Modular Debiasing of Latent User Representations in Prototype-based Recommender Systems

Alessandro B. Melchiorre[1,2][0000−0003−1643−1166] ✉, Shahed Masoudian[1,2][0009−0007−2747−0386], Deepak Kumar[1,2][0000−0002−4828−8328], and Markus Schedl[1,2][0000−0003−1706−3406]

[1] Johannes Kepler University Linz, Austria
[2] Linz Institute of Technology, AI Lab, Austria
{first_name.second_name}@jku.at

Abstract. Recommender Systems (RSs) may inadvertently perpetuate biases based on protected attributes like gender, religion, or ethnicity. Left unaddressed, these biases can lead to unfair system behavior and privacy concerns. Interpretable RS models provide a promising avenue for understanding and mitigating such biases. In this work, we propose a novel approach to debias interpretable RS models by introducing user-specific scaling weights to the interpretable user representations of prototype-based RSs. This reduces the influence of the protected attributes on the RS's prediction while preserving recommendation utility. By decoupling the scaling weights from the original representations, users can control the degree of invariance of recommendations to their protected characteristics. Moreover, by defining distinct sets of weights for each attribute, the user can further specify which attributes the recommendations should be agnostic to. We apply our method to ProtoMF, a state-of-the-art prototype-based RS model that models users by their similarities to prototypes. We employ two debiasing strategies to learn the scaling weights and conduct experiments on ML-1M and LFM2B-DB datasets aiming at making the user representations agnostic to age and gender. The results show that our approach effectively reduces the influence of the protected attributes on the representations on both datasets, showcasing flexibility in bias mitigation, while only marginally affecting recommendation quality. Finally, we assess the effects of the debiasing weights and provide qualitative evidence, particularly focusing on movie recommendations, of genre patterns identified by ProtoMF that correlate with specific genders.

Keywords: Recommender Systems · Debiasing · Interpretability

1 Introduction

Recommender Systems (RSs) typically operate as black boxes trained on large collections of user-item interactions to generate recommendations. Through this

training process, they capture underlying interaction patterns, revealing which users prefer which content, to better model the users' interests. Alas, the observed user behavior might correlate with particular protected user attributes such as gender, age, ethnicity, or religion, even when these are not explicit in the data [12]. When exposed to such data, the RS can encode these correlations in the user representations, potentially leading to biased predictions [17, 46, 33], unfair system behavior across protected groups [11, 20], and the strengthening of per-group "filter bubbles" [13, 29]. Furthermore, they can also raise privacy concerns regarding the disclosure of sensitive information from the representations [36, 4].

Interpretable RS models can be leveraged to understand how these biases manifest in the data [19, 3] and how they are assimilated by the RS [38, 15]. In recent years, several RS models that offer interpretable user representations have emerged. Specifically, each dimension of these representations usually corresponds to an interpretable aspect, such as the user's sentiment towards items' attributes [45], user or item features [16, 37], or similarity to prototypical users/items [1, 34]. These transparent models can assist in defining potential corrective measures. For instance, if a particular dimension strongly correlates with a user's protected attribute, we can choose to weaken it and use the updated representations to generate debiased recommendations. Alternatively, we may also amplify another dimension associated with a different value of the protected attribute, thereby increasing the ambiguity surrounding the true user's attribute. However, determining which dimensions are indicative of the attribute in the first place can still pose challenges.

One solution, explored in recent literature, is to adapt the user representations, predominantly through in-processing techniques [9, 11]. These methods involve (re-)training a RS model to provide relevant recommendations while also optimizing a debiasing objective that attempts to make the predictions invariant to the user's protected attributes, albeit with a trade-off in performance [29, 4, 46, 17]. However, depending on the user's preferences, the context, or bias-utility trade-off considerations, end-users might in practice still prefer to receive *some* recommendations from the original (potentially biased) model. Especially when different users have different attitudes towards their biased representations (e. g., users conforming to stereotypical norms may prefer biased predictions), or when the same user prioritizes having their recommendations unbiased with respect to certain attributes but not others [29]. Accommodating all these scenarios with current approaches can be burdensome, as it requires training a separate RS for every protected attribute and according to each user's request.

In contrast, we propose learning separate user-specific scaling weights that can be applied to the interpretable user representations of a pre-trained RS model. These modular weights automatically adjust the representations to reduce their biases associated with a protected attribute, e. g., gender or age, while still preserving relevant recommendations. This concept aligns with recent efforts in the NLP community focused on modular bias mitigation, enabling end-users to choose whether their results should be biased or unbiased on-demand [25, 32,

26]. Our method allows to flexibly cover different users' needs. By keeping the weights separate from the original representations, users can decide during inference whether their recommendations should be influenced by their protected attributes, applying the scaling weights as needed. Additionally, by training distinct sets of weights for each attribute of interest, users can further specify with respect to which of them the recommendations should be agnostic.

We apply our approach to the recently proposed ProtoMF model [34], a prototype-based RS designed to capture specific item-consumption characteristics of the data through the concept of user/item prototypes [27, 23]. We select this model for its ability to provide relevant and explainable recommendations; nevertheless, our method can be applied to any interpretable RS that provides interpretable user representations. Within ProtoMF, each user is mapped to a representation where each dimension indicates the similarity between the user and a specific user prototype. The application of the scaling weights, hence, tunes these similarities, thereby influencing the impact of the prototype's pattern on the resulting recommendations. Consequently, analyzing which dimensions are attenuated (< 1) or amplified (> 1) by the debiasing strategy aids us in interpreting which consumption patterns might be correlated (and in which way) to a specific protected attribute.

