Exploring Cross-group Discrepancies in Calibrated Popularity for Accuracy/Fairness Trade-off Optimization*

OLEG LESOTA, Johannes Kepler University Linz and Linz Institute of Technology, Austria
STEFAN BRANDL, Johannes Kepler University Linz, Austria
MATTHIAS WENZEL, Johannes Kepler University Linz, Austria
ALESSANDRO B. MELCHIORRE, Johannes Kepler University Linz and Linz Institute of Technology, Austria
ELISABETH LEX, Graz University of Technology, Austria
NAVID REKABSAZ, Johannes Kepler University Linz and Linz Institute of Technology, Austria
MARKUS SCHEDL†, Johannes Kepler University Linz and Linz Institute of Technology, Austria

Popularity bias is an important issue in recommender systems, as it affects end-users, content creators, and content provider platforms alike. It can cause users to miss out on less popular items that would fit their preference, prevent new content creators from finding their audience, and force providers to pay higher royalties for serving expensive popular content. Over the past years, various approaches to mitigate popularity bias in recommender systems have been proposed. Among them, post-processing methods are widely accepted due to their versatility and ease of implementation. While previous studies have investigated the effects of different post-processing techniques on accuracy and fairness of recommendations, the influence of different algorithms on different user groups have not received much attention in this context. Addressing this research gap, we study the effect of a recent mitigation strategy, Calibrated Popularity, in conjunction with a selection of state-of-the-art recommender algorithms including BPR, ItemKNN, LightGCN, MultiVAE, and NeuMF. We show that these algorithms demonstrate different characteristics in terms of the trade-off between accuracy and fairness, both within and between various user groups defined by gender and inclination towards consumption of mainstream items. Finally, we demonstrate how these discrepancies can be exploited to achieve a more effective trade-off between utility and fairness of recommender systems.

1 INTRODUCTION

Recommender systems (RSs) are ubiquitous decision support tools, assisting all kinds of users in their personal and business tasks. They help connect content creators and consumers on streaming platforms, suggest products in online stores, and even influence whether a person finds a fitting job. Considering the important role of RSs, it is crucial to monitor societal and statistical biases they often suffer from.

While not all biases are harmful—recommendation results need to be biased in the sense of personalization to match the end user's preferences—data, algorithmic, and presentation biases may lead to unfair behavior of RSs, i. e., the recommender "systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others" [8]. Popularity bias corresponds to the tendency of some RSs to favor popular items over lesser popular items and is considered to be a harmful phenomenon [1, 7, 13, 16]. Popularity bias has been a long-studied problem in the RSs community (e. g., [3, 4, 17, 20, 28]). A RS with popularity bias creates recommendation lists with highly popular items ranked on top, suppressing the exposure of long-tail items. This often leads to low satisfaction of end users (especially those interested in niche items), unfairly limited exposure of new and niche item producers, and higher expenses for content providers, as serving popular items on online platforms in most cases spells higher royalties. To measure and

Presented at the MORS workshop held in conjunction with the 16th ACM Conference on Recommender Systems (RecSys), 2022, in Seattle, USA.
†This is the corresponding author.

capture different aspects of popularity bias, various metrics have been introduced; for an overview, see Abdollahpouri et al. [4]. In this paper we concentrate on the user side of popularity bias.

Over the past years, researchers proposed a multitude of bias mitigation strategies, working on different stages of the recommendation pipeline. One can distinguish pre-, in- and post-processing methods. The first act before the main RS, often applying transformations to the data the RS is trained on, in an attempt to make its output less biased. In-processing methods usually include additional debiasing training objectives, e. g., through adversarial training [9, 21] or regularization [26]. Post-processing methods act on the output of the RS, usually by re-ranking recommended items to satisfy a certain fairness goal. Post-processing bias mitigation techniques have the advantage of versatility, being independent of the main RS and thus able to work in conjunction with almost any algorithm. In addition, a number of calibration-based post-processing techniques have been shown effective in popularity bias mitigation for matrix factorization algorithms.

