EXPLORING POPULARITY BIAS IN MUSIC STEAMING SERVICES

Vera Crabtree
Ithaca College
vcrabtree99@gmail.com

Sean McQuillan
Ithaca College
mcquill99@gmail.com

Douglas Turnbull
Ithaca College
dturnbull@ithaca.edu

ABSTRACT

Popularity bias is the idea that a music recommender system will unduly favor popular artists when recommending artists to users. In this paper, we attempt to measure popularity bias on three commercial music streaming services (Spotify, Amazon Music, YouTube). We find no significant evidence of popularity bias in the commercial recommendations based on a simulated user experiment.

1. INTRODUCTION

In a recent Rolling Stone article by Emily Blake [1], she found that streaming music services have created steeper long-tail distribution in which a small number of superstar artists receive more attention from listeners when compared to physical album sales.

There could be many reasons this growing inequity is found on music streaming services. First, commercial services employ music recommendation systems to create personalized playlists and radio streams for their listeners. Recent research on fairness in recommender systems has revealed that many recommender system algorithms are subject to *popularity bias* [2, 3]. That is, recommender systems re-enforce a feedback cycle in which popular items get recommended disproportionately more often and thus become even more popular. Another potential cause of popularity bias may be related to the streaming service business practices. For example, record labels can pay streaming services to feature their songs on popular human-curated playlists (e.g., New Music Friday lists on Spotify) [4].

Regardless of the cause, popularity bias, when combined with the concept of *the mere exposure effect* in which listeners prefer familiar songs [5, 6], leads to a rich-get-richer marketplace for music consumption. Songs with unfair initial exposure get picked up by listeners and crowd out other songs which may have been preferred by the listener in a counterfactual setting [7]. This limits consumer awareness and prevents a larger group of artists from being discovered and supported.

2. RELATED WORK

Traditionally, recommender systems are designed to provide the most relevant items to each individual user. However, in recent years, researchers have started to view recommendation as a *mutli-stakeholder* problem in which we consider the needs of two or more groups of individuals [8]. In the context of music recommendation [9], we can think of both listeners, artists, and streaming services as being stakeholders in a multi-sided marketplace. While there are many potential forms of unfair bias, such as gender bias [10, 11], our work specifically focuses on popularity bias in commercial music streaming sites.

We have long known that music consumption follows a long-tail distribution [12] in which a small group of artists receive the vast majority of the attention from listeners. Research by Celma and Cano [13], Kowald et al. [14], and our own work [15], suggests that there is *algorithmic* popularity bias for many state-of-the-art recommender system algorithms. However, Levy and Bosteels [16] conducted an analysis of the effect of music recommendations from Last.fm on user listening habits and found little evidence of *commercial* popularity bias. Building on this initial work, we update and expand the research on popularity bias in commercial streaming services by directly comparing multiple modern services (Spotify, YouTube, Amazon Music) using a *simulator user* experimental design [10].

3. SIMULATED USER EXPERIMENT

We consider some of the most popular streaming services in the United States based on data we gathered from Statista [1] . These services are Spotify, Amazon Music, and YouTube Music.

We created a set U of twelve simulated users $u \in U$ based on real user data from the LFM-1B-Subset [14]. We randomly-selected four users within each of the three subgroups: low mainstream, medium mainstream, and high mainstream users. For each user, we consider the user's *profile* P_u as the top 10 most-listened to artists $a \in P_u$ according to the LFM-1B-Subset. For each of artist, we consider two different popularity measures: a proprietary Spotify popularity score [2] $\phi_s(a)$ which ranges from 0 to 100 and a LFM-1B-Subset score $\phi_{lfm}(a)$ which represents the proportion of users who had listened to artist a out of the 3000 users in the data set.

[1] https://www.statista.com/statistics/758875/consumers-use-music-streaming-download-services/ on May 15, 2021

[2] See the ArtistObject at https://developer.spotify.com/documentation/web-api/reference/#objects-index

For each of our three streaming music services, we create a new *simulated user account* using the service's web or mobile app. From each of these accounts, we "follow" or "like" (depending on the service) the top-ten artists P_u for the user u. We then play each of the artists' top song once all the way through. After "listening" to these ten songs once, the account was logged out of and returned to the next day to analyze the given recommendations.