We evaluate our method on two popular datasets of movie ratings (ML-1M [24]) and music listening records (LFM2B-DB [35]). For both datasets, we learn user representations that are less affected by the user's gender and age, which aligns with the concepts of *representational fairness* [40] or *demographic parity* [2]. Intuitively, if the representations are invariant to these attributes, predictions based on these representations will also be invariant, resulting in less biased recommendations [2]. For instance, the RS will avoid recommending only Romance movies to female users. Our approach is agnostic to the debiasing objective, allowing the scaling weights to be trained with any gradient descent-based signal that ensures representation invariance to a specific user's attribute. In this study, we investigate two debiasing objectives: Maximum Mean Discrepancy [22] and Adversarial Debiasing [43, 21]. To assess the effectiveness of our approach to mitigate bias, we follow the standard evaluation framework for debiasing [14, 32, 26, 29] and report the performance of an external probe network trained to predict the protected attribute from the user representations. Compared to the original user representations, our results show that our proposed method effectively impairs the probe's ability to recover sensitive information, resulting in a substantial reduction in bias while only marginally affecting recommendation performance. Finally, we investigate the effect of the debiasing weights and showcase for ML-1M the genre patterns captured by ProtoMF that are correlated with gender. Our code and settings are publicly available at https://github.com/hcai-mms/modprotodebias.

2 Related Work

Our research is influenced by recent works in NLP. Thus, we review pertinent literature in this field before delving into related research on debiasing RSs.

Bias Mitigation in Natural Language Processing. Extensive research has addressed societal biases within Language Models (LMs), particularly focusing on *attribute erasure* [33]. This involves reducing the influence of protected attributes within LM's embeddings to mitigate empirical biases [33, 40] or achieve representational fairness [14]. Recent studies explore *modular* bias mitigation, enabling end-users to select between biased or bias-mitigated models for individual queries. In particular, Hauzenberger et al. [25] learn a set of sparse additive weights that mitigate societal bias when added to the original model. Kumar et al. [26] leverage adapters [39] to isolate the sensitive information in separate blocks of the LM. Masoudian et al. [32] introduce controllable gates to scale LM's representations to switch between biased/unbiased predictions. Inspired by these studies, our work introduces separate per-user scaling weights to adjust user representations for unbiased recommendations.

Bias Mitigation in Recommender Systems. Being multi-sided platforms, RSs' outcomes may be prone to biases associated with the users [8, 42] and items [5, 44]. While there are several strategies to mitigate these biases and increase the RSs' fairness, recent literature especially focuses on in-processing techniques [11, 12]. Zhu et al. [47] tackle the issue of item under-recommendation from imbalanced train data and propose a regularization objective based on fairness. Similarly from the user side, Li et al. [28] propose a novel RS model to learn user/item representations that avoid unfairly penalizing non-mainstream users. Several studies focus on removing spurious correlations between users' protected attributes and recommendations by leveraging adversarial learning, albeit at some performance trade-off. For instance, Bose and Hamilton [6] and Wu et al. [42], learn user/item representations in graph-based RSs that are invariant to the user's protected attribute. Ganhör et al. [17] adapt Mult-VAE [30] to generate recommendations agnostic to users' gender. Li et al. [29] simultaneously train a set of filters, one for each attribute, as well as the underlying RS, to satisfy different users' fairness demands. Some authors also leveraged interpretable models to assess fairness issues in RSs. Ge et al. [19] employ counterfactual learning to learn the minimal change to the input features of a feature-aware RS to address item exposure unfairness in the recommendations. Fu et al. [15] present a fairness re-ranking approach to decrease performance disparity between active/inactive users in explainable recommendations over knowledge graphs. Our work complements the above studies by addressing the influence of users' protected attributes on the recommendations of a *pre-trained* RS, leveraging modular scaling weights on the interpretable user representations concerning user prototypes.

3 Methodology

Let $\mathcal{U} = \{u_i\}_{i=1}^{N}$ and $\mathcal{T} = \{t_j\}_{j=1}^{M}$ denote the set of N users and M items, respectively. We assume that we only have access to the implicit interaction data $\mathcal{I} = \{(u_i, t_j)\}$, where (u_i, t_j) indicates that user u_i has interacted with item t_j. Additionally, each user u_i is associated with one or more protected attributes $g \in G$. We omit the user and item indexes for brevity. Let $rec(\cdot, \cdot)$ be an interpretable RS model that, beyond scoring each user-item pair $rec(u, t) \in \mathbb{R}$, also maps each user u to an intermediate *interpretable* representation $\boldsymbol{u} \in \mathbb{R}^d$. This representation may align with various aspects, such as user's sentiment towards items' attributes [45], user or item features [16, 37], or similarity to prototypical users and items derived from the dataset [1, 34]. In this work, we focus on the latter and particularly the recently proposed ProtoMF model [34] as it showcased high accuracy in the recommendation task. Nevertheless, our method can be applied to any RS offering interpretable user representations. Within ProtoMF, each dimension $\{i\}_{i=1}^{d}$ in $\boldsymbol{u}$ indicates the similarity of user u to a specific user prototype $\boldsymbol{p}^i$, representing item-consumption characteristics of the data, with similarity values in the range $(0, 2)$. As shown next, the interpretable representation $\boldsymbol{u}$ may encode the protected user attribute g, despite the information not being explicitly provided to the RS. As a consequence, the RS can pick up this information and bias its predictions, as also shown in [17, 46, 29].

