To Compliment a Complement?

Modeling New Avenues for Podcast Content Discovery

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

MSc.

by Bharath Satheesh and Benjamin Lee Dobkin

May 2020

ABSTRACT

[Although podcasts have emerged in recent years as the fastest growing media form in the US, consumption remains significantly skewed in favor of the top 1% most popular podcasts – with many new podcasts largely remaining undiscovered. Among the primary culprits for this friction are the still mostly primitive and generally limited content discovery tools and features offered by today's leading platforms, from Spotify to Apply Podcasts. In this study, we approach this multifaceted challenge in three phases. First, we conduct an extensive qualitative analysis of field study data on CUNY University students, coding interviews surrounding their podcast consumption habits and experiences with the medium more broadly. Second, we provide a quantitative analysis of the students' listening behaviors, particularly with a view to understanding the impact of listening frequency and platform selection upon content and consumption preferences. Ultimately, in observing that the content creator is perhaps the most neglected entity in the podcast value chain and is uniquely positioned to support any content discovery intervention strategy, we develop a model that identifies complementary podcasts based on listeners' subscriptions, content categories, and podcast descriptions. With this proof-of-concept implementation, we propose a new avenue through which to unlock content discovery opportunities for podcast creators and listeners alike.]

CHAPTER 1

INTRODUCTION

The podcast industry has been growing at a phenomenal pace. Several companies like Spotify and Google are investing in the media consumption landscape through new platforms that aim to serve up customers with personalized recommendations. Being a nascent space, the technologies and tools to analyze, search and recommend podcasts are still evolving and that provided us with a great opportunity to understand the various factors that help explain user behavior. Specifically, we focused on how cold start search and recommendations can be enhanced through a better on-boarding process that would inform platforms of prior consumption behavior and hence be able to provide a tailored user experience. Further, through analysis, we found that consumption frequency and platform preferences played a major role in continued podcast consumption and often differed between daily, weekly and monthly users. This paper aims to walk researchers through the correlation between various key factors that directly impacted podcast consumption for 105 college students at CUNY.

At the same time, however, the podcast medium presents a unique challenge not readily solved by typical, leading recommendation engine approaches, which generally surface new content that can be readily substituted for another piece of content that has just been consumed (e.g., the Netflix model). By contrast, podcasts are typically produced and released in a continuous, open-ended fashion, building upon the loyalty of their listeners to stay on top of the latest episodes. In such a context, a standard recommendation engine could conflict with the incentives of both listeners and content creators.

Since observing that podcast creators have recently begun promoting one another's content in the wild, much like they would execute any ad read, we have sought to unpack the incentives that would drive such a seemingly counterintuitive behaviour. Moreover, herein we explore and propose a proof-of-concept but still readily actionable model to demonstrate how this phenomenon could be enabled at scale by identifying complementary podcasts based on listeners' behaviour in a controlled study. This new angle, focused on the content creator, can offer a significant pathway toward alleviating content discovery challenges constantly faced by creators, even without necessitating a significant shift in the technology underpinning today's leading audio platforms.

Note: This is the final report of our Specialization Project, a two-semester project required for the Connective Media Master's program. This project was a two-person research-oriented project done under the guidance of a faculty member at Cornell Tech.

CHAPTER 2

RELATED WORK

2.1 Content Recommendations as Part of On-boarding

We primarily focused on sources of work inside the podcast domain and in general within recommender systems to understand user on-boarding for new podcast platforms. Yang et. al. [1] conducted a 2 2 randomized controlled field experiment (105 urban college students) to compare the effects of intention in-formed 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. They also modified a commercial podcast application (Himalaya) to include a recommender that takes into account users' stated intentions at on-boarding. We add field study data from college students that was also collected at the time and augment our existing quantitative data. This way, we could augment data collected and look at additional factors that affect listening behaviors such as frequency of consumption and preferred mode or platform of consumption. The Berger-Tal O et. al. paper [11] also explains balance between exploratory and exploitative behaviors that we have tried to incorporate into our research.

One of the major challenges of podcast consumption platforms, the cold-start problem, plagues nearly all recommender systems. In particular, new items will often be overlooked, impeding the development of new products online. Further, utilization of the knowledge of recommender systems under resource constraints is extremely important. In our exploration we also considered the Liu J-H et. al. framework [12] that shows that simply pushing new items to active users is not a good strategy. Their paper also indicates that to connect new items with some less active users will statistically yield better performance; namely, these new items will have more chances to appear in other users' recommendation lists.

In order to understand cold-start recommendations we also studied quantitative papers by Sun et. al [13], Yang et. al. [14] and Liu et. al. [15] as we considered Collaborative Filtering (CF) approaches to better serve user onboarding in podcast platforms. CF approaches suffer from the cold-start problem for users and items with few ratings. Hybrid recommender systems that combine collaborative filtering and content-based approaches have been proven as an effective way to alleviate the cold-start issue. In order to augment quantita- tive

evidence, we decided to correlate qualitative participant information with content-based approaches to serve up a better user experience for podcast platform subscribers.

2.2 Analytical Framework for Podcast Recommendations

As we looked at existing recommendation engines in podcasts, most implementations including the Yang Sobolev et. al. [1] paper focused on podcast to podcast recommendation schemes that focus on aspects like host-sentiment and aim to deal with issues like the cold-start problem (as mentioned in 2.1). Given that the Podcasts industry is still growing and has yet to hit a 1 billion USD (United States Dollar) market cap as of 2019, we looked at larger industries with more robust recommendation engines like video content (i.e. YouTube, Netflix etc.) and shopping. We also found the Zhang et. al. paper [10] on Spotify user behavior to be fairly interesting, with emphasis on the correlations between both the length and the downtime of successive user sessions on single devices. In particular, we conduct the first analysis of the device-switching behavior of a massive user base.