Previous studies have shown that not only different RSs vary in the degree they are susceptible to popularity bias, but also different user groups suffer from it to various extents. These findings lead us to the following research questions we tackle in this work: ***RQ1:*** *Is the mitigation technique of post-processing equally effective for all algorithms? Do all algorithms show the same character of trade-off between utility and calibration?* ***RQ2:*** *Are all user groups equally affected by the mitigation procedure? Are the optimal mitigation parameters the same for all user groups?* ***RQ3:*** *To which extent can the trade-off between utility and calibration be softened through using specific mitigation parameters for each user group?*

To answer these questions, we pair a recent mitigation strategy, Calibrated Popularity, with an array of recommender algorithms, analyzing the mitigation effectiveness and utility-fairness trade-off for each of them. We also consider the effect of mitigation on different user groups, and through this characterize the scoring strategy of each recommendation algorithm investigated. Finally, we conduct an experiment to evaluate potential gains of mitigation approaches tailored specifically to each user group.

2 RELATED WORK

Many studies formalize fairness of a RS on the user level through calibration of a certain item attribute (such as genre) [24]. Meaning that recommendation is considered fair only when the distribution of the attribute over the recommended list matches its distribution over some reference list (e. g., each user's consumption history or the whole list of items in the collection). A number of studies follow this approach to investigate and enforce fairness of the recommendations [5, 14]. Lesota et al. [17] study differences between popularity distributions of consumed and recommended items for each user, tackling the problem of measuring popularity bias as miscalibration between the two. They express it in terms of the median as well as several statistical moments and similarity measures. In addition, they combine research strands on popularity bias and gender bias by analyzing how female and male listeners are affected by popularity bias. Abdollahpouri et al. [3] show that state-of-the-art movie recommendation algorithms suffer from popularity bias, and introduce the delta-GAP metric to quantify the level of underrepresentation. Kowald et al. [16] reproduce these results for music domain.

Works on bias mitigation often adopt post-processing strategies. Post-processing is a widely used family of bias mitigation techniques. They operate on the output of a RS, re-ranking the items, striving to create a list that satisfies both utility and fairness objectives. A big advantage of post-processing is its flexibility to be used with almost any RS algorithm. A multitude of post-processing techniques for popularity bias mitigation have been proposed over the years. Abdollahpouri et al. [2] propose an algorithm calibrating proportions of head and tail items from the overall popularity distribution. Zehlike et al. [25] propose FA*IR, a method for boosting exposure of items of some protected

category (of low popularity). Abdollahpouri et al. [4] take a more user-oriented approach, called Calibrated Popularity (CP), calibrating three-bin item popularity distributions between user consumption history and their recommendations. Klimashevskaia et al. [15] take a wide perspective on post-processing popularity bias mitigation techniques and analyze them on both platform-wide and user-preference levels. They show that CP is preferable for providing fairness on per-user level.

Usually bias mitigation algorithms allow to adjust the weight distribution between the utility and fairness objectives. da Silva et al. [5] take this idea further, experimenting with learning personal weighting for every user to ensure proper genre diversity in the recommendation lists. To the best of our knowledge, this approach has not been adopted for popularity bias mitigation. In addition, most studies presenting mitigation techniques limit their demonstration to a narrow scope of algorithms and mainly consider the population of users as a whole. We address these limitations in our research, by (1) conducting a set of bias mitigation experiments on state-of-the-art RSs of different architectures, (2) considering two ways of user grouping as well as the whole populations, and (3) carrying out an experiment with learning bias mitigation weights for every group separately. We investigate two datasets: MovieLens-1M (ML-1M) [10] from the movie domain and LFM-2b [23] from the music domain.