To address this issue, we define a vector of scaling weights $\boldsymbol{\omega}_u \in \mathbb{R}^d$ for each user, which can be plugged in at will. Starting from the $\boldsymbol{u}$ representation obtained from the pre-trained RS model, we derive a new user representation $\tilde{\boldsymbol{u}}$ as follows:

$$\tilde{\boldsymbol{u}} = \boldsymbol{u} \odot \boldsymbol{\omega}_u$$

where $\odot$ is the Hadamard product. We leave the original user representation $\boldsymbol{u}$ (as well the other model parameters) unchanged while we only optimize $\boldsymbol{\omega}$ so that the new representation $\tilde{\boldsymbol{u}}$ remains relevant for the recommendation task while becoming invariant to the protected attribute g. The optimization involves minimizing a recommendation loss $\mathcal{L}_{\text{rec}}$ as well as a debiasing objective $\mathcal{L}_{\text{debias}}$:

$$\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}} \mathcal{L}_{\text{rec}}(\mathcal{I}, \boldsymbol{\omega}) + \lambda \mathcal{L}_{\text{debias}}(\mathcal{I}, \boldsymbol{\omega}, g)$$

where the hyperparameter λ adjusts the strength of the debiasing loss. As recommendation loss $\mathcal{L}_{\text{rec}}$, we adopt the same loss function as the base RS model. In the case of ProtoMF [34], this corresponds to the cross-entropy loss reported below for reference:

$$\mathcal{L}_{\text{rec}} = -\sum_{(u,t)\in\mathcal{I}} \ln p(t|u), \quad p(t|u) = \frac{e^{rec(u,t)}}{\sum_j e^{rec(u,t_j)}} \tag{1}$$

The debiasing objective $\mathcal{L}_{\text{debias}}$ operates on the representations $\tilde{\boldsymbol{u}}$ and the corresponding protected attribute label g to realize invariance. Our approach is agnostic to the debiasing objective, allowing the scaling weights to be trained

with any gradient descent-based signal that ensures representation invariance. In our work, we employ two prominent debiasing strategies: Maximum Mean Discrepancy (MMD) [22, 25] and Adversarial Debiasing (Adv.) [43, 14].

Maximum Mean Discrepancy (MMD) [22] aims to minimize the distribution shift between the representations of a specific protected attribute g. Effectively, given the set of users $\mathcal{U}$ split into two subsets $\mathcal{U}_g^A$ and $\mathcal{U}_g^B$ according to the values of a binary[3] protected attribute g, MMD minimizes the mean distance between the user representations $\tilde{\boldsymbol{u}}$ of the two subgroups:

$$\mathcal{L}_{\text{debias}} = \left\| \frac{1}{|\mathcal{U}_g^A|} \sum_{i \in \mathcal{U}_g^A} \phi(\tilde{\boldsymbol{u}}_i) - \frac{1}{|\mathcal{U}_g^B|} \sum_{j \in \mathcal{U}_g^B} \phi(\tilde{\boldsymbol{u}}_j) \right\|_2^2 \tag{2}$$

where ϕ is a feature map kernel defined as a sum of multiple Gaussian kernels.

Adversarial Debiasing (Adv.) [21, 43] is a common approach in learning input representations that are informative for the task while remaining invariant to specific traits of the data [14, 17]. In our context, each user is passed through an adversarial head $h(\cdot)$ that aims to infer the protected attribute g from $\tilde{\boldsymbol{u}}$ by leveraging the cross-entropy loss $\mathcal{L}_{debias}(\tilde{\boldsymbol{u}}, g) = \mathcal{L}_{\text{CE}}(\tilde{\boldsymbol{u}}, g)$. During training, we aim to learn scaling weights $\boldsymbol{\omega}$ that maintain relevant user recommendations while hindering the adversary's predictive ability. This objective is commonly solved as a minimization task by inserting a gradient reversal layer $grl(\cdot)$ between the adversary and the rest of the model [18, 43]. Essentially, during back-propagation, the $grl(\cdot)$ negates and potentially scales the gradients flowing from the adversary to the weights, pushing the $\boldsymbol{\omega}$ in the opposite direction desired by the adversary. This allows us to formulate the debiasing objective as:

$$\mathcal{L}_{\text{debias}} = \mathcal{L}_{\text{CE}}(grl(\tilde{\boldsymbol{u}}), g) \tag{3}$$

Finally, given the learned scaling weights $\boldsymbol{\omega}$, we derive the adjusted user representations $\tilde{\boldsymbol{u}}$, which are used by the RS model to provide item recommendations that are both relevant and agnostic to the user's protected attribute. Our proposed approach offers a flexible and informative method for debiasing. By keeping the weights separate from the original representations, users can decide during inference whether their recommendations should be influenced by their protected attributes, applying the scaling weights as needed. Using distinct sets of weights for each attribute of interest, users can further specify with respect to which of them the recommendations should be agnostic. Moreover, by analyzing which interpretable dimensions are attenuated ($\omega < 1$) or amplified ($\omega > 1$), we can assess which consumption patterns might be correlated (and in which way) to a specific protected attribute.

