Counteracting Popularity-Bias and Improving Diversity Through Calibrated Recommendations

Andre Sacilotti[a], Rodrigo F. Souza[b] and Marcelo G. Manzato[c]
Mathematics and Computer Science Institute, University of São Paulo, Av. Trab. Sancarlense 400, São Carlos-SP, Brazil

Keywords: Recommender System, Popularity Bias, Fairness, Calibration.

Abstract: Calibration is one approach to dealing with unfairness and popularity bias in recommender systems. While popularity bias can shift users towards consuming more mainstream items, unfairness can harm certain users by not recommending items according to their preferences. However, most state-of-art works on calibration focus only on providing fairer recommendations to users, not considering the popularity bias, which can amplify the long tail effect. To fill the research gap, in this work, we propose a calibration approach that aims to meet users' interests according to different levels of the items' popularity. In addition, the system seeks to reduce popularity bias and increase the diversity of recommended items. The proposed method works in a post-processing step and was evaluated through metrics that analyze aspects of fairness, popularity, and accuracy through an offline experiment with two different datasets. The system's efficiency was validated and evaluated with three different recommendation algorithms, verifying which behaves better and comparing the performance with four other state-of-the-art calibration approaches. As a result, the proposed technique reduced popularity bias and increased diversity and fairness in the two datasets considered.

1 INTRODUCTION

Recommender systems are part of people's routine, influencing decisions they make when accessing online services with the suggestion of specific content for each user. However, some of these systems have certain limitations, such as popularity bias, in which popular items are recommended more often than unpopular items, entailing the long tail effect.

It is known that popularity bias in recommender systems is a particular case of the class imbalance problem in machine learning, which usually leads to unfair classification (Abdollahpouri et al., 2020). Although there are many definitions of fairness, a suitable one is the ability of recommender systems to provide consistent performance across different groups of users (Ekstrand et al., 2018). One of the metrics used to measure the quality of recommendations is calibration, which measures the capability of the system to provide users with proportions of items in their areas of interest that are consistent with their preferences (Steck, 2018).

Hence, when a system recommends items differently from the user's interests, it is considered a poorly calibrated system. When users deal with different levels of calibration, the system is unfair to a group of users (Abdollahpouri et al., 2020). When a system provides recommendations that meet users' preferences, it is an accurate system, as it offers user satisfaction by recommending relevant and exciting items.

To balance the relationship between offering a calibrated system and one that provides accurate recommendations, (da Silva et al., 2021) proposed a system that works in a post-processing stage. It was designed to be independent of any recommendation algorithm and focuses on divergence measures. Similarly, (Steck, 2018) introduces a calibration technique as a post-processing step to bring fairer recommendations to users in a way that suits their interests. (Geyik et al., 2019) also present a framework to return fairer recommendations, initially evaluating the existing bias in the system concerning attributes such as gender and age, and then applying algorithms to reclassify the results.

Despite the promising results, these works still suffer from popularity bias since calibration is based on the items' metadata (e.g., genres) and does not

[a] https://orcid.org/0000-0001-9359-4298
[b] https://orcid.org/0000-0002-9272-4107
[c] https://orcid.org/0000-0003-3215-6918

Sacilotti, A., Souza, R. and Manzato, M.
Counteracting Popularity-Bias and Improving Diversity Through Calibrated Recommendations.
DOI: 10.5220/0011846000003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 1*, pages 709-720
ISBN: 978-989-758-648-4; ISSN: 2184-4992

consider the popularity aspect for calibration. Although the literature reports works that calibrate recommendations considering the popularity bias (Yalcin, 2021; Lesota et al., 2021), we argue the necessity of a model-agnostic calibration approach, which could benefit from a range of high-precision recommendation models available nowadays.

Based on this, we propose a system that works in a post-processing step and is independent of recommendation algorithms. For this, unlike (da Silva et al., 2021) and (Steck, 2018), we propose a calibration approach based on the popularity of items and the proportion of users' preferences based on this aspect of popularity. We hypothesize that this calibration provides fairer recommendations and reduces the popularity bias in collaborative systems. We evaluated the proposed method using metrics that analyze aspects of fairness, popularity, precision and quality from an offline experiment with the MovieLens-20M and Yahoo Movies datasets. The obtained results are promising when compared to existent state-of-art baselines.

The structure for this paper is as follows. In Section 2, we present the related work and compare the existing approaches against our proposal. Section 3 details the calibration framework proposed in this work. Section 4 explains the design of the experiments carried out in the research, and Section 5 presents the results obtained. Finally, in Section 6, we conclude the work, highlighting the effectiveness of fairer recommendations and a reduction in the popularity bias of the proposed system. In addition, we point out some directions for future work. All source code and datasets split to reproduce the reported results are publicly avaliable[1].

2 RELATED WORK

Several state-of-the-art works present approaches to calibrate the recommendation system to bring fairer recommendations and avoid disfavoring less popular items. This way, (Kaya and Bridge, 2019) based their calibration approach on (Steck, 2018)'s work, replacing a calibration based on items' genres with one based on sub-profiles of users and their interests. This work differs from ours, as we focus on each user's interest level by popularity.

(Abdollahpouri et al., 2021) present an approach that utilizes different levels of user interest in popularity. In particular, the authors divided users into three preference groups, which was also adopted in our work, although we used a different methodology for calibration. Our system has a genre-based calibration for when users prefer less popular items so they get more recommendations that interest them. We compared the performance of the two systems in our work.

(Zhu et al., 2020) demonstrate that systems that adopt Bayesian Personalized Ranking make unfair recommendations, as they favor items over others. The work then proposes a calibration model capable of reducing unfairness; however, as opposed to our work, item popularity is not considered.

