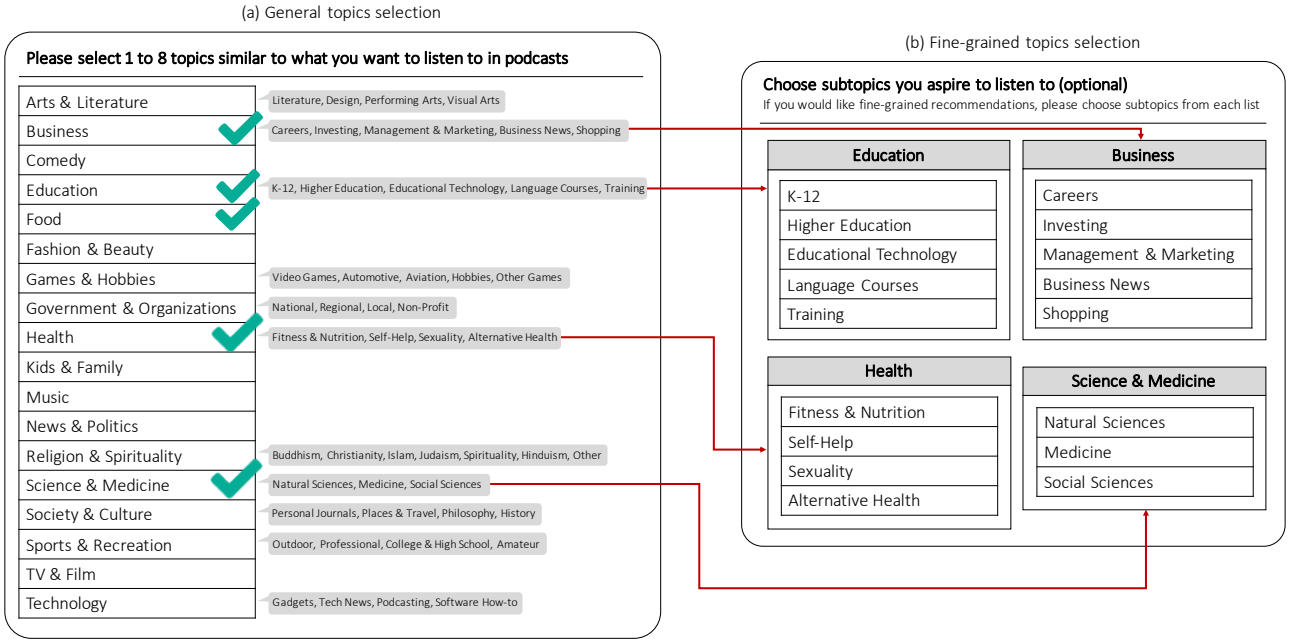


Figure 1: The web user interface designed for participants to indicate their topic-wise intentions. Participants first (a) select up to eight general topics they want to listen to and then (b) optionally select fine-grained topics. The topics are defined using podcast categories in iTunes.

what they might have chosen, and how much of users' intentions were satisfied. As argued by Knijnenburg et al. [26] and Ekstrand et al. [14], future recommenders should be able to satisfy *what users want* and *what they want to achieve*. Our research takes a step further and investigates how intention informed recommenders would in turn affect users' intention-related choices, which closes the feedback loop between choices and recommendation systems.

2.3 Recommendations beyond accuracy

Our work contributes to the increasing recognition and interests in building recommender systems for objectives beyond accuracy [14, 26, 50], such as diversity [21, 53], fairness [13], novelty [45], sustainability [46], and unbiasedness [40, 51]. These objectives were motivated by the observation that recommender systems purely optimized for accuracy may have various negative effects on end users, as reviewed in Section 2.1. These enable recommendations to serve users with different needs and intents. Nevertheless, similar to the limitations discussed in Section 2.2, prior work optimized these systems using hand-crafted or expert-designed metrics (such as categorical accuracy [13]), which may or may not be aligned with users' intentions and goals. Our study reveals the extent to which users' choices are related to their intentions, which can be used to inform future design of recommenders beyond accuracy.

2.4 Web spoken word content

We conducted the field study in the domain of spoken word content (podcasts) — an emerged channel for information and entertainment [37]. In the web community, prior research was mainly focused on building web search engines [5, 17, 19, 33, 35], which index podcast metadata and audio files so as to match given text queries to audio. However, there has been very little work addressing the podcast recommendation problem. The only work we recognized was from Tsagkias et al. [47] that predicted users' podcast preference using hand-crafted preference indicators, which can hardly be applied in the wild because of the heterogeneity of users and content. With the interests from major media companies to serve podcasts, research is needed to build recommenders that better expose users to content beyond passive receiving. Our study contributes a hybrid form of podcast recommender that serves users' intentions, encourages exploration and results in higher user satisfaction. Our paper also presents key guidelines for the design of podcast recommenders, which can be applied to other subscription-based media platforms as well.

3 STUDY DESIGN

Our study design included collecting consumption intentions from all participants and randomly assigning participants to four independent experimental conditions. This design allowed us to conduct within-subject comparisons to understand the discrepancy between users' content consumption and intentions, and between-subject comparisons to measure the effects of different recommendations. Specifically, our study consisted of two phases: an **one-by-one video onboarding** (30 minutes) and a **field study** (four weeks), corresponding to the prominent settings under which podcast listeners are exposed to recommendations in the wild (i.e., when they first begin to use an application, and during the daily usage). Our design for both phases of the study allowed participants to

interact with recommendations naturally. During onboarding, participants were instructed to subscribe to a set of podcast channels they wanted to listen to from a ranked list of candidates; and in the field, participants were provided with a customized commercial podcast mobile app (available on both Android and IOS) to listen to podcasts naturally and without study constraints. The experiment used a full 2×2 factorial design where the two factors were recommendations made in the two study stages, i.e., onboarding (**ONB**) and field (**FIE**) recommendations, and the two interventions within each factor were specific algorithms that presented channels or episodes in different orders. Below, we describe detailed design of each phase.

3.1 Onboarding (ONB)

We onboarded participants one-by-one using remote video conferencing software. Participants were instructed to complete two tasks during onboarding: (1) indicate their topic-wise intentions and interests, and (2) subscribe to channels that they want to listen to in the field. Participants were directed to use a website we developed to complete both tasks.

Indicating topic-wise intentions. We collected participants' listening aspirations in the form of podcast topics (Fig. 1). This topic selection approach is a common practice adopted by major content platforms (e.g., Pinterest and Medium) to elicit user preferences during onboarding. We used podcast categories defined by iTunes[1] as topics, which consists of two levels: general and fine-grained. Through the website, participants first picked 1-8 general topics (Fig. 1-a), and then optionally chose fine-grained topics within the selected general ones (Fig. 1-b). To help participants make sense of general topics, fine-grained topics were shown side-by-side.

Subscribing to channels. Each participant was then asked to subscribe to up to ten podcast channels from a list of recommendations (Fig. 2). The recommendation list was subject to the control or experimental setting, according to the participants' assignments in the study. The control intervention implemented a standard user onboarding strategy that ordered channels based on their popularity on iTunes (**POP**) (Fig. 2-a), whereas the experimental intervention ranked channels by the degree to which they related to participants' aspirations[2] (**ASP**) (Fig. 2-b). The relevance of a channel c for a user u is characterized by a score $s(c|u)$ calculated as follows.

$$s(c|u) = |\mathcal{S}_c \cap \mathcal{A}_u| + \mathbb{1}\left[m_c \in \mathcal{A}_u\right] \tag{1}$$

where $\mathcal{S}_c$ is the set of topics (general and fine-grained) that the channel c belongs to, m_c is the channel's primary topic ($m_c \in \mathcal{S}_c$), and $\mathcal{A}_u$ is the set of topics that users aspired to listen to. Both $\mathcal{S}_c$ and m_c were scraped via iTunes RSS API. As shown in the above equation, when calculating $s(c|u)$, we placed an additional weight on the primary topic.

For both groups, participants were instructed to browse the website freely and make decisions at any point of time. To prepare channels for recommendations, we scraped all top channels returned by the iTunes RSS feed, and made a join with our podcast database. Eventually, 2231 channels were used.

[1]Podcast directory: https://itunes.apple.com/us/genre/podcasts/id26?mt=2
[2]To break ties, channel popularity on iTunes was used.

3.2 Field Study (FIE)

After onboarding, each participant was provided with a podcast mobile app and a pre-registered account to use for four weeks in the field. The app was pre-loaded with the channels for which the participant subscribed during onboarding. We customized a popular commercial app for our study. The app (shown in Fig. 3 and Fig. 4) has three main pages: (1) a **home** page (Fig. 3-b,c) that presented a personalized list of new podcast episodes, and is the default page when opening the app, (2) a **library** page (Fig. 3-a) that showed the channels to which a user has subscribed, and (3) a **discover** page (Fig. 4) that listed channels based on categories and popularity, which were not personalized. In addition, the app allowed users to directly search for content (through the icon at the top-right corner), and users can also consume episodes from a channel's page (by clicking on the channel's thumbnail).

Similar to onboarding, the field intervention was applied to recommendations on the mobile home page, which chronologically listed episodes from a personalized set of channels and was refreshed daily for newly-released episodes. For the control group, the personalized set contained channels to which a user has subscribed (**SUB**); whereas for the experimental intervention, the set additionally mixed five not-subscribed channels (**MIX**). These channels were retrieved by a matrix factorization based recommendation model, which we built as follows:

- **Dataset collection**. We scraped the most recent 500 reviews of 29K popular podcast channels on iTunes to train a recommendation model. In order to conduct recommendations based on users' channel subscriptions, which are binary signals, we disregarded rating scores and treated iTunes reviews as positive-only feedback. The final training dataset contained 702K user-channel interactions from 137K iTunes users.
- **Recommendation model**. We used OpenRec [50] to build a Weighted Regularized Matrix Factorization (WRMF) [22] based recommender, which is a representative implicit-feedback-based recommendation model and is optimized to minimize the following objective function:

$$\min_{x_*, y_*} \sum_{u \in \mathcal{U}, i \in \mathcal{I}} w_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda \|\Theta\|^2 \tag{2}$$

  where Θ is a set of model parameters, x_u and y_i are latent factor representations for user u (among all iTunes users $\mathcal{U}$) and channel i (among all iTunes channels $\mathcal{I}$) respectively, and p_{ui} is a binary indicator for user preferences ($p_{ui} = 1$ if user u subscribed to channel i, and $p_{ui} = 0$ otherwise). In addition, WRMF uses w_{ui} to control models' confidence levels on p_{ui}. We set w_{ui} such that $w_{ui} = 1$ if $p_{ui} = 1$, and $w_{ui} = 0.01$ otherwise. These parameter settings achieved the best validation results in our dataset. When applying the WRMF model, we discarded x_u since it corresponds to users from iTunes, and fixed trained channel representations y_i. For a participant u', we derived an analytic expression of the optimal user representation $x_{u'}$ by differentiating the objective function (eqn. 2):

$$x_{u'} = \frac{1}{\lambda + \sum_i w_{u'i} y_i^T y_i} \sum_i w_{u'i} p_{u'i} y_i \tag{3}$$

- **Not-subscribed channel retrieval**. For any participant u' in the experimental group, we retrieved the top 5 not-subscribed