ML-1M		Users	Interactions	Items	**LFM2B-DB**		Users	Interactions	Items
All		6,034	574,376		All		16,258	2,339,540	
Gender	M	4,326	429,039		Gender	M	12,734	1,981,006	
	F	1,708	145,337			F	3,524	358,534	
Age	< 18	222	15,583	3,125	Age	≤ 18	1,811	232,942	99,824
	18-24	1,100	100,655			19-32	12,613	1,797,291	
	25-34	2,095	222,242						
	35-44	1,192	116,507			33-39	1,126	184,176	
	45-49	550	49,400						
	50-55	496	44,979			> 40	708	125,131	
	> 56	379	25,010						

Table 1: Statistics of the datasets used in our experiments.

4 Experiment Setup

Datasets. We use two standardized datasets containing user-item interactions along with partial user's demographic: **(1) MovieLens-1M**[4] **(ML-1M)** [24] contains the ratings of users on movies as well as user's gender, age group, and occupation. As common [34, 30], we treat high movie ratings (> 3.5 on a 1-5 scale) as positive interactions while discarding the rest, and perform 5-core filtering. **(2) LFM2B-DemoBias (LFM2B-DB)** [35] is a sub-set of the LFM2B[5] dataset, which provides a collection of music listening records of users for whom partial demographic information (i. e., gender, age, country) is available. We follow the same data processing methodology as in Melchiorre et al. [35]. Specifically, we keep user-item interactions with a minimum play count of two and binarize the interactions. Additionally, to accommodate computational constraints, we randomly sample $100,000$ tracks from the large catalog and apply 5-core filtering. Furthermore, we split users into age groups based on their deviation from the mean age ($\mu = 24.87$, $\sigma = 7.30$) by multiples of σ.

Table 1 offers a detailed summary of the dataset statistics, including the breakdown by user attribute. With both datasets, we focus on the gender and age of the user as protected attributes in our experiments.[6]

Data Splits. To train both the underlying RS model and the scaling weights, we employ the leave-k-out strategy [10] for every user. Specifically, for each user, we sort their item interactions according to the timestamps (keeping the earliest interaction if multiple ones with the same item exist). The last 10% interactions of the users are used as test, while the penultimate 10% as validation set. The remaining interactions constitute the training set. During training, for

[3] We consider majority vs. all others subsets for non-binary attributes.
[4] https://grouplens.org/datasets/movielens/1m/
[5] http://www.cp.jku.at/datasets/LFM-2b/
[6] Both datasets provide gender in binary form, neglecting nuanced gender definitions.

each positive user-item interaction, we randomly sample 10 negative items not interacted with by the user. We scale both the adversary's and, later, the probe's loss, ensuring that data points from all user groups contribute equally. This balancing not only aids debiasing [35] but also prevents both classifiers from solely predicting the majority class [14].

ProtoMF Pre-Training. We follow a similar training procedure as the original UI-ProtoMF paper does [34], referred here simply as ProtoMF. We train the model for 50 and 100 epochs on ML-1M and LFM2B-DB, respectively, with the AdamW optimizer [31]. We perform early-stopping if the accuracy on the validation set does not improve for 5 consecutive epochs. After preliminary experiments, we set the number of user prototypes based on the dataset (42 for ML-1M and 64 for LFM2B-DB) and fix the batch size to 256. We then carry out a comprehensive search for optimal embedding sizes and loss-related hyperparameters. Details on the range of hyperparameters explored, as well as those selected for the final models, are provided in the appendix. Once we identify the model that achieves the highest accuracy on the validation set, we freeze its parameters and only update the scaling weights during debiasing.

Evaluation. To assess the effectiveness of our approach to bias mitigation, we follow the standard evaluation framework [14]. Specifically, after freezing the model's parameters (including the $\boldsymbol{\omega}$), we train a probe network to predict the protected attributes from the user representations. We measure the accuracy (Acc) and balanced accuracy (BAcc) when predicting the users' gender and age. Particularly, we focus on the BAcc metric [7] as it is well-suited for imbalanced datasets. BAcc reports the average recall per user group, where a value of $\frac{1}{\#\text{Groups}}$ represents a fully debiased representation which amounts to .50 for gender on both datasets and $\frac{1}{7} = .14$ and $\frac{1}{4} = .25$ for age on ML-1M and LFM2B-DB, respectively. To evaluate recommendation performance, we use Normalize Discount Cumulative Gain (NDCG), specifically NDCG@10. We report performance and bias mitigation results as average computed on the test set for three seeds.

Debiasing and Probing. For the MMD method, we use a batch size of 128 and set the learning rates to $5e^{-5}$ for gender and $5e^{-4}$ for age on both datasets. In the Section 5, we explore different values of λ. Regarding adversarial debiasing, instead, we employ a two-layer neural network with 512 neurons as an adversary network. We investigate the impact of using multiple adversarial networks, i. e., adversarial heads, by averaging their debiasing losses [14]. We use a batch size of 512 for ML-1M and 1024 for LFM2B-DB, with a learning rate of $5e^{-5}$, adjusting λ based on the dataset and attribute.[7] Our probe is a two-layer neural network with 128 neurons in the hidden layer. We set the learning rate and weight decay based on the probe's performance on each dataset and attribute. After debiasing, we train a new probe using the debiased user representations while keeping the scaling weights (and the base model) unchanged. Finally, we

[7] $\lambda = 1$ on LFM2B-DB, $\lambda = 5$ and $\lambda = 10$ on ML-1M for gender and age respectively.

initialize the $\boldsymbol{\omega}$ by sampling from the normal distribution $\mathcal{N}(1, .01^2)$ and train them, as well as the probe, for 25 epochs using the AdamW optimizer [31].