While recommendation algorithms in the advertising space [2], [3], [4] tend to deal with topic modelling [2] from text, optimizing for multi-product utility maximization [3] or willingness-to-pay [4], we were looking for an approach that involved minimal amounts of text based training data. Further, given that listening to podcasts have been free for the most part (the outlier being Luminary Media,) willingness-to-pay does not entire predict listening behavior as well as content based recommendation engines. When we looked at the winners of the Netflix prize [5], we found that user average based ratings, movie average based ratings, SVDs and Pearson correlations were popular. Given our

limited data set as well the fact that we had no time dependence [6] in our data (since most uses were on-boarded together) we decided against this model and went with a simpler approach based on an average of user similarity along with podcast similarity.

We also looked at papers that dealt with explicit complement datasets [7], [8] but realized that podcast complement data was not asymmetric (i.e. a user should ideally listen to a complement podcast as a result of them listening to the original podcast) and that podcast listening behavior is far from transient. The listeners to a podcast are often loyal fans as is demonstrated through high ad conversion rates. Given all the constraints of our data set, we decided to start with a simpler model that indicated a symmetric complement relationship between various podcasts and decided use an unsupervised learning algorithm based on Doc2Vec [9] that analyses podcast description similarities similar to a bag-of-words model that we use to find category based similarity.

CHAPTER 3

METHOD

3.1 Qualitative Coding and Data Analysis

In order to better understand user preferences, we decided to use pre-prepared podcast consumption data and correlate observations from the quantitative data collected in the Yang et. al. paper [1] with partially coded qualitative analyses from a field study conducted for the same paper.

3.1.1 Participants: Qualitative and Quantitative analyses

The Yang et. al. team interviewed 105 undergraduate students from the City University of New York (CUNY) for their study. Given limited access to time or resources, we decided to use this data to analyse podcast consumption behaviour. Further, we also conducted some preliminary checks (i.e., comparing podcast consumption behaviour between collected interviews and popular industry charts) and ruled out the null hypothesis that student data would be invalid in testing with our target audience.

3.1.2 Qualitative Analysis on Field Study Data

Theory: Users were asked a series of questions in the field study, including current podcast consumption frequency, preferred podcasting platform (if they were listeners) and purpose for podcast consumption. We chose to explore both podcast consumption frequency and preferred podcasting platform as they would not only directly impact resulting podcast consumption behaviour but would also give us a good sense of participant familiarity with podcasts. Further, by understanding preferred consumption platform, we could further break down and investigate individual platform features that directly correlate with listening frequency.

Analysis: In order to qualitatively code frequency and platform, we came up with a framework that split up frequency into 5 categories: None, Minimal, Monthly, Weekly and Daily. None meant that the participant had not listened to podcasts previously and were cold users. Minimal signified having listened to podcasts in the past (i.e., the participant had not listened to podcasts in the past 6 months). Monthly, weekly and daily users listened to podcasts at least once a month, week and day respectively.

In order to effectively catalogue platforms, we initially coded interview responses into raw categories. Then we split up the codes into four categories and bucketed our responses into these broad categories. The five categories were as follows: Apple (signifying the Apple Podcast application on the iPhone), Spotify (referring to the Spotify mobile application on iOS and Android devices), Youtube (either on mobile or a PC) and finally Online (referring to podcast users who listened to podcasts online through one medium or another.) For interviews, where we could not gauge interview responses accurately, we left the platform column empty and gave the frequency column a 'None' value. Finally, our codes were tied to participant video ids that were then correlated with dummy email accounts created for the purposes of the study.

3.1.3 Quantitative Analysis

Design Analysis: In order to understand podcast behaviour in relation to listening frequency and platform preferences, we decided to pick out the following variables from the quantitative data available through the Yang paper (per individual participant) over the period of the study:

- Number of podcast searches
- Number of podcast subscriptions
- Number of valid episode listens (both 5 minutes and 10 minutes)
- Average listening time (over all active sessions)
- Initial self-submitted content preferences (selected podcast channels)
- Final self-submitted satisfaction scores (on a scale from 0-10)
- Number of logins over the study period

We also created multiple complex variables by combining existing variables i.e. number of valid listens per login, number of searches per login and average listening time per login. We used 2D graphing and correlation tools to aggregate various variables and create effective visualizations.

3.2 Scoring using Description and Categorical data

We first collected data from 100+ unique user interviews from [1] that contained particular user preferences for podcasts. This again was sub-sampled further since half of this initial podcast data set was based off of popularity (along with user interest) while the other set is based on explicit interest. Once users preferences were loaded, we made sure that the users represent a diverse group or listeners (based on preferences, number of podcasts etc.)

We then tabulated some personal data (that was anonymized to remove sensitive information such as name, email address, age, sex etc.) but we did not fully understand podcast metadata from the title of the podcast; as a result, we looked for external data points that would assist us in creating a richer data set that contained podcast metadata that would be valuable in creating a creator focused recommendation engine.

We compiled a list of potential features that we needed to extract meaningful insights from our podcasts. We then sorted through this list by relevance and then tried to recreate this list in the wild. The final list included categories like content, frequency, length, monetization, production, social/peers, narrative structure, platform, studio/media network, guests, presentation, ad crossover,voice of the host etc. Now given that we identified important podcast features that

we wanted to compare, we started looking out for data that we could use for our implementation. This was extremely hard since:

- We did not receive any funding for this work and hence did not have the means to extract data from Podcast search engines (like Listen Notes.)
- We were pressed for time and bandwidth which made it incredibly difficult to parse Apple Podcasts/ similar pages at scale and compile said data set.
- Even when we did find data, it was not clean, and hence required a lot of manual cleaning effort to make it usable
- Another major challenge was that while we did have access to podcast platform listening data (anonymized), we did not have access to podcast metadata which was crucial for our project.