In a round robin style, we record a set of 10 top recommended artists R_u for each of the generated mixes (e.g., Daily Mixes on Spotify) for each account. We start with the first personalized playlist (e.g., Daily Mix 1) and record the first recommended artist. We then go to the next playlist (e.g., Daily Mix 2) and note the first recommended artist from that playlist. This is repeated for all of the personalized playlists. After one pass through, we return to the first playlist and take note of the second recommended artist not already in R_u. This process was repeated until there were a total of ten recommended artists ($|R_u|$ = 10) for each simulated user account.

We calculate two Group Average Popularity (GAP) statistics, one for artists in the user profile GAP_P and one for the recommend artists GAP_R for an entire set of 12 users:

$$GAP_P(U) = \frac{\sum_{u \in U} \frac{\sum_{a \in P_u} \phi(a)}{|P_u|}}{|U|}$$

$$GAP_R(U) = \frac{\sum_{u \in U} \frac{\sum_{a \in R_u} \phi(a)}{|R_u|}}{|U|}$$

where u is a user, U is our set of 12 users (i.e., |G| = 12), P_u is the users top-ten most played artists (i.e., $|P_u| = 10$), and $\phi(a)$ is one of our two measure of artist popularity (ϕ_S for Spotify or ϕ_{LFM} for LFM-1B-Subset). These GAP statistics effectively measure the average artist popularity for a set of user profiles (i.e., inputs) and a set of recommended artists (i.e., outputs) for a music recommendation system.

To measure the change of the average popularity between the user profile inputs and the recommendation outputs, we compute ΔGAP, which is defined as the popularity lift in artists recommended over artists in the user profiles. ΔGAP is calculated as [3]:

$$\Delta GAP(U) = \frac{GAP(U)_R - GAP(U)_P}{GAP(U)_P}. \qquad (1)$$

We would expect ΔGAP to be 0 when the average popularity of the artists recommended by the music service is equal to the average popularity of artists the users listen to on that service. ΔGAP will be greater than 0 when there is popularity bias in the model since the model is recommending more popular artists than the users listen to.

3.1 Discussion of Results

As shown in table 1, our simulated user experiment reveals that there is a slight *negative* popularity bias according the LFM-1B-Subset based ΔGAP metric for each of the three music streaming services. Furthermore, when using the Spotify-based popularity, there is no observed popularity bias for Spotify or Amazon Music and only a slight popularity bias for YouTube Music.

ϕ_S- **Spotify Popularity**			
	Spotify	Amazon	YouTube
Overall ΔGAP	0.00	-0.13	0.06
Low MS ΔGAP	0.00	-0.29	0.10
Medium MS ΔGAP	0.02	-0.07	0.11
High MS ΔGAP	-0.01	-0.05	-0.01
ϕ_{LFM} - **LFM-1B-Subset Popularity**			
	Spotify	Amazon	YouTube
Overall ΔGAP	-0.22	-0.32	-0.12
Low MS ΔGAP	-0.37	-0.74	0.10
Medium MS ΔGAP	-0.33	-0.21	-0.14
High MS ΔGAP	-0.10	-0.26	-0.19

Table 1. Popularity Bias in three Commercial Streaming Services (Spotify, Amazon Music, YouTube Music). The Overall ΔGAP scores are calculated from 12 *simulated* users. Simulated users are created from randomly selected real users in the LFM-1B-Subset.

This pattern was also consistent across the three types of low, medium, and high mainstream users, the only exception to this being slight popularity bias on YouTube Music. However, the magnitude of this bias is not statistically significant (p=0.09, 1-tailed t-test) suggesting that this difference may be due to random variation. To summarize, we do not find any evidence to support the hypothesis that a subgroup of users is likely to experience popularity bias (i.e., positive ΔGAP) due to personalized recommendation on any of the three steaming services in our study.

4. DISCUSSION

We had expected to find evidence to support the hypothesis that personalized music recommendation plays a role in the accelerating rich-get-richer phenomenon for music consumption as described by Blake's Rolling Stone article [1]. However, our results did not find a evidence of popularity bias in music recommendations from three popular commercial streaming services (Section 3.1.) This result is consistent with the finding from Levy and Bosteels [16] who similarly found little evidence of popularity bias in music recommendations from Last.fm radio listeners.

This unexpected conclusion may be due to various limitations in our experimental design. First, we have a relatively small data set in terms of both simulated users and artists when compared to the amount of user data collected by Spotify, Amazon Music, and YouTube Music. Similarly, our simulated user experiment involved creating simplistic user profiles with a small amount of short-term artist preference information. While we do not have a detailed understanding of the proprietary process for generating recommendations on each commercial music service, we suspect that there are many more inputs beyond artist listening histories (e.g., s contextual information [17]). Finally, we note that while we found no evidence of popularity bias in personalized recommendation, popularity bias on commercial streaming services may result from the inclusion (or exclusion) of songs on popular human curiated playlists [4].

5. ACKNOWLEDGMENTS

This research was supported by NSF grant IIS-1901330/1901168. John Hunter and Sunny Zhang also contributed to this work. Links to the data used in the paper can be found at https://dougturnbull.org/.

6. REFERENCES

[1] E. Blake, "Data shows 90 percent of streams go to the top 1 percent of artists," *Rolling Stone*, 2020. [Online]. Available: https://www.rollingstone.com/pro/news/top-1-percent-streaming-1055005/

[2] D. Jannach, L. Lerche, I. Kamehkhosh, and M. Jugovac, "What recommenders recommend: an analysis of recommendation biases and possible countermeasures," *User Modeling and User-Adapted Interaction*, vol. 25, no. 5, pp. 427–491, 2015.

[3] H. Abdollahpouri, M. Mansoury, R. Burke, and B. Mobasher, "The unfairness of popularity bias in recommendation," *arXiv preprint arXiv:1907.13286*, 2019.

[4] L. Aguiar and J. Waldfogel, "Platforms, power, and promotion: Evidence from spotify playlists," *The Journal of Industrial Economics*, vol. 69, no. 3, pp. 653–691, 2021.

[5] I. Peretz, D. Gaudreau, and A.-M. Bonnel, "Exposure effects on music preference and recognition," *Memory & cognition*, vol. 26, no. 5, pp. 884–902, 1998.

[6] A. C. Green, K. B. Bærentsen, H. Stødkilde-Jørgensen, A. Roepstorff, and P. Vuust, "Listen, learn, like! dorsolateral prefrontal cortex involved in the mere exposure effect in music," *Neurology research international*, vol. 2012, 2012.

[7] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," *science*, vol. 311, no. 5762, pp. 854–856, 2006.

[8] R. Burke, "Multisided fairness for recommendation," *arXiv preprint arXiv:1707.00093*, 2017.

[9] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas, "Music recommender systems," in *Recommender systems handbook*. Springer, 2015, pp. 453–492.

[10] M. Eriksson and A. Johansson, "Tracking gendered streams," *Culture Unbound. Journal of Current Cultural Research*, vol. 9, no. 2, pp. 163–183, 2017.

[11] A. Epps-Darling, R. T. Bouyer, and H. Cramer, "Artist gender representation in music streaming," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (Montréal, Canada)(ISMIR 2020). ISMIR*, 2020, pp. 248–254.

[12] C. Anderson, "The long tail," *Wired Magazine*, 2004. [Online]. Available: http://www.wired.com/wired/archive/12.10/tail.html

[13] Ò. Celma and P. Cano, "From hits to niches? or how popular artists can bias music recommendation and discovery," in *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, 2008, pp. 1–8.

[14] D. Kowald, M. Schedl, and E. Lex, "The unfairness of popularity bias in music recommendation: A reproducibility study," in *European Conference on Information Retrieval*. Springer, 2020, pp. 35–42.

[15] D. R. Turnbull, S. McQuillan, V. Crabtree, J. Hunter, and S. Zhang, "Exploring popularity bias in music recommendation models and commercial steaming services," 2022. [Online]. Available: https://arxiv.org/abs/2208.09517

[16] M. Levy and K. Bosteels, "Music recommendation and the long tail," in *1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain*. Citeseer, 2010.

[17] M. Kaminskas and F. Ricci, "Contextual music information retrieval and recommendation: State of the art and challenges," *Computer Science Review*, vol. 6, no. 2-3, pp. 89–119, 2012.