5 Results and Analysis

Dataset	Attribute	Debiasing	Bias ↓ BAcc	Bias ↓ Acc	NDCG ↑
ML-1M	Gender	None	$.789_{003}$	$.788_{001}$	$.0625_{0000}$
		MMD	$\mathbf{.542_{003}}$	$\mathbf{.497_{025}}$	$.0618_{0000}$
		Adv.	$\mathbf{.542_{008}}$	$.608_{047}$	$\mathbf{.0620_{0000}}$
	Age	None	$.465_{002}$	$.424_{003}$	$.0625_{0000}$
		MMD	$\mathbf{.232_{001}}$	$.207_{004}$	$.0594_{0002}$
		Adv.	$\mathbf{.232_{026}}$	$\mathbf{.193_{011}}$	$\mathbf{.0618_{0001}}$
LFM2B-DB	Gender	None	$.723_{002}$	$.718_{004}$	$.0754_{0000}$
		MMD	$\mathbf{.536_{004}}$	$\mathbf{.397_{078}}$	$.0745_{0001}$
		Adv.	$.600_{011}$	$.636_{065}$	$\mathbf{.0755_{0001}}$
	Age	None	$.581_{005}$	$.504_{008}$	$.0754_{0000}$
		MMD	$\mathbf{.299_{003}}$	$\mathbf{.238_{055}}$	$.0626_{0001}$
		Adv.	$.390_{011}$	$.277_{018}$	$\mathbf{.0755_{0001}}$

Table 2: Debiasing and performance results on both datasets and attributes. We highlight the least biased and best-performing values among Adv. and MMD. Subscripts indicate the standard deviation.

General Results. Table 2 reports the results of the debiasing methods and recommendation utility across datasets and attributes. We highlight in bold the best RS performance (highest NDCG) and best debiasing performance (lowest BAcc of the probing network). Results are computed on the test set as the average of 3 random seeds, with subscripts indicating the standard deviation.

When no debiasing is applied (*None* rows in Table 2), we observe that the users' protected attributes can be predicted with relatively high accuracy by the probe. The BAcc for gender reaches .79 and .72 on ML-1M and LFM2B-DB datasets respectively, compared to a baseline value of .50 of a random predictor. Similar observations can be made for age; on ML-1M the probe's BAcc is .47 against the baseline of .14 and on LFM2B-DB is .58 vs. .25 in a bias-free settings. These results indicate that the user representations $\boldsymbol{u}$ learned by the RS *do* retain information about the user's protected attributes and can potentially bias the recommendations.

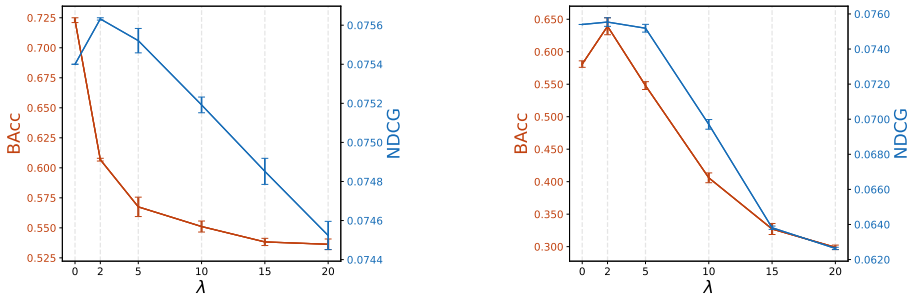
When applying the scaling weights, either learned by MMD or Adv., we observe a substantial decrease in both Acc and BAcc of the probe. This reduction spans across both attributes and datasets, indicating the effectiveness of

Dataset	Attr.	Metric		λ 0	2	5	10	15	20
ML-1M	Gender	Bias ↓	BAcc	$.789_{.003}$	$.633_{.010}$	$.574_{.010}$	$.548_{.002}$	$.548_{.003}$	$\mathbf{.542_{.003}}$
			Acc	$.788_{.001}$	$.631_{.011}$	$.557_{.027}$	$.552_{.012}$	$.514_{.030}$	$\mathbf{.497_{.025}}$
		NDCG ↑		$.0625_{.0000}$	$.0624_{.0000}$	$.0623_{.0000}$	$.0621_{.0001}$	$.0619_{.0001}$	$.0618_{.0000}$
	Age	Bias ↓	BAcc	$.465_{002}$	$.463_{.006}$	$.398_{.004}$	$.320_{.007}$	$.258_{.005}$	$\mathbf{.232_{.001}}$
			Acc	$.424_{003}$	$.356_{.003}$	$.305_{.004}$	$.252_{.005}$	$.215_{.006}$	$\mathbf{.207_{.004}}$
		NDCG ↑		$.0625_{0000}$	$.0626_{.0002}$	$.0619_{.0002}$	$.0609_{.0002}$	$.0600_{.0001}$	$.0594_{.0002}$
LFM2B-DB	Gender	Bias ↓	BAcc	$.723_{.002}$	$.607_{.001}$	$.567_{.008}$	$.551_{.005}$	$.538_{.003}$	$\mathbf{.536_{.004}}$
			Acc	$.718_{.004}$	$.588_{.003}$	$.469_{.065}$	$.428_{.074}$	$\mathbf{.357_{.014}}$	$.397_{.078}$
		NDCG ↑		$.0754_{.0000}$	$.0756_{.0000}$	$.0755_{.0001}$	$.0752_{.0000}$	$.0749_{.0001}$	$.0745_{.0001}$
	Age	Bias ↓	BAcc	$.581_{005}$	$.639_{.013}$	$.548_{.006}$	$.406_{.008}$	$.327_{.008}$	$\mathbf{.299_{.003}}$
			Acc	$.504_{008}$	$.517_{.021}$	$.406_{.010}$	$.232_{.007}$	$\mathbf{.182_{.006}}$	$.238_{.055}$
		NDCG ↑		$.0754_{0000}$	$.0755_{.0002}$	$.0752_{.0002}$	$.0697_{.0003}$	$.0638_{.0001}$	$.0626_{.0001}$

Table 3: Debiasing and performance results on both datasets and attributes using the MMD method across several λ values.

our proposed method in weakening the attribute information in the new user representations. We observe that the efficacy and the impact of the $\boldsymbol{\omega}$ depends on the dataset under scrutiny. On the ML-1M datasets, MMD and Adv. reach similar BAcc values for age and gender, both resulting in a moderate decrease in NDCG. However, Adv. shows higher capability in preserving the recommendation performance compared to MMD. On the LFM2B-DB dataset, the scaling weights learned by MMD display lower bias, although at a larger trade-off in recommendation performance. The Adv. method, on the other hand, appears to fully preserve the initial NDCG while leading to a smaller decrease in BAcc compared to MMD. Considering these results, we derive that (1) the debiased user representations, obtained by either MMD or Adv., exhibit significant decreases in the bias metrics, although the predictions are not yet fully random (e. g., Gender BAcc $> .50$), and (2) there exists a trade-off between bias reduction and recommendation accuracy whose strength depends on the dataset. We investigate the latter aspect below.

Bias vs. Performance Analysis. Considering the MMD method, we report in Table 3 the bias metrics and recommendation accuracy across different values of λ ranging from 0 (no debiasing is applied) to 20 for ML-1M and LFM2B-DB on both attributes. We plot the changes of BAcc and NDCG over the λ's for age and gender on LFM2B-DB in Figure 1. We observe that high λ values lead to a stronger debiasing of the user representations, i. e., lower BAcc, however at the cost of a moderate reduction of NDCG. We also notice that, on both datasets, making the representations agnostic to age leads to a harsher reduction of rec-



(a) Gender (b) Age

Fig. 1: BAcc and NDCG on LFM2B-DB using MMD, over varying λ values. In Fig. 1a, BAcc refers to gender, in Fig. 1b to age.

ommendation accuracy compared to debiasing for gender. Lastly, we observe for $\lambda = 2$ on the age attribute in LFM2B-DB, associated with a milder debiasing, the BAcc even increases, suggesting that without proper debiasing, more bias information can be encoded in the user representations through the $\boldsymbol{\omega}$ by the recommendation loss.

Regarding the Adv. method, preliminary experiments with the Adv. method showed that using a single adversary head while increasing λ led to unstable debiasing behavior, causing the adversary to fail and more bias to be encoded in the scaling weights. To address this, we followed previous research on adversarial debiasing [14, 32, 26] and opted to use multiple adversary heads while fixing λ based on the dataset and attribute (see Section 4). Table 4 displays bias/recommendation performance across different numbers of adversarial heads. Similarly to the MMD method, we observe lowest bias with a stronger debiasing approach, namely 20 heads. By increasing the # of heads, we see a progressive reduction in NDCG on ML-1M for both age and gender, while the recommendation performance on LFM2B-DB remains relatively constant.

In summary, we find that (1) MMD progressively reduces both bias and recommendation performance on both datasets and attributes when increasing λ, and (2) Adv. showcases a similar gradual change on ML-1M while recommendation accuracy remains relatively stable on LFM2B-DB when increasing the number of adversarial heads.

Weights Analysis. We now examine how the scaling weights affect the original user representations. As our method reduces the influence of sensitive information in $\tilde{\boldsymbol{u}}$, we expect that the representations of users with different values of the protected attribute become more similar. We verify this by computing the average user representation for each user group and ranking the prototypes, i. e., the interpretable dimensions, from most to least similar. Given the ranking of two user groups, we compute Spearman's rank correlation [41] where values closer to +1 indicate both groups rank the prototypes similarly while values approach-

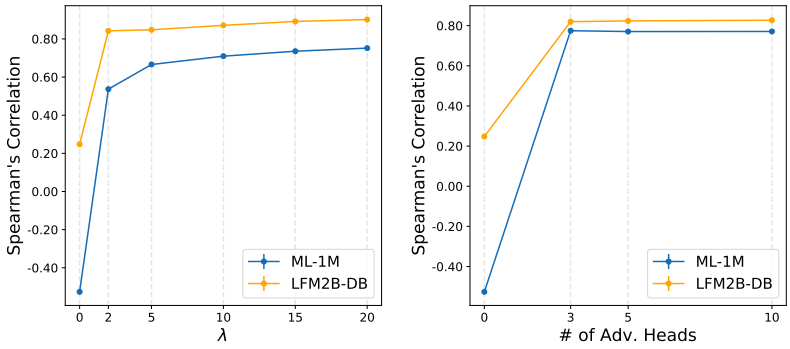
Dataset	Attribute	Metric		# of Adv. Heads 0	3	5	10
ML1M	Gender	Bias ↓	BAcc	$.789_{.003}$	$.642_{.029}$	$.573_{.014}$	$\mathbf{.542_{.008}}$
			Acc	$.788_{.001}$	$.653_{.029}$	$.609_{.009}$	$\mathbf{.608_{.047}}$
		NDCG ↑		$.0625_{.0000}$	$.0621_{.0001}$	$.0621_{.0001}$	$.0620_{.0000}$
	Age	Bias ↓	BAcc	$.465_{002}$	$.310_{.022}$	$.265_{.007}$	$\mathbf{.232_{.026}}$
			Acc	$.424_{003}$	$.267_{.024}$	$.218_{.012}$	$\mathbf{.193_{.011}}$
		NDCG ↑		$.0625_{0000}$	$.0620_{.0002}$	$.0618_{.0001}$	$.0618_{.0001}$
LFM2B-DB	Gender	Bias ↓	BAcc	$.723_{.002}$	$.677_{.011}$	$.608_{.040}$	$\mathbf{.600_{.011}}$
			Acc	$.718_{.004}$	$.705_{.016}$	$\mathbf{.598_{.088}}$	$.636_{.065}$
		NDCG ↑		$.0754_{.0000}$	$.0755_{.0001}$	$.0755_{.0000}$	$.0755_{.0001}$
	Age	Bias ↓	BAcc	$.581_{005}$	$.505_{.020}$	$.455_{.012}$	$\mathbf{.390_{.011}}$
			Acc	$.504_{008}$	$.414_{.016}$	$.375_{.050}$	$\mathbf{.277_{.018}}$
		NDCG ↑		$.0754_{0000}$	$.0754_{.0000}$	$.0754_{.0000}$	$.0755_{.0001}$

Table 4: Debiasing and performance results on both datasets and attributes using the Adv. method across different number of adversarial heads.

ing -1 imply an inverse ranking. Figure 2 shows the results for the two gender groups on both datasets. Plots for age are provided in the appendix. Initially, we observe different prototype rankings between males and females, especially on the ML-1M dataset ($\rho = -.40$). However, as we increase the debiasing strength, the representations of males/females progressively become more aligned, as seen from the correlations plateauing between .70 and .80 across datasets and debiasing strategies. We derive that the scaling weights, while ensuring relevant recommendations for users and mitigating the bias of the protected attribute, lead to an alignment between the representations across user groups.

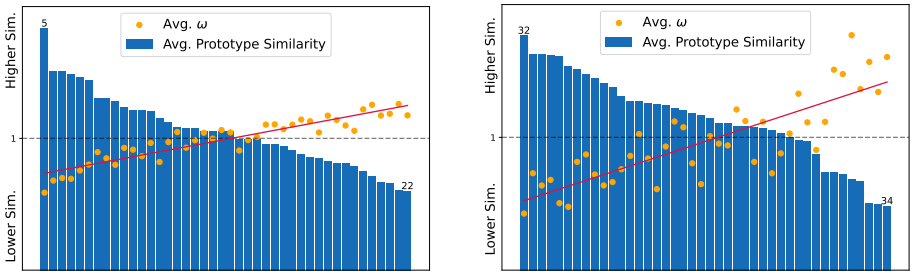
Taking a closer look at the scaling weights, we plot the average user-to-prototype similarities $\boldsymbol{u}$ and average $\boldsymbol{\omega}$ for Female (Fig. 3a) and Male (Fig. 3b) user groups on the ML-1M datasets sorted by most to least similar prototype. We notice a pattern wherein, on average, the scaling values of $\boldsymbol{\omega}$ shrink ($\omega < 1$) the similarities to the prototypes most similar to the user group while they amplify ($\omega > 1$) the similarities to the least similar prototypes. By examining the interpretations of the most and least similar prototypes for each user group, shown in Figure 4a, we infer that the debiased representations for female users show reduced activation towards genre patterns of Romance, Drama, and Comedy and increased activation towards Action and Sci-Fi. Conversely, the debiased representations for the male group display the opposite trend.

Finally, we look at a qualitative example showcasing the application of our learned scaling weights. In Figure 4b, we report the relevant recommendations for an arbitrary female user from ML-1M before and after debiasing. We highlight items dropped from the recommendations in red text and newly recommended items in blue. Additionally, we use green to highlight the cell containing the



(a) MMD (b) Adv.

Fig. 2: Spearman's correlation between avg. male/female prototype rankings on both datasets and both debiasing methods.



(a) Female (b) Male

Fig. 3: Average user-to-prototypes similarities and average $\boldsymbol{\omega}$ values for female users (left) and male users (right), sorted by most to least similar prototypes.

ground truth items. Upon inspection, we observe that the debiased representation indeed affects the recommendations, particularly altering the items at the bottom of the list. We also note a reduction in the number of Romance movies and an increase in Sci-Fi movies before and after debiasing. This diversification also results in better recommendations for the user.

6 Conclusion and Future Directions

This work addresses the pervasive issues of societal bias in RSs from the user perspective. We propose a novel approach that leverages interpretable RS models and introduces per-user scaling weights to mitigate biases in the user representations while preserving recommendation quality. By applying our method to the prototype-based ProtoMF model [34], we demonstrate its effectiveness in reducing bias associated with protected attributes such as gender and age. Our

Prototype 5 (High Sim.)	Prototype 22 (Low Sim.)
Affair to Remember, An *Romance* **Gone with the Wind** *Drama,Romance,War* **Arsenic and Old Lace** *Comedy,Mystery,Thriller* **Love in the Afternoon** *Comedy,Romance* **Swept from the Sea** *Romance*	**Aliens** *Action,Sci-Fi,Thriller,War* **Terminator, The** *Action,Sci-Fi,Thriller* **Star Trek: Wrath of Khan** *Action,Adventure,Sci-Fi* **Night Flier** *Horror* **Blade Runner** *Film-Noir,Sci-Fi*
Prototype 32 (High Sim.)	**Prototype 34 (Low Sim.)**
Star Wars: Episode IV *Action,Adventure,Fantasy,Sci-Fi* **Star Wars: Episode V** *Action,Adv,Drama,Sci-Fi,War* **Star Wars: Episode VI** *Action,Adv.,Romance,Sci-Fi,War* **Back to the Future** *Comedy,Sci-Fi* **Star Trek: Wrath of Khan** *Action,Adventure,Sci-Fi*	**Penny Serenade** *Drama,Romance* **Auntie Mame** *Comedy,Romance* **Charade** *Comedy,Romance,Myst.Thriller* **Love in the Afternoon** *Comedy,Romance* **Crimes of the Hearth** *Comedy,Drama*

(a) Most and least similar prototypes for female (Top) and male (Bottom) users on ML-1M.

Before	After
Wizard of Oz, The *Adv.,Child.,Drama,Musical* **Big** *Comedy,Fantasy* **Breakfast Club, The** *Comedy,Drama* **Vertigo** *Mystery,Thriller* **Raising Arizona** *Comedy* **Ever After** *Drama,Romance* **Clueless** *Comedy,Romance* **Misery** *Horror* **Star Trek: First Contact** *Action,Adv.,Sci-Fi* **Lion King, The** *Animation,Child., Musical*	**Wizard of Oz, The** *Adv.,Child.,Drama,Musical* **Big** *Comedy,Fantasy* **Breakfast Club, The** *Comedy,Drama* **Vertigo** *Mystery,Thriller* **Terminator 2** *Action,Sci-Fi,Thriller* **Raising Arizona** *Comedy* **Alien** *Action,Horror,Sci-Fi,Thriller* **Usual Suspects, The** *Crime,Thriller* **Blade Runner** *Film-Noir,Sci-Fi* **Twelves Monkeys** *Drama,Sci-Fi*

(b) Top-10 recommendations for an arbitrary female user on ML-1M before and after debiasing.

Fig. 4: Examples of qualitative results.

evaluation on ML-1M and LFM2B-DB showcases the flexibility and efficacy of our approach in bias mitigation. Through qualitative analysis, we reveal correlations between consumption patterns and protected attributes, enhancing our understanding of bias in RSs. Moving forward, we envision exploring per-user weights that debias the user representation with respect to a conjunction of different protected attributes [29] simultaneously while also analyzing its effect on recommendation performance. Additionally, investigating end-users' perceptions of biases in recommendations appears as a promising avenue for future work.

Acknowledgments. This research was funded in whole or in part by the Austrian Science Fund (FWF): P36413, P33526, and DFH-23, and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through grant LIT-2021-YOU-215.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Barkan, O., Hirsch, R., Katz, O., Caciularu, A., Koenigstein, N.: Anchor-based collaborative filtering. In: Proc. CIKM (2021)
2. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning: limitations and opportunities. MIT Press (2023)

3. Begley, T., Schwedes, T., Frye, C., Feige, I.: Explainability for fair machine learning. arXiv preprint arXiv:2010.07389 (2020)
4. Beigi, G., Mosallanezhad, A., Guo, R., Alvari, H., Nou, A., Liu, H.: Privacy-aware recommendation with private-attribute protection using adversarial learning. In: Proc. WSDM (2020)
5. Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E.H., Goodrow, C.: Fairness in recommendation ranking through pairwise comparisons. In: Proc. KDD (2019)
6. Bose, A., Hamilton, W.: Compositional fairness constraints for graph embeddings. In: PMLR (2019)
7. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: Proc. ICPR. IEEE (2010)
8. Burke, R., Sonboli, N., Ordonez-Gauger, A.: Balanced neighborhoods for multi-sided fairness in recommendation. In: PMLR (2018)
9. Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., He, X.: Bias and debias in recommender system: A survey and future directions. ACM TOIS **41**(3) (2023)
10. Cremonesi, P., Turrin, R., Lentini, E., Matteucci, M.: An evaluation methodology for collaborative recommender systems. In: Proc. AXMEDIS. IEEE (2008)
11. Deldjoo, Y., Jannach, D., Bellogin, A., Difonzo, A., Zanzonelli, D.: Fairness in recommender systems: research landscape and future directions. User Modeling and User-Adapted Interaction (2023)
12. Ekstrand, M.D., Das, A., Burke, R., Diaz, F., et al.: Fairness in information access systems. Foundations and Trends in Information Retrieval **16**(1-2) (2022)
13. Ekstrand, M.D., Tian, M., Kazi, M.R.I., Mehrpouyan, H., Kluver, D.: Exploring author gender in book rating and recommendation. In: Proc. RecSys (2018)
14. Elazar, Y., Goldberg, Y.: Adversarial removal of demographic attributes from text data. In: Proc. ACL. pp. 11–21 (2018)
15. Fu, Z., Xian, Y., Gao, R., Zhao, J., Huang, Q., Ge, Y., Xu, S., Geng, S., Shah, C., Zhang, Y., de Melo, G.: Fairness-aware explainable recommendation over knowledge graphs. In: Proc. SIGIR (2020)
16. Fusco, F., Vlachos, M., Vasileiadis, V., Wardatzky, K., Schneider, J.: Reconet: An interpretable neural architecture for recommender systems. In: Proc. IJCAI (2019)
17. Ganhör, C., