The Diversity of Music Recommender Systems

Ian Baracskay
iabaracskay@gmail.com
Daniel High School
Clemson, USA

Donald J Baracskay III
3dgeorgia@gmail.com
Northwestern University
Evanston, USA

Mehtab Iqbal
mehtabi@clemson.edu
Clemson University
Clemson, USA

Bart P. Knijnenburg
bartk@clemson.edu
Clemson University
Clemson, USA

ABSTRACT

While the algorithms used by music streaming services to provide recommendations have often been studied in offline, isolated settings, little research has been conducted studying the nature of their recommendations within the full context of the system itself. This work seeks to compare the level of diversity of the real-world recommendations provided by five of the most popular music streaming services, given the same lists of low-, medium- and high-diversity input items. We contextualized our results by examining the reviews for each of the five services on the Google Play Store, focusing on users' perception of their recommender systems and the diversity of their output. We found that YouTube Music offered the most diverse recommendations, but the perception of the recommenders was similar across the five services. Consumers had multiple perspectives on the recommendations provided by their music service—ranging from not wanting any recommendations to applauding the algorithm for helping them find new music.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Recommender Systems, Music Streaming Services

ACM Reference Format:
Ian Baracskay, Donald J Baracskay III, Mehtab Iqbal, and Bart P. Knijnenburg. 2022. The Diversity of Music Recommender Systems. In *27th International Conference on Intelligent User Interfaces (IUI '22 Companion), March 22–25, 2022, Helsinki, Finland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3490100.3516474

1 INTRODUCTION

The number of users using music streaming services has increased from 76.8 million premium subscribers in 2015 to around 400 million in 2021 [19]. The prominence of these streaming services necessitates research on how they can better address the needs of the consumers and retain them.

Among the most important functionalities provided by streaming services is the recommendation of music to users. This work explores whether the *diversity* of the provided recommendations aligns with users' perceptions of the music recommendation algorithms. Currently, Spotify has the most premium subscribers of any music service [7], and a study conducted by Spotify researchers found that users who received a higher diversity of content had greater rates of conversion and retention [1]. Accordingly, this work compares the output diversity of Spotify's recommendations with four of its competitors (i.e., Pandora, YouTube Music, Apple Music, and Last.fm) by providing each service with the same lists of input items.

The aforementioned Spotify work also found that users' organic (i.e., not recommendation-driven) consumption of music was higher in diversity than their recommendation-driven consumption [1]. This begs the question of whether the recommendations provided by music streaming services matches, dampens, or exacerbates the diversity of the items on which the recommendations are based. The current paper addresses this question by providing the five recommenders under investigation with input lists that differ in their level of diversity and then measuring their output diversity.

Finally, we acknowledge that the algorithmic qualities (e.g., accuracy, diversity) of a set of recommendations may not always match users' perceptions of these qualities [21]. Hence, we contextualize our results by analyzing the reviews of the five considered streaming services in the Google Play Store, focusing on users' perception of their recommender systems and the diversity of their output.

2 RELATED WORK

2.1 The Algorithms of Music Streaming Services

Like most recommender systems, existing music streaming services make use of content-based filtering (CB) and collaborative filtering (CF) to recommend songs to their users [18]. For instance, Spotify's algorithm primarily employs CF to suggest songs but also makes use of audio models and natural language processing which both examine the characteristics of songs, indicative of CB [6]. This hybrid approach is commonplace in music streaming services and can help alleviate the cold-start problem [12]. Youtube, Last.FM, and Apple also make use of a hybrid recommender system that

IUI '22, March 22-25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9145-0/22/03.
https://doi.org/10.1145/3490100.3516474

favors CF [2, 16, 20] . Pandora, however, primarily employs CB through their Music Genome Project [4].

2.2 Diversity

In recommender systems, *diversity* refers to how different each recommendation is from the other recommendations. Research has shown that different algorithms result in different levels of recommendation diversity [13]. Diversity is desirable for numerous reasons. Using offline evaluations, Burke et al. show that tailoring recommendations to users' preferred level of diversity increases overall accuracy metrics [8]. Moreover, from the user's subjective perspective, Willemsen et al. demonstrate that diversity is key to overcoming choice overload [21]. This might explain why Spotify users who received a higher diversity of content had greater rates of conversion and retention [1], and why users who consumed more diverse content were more likely to purchase a premium subscription.

3 METHODS

Our study considers both the "objective diversity" of the recommendations and users' perceptions of the recommendations of five music streaming services: Spotify, Pandora, YouTube Music, Apple Music, and Last.fm.

3.1 Measuring The Objective Diversity of the Streaming Services

Three playlists of 20 songs with differing levels of diversity were created to serve as input signal. The "Low Diversity" playlist exclusively consists of songs from the late 70's to early 80's and only contains songs from 5 artists from one specific sub-genre—disco. The "Medium Diversity" playlist consisted of songs of one general genre—rock—but contained songs spanning multiple sub-genres, 50 years of music history, and songs from 10 different artists. The "High Diversity" playlist contained songs from 20 different artists of completely different genres and time periods—varying from Liszt to Drake. These songs were then loaded into a playlist and played in each of the five streaming services. Separate accounts were used for each playlist to avoid recommendation bleed-over.[1] Then the first 30 songs that were algorithmically suggested by each app were recorded. The three playlists and the 15 lists of 30 recommendations can be reviewed at https://usabart.nl/playlists.pdf.