(Beutel et al., 2019) use some metrics to assess the fairness of recommendation systems and propose a pairwise regularization approach to improve fairness during the recommendation algorithm training process. Although the promising results, the work differs from ours because it is not a post-processing step and does not analyze the level of user interest regarding popularity. There are some other calibration approaches to reduce the impact of biases, such as the one made by (Wang et al., 2021) to propose a system that models the causal effect on the user representation during item score prediction and that can work with several recommendation algorithms. In addition, there are approaches based on optimal solutions for methodologies based on heuristics (Seymen et al., 2021), but they are out of the scope of this work.

In an approach adopted by (Yalcin and Bilge, 2021) and (Borges and Stefanidis, 2021), the system reclassifies the recommended items penalizing the most popular items, reducing bias and accuracy. This approach is different from our work, as in our study we used the popularity aspect to rank items according to the user's preferences.

There is also an approach that adopts Inverse Propensity Weighting (IPW), where the impact of popular items is reduced in the training phase through the analysis of a cause and effect relationship to reduce the problem of popularity bias (Wei et al., 2021). This method is compatible with many models, as is our proposal, but it is applied in the training phase. (Boratto et al., 2021) calibrate the system based on the long tail to measure how the system treats its items equally in this distribution, proposing a metric to minimize the biased correlation between the item and its popularity. (Zhang et al., 2021) attempt to remove the bias at the same moment recommendations are generated, unlike our approach, whose bias removal is accomplished in a post-processing step. The performance of these last three works (Wei et al., 2021) (Boratto et al., 2021) (Zhang et al., 2021) were compared against our proposal and detailed in the experiments.

[1] https://github.com/Andre-Sacilotti/recommender-popularity-bias-calibration

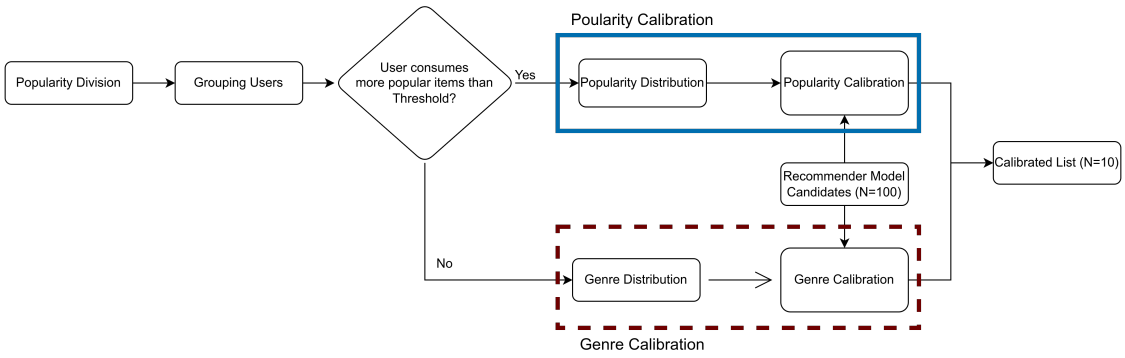

Figure 1: Flowchart showing our personalized method highlighting the genre calibration and our popularity calibration steps.

3 CALIBRATION FRAMEWORK

This section formalizes the problem we are dealing with and the proposed solution. Figure 1 presents the flowchart adopted in this work's calibration proposal, which will be more detailed in this section. In addition, we provide the Table 1, that contains the notation used in the following sections.

Problem Statement. Suppose we have a set of items $I = \{i_1, i_2, ..., i_{|I|}\}$, a set of users $U = \{u_1, u_2, ..., u_{|U|}\}$ and a set of candidate items for each user $CI_u = \{i_1, i_2, ..., i_N\}$, where N is the number of items suggested by the recommender system. We have two valuable pieces of information to know the user's preferences: (1) genres or metadata of items interacted by user u; and (2) popularity of items interacted by user u. Our task is to exploit users' preferences related to popularity and genres to increase the fairness and accuracy of recommendations and reduce popularity bias by increasing the long tail coverage of all users based on CI_u generated by any recommendation model.

Approach. We propose the personalized calibration approach as shown in Figure 1. In particular, our approach is divided into two methods: **popularity calibration** (highlighted in continuous blue line in Figure 1), which extends the genre calibration previously proposed by (Steck, 2018) (highlighted in dotted red line in Figure 1); and **personalized calibration** (Figure 1 as a whole), which uses genre calibration and popularity calibration in a unified model to provide recommendations calibrated according to popularity and genres.

To calibrate the recommendation list based on the popularity of the items consumed by the user in the past, we introduce a popularity division to group items based on how much users access them. In addition, to provide personalized calibration based on popularity and genres, we first group users based on their consumption. In this way, if the user consumes popular items below an established limit, we perform a genre calibration; otherwise, we perform a popularity calibration to meet the preference level for this aspect.

3.1 Popularity Division

The popularity division, introduced in this paper and shown in Figure 1, is performed based on the long tail concept in recommender systems. We propose to divide this curve into three parts. The **Head (H)**, with items representing the top 20% of the total of the past interactions. Then, we get the **Tail (T)** with items that sum the less 20% of interactions. Finally, the **Mid (M)** group contains items that are neither **Head (H)** nor **Tail (T)**. It is worth mentioning that this division by percentage was chosen based on Pareto's principle.

3.2 Grouping Users

As shown in Figure 1, our unified model switches between popularity and genre calibration. To make this decision, it is required to group users according to their interests in unpopular/popular items. So, we defined the threshold as the mean of all ratios, which is a value that can be easily computed on every dataset, as shown in Equation 1:

$$G_{threshold} = \frac{\sum_{u}^{U} \frac{\sum_{i}^{I_u} \mathbf{1}(i)}{|I_u|}}{|U|} \quad (1)$$

where $\mathbf{1}(i)$ is an indicator function that returns 1 if the item i, interacted by user u, is in the **H** popularity category. Finally, we assume that if the ratio of items in the category **H** is lower than $G_{threshold}$, then we should get a recommendation list calibrated by genre; otherwise, by popularity.

