

Better than Bieber?

Measuring Song Quality Using Human Feedback

Sasha Stoikov¹

¹Cornell Financial Engineering Manhattan, New York, United States

Abstract

Music recommendation algorithms on streaming platforms tend to reinforce popularity biases over time. This leads to a perceived unfairness for a majority of artists. In this article we address the fairness problem through an assessment of actual user preferences collected on a music app designed to capture the intensity of user sentiment. Users of the app provide explicit feedback on songs, designating them as dislike, like, and superlike (a function which saves songs to a playlist). We find that while the like rate increases monotonically with artist popularity, this does not hold true for superlike rates — those are highest for artists with much lower monthly streams. These findings have implications for the growth potential in the music market and the legitimacy of current recommendation algorithms.

Keywords

Algorithmic auditing, Fairness, Music Platforms, Recommendation Systems, Information Retrieval

1. Introduction

Why are some songs played billions of times, while others remain unheard? Algorithms on streaming platforms have the power to make or break songs, in ways that may seem unfair and opaque, particularly to the artists whose songs have not gone viral. Recommendation systems are prone to feedback loops where popular songs are recommended disproportionately more often than undiscovered ones. Simulation studies have shown that systems trained on data produced by users exposed to recommendations can lead to less diversity and lower utility from the perspective of users of these platforms [1] [2]. From the perspective of artists, field studies have found that popularity bias and lack of transparency in music recommendation engines are perceived as a major source of unfairness [3] [4]. Since most streaming platforms pay a fraction of a penny per stream, exposure is at the heart of the cultural and economic capital of an artist. For this reason, exposure fairness, ranking fairness and popularity bias have been studied extensively in the literature [5] [6] [7] [8] [9] [10] [11] [12]. Other notions of fairness related to gender, diversity and genres have also received attention [13] [14] [15] [16] but there is no consensus among artists as to how they should be addressed [3].

The concept of algorithmic fairness has been studied by scholars in other fields like college admissions or mortgage applications. [17] [18] define algorithmic unfairness in terms of instances where a candidate of higher quality is ranked lower than someone with lower


HCMIR23: 2nd Workshop on Human-Centric Music Information Research, November 10th, 2023, Milan, Italy

✉ sfs33@cornell.edu (S. Stoikov)

ORCID 0000-0002-0540-5846 (S. Stoikov)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

quality. However, in a domain as subjective as music, the notion of quality needs to be properly defined. To determine if the popularity of a given artist is fair, exposure data obtained on streaming platforms is not enough: explicit human feedback is required. What percentage of people like the top hits? What percentage of people love them? If a lesser known artist is more broadly liked and more deeply loved, it may be possible to establish that they have been unfairly under-exposed. In this paper, we address these questions by proposing a measure of quality for music, using an explicit music ratings dataset [19] collected by Piki, a music ratings app.

User generated datasets fall into two broad categories: explicit datasets, where the users express their opinions on the quality of an item and implicit datasets, where user behavior is surveilled by a platform and the opinions are implied by the behavior of the user. Well-known explicit datasets include the Yelp restaurant dataset [20] and the Movielens dataset [21]. Since feedback is voluntary and since users typically have access to a search bar, this kind of dataset is often subject to self-selection bias: if there is no obligation to rate, users tend to rate only when they feel very strongly about an item, and rate more items they like than items they dislike. Likewise, implicit datasets commonly used to train modern recommendation systems on streaming apps like Spotify [22] will tend to collect more positive signals, i.e. streams of a song, than negative signals, say song skips, which are not necessarily expressing an opinion about a song. The Piki music dataset studied in this paper has been shown to mitigate these positive biases, which have been shown to increase the accuracy of recommendations algorithms [23].

The paper is organized as follows. In section 2, we describe the Piki music interface and dataset. In section 3, we aggregate this data into two quality metrics, validate their consistency across populations and compare them for a range of artist popularity and song release dates. We conclude in section 4.

2. Data collection

The circumstances motivating Piki users to give ratings are very different from those of users of other ratings or streaming apps.