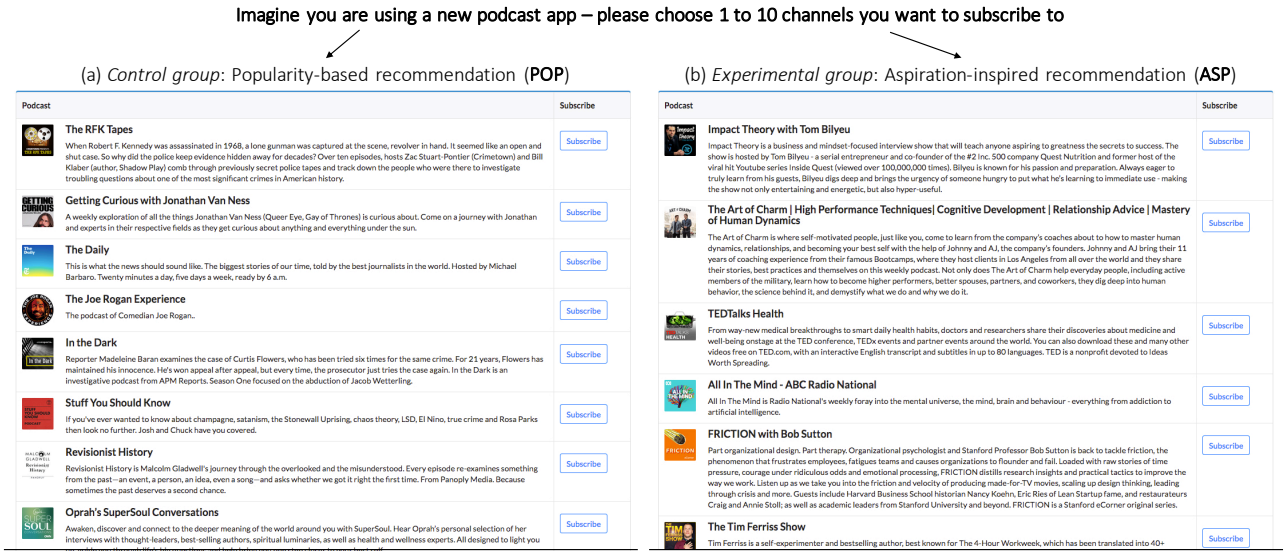
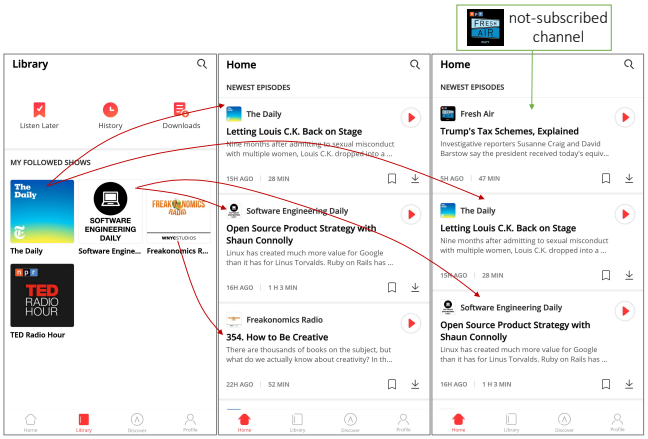


Figure 2: The web user interface designed for participants to subscribe to channels during onboarding. The interface presented a list of podcast shows, and participants were instructed to subscribe to up to ten of them. For the control group (POP), channels were ordered by their popularity on iTunes, whereas for the experimental group (ASP), the ordering was determined by channels' alignment to participants' topic-wise intentions. Both groups shared the same set of candidate content.

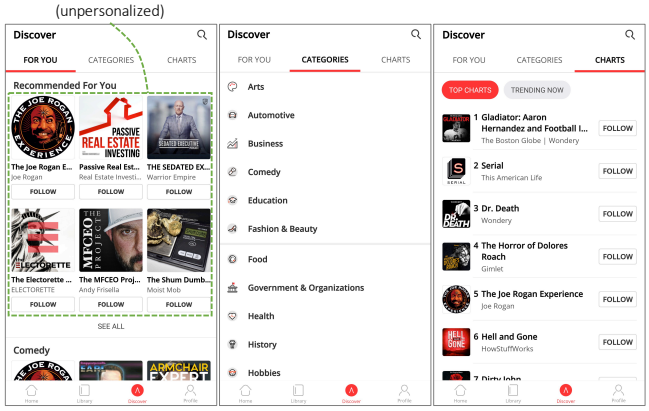


(a) Library showing subscribed channels

(b) *Control group*: episodes from subscribed channels (**SUB**)

(c) *Experimental group*: episodes from mixed channels (**MIX**)

Figure 3: The library and home pages of the customized podcast mobile app. The library page showed the channels to which a user has subscribed, and the home page chronologically presented a list of episodes. For the control group (SUB), the episodes were retrieved from subscribed channels, whereas for the experimental group (MIX), those episodes were mixed with the ones from selected not-subscribed channels based on a CF recommendation model.



(a) Randomly sampled popular channels

(b) Categorical listing of channels

(c) Popular/Trending channels

Figure 4: The discover page of the customized podcast mobile app. The page grouped channels into topic-wise categories and presented a trending chart that ordered channels according to their popularity on iTunes. This page allowed users to readily explore and subscribe to new channels.

channels that had the highest dot product scores (i.e., $x_{u'}y_i, i \in \{i|p_{u'i} = 0 \wedge i \in \mathcal{I}\}$). Although the recommendation model was fixed throughout the study, retrieved channels were adaptively updated whenever participants subscribed to new shows.

3.3 Post-study Survey

After participants finished the 4-weeks field study, we conducted a post-study survey through email to elicit user satisfaction. The survey questions follow a template: "How satisfied were you with

Total number of participants: 105, **unreported:** 26	
Gender:	Female: 50 Male: 29
Age (years):	Max: 43 Min: 17 Mean: 21
Device:	IOS: 49 Android: 30
Major:	Computing and Information Science: 20 Arts & Sciences: 22 Life Sciences: 10 Medicine: 2 Business: 16 Engineering: 9

Table 1: Participants' demographic information including gender, age, primary mobile device, and college major.

___?", and the aspects we surveyed include *the app*, *the experiment*, *your current podcast channel subscriptions*, and *the home feed in the app*. For each question, participants were instructed to give a likert-scale rating (i.e., *not at all satisfied*, *slightly satisfied*, *neutral*, *very satisfied*, and *extremely satisfied*).

3.4 Participant Recruitment

We recruited 105 full-time undergraduate students who were studying in New York City and were from diverse background. The demographic information of the participants is summarized in Table. 1. Participants were compensated with $30 after completing both phases of the study. To encourage app usage in the field, we provided an additional $20 bonus for those who used the mobile app for at least five days a week, and reminded all participants to listen to new episodes weekly. Finally, participants were randomly assigned to one of the 2×2 conditions (POP-SUB: 25, POP-MIX: 26, ASP-SUB:29, ASP-MIX:25), and two research personnel who were blind to condition assignments managed and executed participants onboarding and the field study. The study was approved by the Institutional Review Board (IRB) under the protocol #1507005739.

4 STUDY RESULTS

Our study recorded the choices that participants made at onboarding and in the field including both channel subscriptions and episode listening. In addition, we recorded satisfaction ratings that participants gave to questions in the final survey. Eventually, 99 out of 105 participants completed the study (POP-SUB: 24, POP-MIX: 23, ASP-SUB:28, ASP-MIX:24). We summarize and present our study results in four dimensions: general usage patterns (Section 4.1), choices related to topic-wise intentions (Section 4.2), exploratory choices (Section 4.3), and user satisfaction (Section 4.4).