We used a Matrix Factorization (MF) algorithm trained on the Spotify Million Playlist Challenge Dataset [5] to calculate a similarity coefficient for each playlist and each set of recommendations. Particularly, the MF algorithm was used to extract 50 latent features for each item. Then cosine similarity was calculated between the feature sets of each pair of songs in the set and averaged over all pairs. The first column of Table 1 shows the similarity coefficients of our low, medium, and high diversity input playlists to be 0.510, 0.259, and 0.084, respectively, demonstrating that the construction of these lists was successful in creating lists of differing levels of diversity.

[1]Because Apple Music is a paid service, only one account was made, but the playlists were made separately to avoid any bleed-over.

3.2 Analyzing Users' Perceptions

The field of recommender systems has largely acknowledged that the objective algorithmic qualities of a recommender system do not paint a full picture of how users experience the system and the recommendations it provides [10, 11, 15]. While user experiments constitute the "gold standard" to evaluate the user experience of recommender systems [9], we leverage existing online discourse around the five streaming services to efficiently gain a preliminary understanding of users' perception of each services' recommender system, focusing on the perception of recommendation diversity. To this effect, we crawled the 200 most recent reviews for each streaming service from the Google Play Store, capturing the review date, star rating, and the content of each review. Reviews that mention the recommendations and/or the recommender system were separated for further qualitative analysis.

4 RESULTS

4.1 Objective Diversity

Table 1 shows the diversity of the first 30 recommendations produced by each streaming service based on each of the three input playlists. On average, Spotify's recommendations are the least diverse, with similarity coefficients of 0.681 for the low diversity list, 0.538 for the medium diversity list, and 0.217 for the high diversity list. These output lists are consistently less diverse than the input playlists themselves. In contrast, YouTube Music produces the most diverse suggestions, with similarity coefficients of 0.446, 0.329, and 0.092 for the low-, medium- and high-diversity input lists, respectively. These similarity scores very closely match the scores of the input playlists.

Most services tend to "match" the level of diversity of the playlist given to the service, with similarity coefficients decreasing between the low-, medium-, and high-diversity input lists (except for the higher similarity coefficient of the Last.fm recommendations based on the high diversity input list). In effect, each service is able to cater to users' preferred level of diversity, as expressed in the songs they listen to organically (cf. [8]).

4.2 Users' Perceptions

Among the 1000 analyzed reviews, the most common review topics[2] were issues with the app (312), followed by discussion of features other than the recommender (239), and in-stream advertisements (47). Only 27 reviews discussed the recommender system and/or the recommendations provided by the music streaming service.

The small number of reviews about the recommendations prevents us from comparing users' perceptions between apps. Instead, we examined the reviews that explicitly discussed the recommendations to present general trends across all apps. 8 of the 27 reviews opined that they did not want the streaming service to give them *any* music suggestions—these users simply wanted to listen to the music they manually selected. One such review stated: "I cannot even play the songs I want as Spotify plays random songs that it recommended." In contrast, another 8 reviews praised the effectiveness of the algorithms in helping them find new music or re-find

[2]Note that a large number of reviews did not contain content that could easily be categorized.

Table 1: Similarity Coefficients of Music Services

	Input playlists	Recommendation output					
		Spotify	Pandora	Apple	YouTube	Last.fm	Average
Low	0.510	0.681	0.606	0.448	0.446	0.427	0.522
Medium	0.289	0.538	0.282	0.313	0.239	0.236	0.322
High	0.084	0.217	0.195	0.101	0.092	0.266	0.174
Average		0.479	0.361	0.287	0.259	0.310	

old hits they had heard before. Additionally, 6 reviews mentioned that the recommendations felt repetitive, and that the algorithm made them listen to many of the same songs/artists repeatedly. The latter suggests a dissatisfaction with a lack of diversity.

5 DISCUSSION AND CONCLUSION

While the recommendations provided by the streaming services studied in this paper differ in terms of diversity, each service seemed to attempt to match the level of diversity of the recommendations to the diversity of the input playlist. This could be the outcome of a deliberate attempt to match the user's preference for diversity (cf. [8]), but it could also be a result of a standard recommendation algorithm. While *accuracy* is often seen as antithetical to diversity [3], *precision* and *recall* may favor recommendations that match the diversity of the recommendations to the user's profile: if a user's profile contains more diverse items, precision and recall will improve with more diverse recommendations.