1. Piki users are presented with sets of 30-second music video clips which they rate while they are listening to the music. This is in contrast to restaurant and movie rating apps where there may be a significant time lag between the experience and the rating. Note that some of the songs have music videos while others only have audio. The start time is in the middle of the songs and may be different from song to song.
2. Users must provide explicit ratings, disliking, liking, or superliking a song – until they do so, the clip plays in a loop. Immediately after a rating is provided, the user is presented the next song. This is in contrast with streaming apps where implicit actions such as listens, skips, shares and saves to a playlist are a noisy reflection of a user’s tastes.
3. Users are paid small amounts of cash, to rate a large amount of songs. This incentivizes them to rate, even if some of the songs are not to their liking. Since streaming apps aim to maximize user retention, they are likely to recommend songs that are predictably likely to please, not more surprising songs that are outside of their comfort zone.
4. Users don’t have access to a search bar, which could naturally lead to users self-selecting artists that first come to mind, often celebrities, thus leading to a popularity bias. The

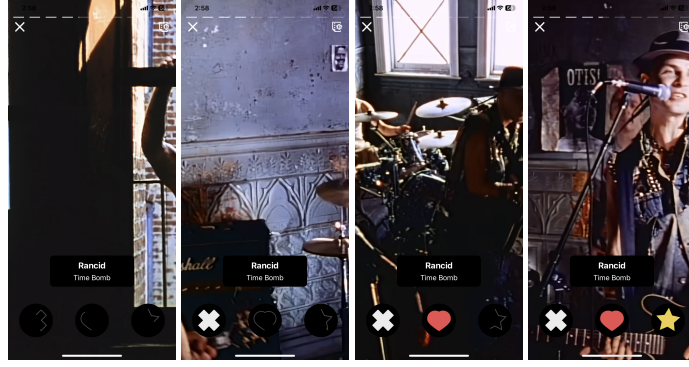


Figure 1: The dislike, like and superlike buttons are progressively unlocked after 4, 6 and 10 seconds

distribution of songs in Piki are approximately 16% random (from a list of songs with high superlike rates), 76% personalized (a collaborative filtering algorithm) and 8% hyper-personalized (an algorithm that presents songs by artists the user has already superliked).

5. A system of timers incentivizes users to rate more uniformly, despite the very diverse behaviors of people rating a very subjective media like music. Note that the dislike, like and superlike buttons appear sequentially in that order, several seconds after the song begins playing. Figure 1 illustrates the way the timers work. The timer on the dislike button ensures that each song is given a fair listen by the rater. The timers on the like and superlike buttons ensure that users are willing to invest time in their most preferred songs. The timers may also be slowed down to throttle users who dislike indiscriminately rate to get rewards.
6. Superliked songs are saved to a playlist that the user may export to other applications. This indicates that they are invested in listening to the song again soon.

The Piki dataset [19] has a similar structure to datasets produced by explicit ratings apps like Movielens and Yelp, except for each user id and song id, there are 3 possible ratings (dislike, like and superlike), instead of 5. The dataset consists of 1.5 million ratings on 231,800 songs rated by 7,519 users. The catalog was built by querying top songs from artists popular on streaming and live music apps. The songs presented on Piki have a diversity in artist Spotify popularity (Figure 2) and release decade (Figure 3).

3. Song quality measures

Since there are 3 possible ratings and they appear in sequential order, we define two natural conditional probabilities, the like rate and the superlike rate. If n_D^i , n_L^i and n_S^i are the number of dislikes, likes and superlikes for a song i , we define the like rate for song i as:

$$L^i = \frac{n_L^i + n_S^i}{n_D^i + n_L^i + n_S^i} = P(\text{like}|\text{listen})$$

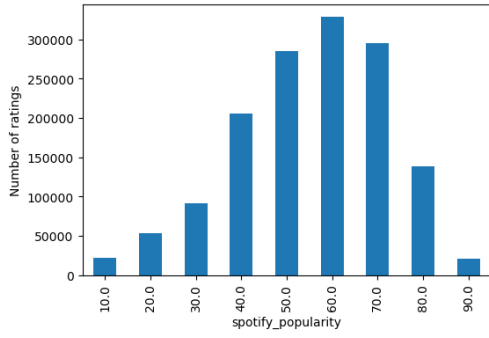


Figure 2: Ratings distribution by popularity

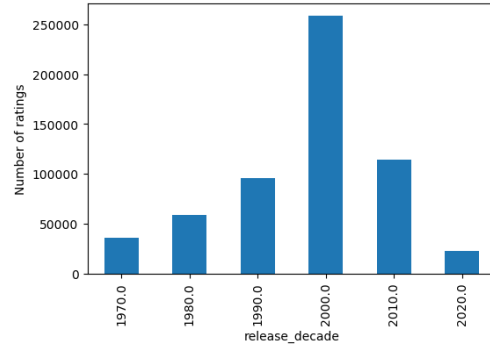


Figure 3: Ratings distribution by decade of release

this quantity represents the conditional probability that a song is liked (or superliked), given that it is listened to.