3 METHODOLOGY

We base our study on the common assumption that consumers prefer calibrated recommendations [24], i. e., the distribution of item popularity in a user's recommendation list should match that of their interaction history. We investigate the trade-off between popularity calibration and utility of recommendations considering different recommender algorithms, user groups, and settings of the mitigation technique.

Item Popularity. Following common practice [4, 17], we define popularity of each item through the number of interactions with it. We distinguish Popular, Niche, and Mid categories of items. Popular items are represented by items most interacted with and jointly receiving 20% of all user-item interactions. Similarly, Niche items are the least interacted with, receiving 20% of aggregated user-item interactions. The rest of items falls into the category Mid.

User Groups. This work concerns both overall user population as well as specific user groups. We investigate two ways of user grouping: by users' gender[1] and by their inclination towards consumption of popular items. With the latter, similar to [4], we define three user groups: HighPop, MidPop, and LowPop, based on the proportion of popular items in their consumption histories. The groups are defined by sorting users in descending order with respect to the proportion of popular items they consume, and then selecting the top 20%, mid 60%, and bottom 20% of users, respectively.

Metrics. In this work, we consider the trade-off between recommenders' utility expressed through NDCG @10 metric and their proneness to popularity bias. Following previous work, we define the latter on per-user level as Jensen-Shannon Divergence between the popularity distribution of a user's already consumed items and the top 10 recommended items. If H_u and R_u are item popularity (probability) distributions of consumption history and recommendation for user u, respectively, we calculate Jensen-Shannon Divergence as:

$$JSD(H_u, R_u) = \frac{1}{2}\left(\sum_c H_u(c) log_2 \frac{2H_u(c)}{H_u(c) + R_u(c)} + \sum_c R_u(c) log_2 \frac{2R_u(c)}{H_u(c) + R_u(c)}\right) \quad (1)$$

where $H_u(c)$ is the proportion of items of popularity category c in the consumption history of user u. JSD can be seen as symmetrical version of Kullback–Leibler divergence. Note that using log_2 we ensure that the value of JSD is bound

[1]Due to limitations of considered datasets we have to treat gender as a binary concept (male versus female).

between 0 and 1 [19]. We express the degree of exposure to popularity bias of a user group g as JSD_g, defined as the average JSD over all users in g.

Bias Mitigation Technique. Calibrated Popularity is a recent post-processing technique for popularity bias mitigation [4]. It re-ranks a recommendation list L'_u of m items initially recommended to each user, to create a personalized popularity-aware recommendation list L^*_u of n items ($n << m$):

$$L^*_u = \underset{L_u, |L_u|=n}{\arg\max} \; (1-\lambda) \cdot Rel(L_u) - \lambda \cdot JSD(H_u, P(L_u)) \tag{2}$$

where $Rel(L_u)$ is the sum of relevance scores and $P(L_u)$ the item popularity distribution of the n candidate item list. To ensure consistency of the mitigation procedure across different recommenders, we re-scale the relevance scores constituting $Rel(L_u)$ to the interval $[0, 1]$ where needed. The parameter λ allows to prioritize between the utility (first term) and bias mitigation (second term) objectives. Similar to [4] the final recommendation list L^*_u is created through the process of greedy optimization.

Choosing Optimal Mitigation Parameters. To illustrate potential gains of choosing group-specific mitigation parameters for different user groups we introduce a way of selecting an optimal value of parameter λ taking into account both utility and fairness of the recommendation. For ease of notation, we denote $JSF(H_u, R_u) = 1 - JSD(H_u, R_u)$ a fairness measure; similar to NDCG, higher values are better. We define the optimal value of λ for a user group g as follows:

$$\lambda_g = \underset{\lambda \in [0,1]}{\arg\max} \frac{\text{NDCG}_g \cdot JSF_g}{\text{NDCG}_g + JSF_g} \tag{3}$$

In other words, for a given group g we select λ to maximize the harmonic mean between utility (NDCG) and fairness (JSF) for the group. The selection is done by conducting a grid search on the interval $[0, 1]$.