After reaching out to multiple sources within the industry and trying out various APIs that allowed free data extraction, the founder of Listen Notes directed us to their online Kaggle competition that contained a fraction of the top podcasts that were listened to at the time (2017.) Given that this data was relatively clean and contained more than 3 features/ podcast, we decided use this data as our original podcast data dictionary.

Given that the data dictionary was not clean, often times there was data match errors between our user data and the dictionary while meant that we had to implement fuzzy matching for the titles. We did this using the python package *fuzzywuzzy*. We then created a new data-table that contained meta-data (from the dictionary) about all the podcasts that users listened to along with an additional column that contained information about the user (i.e. user_id.) Some of the columns included 'title', 'image', 'description', 'language', 'categories', 'website', 'author', and 'iTunes_id'.

We then selected the most important feature vectors based off our data and our earlier analyses (described in 4.) This turned out to be the description of a podcast and the categories that a podcast belonged to.

User-based Category Similarity Scoring

For categories, we then created a BoW model that had binary variables for whether a podcast contained a certain category (i.e. did podcast A contain the category "comedy"?) We then compiled the list of categories by user, rather than by podcast. This was done because we were more interested in finding user-user similarities rather than podcast-podcast similarity. This was done based off our research on collaborative filtering models (both in content and shopping) as well our general understanding that user-user podcast appetite was diverse and that users tended to listen to a diverse set of podcasts rather than consume multiple podcasts in few genres.

Once we created user based category vectors, we could then look at a new podcast from a content creator and match them to audiences that not only already subscribed to their existing content genres, but also subscribed to other categories of podcasts (i.e. had a diverse appetite for content consumption.) We also grappled with identifying what would serve as similarities between two category vectors. For instance, we noticed that users that listened to more podcasts would be favored over users that listened to fewer podcasts since penalties for an similarity loss function was defined by the number of similarities or differences between podcast categories. I.e. the fewer podcasts a user listened to, the lower their overall penalties would be (regardless of the distance.) As a result, we then calculated penalties by diving final penalties by the l1 norm of the category vector.

By finding the *n* users that were most likely to subscribe to a creator, the creator could then look at the other types of category content that were popular amongst this user segment. While understanding popular podcast categories amongst users is important to know if you're a creator trying to make new podcasts, it really does not add value to existing popular podcasts and their creators. As a result, we also translated use-level categorization into podcasts that larger creators could endorse as potential complements.

While we took note of the fact that endorsement behavior is still growing, we still needed to match our test set with some amount of ground truth to understand if our recommendations held any value. In order to validate our model, we:

- Compiled potential test sets of podcasts that would show up as complements and looked at the intersection of our compiled set with our test results and compared the intersection to find a match score
- Tested results against similar podcasts (Politics, News and Comedy etc.) and validated that a high fraction of users intersected
- Used podcast search engines to come up with complement collaborations in the wild to see if we caught true-positives in our model; we used Listen Notes to come up references to other podcasts in their transcription based search results.

Once we compiled a scoring chart based on categories, we found that due to miscategorization within our data set, few podcasts had incorrect categories and as a result, we decided to use podcast descriptions to augment our matching process. Unfortunately, descriptions are never uniform and are often short. They also refer to proper nouns (i.e. Ben Shapiro.)

Description-based Similarity Scoring

We initially cleaned the data with stop-words from *nltk* and removed additional punctuation and restricted our data to podcasts produced in English. The next step was to understand how we could potentially make sense of short description vectors and as to what model would be best to fit descriptions. After comparison shopping multiple models, we decided to settle on Gensim's Doc2Vec model to train our data. We initially found that standard hyper-parameters did not seem to match up similar podcast descriptions and as a result, we decided to do some hyper-parameter tuning to adjust our initial α and step size along with vector length.

An interesting observation during this process (apart from how alpha values and step sizes behaved) was the min-word counts. The Min-word parameter gets rid of words if they don't occur frequently. Initially, our thought was to set a high min-word count to get rid of filler words but through our testing we found that "important" words that reflected podcast category seldom showed up more than once in a podcast description.

This meant that we also got rid of key words when we set the min-count to more than 0. Given our Doc2Vec implementation, we could then go ahead and test description similarity quality by looking at the most similar descriptions

given our test set. This was also challenging since we were looking for keywords rather than writing style and the model would return multiple results that turned out to be false-positives.

That being said, we still managed to get rid of data set inaccuracies (for the most part) when we used descriptions so we decided to use a scoring model from Doc2Vec for descriptions. Now, given we had 2 scoring sets (for simi- larity based on user-category relationships as well as for descriptions) we then normalized both scores (to range from 0 to 1) and then identified various loss functions to sum them together to create a final podcast index for a content creator.

We decided to put more weight on categories than descriptions since they held up better against test data and also validated our assumptions better. We also tested different weights on our loss function to see how category-description interdependence affected results in comparison to category based ranking. We once again validated our final combined scoring model both against ground truth and our expectation of what that would look like (from our hypothesis.) We also came up with suiting next feature vectors (like host specific qualities, reaching out to external online sources for verification etc.) in order to make our model more robust against sampling bias and data sparsity. Given weights x_1 and x_2 and normalized category and description scores c and d, we have the following loss function that ranks various podcasts in 3.1

$$L(x_1, x_2) = x_1 * (cd) + x_2 * c \quad \forall x_1, x_2 \in [0, 10] \tag{3.1}$$

Our code as well as implementations can be found attached as a Jupyter Notebook file that will soon be available for download.

CHAPTER 4

RESULTS

4.1 Extended Qualitative Analyses

Why and when do people listen to podcasts?