4.1 General usage patterns

To understand the usability and user experience with our podcast content platform, we investigate a type of commonly used metrics, user activity level [28]. We count the number of subscriptions that each user made in the field, and the amount of time that each user spent listening to episodes. The distributions of these measures over users are illustrated in Fig. 5. Overall, participants were fairly active in using the mobile app in the field with 8.8 average number of subscriptions and 4.58-hour average listening time. Participants' activity level is also distributed within a range and has rare outliers (Fig. 5-b,c). In Fig. 5-a, we also plot the distribution of the number of onboarding subscriptions, which is shown to spread from one to ten (maximum allowance) with an average of 7.4. To test whether two experimental factors affect the three measures in Fig. 5, we conduct a general nonparametric factorial analysis using the Aligned Rank Transform (**ART**) [49] (by treating the three measures as responses). We use ART because our study contains more than one factor, and all the measures are not normally distributed over users[3]. In the rest of this paper, if not specified, the ART is used to conduct statistical significance tests (notations: $***$: $p < 0.001$, $**$: $p < 0.01$, $*$: $p < 0.05$). The ART reports no significant effect from ONB, FIE, or ONB×FIE for all the three measures. However, as shown in Section 4.2 and 4.3, ONB and FIE have significant effects on users' podcast consumption patterns, although they are not captured in the general user activity measures. We discuss the limitations of these traditional measures in Section 5.5.

In addition to aggregate users' activities (subscriptions and listening time) on a per-user basis, we also cluster the activities into hour of day (Fig. 6-a), day of week (Fig. 6-b), and distinct channels (Fig. 7). Temporal distributions of listening instances (Fig. 6) reveal several diurnal and weekly listening patterns, such as decreased listening during night and over weekends. However, no statistical evidence shows significant effects of experimental factors on these temporal patterns. Regarding the channel-wise user activity distributions (Fig. 7), they demonstrate that (1) during onboarding (Fig. 7-a), participants' channel subscriptions manifested significant popularity bias under the POP treatment, i.e., the majority of user subscriptions were concentrated on a small number of channels, whereas under the ASP treatment, subscriptions were spread out to more channels and tended to be uniformly distributed, and (2) in the field, users interacted with a broader set of podcast channels than during onboarding, but both experimental factors have no significant effect on the number of interactions that each channel received. In the Appendix, we additionally visualize top channels subscribed by participants.

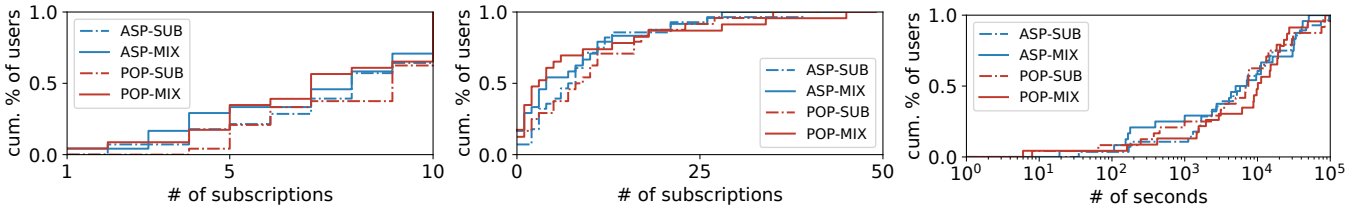
4.2 Choices related to topic-wise intentions

The distribution of the topics that participants intended to consume is shown in Fig. 8, which shows the diversity of the topics of interest chosen by participants — the intended topics in the population were spread across 53 distinct general and fine-grained categories, and most of the topics were selected by less than half of the population. Such a wide range of selected topics is partially attributable to the diverse background of our recruited participants (Table. 1). To show how users' choices related to topic-wise intentions may be modulated by two stages of recommendations, we define a topic-wise intention ratio $r_{\text{topic}}(c|u)$ of a channel c for the user u as follows:

$$r_{\text{topic}}(c|u) = \frac{|\mathcal{S}_c \cap \mathcal{A}_u|}{|\mathcal{S}_c|} \tag{4}$$

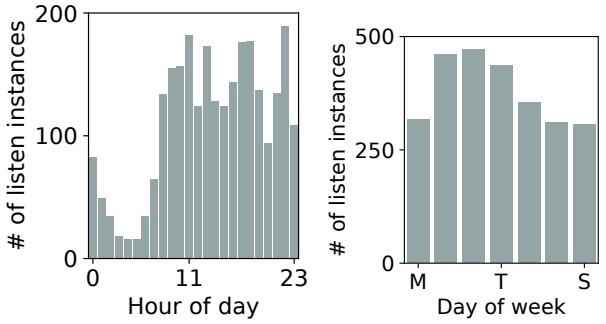
where we use the notations from eqn. 1. The value of $r_{\text{topic}}(c|u)$ corresponds to the proportion of a channel's content that aligns with a user's intended topics. Then using $r_{\text{topic}}(c|u)$, we calculate the average alignment of a user u's subscriptions, $P^{\text{sub}}_{\text{topic}}(u)$, as the

[3]The normality test is conducted via the Shapiro-Wilk normality test



(a) Onboarding subscriptions. (b) Field subscriptions. (c) Listening time.

Figure 5: The cumulative distributions of users over number of subscriptions and listening time. These figures show the extent to which participants were actively subscribing and listening to podcasts throughout the study. A vertical line in these figures represents a group of users with a similar activity level. We note that these commonly-used aggregated measures are not statistically different across the four groups. In other words, they do not reflect the different composition of content consumption across these groups (Section 4.2 and 4.3). These differences are critical to understand the effects of recommendations on individual growth and experience.



(a) Hour of day distribution. (b) Day of week distribution.

Figure 6: The distribution of podcast listening instances over hour of day and day of week. The aggregation is across all participants. Again we note that no statistical difference is observed across the four groups.

average $r_{\text{topic}}(c|u)$ over all followed channels $\mathcal{F}_u$, i.e.,

$$P^{\text{sub}}_{\text{topic}}(u) = \frac{\sum_{c\in\mathcal{F}_u} r_{\text{topic}}(c|u)}{|\mathcal{F}_u|} \tag{5}$$

and calculate the average alignment of a user u's listening, $P^{\text{listen}}_{\text{topic}}(u)$, as the weighted average of $r_{\text{topic}}(c|u)$ over listened channels $\mathcal{L}_u$ with the weight proportional to the listening duration d_c, i.e.,

$$P^{\text{listen}}_{\text{topic}}(u) = \frac{\sum_{c\in\mathcal{L}_u} r_{\text{topic}}(c|u)d_c}{\sum_{c\in\mathcal{L}_u} d_c} \tag{6}$$

We show the cumulative distributions of users over $P^{\text{sub}}_{\text{topic}}(u)$ and $P^{\text{listen}}_{\text{topic}}(u)$ in Fig. 9, and the 2×2 groupwise averages in Fig. 10. These graphs and corresponding ART tests demonstrate that under all scenarios, the ASP intervention significantly improves the ratio of content consumption that matches users' topic-wise intentions — during onboarding, ASP increases $\overline{P}^{\text{sub}}_{\text{topic}}$ by 72.1% (ONB:**), and in the field, ASP improves $\overline{P}^{\text{sub}}_{\text{topic}}$ and $\overline{P}^{\text{listen}}_{\text{topic}}$ by 36.5% (ONB:***) and 24.9% (ONB:*) respectively. It is worth noting that although improvements are larger at onboarding when the intervention is directly applied, ASP is shown to have significant indirect effects on users' content consumption in the field as well. The statistical test does not show significant effects from the FIE factor and the interaction (i.e., ONB×FIE).

4.3 Exploratory choices

To investigate how participants' exploratory choices were affected by recommendations, we divided their podcast listening into subscribed listening (exploitation) and not-subscribed listening (exploration). We define the exploratory ratio $r_{\text{explore}}(c|u)$ as a counterpart for $r_{\text{topic}}(c|u)$ (Section 4.2). This exploratory ratio is calculated as follows.

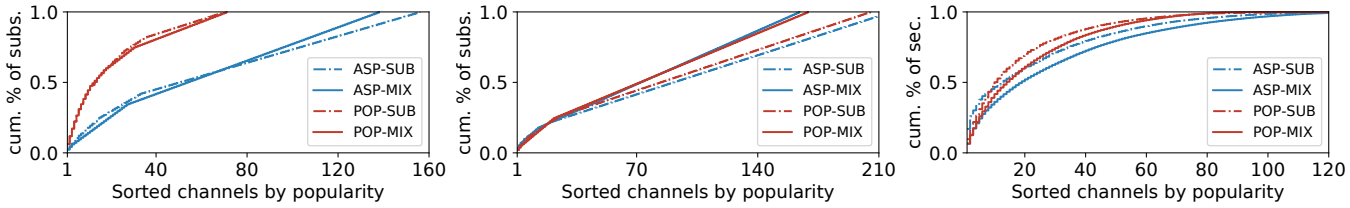
$$r_{\text{explore}}(c|u) = 1 - \mathbb{1}\left[c \in \mathcal{F}^t_u\right] \tag{7}$$

where $\mathbb{1}$ is an indicator function, and $\mathcal{F}^t_u$ is the set of channels that the user u subscribed to at time t when the channel c was consumed. Essentially, $r_{\text{explore}}(c|u) = 1$ if the channel was not subscribed when consumed, otherwise $r_{\text{explore}}(c|u) = 0$. We then substitute $r_{\text{topic}}(c|u)$ in eqn. 6 with $r_{\text{explore}}(c|u)$ and derive an exploratory measure of a user u' s listening, denoted as $P^{\text{listen}}_{\text{explore}}(u)$. From another angle, $P^{\text{listen}}_{\text{explore}}(u)$ can be viewed as the percentage of time that the user u explored new information channels.

We show the distributions of users over $P^{\text{listen}}_{\text{explore}}(u)$ and the groupwise average scores in Fig. 11. Both figures and ART statistical tests demonstrate that the MIX intervention significantly increases $\overline{P}^{\text{listen}}_{\text{explore}}$ by 127.5% (FIE:*). In other words, the MIX feeds significantly encouraged participants to explore beyond existing and potentially narrow information channels. The onboarding recommendations (ONB) and the interaction between the two stages of recommendations (ONB×FIE) do not have significant effects.

4.4 User satisfaction

Four satisfaction indicators were surveyed and reported by participants after the study was over (Section 3.3). Among 99 valid participants, 89 of them responded to our email survey (Response rate: 89.9%). Both experimental factors and their interaction do not have significant effects on whether or not a participant responded to the survey. To quantitatively analyze survey results, following



(a) Onboarding subscriptions. **(b) Field subscriptions.** **(c) Listening time.**

Figure 7: The cumulative distributions of subscriptions and listening time over channels ordered by popularity. The popularity is defined as the number of subscription (a, b) and the amount of listening (c). These figures show the extent to which participants' content consumption was concentrated on a small set of popular items. A linear line in the figure represents uniformly distributed consumption over all channels. During onboarding, the POP intervention resulted in significant popularity bias in participants' subscriptions, but in the field, no significant effect from experimental factors is observed.

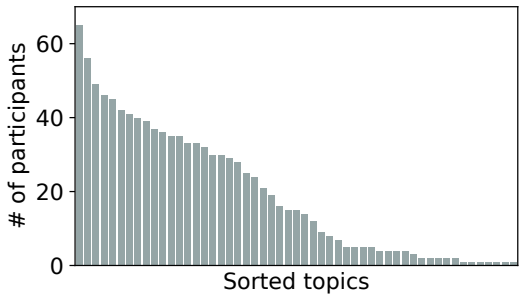


Figure 8: The distribution of user intentions over podcast topics (categories). Topics are sorted by their popularity in a descending order. Participants' intended topics were diversely spread across 53 categories, with most of the topics liked by less than half of the participants.

the common practice [21], we convert the five options in each survey question, i.e., *not at all satisfied*, *slightly satisfied*, *neutral*, *very satisfied*, and *extremely satisfied*, to 1–5 numerical ratings.

We found that satisfactions for all indicators are highly correlated. Therefore, we aggregated them into one factor by taking the average of the ratings. The distributions of the aggregated satisfaction ratings and the groupwise average values are shown in Fig. 12. Participants' satisfaction is significantly affected by the interaction between two factors (ONB×FIE: $*$); and the post-hoc differences of differences test [6, 30] confirms the effects of one factor given the other. In other words, if participants were onboarded with the popularity-based recommender, applying the CF-based recommender to populate users' home feeds significantly degraded users' satisfaction, whereas if participants were initially presented with a channel list ranked by their intentions, the CF-based recommender used in the field showed positive effects. These findings have important implications as to how the reinforcing nature of recommendations may improve or degrade utility and user experience, as discussed in Section 5.

5 IMPLICATIONS AND DISCUSSIONS

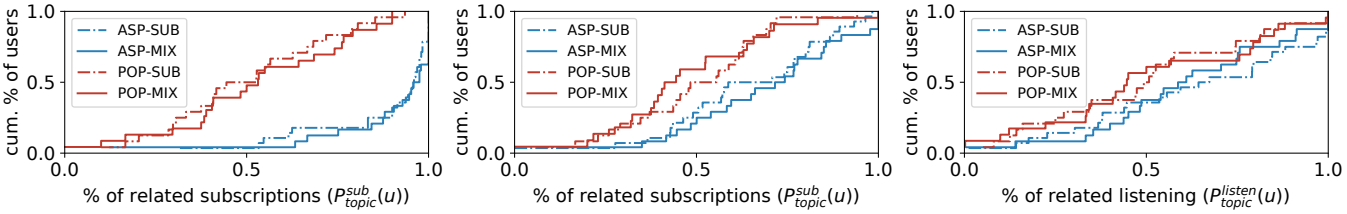
Our study results indicate significant interactions between recommendations and intentions. We discuss our findings in light of theoretical and empirical research on human decision making and suggest directions for designing better recommendation systems that benefit end users.

5.1 Employing planning and intentions

Individuals face self-control challenges when making decisions about content consumption, just as they do when managing diet or finance [32]. People have troubles translating their intentions and goals into actions when facing real-world decision-making problems. For example, prior research showed that people rent documentaries in line with their "aspirational self" but were less likely to actually consume this type of movie compared to more affective movies such as action films [32]. Filter bubbles [36] are another example in which users' long-term interests do not match with short-term consumption of news. To help people choose according to their long-term interests, our study suggests to employ a deliberative thinking via planning in the form of preference elicitation. As shown in Section 3.1, our onboarding system leveraged a preference elicitation-based interaction technique and an intention-aware recommender system that allowed for the explicit inclusion of user intentions. Such a design was shown to have significantly positive effects as users subscribed according to their elicited intentions during onboarding and later followed up on their plans when listening in the wild. Similar strategies were examined in behavioral science literature suggesting that people planning ahead are more likely to act on their intentions and to exhibit aspirational behavior in line with their long term interests [18].

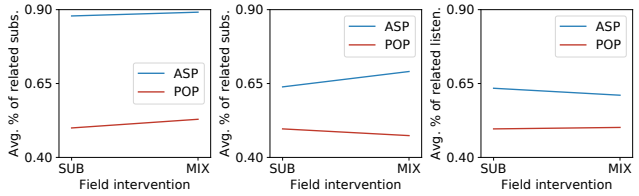
5.2 Encouraging exploration

Classical recommendation systems based on collaborative filtering and click-based metrics are often criticized since they are likely to be overly optimized to reinforce past behavior and preferences [24]. As a result, measures such as novelty and diversity are increasingly explored both in research papers and industry practice in recent years [21, 45, 53]. Finding the right mix of novel and familiar items can be challenging as it is not clear to what extent a certain quality characteristic like novelty is truly desired in a given application for a specific user and at a certain time. In the social and behavioral science literature this is often formulated as the exploration-exploitation trade-off [2, 4, 23]. Our results demonstrate



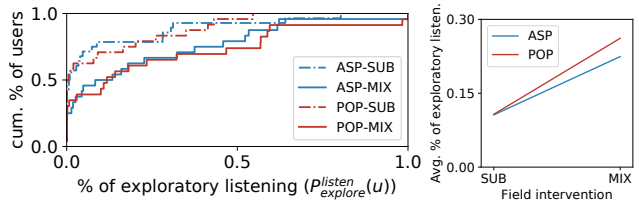
(a) Onboarding subscriptions (ONB: $**$). **(b) Field subscriptions (ONB: $***$).** **(c) Listening time (ONB: $*$).**

Figure 9: Cumulative distributions of users over the percentage of the topicwise intention-related subscriptions and listening. In the above figures, an $x = 1.0$ curve denotes that all users' consumption is related to their topicwise intentions, while an $x = 0.0$ curve denotes that none are related. The ASP intervention during onboarding is shown to significantly increase the topic-related onboarding subscriptions, topic-related field subscriptions, and topic-related field listening. The FIE factor and the interaction ONB×FIE have no significant effect.



(a) Onboard. subs. (ONB:$**$) **(b) Field subs. (ONB:$***$)** **(c) Listening time (ONB:$*$)**

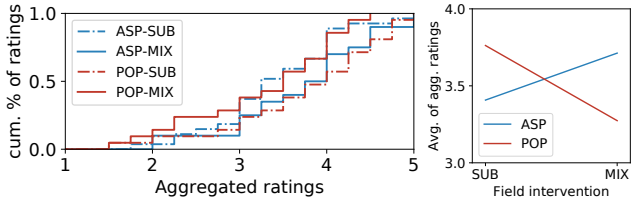
Figure 10: The groupwise average percentage of the topicwise intention-related subscriptions and listening. The ASP intervention significantly improves the topic-relatedness of onboarding subscriptions, field subscriptions, and field episode listening by 72.1%, 36.5%, and 24.9% respectively. The FIE and the interaction (ONB×FIE) have no significant effect.



(a) The cumulative distribution over users (FIE: $*$). **(b) The groupwise average (FIE: $*$).**

Figure 11: The percentage of subscriptions and listening from not-subscribed channels — (a) cumulative distributions over users, and (b) groupwise average. In (a), a $x = 1.0$ curve denotes that users do not listen to episodes from subscribed channels, while a $x = 0.0$ curve denotes that all listening comes from subscribed channels. The MIX intervention is shown to significantly increase the exploration rate by 127.5%. The ONB factor and the interaction (ONB×FIE) have no significant effect.

that introducing recommendation systems in content platforms that were mainly driven by user intentions provided benefits in the form of user exploration, because recommendations helped people find



(a) The cumulative distributions of aggregated ratings (ONB×FIE:$*$). **(b) The average ratings (ONB×FIE:$*$).**

Figure 12: Participants' satisfaction (the averaged ratings of all indicators) — (a) cumulative distributions of aggregated ratings, and (b) groupwise average ratings. The interaction between two factors (ONB×FIE) significantly affects satisfaction — MIX improves satisfaction if participants were onboarded with the ASP, otherwise MIX shows negative effects. No single factor alone has a significant effect.

choice alternatives that they were not aware of. However, how recommendation systems may influence the explore-exploit dilemma in the long term is an open question for future research.