One limitation of our work is that we do not know the exact nature of the underlying algorithms of each recommendation service. However, we note that the most diverse service (YouTube) and the least diverse service (Spotify) both use a hybrid recommender with an emphasis on collaborative filtering [6, 20]. Additionally, Pandora, which uses a primarily content-based algorithm [4], did not produce recommendations that were discernibly more or less diverse than the other services. Thus, it is likely that the particular way suggestions are made is more important than the general mechanism by which recommendations are calculated. Future work could delve deeper into how different recommendation techniques may result in different levels of diversity [13]. Additionally, this study only tested one playlist for each level of diversity, future work could use multiple playlists of each level of diversity or use playlists made by real users rather than artificially designed ones.

Our study of Google Play Store reviews did not return enough reviews about the recommender system or the recommendations to allow for a comparison between services. Future work could delve deeper into existing reviews and/or supplement these results with user surveys. The lack of reviews about recommendations suggests that consumers are either unaware of the recommendation algorithms, or did not feel that the quality of the recommendations was a reason to leave a review. Among the reviews that did mention recommendations, a notable proportion expressed a desire to receive no recommendations at all. The idea that consumers would not want an algorithm to control their listening habits is intuitive (cf. [14]), but this desire is not recognized by any of the studied streaming services, who tend to consider their algorithm as a primary way to improve retention [17]. While consumers can organically search for songs and playlists, each service provides suggestions with no way to disable this—even for premium subscribers. Future work could study the benefits of allowing users to disable the recommender feature. Additionally, the disagreement between reviews on whether the algorithm was helpful or repetitive could represent an inadequacy in the algorithm to suggest at the preferred level of diversity for each consumer.

ACKNOWLEDGMENTS

This work was partially supported by NSF award IIS-2045153. Special Thanks to Mrs. Tidwell

REFERENCES

[1] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. *Algorithmic Effects on the Diversity of Consumption on Spotify.* Association for Computing Machinery, New York, NY, USA, 2155–2165. https://doi.org/10.1145/3366423.3380281

[2] Apple Inc. (n.d.). Apple Music. https://www.apple.com/apple-music/

[3] Keith Bradley and Barry Smyth. 2001. Improving recommendation diversity. In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*, Vol. 85. 141–152.

[4] Oscar Celma and Paul Lamere. 2011. Music recommendation and discovery revisited. In *Proceedings of the fifth ACM conference on Recommender systems (RecSys '11).* Association for Computing Machinery, New York, NY, USA, 7–8. https://doi.org/10.1145/2043932.2043936

[5] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. 2018. Recsys Challenge 2018: Automatic Music Playlist Continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18).* Association for Computing Machinery, New York, NY, USA, 527–528. https://doi.org/10.1145/3240323.3240342

[6] Sophia Ciocca. 2017. How Does Spotify Know You So Well? https://medium.com/s/story/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe

[7] Stuart Dredge. 2020. *How many users do Spotify, Apple Music and streaming services have?* https://musically.com/2020/07/28/spotify-apple-how-many-users-big-music-streaming-services/

[8] Farzad Eskandanian, Bamshad Mobasher, and Robin Burke. 2017. A Clustering Approach for Personalizing Diversity in Collaborative Recommender Systems. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17).* Association for Computing Machinery, New York, NY, USA, 280–284. https://doi.org/10.1145/3079628.3079699

[9] Bart P. Knijnenburg and Martijn C. Willemsen. 2015. Evaluating Recommender Systems with User Experiments. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, 309–352. https://doi.org/10.1007/978-1-4899-7637-6_9

[10] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[11] Joseph A. Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (April 2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x

[12] Le Duong Lam, Vu. 2008. Adressing cold-start problem in recommendation systems. In *ICUIMC '08: Proceedings of the 2nd international conference on Ubiquitous information management and communication.* 208–211. https://doi.org/10.1145/1352793.1352837

[13] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society* 21, 7 (2018), 959–977.

[14] Eli Pariser. 2012. *The filter bubble: how the new personalized Web is changing what we read and how we think.* Penguin Books, New York, N.Y.
[15] Hu Pu, Chen. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/doi.org/10.1145/2043932.2043962
[16] Róbert Pálovics and András A. Benczúr. 2015. Temporal influence over the Last.fm social network. *Social Network Analysis and Mining* 5, 1 (Jan. 2015), 4. https://doi.org/10.1007/s13278-014-0244-y
[17] Nick Seaver. 2018. Captivating algorithms: Recommender systems as traps. *Journal of Material Culture* (Dec. 2018), 1359183518820366. https://doi.org/10.1177/1359183518820366
[18] Yading Song, Simon Dixon, and Marcus Pearce. 2012. A Survey of Music Recommendation Systems and Future Perspectives. 9th International Symposium on Computer Music Modeling and Retrieval, London, UK.
[19] Statista. 2021. *Global streaming music subscribers 2020.* https://www.statista.com/statistics/669113/number-music-streaming-subscribers/
[20] John Paul Titlow. 2013. Inside Google's Infinite Music Intelligence Machine. https://www.fastcompany.com/3014183/inside-googles-infinite-music-intelligence-machine
[21] Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. 2016. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction* 26, 4 (Oct. 2016), 347–389. https://doi.org/10.1007/s11257-016-9178-6