Motivations for podcast consumption: The interviews underscored a broad set of motivations for listening to podcasts. More than half of active podcast listeners reported that they primarily listen for entertainment purposes, often citing longer-form narrative podcasts such as Serial or comedy as an appealing genre. Typically, this sense of entertainment was driven by a strong affinity for the host, either in the context of the podcast alone or broader fandom for the host that extended beyond the podcast medium alone. "I was already a huge fan of this comedian and then found out he had a podcast…," and "I just really enjoy the way the host tells stories…" were common refrains for this group.

Other participants shared a professionally-oriented interest that has shaped their podcast discovery and consumption. These users were primarily leveraging podcasts as yet another content medium through which to deepen their knowledge and expertise in a particular domain relevant to their career goals. Exemplifying this segment, a medical student shared, "I often look for podcasts to further explore the applications of the medicine I have been learning in the classroom." In a related manner, a smaller segment of the CUNY student population divulged that podcasts offered an encyclopedia-esque ecosystem through which to learn about any potential topic of interest. Capturing the sentiment of this cohort of curious learners, one student shared that "podcasts are a wonderful way to learn more on topics that I've always been curious about, especially when my time might otherwise be limited, and I can do other things at the same time too while I listen."

Occasions for podcast consumption: Interview participants revealed that they consume podcasts across a variety of settings and occasions. A significant majority reported that they typically listen while commuting or waiting in line - instances that contain time that otherwise is generally unoccupied. Others shared that they often catch up on their podcast feeds while completing a parallel task that may not be especially demanding or require all of their attention. Notably, this latter mode of consumption often coincided with those indicating that they listen to podcasts for entertainment purposes. As one interviewee aptly stated, "podcasts are unique in that they provide entertainment value while still allowing you to be alert of your surroundings or something else you might be doing at the same time." By contrast, a smaller portion of participants who do not actively listen to podcasts offered a different view, claiming that they did not incorporate podcast consumption into their day-to-day life because they "didn't feel that divided attention was possible."

How people discover podcasts

Content discovery triggers outside of consumption applications: Through the 112 user interviews, there is substantial evidence to believe that users are driven to platforms (at least initially) based on external content discovery triggers such as celebrity presence, friends etc. We observed this in particular for hesitant podcast listeners (listening frequency of 1/week.) Further, for hesitant listeners or family members who wanted to listen to podcasts, specific speakers or hosts all turned out to be important with a lot of emphasis being put on listening to story format podcasts, hosted by "passionate" host-speakers about topics personal or relevant to the target listeners.

There was also some significant correlation between listener frequency and the fact that at least one friend or family member also listened to podcasts. We

however, do not have sufficient evidence to show that listening patterns vary as the number of friends (who are podcast listeners) vary. One participant suggested that the research team provided them with a diverse array of podcasts but did not get any new or relevant podcasts since they wanted to learn more about "internet culture" rather than broader technology podcasts. They also suggested explicit measurement tools such as logging in with social media accounts or adding specific information such as occupation, demographic information etc.

In-app content exploration triggers: Multiple participants in the study expressed interest in a podcast platform that allowed previewing podcasts or annotating/ tagging podcasts for better selection. Further several complaints regarding podcast search and better phenotyping features. 14 participants explicitly stated that they gave low scores or average scores to podcast recommendation because of lack of clarity or time. They claimed that they did not have the time to listen to the prescribed podcasts and so they did not find additional value within the recommended set.

In addition to phenotyping podcasts, another important note involves separating podcasts from individual episodes. While several participants were interested in podcasts related to specific fields, they also wanted speaker or guest specific podcasts as a great step into consuming this media form. There was also a clear indication for the need for podcast preview (15-20secs of content) to find speaker compatibility. However, there is no specific evidence pointing to formats for these snippets (i.e. introductions, curated highlights etc.) This problem is both interesting and builds off prior work conducted by Yang, et. al.

Opportunities for the podcast space

Desired improvements to existing content consumption platforms: Most (over 50) participants used Spotify or Apple Podcasts as the default podcast content consumption platform. This however could be biased against Google Podcast users or an older demographic that may not be students. That being said, most participants found their music and video content from Youtube and Spotify. There is no direct link between their favourite shows on Youtube and their podcast subscriptions on Apple Podcasts. Further, as mentioned in the other sections, the search feature in most (if not all) current podcast platforms is lacking and multiple study participants suggested an improved searching experience rather than the primitive, often clunky features that exist today. Several participants cited features that could aid in addressing this content discovery capability gap, highlighting the utility of short podcast descriptions and preview audio snippets, when available.

Non-listeners' associations with the term 'podcasts': More than one-third of interviewees actually were not active podcast listeners. At the close of each interview, they were typically asked what first thoughts would come to mind upon hearing the word 'podcast'. Perhaps exhibiting the proliferation of podcasts and the growth of the medium as a whole in recent years, the responses actually echoed many of the core value propositions voiced by active listeners as well. From the perspective of the non-listeners, podcasts are a "conversation medium", "another way to consumer information or keep up with the news", and "maybe another way to learn something new or resourceful". Aligning with those listening to podcasts while on-the-go, one participant highlighted, "I've only ever really listened to podcasts while traveling or being otherwise stuck somewhere, like a car, for a while."