Table 1: Notations used in this work.

Notation	Explanation
U	Set of users
I	Set of items
u	A specific user
i	A specific item
CI_u	Candidate items for user u
I_u	Set of items interacted by u
$w_r(u,i)$	Rating user u gave to item i
$w_p(u,i)$	Rank position of i in the recommended list for u
λ	Constant used to balance the trade-off between recommendation scores and fairness
t	A popularity category

3.3 Popularity Distributions

This section describes how to calculate the distribution used to calibrate recommendation lists based on the item's popularity. In this work, we adapted the formulation proposed by (Steck, 2018).

His work assumes items can have more than one genre, which is not valid in our context, where an item has only a level of popularity. So instead, we calculate the sums of weights of every popularity type over the sum of all weights.

The $p(t|u)$ is defined as the target distribution that was based on the popularity of the items that the user interacted with in the past. In Equation 2 we used the weights $w_r(u,i)$ as the explicit or implicit rating the user u gave to the item i:

$$p(t|u) = \frac{\sum_{i \in I_u} w_r(u,i) \cdot p(t|i)}{\sum_{i \in I_u} w_r(u,i)} \tag{2}$$

where $p(t|i)$ is defined as 1 if the item i is in the popularity category t. Then to deal with the recommended list distribution, Equation 3 defines $q(t|u)$:

$$q(t|u) = \frac{\sum_{i \in R^*_u} w_p(u,i) \cdot p(t|i)}{\sum_{i \in R^*_u} w_p(u,i)} \tag{3}$$

In this case, we use the weights $w_p(u,i)$ as the rank position of the item i in the reordered recommended list to the user u.

3.4 Fairness Measure

In our context, the system is fair when it meets the popularity proportions expected by the user. Therefore, if the user consumes fewer popular items than the established limit, we will perform a genre-based calibration because, in this case, the user does not care about the popularity of the items. If the user consumes more popular items than the established limit, we will calibrate based on popularity, respecting the user's level of interest in this aspect. Several metrics assess fairness in recommender systems (Verma et al., 2020). However, in our case, we use the Kullback-Leibler for the same reasons pointed out by (Steck, 2018) and exploited by (da Silva et al., 2021).

The Kullback-Leibler quantifies the inequality in the interval $[0,\infty]$, where 0 means both distributions are almost the same, and higher values indicate unfairness. Also, we adopted the regularization proposed by (Steck, 2018), which defined the $\alpha = 0.01$ as a regularization variable to avoid zero division when $q(t|u)$ goes to zero.

$$D_{KL}(p\|q) = \sum_t p(t|u) \cdot log \frac{p(t|u)}{(1-\alpha) \cdot q(t|u) + \alpha \cdot p(t|u)} \tag{4}$$

Although there are other divergence metrics, like Hellinger and Person Chi-Square, proposed by (Cha, 2007) and exploited by (da Silva et al., 2021), we use only the Kullback-Leibler due to its simplicity.

3.5 Calibration

This section explains the final calibration process shown in Figure 1. We call **calibration** refereeing as the process to find the optimal set R^*_u, using the maximum marginal relevance, as shown in Equation 5, where D_{KL} is the fairness function. In this formulation, when $\lambda = 0$, it focuses only on the recommendation scores, and when $\lambda = 1$, we focus on fair items concerning the user's profile.

$$R^*_u = \max_{CI_u} (1-\lambda) \cdot \sum_{i \in CI_u} w_r(u,i) - \lambda \cdot D_{KL}(p, q(CI_u)) \tag{5}$$

It is worth mentioning that such formulation is similar to the calibration approach proposed by

(Steck, 2018). Consequently, although we focus on popularity in this work, it is possible to adopt a greedy approach to solve Equation 5, whose details can be found in (Steck, 2018).

4 EXPERIMENTAL SETUP

This section describes the steps to reproduce our comparisons, including dataset pre-processing, baseline parameters, training, and evaluation methodology.

4.1 Datasets

In this study, we used two datasets from the movies domain:

Yahoo Movies[2]: This dataset is a user-movie rating, where the user gives a rating from one to five to the movies they watched. In the pre-processing step, we only removed movies with no genres in the metadata. Instead of binarizing the rating as done by (Steck, 2018), we used the explicit feedback as the weight $w_r(u,i)$ in Equation 2.

MovieLens-20M[3]: In this dataset, similarly to (Steck, 2018) and unlike the Yahoo Movies dataset, we binarized the ratings by keeping interactions whose rating was above 4. Also, due to hardware limitations, we reduced the dataset's size by removing movies with less than ten interactions and users with less than 190 movies.

Table 2 summarizes important statistics about the processed datasets. For reproducibility, we provided the train-test split and folds used in our experiments[4].

Table 2: Statistics of the datasets after all pre-processing steps.

Dataset	# Users	# Ratings	# Items
Yahoo Movies	7,642	221,367	10,825
MovieLens 20M	11,530	3,786,788	10,347

4.2 Metrics

In our experiments, we evaluated the effects of different calibrations in terms of precision, fairness, and popularity bias, as detailed next:

1. **Precision and Quality**: In these topics, we used the Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP) metrics to measure the rank quality of the item in the re-ranked list. MAP and MRR range in the interval $[0,1]$ where higher is better.
2. **Fairness**: In the fairness topic, we used a metric proposed by (da Silva et al., 2021), called Mean Rank Miscalibration (MRMC), which covers the interval $[0,1]$, where lower is better. Initially, it was used to compute the fairness in genres on the recommendation list, but we also used it to calculate the popularity miscalibration in our work.
3. **Popularity Bias**: We used the metrics long-tail coverage (LTC) (Abdollahpouri et al., 2018) and group average popularity (ΔGAP) (Abdollahpouri et al., 2019b) to measure popularity bias. The LTC metric indicates the fraction of items that users see in the recommendation lists and varies in the interval $[0,1]$, where 0 means all recommended items are the most popular and 1 means all items recommended to a user are in the less popular categories. Thus, the closer to 1, the more diverse content will be recommended (Abdollahpouri et al., 2018). The ΔGAP ranges in the interval $[-1,\infty]$, where negative values mean recommendations are less popular than expected by the users' preferences, and positive values mean recommendations are more popular than expected. We also adopted three divisions of user groups, based on (Abdollahpouri et al., 2019b) for the ΔGAP: **BlockBuster (BB)** whose users' consumption is at least 50% of the most popular items, **Niche (N)** where users' consumption is at least 50% of the lowest popularity items and **Diverse (D)** whose users' preferences diverge from the other two groups.

4.3 Experiments

We executed three times the calibration process in the experiments involving the MovieLens and Yahoo Movies datasets. We got the mean of the values outputted by the metrics to guarantee the stability of the results. Also, the train and test sets were chosen by randomly splitting the dataset in 70/30% of interactions, respectively (Abdollahpouri et al., 2019a; da Silva et al., 2021).

The process of calibration does not depend on the recommender system algorithm. It acts as a post-processing step where, after the model predicts the candidate items for a user, we apply the calibration technique described in Equation 5 to find the best list of items for that user. Consequently, to understand how the calibration approaches perform under different recommender algorithms, we used three

[2]https://webscope.sandbox.yahoo.com/

[3]https://grouplens.org/datasets/movielens/20m/

[4]https://drive.google.com/drive/folders/1wlMyypxzpTo86nydWucz7oMXCocUR5K0?usp=sharing

well-known models described below, based on (Steck, 2018) and (da Silva et al., 2021) works. For some models, we used the implementation provided by (Hug, 2020).

1. **SVD++:** Singular Value Decomposition extension (Koren, 2008) to work with implicit feedback. Similarly to (da Silva et al., 2021), we used $ne = 20$ as the number of epochs, $\gamma_u = \gamma_i = 0.005$ as the learning rate for users and items, $\lambda_u = \lambda_i = 0.02$ as regularization constants, and $f = 20$ factors.
2. **NMF:** Non-negative Matrix Factorization proposed by (Luo et al., 2014). Similarly to (da Silva et al., 2021), we used $ne = 50$, $\gamma_u = \gamma_i = 0.005$, $\lambda_u = \lambda_i = 0.06$ and $f = 15$.
3. **VAE:** Variational Autoencoder for Collaborative Filtering, proposed by (Liang et al., 2018). We used Multi-VAE with annealing and optimized the best β parameter for each dataset with 400 epochs of training. We used the implementation made by Microsoft[5].

The experiment for each recommender system consists of using the training data to feed the model to learn the user's preferences based on items interacted in the past, represented as I_u. After the training step, we predict all missing ratings, and for every user, we select the top 100 items with the highest predicted rating, represented as CI_u. We use the weight $w_r(u,i)$ as the rating the algorithm predicted for the candidate item. Finally, the final recommendation list R^*_u is created with the top 10 items given by the calibration process.

We analyzed three types of calibration. First is the genre calibration, proposed by (Steck, 2018). Then, regarding our proposal, we separately analyze the performance of **popularity calibration** and **personalized calibration**. Both calibrations were described in Section 3.

For the trade-off between similarity and fairness metrics, in Equation 5, we adopted the values described by (Steck, 2018), ranging from $\lambda \in [0, 0.1, 0.2, \ldots, 1]$. Our evaluation consists of three recommenders, four types of calibration, and eleven trade-off weights, resulting in $3 \times 11 \times 4 = 132$ combinations of the recommendation list to be evaluated for each dataset.

4.4 Baselines

To compare the efficiency of our proposed method regarding the popularity bias, we selected four state-of-the-art methods specialized in popularity debiasing. To a fair comparison, we applied the same train-test split methodology and result stability to calibration approaches.

1. **PairWise:** Proposed by (Boratto et al., 2021), this method acts as an in-processing step for popularity debiasing. For the Yahoo Movies dataset, we applied $epoch = 100$, $batch = 1024$, and we chose the best α ranging in the interval $[0,1]$. For the MovieLens dataset, we used $batch = 2048$ and $epoch = 20$. We followed the authors' implementation[6] and obeyed all instructions.
2. **MF MACR:** Proposed by (Wei et al., 2021), this method acts as a post-processing step for popularity debiasing. For the Yahoo Movies dataset, we chose the fine-tuned parameters as $epoch = 100$, $batch = 1024$, $lr = 0.01$, $reg = 0.01$, $alpha = 0.001$, $beta = 0.001$ and $c = 20$. For the Movielens 20M dataset, we used $batch = 2048$. We followed the authors' implementation[7] and obeyed all instructions.
3. **PDA:** Proposed by (Zhang et al., 2021), this method implements a new training and inference paradigm via causal intervention for popularity debiasing. For both datasets we used $epoch = 2000$, $batch = 2048$, $lr = 0.01$, $reg = 0.01$ and $pop_{exp} = 0.16$. We followed the authors' implementation[8] and obeyed all instructions.
4. **CP:** Proposed by (Abdollahpouri et al., 2021), this method implements a calibration technique for popularity, similar to our proposed popularity calibration, but using the Jensen-Shannon divergence metric for comparing the profile and recommendation distributions. We followed the authors' method for both datasets to split the popularity into groups and exploited the parameter $\lambda \in [0,1]$. This method is compared against our proposals using the same set of three recommender algorithms (SVD++, NMF and VAE).

4.5 Statistical Analysis

Comparing the three types of calibration (genre, popularity, personalized) resulted in 33 combinations for each metric and each experiment. Based on this, we chose to evaluate the statistical significance with the Wilcoxon test, as we are comparing the difference between two dependent samples that vary in terms of the same trade-off ($\lambda = [0, 0.1, \cdots, 1]$).

[5] github.com/microsoft/recommenders/blob/main/examples/02_model_collaborative_filtering/multi_vae_deep_dive.ipynb

[6] https://github.com/biasinrecsys/wsdm2021

[7] https://github.com/weitianxin/MACR

[8] https://github.com/zyang1580/PDA

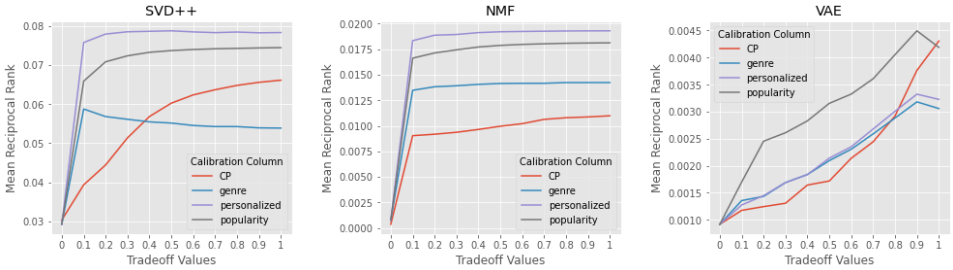

Figure 2: MRR results over the selected recommenders and three types of calibration on the Yahoo Movies dataset. The results are statistically significant (p-value < 0.01).

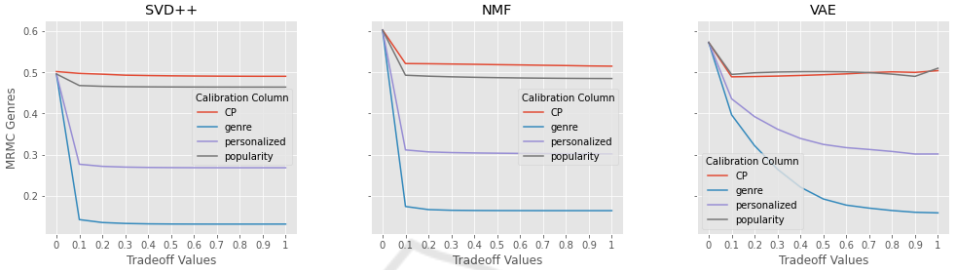

Figure 3: MRMC of genres results over the selected recommenders and three types of calibration on the Yahoo Movies dataset.

However, in comparing the baselines with the personalized calibration, we have 3 combinations of results for each metric in each experiment. In addition, we are dealing with independent samples as we are testing the models with the trade-off parameters that resulted in the best results. Consequently, we adopted the Student's t-test, which can also give more statistical power in extremely small samples ($2 \leq n \leq 5$) as studied by (de Winter, 2013), giving more confidence to the statistical significance.

5 EXPERIMENTAL RESULTS

This section presents the results of our experiments for both datasets: Yahoo Movies and MovieLens 20M.

For both datasets, we present the results for Mean Reciprocal Rank, Genre Mean Rank Miscalibration, Popularity Mean Rank Miscalibration, Long Tail Coverage, and a comparison against the baselines. It is worth mentioning that the comparisons under a varying number of trade-off values are accomplished only with our proposals (Popularity Calibration and Personalized Calibration), genre calibration and CP (Abdollahpouri et al., 2021). For the comparison against the other baselines, these models were set using the trade-off value that achieved the highest LTC value.

5.1 Yahoo Movies

5.1.1 Mean Reciprocal Rank

Figure 2 shows the results of **MRR** on the Yahoo Movies dataset. We observe that for the three recommenders, at least one of our proposed methods overcomes the CP method. We also notice that with an increase in the trade-off (λ parameter in Equation 5), the MRR tends to achieve higher values than the MRR obtained in a recommendation list without any calibration ($\lambda = 0$). In particular, the **SVD++** model performed best when compared with other recommenders. Indeed, it benefited most from the calibration approaches, particularly our two calibration techniques based on popularity.

5.1.2 Genre Mean Rank Miscalibration

Figure 3 shows the result of **MRMC** related to genres on the Yahoo Movies dataset. Notably, when $\lambda \geq 0.1$, all methods increased the fairness related to genres, whereas calibration by genre solely performed the best as it was initially designed to provide fairness according to the genres.

5.1.3 Popularity Mean Rank Miscalibration

Figure 4 shows the results of **MRMC** related to popularity on the Yahoo Movies dataset. Our proposed method, Popularity Calibration, outperformed fair-

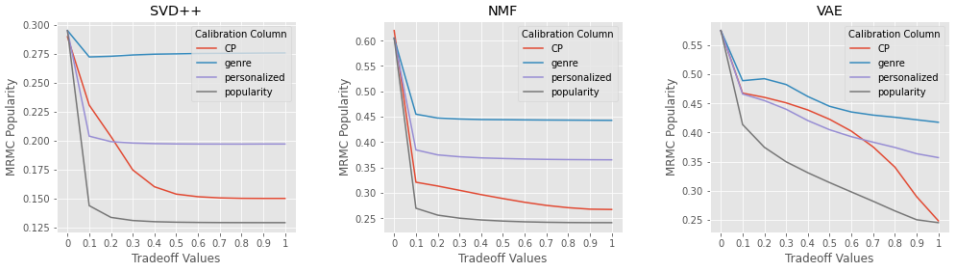

Figure 4: MRMC of popularity results over the selected recommenders and three types of calibration on the Yahoo Movies dataset. The results are statistically significant (p-value < 0.01).

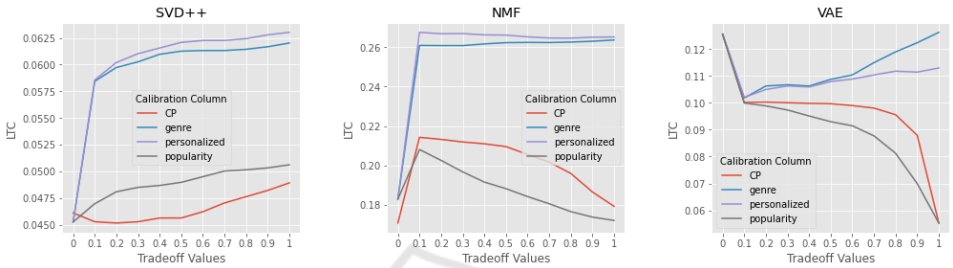

Figure 5: LTC results over the selected recommenders and four types of calibration on the Yahoo Movies dataset. The results are not statistically significant.

ness associated with popularity in all recommenders, including the CP (Abdollahpouri et al., 2021) calibration method.

5.1.4 Long Tail Coverage

Figure 5 shows the result of **LTC** on the Yahoo Movies dataset. We observed that genre and personalized calibrations were the best performers in increasing the discovery of less popular items in all recommenders. It demonstrates that, although we achieved better results on MRR, as shown in Figure 2, our personalized calibration was still able to provide diversity by covering the long tail as much as genre calibration. In turn, although it achieved better results on MRR, popularity calibration performed worse on the long tail coverage, as it does not consider the genres for calibrating the recommendation list.

5.1.5 Baselines Comparison

Table 3 compares our proposed Personalized Calibration method against the four popular debiasing methods described in Subsection 4.4. An important aspect is the ability of our approach to be used with any recommender model, depending on application requirements. NMF and SVD++ associated with our calibration method achieved the best results in LTC and MRMC of Genres and Popularity, respectively, indicating better diversity and fairness.

As explained in Section 4.2, the ΔGAP metric indicates how much the recommendation average popularity increases or decreases based on the users' profile. So, when $\Delta GAP = 0$, the system balances the items' popularity and the user profile. Still, when $\Delta GAP > 0$, it indicates the system is recommending more popular movies than expected, in other words, increasing the popularity bias in the system.

Analyzing the ΔGAP for the BlockBuster (BB) and Diverse (D) groups, we note that PDA is the only one that recommended more popular items than expected (popularity bias). On the other hand, the other methods, including ours, recommended less popular movies. For users who like unpopular items (the Niche (N) group), PDA and PairWise recommended too many popular items, whereas the other methods, including ours, suggested less popular items to these users.

Regarding MAP and MRR, we acknowledge the better results of PairWise and PDA against all our combinations. However, these methods were responsible to recommend the most biased and unfair popular items, in particular for the Niche group of users. Our methods, on the other hand, were capable to recommend items calibrated by popularity and genres, resulting in fairer suggestions.

Table 3: Comparison of our proposed Personalized Calibration against baselines in the Yahoo Movies dataset. The results of comparing our proposed methods with other methods are statistically significant. (LTC: p-value < 0.01; MRMC Genre: p-value < 0.01; GAPBB: p-value < 0.01; GAPN: p-value < 0.05; GAPD: p-value < 0.01).

Algorithm	MAP	MRR	MRMC Genre	MRMC Pop.	LTC	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}
MF MACR (Wei et al., 2021)	0.01	0.02	0.35	0.40	0.13	-0.93	-0.57	-0.86
PairWise (Boratto et al., 2021)	0.04	0.09	0.50	0.33	0.14	-0.66	1.33	-0.36
PDA (Zhang et al., 2021)	**0.16**	**0.35**	0.43	0.29	0.12	0.09	2.7	1.13
SVD++ + CP ($\lambda = 0.9$) (Abdollahpouri et al., 2021)	0.016	0.042	0.48	0.22	0.05	-0.768	-0.174	-0.546
NMF + CP ($\lambda = 0.1$) (Abdollahpouri et al., 2021)	0.004	0.011	0.51	0.31	0.21	-0.937	-0.708	-0.860
VAE + CP ($\lambda = 0.2$) (Abdollahpouri et al., 2021)	0.001	0.001	0.48	0.46	0.10	-0.988	-0.863	-0.972
SVD++ + Personalized ($\lambda = 0.8$)	0.028	0.073	**0.26**	**0.19**	0.06	-0.749	-0.188	-0.611
NMF + Personalized ($\lambda = 0.8$)	0.006	0.018	0.31	0.38	**0.27**	-0.927	-0.718	-0.899
VAE + Personalized ($\lambda = 1$)	0.001	0.003	0.30	0.35	0.11	-0.956	-0.803	-0.941

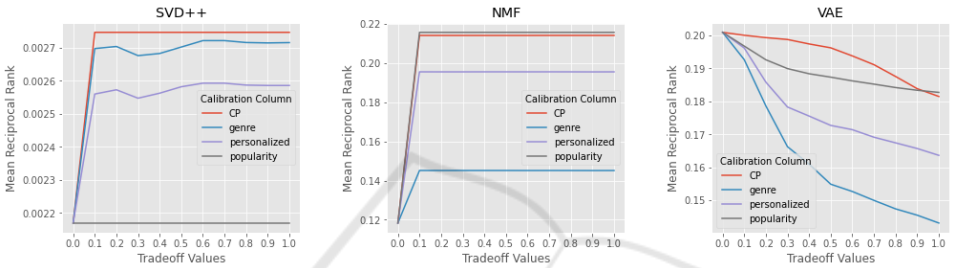

Figure 6: MRR results over the selected recommenders and three types of calibration on the MovieLens 20M dataset. The results are statistically significant (p-value < 0.01).

5.2 MovieLens 20M

5.2.1 Mean Reciprocal Rank

Figure 6 shows the comparison of **MRR**. Our method achieves a lower but competitive MRR increase according to the trade-off values. This behavior was expected, as our methods propose increasing fairness and reducing popularity bias.