4 EXPERIMENT SETUP

Datasets. We investigate two datasets, *MovieLens-1M (ML-1M)* [2] in the movie domain and *LFM-2b* [3] in the music domain. The former provides ratings for 6K users and almost 4K movies. The latter is a Last.fm music listening dataset, which we modify to fit our experimental setup. Firstly, we consider only listening events in the year 2019 of users with meta-information regarding age, gender and country. Secondly, all user–item interactions with a playcount (*PC*) of < 2 are removed to reduce the number of spurious interactions and noise. Thirdly, we only consider tracks that were listened to by at least 5 different users (constraint 1) and we only consider users who have listened to at least 5 different tracks (constraint 2). Lastly, we treat each user–item interaction in a binary way – 1 if the user has listened to a track, 0 otherwise. Furthermore, we sample 100K tracks uniformly-at-random to ensure items of different characteristics are equally likely to be included in the final subset. We then reinforce constraints 1 and 2. This ultimately results in almost 10K users retained with a total of 10.7M listening events. See Table 1 for details.

Algorithms and Baselines. We study popular collaborative filtering algorithms (i. e., neighborhood-based, neural matrix factorization, autoencoders, and graph convolution networks), briefly described in the following. For consistency, we use the algorithm implementations from the *Recbole* framework [27].[4] These are: **Bayesian Personalized Ranking (BPR)** [22] adopts an optimization function that ranks the items consumed by the users according to their preferences, by defining an implicit order between pairs of items. **Item k-Nearest Neighbors (KNN)** [6] recommends items based on item-to-item similarity. Specifically, an item is recommended to a user if it is similar (in terms of ratings or interactions)

[2] https://grouplens.org/datasets/movielens/1m
[3] http://www.cp.jku.at/datasets/LFM-2b
[4] https://recbole.io/

Table 1. Statistics of the LFM-2b and MovieLens-1M datasets, broken down into investigated user gender groups.

Dataset	Demographic	Users	Items (Tracks / Movies)	Listening Events / Interactions	Sparsity
LFM-2b	All	9,759	99,922	10,746,088	99.8063%
	Female	1,820	70,780	1,856,757	99.8359%
	Male	7,939	99,890	8,889,331	99.7995%
MovieLens-1M	All	6,040	3,706	1,000,209	95.5316%
	Female	1,709	3,481	246,440	96.1090%
	Male	4,331	3,671	753,769	95.3038%