4.2 Quantitative Analyses – Scoping the Content Discovery Problem

	asp_norec		asp_rec		pop_norec		pop_rec			
Frequency_Final	Number of Participants	Percent of Participants	Number of Participants	Percent of Participants	Number of Participants	Percent of Participants	Number of Participants	Percent of Participants	Grand total of Number of Participants	Grand total of Percent of Participants
None	7	21%	11	33%	9	27%	6	18%	33	100%
Minimal	2	22%	-	-	2	22%	5	56%	9	100%
Monthly	7	35%	3	15%	5	25%	5	25%	20	100%
Weekly	3	20%	3	20%	3	20%	6	40%	15	100%
Daily	7	29%	9	38%	5	21%	3	13%	24	100%
Grand total	26	26%	26	26%	24	24%	25	25%	101	100%

Figure 1: Breakdown of Consumption Frequency Behaviours across Experimental Treatments

The Yang [1] field study revealed that of the 79 participants who reported demographic information, there were 50 females and 29 males, with a mean age of 21. Further, 49 participants were iOS users, while 30 were Android users, and the population stemmed from a variety of academic majors.

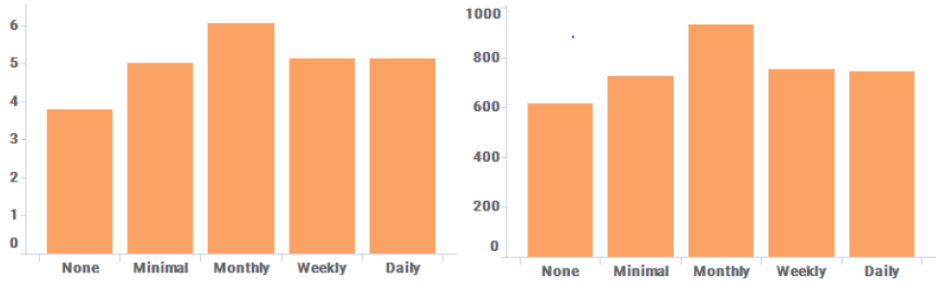


Figure 2: Total Searches and Average Listening Time, Respectively, vs. Podcast Consumption Frequency

Before exploring potential segmentation of the results, we demonstrate (Figure 1) that the participants in each frequency band were relatively evenly distributed across the four groups, thereby enabling us to pursue direct segmentation of the Yang et. al. [1] results according to the frequency with which participants' previously consumed podcasts.

Podcast Consumption Frequency Shapes Distinct Content Preferences: We correlated initial selection of podcasts in the field study to frequency of podcast consumption (based on our daily, weekly, monthly etc. buckets.) We then compared the top 15 podcasts based on the number of participants who chose the podcast and compared to the top 15 podcasts that were selected by daily users, weekly users and monthly users. We then created comparative Venn diagrams to find the intersection of podcasts between these various groups. Our diagrams showed us that (1) half the popular podcasts (7/15) listened to by weekly listeners coincided with those listened to by daily and monthly listeners (2) that there is greater correlation between podcasts chosen by weekly listeners to overall popular podcasts (9 out of 15) than to those selected by daily and monthly listeners (6 out of 15.)

Controlling for Podcast Consumption Frequency Reveals Some Novelty Bias in User Reception of a Podcast Recommendation Engine: Beyond the exploration of users' content preferences according to the frequency with which they consume podcasts, we also examined the effect of prior podcast exposure and consumption frequency upon key listening behaviours quantified by Yang et. al. [1]- total searches, total average listening time per listen, and average total listening time overall. As shown in Figure 2, minimal and monthly podcast listeners exhibit increasing average listening times, respectively, relative to non-listeners. However, the average listening time then decreases from the peak of 931 seconds per listen to 753 and 744 seconds per listen across weekly and daily podcast listeners, respectively.

These patterns suggest that the Yang et. al. preference measurement [1] and content recommendation engine in Himalaya media yielded successively enhanced content engagement and listening times for light to moderate listeners

- but less so for the most active podcast consumers. Of note, however, the average total listening time breakdown shown in Figure 2 as well suggests that the average listening time per listen may have been negatively skewed slightly as a result of the most active consumers listening to podcasts across a greater number of shorter stints rather than extended, uninterrupted sessions.

Ineffective Search Capabilities Appear to Sit at the Heart of the Podcast Content Discovery Challenge: Delving into the relationship between total searches, binned by quartile, and total listens as well as total valid listens - defined as any listening session that lasted at least five minutes, Figure 3 below illuminates the ineffectiveness today's podcast content search capabilities - at least those offered by Himalaya Media's app. Across the bottom three quartiles of participants in terms of total searches, both the average total listens and the average total 'valid' listens do not vary materially. Only in the top quartile of searches do we observe any material increase in the number of listens, as well as valid listens. To further underscore the deficiencies of search capabilities, Figure 4 also exhibits the average number of total invalid listens per search activity quartile. Therein, we observe that the number of invalid listens, on average, increased significantly from the second lowest quartile through the top two quartiles in terms of the total number of searches executed. Such a stark image especially underscores the opportunity to develop better content discovery capabilities for podcast consumers.

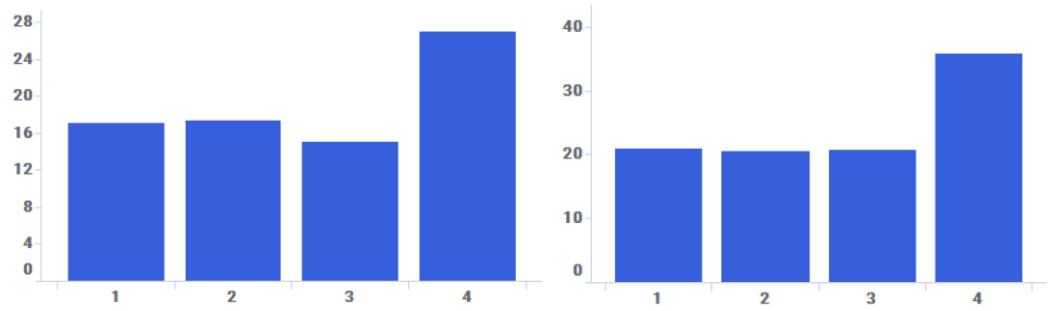


Figure 3: Average Total Listens and Average Total Valid Listens (>5 minutes), Respectively, vs. Total Searches by Quartile

Further, we see, in Figure 4, that the group who received popular impersonalized recommendations did search marginally more as compared to those who received personalized recommendations; however, they experienced nearly three-fold more 'invalid' listens wherein they abandoned a podcast in the first five minutes of listening time. This result further confirms the value and critical importance of robust recommendation engines in the podcast consumption space.