The superlike rate for song i is:

$$S^i = \frac{n_S^i}{n_L^i + n_S^i} = P(\text{superlike}|\text{like})$$

represents the conditional probability that a song is superliked, given that it is liked. Since superlikes require users to invest more time and the songs are saved to a playlist, we can assume that they indicate a more significant valuation on the part of the user.

In Figure 4, we compute the like rates of songs based on their first 100 ratings and plot them against the like rates based on the next 100 ratings for that same song (note that we only considered songs with more than 200 ratings of which there are 669). This shows that despite the diversity of tastes of listeners, these two quality metrics are consistent: if a song scores highly among 100 listeners, it will score highly among the next 100 listeners.

Having shown consistency in the like and superlike rates for individual songs, we now aggregate like and superlike rates by grouping by decade of release and Spotify popularity. Both the like rates and superlike rates are increasing with age (see Figure 5), possibly due to a survivorship bias. Notice in Figure 6, the like rate is monotonic in artist popularity: this seems to justify the popularity of the most streamed artists with more than 70 million monthly listeners (the likes of Justin Bieber, Taylor Swift, Drake, Doja Cat, Bad Bunny and Ed Sheeran), according to this quality metric. More surprisingly, the superlike rates seem to peak at a popularity of 70, which typically corresponds to artists with a few million monthly listeners (the likes of Krewella, Louis Tomlinson, The Marias, Khruangbin, Sting, Raekwon and Kaytranada). These results seem to indicate that like rates measure a type of familiarity which grows with popularity, while superlike rates measure a type of excitement that comes from deeper discovery, which is likely to happen for lesser known artists. If that is the case, many of these middle class artists may be more deserving of exposure than many artists in the 90-100 popularity range.



Figure 4: Metrics consistency

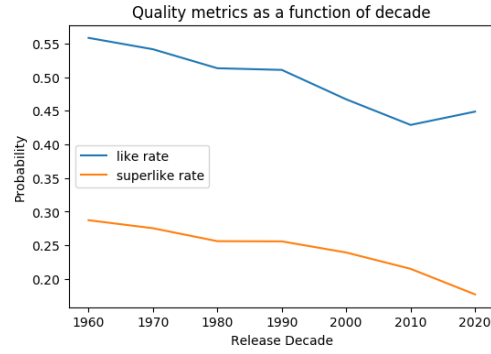


Figure 5: Metrics by decade of release



Figure 6: Metrics by artist popularity

4. Conclusion

Many fairness metrics have been proposed in the information retrieval literature [5] [11] and, from the point of view of algorithmic design, fairness to producers is often framed as a constraint on the utility of the consumers. In this paper, we show empirically how fairness to artists may not need to come at the expense of music listeners. The fact that some under-exposed artists are superliked at higher rates than top celebrities, indicates that an algorithmic system that gives these artists more exposure may increase the utility of its users.

Establishing fairness of exposure on streaming apps requires transparency. Artists will only feel the system is fair, if they understand why some songs go viral, while others don't. We propose achieving this transparency by defining metrics based on human feedback, with strict principles on the interfaces collecting the feedback and consistency tests on the quality metrics. Only then can we deploy them as auditing or regulatory tools to keep algorithms accountable.

The notion of quality metrics for works of art is likely to be controversial. The old saying that “in matters of taste, there can be no disputes” will always rear its head. Despite the challenge in defining tools for assessing creative works, fairness for the creative class can only be achieved if artists and platforms agree on transparent notions of music quality. We argue in this paper, that these metrics need to focus on aggregating critical human opinions, rather than simply

measuring past exposure and replicating it into the future.

References

- [1] A. J. B. Chaney, B. M. Stewart, B. E. Engelhardt, How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility, in: Proceedings of the 12th ACM Conference on Recommender Systems, 2018, pp. 224–232. URL: <http://arxiv.org/abs/1710.11214>. doi:10.1145/3240323.3240370, arXiv:1710.11214 [cs, stat].
- [2] A. Ferraro, D. Bogdanov, X. Serra, J. Yoon, Artist and style exposure bias in collaborative filtering based music recommendations, 2019. arXiv:1911.04827.
- [3] K. Dinnissen, C. Bauer, Fairness in Music Recommender Systems: A Stakeholder-Centered Mini Review, *Frontiers in Big Data* 5 (2022) 913608. URL: <https://www.frontiersin.org/articles/10.3389/fdata.2022.913608/full>. doi:10.3389/fdata.2022.913608.
- [4] A. Ferraro, X. Serra, C. Bauer, What is fair? Exploring the artists' perspective on the fairness of music streaming platforms, 2021. URL: <http://arxiv.org/abs/2106.02415>, arXiv:2106.02415 [cs].
- [5] A. Raj, M. D. Ekstrand, Comparing fair ranking metrics, 2022. arXiv:2009.01311.
- [6] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The unfairness of popularity bias in recommendation, 2019. arXiv:1907.13286.
- [7] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, B. Carterette, Evaluating Stochastic Rankings with Expected Exposure, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 275–284. URL: <http://arxiv.org/abs/2004.13157>. doi:10.1145/3340531.3411962, arXiv:2004.13157 [cs].
- [8] A. Singh, T. Joachims, Fairness of Exposure in Rankings, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2219–2228. URL: <http://arxiv.org/abs/1802.07281>. doi:10.1145/3219819.3220088, arXiv:1802.07281 [cs].
- [9] A. J. Biega, K. P. Gummadi, G. Weikum, Equity of Attention: Amortizing Individual Fairness in Rankings, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 405–414. URL: <http://arxiv.org/abs/1805.01788>. doi:10.1145/3209978.3210063, arXiv:1805.01788 [cs].
- [10] G. K. Patro, L. Porcaro, L. Mitchell, Q. Zhang, M. Zehlike, N. Garg, Fair ranking: a critical review, challenges, and future directions, 2022. URL: <http://arxiv.org/abs/2201.12662>, arXiv:2201.12662 [cs].
- [11] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, Y. Zhang, Fairness in recommendation: A survey, 2022. arXiv:2205.13619.
- [12] Z. Zhu, J. Kim, T. Nguyen, A. Fenton, J. Caverlee, Fairness among New Items in Cold Start Recommender Systems, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Virtual Event Canada, 2021, pp. 767–776. URL: <https://dl.acm.org/doi/10.1145/3404835.3462948>. doi:10.1145/3404835.3462948.
- [13] A. B. Melchiorre, N. Rekabsaz, E. Parada-Cabaleiro, S. Brandl, O. Lesota, M. Schedl, Investigating gender fairness of recommendation algorithms in the music domain, *Information*

- Processing & Management 58 (2021) 102666. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0306457321001540>. doi:10.1016/j.ipm.2021.102666.
- [14] A. Ferraro, X. Serra, C. Bauer, Break the Loop: Gender Imbalance in Music Recommenders, in: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, ACM, Canberra ACT Australia, 2021, pp. 249–254. URL: <https://dl.acm.org/doi/10.1145/3406522.3446033>. doi:10.1145/3406522.3446033.
 - [15] D. Shakespeare, L. Porcaro, E. Gómez, C. Castillo, Exploring Artist Gender Bias in Music Recommendation, 2020. URL: <http://arxiv.org/abs/2009.01715>, arXiv:2009.01715 [cs].
 - [16] A. Epps-Darling, R. T. Bouyer, H. Cramer, Artist gender representation in music streaming, In *Proceedings of the 21st ISMIR Conference*. International Society for Music Information Retrieval (2020).
 - [17] M. Kearns, A. Roth, Z. S. Wu, Meritocratic fairness for cross-population selection, in: D. Precup, Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1828–1836. URL: <https://proceedings.mlr.press/v70/kearns17a.html>.
 - [18] M. Joseph, M. Kearns, J. Morgenstern, A. Roth, Fairness in Learning: Classic and Contextual Bandits, 2016. URL: <http://arxiv.org/abs/1605.07139>, arXiv:1605.07139 [cs, stat].
 - [19] The piki dataset (2023). URL: <https://github.com/sstoikov/piki-music-dataset>.
 - [20] The yelp dataset (2023). URL: <https://www.yelp.com/dataset>.
 - [21] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, *ACM Trans. Interact. Intell. Syst.* 5 (2015). URL: <https://doi.org/10.1145/2827872>. doi:10.1145/2827872.
 - [22] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, *ICDM '08*, IEEE Computer Society, USA, 2008, p. 263–272. URL: <https://doi.org/10.1109/ICDM.2008.22>. doi:10.1109/ICDM.2008.22.
 - [23] S. Stoikov, H. Wen, Evaluating music recommendations with binary feedback for multiple stakeholders, in: H. Abdollahpouri, M. Elahi, M. Mansoury, S. Sahebi, Z. Nazari, A. Chaney, B. Loni (Eds.), *Proceedings of the 1st Workshop on Multi-Objective Recommender Systems (MORS 2021) co-located with 15th ACM Conference on Recommender Systems (RecSys 2021)*, Amsterdam, The Netherlands, September 25, 2021, volume 2959 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: <http://ceur-ws.org/Vol-2959/paper9.pdf>.