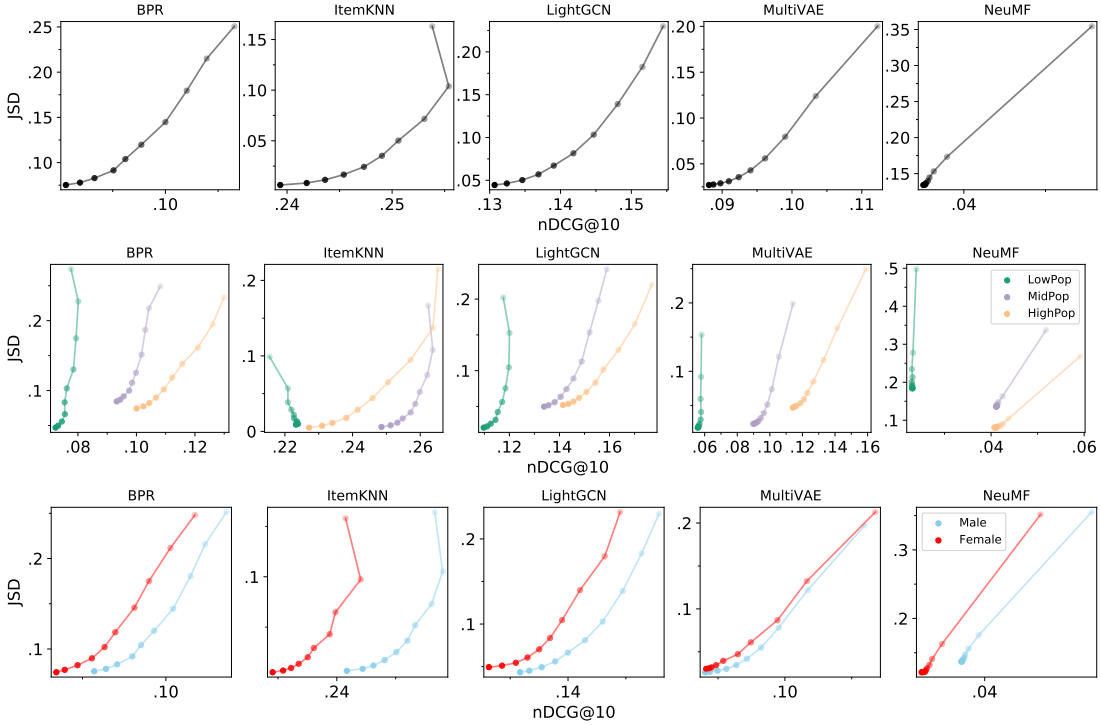

Fig. 1. Trade-off between utility and fairness in LFM-2b dataset. Paler points denote smaller weight for the fairness objective (λ). Points with the highest JSD for each curve correspond to $\lambda = 0$.

to the items previously interacted with by the user. **Light Graph Convolution Network (LightGCN)** [11] learns user and item embeddings by linearly propagating them through the user-item interaction graph. It uses the weighted sum of the embeddings learned at all layers as the final embedding. **Neural Matrix Factorization (NeuMF)** [12] builds on the basic matrix factorization approach but replaces the inner product with a neural architecture that can learn an arbitrary function from the interaction data. **Variational Autoencoders (MultiVAE)** [18] estimates a probability distribution over all items, given the user's interaction vector.

Training and Evaluation. To evaluate the aforementioned algorithms, we partition the interactions of each user in train/validation/test groups with a 60-20-20 ratio split. Therefore, 60% of all users' interactions are used to train the algorithms. We maximize the NDCG @10 metric over the validation set. All results are reported for the test set.

Popularity Bias Mitigation. We conduct a series of tests comparing utility and fairness of recommendation lists produced by the above mentioned algorithms and re-ranked by the CP post-processing mitigation technique with

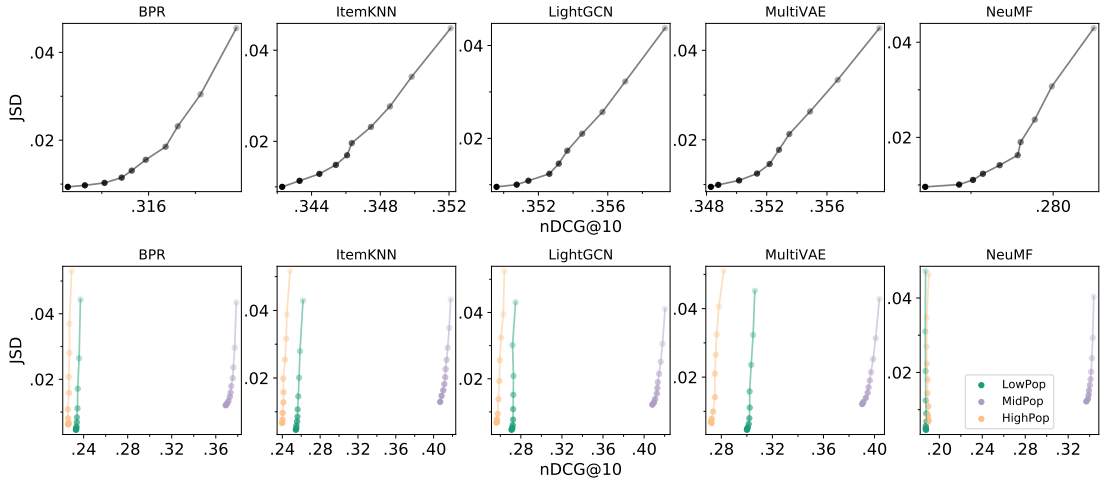

Fig. 2. Trade-off between utility and fairness in ML-1M dataset. Paler points denote smaller weight for the fairness objective (λ). Points with the highest JSD for each curve correspond to $\lambda = 0$.

