Information Processing and Management 58 (2021) 102666

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Investigating gender fairness of recommendation algorithms in the music domain

Alessandro B. Melchiorre [a,b], Navid Rekabsaz [a,b], Emilia Parada-Cabaleiro [a,b], Stefan Brandl [a], Oleg Lesota [a,b], Markus Schedl [a,b,*]

[a] *Johannes Kepler University Linz, Institute of Computational Perception, Multimedia Mining and Search Group, Altenberger Straße 69, 4040 Linz, Austria*

[b] *Linz Institute of Technology, AI Lab, Human-centered AI Group , Altenberger Straße 69, 4040 Linz, Austria*

ARTICLE INFO

Keywords:
Recommender systems
Music
Bias
Neural networks
Demographics

ABSTRACT

Although recommender systems (RSs) play a crucial role in our society, previous studies have revealed that the performance of RSs may considerably differ between groups of individuals with different characteristics or from different demographics. In this case, a RS is considered to be *unfair* when it does not perform equally well for different groups of users. Considering the importance of RSs in the distribution and consumption of musical content worldwide, a careful evaluation of fairness in the context of music RSs is crucial. To this end, we first introduce *LFM-2b*, a novel large-scale real-world dataset of music listening records, comprising a subset to investigate bias of RSs regarding users' demographics. We then define a notion of fairness based on the performance gap of a RS between the users with different demographics, and evaluate a variety of collaborative filtering algorithms in terms of accuracy and beyond-accuracy metrics to explore the fairness in the RS results toward a specific gender group. We observe the existence of significant discrepancies (unfairness) between the performance of algorithms across male and female user groups. Based on these discrepancies, we explore to what extent recommender algorithms lead to intensifying the underlying population bias in the final results. We also study the effect of a resampling strategy, commonly used as debiasing method , which yields slight improvements in the fairness measures of various algorithms while maintaining their accuracy and beyond-accuracy performance.

1. Introduction

Recommender systems (RSs) play an indisputable role in our lives by taking part in a variety of day-to-day decisions, influencing which content we are exposed to on digital platforms. Such decisions include selecting a book to buy, a movie to watch, a hotel to book, or a song to listen to. While RSs, therefore, offer great support in accessing otherwise barely manageable amounts of data, several studies revealed that their performances may differ between groups of users depending on their characteristics (e.g., gender, race, ethnicity, age, country of origin, or personality Datta, Tschantz, & Datta, 2015; Lambrecht & Tucker, 2019; Melchiorre, Zangerle, & Schedl, 2020; Schedl, Hauger, Farrahi and Tkalcic, 2015). Some of these performance disparities can disadvantage certain user groups in accessing opportunities, and therefore disregard the principle of fairness, namely the "absence of any prejudice

* Correspondence to: Altenberger Strasse 69 / S3 407, 4040 Linz, Austria.
E-mail addresses: alessandro.melchiorre@jku.at (A.B. Melchiorre), navid.rekabsaz@jku.at (N. Rekabsaz), emiliar.parada-cabaleio@jku.at (E. Parada-Cabaleiro), stefan.brandl@jku.at (S. Brandl), oleg.lesota@jku.at (O. Lesota), markus.schedl@jku.at (M. Schedl).

https://doi.org/10.1016/j.ipm.2021.102666
Received 2 December 2020; Received in revised form 28 April 2021; Accepted 16 June 2021
Available online 8 July 2021

or favoritism toward an individual or a group based on their intrinsic or acquired traits" (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019). In this article, we set out to study *group fairness* from the point of view of the *users* of RSs.

This form of unfairness in RSs can be traced back to biases resulting from the (interplay between) data, model/algorithm, and users (Chen, Dong, Wang, Feng, Wang, & He, 2020). Here, we focus on biases resulting from an imbalanced number of data points regarding various demographics. These imbalances can be classified as *population bias* (Olteanu, Castillo, Diaz, & Kiciman, 2019), which is part of a more general *data bias* (Baeza-Yates, 2018). In the presence of population bias, a machine learning model (such as those created by a RS algorithm) captures the interaction patterns of the majority group more prominently, which can lead to better model performance for the majority group in comparison to the one for the minority group (an unfair system) (Hardt, Price, Price, & Srebro, 2016). We refer to the cause of this unfairness as *model/algorithmic bias*. Clearly, data is a major reason for model bias, but different models still can lead to different degrees of unfairness or even intensify data bias. As shown in previous studies in the contexts of text classification (De-Arteaga et al., 2019) and passage retrieval (Rekabsaz & Schedl, 2020), a machine learning model may even compound the discrepancies in the collection, such that the distribution of the model's results is even more unfair in comparison with the existing discrepancies in the underlying data. This phenomena is referred to as *compounding imbalances* (De-Arteaga et al., 2019; Hellman, 2018). In this case, the system is not only unfair toward minority groups but even intensifies the existing imbalances.

1.1. Research questions

In the work at hand, we comprehensively study both population and model bias in the context of unfairness with regard to *gender* in one important domain of RS research, *music* recommendation (Schedl, Knees, McFee, Bogdanov and Kaminskas, 2015). Concretely, we explore the following research questions: **RQ1:** Do recommender algorithms of various categories yield different performance scores (in terms of accuracy and beyond-accuracy metrics) for different user groups with respect to gender? If so, how can these differences be characterized? **RQ2:** What is the effect of a resampling strategy, commonly used as debiasing method, on the performance and fairness of algorithms? **RQ3:** Do RS algorithms compound population bias? If so, how can this be characterized?

These questions can only be approached using a music recommendation dataset that contains gender information about users. Existing (publicly available) datasets of this kind are exclusively composed of data from the music streaming platform *Last.fm*, and include *Last.fm 1K* (Celma, 2010), *LFM-1b* (Schedl, 2016), and *Music Listening Histories Dataset (MLHD)* (Vigliensoni & Fujinaga, 2017). Most of them are either small (Celma, 2010) or do not contain up-to-date listening information (Schedl, 2016; Vigliensoni & Fujinaga, 2017). Therefore, we introduce – as an additional contribution – the *LFM-2b* dataset, a novel up-to-date large-scale real-world collection of music listening records, gathered from *Last.fm*, which considerably extends *LFM-1b*. Unlike existing datasets, *LFM-2b* provides more than two billion listening records for more than 120,000 users who listened to more than 50 million unique tracks in total. Another remarkable difference to *LFM-1b* is the large temporal coverage of listening records (2005–2020), which allows to track users' listening behavior over considerable periods of time. *LFM-2b* contains a dedicated subset, *LFM-2b-DemoBias*, which we especially developed to study and evaluate fairness and biases in music RS in terms of users' gender, age, and country of origin. Because of these characteristics as well as a higher average number of listening records per user in comparison to *LFM-1b* (Table 1, rows *LFM-1b*, *LFM-2b*, and *LFM-2b/1b*$_{diff}$), *LFM-2b-DemoBias* is well-suited to approach the research questions under investigation. All data is publicly available.

In approaching **RQ1**, we study gender-related unfairness on a variety of RS algorithms. Using *LFM-2b-DemoBias*, we train the algorithms and compute evaluation metrics separately on the subsets of male and female users in the test set, using accuracy-based (recall and NDCG) as well as beyond-accuracy metrics (diversity and coverage). We define unfairness of a RS algorithm regarding an evaluation metric as the mean absolute differences of average evaluation scores across pairs of sensitive attributes (genders in our experiments).[1] This notion of fairness closely corresponds to the *equality of opportunity* metric of Hardt et al. (2016).

To address **RQ2**, we repeat the above-mentioned experiments by considering a debiasing method where – following Geyik, Ambler, and Kenthapadi (2019) – data points of the minority group (female) in training data are resampled up to the number of data points of the majority group (male).

To answer **RQ3**, we extend the concept of compounding imbalances, introduced by De-Arteaga et al. (2019) on the true-positive metric, to any arbitrary metric and refer to it as *compounding factor*. The compounding factor regarding a model and an evaluation metric is defined as the divergence of the distribution of the metric's results over the users' groups, from the population distribution in the dataset. A higher compounding factor indicates that the model intensifies the existing bias in data, toward the majority group.

[1] In the experiments, gender is considered as a binary construct due to practical constraints. In particular, in *LFM-2b* the users are either categorized to one of these two genders, or their gender is not specified. We are however fully aware that a gender binary model is not representative of all individuals; yet, working with in-the-wild data (as in our case) entails an unavoidable caveat, derived from the still predominant belief that human beings can be sorted into two discrete categories (Hyde, Bigler, Joel, Tate, & van Anders, 2019). All introduced metrics, however, are defined for generic non-binary settings, and can be applied to gender or any other sensitive attribute.

1.2. Major contributions

The main contributions of the work at hand can be summarized as follows:

- We introduce a large-scale real-world dataset of music listening records that can be used to study fairness and bias of music RSs.
- We point out, through the use of specific fairness measures, which are the recommender algorithms that more robustly handle the gender bias existing in real-world data.
- We identify to what extent the recommender algorithms compound population bias, and how a data debiasing method mitigates this issue.

1.3. Structure of the article

In the remainder of this article, we present related literature and datasets (Section 2), describe the *LFM-2b* and *LFM-2b-DemoBias* datasets we created and released together with this article (Section 3). We then introduce and formulate our notion of fairness and compounding factor (Section 4). Subsequently, we outline the experiment setup by describing the data processing, presenting the considered algorithms, and discussing the experiment procedures (Section 5). Finally, the outcomes of our investigation are discussed (Section 6), and the conclusions and limitations of the presented work, as well as the future research directions in the study of demographic bias of music RSs, are presented (Section 7).

2. Related work

Related literature can be categorized into RSs algorithms (Section 2.1), studies of bias, fairness, and debiasing in the context of RSs (Section 2.2), and existing datasets for music recommendation (Section 2.3).

2.1. Recommender system algorithms

RSs are systems that match users to items[2] in order to maximize some metric (or a combination of metrics). In the case of music recommendation (Schedl, Knees et al., 2015), the users are typically people using some recommendation service and the recommended items are most commonly tracks/recordings or artists. The goal of such RSs is then to provide satisfying track or artist recommendations to the users. The underlying data leveraged by the RS algorithm usually include a user–item interaction matrix, where a cell's value indicates a user's rating or playcount value (i.e., how often the user listened to the track), sometimes complemented by side information on the user (e.g., age or gender) or the item (e.g., audio features, editorial and user-generated metadata such as genre, artist biographies, or tags contributed by the platform users, etc.), as well as contextual information (e.g., timestamp or location).

The most widely adopted strategies in RSs are *collaborative filtering* (CF) (Aggarwal, 2016b; Koren & Bell, 2015) and *content-based filtering* (CBF) (Deldjoo, Schedl, Cremonesi, & Pasi, 2020; Lops, de Gemmis, & Semeraro, 2011), as well as combinations thereof (*hybrid recommender systems*) (Aggarwal, 2016a; Beliakov, Calvo, & James, 2011), sometimes incorporating contextual information into a *context-aware recommender system* (CARS) (Adomavicius, Mobasher, Ricci, & Tuzhilin, 2011).

Pure CF approaches only exploit historical data of user–item interactions to predict the level of preference of unknown items for an active user, and recommend to them the items with the highest prediction scores (Aggarwal, 2016b; Koren & Bell, 2015). CF systems can further be categorized into memory-based and model-based. The former computes similarities between users (or items) directly on the high-dimensional vectors of the user–item interaction matrix to identify nearest neighbors (e.g., users whose music taste is most similar to that of the active user); the latter learns a model from the user–item interactions, which is used to effect predictions or recommendations. Often a latent factor model is used, which can be regarded as a low-dimensional embedding of the user–item interaction space. This can be achieved, e.g., by matrix factorization. Similarities between users and items are then computed in this latent factor space rather than in the original space.

CBF systems, in contrast, leverage content descriptors of the items, the active user already interacted with to identify items similar to those preferred by the user (Deldjoo et al., 2020; Lops et al., 2011). In the music domain, such content descriptors may include tempo, rhythm patterns, or genre information. In addition to content information, CBF systems therefore only require the active user's interaction data (but not those of other users) to make recommendations.

CARS leverage contextual information, such as time, location, or activity of the users, to build a recommendation model that considers the current situation of the active user when effecting recommendations (Adomavicius et al., 2011). CARS most commonly extend CF or CBF approaches.

Hybrid RSs combine at least two individual approaches, using a fusion or aggregation method (Burke, 2002; Çano & Morisio, 2017). Most recent deep learning-based approaches for music recommendation belong to this category and leverage co-listening information (CF) as well as content features (CBF), e.g. Huang et al. (2020), Oramas, Nieto, Sordo, and Serra (2017) and van den Oord, Dieleman, and Schrauwen (2013).

[2] We denote the recommended objects as "items" irrespective of the fact that some RSs recommend users to users, e.g., on online dating platforms.

In the study at hand, we focus our investigation on CF algorithms for several reasons. First, they are more widely adopted than CBF. Second, they typically yield better performance than CBF systems, and are therefore used in (or as part of) almost all state-of-the-art systems. Third, investigating hybrid systems instead would render it hard to disentangle unfairness aspects originating from the collaborative information from those originating from content information. Fourth, the output of CF systems is particularly sensitive to data biases, as shown for instance in Melchiorre et al. (2020) for differences in interaction data resulting from different personality traits of users, in Lambrecht and Tucker (2019) for gender-specific differences in job advertisements, in Bauer and Schedl (2019) for differences in terms of users' inclination to listen to mainstream music, and in Abdollahpouri, Mansoury, Burke, and Mobasher (2019) and Kowald, Schedl, and Lex (2020) for differences resulting from varying item popularity in the music and movie domains.

2.2. Investigating and alleviating bias in recommender systems

Fairness and bias in RSs have been the focus of numerous recent studies (Lin, Sonboli, Mobasher, & Burke, 2019a; Mansoury, Abdollahpouri, Pechenizkiy, Mobasher, & Burke, 2020a; Mansoury, Mobasher, Burke, & Pechenizkiy, 2019). The concept of fairness requires systems not to discriminate against either a group (Pedreshi, Ruggieri, & Turini, 2008) or individuals (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012) in terms of recommendation quality.

Assessing bias and fairness. Various studies approach fairness in RSs from the points of view of different involved stakeholders, namely RS users (Ekstrand et al., 2018), items/item providers (Beutel et al., 2019; Biega, Gummadi, & Weikum, 2018; Borges & Stefanidis, 2019), or all, known as multi-sided fairness (Burke, 2017; Patro, Biswas, Ganguly, Gummadi, & Chakraborty, 2020). The work at hand focuses on studying fairness and bias from the perspective of user groups, defined by common gender.

Fairness has been investigated in various domains. For instance, in *job* recommendation, studies have found that highly paid jobs are more often recommended to men than to women, both on Facebook by Lambrecht and Tucker (2019) and on Google by Datta et al. (2015). Lambrecht and Tucker identify as the reason the cost-minimizing strategy of advertising algorithms. More precisely, platform owners charge for showing a job advertisement to users; and the cost is different for different demographic target groups. Since young women belong to a particularly expensive group, well-paid ads that are meant to be gender-neutral are, in fact, more frequently presented to male users by algorithms adopting a cost-minimizing strategy, because the latter group is "less expensive". In the domain of *books*, Ekstrand, Tian, Imran, Mehrpouyan and Kluver (2018) investigate the disparity of book authors' gender distribution in the user profiles and in the recommendation lists. They find that particularly CF algorithms often create recommendation results that are biased toward male authors. In the movie domain, Lin, Sonboli, Mobasher, and Burke (2019b) study how different recommender system algorithms amplify or dampen preferences for specific item categories (e. g., Action versus Romance) for male and female users. They show, for instance, that neighborhood-based models intensify the preferences, for all users, toward the preferred item category of the dominant group (males), while some other algorithms, such as SVD++ and BiasedMF, dampen these preferences. Similarly, Mansoury, Abdollahpouri, Pechenizkiy, Mobasher, and Burke (2020b) assess the amplification of popularity bias in recommender systems due to the feedback loop, i. e., recommending popular items makes the popular items even more popular, showing that the bias amplification is stronger for the minority group (i. e., females). To which extent different recommender system algorithms reflect the user group preferences for item categories in input has also been investigated by Mansoury et al. (2019), who follow a similar approach to Lin et al. (2019b). In the *music* domain, which is also the target domain of our study, bias in recommender systems (Ekstrand, Tian, Azpiazu et al., 2018; Melchiorre et al., 2020; Schedl, Hauger et al., 2015; Shakespeare, Porcaro, Gómez, & Castillo, 2020) and gender representation in music streaming and broadcasting services (Epps-Darling, Bouyer, & Cramer, 2020; Watson, 2020), have recently been investigated. In particular, Schedl, Hauger et al. (2015) show that precision and recall obtained by simple CF and CARS algorithms substantially diverge for users of different gender, age, and country. Melchiorre et al. (2020) show that state-of-the-art CF algorithms yield different performance scores (recall and NDCG) for different user groups with respect to their personality traits, in particular for users with high versus low openness and neuroticism. Ekstrand, Tian, Azpiazu et al. (2018) reveal performance disparities (with respect to the NDCG metric) of simple CF algorithms in the music and movie domains, resulting in unfairness with regards to age and gender. They also find that biases do not necessarily correlate with user group size.

