Exploring Artist Gender Bias in Music Recommendation

Dougal Shakespeare[1], Lorenzo Porcaro[1], Emilia Gómez[1,2], Carlos Castillo[3]
[1]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2]Joint Research Centre, European Commission, Seville, Spain
[3]Web Science and Social Computing Group, Universitat Pompeu Fabra, Barcelona, Spain
dougalian.shakespeare01@estudiant.upf.edu
{lorenzo.porcaro,emilia.gomez,carlos.castillo}@upf.edu

ABSTRACT

Music Recommender Systems (mRS) are designed to give personalised and meaningful recommendations of items (i.e. songs, playlists or artists) to a user base, thereby reflecting and further complementing individual users' specific music preferences. Whilst accuracy metrics have been widely applied to evaluate recommendations in mRS literature, evaluating a user's item utility from other impact-oriented perspectives, including their potential for discrimination, is still a novel evaluation practice in the music domain. In this work, we center our attention on a specific phenomenon for which we want to estimate if mRS may exacerbate its impact: *gender bias*. Our work presents an exploratory study, analyzing the extent to which commonly deployed state of the art Collaborative Filtering (CF) algorithms may act to further increase or decrease artist gender bias. To assess group biases introduced by CF, we deploy a recently proposed metric of bias disparity on two listening event datasets: the LFM-1b dataset, and the earlier constructed Celma's dataset. Our work traces the causes of disparity to variations in input gender distributions and user-item preferences, highlighting the effect such configurations can have on user's gender bias after recommendation generation.

CCS CONCEPTS

• **Social and professional topics** → *Socio-technical systems*; **Gender**; • **Information systems** → **Collaborative filtering**; **Recommender systems**.

KEYWORDS

gender bias, bias disparity, music recommendation

1 INTRODUCTION

Impact-oriented Recommender System (RS) research is gaining attention as a novel paradigm for understanding not only how users interact with recommendations, but also for shedding light on how these interactions can influence users' behaviours in the short- and the long-term [25]. An outstanding issue when studying the possible impact of RS is the heterogeneity of evaluation procedures described in the literature. Evaluating recommender systems is a non-trivial task because of the multiple facets that a *good* recommendation can have, and the multiple players influencing these aspects [20]. Even if the need for going beyond the evaluation in terms of accuracy metrics has been well-recognized by the RS community [32], shared practices for evaluating the impact of recommendations still are missing.

Notwithstanding, recent years have seen a rise in awareness in the scientific community about the implications of socio-technical systems' design and implementation responsible of reinforcing bias and discrimination [4, 42]. Music Information Retrieval (MIR) research is still in its early-stage with regards to the analysis of the ethical dimensions and impact of music technology [19, 22, 37, 39], and several challenges still need to be tackled when approaching MIR research from a socio-technical perspective. A common issue is the availability of data, often limited in terms of size, user information or musical information, and as in many other fields, a chronic shortage of gender-disaggregated data [35]. The difficulties in our research to retrieve the artists' gender are just one example of this limitation, as presented in Section 3 and 4.

We center our attention on a specific phenomenon that recommender systems may exacerbate: *gender bias*. In its broader sense, gender discrimination is a disadvantage for a group of people based on their gender. Far from being an emerging problem, gender discrimination has its roots in cultural practices historically related with socio-political power differentials [12]. Nonetheless, the modern day prevalence of gender discrimination is not to be understated: recent reports find the disproportionate treatment of female artists to be prevalent in the Western music industry to this day [1]. Whilst the cause of such treatment is multifaceted, our work traces the influence of one factor evidenced to be present in the works of Millar [33] that is, the pre-existing gender bias of a music listener.

In this exploratory study, we assess the extent to which Collaborative Filtering (CF) algorithms commonly deployed in mRS may exacerbate pre-existing users' gender biases thereby affecting an artist genderâĂŹs exposure and proportional representation. We focus on the measurement of bias disparity in recommender systems, defined as "*[...] the case where the recommender system introduces bias in the data, by amplifying existing biases and reinforcing stereotypes.*" [41]. Building on existing literature [29, 31, 41, 43], we first reproduce the study presented by Lin et al. [29], in which preference bias amplification in collaborative recommendation is analyzed using the MovieLens dataset[21], a dataset of user activity with a movie recommendation system. In our work, we focus on the music domain making use of two Last.fm[2] listening event datasets publicly available: 1) Celma's LFM-360k dataset [10]; 2) Schedl's LFM-1b dataset [38]. Our goal is twofold: on one hand, reproducing and verifying whether previous results [29] hold across different datasets. On the other hand, we aim at highlighting which aspects specific to the music domain can be extracted by this analysis, connecting with existing literature on gender bias in music preferences [3, 33].

[1]http://assets.uscannenberg.org/docs/aii-inclusion-recording-studio-2019.pdf
[2]https://www.last.fm

The paper is structured as follows. Section 2 provides an overview of previous works related to bias in Information Technology, focusing on gender bias, but also how this bias has been approached in music-related fields. We then introduce the considered datasets, LFM-1b and LFM-360K respectively in Section 3 and 4. In Section 5, the recommendation models used and the experimental settings are presented, followed by Section 6 which details the results obtained. Lastly, in section 7 conclusions and future work are discussed.

2 RELATED WORK

The notion of bias has been extensively explored in the Information Retrieval domain [4, 5, 7, 11, 24]. Typically, metrics aim to capture *relative bias* (i.e. bias pre-existing in data, for example in user listening histories in LFM-1b), and *algorithmic bias* (i.e. how filtering algorithms can result in unfair item and user treatment) to measure disproportionate unfair treatment of a protected group.

One of the most well-studied biases in RS literature is popularity bias, with the music domain being no exception to this phenomenon [6, 10, 28]. This describes the scenario in which a few popular items are recommended frequently, while the majority of items in the long-tail do not get proportional attention. Highlighted in literature as a prominent issue for CF algorithms [1, 10, 34], Kowald et al. in [28] find that from a user's perspective the groups who do not favor popular items may receive worsened recommendations in terms of accuracy and calibration. Moreover, Ferraro et al. in [18] study the effect of musical styles with respect to popularity bias, showing that CF approaches increase users' exposure to popular musical styles.