different settings. For every algorithm[5], we re-rank the list of top 100 recommendations to create the final list of 10 items for each user. We test it for ten values of the weighting parameter λ from 0 (no mitigation) to 0.9 (weight 0.9 to the fairness objective and 0.1 to utility) with step size 0.1. We aim to exploit potential differences in the way various user groups respond to the popularity bias mitigation to achieve a better trade-off between utility and fairness. To this end, we split users uniformly-at-random into train and test sets of the same size striving to ensure all user groups are represented in both. We search for optimal values of λ for each user group and the whole population using the criterion in Equation 3 on the train set. We then compute the new recommendation lists for the test users, applying mitigation with the weights learned for their corresponding user groups.

5 RESULTS

We approach **RQ1** by analyzing popularity calibration over various factors, shown in Figures 1 and 2. The figures report NDCG (utility measure) against JSD (popularity bias measure) for every recommendation algorithm and the ten values of λ, respectively, for LFM-2b and ML-1M dataset. On the plots, the opacity of the points corresponds to the value of λ, such that the palest show the results of $\lambda = 0$. Let us first look at the the top row of the plots in each figure which shows the results achieved on the whole population of the dataset. On LFM-2b we notice differences in behavior of recommended lists produced by different recommendation algorithms. In particular, KNN shows an increase in NDCG combined with an improvement in fairness at $\lambda = 0.1$. BPR and LightGCN show steady progress towards debiased results of lower utility with growing λ. At the same time, MultiVAE and NeuMF demonstrate a notably larger drop in utility over the first step (from $\lambda = 0$ to 0.1). Considering that every point on each plot corresponds to a new set of items recommended to most of the users, we can examine the quality of the top 100 recommendation lists produced by different algorithms. In this regard, a smooth decay of NDCG and JSD signify a better overall quality of the top 100 list as it allows to debias the recommendation gradually without a sudden drop in utility. A sudden drop in both metrics however would mean that the achieved utility to a large degree comes from concentration of the popular items at the

[5]We re-scale relevance scores provided by the algorithms to the interval [0, 1] in order to ensure comparability of mitigation results, see Equation 2.

Table 2. Results of popularity bias mitigation with different settings.

		LFM-2b				ML-1M			
Algorithm	Metric	$\lambda = 0$	λ_{all}	λ_{pop}	λ_{gender}	$\lambda = 0$	λ_{all}	λ_{pop}	λ_{gender}
BPR	NDCG ↑	**0.102**	**0.102**	**0.102**	**0.102**	**0.315**	0.314	0.314	0.314
	JSD ↓	0.253	0.253	**0.252**	0.253	0.044	0.044	**0.041**	0.044
KNN	NDCG ↑	0.252	**0.254**	**0.254**	**0.254**	0.354	0.354	0.354	0.354
	JSD ↓	0.163	0.138	**0.129**	0.139	0.044	0.044	0.044	0.044
LightGCN	NDCG ↑	0.152	0.152	0.152	0.152	0.352	0.352	0.352	0.352
	JSD ↓	0.230	0.229	0.229	0.229	0.045	0.045	0.044	0.044
MultiVAE	NDCG ↑	**0.110**	**0.110**	**0.110**	**0.110**	0.357	0.357	0.357	0.357
	JSD ↓	0.200	0.200	**0.196**	0.200	0.046	0.046	0.046	0.046
NeuMF	NDCG ↑	**0.050**	**0.050**	0.049	**0.050**	**0.277**	0.273	0.275	0.275
	JSD ↓	0.358	0.358	**0.292**	0.358	0.043	**0.016**	0.020	0.024

top, and the rest of the top 100 contains less relevant items. Considering this, we observe that on the ML-1M dataset all five recommenders show a very similar behavior.