Unfairness and bias caused by algorithms are studied in various related tasks to recommendation, indeed, fairness-aware recommendation algorithms have been presented in the literature (Steck, 2018; Yao & Huang, 2017). Rekabsaz and Schedl (2020) show that ranking models based on neural networks increase the gender bias toward male in retrieval results in comparison with the classical exact-matching models. In the direction of analyzing algorithmic bias, De-Arteaga et al. (2019) discuss compounding imbalances, a concept related to compounding injustices (Hellman, 2018) in political philosophy. De-Arteaga et al. show that if a classifier performs with a lower sensitivity, i.e. true-positive rate (TPR), on the minority group in comparison with the majority group, the imbalance between the groups in the final TPRs becomes larger than the initial imbalance in the underlying dataset. In this case, the model (classifier) intensifies the existing imbalances in the dataset.

Besides fairness, recent work sheds light on various types of biases involved in RSs. For instance, recent research reveals a popularity bias in current recommendation algorithms. In particular, it was shown that users are recommended items that do not match their preference toward a certain popularity level (niche songs/artists are undervalued) (Abdollahpouri et al., 2019; Kowald et al., 2020).

Debiasing and improving fairness. State-of-the-art *debiasing* methods which are applicable to RSs are commonly categorized into four approaches (Chen et al., 2020): (1) rebalancing, (2) regularization, (3) counterfactual intervention, and (4) adversarial training. In the first category, the data or recommendation results are rebalanced in order to satisfy a certain fairness measure (e.g., demographic parity). In such methods, debiasing is approached as a pre- or post-processing step. Common pre-processing approaches are: re-labeling training data to achieve an equal number of relevant labels across the groups (Pedreshi et al., 2008), or resampling data to have an equal number of training data (Geyik et al., 2019). Post-processing methods typically aim to change the output list of RSs in a way that the results over each recommendation or the expectation of results satisfy targeted fairness measures (Biega et al., 2018; Zehlike et al., 2017).

In the second category, debiasing is done by steering the optimization process of the recommendation model during training through including a regularization term for fairness. Zemel, Wu, Swersky, Pitassi, and Dwork (2013) propose a general framework which seeks to learn representations that contain sufficient information for the task in hand but are invariant regarding sensitive attributes. Kamishima et al. adopted this framework to the context of RS and later generalized it to implicit feedback-based recommender systems (Kamishima & Akaho, 2017; Kamishima, Akaho, Asoh, & Sakuma, 2012). Regularization-based approaches are also studied in the context of biases in RSs, for instance by Abdollahpouri, Burke, and Mobasher (2017) to address popularity bias.

Regarding the third category, Kusner, Loftus, Russell, and Silva (2017) introduce counterfactual fairness to RSs. In this method, fairness criteria are satisfied when the evaluation of an individual in the counterfactual world – where the individual's sensitive attribute is changed by intervention – and in the real world are identical.

Finally, the category of debiasing through adversarial learning approaches the topic by creating fair representations through a min–max game, which are agnostic to sensitive attributes (Bose & Hamilton, 2019). In this direction, recently, Beigi et al. (2020) propose an adversarial training method which seeks to protect users' sensitive attributes from an attacker who has access to users' item list and recommendations.

In the study at hand, we investigate a method from the first category of approaches, namely rebalancing. In particular, we investigate the use of a rebalancing technique on the training data of the RS algorithm to achieve statistical parity. For this purpose, we resample the data points of the users of the minority group (female) in training data to achieve an equal number of users with the ones of the majority group (male).

2.3. Datasets for music recommendation experiments

Investigating (music) recommendation algorithms in such a way that insights gained can generalize to real-world applications requires access to suitable datasets containing data obtained in-the-wild. Although many corpora have been publicly released in the last decade for the study of music RSs, the majority of these – unlike the proposed *LFM-2b* dataset – do not include users' demographic information. This omission of users' demographics is particularly the case for corpora containing data from *Spotify* (Brost, Mehrotra, & Jehan, 2019; Pichl, Zangerle, & Specht, 2015; Zamani, Schedl, Lamere, & Chen, 2019), *Yahoo!* (Dror, Koenigstein, Koren, & Weimer, 2011), *Echo Nest* (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011), or *Art of the Mix* (McFee & Lanckriet, 2012).

However, to investigate bias in music RSs in general, and the so-called population bias (Olteanu et al., 2019) in particular, users' demographic information becomes essential. The population bias, an aspect of which we study here, is a type of bias contained in the data itself that arises from the distortion of a given population with respect to a target population. This is typical, for instance, of social media, since some platforms are more frequently used by a specific group, e. g., females on Pinterest, while others by another group, e. g., males on Twitter.[3] To investigate how state-of-the-art recommender algorithms might increase or mitigate a bias already present in the data, a dataset containing users' demographic information becomes indispensable. From the publicly available datasets already presented in the literature, those containing such information are: (1) datasets collected from the music platform *Last.fm*, i. e., *Last.fm 360K* and *Last.fm 1K* (Celma, 2010), *LFM-1b* (Schedl, 2016, 2019), and *Music Listening Histories Dataset (MLHD)* (Vigliensoni & Fujinaga, 2017), which include users' gender, country, and age gathered at the time of their registration; (2) datasets created from data shared on other social media sites, such as Twitter, containing music-related hashtags mapped onto musical metadata through open music encyclopediae such as MusicBrainz,[4] i. e., *MusicMicro* (Schedl, 2013), *Million Musical Tweets Dataset (MMTD)* (Hauger, Schedl, Košir, & Tkalčič, 2013), and *#nowplaying-RS* (Poddar, Zangerle, & Yang, 2018). Nevertheless, none of the datasets containing demographic information has been developed by having in mind the evaluation of population bias beforehand, which impairs a clear understanding of their real potential for bias-related research: considering that users lacking on demographics would be discarded for such a study, the actual value of these datasets for the assessment of bias in music RSs is unknown. Furthermore, although the use of online music platforms and social media – namely the main sources for retrieving users' demographic information in this context – has particularly increased in the last years,[5] up-to-date datasets of this nature have not been recently presented. Therefore, we introduce the *LFM-2b* dataset, an up-to-date large-scale corpus containing listening histories from *Last.fm* users collected over the last 15 years (from 2005 until 2020). In addition to this, the *LFM-2b-DemoBias*, i. e., a subset of the former demographics specially tailored to assess population bias in music RSs, is also presented.

[3] https://sproutsocial.com/insights/new-social-media-demographics/.

[4] https://musicbrainz.org/.

[5] https://ourworldindata.org/rise-of-social-media/.

Table 1
Descriptive statistics for the *LFM-1b*, the *LFM-2b*, the $LFM\text{-}2b\backslash 1b_{diff}$, the *LFM-2b-DemoBias*, and the demographic groups Gender: F(emale), M(ale); Country: USA (US), Russia (RU), Germany (DE), UK, Poland (PL), and Brazil (BR), Other (those with LEs than 2000 users); and Age: users under 30 years old and users at least 30 years old (<30 and ≥ 30, respectively). Number of Users, Tracks, Artists, and Listening Events (LEs) are given across groups (All) and for each class. Mean and standard deviation (indicated after ±) of the number of Tracks, Artists, and LEs per User, are also indicated.

Dataset			Users	Tracks	Artists	LEs	Tracks/User	Artists/User	LEs/User
LFM-1b			120,322	31,413,999	3,116,790	1,088,161,692	2659 ± 3600	541 ± 680	9,044 ± 16,244
LFM-2b			120,322	50,813,373	5,217,014	2,014,164,872	4316 ± 7465	881 ± 1566	16,740 ± 33,722
LFM-2b $1b_{diff}$			54,202	31,169,391	3,330,322	926,003,180	4476 ± 8000	975 ± 1843	17,084 ± 34,678
LFM-2b-DemoBias	All		60,972	44,917,375	4,571,252	1,678,139,732	6692 ± 9136	1256 ± 1852	27,523 ± 41,856
	Gender	All	55,771	42,241,320	4,278,259	1,551,296,848	6725 ± 9088	1258 ± 1849	27,815 ± 41,713
		F	15,802	12,888,611	1,422,563	326,594,434	4622 ± 6053	971 ± 1323	20,668 ± 30,104
		M	39,969	37,873,456	3,789,381	1,224,702,414	7556 ± 9915	1372 ± 2008	30,641 ± 45,182
	Country	All	55,190	43,217,992	4,278,259	1,572,924,251	6905 ± 9245	1290 ± 1808	28,500 ± 42,596
		US	10,255	12,306,070	1,234,159	249,308,203	6391 ± 9009	1196 ± 1686	24,311 ± 41,219
		RU	5,024	9,543,224	1,167,304	147,358,070	7438 ± 10,250	1383 ± 1966	29,331 ± 45,396
		DE	4,578	7,493,831	750,767	119,863,992	6332 ± 8629	1238 ± 1712	26,183 ± 41,828
		UK	4,534	7,051,041	773,461	111,651,943	6477 ± 8710	1304 ± 1856	24,625 ± 41,000
		PL	4,408	6,101,991	662,239	138,766,755	6351 ± 8163	1127 ± 1645	31,481 ± 38,963
		BR	3,886	4,969,966	495,944	114,103,850	5792 ± 7361	964 ± 1280	29,363 ± 40,601
		Other	22,505	26,222,237	2,656,266	691,871,438	7524 ± 9766	1409 ± 1927	30,743 ± 43,813
	Age	All	46,120	37,944,004	3,863,504	1,341,935,622	6944 ± 9166	1285 ± 1862	29,097 ± 42,390
		≥30	8,892	16,611,760	1,630,651	245,663,773	8105 ± 11,804	1550 ± 2202	27,628 ± 48,742
		<30	37,228	30,394,356	3,178,611	1,096,271,849	6666 ± 8390	1222 ± 1765	29,448 ± 40,719

3. *LFM-2b* Dataset

In this section, the *LFM-2b* and the *LFM-2b-DemoBias* datasets are introduced. Aspects such as data acquisition procedures, accessibility, as well as the main characteristics of each collection, will be discussed in the following.

3.1. Data acquisition and accessibility

The *LFM-2b*(illion) dataset is a large collection of music listening events (LEs), i. e., users' interactions with the music online platform *Last.fm*,[6] enriched by users' demographic information (i. e., users' age, country, and gender), music-related metadata (e. g., artist and track names), and timestamps (specific time when a particular track was listened to by a given user). Following the methodology applied in the acquisition of the *LFM-1b* dataset (Schedl, 2016), the *LFM-2b*[7] was collected from the web streaming service *Last.fm* using the *Last.fm* API.[8] The *LFM-2b* (encompassing more than 2 billion of LEs) is an extension of the former, containing the same 120,322 users but with listening histories extended over 15 years: from 14 February 2005, until 20 March 2020; which yields $2014,164,872$ LEs in total.[9]

In order to enable reproducibility of our results and to foster further experiments on bias and fairness in the music recommendation domain, the *LFM-2b* dataset is stored in form of tabular data encoded in UTF-8. LEs are codified in a unique file containing for each row a LE, for each column users' demographics, music metadata, and the timestamp. Users' demographics are: user ID (unique for each user), gender information (female or male), country, and age. Musical attributes are: track ID (unique for each track) and track identifier (track and artist names combination). Note that, with track we refer to each unique musical item produced by a specific artist. Furthermore, user-track-playcount matrix (UTM) and user-artist-playcount matrix (UAM), i. e., two 2-dimensional matrices containing the interactions between unique user-track and user-artist, respectively, are also provided as sparse matrices in Python NumPy format.[10]

3.2. From LFM-1b to LFM-2b

In Table 1, descriptive statistics for the *LFM-1b*, the *LFM-2b*, the $LFM\text{-}2b\backslash 1b_{diff}$ (set difference between *LFM-1b* and *LFM-2b*), and the *LFM-2b-DemoBias* (Demographic Bias subset), are displayed.[11] For each dataset, the number of Users, Tracks, Artists, and

[6] https://www.last.fm/.

[7] The LFM-2b dataset used in our study is considered derivative work according to paragraph 4.1 of *Last.fm*'s API Terms of Service (https://www.last.fm/api/tos). The *Last.fm* Terms of Service further grant us a license to use this data (according to paragraph 4).

[8] https://www.last.fm/api.

[9] Note that 66,120 users included in LFM-1b have become inactive since 2014. Therefore, the extension provided in LFM-2b concerns only listening events for 54,202 users of the LFM-1b dataset.

[10] The *LFM-2b* is freely accessible at http://www.cp.jku.at/datasets/LFM-2b/.

[11] Please note that the number of artists and tracks reported here slightly differs from the numbers reported in Schedl (2016) because of some errors in the original *LFM-1b* dataset, which we corrected when creating *LFM-2b*.

Listening Events (LEs), as well as the mean and standard deviation (indicated after ±) of users' interactions in terms of (unique) tracks per user, (unique) artists per user, and total LEs per user (see Tracks/User, Artists/User, and LEs/User, respectively) are reported.

Although there is only a difference of 6 years in the length of the listening histories collected for the *LFM-1b*[12] w. r. t. the *LFM-2b* dataset, the latter represents a considerably larger range of listened tracks and artists: 31,413,999 versus 50,813,373, and 3,116,790 versus 5,217,014 for *LFM-1b* versus *LFM-2b* (see Track and Artist, respectively, in Table 1). Similarly, the *LFM-2b* contains approximately double the number of LEs than the *LFM-1b*: 1,088,161,692 versus 2,014,164,872 (see LEs in Table 1). When evaluating the set difference between *LFM-1b* and *LFM-2b*, i. e., the LEs from *LFM-2b* collected only during the last 6 years (see *LFM-2b*$\backslash$*1b*$_{diff}$), we observe that the number of LEs from *LFM-2b*$\backslash$*1b*$_{diff}$ is comparable to the one from *LFM-1b* (collected during 9 years), which indicates that the users have increased their music consumption within the platform in the last 6 years: 1,088,161,692 versus 926,003,180, respectively, for *LFM-1b* versus *LFM-2b*$\backslash$*1b*$_{diff}$ (see LEs in Table 1). This goes along with the general raise in social media usage displayed in recent years, which emphasized the importance of using up-to-date datasets in the evaluation of users' music consumption.