Figure 4: From Left, Average Total Searches vs. Experimental Group, Average Invalid Listens (<5 minutes) vs. Experimental Group, Average Invalid Listens vs. Total Searches by Quartile

User Behaviour Metrics on a per Login Basis Reveal Defining User Characteristics: Investigating searches, valid listens, and listening time on a per login basis across the frequency with which listeners consume podcasts, Figure 5 below initially echoes the novelty effect described (Figure 2) wherein the Yang et. al. field study [1] drove increased content searches among light and moderate consumers. At the same time, however, the increased volume of searches did not drive a

corresponding increase in valid listens - further highlighting the inability of the search capability to recommend personalized content.

With that context in mind, the view of average listening time per login across this same segmentation of consumption frequency reveals that weekly and daily users simply consume shorter-form podcasts, on average. Further work in this area may therefore require a more refined definition of a 'valid' listen - rather than simply using a single cut-off time across all segments of consumers and podcast formats.

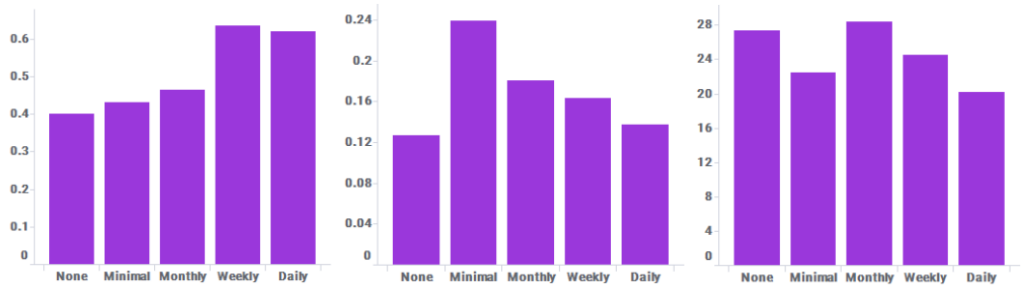


Figure 5: From Left, Average Searches Per Login, Average Valid Listens Per Login, and Average Listening Time per Login, Respectively, vs. Podcast Consumption Frequency

Figure 6 examines searches, valid listens, and average listening time per login across the four experimental treatment groups implemented in the Yang et. al. study [1]. While we observe a greater number of searches per login among the 'asp_norec' cohort who received personalized recommendations and the 'pop_rec' group who received popular recommendations for subscriptions initially and on an ongoing basis, the number of valid listens per login for the latter group is nearly half that of the other three groups. This finding suggests that the recommendation engine explored by Yang et. al. informed more tailored and precise searches for the 'asp_norec' and 'asp_rec' segments, as well as greater average listening times per login.

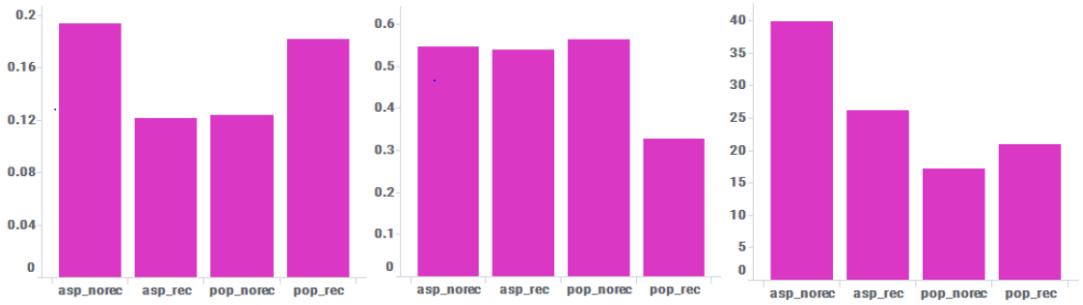


Figure 6: From Left, Average Searches Per Login, Average Valid Listens per Login, and Average Listen Time per Login, Respectively, vs. Experimental Group Assignment

4.3 Quantitative Analyses – A Model for Podcast-Podcast Promotion

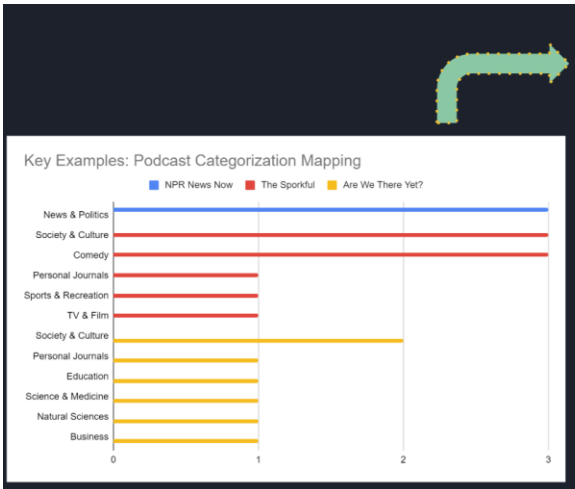
As outlined above in the discussion of the methods, we vectorized the categories of each podcast at the individual user level – based on their podcast subscriptions at the start of the Yang-Sobolev study. [1] In figure 7 below, we detail three example outputs of the model for the following podcasts, respectively: NPR News Now, The Sporkful, and Are We There Yet? First, in the left portion of the figure, we detail the frequency with which each given category was observed at the user level. As noted in the methods details, this categorization mapping directly shapes the spread of the categories evident across the top-10 closest peer podcasts provided in the model output (as highlighted in the three examples shown in the table in Figure 7).