Approaching **RQ2**, we look at the rest of the plots in Figures 1 and 2 which as before show the trade-off between utility and fairness, but here according to specific user groups. The second rows of the plots correspond to the user grouping according to their inclination towards the consumption of popular music (aka. mainstreamness). The plots on the third row of Figure 1 show the results of the users grouped by genders[6]. Analyzing the results according to the mainstreamness groups for LFM-2b, we observe initially all algorithms provide the best utility to the HighPop group, the group most exposed to popularity bias varies from model to model. BPR, LightGCN and KNN show that LowPop user group benefit from the mitigation method in terms of utility. KNN also shows the same for the MidPop group. In most cases, the HighPop group experiences the largest drop in utility through popularity bias mitigation. These findings support our hypothesis that selecting mitigation parameters separately for each user group potentially improves the trade-off between utility and fairness. On the ML-1M dataset we see that all three groups maintain a certain level of utility, while steadily decreasing the bias metric, showing that all five algorithms on this dataset can successfully achieve good results in terms of bias mitigation while maintaining utility. Considering the results on genders, we do not observe a notable difference in the overall patterns between the genders.

Finally addressing **RQ3**, Table 2 shows the results of the experiment with group-specific values of λ. For every algorithm and dataset, we report utility and bias metric under four conditions: $\lambda = 0$ no mitigation, λ_{all} where one optimal parameter value is selected for the whole population, λ_{pop} where specific optimal parameter value selected for each popularity inclination user group, and finally λ_{gender} indicating the selection of specific parameter value for each gender. We observe that, except for NeuMF on ML-1M, group-specific λs provide the best result on bias metric and trade-off between utility and fairness, namely a lower value of *JSD* together with utility staying on the same level or slightly decreasing. Among all algorithms LightGCN has shown the lowest sensitivity to the mitigation, as its values do not notably drop in either utility or bias metrics. All algorithms show low bias metrics without any mitigation on ML-1M, nevertheless leveraging group specific λ allows to improve trade-off between utility and fairness for BPR and NeuMF.

[6]We do not report this grouping for ML-1M as all five algorithms display the same behavior for both genders: steady decrease of bias metric with only slight decrease in utility.

6 CONCLUSION AND FUTURE WORK

We explore the effectiveness of a post-processing popularity bias mitigation technique, Calibrated Popularity, applied to an array of state-of-the-art recommendation algorithms. We conduct experiments on two datasets from the music and movie domain, of different size and sparsity, considering how various user groups (defined by gender and mainstreaminess) are affected by bias mitigation. The larger music dataset LFM-2b exposes discrepancies in behavior of different algorithms. NeuMF and MultiVAE show a notable drop in utility and bias metrics even with light bias mitigation settings. KNN shows an increase in utility with moderate bias mitigation applied. We also show that for BPR, KNN, and LightGCN users least interested in popular items can benefit in terms of utility from popularity bias mitigation as opposed to other users. Our experiments show that different user groups respond to the mitigation differently depending on their inclination towards consumption of popular content. We also found responses of different gender groups relatively similar. Finally, we conduct an experiment showing that selecting mitigation parameters individually for every user group (by interest towards popular items) leads to a better trade-off between utility and fairness overall. In future work, other mitigation strategies as well as criteria for selecting optimal mitigation parameters could be tested. Also, additional user groups could be addressed, including choosing individual settings for each user.

ACKNOWLEDGMENTS

This work received financial support by the Austrian Science Fund (FWF): P33526 and DFH-23; and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through grants LIT-2020-9-SEE-113 and LIT-2021-YOU-215.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*. 42–46.

[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing popularity bias in recommender systems with personalized re-ranking. In *The thirty-second international flairs conference*.

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Unfairness of Popularity Bias in Recommendation. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, September 20, 2019 (CEUR Workshop Proceedings, Vol. 2440)*. CEUR-WS.org. http://ceur-ws.org/Vol-2440/paper4.pdf

[4] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward C. Malthouse. 2021. User-centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, Judith Masthoff, Eelco Herder, Nava Tintarev, and Marko Tkalcic (Eds.). ACM, 119–129. https://doi.org/10.1145/3450613.3456821

[5] Diego Corrêa da Silva, Marcelo Garcia Manzato, and Frederico Araújo Durão. 2021. Exploiting personalized calibration and metrics for fairness recommendation. *Expert Systems with Applications* 181 (2021), 115112. https://doi.org/10.1016/j.eswa.2021.115112

[6] Mukund Deshpande and George Karypis. 2004. Item-Based Top-N Recommendation Algorithms. *ACM Trans. Inf. Syst.* 22, 1 (jan 2004), 143–177. https://doi.org/10.1145/963770.963776

[7] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on Fairness, Accountability and Transparency*. 172–186.

[8] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (1996), 330–347. https://doi.org/10.1145/230538.230561

[9] Christian Ganhör, David Penz, Navid Rekabsaz, Oleg Lesota, and Markus Schedl. 2022. Unlearning Protected User Attributes in Recommendations with Adversarial Training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2142–2147. https://doi.org/10.1145/3477495.3531820

[10] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[11] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. *LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation.* Association for Computing Machinery, New York, NY, USA, 639–648. https://doi.org/10.1145/3397271.3401063

[12] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 173–182. https://doi.org/10.1145/3038912.3052569

[13] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (2015), 427–491.

[14] Michael Jugovac, Dietmar Jannach, and Lukas Lerche. 2017. Efficient optimization of multiple recommendation quality factors according to individual user tendencies. *Expert Systems with Applications* 81 (2017), 321–331. https://doi.org/10.1016/j.eswa.2017.03.055

[15] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Christoph Trattner, and Lars Skjærven. 2022. *Mitigating Popularity Bias in Recommendation: Potential and Limits of Calibration Approaches.* 82–90. https://doi.org/10.1007/978-3-031-09316-6_8

[16] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12036)*. Springer, 35–42. https://doi.org/10.1007/978-3-030-45442-5_5

[17] Oleg Lesota, Alessandro B. Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected?. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge (Eds.). ACM, 601–606. https://doi.org/10.1145/3460231.3478843

[18] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 689–698. https://doi.org/10.1145/3178876.3186150

[19] J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151. https://doi.org/10.1109/18.61115

[20] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 2145–2148. https://doi.org/10.1145/3340531.3412152

[21] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada). 306–316.

[22] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) *(UAI '09)*. AUAI Press, Arlington, Virginia, USA, 452–461.

[23] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis. In *ACM SIGIR Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 337–341. https://doi.org/10.1145/3498366.3505791

[24] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 154–162. https://doi.org/10.1145/3240323.3240372

[25] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1569–1578. https://doi.org/10.1145/3132847.3132938

[26] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating Bias in Search Results Through Contextual Document Reranking and Neutrality Regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2532–2538. https://doi.org/10.1145/3477495.3531891

[27] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Kaiyuan Li, Yushuo Chen, Yujie Lu, Hui Wang, Changxin Tian, Xingyu Pan, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2020. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. *CoRR* abs/2011.01731 (2020). arXiv:2011.01731 https://arxiv.org/abs/2011.01731

[28] Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. 2021. Popularity Bias in Dynamic Recommendation. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 2439–2449. https://doi.org/10.1145/3447548.3467376