By calculating the differences in the coefficient of variation (CV$_{diff}$), i. e., the difference in the ratio of the standard deviation to the mean (between *LFM-1b* and *LFM-2b*$\backslash$*1b*$_{diff}$) for each type of interaction, a general increment in the variability of the users' consumption habits is revealed. The smallest increment is shown for the interactions in terms of LEs (see LEs/User for *LFM-1b* versus *LFM-2b*$\backslash$*1b*$_{diff}$ in Table 1), yielding CV$_{diff}$ = 20%. The largest increment is found for the interactions in terms of artist (see Artists/User for *LFM-1b* versus *LFM-2b*$\backslash$*1b*$_{diff}$ in Table 1), yielding CV$_{diff}$ = 70%. In between, the interactions in terms of track (see Tracks/User for *LFM-1b* versus *LFM-2b*$\backslash$*1b*$_{diff}$ in Table 1), yielding CV$_{diff}$ = 50%. Overall, this indicates that in the last 6 years, users' listening behavior changed especially concerning artist variability, meaning that many users increased substantially the number of artists they listen to. It is also evidenced by comparing the coefficient of variation (CV) for the interaction Artists/User of each collection: CV = 190% versus CV = 120% for Artists/User in *LFM-2b*$\backslash$*1b*$_{diff}$ versus *LFM-1b*, respectively. Differently, the amount of interactions within the platform remained more stable across users: CV = 200% versus CV = 180% for LEs/User in *LFM-2b*$\backslash$*1b*$_{diff}$ versus *LFM-1b*, respectively. As expected when working with data collected in-the-wild, both *LFM-1b* and *LFM-2b*$\backslash$*1b*$_{diff}$ display great differences across users' consumption behavior, i. e., some users have a much lower number of interactions than others. Yet, *LFM-2b*$\backslash$*1b*$_{diff}$ indicates that this differences across users have become more salient in terms of artist, which indicates an increased interest of users toward musical diversity. All in all, *LFM-2b*, being the union between *LFM-1b* and *LFM-2b*$\backslash$*1b*$_{diff}$, presents a considerably higher amount of LEs per user w. r. t. *LFM-1b*. Furthermore, it is more up-to-date and shows also a much higher artist variety; thus, being particularly suitable for assessing the performance of music recommender systems in general and for studying their bias in particular.

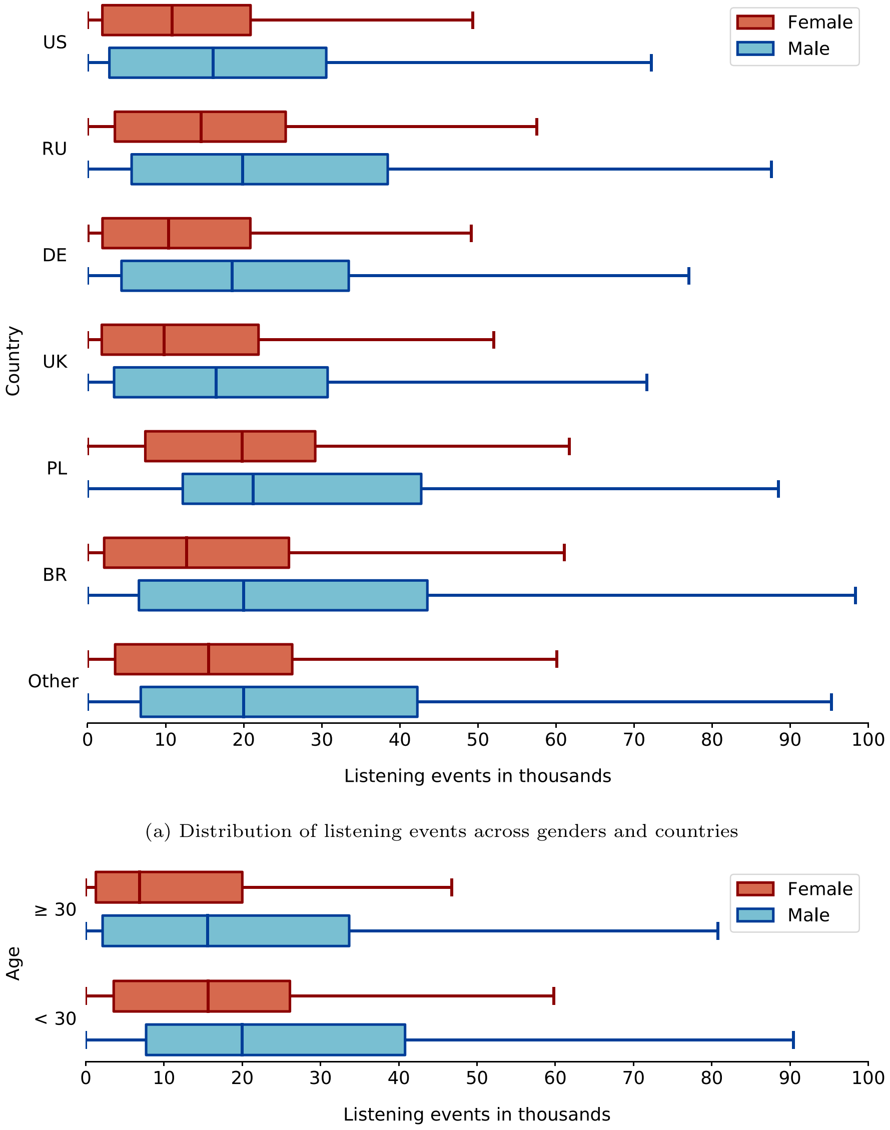
3.3. *LFM-2b-DemoBias: a collection for studying fairness*

In order to examine data and algorithmic/model bias in music RSs, along to the *LFM-2b* dataset, we introduce the *LFM-2b-DemoBias* (Demographic Bias) subset, which contains music LEs for users with valid demographic information in terms of age, gender, and country. Since bias might be independently investigated for gender, age, and country, we intentionally consider in the *LFM-2b-DemoBias* users who have at least one of the demographic attributes, instead of all three. Note that all the filtered collections, i. e., the subsets at the intersection between several demographic attributes (e. g., LEs for users with a valid gender and age information) as well as the playcount matrices, can be inferred from the *LFM-2b* directly.

The *LFM-2b-DemoBias* encompasses a total of 60,972 users (see All for *LFM-2b-DemoBias* in Table 1), from which 55,771 provide gender information (see All for Gender in Table 1), 55,190 country information (see All for Country in Table 1), and 46,120 age information (see All for Age in Table 1). Within each demographic group, i. e., gender, country, and age, there is an unbalanced distribution of users between sample: for instance, 15,802 female versus 39,969 male (see F and M for User in Table 1); 10,255 users from the USA, i. e., the first ranked country, which consists of more than twice as many users in comparison to any other country. (see US for User in Table 1); 37,228 users under thirty years versus 8,892 users above (see <30 and ≥ 30 for User in Table 1). As expected, this is similarly displayed when evaluating the collection at the track and artist level (see differences between sample within the Gender, Country, and Age groups for Track and Artist in Table 1). We observe, indeed, a great diversity between countries concerning the unique artists listened to: despite their differences in number of users, the USA and Russia both show a high number of unique artists (see 1,234,159 and 1,167,304, for Artists in US and RU in Table 1); Germany, the UK, and Poland – while similar in number of users – show a considerably lower diversity (750,767, 773,461, and 662,239 Artists in Table 1). Such unbalanced distributions, characteristic of datasets collected in-the-wild, indicates the population bias of *LFM-2b*, in which male, US, and young listeners represent the dominant group of users. In other words, the population of the *LFM-2b-DemoBias* is distorted w. r. t. the real world population, which does not contain such a bias. The population bias shown in *LFM-2b-DemoBias* makes this collection particularly suited to investigate model/algorithm bias, since it enables to assess to which extent a given recommender algorithm might create a model that reflects, amplifies, or alleviates this data bias.

Since the aim of the study at hand is to investigate *gender* fairness, we further inspect the interaction of the gender attribute in *LFM-2b-DemoBias* with the other two demographics, i. e., country and age. Therefore, in Fig. 1, the distribution of listening events across users containing information for at least two demographic attributes, i. e., gender and country (see Fig. 1a), or gender and

[12] The listening histories of *LFM-1b* cover more than 9 years: from February 2005 until August 2014.



(a) Distribution of listening events across genders and countries

(b) Distribution of listening events across genders and ages

Fig. 1. Distribution of thousands of listening events per user across countries (a) and ages (b) for female and male (indicated in red and blue, respectively). The median and the four quartiles are indicated: first quartile, second quartile, median, third quartile, and fourth quartile (from left to right). Countries with more than 2000 users are indicated individually, with LEs than 2000 aggregated (and denoted as "Other"); for the countries' abbreviations and number of users per country, see caption of Table 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

age (see Fig. 1b) is displayed. Although the US is the country with more users, their interactions within the platform are generally lower than in other countries with much users, such as Poland. This is particularly clear for male users: those from the US show the lowest median across male (16,108), those from Poland the highest (21,223); see median for male US and male PL, respectively, in Fig. 1 (upper plot). This higher consumption from Polish users is also clearly displayed for female, whose median (19,813) is not only almost as high as that displayed by male users, but is as high as the median for all the other countries with less than 2000 users (i. e., other), and overtakes that displayed by male users from many countries, including the US; see median for female PL and male other and US, in Fig. 1 (upper plot). Concerning age, similar trends between female and male can be observed for the two displayed age groups, i. e., above and below the thirties: for both, female and male, only one fourth of the users per group (under and over 30) produces more than half of the LE; see the fourth quartile for each gender and group, in Fig. 1 (lower plot). This skewed distribution was already shown in the *LFM-1b* dataset, and it is also clearly shown for the demographic information of country (see Fig. 1, upper plot).

4. Measuring user group fairness in recommendation

In this section, we define and formulate our approach to quantify fairness/bias of a RS, as well as the compounding factor. We first explain *RecGap*, a metric of measuring unfairness of RSs from the point of view of user groups (Section 4.1). The proposed *RecGap* metric is closely related to the *Gap* metric introduced by De-Arteaga et al. (2019) while expands *Gap* to any arbitrary evaluation measure, making it suitable for RSs. We then describe the concept of compounding imbalances in RSs, and suggest a metric to capture this concept (Section 4.2). The metrics discussed in this work are formulated for an arbitrary number of user groups (non-binary setting) for the sensitive attribute. The metrics are defined for the set of user groups G, which, in the case of the binary setting of the sensitive attribute gender, is equal to $G = \{female, male\}$.

4.1. RecGap : Recommendation unfairness metric

Our notion of fairness of a RS closely relates to the *equal opportunity* metric of Hardt et al. (2016). We consider a RS to be fair if it performs equally good across the groups of users, according to any arbitrary evaluation metric. To measure the fairness of a recommender algorithm, first an RS model is trained – in the case that the algorithm requires training – and then its recommendation predictions are evaluated separately for each groups of users in the test set. The fairness of the RS model regarding the given evaluation metric is defined as the mean absolute differences between the average evaluation scores (across all users) in each pair of groups. We refer to this metric as *RecGap*, formulated as follows:

$$RecGap^{\mu} = \frac{\sum_{\langle g,g'\rangle \in G^{pair}} \left| \frac{\sum_{u\in U_g} \mu(u)}{\left|U_g\right|} - \frac{\sum_{u'\in U_{g'}} \mu(u')}{\left|U_{g'}\right|} \right|}{|G^{pair}|} \tag{1}$$

where G^{pair} is the set of groups pairs,[13] U_g denotes the set of users in group g, and $\mu(u)$ returns the evaluation results of the metric μ for user u (assuming the availability of corresponding ground-truth data). Intuitively, the *RecGap* metric quantifies the disparity between average results, in terms of $\mu(u)$, of different user groups.

The metric μ can be any evaluation measure that can be calculated on user level (and aggregated on user-group level) such as NDCG, diversity, recall, etc. According to this formulation, *RecGap* uses the average values across users (as common in evaluation metrics), and does not impose the groups to have an equal number of users. This characteristic of *RecGap* is particularly important in real-world scenarios, as the number of users across sensitive attributes are typically different. $RecGap^{\mu}$ can have zero or positive values, where zero indicates that the recommendation algorithm is fair, while a higher positive value shows the existence of a higher degree of unfairness. Interpreting the magnitude of $RecGap^{\mu}$ depends on the considered μ function and the number of groups. As a simple example for the recall metric, let us consider a binary gender setting with the average recall scores of 0.5 and 0.3 for male and female, respectively. In the example, $RecGap^{\text{Recall}}$ is equal to 0.2, which implies that the average male user receives in their top-10 results an increase of 20 percentage of the relevant items, when compared with the ones provided to the average female user.

In the particular case of G being a binary set, $RecGap^{\mu}$ is simplified to the absolute difference of the groups' means of evaluation scores. In this case, the magnitude of $RecGap^{\mu}$ can be assessed by applying hypothesis testing, which examining whether the difference between groups is statistically significant, indicates to which extent the $RecGap^{\mu}$ is meaningful or not. Finally, the direction of bias/unfairness in *RecGap* can simply be identified by comparing the degrees of the groups' mean values: the model is biased toward the group with the higher mean score (considering that higher μ values means higher gains).

4.2. Compounding imbalances in recommender systems

Our definition of compounding imbalances follows the work of De-Arteaga et al. (2019) by extending the concept of compounding to RSs according to any arbitrary evaluation metric. Considering a recommendation algorithm with some degree of unfairness across user groups ($RecGap \neq 0$), we aim to quantify the *extent to which the algorithm compounds the initial imbalances in the data given any arbitrary evaluation metric*. More concretely, similar to De-Arteaga et al. (2019), we expect that the gain of each group according to an evaluation metric to be proportional to its population, where otherwise the population bias is compounded by the algorithm.

To define such a metric, we first introduce the population distribution B as the distribution of the portions of users in groups. For instance in the case of a binary gender setting, B can be $B = [0.8, 0.2]$, meaning that 80% of the users are male and the rest are female. Given a recommendation algorithm and the evaluation metric μ, we define the *metric scores distribution* over user groups, denoted by $C^{\mu} = \{c^{\mu}_g | g \in G\}$. Each c^{μ}_g element of C^{μ} is a probability, defined as the sum of the evaluation scores of all users in group g divided by the sum for all users in all groups:

$$c^{\mu}_g = \frac{\sum_{u\in U_g} \mu(u)}{\sum_{g'\in G} \sum_{u'\in U_{g'}} \mu(u')} \tag{2}$$

The C^{μ} distribution contains the portion of the score of the metric μ regarding each group. Following the above example of the binary genders for a RS biased toward males according to NDCG ($RecGap^{\text{NDCG}} \neq 0$), the distribution of NDCG scores might results in

[13] $G^{pair} = \{\langle g, g'\rangle | g, g' \in G, g \neq g'\}$.

Table 2
Statistics of the *LFM-2b-DemoBias*$_{Sub}$ dataset. Number of Users, Tracks, Artists, and LEs are reported across F(emale) and M(ale) separately and also together (All). Mean and standard deviation (indicated after ±) of the interactions of users with tracks, artists, and listening events are indicated in the last three columns, respectively.

Gender	Users	Tracks	Artists	LEs	Tracks/User	Artists/User	LEs/User
All	19,972	99,831	40,182	19,906,272	142 ± 172	128 ± 150	997 ± 1571
F	4,415	70,980	32,414	3,397,310	101 ± 121	93 ± 110	769 ± 1158
M	15,557	99,810	40,176	16,508,962	153 ± 182	138 ± 158	1061 ± 1664

$C^{\mathrm{NDCG}} = [0.85, 0.15]$, namely the value 0.85 for male and 0.15 for female. Now, comparing C^{NDCG} with the population distribution $B = [0.8, 0.2]$, we observe that the bias of the population toward male (80% to 20%) has now intensified (toward male) in the results of the recommnender (85% to 15%). In other words, the recommender algorithm compounds the population bias.

To characterize the differences between these distributions as one single number, we introduce the *compounding factor* metric, defined as the Kullback–Leibler (KL) divergence between the distributions. The compounding factor of a recommendation algorithm regarding the metric μ ($CompFct^{\mu}$) is therefore formulated as:

$$CompFct^{\mu} = \mathrm{KL}\left(B \middle\| C^{\mu}\right) \tag{3}$$

The value of $CompFct^{\mu}$ metric shows the extent of the divergence of C^{μ} from B, where higher values indicate higher degrees of bias compounding by the algorithm. This value in the example above, given $B = [0.8, 0.2]$ and $C^{\mathrm{NDCG}} = [0.85, 0.15]$, is $CompFct^{\mathrm{NDCG}} = 0.0101$.