NPR News Now, as shown in the first example, belongs to the 'News and Politics' content category, and as the mapping illustrates, the individuals in this study population also generally listened to multiple 'News and Politics' podcasts alongside NPR News Now. Thus, in the final output, the model indicates a number of other News and Politics podcasts as the most relevant peer podcasts,

including FiveThirtyEight Politics, Pod Save America, the Ben Shapiro Show, The Daily, FT News, and NPR Politics Podcast.

By contrast, the second example provides a distinct dynamic, as evidenced by The Sporkful podcast, a WNYC podcast for listeners passionate about culinary adventures. As noted in the category mapping, The Sporkful, belongs to the Comedy, Food, Society and Culture, and Arts categories, was a lone food-centric podcast consumed alongside those of many other categories, including personal journals, sports and recreation, and tv and film. In turn, our model proposes a broader set of complementary podcasts, including This American Life (News and Politics, Personal Journals, Arts, Society and Culture) and Brainstuff (Technology, Society and Culture, Natural Sciences, and Science and Medicine), among others.

Further, the final example highlights 'Are We There Yet?' (Science and Medicine, and Natural Sciences), an NPR-produced podcast about the latest in space-related news and developments, featuring interviews with astronauts, engineers, and other visionaries. However, of note, the closest peer podcasts surfaced by the model extend into other categories as well, via podcasts such as Radio lab that are also classified under the Education and Society and Culture Categories, and even Philosophize This!, categorized under Philosophy, Society and Culture, Education, and Comedy (expanding beyond Science and Medicine, and Natural Sciences altogether).



NPR News Now	*The Sporkful*	*Are We There Yet?*
FiveThirtyEight Politics	Poop Culture	Science… Sort Of
Pod Save America	Philosophize This!	Radiolab
The Ben Shapiro Show	Getting Curious with Jonathan Van Ness	BrainStuff
The Daily	This American Life	Philosophize This!
FT News	Podcast Podcast	Invisibilia
Lovett or Leave It	The Fighter & The Kid	Sword and Scale
Crimetown	Brainstuff	The Dave Ramsey Show
NPR Politics Podcast	The Mustards	Big Picture Science
This American Life	Pardon My Take	Casefile True Crime
Louder with Crowder	The minimalists Podcast	Where Should We Begin?

Figure 7: Key Example Outputs Demonstrating the Dynamics of Our Proposed Model Identifying Peer Podcast-Podcast Promotion Opportunities

CHAPTER 5

DISCUSSION

While our investigation builds on the recent field study published by Yang et. al. [1] earlier this year, we highlight the power of extracting additional features from largely unstructured qualitative interviews conducted in conjunction with the study - but whose data was neither previously extracted nor published. Initially, we demonstrate that the frequency with which listeners consume podcasts manifests itself in distinct podcast content preferences. We therefore hypothesize that prior consumption frequency behaviours can be used as a critical feature to streamline future preference measurement and content recommendation engines. Moreover, we show that field studies such as that of Yang et. al. also demonstrate a novelty effect, wherein the impact of preference measurement and tailored recommendations is more pronounced for those participants who have relatively less exposure to podcast content, or otherwise just consume podcasts relatively infrequently. In turn, the evaluation of key metrics such as searches,

valid listens, and overall listening time must include a control for prior consumption behaviours and podcast ecosystem exposure, which we have established as a valid source of potential omitted variable bias in this space.

Further, in exploring the efficacy of search capabilities, we demonstrate that encouraging more searches does not enable the discovery of sufficiently tailored content. In fact, the results above illustrate that enhanced search activity alone generally drives a higher volume of 'invalid' listens and relatively poor satisfaction outcomes, as confirmed by the Yang et. al. field study [1] outcomes. At the same time, when paired with a personalized preference measurement and recommendation engine, search weaknesses are less pronounced and do not drive the same volume of invalid listens. This crucial nuance further highlights the promise of increasingly sophisticated recommendation engines in the podcast space and suggests the potential promise of implementing a more seamless pairing between a recommendation engine and increasingly precise search capabilities.

At the same time, upon hypothesizing that the content discovery challenge in the podcast medium could ultimately be better alleviated by surfacing complementary, rather than substitute content, we sought to pursue a distinct model from those popularized among other media forms where the consumption of a given creator's work tends to take place in a closed-ended timeframe (rather than by an open-ended subscription). Figure 7 demonstrates how the model we have developed in fact not only highlights the heterogeneity of content preferences at the individual level, but also reveals listener 'phenotypes' that extend a definition of complementary podcasts across content categories.

At first, the NPR News Now example highlights a listener phenotype consisting of a deep appetite for a single category of content – News and Politics. In capturing this signal from the study population, the model in turn returned top complementary podcasts largely in the same category (e.g., NPR Politics Podcast and Pod Save America). However, the Sporkful example proceeds to revel a distinct phenotype, wherein food-oriented listeners of this podcast expressed an affinity for a number of other categories. Unlike listeners of NPR News Now, Sporkful listeners in this study were not particularly inclined to limit themselves to podcasts about culinary experiences – instead opting to extend their listening to comedy and sports and recreation, among other areas.