5.3 Understanding user satisfaction

Our results show that users were satisfied when CF-based recommendations (MIX) were delivered based on intention-driven subscriptions (ASP) at onboarding. This can be explained by the benefits of reinforcing users' long-term interests. Whereas when users' initial subscriptions were only driven by channels' popularity and not aspirational, CF-based recommendations ignored their intentions and left them dissatisfied. Another possible explanation is the explainability and trust of recommendations. People are more likely to follow recommendations they trust, and explaining recommendations is shown to increase the trust [54]. Since the ASP-MIX hybrid recommender systems were informed by stated users' preferences, recommendations were implicitly explained and were easier to be perceived and understood by users. Whereas when the POP-MIX systems were used, the explainability and trust of field recommendations was expected to be low.

Additionally, as shown in Section 4.4, we also observe high user satisfaction under the POP-SUB interventions, in which users were

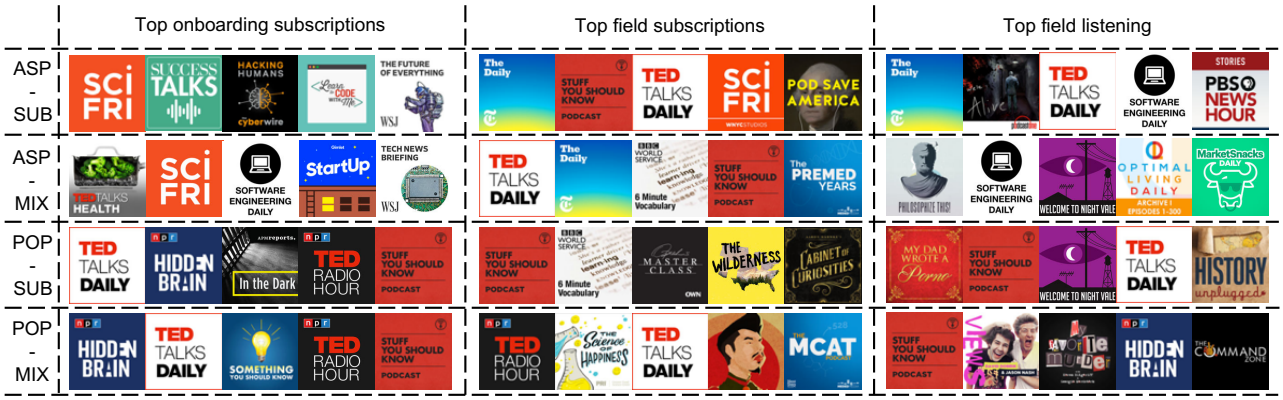


Figure 13: Top five most interacted content source during onboarding and in the field, categorized by 2×2 groups. Each square icon represents a podcast channel. These qualitative results further demonstrate how users' content consumption in the field was jointly affected by users' intentions and recommendation systems.

left in their information bubble populated with self-chosen popular items. This may be explained by people's inherent motivation to chase popular items [51] even if these items were misaligned with users' stated intentions; such content satisfies an important, if implicit, aspect of people's information needs and desires.

5.4 Optimizing for multiple objectives

Our study reveals benefits of jointly optimizing people's information consumption for multiple objectives. For example, for podcasts and other subscription-based media, service providers should consider a hybrid form of recommender that contains an intention-aware recommender for onboarding and a CF-based recommender for field listening. This combination can support users' intentions while encouraging them to explore beyond existing channels. As a result, users are likely to be more satisfied. More generally, with a global view of the recommendations that people are increasingly exposed to, we can jointly optimize recommendation systems to support an individual's aspirations and satisfaction in other domains such as diet and time management.

5.5 Limitations of intention-agnostic metrics

Commonly-used metrics that quantify user experiences are often agnostic to people's intentions. As a result, these metrics mainly reflect the extent to which recommendations engage people but overlook the utility of those engagements. For example, in our study, total listening time and total number of subscriptions show that people were equally active across different groups (Section 4.1), but in reality, people in certain groups were less exposed to new information, guided away from their aspirations, and less satisfied. Therefore, when probing and evaluating the performance of recommendation systems, it is important to condition metrics on individual intentions.

6 CONCLUSIONS AND FUTURE WORK

We presented a randomized controlled field experiment that studied the effects of recommendations on people's content choices related to intentions. Our study revealed how recommendations (1) modulated people's choices of topically relevant content, (2) affect the likelihood that people explore beyond their existing information sources, and (3) jointly affected user satisfaction. We discussed the implications and applications of our study findings on the design, evaluation and understanding of recommendation systems. Our study confirms the suspected importance of recommendations beyond discovering relevant information; in particular, that these systems implicitly alter online behavior in a manner that can have profound implications for individuals and society [1]. Future work is needed to study the generalization of these effects to wider demographic groups, and explore broader and longer term effects of recommendations through offline evaluation, simulations, and larger scale field experiments.

A QUALITATIVE USAGE RESULTS

We show the channels that were most-subscribed and listened during onboarding and in the field (Fig. 13). During onboarding, the subscriptions made in ASP-$*$ groups were much more diverse compared to the POP-$*$ groups. The subscriptions from POP-$*$ groups were mostly concentrated on trendy channels such as TED Talks Daily, TED Radio Hour, and Hidden Brain. However, in the field, all groups manifested diverse content consumption patterns, and the top subscribed and listened channels contained both trendy and long-tail items. These qualitative results further illustrate how users' podcast content consumption was driven by users' intentions and at the same time affected by recommendation systems.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant IIS-1700832 and by Yahoo Research (via the Connected Experiences Laboratory at Cornell Tech). The work was further supported by the small data lab at Cornell Tech, which received funding from NSF, NIH, RWJF, UnitedHealth Group, Google, and Adobe. The City University of New York provided generous support in recruiting participants. We thank the anonymous reviewers for their insightful comments and suggestions.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2011. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 217–253.

[2] Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, and Raghu Ramakrishnan. 2013. Content recommendation on web portals. *Commun. ACM* 56, 6 (2013), 92–101.

[3] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[4] Oded Berger-Tal, Jonathan Nathan, Ehud Meron, and David Saltz. 2014. The exploration-exploitation dilemma: a multidisciplinary framework. *PloS one* 9, 4 (2014), e95693.

[5] Jana Besser, Martha Larson, and Katja Hofmann. 2010. Podcast search: User goals and retrieval technologies. *Online information review* (2010).