Bias Disparity is a metric deployed to assess bias propagation across user's and item's group, measuring the deviation of the recommender output from the input preference, as detailed in Section 5.1. A first application to the RS domain was described by Tsintzou et al. [41], but the metric has recently gained more traction in its application to different domains. In Lin et al. [29], bias disparity is applied to measure the extent to which state of the art CF algorithms can exacerbate pre-existing biases in the MovieLens dataset. Their findings show significant differences in bias propagation across memory- and model-based CF algorithms.

Gender treatment and issues of proportional treatment in RS have been considered in a range of literature, for which we highlight some examples. Ekstrand et al. [17] examined gender distribution of item recommendations in the book RS domain. Results prove that commonly deployed CF models differ in the gender distributions of generated item recommendation lists, such that neighbour-based approaches are shown to proportionality reflect user-item preferences in their reading histories, whereas model-based matrix factorisation favor books whose author is of male gender. Furthermore, Ekstrand et al. in [16] study the effect of recommendation algorithms on the utility for users of different gender groups, finding difference in effectiveness across gender groups. Such work highlights that the effect in utility does not exclusively benefit large groups, implying that there may be other underlying latent factors that influence recommendation accuracy. To address such issues of disproportionate gender treatment in recommendations, Edizel et al. in [15] have recently proposed a novel means of mitigating the derivation of sensitive features (such as gender) in the latent space, using fairness constraints based on the predictability of such features. A similar approach proposing fairness-aware tensor-based recommendation is also presented by Zhu et al. in [44].

In the music domain, Aguiar et al. [2] propose a methodology to assess the extent to which artists ranked in Spotify playlists are affected by gender after accounting for plausible determinants of inclusion on playlists such as country, song characteristics (e.g. bpm, key signature), and past streaming success. The authors find that there is some evidence consistent with the presence of bias (both for and against female artists), however they do not draw subsequent relations between this and the disproportionate low streaming share of female artists on the platform. In the work by Anglata-Tort et al. [3], through the analysis of UK top 5 music charts between the years 1960-1995, authors show how popular music is affected by a large gender inequality, showing the presence of an existing bias in the listening preferences towards male artists. Similarly, Millar in [33], surveying a population of Australian young adults, shows how music preferences are affected by gender bias, evidencing differences between male and female listeners. In contrast, in our work we apply an auditing strategy for bias propagation showing under which conditions input preferences are reflected in RS output, inferring music preferences from the users' listening history grouped with respect to the artists' gender.

3 THE LFM-1B DATASET

The LFM-1b dataset consists of more than one billion listening events created by over 120,000 users of the music streaming platform Last.fm [38]. In our analysis, we consider user-artist playcounts formed by aggregating user-song listening events by common artists. We then scale logarithmically the number of listens, as done in [13, 26]. We work with a filtered version of the dataset in which: a) we remove users who listened to less than 10 unique artists, and artists listened to by less than 10 users; b) we discard users whose listening history contains more than 25% of artists with unknown gender, to mitigate the impact of artists with missing gender in the dataset.

User gender is represented in the dataset with three categories: *male*, *female* and *N/A*. We choose to focus only on users with self-declared gender, working with two final categories of user gender: male and female. As shown in Table 1, distributions are highly imbalanced towards men – 72% of the users are men.

Artist gender is not represented in the LFM-1b dataset, consequently we retrieve this information from the open music encyclopedia MusicBrainz[3] (MB) [40]. Code repositories to implement the following approach are made openly available[4] alongside the acquired results of the data wrangling[5] to elicit reproducibility .

We identify five discrete categories of gender defined in the MB database: *male*, *female*, *other*, *N/A* and *undef*. In the case of artists of gender *N/A* and *undef*, these are differentiated by artists for which gender is not applicable and identifiable respectively. For bands, we compute gender counts of all members and then compute an overall classification based on whichever count has a majority. In the case of artists with gender ties (e.g, a band consisting of 2 males

[3] https://musicbrainz.org/
[4] https://github.com/dshakes90/LFM-1b-MusicBrainz-Gender-Wrangler
[5] https://zenodo.org/record/3964506#.XyE5N0FKg5n

	LFM-1b		LFM-360k	
	male	*female*	*male*	*female*
Users	31.4K	11.5K	94.3K	30.8K
%	*71.67*	*28.33*	*75.40*	*24.60*
Artists	127K	27.3K	50.4K	10.5K
%	*82.30*	*17.70*	*82.83*	*17.17*
Top-head	25.7K	4.8K	10.1K	1.5K
%	*85.21*	*15.79*	*86.99*	*13.01*
Long-tail	100K	22.2K	38.7K	8.5K
%	*81.87*	*18.13*	*81.95*	*18.05*

Table 1: Users' and artists' distributions after the filtering process. "Top-head" artists are the top 20% of artists by play counts, while the remaining 80% are the "long-tail."

LFM-1b				
No.	Male artist	Plays	Female artist	Plays
1	Radiohead	2.6M	Lana Del Rey	1.2M
2	The Beatles	2.5M	Lady Gaga	1.1M
3	Pink Floyd	2.1M	Rihanna	0.8M
4	Daft Punk	2.0M	Björk	0.7M
5	Metallica	1.9M	Madonna	0.6M
LFM-360k				
1	Radiohead	6.2M	Björk	1.3M
2	The Beatles	5.4M	Avril Lavigne	1.1M
3	In Flames	4.9M	Madonna	1.1M
4	Metallica	4.3M	Britney Spears	0.9M
5	Muse	4.2M	Regina Spektor	0.9M

Table 2: Top 5 artists ordered by total play counts in LFM-1b and LFM-360k datasets.

and 2 females), we discard such artists from our final analysis as gender is in this instance, deemed ambiguous. After applying this methodology, we are able to identify 27% of artists with a known-gender. Distributions are observed to be highly imbalanced such that artists of male gender consist of the majority (82%) of artists for which gender can be identified, as shown in Table 1.

In our final analysis, we further filter artists not identified as male or female according to the procedure described above. Artists of gender *other* are discarded as we deem such data to be too sparse to be informative in the analysis of users' listening preferences. We note this group merits further future evaluation, perhaps relying on qualitative methods, and limitations of this binary approach are discussed in Section 7. Table 2 presents the top 5 artists based on the total sum of play counts in the filtered LFM-1b dataset. We observe a trend for male artists' popularity, having approximately twice as much play counts as top-rated female artists/bands. We also observe a trend for the top male artists on the platform to be more commonly composed of bands in comparison to the top-rated female artists.

4 THE LFM-360K DATASET

The LFM-360k dataset [10] consists of approximately 360,000 users listening histories from Last.fm collected during Fall 2008, presenting a snapshot of listening activity for an earlier period in comparison to the LFM-1b dataset. With respect to user gender distributions the proportion of users with a self-declared gender rises to 91% whereas similarly to the LFM-1b dataset, artist gender is not defined. To resolve this, we implement the same pre-processing methodology with the MB database as described for the LFM-1b dataset. After further applying the filtering criteria previously detailed, we are able to identify 31% of artists with a known gender, a proportion notably higher than that of what we were able to identify for the LFM-1b dataset. As presented in Table 1, artist gender distributions in the filtered dataset are once again highly imbalanced towards artists classified as men. For users with identified gender, we again observe a high imbalance towards male users (75%) comparable to rates observed in the LFM-1b dataset. When comparing the two datasets we observe several additional differences and similarities which may impact the propagation of a gender bias in artist recommendations. First, the number of users is significantly larger than that of the LFM-1b, whilst the number of artists is much smaller. Second, sparsity is higher in the LFM-360k dataset in comparison to the LFM-1b. Third, with regard to the top 5 artists of male and female gender in the dataset we observe significantly higher play-counts for artists classified as male in comparison to the LFM-1b dataset, as shown in Table 1. With regard to similarities across the two datasets, we observe that top 5 popular male artists are more commonly bands in comparison to the top 5 female artists. In addition, we observe that the long-tail of both datasets contains significantly higher distribution of female artists, in comparison to the top head reinforcing the conclusion that female artists are significantly more likely to be less popular on the Last.fm platform and hence, more likely to be less recommended as a result of this popularity bias.

5 METHODOLOGY

5.1 Evaluation Metrics

In this section, we formally outline the metrics of preference ratio, bias disparity, as well as accuracy and beyond-accuracy metrics considered during the evaluation.

Preference ratio (PR). Let U be the set of n users, I be the set of m items and S be the nxm input matrix, where $S(u,i) = 1$ if user u has selected item i, and zero otherwise. Given matrix S, the input preference ratio for user group G on item category C is the fraction of liked items by group G in category C, formally defined as the following:

$$PR_S(G,C) = \frac{\sum_{u \in G} \sum_{i \in C} S(u,i)}{\sum_{u \in G} \sum_{i \in I} S(u,i)} \tag{1}$$

Bias disparity (BD). It is defined to be the relative difference between the preference bias for input S and output of a recommendation algorithm R. Formally we define the metric as the following:

$$BD(G,C) = \frac{PR_R(G,C) - PR_S(G,C)}{PR_S(G,C)} \tag{2}$$

In our analysis, we generate a set of r ranked items, R_u which have the highest predicted ratings for a given user u, limiting the value of r to 5.

Accuracy and beyond-accuracy metrics. To evaluate the RS performance, we additionally deploy two accuracy metrics: *Precision*, *nDCG*, and three beyond-accuracy metrics: *coverage*, *spread* and *long-tail percentage*. We refer to the metrics formulation as detailed in the work by Noia et al. [14]. *Precision* (*p@n*) captures the proportion of relevant items in top-N recommendations, such that relevance is a binary function that represents the relevance of item i for a user u. In our work, we consider relevant a recommendation which is greater or equal to the average scaled listening count for a user, after discarding outliers in the data computed using the interquartile range. Although *p@n* is useful for analysing generated item recommendations, it does not capture accuracy aspects relating to the rank of a recommendation. Hence, in our work we also deploy the metric *nDCG*, a rank sensitive metric used to evaluate the accuracy of a RS. With respect to metrics beyond accuracy, we utilise both *spread* and *coverage* to capture a recommender systems ability to recommend a broad range of unique items. Such approaches are important to consider in our work to potentially reason and explain bias propagation across artist genders. The metric *long-tail percentage* is used to capture the proportion of item recommendations which exist in the long tail. In our work, we define the long tail as the 80% of least popular items in the system. We use the metric to capture a filtering algorithms capacity to display the popularity bias.

5.2 Recommendation Algorithms

We test several commonly deployed memory- and model-based CF algorithms, following a similar approach to previous work [28, 29]. Using Surprise [23], a Python library for recommender systems, we formulate our music recommendations as a rating prediction problem where we predict the preference of a target user u for a target artist a. We then evaluate RS recommending the top-5 artists with the highest predicted preferences.

We consider two types of CF algorithms: (1) KNN-based approach: *UserKNNAvg* [27], and (2) factorisation-based approach: *Non-Negative Matrix Factorization* (*NMF*) [30]. Hyperparameters of *UserKNNAvg* and *NMF* are tuned to give the best performance we can achieve with respect to the rank aware metric, *nDCG*. In addition, we consider two *MostPopular* and *UserItemAvg* algorithms which respectively, recommend the most popular and highest rated artists. We consider these algorithms for a baseline comparison.

A variation of the *leave-l-out* evaluation detailed in [9] is performed whereby we translate the approach to evaluate a *top-n* RS. Drawing influence from the methodology of Said et al. [36] we define 3 parameters: (1) n, the size of the recommendation list generated, (2) N, the number of items selected for each user to appear in the test set. N is constrained to be $> n$ to allow for variance in item recommendations across tested algorithms. (3) M, the minimum number of unique artists listened to by a user. M is constrained to be $> N$ to ensure a non-empty test set is able to be formed for each user. We construct three folds, randomly selecting for each user, N items in their listening history to belong to the foldâĂŹs test set and then subsequently removing these listening events from the folds training set. For each of the algorithms tested, we compute all evaluation metrics and preference ratios over each fold and then subsequently report average performance. In our work we set N = 10, M = 20 and n = 5, thereby generating top-5 recommendation lists. We consider a user's test set of size N as the sample space for recommendations to be formed.

5.3 Experimental Design

We set up two experimental designs to evaluate variations in gender bias disparity across recommended artists and user groups for the two datasets. For all experiments detailed, code repositories are made openly available[6]. Experiment 1 is a real-world scenario in which male and female gender distributions are representative of those in both datasets. Experiment 2 is an extreme scenario in which all users have high levels of preference ratio, representing extreme listening preferences towards artists of a specific gender.

Experiment 1. We generate recommendations for a sample of all users for which gender can be identified. In the LFM-1b dataset, we limit the size of this sample to be 30% randomly chosen of all male and female users in the whole dataset (approx 12,000 users), due to computational constraints. The size of the user sample for the LFM-360k dataset was also constrained to be approximately the same size as samples for the LFM-1b dataset. User and artist gender distributions in both samples are representative of overall gender distributions in the entirety of both datasets. We therefore use this experiment to consider the case of gender bias propagation under a real world scenario, assessing the extent to which gender bias disparity may differ across datasets.

Experiment 2. We generate recommendations only for a sample of male and female users which have high preference ratios in the dataset, thereby simulating an extreme scenario under which all users are highly biased towards one artist gender group in their listening preferences. For the LFM-1b dataset, we select the top 30% of both male and female user groups with the highest maximum input preference ratios, maintaining both the proportions of male and female users in the datasets, and the sample size of experiment 1. For the LFM-360k dataset, we sample users from both male and female user groups maintaining the distribution of male and female users in the original dataset. The final user sample has approximately the same sample size as that of the LFM-1b user sample.

Figure 1 represents the distributions of users' input preference ratio towards male and female artist groups. For both datasets considered in this study, it shows that only around 20% of users have a preference ratio towards male artists lower than 0.8. On the contrary, 80% of users have a preference ratio lower than 0.2 towards female artists. Due to the disproportionate amount of users with extreme preferences for male artists across both datasets, a random sampling methodology proposed does little to assess extreme preference towards female artists, resulting in a situation very similar to experiment 1. To resolve this, we further limit our sample space to only users who have extreme preference for female artists, with input preference ratio towards female artists > 0.6. This results in a sample size reduction to 100 users for the LFM-1b dataset, and 400 users for the LFM-360k dataset. Although reduced in size in

[6] https://github.com/dshakes90/Last-fm-Gender-Bias-Analysis

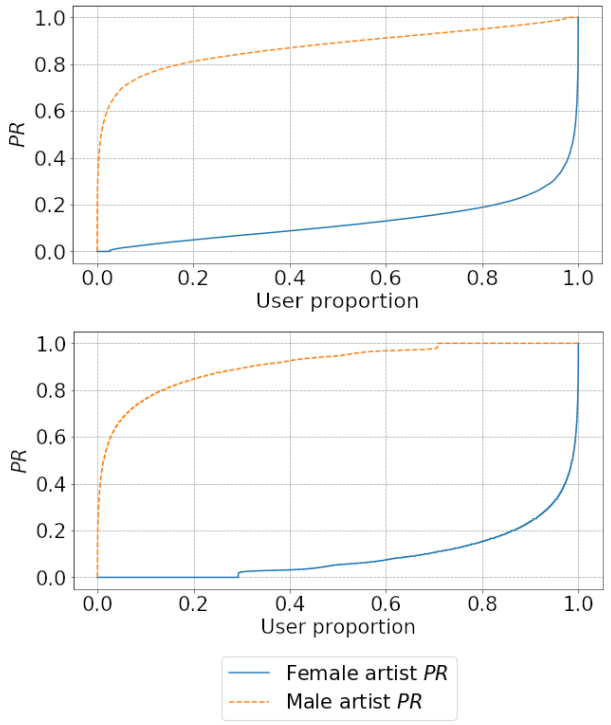

Figure 1: Input Preference Ratio (PR) distributions: LFM-1b (*top*) and LFM-360k (*bottom*).

comparison to experiments 1, we believe such experimental designs to be fundamental to measure the extent to which the treatment of users with extreme preferences differs across artist genders. Experiment 2 represents a situation opposite to the one proposed in experiment 1, thanks to which we can assess if bias propagation is not embedded in the gender *per se*, but is a result of pre-existing bias.

6 RESULTS

6.1 Experiment 1 - Whole population

We report in Figure 2 preference ratio, and in Figure 3 bias disparity results obtained with the LFM-1b dataset. Figure 4 and Figure 5 present preference ratio and bias disparity results respectively for the LFM-360K dataset. The dotted lines in Figure 2 and Figure 4 represent input preference ratios whereas the plot's bars display output preference ratios computed from generated recommendation lists. With regard to pre-existing bias, users in both datasets display high and low input preference ratios for male and female artists respectively, thereby in line with the findings of Millar [33]. In addition, for both artist genders input preference ratios can be seen to be higher by users who share the same gender as the artist. With regard to bias propagation after recommendation, all recommendation models tested result in a positive bias disparity for male artists for which there is minimal variance in treatment across user

	Most Popular	*UserItem Avg*	*UserKNN Avg*	*NMF*
precision	0.010	0.595	0.676	***0.734**
nDCG	0.012	0.663	0.793	***0.880**
coverage	1.7E-04	0.364	***0.558**	0.552
spread	2.322	11.85	***12.84**	12.72
longtail %	0	0.027	0.053	***0.054**

Table 3: Experiment 1 evaluation results on the LFM-1b dataset. Values in bold represent the top value, while marked with * are results where the difference is statistically significant, according to a t-test with $\alpha = 0.05$.

genders. The popularity-based algorithm results in the highest levels of bias disparity for both male and female users, whilst the *NMF* and *UserKNNAvg* algorithms tested result in the lowest absolute levels of bias disparity with marginal difference in bias propagation across the two algorithms. Whatsmore, our findings show male users to be more affected by bias propagation in the LFM-1b dataset whilst for LFM-360K, we observe bias propagation to be greater for female users thereby inline with the findings of Lin et al. [29]. With regard to bias disparity for female artists, negative levels are observed for all algorithms tested. The *MostPopular* algorithm results in the lowest levels of bias disparity due to female artists having significantly lower popularity for both datasets tested, as shown in Table 1. We observe bias propagation to be greater for recommendations generated using the LFM-1b dataset reflected in the lower *long-tail percentage* attained. This suggests that users in the LFM-1b dataset may be more subject to a popularity bias in comparison to LFM-360k which may translate to increased levels of gender bias disparity due to female artists proportionally residing less in the top-head. Together, our findings suggest that differences in bias propagation across the two datasets may be traced to pre-existing bias entering the system in the form of listening events.

6.2 Experiment 2 - Extreme preferences

Considering users with extreme preferences for female artists we observe the inverse scenario of experiment 1, such that bias disparity is positive for female artists and negative towards male artists, as shown in Figure 3 and Figure 5. For both datasets, we comment that one cause of such disparity is a dramatic imbalance in users' listening preference, which then subsequently propagates through to other users' recommendations. Our findings show that such bias propagation is not reserved for male artists on the platform and can, under extreme scenarios emerge in the opposite manner. For both memory- and model-based approaches tested we observe significant differences in bias disparity: *NMF* results in the smallest absolute bias disparity increase thereby reflecting a users' input preference, whereas the neighbour-based *UserKNNAvg* increases absolute bias disparity levels towards whichever user-artist preference is in the majority. The tendency of *NMF* to propagate less bias, positively or negatively speaking, in comparison to the other models is also reflected in the results obtained from the beyond-accuracy metrics evaluation. Indeed, for experiment 2 *NMF* achieves the high levels of coverage, recommending wider subsets of artists, and at the

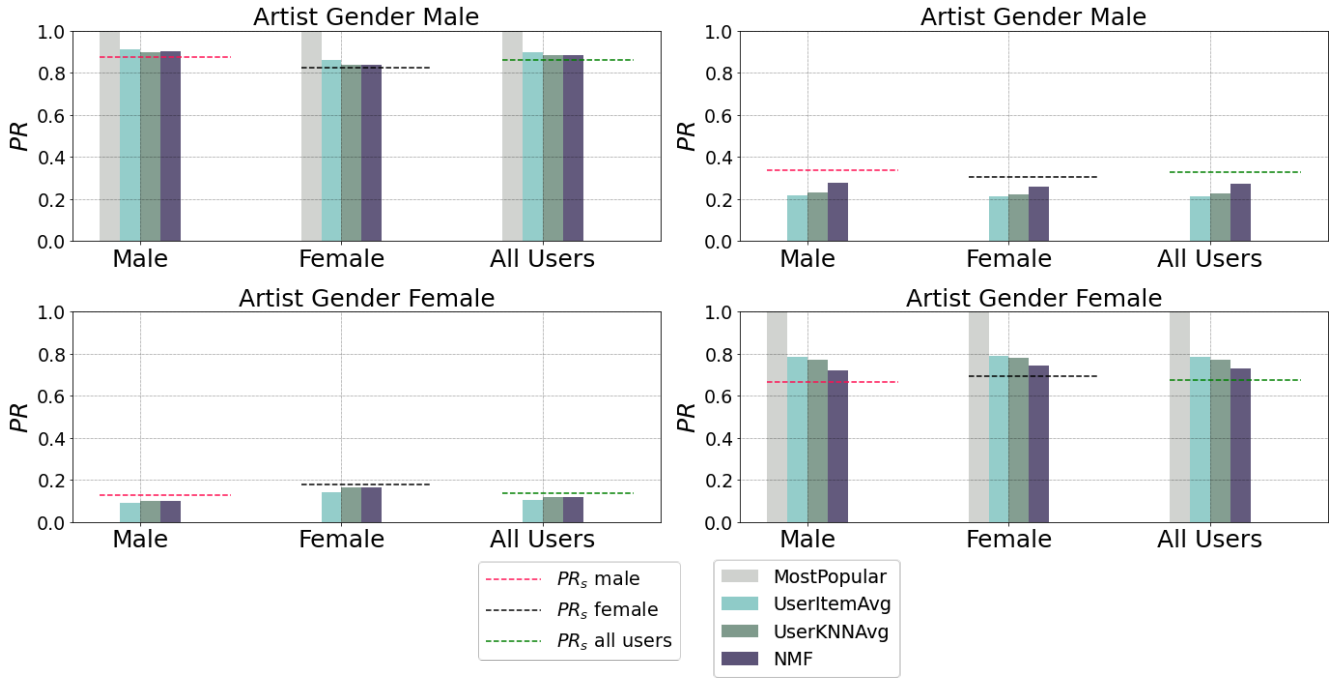

Figure 2: Preference Ratio (PR) results for LFM-1b dataset for experiment 1 (*left column*), and experiment 2 (*right column*).