Further, the 'Are We There Yet' example highlights yet another phenotype, wherein a corpus of complementary podcasts has been isolated for their intent to inform listeners. While 'Are We There Yet' sits squarely in the Science and Medicine and Natural Sciences categories, its listeners appear to generally be keen to learn something new with each podcast they consume. The top complementary podcasts in this case consistently reflect this ethos, from Radiolab to BrainStuff, to Philosophize This, to Big Picture Science, and more. Altogether, these key example outputs underscore the need for future recommendation engines in the podcast space to more precisely consider distinctions between substitute and complement content across listener phenotypes, and to extend the definition of complementary podcasts beyond typical category-based conventions. Moreover, the model highlights the significant opportunity for content creators to extend the scope of their promotional channels to include complementary fellow creators, thereby also efficiently unlocking additional chances to be discovered by new potential listeners.

At the same time, there are a few key limitations of the study to note, as alluded to in part above among the methods. First, it is worth highlighting that for many podcast platforms, content categorization is baked into the user interface and inevitably steers discovery in a material way as a result. In turn, in the Yang-Sobolev field study as well, this was undoubtedly a factor that shaped how users make their initial selections of podcasts to which they wanted to subscribe. Moreover, it is important to note that we included all 105 CUNY students, across all experimental treatments, in the training dataset for the complementary podcast identification model. Of course, whether participants were served up recommended podcasts based on their preferences or popular rankings inevitably had an effect, to some degree, on the selection of podcasts to which they subscribed. In future work, we would seek to address this by attempting to mimic a cold start in full across an entire training set population, such that neither the user interface, nor any experimental treatment, could levy a baked-in impact on the data. In this case, we had no other recourse, as the dataset itself was already of a limited size. Across future work, we would also certainly seek to scale up the training data set and/or relevant study population used to train the model.

CHAPTER 6
CONCLUSION

We presented a series of analyses that studied the effects of search and recommendations with implicit user preferences like podcast consumption frequency and intrinsic platform preferences. Our study revealed how (1) frequency modulated people's content preferences, (2) content frequency controlled for existing novelty bias in user reception of a podcast recommendation engine (3) ineffective search capabilities appear to sit at the

heart of the podcast content discovery challenge, and (4) that user behaviour metrics on a per-login basis reveal important user characteristics. We discussed the implications and applications of our study findings on the design, evaluation and understanding of correlation between qualitative and qualitative evidence that we put together. Our study confirms the suspected importance of recommendations [1] and search and show us the criticality of an efficient on-boarding process in building better user experiences for podcast listeners.

We then followed with the development of a proof-of-concept model for complementary podcast identification, ultimately offering a potential avenue to alleviate content discovery challenges faced by content creators. Even with a limited training dataset focused around two key features – podcast category and description, our model surfaced significantly distinct listener phenotypes and identified numerous instances of listener preference heterogeneity transcending category definitions. Ultimately, we intend to continue building on the promise of this model with more robust training data, seeking to more fully validate the viability of podcast-podcast promotion as a solution to content discovery friction at scale.

ACKNOWLEDGEMENTS

We'd like to thank Michael Sobolev and Longqi Yang for their work, data and advice as we persevered through this analysis and report. In particular, we extend our gratitude to Michael for taking us on as mentees through the year-long Specialization project and for giving us great advice as well as a clear direction as we refined and built upon our hypotheses and ideas.

BIBLIOGRAPHY

[1] L. Yang, M. Sobolev, Y. Wang, J. Chen, D. Dunne, C. Tsangouri, et al., 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.

[2] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products." Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2015. https://doi.org/10.1145/2783258.2783381.

[3] Q. Zhao, et al., Recommendation based on multiproduct utility maximization. No. SP II 2016-503. WZB Discussion Paper, 2016.

[4] M. Zhang and J. Bockstedt, "Complements and Substitutes in Product Recommendations: The Differential Effects on Consumers' Willingness-to-pay." IntRS@ RecSys. 2016.

[5] Bennett, James, and Stan Lanning. "The netflix prize." Proceedings of KDD cup and workshop. Vol. 2007. 2007.

[6] Bell, Robert M., Yehuda Koren, and Chris Volinsky. "The bellkor 2008 solution to the netflix prize." Statistics Research Department at ATT Research 1 (2008).

[7] Yu, Hang, et al. "Complementary Recommendations: A Brief Survey." 2019 International Conference on High Performance Big Data and Intelligent Systems (HPBDIS). IEEE, 2019.

[8] T. Zhao, et al., "Improving recommendation accuracy using net- works of substitutable and complementary products." 2017 Interna- tional Joint Conference on Neural Networks (IJCNN). IEEE, 2017. https://doi.org/10.1109/IJCNN.2017.7966315.

[9] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation." arXiv preprint arXiv:1607.05368 (2016). https://doi.org/10.18653/v1/W16-1609.

[10] B. Zhang, et al., "Understanding user behavior in spotify." 2013 Proceedings IEEE INFOCOM. IEEE, 2013.

[11] Berger-Tal, Oded, et al. "The exploration-exploitation dilemma: a multidis-

ciplinary framework." PloS one 9.4 (2014): e95693.

[12]Liu, Jin-Hu, et al. "Promoting cold-start items in recommender systems." PloS one 9.12 (2014): e113457.

[13]Sun, Mingxuan, Fei Li, and Jian Zhang. "A multi-modality deep network for cold-start recommendation." Big Data and Cognitive Computing 2.1 (2018): 7.

[14]S. Longqi Yang and Y. Wang, Drew Dunne, Michael Sobolev, Mor Naa- man, and Deborah Estrin. 2019. More than Just Words: Modeling Non- textual Characteristics of Podcasts. In The Twelfth ACM International Con- ference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia. ACM, New York, NY, USA, 9. 3290993 https://doi.org/10.1145/3289600.3290993.

[15]Liu JH, Zhou T, Zhang ZK, Yang Z, Liu C, Li WM. Promoting cold-start items in recommender systems. PLoS One. 2014;9(12):e113457. Published 2014 Dec 5. doi:https://doi.org/10.1371/journal.pone.0113457.