[6] Robert J Boik. 1979. Interactions, partial interactions, and interaction contrasts in the analysis of variance. *Psychological Bulletin* 86, 5 (1979), 1084.

[7] Danah Boyd. 2010. Streams of content, limited attention: The flow of information through social media. *Educause Review* 45, 5 (2010), 26.

[8] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2017. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. *arXiv preprint arXiv:1710.11214* (2017).

[9] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 125–150.

[10] Justin Cheng, Caroline Lo, and Jure Leskovec. 2017. Predicting intent using activity logs: How goal specificity and temporal range affect user behavior. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 593–601.

[11] Elizabeth M Daly, Werner Geyer, and David R Millen. 2010. The network effects of recommending social connections. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 301–304.

[12] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 581–590.

[13] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Conference on Fairness, Accountability and Transparency*. 172–186.

[14] Michael D Ekstrand and Martijn C Willemsen. 2016. Behaviorism is not enough: better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 221–224.

[15] Andrew J Elliot and Judith M Harackiewicz. 1994. Goal setting, achievement orientation, and intrinsic motivation: A mediational analysis. *Journal of personality and social psychology* 66, 5 (1994), 968.

[16] 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.

[17] Marguerite Fuller, E Tsagkias, Eamonn Newman, Jana Besser, Martha Larson, Gareth JF Jones, M Rijke, et al. 2008. Using term clouds to represent segment-level semantic content of podcasts. (2008).

[18] Peter M Gollwitzer. 1999. Implementation intentions: strong effects of simple plans. *American psychologist* 54, 7 (1999), 493.

[19] Masataka Goto and Jun Ogata. 2011. PodCastle: Recent advances of a spoken document retrieval service improved by anonymous user contributions. In *Twelfth Annual Conference of the International Speech Communication Association*.

[20] Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. 2013. Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. *Management Science* 60, 4 (2013), 805–823.

[21] Cheng-Kang Hsieh, Longqi Yang, Honghao Wei, Mor Naaman, and Deborah Estrin. 2016. Immersive recommendation: News and event recommendations using personal digital traces. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

[22] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. Ieee, 263–272.

[23] Michael Inzlicht, Brandon J Schmeichel, and C Neil Macrae. 2014. Why self-control seems (but may not be) limited. *Trends in cognitive sciences* 18, 3 (2014), 127–133.

[24] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender systems-beyond matrix completion. *Commun. ACM* 59, 11 (2016), 94–102.

[25] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A Konstan, Loren Terveen, and F Maxwell Harper. 2017. Understanding how people use natural language to ask for recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 229–237.

[26] Bart P Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. 2016. Recommender systems for self-actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 11–14.

[27] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18, 1 (2009), 140–181.

[28] Mounia Lalmas and Liangjie Hong. 2018. Tutorial on Metrics of User Engagement: Applications to News, Search and E-Commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 781–782.

[29] Uichin Lee, Zhenyu Liu, and Junghoo Cho. 2005. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*. ACM, 391–400.

[30] Leonard A Marascuilo and Joel R Levin. 1970. Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of type IV errors. *American Educational Research Journal* 7, 3 (1970), 397–421.

[31] Katherine L Milkman, Dolly Chugh, and Max H Bazerman. 2009. How can decision making be improved? *Perspectives on psychological science* 4, 4 (2009), 379–383.

[32] Katherine L Milkman, Todd Rogers, and Max H Bazerman. 2008. Harnessing our inner angels and demons: What we have learned about want/should conflicts and how that knowledge can help us reduce short-sighted decision making. *Perspectives on Psychological Science* 3, 4 (2008), 324–338.

[33] Junta Mizuno, Jun Ogata, and Masataka Goto. 2008. A similar content retrieval method for podcast episodes. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*. IEEE.

[34] 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.

[35] Jun Ogata and Masataka Goto. 2009. PodCastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription. In *Tenth Annual Conference of the International Speech Communication Association*.

[36] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

[37] Edison Research. 2017. The Podcast Consumer 2017. http://www.edisonresearch.com/the-podcast-consumer-2017/

[38] William W Ronan, Gary P Latham, and SB Kinne. 1973. Effects of goal setting and supervision on worker behavior in an industrial situation. *journal of Applied Psychology* 58, 3 (1973), 302.

[39] Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. ACM, 13–19.

[40] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR. org.

[41] Amit Sharma, Jake M Hofman, and Duncan J Watts. 2015. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. ACM, 453–470.

[42] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 923–932.

[43] Jessica Su, Aneesh Sharma, and Sharad Goel. 2016. The effect of recommendations on network structure. In *Proceedings of the 25th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1157–1167.

[44] Jaime Teevan, Susan T Dumais, and Daniel J Liebling. 2008. To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 163–170.

[45] Choon Hui Teo, Houssam Nassif, Daniel Hill, Sriram Srinivasan, Mitchell Goodman, Vijai Mohan, and SVN Vishwanathan. 2016. Adaptive, personalized diversity for visual discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 35–38.

[46] Sabina Tomkins, Steven Isley, Ben London, and Lise Getoor. 2018. Sustainability at scale: towards bridging the intention-behavior gap with sustainable recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 214–218.

[47] Manos Tsagkias, Martha Larson, and Maarten De Rijke. 2010. Predicting podcast preference: An analysis framework and its application. *Journal of the American Society for information Science and Technology* 61, 2 (2010), 374–391.

[48] Ryen W White and Steven M Drucker. 2007. Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 21–30.

[49] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 143–146.

[50] Longqi Yang, Eugene Bagdasaryan, Joshua Gruenstein, Cheng-Kang Hsieh, and Deborah Estrin. 2018. OpenRec: A Modular Framework for Extensible and

Adaptable Recommendation Algorithms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 664–672.

[51] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 279–287.

[52] Longqi Yang, Cheng-Kang Hsieh, Hongjian Yang, John P Pollak, Nicola Dell, Serge Belongie, Curtis Cole, and Deborah Estrin. 2017. Yum-me: a personalized nutrient-based meal recommender system. *ACM Transactions on Information Systems (TOIS)* 36, 1 (2017), 7.

[53] Longqi Yang, Michael Sobolev, Christina Tsangouri, and Deborah Estrin. 2018. Understanding user interactions with podcast recommendations delivered via voice. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 190–194.

[54] Mike Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2016. Making sense of recommendations. *Preprint at http://scholar. harvard. edu/files/sendhil/files/recommenders55_01. pdf* (2016).