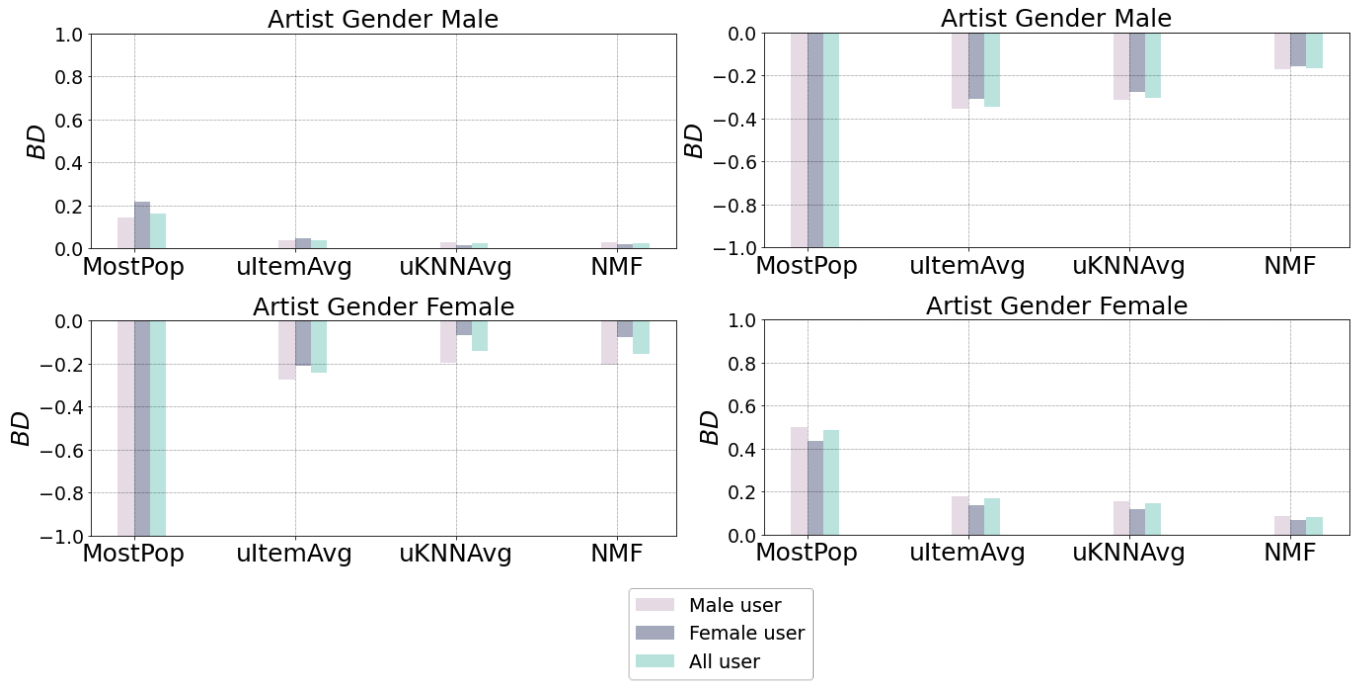

Figure 3: Bias Disparity (BD) results for LFM-1b dataset for experiment 1 (*left column*), and experiment 2 (*right column*).

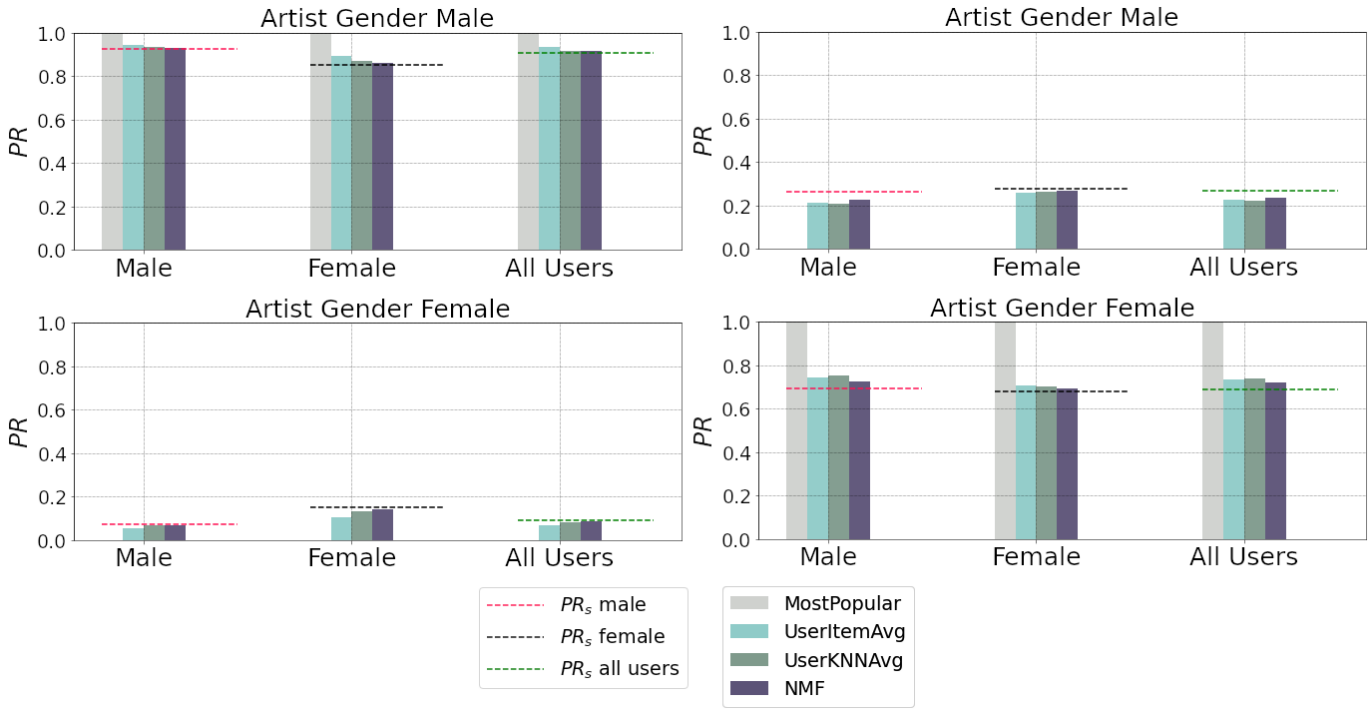

Figure 4: Preference Ratio (PR) results for LFM-360k dataset for experiment 1 (*left column*), and experiment 2 (*right column*).

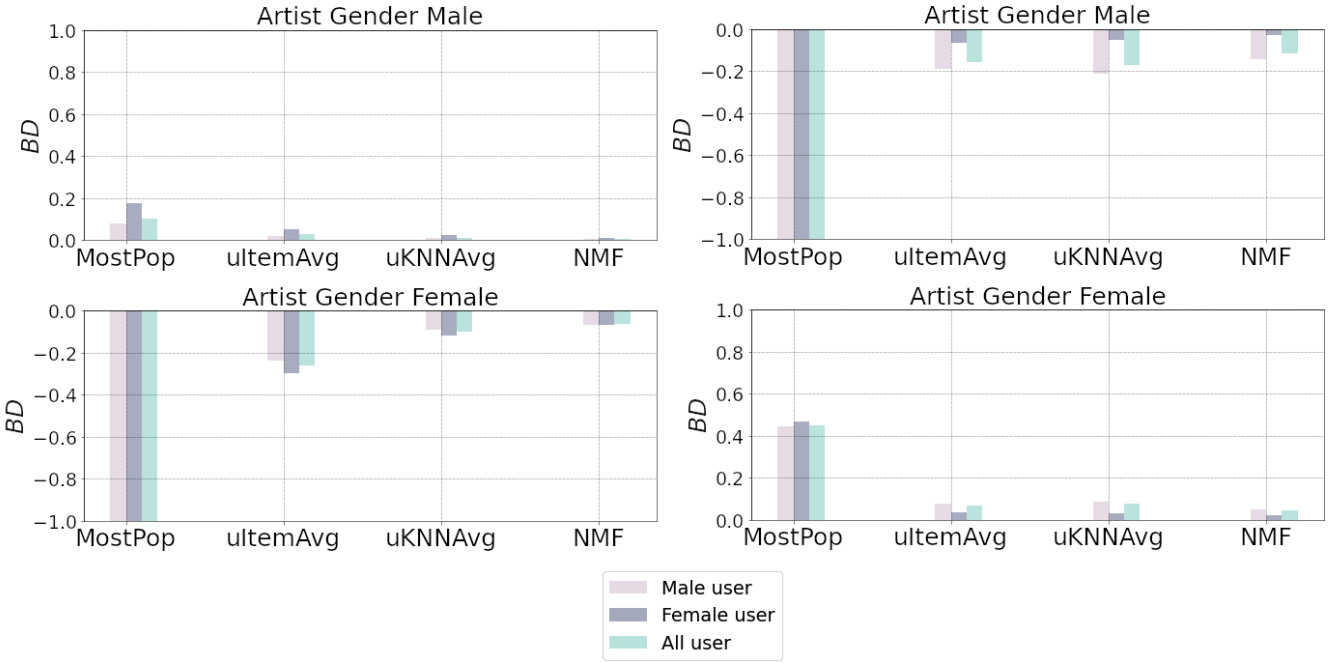

Figure 5: Bias Disparity (BD) results for LFM-360k dataset for experiment 1 (*left column*), and experiment 2 (*right column*).

same time high levels of recommendation spread. Together these results suggest that the model-based algorithm considered in this study is capable of achieving a higher level of diversification in the outcomes in comparison to the memory-based model. Translated to our scenario, it means that *NMF* is the algorithm that focuses less on recommending a specific gender group, avoiding the exacerbation of pre-existing bias in the dataset that other recommendation algorithms exhibit. Again, the effect of bias propagation is seen to be more amplified in the case of the LFM-1b dataset.

7 CONCLUSIONS AND FUTURE WORK

Studies of gender bias in music preferences, conducted in a field such as Music Psychology and Gender Studies, have already evidenced how socio-cultural factors are responsible for disparate treatment of not-male artists. In the field of MIR, relatively little research has analyzed how existing technology can have a role in mitigating or amplifying this bias. In line with the studies on bias disparity in the RS literature, focusing on the musical domain we show how recommendation outcomes can actually impact gender bias in music preferences. Using a binary gender classification, where users and artists are classified as male or female, we have shown how at different levels recommender systems can propagate a pre-existing bias. In addition, simulating an "upside down" world where users have a much higher preference towards female artists, still we find evidence of an exacerbation of that bias. Our results show that gender bias can be propagated by CF-based recommendations, according to the bias present in the data. Hence, RS can have a role in propagating bias, but at least in our exploratory study, we have not found evidence about if they cause the emergence of new forms of biases.

The limitations of our work are several. First, it is important to remark that the binary classification of gender is an oversimplification of gender representation. The state of the art perspective of gender from both natural and social science domains is often non-binary, where male and female are just one of the many genders in which an individual may choose to identify by. Binary definitions of gender have been widely critiqued to be socially constructed through routine gendered performances [8, 12] thereby, considering gender to be only binary in this work is both limiting and to some degree, reinforcing of such binary logic. Second, the evaluation of RS is computed such that the impact of the outcome can be intended in the short- but not in the long-term. Using longitudinal data or simulation frameworks, we believe that a better comprehension of the phenomenon can be achieved, complementing the results we have presented. Lastly, Last.fm users tend to come mostly from Western countries, consequently our results cannot be generalized to represent a global scenario. This issue is well known in the MIR domain [39], and we do believe that to consider a multicultural perspective is undoubtedly a necessary step to give robustness to MIR studies dealing with socio-cultural and socio-technical phenomena.

8 ACKNOWLEDGMENTS

This work is partially supported by the European Commission under the TROMPA project (H2020 770376).

REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *CEUR Workshop Proceedings* 2440 (2019). arXiv:1907.13286

[2] Luis Aguiar, Joel Waldfogel, and Sarah Waldfogel. 2018. *Playlisting Favorites: Is Spotify Gender-Biased?* Technical Report November. https://ec.europa.eu/jrc/sites/jrcsh/files/jrc113503.pdf

[3] Manuel Anglada-Tort, Amanda E Krause, and Adrian C North. 2019. Popular music lyrics and musicians' gender over time: A computational approach. *Psychology of Music* (2019). https://doi.org/10.1177/0305735619871602

[4] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61. https://doi.org/10.1145/3209581

[5] Solon Barocas and Andrew D. Selbst. 2014. Big Data's Disparate Impact. *California Law Review* 671 (2014), 671–732.

[6] Christine Bauer and Markus Schedl. 2019. Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PLOS ONE* i (2019), 1–36.

[7] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15, 3 (2013), 209–227. https://doi.org/10.1007/s10676-013-9321-6

[8] Judith Butler. 2006. *Gender Trouble.* Taylor and Francis.

[9] RocÃŋo CaÃśamares, Pablo Castells, and Alistair Moffat. 2020. Offline evaluation options for recommender systems. *Information Retrieval Journal* 23 (03 2020). https://doi.org/10.1007/s10791-020-09371-3

[10] Òscar Celma. 2010. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space.* Springer-Verlag Berlin Heidelberg.

[11] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and addressing algorithmic bias in practice. *Interactions* 25, 6 (2018), 58–63. https://doi.org/10.1145/3278156

[12] Simone de Beauvoir. 1949. *The Second Sex.* Vintage Classics.

[13] Sarah Dean, Sarah Rich, and Benjamin Recht. 2020. Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information. In *Proceedings of the 3rd ACM Conference on Fairness, Accountability and Transparency (ACM FAccT 2020).* Barcelona, Spain, 436–445. https://doi.org/10.1145/3351095.3372866

[14] Tommaso Di Noia, Jessica Rosati, Paolo Tomeo, and Eugenio Di Sciascio. 2017. Adaptive multi-attribute diversity for recommender systems. *Information Sciences* 382-383 (2017), 234–253. https://doi.org/10.1016/j.ins.2016.11.015

[15] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. 2019. FaiRecSys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics* 9, 2 (2019), 197–213. https://doi.org/10.1007/s41060-019-00181-5

[16] Michael D. Ekstrand, Mucun Tian, Jennifer D. Ekstrand, Oghenemaro Anuyah, David Mcneill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In? Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st ACM Conference on Fairness, Accountability and Transparency (ACM FAccT 2018)*, Vol. 81. 172–186. https://doi.org/10.18122/B2GM6F

[17] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring Author Gender in Book Rating and Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18).* 242–250. http://dl.acm.org/citation.cfm?doid=3240323.3240373

[18] Andres Ferraro, Dmitry Bogdanov, Xavier Serra, and Jason Yoon. 2019. Artist and style exposure bias in collaborative filtering based music recommendations. In *1st Workshop on Designing Human-Centric MIR Systems (wsHCMIR19), co-located at 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019).* arXiv:1911.04827 http://arxiv.org/abs/1911.04827

[19] Emilia Gomez, Andre Holzapfel, Marius Miron, and Bob L. Sturm. 2019. Fairness, Accountability and Transparency in Music Information Research (FAT-MIR). https://doi.org/10.5281/zenodo.3546227

[20] Asela Gunawardana and Guy Shani. 2015. *Evaluating Recommender Systems.* Springer US, Boston, MA, 265–308. https://doi.org/10.1007/978-1-4899-7637-6_8

[21] F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (2015), 1–19. https://doi.org/10.1145/2827872

[22] Andre Holzapfel, Bob L. Sturm, and Mark Coeckelbergh. 2018. Ethical Dimensions of Music Information Retrieval Technology. *Transactions of the International Society for Music Information Retrieval* 1 (2018), 44–55.

[23] Nicolas Hug. 2017. Surprise, a Python library for recommender systems. http://surpriselib.com.

[24] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (2015), 427–491. https://doi.org/10.1007/s11257-015-9165-3

[25] Dietmar Jannach, Oren Sar Shalom, and Joseph A Konstan. 2019. Towards More Impactful Recommender Systems Research. In *Proceedings of the ImpactRS Workshop, 13th ACM Conference on Recommender Systems (RecSys 2019).* 15–17.

[26] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. 2010. Comparison of implicit and explicit feedback from an online music recommendation service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, HetRec 2010, Held at the 4th ACM Conference on Recommender Systems (RecSys 2010)*. 47–51. https://doi.org/10.1145/1869446.1869453

[27] Yehuda Koren. 2010. Factor in the Neighbors: Scalable and Accurate Collaborative Filtering. *ACM Trans. Knowl. Discov. Data* 4, 1, Article 1 (Jan. 2010), 24 pages. https://doi.org/10.1145/1644873.1644874

[28] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Advances in Information Retrieval*, Joemon M Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 35–42.

[29] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. 2019. Crank up the volume: Preference bias amplification in collaborative recommendation. In *CEUR Workshop Proceedings*, Vol. 2440. arXiv:1909.06362

[30] Xin Luo, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. 2014. An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems. *IEEE Transactions on Industrial Informatics* 10, 2 (2014), 1273–1284.

[31] Masoud Mansoury, Bamshad Mobasher, Robin Burke, and Mykola Pechenizkiy. 2019. Bias disparity in collaborative recommendation: Algorithmic evaluation and comparison. In *CEUR Workshop Proceedings*, Vol. 2440. arXiv:1908.00831

[32] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. Association for Computing Machinery, New York, NY, USA, 1097–1101. https://doi.org/10.1145/1125451.1125659

[33] Brett Millar. 2008. Selective hearing: Gender bias in the music preferences of young adults. *Psychology of Music* 36, 4 (2008), 429–445. https://doi.org/10.1177/0305735607086043

[34] Yoon Joo Park and Alexander Tuzhilin. 2008. The Long Tail of Recommender Systems and How to Leverage It. *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)* (2008), 11–18. https://doi.org/10.1145/1454008.1454012

[35] Caroline Criado Perez. 2019. *Invisible Women: Exposing data bias in a world designed for men.* Random House.

[36] Alan Said, Alejandro Bellogín Kouki, and A. P. deVries. 2013. A Top-N Recommender System Evaluation Protocol Inspired by Deployed Systems.

[37] Justin Salamon. 2019. What's Broken in Music Informatics Research? Three Uncomfortable Statements. In *Proceedings of the 36th International Conference on Machine Learning*. 2012–2014.

[38] Markus Schedl. 2016. The LFM-1b Dataset for Music Retrieval and Recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (New York, New York, USA) *(ICMR âĂŹ16)*. Association for Computing Machinery, New York, NY, USA, 103âĂŞ110. https://doi.org/10.1145/2911996.2912004

[39] Xavier Serra, Michela Magas, Emmanouil Benetos, Magdalena Chudy, Simon Dixon, Arthur Flexer, Emilia Gómez, Fabien Gouyon, Perfecto Herrera, Sergi Jorda, Oscar Paytuvi, Geoffroy Peeters, Jan Schlüter, Hugues Vinet, and Gerhard Widmer. 2013. *Roadmap for Music Information ReSearch.*

[40] Aaron Swartz. 2002. MusicBrainz: A Semantic Web Service. *IEEE Intelligent Systems* 17, 1 (Jan. 2002), 76âĂŞ77. https://doi.org/10.1109/5254.988466

[41] Virginia Tsintzou, Evaggelia Pitoura, and Panayiotis Tsaparas. 2018. Bias Disparity in Recommendation Systems. *CoRR* abs/1811.01461 (2018). arXiv:1811.01461 http://arxiv.org/abs/1811.01461

[42] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. *Discriminating Systems: Gender, Race and Power in AI.* AI Now Institute. https://ainowinstitute.org/discriminatingsystems.html

[43] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (2017), 2979–2989. https://doi.org/10.18653/v1/d17-1323

[44] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-Aware Tensor-Based Recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM âĂŹ18)*. Association for Computing Machinery, New York, NY, USA, 1153âĂŞ1162. https://doi.org/10.1145/3269206.3271795