5. Experiment setup

In this section, we explain our experiment setups. Overall, we conduct all the experiments on the data of the Gender subgroup in *LFM-2b-DemoBias* considering two experiment scenarios. In the first one, i. e., Standard, we carry out the experiments on the dataset without any intervention. This scenario reflects how the recommender systems algorithms under investigation would behave in the wild and what would be their unfairness treatments of users on the basis of gender. The second scenario, i. e., Resampled, considers instead a debiasing procedure applied on the original dataset. The debiasing attempts to reduce the difference in treatment of males and females and corresponds to a scenario where the acknowledged gender-based differences are addressed. In each scenario, we study a variety of core collaborative filtering (CF) algorithms of a different nature (e. g., matrix factorization Billsus, Pazzani, et al., 1998 and autoencoders Zhang, Yao, Sun, & Tay, 2019), as well as several evaluation metrics (both accuracy-related and beyond-accuracy). In the subsequent subsections, we detail the processing steps performed to obtain the dataset used in the experiments (Section 5.1), all algorithms investigated (Section 5.2), the procedure of training and evaluation of algorithms (Section 5.3), the experiments scenarios (Section 5.4), the used metrics (Section 5.5), the used significance tests (Section 5.6), and our approach to hyper-parameter tuning (Section 5.7). Our code for reproducing the experiments is publicly available at https://github.com/CPJKU/recommendation_systems_fairness.

5.1. Data processing and preparation

In our experiments, we focus on the Gender subgroup of the *LFM-2b-DemoBias* (see the part related to Gender in Table 1). We process the data of this subgroup according to the following filtering criteria. First, we consider only user–track interactions with a playcount (PC) > 1. This removes possibly-noisy user–track interactions likely introduced by single interactions. Second, we consider only tracks listened to by at least 5 different users and users that listened to at least 5 different tracks. These thresholds, commonly used in previous work (Bauer & Schedl, 2019; Liang, Krishnan, Hoffman, & Jebara, 2018; Ning & Karypis, 2011; Schedl, 2017), are necessary for a meaningful use of collaborative filtering algorithms. Third, we consider only LEs collected within the last 5 years. This makes our study focused on the users' most recent listening behaviors which, as shown in Section 3, have considerably increased in the last years. Finally, we transform the user–track interactions to binary values, namely by setting a user–track interaction to 1 if the user has listened to the track at least once, to 0 otherwise.

The resulting dataset consists of 23,272 users – a favorable setting for studying fairness from user perspective – but also a very high number of items/tracks (1,606,686 in total), which makes it impractical for large-scale recommendation experiments. We address this issue by randomly sampling 100,000 tracks. Note that applying random sampling guarantees that tracks covering different levels of popularity are included in the final dataset.[14] We refer to this final subset as *LFM-2b-DemoBias*$_{Sub}$. The statistics of the dataset are reported in Table 2. We should note that even with this reduction, the *LFM-2b-DemoBias*$_{Sub}$ dataset is still larger than the Million Song Dataset (Bertin-Mahieux et al., 2011) (containing 41,000 items), which is commonly used in research on music RSs.

[14] We consider the *LFM-2b-DemoBias*$_{Sub}$ dataset a reliable representation of the Gender subgroup of *LFM-2b-DemoBias*. This can be seen by the small differences in the coefficients of variation (CV_{diff}) between the *LFM-2b-DemoBias*$_{Sub}$ and the Gender group of *LFM-2b-DemoBias* across interactions: $\mathrm{CV}_{diff} = 10\%$ for Tracks/User; $\mathrm{CV}_{diff} = 20\%$ for Artists/User; $\mathrm{CV}_{diff} = 0\%$ for LEs/User.

5.2. Recommender system algorithms

We investigate to which extent different recommendation algorithms for implicit data yield different results, depending on users' demographic traits. The selected algorithms cover different types of collaborative filtering approaches. In particular, we study algorithms based on non-personalized recommendation, matrix factorization, k-nearest neighborhood, and autoencoders, which have been central in RS research. All the algorithms are applied to a user–item interaction matrix, where users are represented in the rows and items in the columns. If a user interacted with an item, the corresponding value in the matrix is 1 and otherwise 0. In the following, we provide a brief explanation of each of the RS algorithms studied in this work:

- *Popular Items* (POP) provides a simple non-personalized baseline. It recommends to the users the same set of top-n tracks, where the tracks are sorted by overall popularity (how many users listened to that track).
- *Item k-Nearest Neighbors* (ItemKNN) (Sarwar, Karypis, Konstan, & Riedl, 2001) is a basic memory-based recommendation approach based on computing item–item similarity. In this approach, an item is recommended to a user if the item is similar to the items previously selected by the user. In the case of CF systems, items selected by the same group of users are considered more similar than items with non-overlapping user groups.
- *Alternating Least Squares* (ALS) (Hu, Koren, & Volinsky, 2008) falls in the category of matrix factorization approaches, a widespread family of algorithms since the Netflix challenge (Billsus et al., 1998). ALS employs an alternating training procedure to obtain a set of user and item embeddings, in such a way that the dot product of the embeddings approximates the original user–item matrix.
- *Bayesian Personalized Ranking* (BPR) (Rendle, Freudenthaler, Gantner, & Schmidt-Thieme, 2012) provides an optimization function that, instead of predicting the rating for a specific pair of user and item, ranks the items consumed by the users according to their preferences (hence, personalized ranking). To this end, BPR defines an implicit order between pairs of items. BPR maximizes the difference between the rating prediction of items that have interactions with the user and the ones with no interaction. We apply the BPR objective function on matrix factorization embeddings.
- *Sparse Linear Methods* (SLIM) (Ning & Karypis, 2011) is a linear model that aims to compute top-n recommendations, by factorizing the item–item co-occurrence matrix under the non-negativity, L_1, and L_2 constraints. The learned item coefficients are used to sparsely aggregate past user interactions and predict the recommended items of the user.
- *Variational Autoencoders* (MultiVAE) (Liang et al., 2018) is a variational autoencoder architecture that first projects the sparse user's interaction vector to a latent distribution space, used afterwards to generate a probability distribution over all the items. MultiVAE employs multinomial likelihood and a different regularization procedure involving linear annealing.

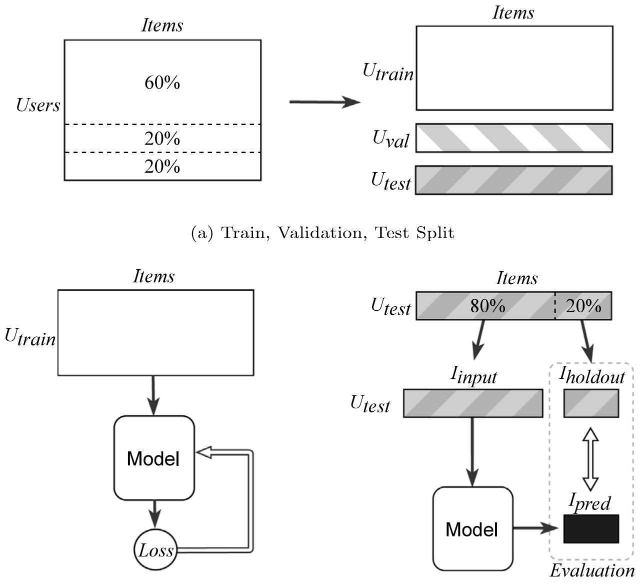
5.3. Experiment procedure

In this section, we describe in detail the procedure of training and evaluation of the algorithms, accompanied with an applied cross validation method.

For the recommendation task at hand, different evaluation objectives and data splits have been proposed in the RS literature (Meng, McCreadie, Macdonald, & Ounis, 2020; Sun et al., 2020). In our experiments, we employ a *User Split* strategy (Meng et al., 2020) of the dataset, commonly used for autoencoder-like algorithms (Liang et al., 2018; Sachdeva, Manco, Ritacco, & Pudi, 2019; Steck, 2019), among others. The splitting strategy is shown in Fig. 2a. The 19,972 users of the *LFM-2b-DemoBias*$_{Sub}$ dataset are partitioned in train, validation, and test set using a common 60-20-20 ratio split. The users in the training set, along with all their interactions, are used to train the algorithms in analysis (Fig. 2b). The evaluation procedure (either validation or testing) is carried out by feeding 80% of the users' items sampled uniformly at random to the models and using the remaining 20% as ground truth for calculating the metrics (Fig. 2c). Intuitively, the evaluation procedure forces the models to learn the broad music tastes of the users instead of only predicting what the user is going to listen next (e. g., as in leave-k-out strategies Meng et al., 2020). This experiment setup is also referred to as *strong generalization* (Liang et al., 2018; Marlin, 2004) since we evaluate the recommender systems on novel users not encountered during training.[15]

In order to provide evaluation for all users in the dataset and also avoid possible biases introduced by the user-sampling strategy mentioned above, we follow the standard practice in machine learning and perform 5-fold cross validation as shown in Fig. 3a. In more detail, we split the users in 5 equal-sized groups and use 3 groups for training (60% of the users), 1 for validation (20% of the users), and 1 for testing (20% of the users). For each fold, we follow the training and evaluation procedure described above. We switch the groups in a round-robin fashion until each one of the user groups is used as a test set. Applying cross-validation provides better estimates for the metrics and also allows testing all the users in our dataset, leading to a better comparison based on the gender attribute.

[15] In contrast, *weak generalization*, instead, refers to the setup where part of the user listening history is used for training (e. g., leave-k-out splitting strategies). In these cases, the model is already exposed to the consumption patterns of these users, possibly leading to better prediction performance during inference and, hence, higher accuracy metrics.

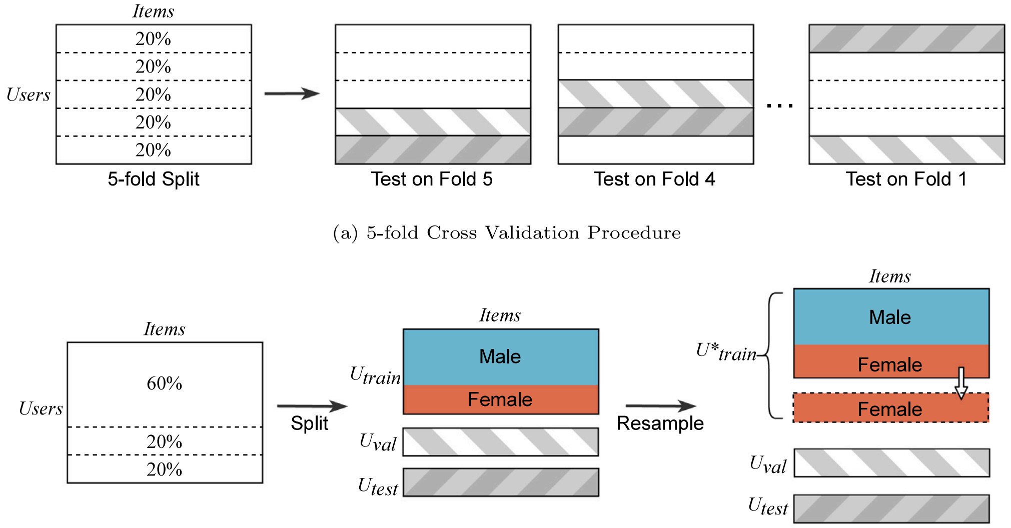


(a) Train, Validation, Test Split

(b) Training Procedure

(c) Evaluation Procedure

Fig. 2. User-based split. As shown in Fig. 2a, 60% of users in the datasets are used as training data, 20% as validation data, and 20% as test data. The users, and all their items, are used to train the model under investigation (cf. Fig. 2b). For the evaluation procedure (cf. Fig. 2c), 80% of the test user's items randomly selected are provided as input to the model. Subsequently, using the remaining 20% of the data as holdout set ($I_{holdout}$), the metrics are computed on the model output (predicted items I_{pred}) and the holdout set.



(a) 5-fold Cross Validation Procedure

(b) Resampling Procedure

Fig. 3. 5-fold cross validation and the resampling procedure (cf. Figs. 3a and 3b, respectively). The 5-fold cross validation cyclic procedure guarantees that every user ends up in a test set once (shown with the darker texture for each of the five iterations) and in a validation set once (shown with the lighter texture). In Fig. 3b, the resampling procedure is illustrated for a specific fold: the female users in the training data are resampled until they match in frequency the male users.

5.4. Experiment scenarios

We perform all the experiments bearing in mind two scenarios: Standard and Resampled. In the Standard scenario, we train the system without any intervention on the data, which corresponds exactly to the procedures described in Section 5.3. For the Resampled scenario, we attempt to debias the recommendation algorithm by intervening on the underlying dataset using resampling, as shown in Fig. 3b. Firstly, following the procedure outlined in Section 5.3, we split the users in training, validation, and test, as done in the Standard scenario. Secondly, we resample the users of the minority group (female) in the training set until they match the number of training data points of the majority group (male). Note that when a user is resampled, her listening history is duplicated and used fully in the training procedure. Following Geyik et al. (2019), by providing a balanced representation of male and female users during training, we aim to promote equally good recommendations at inference time. Furthermore, since the validation and test sets are left untouched, the evaluation results from the Standard and Resampled scenarios are comparable.

5.5. Evaluation metrics

We evaluate the performance of the algorithms using two accuracy-based metrics: recall and Normalized Discounted Cumulative Gain (NDCG). We also evaluate the results using two beyond-accuracy metrics: diversity and coverage. All metrics are calculated over a ranking result up to the position K. We provide a brief explanation of the metrics in what follows.

Recall@K for user u is defined as:

$$Recall@K(u) = \frac{1}{\min(K, N_u)} \sum_{i=1}^{K} rel(i) \tag{4}$$

where N_u is the number of items in the test set which are relevant to u, and $rel(i)$ is an indicator function signaling whether the recommended track at rank i is relevant to u (i.e., $rel(u) = 1$) or not relevant to u (i.e., $rel(u) = 0$). Recall@K quantifies the ability of retrieving relevant items for the user in analysis. It ranges from 0, where no relevant items for the user are retrieved, to 1, where all relevant items are present in the first k position.

NDCG@K is defined as

$$NDCG@K(u) = \frac{DCG@K(u)}{IDCG@K(u)} \tag{5}$$

where $IDCG@K(u)$ is the ideal $DCG@K$ for user u, obtained when all items in u's test set are ranked at the top , and $DCG@K(u)$ is the discounted cumulative gain at position K for user u, given by

$$DCG@K(u) = \sum_{i=1}^{K} \frac{rel(i)}{\log_2(i+1)} \tag{6}$$

where $rel(i)$ is the same indicator function as above. Compared to Recall@K, NDCG@K is an accuracy metric that not only quantifies the ability of retrieving relevant items, but also the ability of ranking them. A recommender system algorithm that provides relevant items at the top of the list will score higher in NDCG@K than an algorithm for which the relevant items are at the bottom. This behavior is enforced by "discounting" items according to their position, i. e., computing the Discounted Cumulative Gain (DCG). To normalize the score between 0 and 1, the DCG is then compared to the so-called "ideal ranking" obtained by placing all the relevant items at the top of the list.

Diversity is calculated for each user as normalized Shannon entropy on the *artist* level:

$$Diversity@K(u) = -\frac{1}{\log_2 |A_u|} \sum_{i=1}^{|A_u|} p(a_i)\log_2 p(a_i) \qquad a_i \in A_u \tag{7}$$

where A_u is the set of unique artists whose tracks were recommended (in top K) to the user u, and $p(a_i)$ is the proportion of the tracks by the artist a_i in top K of the recommendation list. $Diversity@K(u)$ is therefore equal to 1 if every track among the top K tracks recommended to the user u has a different artist. $Diversity@K(u)$ becomes 0 if all top K recommended tracks come from the same artist.

Finally, $Coverage@K$ is defined as the fraction of tracks in the test set that are included in the top K recommendation list of at least one user.

In our experiments, we compute the metrics for $K = \{5, 10, 50\}$. This aims to model different user needs, ranging from a user interested in only a few top recommendations, to a user who inspects a longer list of recommended items. When discussing results in detail (in Section 6), we focus on the setting $K = 10$ because this is the number of tracks Last.fm's recommender displays to their users by default. Results for $K = \{5, 50\}$ are provided in Appendix A.

5.6. Significance test

We test the significance of the differences of results in two settings. The first setting (considered to examine the results of *RecGap*) regards the differences of one recommendation algorithm between two groups with different sizes, i. e., females versus males. The second setting concerns the differences of one recommendation algorithm between two application scenarios, i. e., the Standard and the Resampled (see Section 5.4). In both cases, we perform the Mann–Whitney U test, also known as the Wilcoxon rank-sum test (McKnight & Najab, 2010).[16] In addition, pairwise comparisons across the different models for each scenario are carried out. For these, Dunn test and Bonferroni correction for the p-values adjustment is applied.

[16] Note that we consider a non-parametric test since our data do not follow a normal distribution as Kolmogorov–Smirnov test rejects the null-hypothesis for all the considered samples.

Table 3
Overall results of accuracy (NDCG and recall) and beyond-accuracy (diversity and coverage) metrics on the RS algorithms: POP, ItemKNN, BPR, ALS, SLIM, and MultiVAE; considering two experiment scenarios: Standard and Resampled for all users together (female and male). All results are rounded to the third digit. Statistically significant differences between the Standard and Resampled datasets for each model and metric are indicated with an asterisk (∗) on the highest value between Standard and Resampled. Highest values for each metric are shown in bold.

Model	Scenario	NDCG@10	Recall@10	Diversity@10	Coverage@10
POP	Standard	.046	.043	.983	.000
	Resampled	.045	.043	.987*	.000
ItemKNN	Standard	.301*	.284*	.984	**.157***
	Resampled	.292	.275	.985*	.128
BPR	Standard	.127	.117	.988	.126
	Resampled	.123	.115	.988	.124
ALS	Standard	.241	.220	.979*	.055*
	Resampled	.238	.218	.978	.052
SLIM	Standard	**.364**	**.340**	**.991**	.115*
	Resampled	.359	.333	.990	.110
MultiVAE	Standard	.192*	.175*	.985	.046
	Resampled	.183	.167	.985	.048

Considering that we carry out the experiments through 5-fold cross-validation, in each experiment, five independent statistical tests (one for each fold) are performed. Subsequently, the resulting p-values were combined by applying the weighted Stouffer's Z-method (Mosteller, Bush, & Green, 1954; Stouffer, Suchman, DeVinney, Star, & Williams Jr, 1949).[17] We select the weighted Z-method as it is robust to asymmetry problems (Whitlock, 2005) and less sensitive to a single low p-value, i. e., in order to achieve a low combined p-value, several consistently low p-values are required (Darlington & Hayes, 2000). Furthermore, the weighted Z-method is also suitable when the combined p-values come from multiple tests of the same hypothesis (Whitlock, 2005), as in our study. In all our experiments, we consider the results with $p < 0.01$ as significant.

5.7. Hyper-parameters and training

We select the hyper-parameters of the algorithms under investigation by performing a grid search over different parameters and finding the best set of parameters according to the NDCG@50 results on the validation set. After validation, the best-performing model is selected and finally evaluated on the test set. We reselect the hyper-parameters for each fold.

In the following, we report the range of hyper-parameters relevant to each model.

For ItemKNN, we select the number of neighbors from {3, 5, 10}, and the similarity function among the cosine metric, Pearson correlation coefficient, and Jaccard coefficient. We also examine the effect of removing normalization in the above-mentioned similarity functions (by dismissing the denominator). We select the value for shrinkage (Bell & Koren, 2007) between {0, 10, 100}.

For ALS and BPR, we select the embedding size among {10, 100, 1000}, the number of iterations for training between {500, 1000}, and the regularization factor from {$1e-3$, $1e-4$}. For BPR, we further tune the learning rate with the values {$1e-3$, $1e-4$}.

For SLIM, we explore different α values (sum of the L_1 and L_2 coefficients) and L_1 *ratios* (ratio of L_1 coefficient in α). We search the α value among {0.5, 0.1, 0.01, 0.001}, and L_1 *ratio* from {0.1, 0.01}. We set the number of iterations to 500.

For MultiVAE, we explore different (symmetric) architectures and annealing procedures. We set the total number of epochs to 100 and the learning rate to $1e-3$. We examine various architectures, namely *I-500-I*, *I-1000-I*, and *I-1000-500-1000-I*. In the architectures, I is the total number of tracks, the numbers in the middle denote the dimensions of the latent embeddings, and the numbers in between (when existing) are the intermediary dimensions of the feed-forward networks with hyperbolic tangent non-linearity.[18] Regarding the annealing procedure described in the original paper (Liang et al., 2018) , we linearly anneal the regularization parameter by choosing the beta steps from the values {5000, 100 000}. We set the annealing cap to 1, i. e., the regularization is performed until its maximum value.

6. Results and discussion

In this section, we report the results of our experiments and discuss the findings. We first present the overall performance of the recommendation algorithms/models (Section 6.1), and then report the results of measuring fairness in recommendations (Section 6.2). In the current section, we only report the results of the ranking up to the position $K = 10$. The results regarding other positions (5 and 50) are reported in Appendix A. Furthermore, we carry out experiments also on the *LFM-1b* dataset with the exact experiment settings, whose results are reported in Appendix B.

[17] Note that, unlike the other metrics, coverage is defined only at the level of group of users instead of single users, meaning that one global measure is obtained as output for each cross-validation. Hence, Mann–Whithney U test was applied across folds without the need of combining p-values.

[18] We do not observe any further improvement by increasing the layers of the feed-forward networks and/or the size of dimensions.

Table 4
NDCG@10 results on the users of the male/female (M/F) groups for the Standard and Resampled scenarios. The value of *RecGap* shows the degree of favorable treatment toward (m)ales or (f)emales. Highest values are shown in bold. Statistically significant differences between the results of female and male are shown with † symbol. Score Dist. columns shows the metric score distributions across males and females. The value of *CompFct* shows the effect of compounding imbalances in data. The population distribution to calculate *CompFct* is $B = [0.779, 0.221]$.

Model	Scenario	All	M/F	*RecGap*	Score Dist. (M/F)	*CompFct*
POP	Standard	.046	.045/.049	.004 (f)	76.2/23.8	.001
	Resampled	.045	.044/.051	.007 (f) †	75.1/24.9	.003
ItemKNN	Standard	.301	.313/.259	.054 (m) †	81.0/19.0	.004
	Resampled	.292	.304/.250	.054 (m) †	81.1/18.9	**.005**
BPR	Standard	.127	.129/.117	.012 (m) †	79.6/20.4	.001
	Resampled	.123	.124/.116	.008 (m)	79.0/21.0	.001
ALS	Standard	.241	.251/.205	.046 (m) †	81.2/18.8	**.005**
	Resampled	.238	.248/.204	.044 (m) †	81.1/18.9	**.005**
SLIM	Standard	.364	.378/.315	**.063** (m) †	80.8/19.2	.004
	Resampled	.359	.372/.312	**.060** (m) †	80.8/19.2	.004
MultiVAE	Standard	.192	.197/.173	.024 (m) †	80.0/20.0	.002
	Resampled	.183	.188/.166	.023 (m) †	80.0/20.0	.002

6.1. Performance evaluation results

Table 3 shows the evaluation results of experiments, averaged over all users, for the two experiment scenarios, namely Standard and Resampled. We conduct significance tests (see Section 5.6) between Standard and Resampled for the results of each algorithm and metric.

Overall, SLIM shows the best performance in terms of the accuracy-based metrics (NDCG and recall) as well as for diversity across the two experiment scenarios. For all the pairwise comparisons, the difference between the scores achieved by SLIM in comparison with the other models is statistically significant. The lowest difference is observed between SLIM and ItemKNN in the Resampled scenario for diversity ($p = .0002$). On the other hand, ItemKNN has the highest score in terms of coverage. Except for the comparisons with SLIM and BPR (in both the Resampled and the Standard scenarios), all the other pairwise comparisons between ItemKNN and the other models show statistically significant differences. As expected, the non-personalized approach (POP) shows the lowest performance on the accuracy-based metrics and a value of 0.0 on coverage, as POP recommends the same set of items to all users.[19] Matrix factorization approaches (ALS and BPR) perform generally inferior than the memory-based ItemKNN in terms of the accuracy-based metrics, while BPR has a higher diversity in comparison with ItemKNN. Finally, MultiVAE generally performs weaker especially on the accuracy-based metrics. This observation is in contrast to the results reported in previous studies on smaller datasets (Dacrema, Boglio, Cremonesi, & Jannach, 2019). We suspect that the lower performance of MultiVAE is due to the large number of items in our datasets, which makes it harder for the algorithm to provide effective distributions of output predictions. We consider further analysis of this behavior of MultiVAE as future work.

Comparing the results of Standard with the corresponding ones of Resampled, we observe an overall decrease in performance, while in the majority of the cases no significant differences are observed. The cases where debiasing significantly harms the performance are for ItemKNN on all metrics, MultiVAE on NDCG and recall, and ALS on diversity and coverage.[20]

In the rest of this section, we study the results of fairness and compounding factors on these algorithms.

6.2. Fairness gap in recommender systems

In this section, we first discuss the evaluation results of *RecGap*, and then analyze the outcomes of the compounding factor. We calculate the *RecGap* metric, as explained in Section 4, on Standard and Resampled by evaluating the test set results separately on male and female user groups. Tables 4–7 repeat the overall evaluation results, but also report the results on each user group, as well as the calculated *RecGap* metric. The higher absolute values of *RecGap* indicate a higher degree of unfairness, while (m) or (f) indicates the direction of favored treatment, respectively, wherever it is toward (m)ales or (f)emales. We also report the significance of differences between the evaluation results related to the group of male versus female, shown by the dagger sign. Highest absolute values for each scenario and metric are shown in bold.

As shown, the majority of algorithms show the existence of a significant gap (*RecGap*) in performances in favor of the male user group, in particular on NDCG, recall, and coverage.[21] The diversity metric in general shows a very low gap with a slight tendency toward the female group.

[19] In the case of POP, the number of distinct items in all users' recommendation lists is obviously 10, resulting in a Coverage@10 value of only $\frac{10}{\sim 100{,}000}$.

[20] Notably, the debiasing results of POP on diversity shows a significant improvement.

[21] The only exception is POP which shows counter-bias, namely a *RecGap* value leaned toward female users. This case is discussed later in the section.

Table 5
Recall@10 results. Details are identical to Table 4.

Model	Scenario	All	M/F	*RecGap*	Score Dist. (M/F)	*CompFct*
POP	Standard	.043	.041/.051	.010 (f) †	74.0/26.0	**.006**
	Resampled	.043	.041/.053	.013 (f) †	72.9/27.1	**.009**
ItemKNN	Standard	.284	.294/.247	.047 (m) †	80.7/19.3	.004
	Resampled	.275	.285/.239	.046 (m) †	80.8/19.2	.004
BPR	Standard	.117	.119/.111	.009 (m) †	79.2/20.8	.001
	Resampled	.115	.116/.112	.003 (m)	78.4/21.6	.000
ALS	Standard	.220	.228/.191	.037 (m) †	80.8/19.2	.004
	Resampled	.218	.225/.191	.034 (m) †	80.6/19.4	.003
SLIM	Standard	.340	.351/.299	**.052** (m) †	80.5/19.5	.003
	Resampled	.333	.345/.294	**.050** (m) †	80.5/19.5	.003
MultiVAE	Standard	.175	.178/.161	.018 (m) †	79.6/20.4	.001
	Resampled	.167	.171/.155	.016 (m) †	79.6/20.4	.001

Table 6
Diversity@10 results. Details are identical to Table 4.

Model	Scenario	All	M/F	*RecGap*	Score Dist. (M/F)	*CompFct*
POP	Standard	.983	.983/.983	.000 (m)	77.9/22.1	.000
	Resampled	.987	.987/.987	.000 (m)	77.9/22.1	.000
ItemKNN	Standard	.984	.983/.985	.002 (f) †	77.9/22.1	.000
	Resampled	.985	.985/.987	.002 (f) †	77.9/22.1	.000
BPR	Standard	.988	.988/.989	.002 (f) †	77.9/22.1	.000
	Resampled	.988	.988/.990	.002 (f) †	77.9/22.1	.000
ALS	Standard	.979	.979/.980	.002 (f) †	77.9/22.1	.000
	Resampled	.978	.977/.980	**.003** (f) †	77.8/22.2	.000
SLIM	Standard	.991	.991/.991	.000 (f)	77.9/22.1	.000
	Resampled	.990	.990/.991	.001 (f) †	77.9/22.1	.000
MultiVAE	Standard	.985	.984/.987	**.003** (f)†	77.8/22.2	.000
	Resampled	.985	.984/.988	**.003** (f) †	77.8/22.2	.000

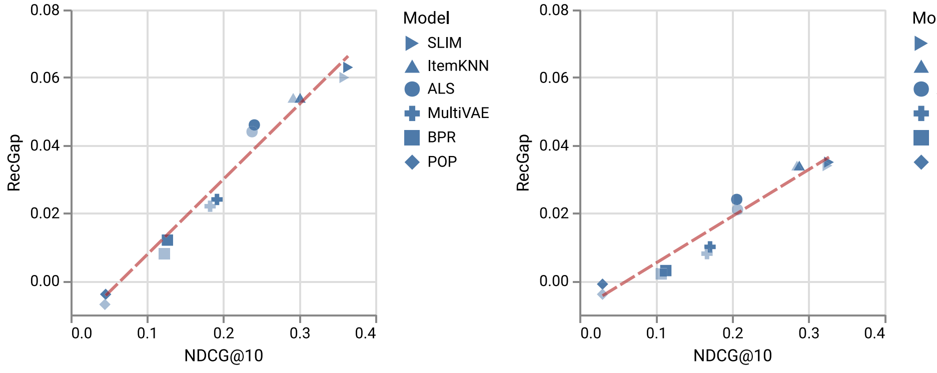
Table 7
Coverage@10 results. Details are identical to Table 4.

Model	Scenario	All	M/F	*RecGap*	Score Dist. (M/F)	*CompFct*
POP	Standard	.000	.000/.000	.000 (m)	79.6/20.4	.001
	Resampled	.000	.000/.000	.000 (m)	80.1/19.9	.002
ItemKNN	Standard	.157	.135/.051	**.084** (m) †	90.3/9.7	.097
	Resampled	.128	.111/.046	.065 (m) †	89.5/10.5	.082
BPR	Standard	.126	.115/.046	.069 (m) †	89.8/10.2	.087
	Resampled	.124	.113/.047	**.066** (m) †	89.4/10.6	.079
ALS	Standard	.055	.052/.028	.024 (m) †	86.6/13.4	.041
	Resampled	.052	.049/.028	.021 (m) †	85.9/14.1	.034
SLIM	Standard	.115	.100/.045	.056 (m) †	88.8/11.2	.069
	Resampled	.110	.096/.044	.052 (m) †	88.6/11.4	.066
MultiVAE	Standard	.046	.041/.016	.026 (m) †	90.4/9.6	**.098**
	Resampled	.048	.044/.015	.030 (m) †	91.4/8.6	**.120**

Across the algorithms, SLIM has the highest degree of unfairness on the accuracy-based metrics. This is particularly concerning as SLIM performs the best across all algorithms, making it a strong candidate for a potential RS when not taking into account the fairness measure. In fact, we can even observe an inverse relationship between the accuracy-based and fairness metric, namely *RecGap* becomes larger for the better performing algorithms such as ItemKNN and SLIM, while it decreases for the algorithms that perform worse in terms of accuracy, reaching the minimum with BPR and POP.

Looking at the effect of the debiasing method, we observe that *RecGap* only slightly decreases on Resampled in comparison with Standard on NDCG, recall, and coverage. This decrease, while marginal, is still valuable considering the fact that the debiasing method does not deteriorate the performance of the majority of the algorithms, such as SLIM, ALS, and BPR. These results indicates the need for further studying other algorithmic debiasing methods on this dataset, which we consider as a future direction.

We now inspect the relationship between the results of a performance measure, NDCG, with its corresponding *RecGap*. Fig. 4a depicts the NDCG@10 and its *RecGap* for different recommendation algorithms, evaluated on the *LFM-2b-DemoBias*$_{Sub}$ dataset. The darker mark for each recommendation algorithm indicates the result of the Standard scenario, and the lighter mark, the ones in the



(a) $LFM\text{-}2b\text{-}DemoBias_{Sub}$ (b) $LFM\text{-}1b\text{-}DemoBias_{Sub}$

Fig. 4. Performance (NDCG@10) versus unfairness (*RecGap* corresponding to NDCG@10) of different recommendation algorithms on *LFM-2b-DemoBias*$_{Sub}$ and *LFM-1b-DemoBias*$_{Sub}$ datasets. The experiments and results related to *LFM-1b-DemoBias*$_{Sub}$ are explained in Appendix B. For each recommendation algorithm, the darker and lighter marks represent experiments on the Standard and Resampled scenarios, respectively. Linear fit corresponds to the Standard scenario.

Resampled scenario. The dashed line shows the linear fit over the results of Standard, indicating the correlation between NDCG@10 and *RecGap*. As shown, the performance metric and *RecGap* highly correlate, such that the algorithms with higher NDCG also show higher unfairness. In other words, the recommendation algorithm achieve better performances by more strongly improving the majority group, and hence increasing the gap. Our results highlight the importance of studying fairness of recommendation algorithms in parallel and side-by-side with their performances.

For completeness, we also report the results of the same experiments on *LFM-1b-DemoBias*$_{Sub}$ (see Appendix B for details). As shown, the same pattern of correlation can be observed on this dataset. However, since the NDCG@10 results of the algorithms on *LFM-1b-DemoBias*$_{Sub}$ are consistently lower than the corresponding ones in *LFM-2b-DemoBias*$_{Sub}$, the corresponding correlation coefficient is relatively smaller.

As for the results of the compounding factor, namely the metric scores distributions and the corresponding *CompFct*, they are provided in the last two columns of Tables 4–7, respectively. These results provide a complementary view on *un*fairness in RSs by quantifying in what extent the data/population bias is intensified by model bias.

The distribution of male and female users in the population distribution is $B = [0.779, 0.221]$. Based on this population distribution, if the value of male in a metric gain distribution becomes higher than 77.9%, the corresponding model has compounded the population bias toward the male group, which is accordingly reflected in *CompFct*. Highest absolute values of *CompFct* across each algorithm, metric, and dataset are shown in bold. As expected, the majority of algorithms (with the exception of POP) compound the existing data bias in the final results. Similar to *RecGap*, the debiasing approach generally decreases the absolute values of *CompFct* (except in POP and ItemKNN), while such decreases are marginal. These results highlight how the existence of unfairness in RS models amplify the underlying biases in data, and motivate future work for addressing this issue.

Extending the previous description of results, we analyze the differences between the recommendation algorithms in terms of the *RecGap* measure, and provide possible explanations for these differences. POP expectedly does not perform well because it only considers popular items and ignores the subtleties of personalized recommendation algorithms. Our results imply that female users, on average, consume slightly more popular tracks compared to male users, which is inline with previous findings (Schedl, Hauger et al., 2015). One particular observation about POP is that *RecGap* results in the Resampled scenario slightly increases, while decreasing in other algorithms. We argue that the cause of this is due to the increase in popularity of the items listened by female users, resulted from the resampling of female users. Such increase eventually leads to slight improvements of POP in terms of accuracy metrics (NDCG@10 and Recall@10) for the female group, and consequently a marginal increase in *RecGap* toward female.

For the algorithms that create personalized models, the observations show considerably different characteristics. The most unfair algorithms in terms of accuracy are SLIM and ItemKNN (highest *RecGap*). Both of these algorithms rely on an item–item similarity measure, computed from the user–item interaction matrix. This suggests the possible effect of item–item similarity measure on reflecting the preferences of the majority group. We consider further in-depth analyzes of such an effect as a future direction.

Finally, we compare the results of the two investigated matrix factorization approaches, namely ALS and BPR. The two algorithms yield substantially different *RecGap*, where BPR provides the most fair results among the personalized models (though also the second poorest in terms of NDCG and Recall), while ALS's results are much highly unfair. Based on these experiments, we do not observe any direct effect of matrix factorization method on the fairness of recommendation algorithms.

7. Conclusions, limitations, and future work

In this work, we study the effect of population and model/algorithm bias regarding genders in the context of music recommendation. To this end, we first introduce *LFM-2b*, a novel large-scale real-world dataset of music listening records, which comprises

LFM-2b-DemoBias, a subset containing the listening records of users for which demographic information in terms of gender, age, and country of origin is available. Using *LFM-2b-DemoBias*, we explore the group fairness of RS algorithms regarding users' genders, according to the discrepancies in the evaluation measures of algorithms. We study different collaborative filtering algorithms common in the literature, and consider accuracy and beyond-accuracy metrics. In addition, we formulate the compounding factor for RSs, and study in what extent the RS algorithms intensify the underlying biases in data. Furthermore, we exploit a debiasing method applied to data, which aims to mitigate the model bias. In the following, we summarize our findings regarding the considered research questions:

RQ1: *Do recommender algorithms of various categories yield different performance scores (in terms of accuracy and beyond-accuracy metrics) for different user groups with respect to gender? If so, how can these differences be characterized?* Our research outcomes show that most of the collaborative filtering algorithms considered in our study tend to be unfair toward the female group (minority), in terms of NDCG, recall, and coverage metrics, particularly on shorter recommended lists, i. e., with ranking list up to positions 5 and 10. Furthermore, we notice a (reverse) relation between the accuracy-based and fairness metrics: better performing algorithms, such as `SLIM` and `ItemKNN`, show larger degrees of unfairness compared to the less accurate algorithms such as `BPR` and `POP`.

RQ2: *What is the effect of a resampling strategy, commonly used as debiasing method, on the performance and fairness of algorithms?* Overall, the studied debiasing approach marginally improves the fairness of recommendation results (by reducing *RecGap*) across the various RS algorithms. Applying debiasing only slightly deteriorates the performance of RS algorithms in terms of accuracy and beyond-accuracy metrics (no significant changes are observed in the majority of cases), which indicates the benefit of using the debiasing method.

RQ3: *Do RS algorithms compound data bias? If so, how can this be characterized?* The algorithms that lead to unfair results in RSs also compound the bias in data. In such cases, the distributions of final gains of the algorithms are even more biased than the distribution of genders in data. We observe that this compounding of imbalances is particularly high for `ALS` and `ItemKNN` on the accuracy-based metrics, and for `MultiVAE` on coverage.

Our findings translate to reusable insights for the music information retrieval (MIR) and music recommender systems (MRS) communities. First, based on our experimental results, we believe that developing, refining, and adopting debiasing strategies is urgently needed for MIR and MRS tasks that involve personalization, to account not only for differences in performance related to gender – which is shown in this paper – but also for other user- and data-specific biases (e. g., according to age, experience, or popularity). While there already exist several debiasing approaches, to the best of our knowledge, the vast majority of them still lack validation and adoption in the MIR and MRS communities. Second, when conducting evaluation experiments to assess performances of (newly proposed) algorithms, which is still the focus of the (technically oriented) MIR and MRS communities, results should be reported for different user groups. While this is commonly done in MIR and MRS research that explicitly aims at comparing different group-specific characteristics, it is often neglected in more technically driven work. This typically requires adapting the experimental setup, making it vital for MIR and MRS researchers to internalize the corresponding awareness of gender (and other) biases during experimental design. Similarly, user studies conducted in MIR and MRS should critically reflect on potential effects of unequal gender distribution among participants. Even if studies presented in MIR and MRS literature mention the gender distribution of subjects, often the results are not discussed under this perspective, in particular the extent to which differences in results are possibly caused by gender-related aspects. Finally, from a user perspective, methods to increase transparency of recommendation algorithms and explainability of recommendations should be more widely adopted, not only to improve trust in the MRS (which is the motivation commonly mentioned by system providers), but also to raise awareness of potential fairness problems, e. g., through explanations of the form "You are being recommended song X because other female listeners like it".

As for limitations of the current study, we acknowledge that the assumption of gender as a binary construct is an over-simplification, and does not reflect the complexity of fairness and bias regarding gender. This decision, however, enables us to take practical steps. In addition, we have centered our study on the evaluation of gender-related bias, while other demographics such as age and country of origin, are not considered. Furthermore, the used datasets contain logs of user interactions with the online platform *Last.fm*. As such, they can only capture the listening events of people using the platform. All contained information (demographics and listening records) is self-reported by the users, which may be prone to errors and may not necessarily reflect the truth.

These limitations will be addressed in future work by extending our framework to a non-binary setting, which enables the study of the mentioned cases. In addition, we will consider datasets that originate from other platforms, within and beyond the music domain. Finally, a natural future direction of this work is the study of additional algorithmic debiasing methods. In particular, we will explore how/whether fairness of RS algorithms can be achieved without negatively impacting their average performance.

CRediT authorship contribution statement

Alessandro B. Melchiorre: Methodology, Software, Investigation, Writing - original draft. **Navid Rekabsaz:** Methodology, Conceptualization, Writing - original draft, Writing - review & editing. **Emilia Parada-Cabaleiro:** Conceptualization, Formal analysis, Writing - original draft, Writing - review & editing. **Stefan Brandl:** Data curation, Visualization, Writing - original draft, Writing - review & editing. **Oleg Lesota:** Software, Visualization, Writing - original draft, Writing - review & editing. **Markus Schedl:** Supervision, Funding acquisition, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table A.8
Overall results of accuracy (NDCG and recall) and beyond-accuracy (diversity and coverage) metrics; on the different evaluated models: POP, ItemKNN, BPR, ALS, SLIM, and MultiVAE; considering two settings: Standard (Standard) and Resampled (Resampled); for all users together (female and male). Statistically significant differences between the Standard and Resampled scenarios for each model and metric are indicated with an asterisk (∗) on the highest value between Standard and Resampled. Highest values for each metric are shown in bold.

Model	Scenario	NDCG@5	Recall@5	Diversity@5	Coverage@5
POP	Standard	.051	.048	.968	.000
	Resampled	.051	.050	.994*	.000
ItemKNN	Standard	.332	.316*	.984	**.093***
	Resampled	.323	.306	.986*	.076
BPR	Standard	.142	.131	.984	.081
	Resampled	.136	.127	.985*	.079
ALS	Standard	.275	.256	.978*	.038*
	Resampled	.272	.253	.975	.036
SLIM	Standard	**.405**	**.382**	**.992**	.069*
	Resampled	.401	.378	**.992**	.067
MultiVAE	Standard	.219*	.203*	.983	.030
	Resampled	.209	.193	.984	.031

Table A.9
Overall results of accuracy (NDCG and recall) and beyond-accuracy (diversity and coverage) metrics; on the different evaluated models: POP, ItemKNN, BPR, ALS, SLIM, and MultiVAE; considering two settings: Standard (Standard) and Resampled (Resampled); for all users together (female and male). Statistically significant differences between the Standard and Resampled scenarios for each model and metric are indicated with an asterisk (∗) on the highest value between Standard and Resampled. Highest values for each metric are shown in bold.

Model	Scenario	NDCG@50	Recall@50	Diversity@50	Coverage@50
POP	Standard	.041	.051	.987*	.001
	Resampled	.041	.052	.986	.001
ItemKNN	Standard	.273*	.289*	.986*	**.454***
	Resampled	.263	.275	.987	.387
BPR	Standard	.119	.139	**.992**	.308
	Resampled	.115	.134	**.992**	.298
ALS	Standard	.206	.218	.986*	.129*
	Resampled	.203	.214	.985	.124
SLIM	Standard	**.319**	**.331***	.990	.315*
	Resampled	.313	.322	.990	.296
MultiVAE	Standard	.169	.184	.984	.117
	Resampled	.163	.181	.985	.127

Acknowledgments

This research is supported by Know-Center Graz, through project "Theory-inspired Recommender Systems".

Appendix A. Evaluation results using other thresholds *K* for NDCG, recall, diversity, and coverage

See Tables A.8–A.17.

Table A.10
NDCG@5 results on the users of the male/female (M/F) groups for the Standard and Resampled scenarios. The value of *RecGap* shows the degree of favorable treatment toward (m)ales or (f)emales. Highest values are shown in bold. Statistically significant differences between the results of female and male are shown with † symbol. Score Dist. columns shows the metric gain distributions across males and females. The value of *CompFct* shows the effect of compounding imbalances in data. The population distribution to calculate *CompFct* is $B = [0.779, 0.221]$.

Model	Scenario	NDCG@5				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.051	.051/.052	.001 (f)	77.4/22.6	.000
	Resampled	.051	.050/.055	.006 (f)	76.0/24.0	.001
ItemKNN	Standard	.332	.346/.284	.063 (m) †	81.1/18.9	.005
	Resampled	.323	.337/.274	.063 (m) †	81.2/18.8	.005
BPR	Standard	.142	.145/.129	.016 (m) †	79.8/20.2	.002
	Resampled	.136	.139/.127	.012 (m)	79.4/20.6	.001
ALS	Standard	.275	.288/.230	.058 (m) †	81.5/18.5	**.006**
	Resampled	.272	.284/.228	.056 (m) †	81.4/18.6	**.006**
SLIM	Standard	.405	.421/.347	**.073** (m) †	81.0/19.0	.004
	Resampled	.401	.416/.345	**.071** (m) †	80.9/19.1	.004
MultiVAE	Standard	.219	.225/.195	.031 (m) †	80.3/19.7	.003
	Resampled	.209	.215/.186	.029 (m) †	80.3/19.7	.003

Table A.11
Recall@5 results. Details are identical to Table A.10.

Model	Scenario	Recall@5				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.048	.047/.052	.005 (f)	76.1/23.9	.001
	Resampled	.050	.047/.058	.010 (f) †	74.3/25.7	.005
ItemKNN	Standard	.316	.329/.269	.060 (m) †	81.2/18.8	.005
	Resampled	.306	.319/.261	.058 (m) †	81.2/18.8	.005
BPR	Standard	.131	.135/.120	.014 (m) †	79.8/20.2	.002
	Resampled	.127	.129/.120	.009 (m)	79.1/20.9	.001
ALS	Standard	.256	.268/.214	.053 (m) †	81.5/18.5	**.006**
	Resampled	.253	.265/.213	.052 (m) †	81.4/18.6	**.006**
SLIM	Standard	.382	.397/.328	**.070** (m) †	81.0/19.0	.004
	Resampled	.378	.393/.325	**.068** (m) †	81.0/19.0	.004
MultiVAE	Standard	.203	.209/.181	.029 (m) †	80.3/19.7	.003
	Resampled	.193	.199/.173	.025 (m) †	80.1/19.9	.002

Table A.12
Diversity@5 results. Details are identical to Table A.10.

Model	Scenario	Diversity@5				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.968	.968/.968	.000 (m)	77.9/22.1	.000
	Resampled	.994	.994/.993	.001 (m) †	77.9/22.1	.000
ItemKNN	Standard	.984	.983/.985	.001 (f)	77.9/22.1	.000
	Resampled	.986	.985/.987	.001 (f) †	77.9/22.1	.000
BPR	Standard	.984	.984/.985	.001 (f)	77.9/22.1	.000
	Resampled	.985	.985/.987	.002 (f) †	77.9/22.1	.000
ALS	Standard	.978	.977/.979	.001 (f)	77.9/22.1	.000
	Resampled	.975	.975/.977	.002 (f) †	77.9/22.1	.000
SLIM	Standard	.992	.992/.992	.000 (f)	77.9/22.1	.000
	Resampled	.992	.992/.992	.000 (f)	77.9/22.1	.000
MultiVAE	Standard	.983	.983/.985	**.002** (f) †	77.9/22.1	.000
	Resampled	.984	.983/.986	**.003** (f) †	77.8/22.2	.000

Table A.13
Coverage@5 results. Details are identical to Table A.10.

Model	Scenario	Coverage@5				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.000	.000/.000	.000 (m)	80.4/19.6	.003
	Resampled	.000	.000/.000	.000 (m)	79.9/20.1	.002
ItemKNN	Standard	.093	.079/.029	**.050** (m) †	90.6/9.4	.103
	Resampled	.076	.065/.026	.039 (m) †	89.7/10.3	.086
BPR	Standard	.081	.072/.026	.046 (m) †	90.6/9.4	.103
	Resampled	.079	.071/.027	**.043** (m) †	90.1/9.9	.093
ALS	Standard	.038	.035/.018	.018 (m) †	87.5/12.5	.051
	Resampled	.036	.033/.018	.015 (m) †	86.8/13.2	.042
SLIM	Standard	.069	.060/.025	.034 (m) †	89.2/10.8	.076
	Resampled	.067	.058/.025	.033 (m) †	89.0/11.0	.073
MultiVAE	Standard	.030	.027/.009	.017 (m) †	90.9/9.1	**.110**
	Resampled	.031	.028/.009	.019 (m) †	91.7/8.3	**.128**

Table A.14
NDCG@50 results on the users of the male/female (M/F) groups for the Standard and Resampled scenarios. The value of *RecGap* shows the degree of favorable treatment toward (m)ales or (f)emales. Highest values are shown in bold. Statistically significant differences between the results of female and male are shown with † symbol. Score Dist. columns shows the metric gain distributions across males and females. The value of *CompFct* shows the effect of compounding imbalances in data. The population distribution to calculate *CompFct* is $B = [0.779, 0.221]$.

Model	Scenario	NDCG@50				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.041	.040/.047	.008 (f)	74.7/25.3	**.004**
	Resampled	.041	.039/.049	.010 (f) †	73.7/26.3	**.007**
ItemKNN	Standard	.273	.279/.252	.027 (m) †	79.6/20.4	.001
	Resampled	.263	.269/.241	.027 (m) †	79.7/20.3	.001
BPR	Standard	.119	.118/.122	.004 (f) †	77.3/22.7	.000
	Resampled	.115	.114/.121	.007 (f) †	76.8/23.2	.000
ALS	Standard	.206	.209/.192	.018 (m) †	79.4/20.6	.001
	Resampled	.203	.206/.192	.014 (m) †	79.1/20.9	.001
SLIM	Standard	.319	.325/.297	.028 (m) †	79.4/20.6	.001
	Resampled	.313	.318/.293	.026 (m) †	79.3/20.7	.001
MultiVAE	Standard	.169	.169/.167	.002 (m)	78.1/21.9	.000
	Resampled	.163	.164/.161	.003 (m)	78.2/21.8	.000

Table A.15
Recall@50 results. Details are identical to Table A.14.

Model	Scenario	Recall@50				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.051	.047/.065	.018 (f) †	71.8/28.2	**.014**
	Resampled	.052	.047/.068	.021 (f) †	71.0/29.0	**.018**
ItemKNN	Standard	.289	.290/.285	.005 (m)	78.2/21.8	.000
	Resampled	.275	.277/.269	.008 (m)	78.4/21.6	.000
BPR	Standard	.139	.134/.155	**.021** (f) †	75.3/24.7	.003
	Resampled	.134	.129/.154	**.025** (f) †	74.7/25.3	.004
ALS	Standard	.218	.218/.220	.002 (f)	77.7/22.3	.000
	Resampled	.214	.212/.221	.009 (f)	77.2/22.8	.000
SLIM	Standard	.331	.332/.330	.002 (m)	78.0/22.0	.000
	Resampled	.322	.322/.321	.000 (m)	77.9/22.1	.000
MultiVAE	Standard	.184	.181/.198	.017 (f)	76.3/23.7	.001
	Resampled	.181	.177/.195	.018 (f)	76.2/23.8	.001

Table A.16
Diversity@50 results. Details are identical to Table A.14.

Model	Scenario	Diversity@50				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.987	.987/.987	.000 (m)	77.9/22.1	.000
	Resampled	.986	.986/.986	.000 (m)	77.9/22.1	.000
ItemKNN	Standard	.986	.986/.988	.002 (f) †	77.9/22.1	.000
	Resampled	.987	.987/.989	.002 (f) †	77.9/22.1	.000
BPR	Standard	.992	.991/.993	.002 (f) †	77.9/22.1	.000
	Resampled	.992	.991/.993	.002 (f) †	77.9/22.1	.000
ALS	Standard	.986	.985/.988	.002 (f) †	77.8/22.2	.000
	Resampled	.985	.984/.988	**.003** (f) †	77.8/22.2	.000
SLIM	Standard	.990	.990/.991	.002 (f) †	77.9/22.1	.000
	Resampled	.990	.990/.992	.002 (f) †	77.9/22.1	.000
MultiVAE	Standard	.984	.983/.987	**.003** (f) †	77.8/22.2	.000
	Resampled	.985	.984/.987	**.003** (f) †	77.8/22.2	.000

Table A.17
Coverage@50 results. Details are identical to Table A.14.

Model	Scenario	Coverage@50				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.001	.001/.001	.000 (m)	79.6/20.4	.001
	Resampled	.001	.001/.001	.000 (m)	80.0/20.0	.002
ItemKNN	Standard	.454	.406/.165	**.241** (m) †	89.6/10.4	**.084**
	Resampled	.387	.346/.150	**.196** (m) †	89.1/10.9	.074
BPR	Standard	.308	.292/.148	.145 (m) †	87.5/12.5	.051
	Resampled	.298	.282/.149	.133 (m) †	87.0/13.0	.045
ALS	Standard	.129	.123/.073	.050 (m) †	85.6/14.4	.031
	Resampled	.124	.117/.072	.045 (m) †	85.1/14.9	.026
SLIM	Standard	.315	.286/.143	.142 (m) †	87.5/12.5	.051
	Resampled	.296	.268/.139	.130 (m) †	87.2/12.8	.047
MultiVAE	Standard	.117	.109/.047	.062 (m) †	89.1/10.9	.074
	Resampled	.127	.120/.046	.074 (m) †	90.2/9.8	**.094**

Table B.18
Statistics of the *LFM-1b-DemoBias*$_{Sub}$ dataset. Number of Users, Tracks, Artists, and LEs are reported across F(emale) and M(ale) separately and also together (All). Mean and standard deviation (indicated after ±) of the interactions of users with tracks, artists, and listening events are indicated in the last three columns, respectively.

Gender	Users	Tracks	Artists	LEs	Tracks/User	Artists/User	LEs/User
All	50,578	99,792	43,764	27,927,919	81 ± 82	71 ± 68	552 ± 809
F	14,143	82,637	38,610	7,093,595	69 ± 64	60 ± 55	502 ± 650
M	36,435	99,640	43,717	20,834,324	86 ± 87	75 ± 72	572 ± 862

Table B.19
Overall LFM-1b results of accuracy (NDCG and recall) and beyond-accuracy (diversity and coverage) metrics; on the different evaluated models: POP, ItemKNN, BPR, ALS, SLIM, and MultiVAE; considering two scenarios: Standard (Standard) and Resampled (Resampled); for all users together (female and male). Statistically significant differences between the Standard and Resampled scenarios for each model and metric are indicated with an asterisk (∗) on the highest value between Standard and Resampled. Highest values for each metric are shown in bold.

Model	Scenario	NDCG@10	Recall@10	Diversity@10	Coverage@10
POP	Standard	.030	.029	**1.00**	.000
	Resampled	.030	.029	.999	.000
ItemKNN	Standard	.288	.269	.979*	**.241***
	Resampled	.285	.265	.981	.197
BPR	Standard	.113*	.105*	.983*	.144*
	Resampled	.107	.101	.985	.155
ALS	Standard	.206	.184	.981*	.054*
	Resampled	.207	.184	.980	.053
SLIM	Standard	**.327**	**.302**	.986	.137*
	Resampled	.325	.299	.986	.133
MultiVAE	Standard	.171	.157	.984*	.077*
	Resampled	.167	.154	.983	.072

Table B.20
LFM-1b NDCG@10 results on the users of the male/female (M/F) groups for the Standard and Resampled scenarios. The value of *RecGap* shows the degree of favorable treatment toward (m)ales or (f)emales. Highest values are shown in bold. Statistically significant differences between the results of female and male are shown with † symbol. Score Dist. columns shows the metric gain distributions across males and females. The value of *CompFct* shows the effect of compounding imbalances in data. The population distribution to calculate *CompFct* is $B = [0.720, 0.280]$.

Model	Scenario	NDCG@10				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.030	.030/.030	.001 (f)	71.7/28.3	.000
	Resampled	.030	.028/.032	.004 (f) †	69.4/30.6	**.002**
ItemKNN	Standard	.288	.298/.264	.034 (m) †	74.4/25.6	**.002**
	Resampled	.285	.294/.261	**.034** (m) †	74.4/25.6	**.002**
BPR	Standard	.113	.114/.110	.003 (m)	72.6/27.4	.000
	Resampled	.107	.107/.105	.002 (m)	72.5/27.5	.000
ALS	Standard	.206	.213/.189	.024 (m) †	74.3/25.7	**.002**
	Resampled	.207	.213/.192	.021 (m) †	74.0/26.0	.001
SLIM	Standard	.327	.337/.302	**.035** (m) †	74.2/25.8	**.002**
	Resampled	.325	.334/.300	**.034** (m) †	74.1/25.9	**.002**
MultiVAE	Standard	.171	.173/.164	.010 (m) †	73.2/26.8	.000
	Resampled	.167	.170/.162	.008 (m) †	73.0/27.0	.000

Appendix B. LFM-1b evaluation results for NDCG, recall, diversity, and coverage

In this appendix we report the results of the experiments on the *LFM-1b* dataset. We follow a similar data processing procedures as detailed in Section 5.1. We only consider user-track interactions with a playcount (PC) > 1, only tracks listened to by at least 5 different users and users that listened to at least 5 different tracks. However, we do not impose any constraints on the recency of the LEs. Furthermore, we also randomly sample 100,000 tracks. We refer to this final subset as *LFM-1b-DemoBias*$_{Sub}$. The statistics for the datasets are reported in Table B.18.

We follow the same exact experiment procedures as in Sections 5.2–5.7 (see Table B.19). The results are reported for K = 10 in Tables B.20–B.23.

Table B.21
LFM-1b Recall@10 results. Details are identical to Table B.20.

Model	Scenario	Recall@10				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.029	.029/.031	.002 (f)	70.8/29.2	.001
	Resampled	.029	.027/.034	.006 (f) †	67.7/32.3	**.006**
ItemKNN	Standard	.269	.276/.248	.028 (m) †	74.1/25.9	**.002**
	Resampled	.265	.273/.245	**.028** (m) †	74.1/25.9	.002
BPR	Standard	.105	.105/.105	.000 (f)	72.0/28.0	.000
	Resampled	.101	.101/.101	.000 (f)	72.0/28.0	.000
ALS	Standard	.184	.188/.171	.018 (m) †	74.0/26.0	.001
	Resampled	.184	.188/.174	.014 (m) †	73.6/26.4	.001
SLIM	Standard	.302	.310/.281	**.029** (m) †	74.0/26.0	.001
	Resampled	.299	.307/.279	**.028** (m) †	73.9/26.1	.001
MultiVAE	Standard	.157	.159/.152	.007 (m) †	73.0/27.0	.000
	Resampled	.154	.155/.150	.005 (m) †	72.7/27.3	.000

Table B.22
LFM-1b Diversity@10 results. Details are identical to Table B.20.

Model	Scenario	Diversity@10				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	1.00	1.00/1.00	.000 (f)	72.0/28.0	.000
	Resampled	.999	.999/.999	.000 (f)	72.0/28.0	.000
ItemKNN	Standard	.979	.979/.978	**.001** (m)	72.1/27.9	.000
	Resampled	.981	.981/.980	.000 (m)	72.0/28.0	.000
BPR	Standard	.983	.983/.983	**.001** (f)	72.0/28.0	.000
	Resampled	.985	.984/.985	.001 (f) †	72.0/28.0	.000
ALS	Standard	.981	.980/.981	**.001** (f) †	72.0/28.0	.000
	Resampled	.980	.979/.981	**.002** (f) †	72.0/28.0	.000
SLIM	Standard	.986	.986/.986	**.001** (m) †	72.0/28.0	.000
	Resampled	.986	.986/.986	.000 (m)	72.0/28.0	.000
MultiVAE	Standard	.984	.983/.984	**.001** (f) †	72.0/28.0	.000
	Resampled	.983	.983/.984	**.002** (f) †	72.0/28.0	.000

Table B.23
LFM-1b Coverage@10 results. Details are identical to Table B.20.

Model	Scenario	Coverage@10				
		All	M/F	*RecGap*	Score Dist.	*CompFct*
POP	Standard	.000	.000/.000	.000 (m)	73.2/26.8	.000
	Resampled	.000	.000/.000	.000 (m)	73.5/26.5	.001
ItemKNN	Standard	.241	.206/.103	**.103** (m) †	83.8/16.2	**.063**
	Resampled	.197	.170/.091	**.079** (m) †	82.8/17.2	.051
BPR	Standard	.144	.133/.080	.054 (m) †	81.2/18.8	.035
	Resampled	.155	.142/.087	.055 (m) †	80.8/19.2	.032
ALS	Standard	.054	.051/.040	.011 (m) †	76.8/23.2	.009
	Resampled	.053	.050/.040	.010 (m) †	76.4/23.6	.007
SLIM	Standard	.137	.121/.072	.048 (m) †	81.1/18.9	.035
	Resampled	.133	.117/.071	.046 (m) †	81.0/19.0	.034
MultiVAE	Standard	.077	.069/.036	.033 (m) †	83.2/16.8	.056
	Resampled	.072	.065/.034	.031 (m) †	83.1/16.9	**.055**

References

Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). Controlling popularity bias in learning-to-rank recommendation. In *RecSys '17*, *Proceedings of the eleventh ACM conference on recommender systems* (pp. 42–46). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3109859.3109912.

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2019). The unfairness of popularity bias in recommendation. In R. Burke, H. Abdollahpouri, E. C. Malthouse, K. P. Thai, & Y. Zhang (Eds.), *CEUR workshop proceedings*: *vol. 2440*, *Proceedings of the workshop on recommendation in multi-stakeholder environments co-located with the 13th conference on recommender systems (RecSys)*. Copenhagen, Denmark: ACM, URL http://ceur-ws.org/Vol-2440/paper4.pdf.

Adomavicius, G., Mobasher, B., Ricci, F., & Tuzhilin, A. (2011). Context-aware recommender systems. *AI Magazine*, *32*(3), 67–80. http://dx.doi.org/10.1609/aimag.v32i3.2364.

Aggarwal, C. C. (2016a). Ensemble-based and hybrid recommender systems. In *Recommender systems* (pp. 199–224). New York, NY, USA: Springer, http://dx.doi.org/10.1007/978-3-319-29659-3_6.

Aggarwal, C. C. (2016b). Neighborhood-based collaborative filtering. In *Recommender systems* (pp. 29–70). New York, NY, USA: Springer, http://dx.doi.org/10.1007/978-3-319-29659-3_2.

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61. http://dx.doi.org/10.1145/3209581.

Bauer, C., & Schedl, M. (2019). Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PLOS ONE*, *14*(6), 1–36. http://dx.doi.org/10.1371/journal.pone.0217389.

Beigi, G., Mosallanezhad, A., Guo, R., Alvari, H., Nou, A., & Liu, H. (2020). *WSDM '20*, *Privacy-aware recommendation with private-attribute protection using adversarial learning* (pp. 34–42). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3336191.3371832.

Beliakov, G., Calvo, T., & James, S. (2011). Aggregation of preferences in recommender systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 705–734). New York, NY, USA: Springer, http://dx.doi.org/10.1007/978-0-387-85820-3_22.

Bell, R. M., & Koren, Y. (2007). Improved neighborhood-based collaborative filtering. In *KDD cup and workshop at the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 7–14). Citeseer.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., et al. (2019). Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th international conference on knowledge discovery and data mining (SIGKDD)* (pp. 2212–2220). Anchorage, AK, USA: ACM.

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st SIGIR international conference on research and development in information retrieval* (pp. 405–414). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3209978.3210063.

Billsus, D., Pazzani, M. J., et al. (1998). Learning collaborative information filters.. In *Icml, Vol. 98* (pp. 46–54).

Borges, R., & Stefanidis, K. (2019). Enhancing long term fairness in recommendations with variational autoencoders. In *Proceedings of the 11th international conference on management of digital ecosystems* (pp. 95–102). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3297662.3365798.

Bose, A., & Hamilton, W. (2019). Compositional fairness constraints for graph embeddings. In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of machine learning research*: *vol. 97*, *Proceedings of the 36th international conference on machine learning* (pp. 715–724). Long Beach, California, USA: PMLR, URL http://proceedings.mlr.press/v97/bose19a.html.

Brost, B., Mehrotra, R., & Jehan, T. (2019). The music streaming sessions dataset. In *WWW '19*, *The world wide web conference* (pp. 2594–2600). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3308558.3313641.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, *12*(4), 331–370.

Burke, R. (2017). Multisided fairness for recommendation. CoRR abs/1707.00093. arXiv:1707.00093. URL http://arxiv.org/abs/1707.00093.

Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, *21*(6), 1487–1524. http://dx.doi.org/10.3233/IDA-163209.

Celma, O. (2010). *Music recommendation and discovery - the long tail, long fail, and long play in the digital music space*. Berlin, Germany: Springer, http://dx.doi.org/10.1007/978-3-642-13287-2.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2020). Bias and debias in recommender system: A survey and future directions. CoRR abs/2010.03240. arXiv:2010.03240. URL https://arxiv.org/abs/2010.03240.

Dacrema, M. F., Boglio, S., Cremonesi, P., & Jannach, D. (2019). A troubling analysis of reproducibility and progress in recommender systems research. arxiv preprint arXiv:1911.07698.

Darlington, R. B., & Hayes, A. F. (2000). Combining independent p values: Extensions of the stouffer and binomial methods.. *Psychological Methods*, *5*(4), 496.

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, *2015*(1), 92–112. http://dx.doi.org/10.1515/popets-2015-0007.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., et al. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the conference on fairness, accountability, and transparency, Atlanta, GA, USA* (pp. 120–128). http://dx.doi.org/10.1145/3287560.3287572.

Deldjoo, Y., Schedl, M., Cremonesi, P., & Pasi, G. (2020). Recommender systems leveraging multimedia content. *ACM Computing Surveys*, *53*(5), 106:1–106:38. http://dx.doi.org/10.1145/3407190.

Dror, G., Koenigstein, N., Koren, Y., & Weimer, M. (2011). The yahoo! music dataset and KDD-cup'11. In *KDDCUP'11*, *Proceedings of the 2011 international conference on KDD Cup 2011 - Volume 18* (pp. 3–18). JMLR.org.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, Cambridge, MA, USA (pp. 214–226). http://dx.doi.org/10.1145/2090236.2090255.

Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., et al. (2018). All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In S. A. Friedler, C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency, Vol. 81* (pp. 172–186). New York, NY, USA: PMLR, URL http://proceedings.mlr.press/v81/ekstrand18b.html.

Ekstrand, M. D., Tian, M., Kazi, M. R. I., Mehrpouyan, H., & Kluver, D. (2018). Exploring author gender in book rating and recommendation. In S. Pera, M. D. Ekstrand, X. Amatriain, & J. O'Donovan (Eds.), *Proceedings of the 12th conference on recommender systems (RecSys)* (pp. 242–250). Vancouver, BC, Canada: ACM, http://dx.doi.org/10.1145/3240323.3240373.

Epps-Darling, A., Bouyer, R. T., & Cramer, H. (2020). Artist gender representation in music streaming. In *Proceedings of the international society for music information retrieval conference* (pp. 248–254). Online event: ISMIR.

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *Proceedings of the 25th international conference on knowledge discovery & data mining (SIGKDD)* (pp. 2221–2231). Anchorage, AK, USA: ACM, http://dx.doi.org/10.1145/3292500.3330691.

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Proceedings of the advances in neural information processing systems (NIPS), Vol. 29* (pp. 3315–3323). Barcelona, Spain: Curran Associates, Inc., http://dx.doi.org/10.5555/3157382.3157469, URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Hauger, D., Schedl, M., Košir, A., & Tkalčič, M. (2013). The million musical tweet dataset: what we can learn from microblogs. In *Proceedings of the international society for music information retrieval conference*, Curitiva, Brazil (pp. 189–194).

Hellman, D. (2018). Indirect discrimination and the duty to avoid compounding injustice. In *Foundations of indirect discrimination law* (pp. 2017–2053). Hart Publishing Company, URL https://ssrn.com/abstract=3033864.

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE international conference on data mining, 2008. ICDM'08* (pp. 263–272). IEEE.

Huang, Q., Jansen, A., Zhang, L., Ellis, D. P. W., Saurous, R. A., & Anderson, J. R. (2020). Large-scale weakly-supervised content embeddings for music recommendation and tagging. In *Proceedings of the international conference on acoustics, speech and signal processing (ICASSP)* (pp. 8364–8368). Barcelona, Spain: IEEE, http://dx.doi.org/10.1109/ICASSP40776.2020.9053240.

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, *74*(2), 171–193. http://dx.doi.org/10.1037/amp0000307.

Kamishima, T., & Akaho, S. (2017). Considerations on recommendation independence for a find-good-items task. In *Proceedings of the FATREC workshop on responsible recommendation proceedings*. http://dx.doi.org/10.18122/B2871W.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Enhancement of the neutrality in recommendation. In *Proceedings of the workshop on human decision making in recommender systems* (pp. 8–14). URL http://ceur-ws.org/Vol-893/paper2.pdf.

Koren, Y., & Bell, R. M. (2015). Advances in collaborative filtering. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 77–118). New York, NY, USA: Springer, http://dx.doi.org/10.1007/978-1-4899-7637-6_3.

Kowald, D., Schedl, M., & Lex, E. (2020). The unfairness of popularity bias in music recommendation: A reproducibility study. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, & F. Martins (Eds.), *Lecture notes in computer science*: *vol. 12036*, *Proceedings of the 42nd European conference on advances in information retrieval research (ECIR)* (pp. 35–42). Lisbon, Portugal: Springer, http://dx.doi.org/10.1007/978-3-030-45442-5_5.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems, Vol. 30* (pp. 4066–4076). Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, *65*(7), 2966–2981. http://dx.doi.org/10.1287/mnsc.2018.3093.

Liang, D., Krishnan, R. G., Hoffman, M. D., & Jebara, T. (2018). Variational autoencoders for collaborative filtering. In P. Champin, F. L. Gandon, M. Lalmas, & P. G. Ipeirotis (Eds.), *Proceedings of the 2018 world wide web conference on world wide wb, WWW 2018, Lyon, France* (pp. 689–698). ACM, http://dx.doi.org/10.1145/3178876.3186150.

Lin, K., Sonboli, N., Mobasher, B., & Burke, R. (2019a). Crank up the volume: preference bias amplification in collaborative recommendation. In *Proceedings of CEUR workshop*.

Lin, K., Sonboli, N., Mobasher, B., & Burke, R. (2019b). Crank up the volume: preference bias amplification in collaborative recommendation. CoRR abs/1909.06362. arXiv:1909.06362. URL http://arxiv.org/abs/1909.06362.

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 73–105). New York, NY, USA: Springer, http://dx.doi.org/10.1007/978-0-387-85820-3_3.

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020a). Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 2145–2148).

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020b). Feedback loop and bias amplification in recommender systems. In *CIKM '20*, *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 2145–2148). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3340531.3412152.

Mansoury, M., Mobasher, B., Burke, R., & Pechenizkiy, M. (2019). Bias disparity in collaborative recommendation: Algorithmic evaluation and comparison. In *Proceedings of CEUR workshop*.

Marlin, B. (2004). *Collaborative filtering: A machine learning perspective*. Toronto: University of Toronto.

McFee, B., & Lanckriet, G. R. (2012). Hypergraph models of playlist dialects.. In *Proceedings of the international society for music information retrieval conference* (pp. 343–348). Porto, Portugal: ISMIR.

McKnight, P. E., & Najab, J. (2010). Mann-whitney u test. *The Corsini Encyclopedia of Psychology*, 1.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. CoRR abs/1908.09635. arXiv:1908.09635. URL http://arxiv.org/abs/1908.09635.

Melchiorre, A. B., Zangerle, E., & Schedl, M. (2020). Personality bias of music recommendation algorithms. In R. L. T. Santos, L. B. Marinho, E. M. Daly, L. Chen, K. Falk, N. Koenigstein, & E. S. de Moura (Eds.), *Proceedings of the 14th conference on recommender systems (RecSys)* (pp. 533–538). Virtual Event, Brazil: ACM, http://dx.doi.org/10.1145/3383313.3412223.

Meng, Z., McCreadie, R., Macdonald, C., & Ounis, I. (2020). Exploring data splitting strategies for the evaluation of recommendation models. In *RecSys '20*, *Proceedings of the fourteenth ACM conference on recommender systems* (pp. 681–686). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3383313.3418479.

Mosteller, F., Bush, R. R., & Green, B. F. (1954). *Selected quantitative techniques*. Addison-Wesley.

Ning, X., & Karypis, G. (2011). SLIM: Sparse linear methods for top-n recommender systems. In D. J. Cook, J. Pei, W. Wang, O. R. Zaïane, & X. Wu (Eds.), *Proceedings of the 11th IEEE international conference on data mining, ICDM 2011, Vancouver, BC, Canada* (pp. 497–506). IEEE Computer Society, http://dx.doi.org/10.1109/ICDM.2011.134.

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 1–13. http://dx.doi.org/10.3389/fdata.2019.00013.

Oramas, S., Nieto, O., Sordo, M., & Serra, X. (2017). A deep multimodal approach for cold-start music recommendation. In B. Hidasi, A. Karatzoglou, O. S. Shalom, S. Dieleman, B. Shapira, & D. Tikk (Eds.), *Proceedings of the 2nd workshop on deep learning for recommender systems (DLRS@RecSys)* (pp. 32–37). Como, Italy: ACM, http://dx.doi.org/10.1145/3125486.3125492.

Patro, G. K., Biswas, A., Ganguly, N., Gummadi, K. P., & Chakraborty, A. (2020). Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of the web conference* (pp. 1194–1204). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3366423.3380196.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th international conference on knowledge discovery and data mining (SIGKDD)* (pp. 560–568). Las Vegas, NV, USA: ACM, http://dx.doi.org/10.1145/1401890.1401959.

Pichl, M., Zangerle, E., & Specht, G. (2015). Towards a context-aware music recommendation approach: What is hidden in the playlist name?. In *2015 IEEE international conference on data mining workshop (ICDMW)* (pp. 1360–1365). IEEE, http://dx.doi.org/10.1109/ICDMW.2015.145.

Poddar, A., Zangerle, E., & Yang, Y. (2018). nowplaying-RS: a new benchmark dataset for building context-aware music recommender systems. In *Proceedings of the 15th sound and music computing conference*.

Rekabsaz, N., & Schedl, M. (2020). Do neural ranking models intensify gender bias? In *Proceedings of the 43rd SIGIR international conference on research and development in information retrieval* (pp. 2065–2068). Virtual Event, China: ACM, http://dx.doi.org/10.1145/3397271.3401280.

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. arxiv preprint arXiv:1205.2618.

Sachdeva, N., Manco, G., Ritacco, E., & Pudi, V. (2019). Sequential variational autoencoders for collaborative filtering. In *WSDM '19*, *Proceedings of the 12th ACM international conference on web search and data mining* (pp. 600–608). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3289600.3291007.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *WWW '01*, *Proceedings of the 10th international conference on world wide web* (pp. 285–295). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/371920.372071.

Schedl, M. (2013). Leveraging microblogs for spatiotemporal music information retrieval. In *Proceedings of the European conference on information retrieval* (pp. 796–799). Springer.

Schedl, M. (2016). The LFM-1b dataset for music retrieval and recommendation. In *ICMR '16*, *Proceedings of the 2016 ACM on international conference on multimedia retrieval* (pp. 103–110). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2911996.2912004.

Schedl, M. (2017). Investigating country-specific music preferences and music recommendation algorithms with the LFM-1b dataset. *International Journal of Multimedia Information Retrieval*, *6*(1), 71–84. http://dx.doi.org/10.1007/s13735-017-0118-y.

Schedl, M. (2019). Deep learning in music recommendation systems. *Frontiers in Applied Mathematics and Statistics*, *5*, 44. http://dx.doi.org/10.3389/fams.2019.00044, URL https://www.frontiersin.org/article/10.3389/fams.2019.00044.

Schedl, M., Hauger, D., Farrahi, K., & Tkalcic, M. (2015). On the influence of user characteristics on music recommendation algorithms. In A. Hanbury, G. Kazai, A. Rauber, & N. Fuhr (Eds.), *Lecture notes in computer science: vol. 9022*, *Proceedings of the 37th European conference on advances in information retrieval research (ECIR)*, Vienna, Austria (pp. 339–345). http://dx.doi.org/10.1007/978-3-319-16354-3_37.

Schedl, M., Knees, P., McFee, B., Bogdanov, D., & Kaminskas, M. (2015). Music recommender systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 453–492). New York, NY, USA: Springer, http://dx.doi.org/10.1007/978-1-4899-7637-6_13.

Shakespeare, D., Porcaro, L., Gómez, E., & Castillo, C. (2020). Exploring artist gender bias in music recommendation. In *2nd workshop on the impact of recommender systems (ImpactRS20), Co-located at RecSys2020*.

Steck, H. (2018). Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 154–162).

Steck, H. (2019). Embarrassingly shallow autoencoders for sparse data. In *WWW '19*, *Proceedings of the the world wide web conference* (pp. 3251–3257). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3308558.3313710.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). The american soldier: Adjustment during army life. *Studies in Social Psychology in World War Ii*.

Sun, Z., Yu, D., Fang, H., Yang, J., Qu, X., Zhang, J., et al. (2020). Are we evaluating rigorously? Benchmarking recommendation for reproducible evaluation and fair comparison. In *RecSys '20*, *Fourteenth ACM conference on recommender systems* (pp. 23–32). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3383313.3412489.

van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Proceedings of the 27th annual conference on advance in neural information processing systems (NIPS)*, Lake Tahoe, NV, USA (pp. 2643–2651). URL http://papers.nips.cc/paper/5004-deep-content-based-music-recommendation.

Vigliensoni, G., & Fujinaga, I. (2017). The music listening histories dataset.. In *Proceedings of the international society for music information retrieval conference* (pp. 96–102). Suzhou, China: ISMIR.

Watson, J. (2020). Programming inequality: Gender representation on Canadian country radio (2005–2019). In *Proceedings of the international society for music information retrieval conference* (pp. 392–399). Online event: ISMIR.

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, *18*(5), 1368–1373.

Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. In *Advances in neural information processing systems* (pp. 2921–2930).

Zamani, H., Schedl, M., Lamere, P., & Chen, C.-W. (2019). An analysis of approaches taken in the ACM RecSys challenge 2018 for automatic music playlist continuation. *ACM Transactions on Intelligent Systems and Technology*, *10*(5), http://dx.doi.org/10.1145/3344257.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). FA*IR: A fair top-k ranking algorithm. In *Proceedings of the conference on information and knowledge management* (pp. 1569–1578). Singapore, Singapore: ACM, http://dx.doi.org/10.1145/3132847.3132938.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In S. Dasgupta, & D. McAllester (Eds.), *Proceedings of the 30th International conference on machine learning, Vol. 28* (pp. 325–333). Atlanta, GA, USA: PMLR, URL http://proceedings.mlr.press/v28/zemel13.html.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, *52*(1), 1–38.