5.2.2 Genre Mean Rank Miscalibration

Figure 7 shows the comparison of **MRMC** related to genres. Our methods could not achieve the best values compared to genre calibration. However, they were still able to increase fairness in genres compared to the CP method.

5.2.3 Popularity Mean Rank Miscalibration

Figure 8 shows the **MRMC** results related to popularity. Our proposed methods increased the popularity fairness in NMF and VAE, whereas CP achieved the best fairness with the SVD++ model.

5.2.4 Long Tail Coverage

Figure 9 shows the results of **LTC** on the MovieLens 20M dataset. It is possible to note a significant increase of SVD++ associated with Personalized Calibration compared to genre calibration. For the other models, there was no statistical significance. Similar to what occurred in the Yahoo Movies dataset for this metric, although we obtained very similar results as the genre calibration, we notice our method also achieved the best results in MRR, counteracting the inverse relationship between diversity and precision (Landin et al., 2018).

5.2.5 Baselines Comparison

Table 4 compares our proposed Personalized Calibration method against the four state-of-the-art popularity debiasing methods described in Subsection 4.4. SVD++ and NMF associated with our calibration method achieved the best results in MRMC of Genres and Popularity, indicating better fairness of genres and popularity; SVD++, in particular, obtained the highest LTC, indicating better discovery of unpopular items. Regarding MAP and MRR, PDA achieved the best results than competitors, but at the cost of suggesting popular items regardless of the users' preferences.

Analyzing the ΔGAP, NMF+CP (Abdollahpouri et al., 2021) and the proposed NMF+Personalized were the combinations that provided the fairest recommendations according to each group of user

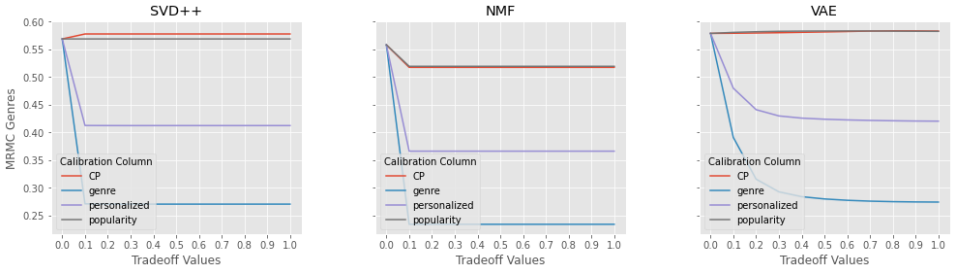

Figure 7: MRMC of genres results over the selected recommenders and three types of calibration on the MovieLens 20M dataset.

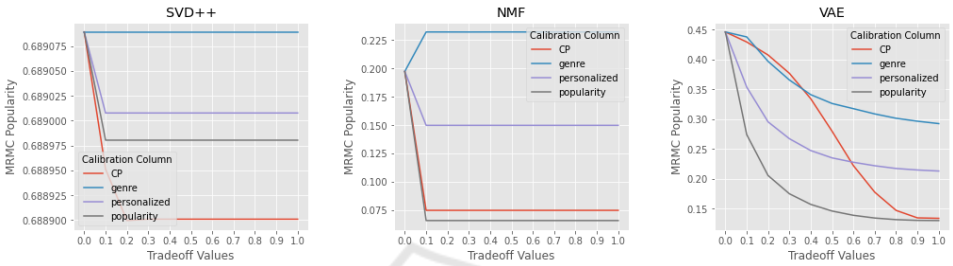

Figure 8: MRMC of popularity results over the selected recommenders and three types of calibration on the MovieLens 20M dataset. The results are statistically significant (p-value <0.01).

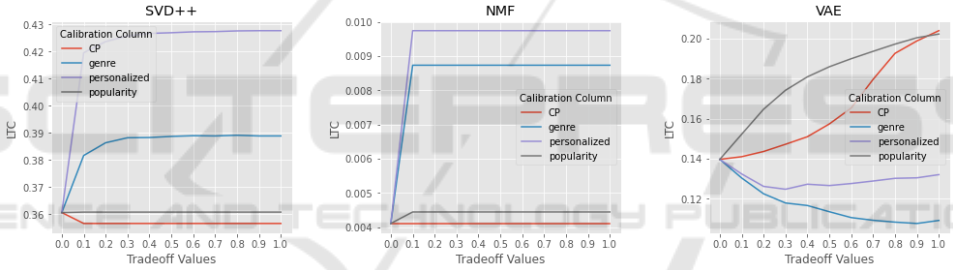

Figure 9: LTC results over the selected recommenders and three types of calibration on the MovieLens 20M dataset. The results are not statistically significant.

Table 4: Comparison of our proposed method with baseline methods in MovieLens 20M dataset. The results comparing our proposed methods with other methods are statistically significant. (LTC: p-value < 0.10; GAPBB: p-value < 0.01; GAPN: p-value < 0.05; GAPD: p-value < 0.01), except comparing with MF MACR where the LTC result is not significant.

Algorithm	MAP	MRR	MRMC Genre	MRMC Pop.	LTC	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}
MF MACR (Wei et al., 2021)	0.01	0.02	**0.35**	0.40	0.13	-0.930	-0.570	-0.860
PairWise (Boratto et al., 2021)	0.04	0.09	0.50	0.33	0.14	-0.660	1.330	-0.360
PDA (Zhang et al., 2021)	**0.16**	**0.35**	0.43	0.29	0.12	0.090	2.700	1.130
SVD++ + CP ($\lambda = 0.9$) (Abdollahpouri et al., 2021)	0.001	0.002	0.56	0.68	0.35	-0.991	-0.976	-0.987
NMF + CP ($\lambda = 0.1$) (Abdollahpouri et al., 2021)	0.084	0.213	0.51	**0.11**	0.01	-0.136	0.145	-0.118
VAE + CP ($\lambda = 1$) (Abdollahpouri et al., 2021)	0.072	0.181	0.58	0.13	0.20	0.425	0.401	0.587
SVD++ + Personalized ($\lambda = 0.9$)	0.001	0.002	0.41	0.69	**0.42**	-0.992	-0.970	-0.985
NMF + Personalized ($\lambda = 0.9$)	0.077	0.195	**0.35**	**0.11**	0.01	-0.171	0.238	-0.212
VAE + Personalized ($\lambda = 0.1$)	0.079	0.196	0.47	0.35	0.14	0.850	1.553	1.386

(Δ*GAP* close to zero). However, although NMF+CP was able to achieve higher MAP and MRR than our proposal, we obtained higher MRMC Genre, indicating that our method can provide genre and popularity calibration at the same time[9].

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed a personalized calibration technique, which uses popularity and genre calibrations in a switch-based approach to provide fairer recommendations to users according to their interests. Our main contribution is the possibility to calibrate recommendations generated by any recommender model, whose choice could be according to application requirements.

We showed that the calibration of items based on the popularity aspect is a way to improve a recommendation system to bring fairer recommendations to users to meet their preferences and reduce the impact of popularity bias in the system. We presented a calibration approach that works in the post-processing step and is independent of any recommendation algorithm.

Our experiments showed that our proposal could reduce the popularity bias, recommending less popular items by covering the long tail and consequently increasing diversity and fairness related to genres and popularity in recommendations. Although we achieved better results in precision against genre calibration, other methods provided more accurate recommendations, but at the cost of higher popularity bias.

In future work, we plan to analyze the effect of our calibration with other recommendation models, particularly considering the aspects of precision and popularity bias. We will also conduct online experiments to verify the performance of the proposed calibration system with real users. In addition, we plan to evaluate our approaches with other metadata and different contexts.

ACKNOWLEDGEMENTS

The authors would like to thank the financial support from FAPESP, process number 2022/07016-9.

[9]In this comparison, we selected the $\lambda \neq 0$ with the highest LTC value.

REFERENCES

Abdollahpouri, H., Burke, R., and Mobasher, B. (2018). Popularity-aware item weighting for long-tail recommendation. arXiv.

Abdollahpouri, H., Mansoury, M., Burke, R., and Mobasher, B. (2019a). The impact of popularity bias on fairness and calibration in recommendation.

Abdollahpouri, H., Mansoury, M., Burke, R., and Mobasher, B. (2019b). The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*.

Abdollahpouri, H., Mansoury, M., Burke, R., and Mobasher, B. (2020). The connection between popularity bias, calibration, and fairness in recommendation. In *Fourteenth ACM conference on recommender systems*, pages 726–731.

Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., and Malthouse, E. (2021). User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 119–129.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al. (2019). Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2212–2220.

Boratto, L., Fenu, G., and Marras, M. (2021). Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management*, 58(1):102387.

Borges, R. and Stefanidis, K. (2021). On mitigating popularity bias in recommendations via variational autoencoders. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 1383–1389, New York, NY, USA. Association for Computing Machinery.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307.

da Silva, D. C., Manzato, M. G., and Durão, F. A. (2021). Exploiting personalized calibration and metrics for fairness recommendation. *Expert Systems with Applications*, 181:115112.

de Winter, J. (2013). Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, 18.

Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., and Pera, M. S. (2018). All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In Friedler, S. A. and Wilson, C., editors, *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 172–186. PMLR.

Geyik, S. C., Ambler, S., and Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 2221–2231.

Hug, N. (2020). Surprise: A python library for recommender systems. *Journal of Open Source Software*, 5(52):2174.

Kaya, M. and Bridge, D. (2019). A comparison of calibrated and intent-aware recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 151–159.

Koren, Y. (2008). Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 426–434, New York, NY, USA. Association for Computing Machinery.

Landin, A., Suárez-García, E., and Valcarce, D. (2018). When diversity met accuracy: A story of recommender systems. *Proceedings*, 2(18).

Lesota, O., Melchiorre, A., Rekabsaz, N., Brandl, S., Kowald, D., Lex, E., and Schedl, M. (2021). Analyzing item popularity bias of music recommender systems: Are different genders equally affected? In *Fifteenth ACM Conference on Recommender Systems*, pages 601–606.

Liang, D., Krishnan, R. G., Hoffman, M. D., and Jebara, T. (2018). Variational autoencoders for collaborative filtering.

Luo, X., Zhou, M., Xia, Y., and Zhu, Q. (2014). An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10(2):1273–1284.

Seymen, S., Abdollahpouri, H., and Malthouse, E. C. (2021). A constrained optimization approach for calibrated recommendations. In *Fifteenth ACM Conference on Recommender Systems*, pages 607–612.

Steck, H. (2018). Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 154–162, New York, NY, USA. Association for Computing Machinery.

Verma, S., Gao, R., and Shah, C. (2020). Facets of fairness in search and recommendation. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 1–11. Springer.

Wang, W., Feng, F., He, X., Wang, X., and Chua, T.-S. (2021). Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1717–1725.

Wei, T., Feng, F., Chen, J., Wu, Z., Yi, J., and He, X. (2021). Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1791–1800, New York, NY, USA. Association for Computing Machinery.

Yalcin, E. (2021). Blockbuster: A new perspective on popularity-bias in recommender systems. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 107–112. IEEE.

Yalcin, E. and Bilge, A. (2021). Investigating and counteracting popularity bias in group recommendations. *Information Processing & Management*, 58(5):102608.

Zhang, Y., Feng, F., He, X., Wei, T., Song, C., Ling, G., and Zhang, Y. (2021). Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Zhu, Z., Wang, J., and Caverlee, J. (2020). Measuring and mitigating item under-recommendation bias in personalized ranking systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 449–458, New York, NY, USA. Association for Computing Machinery.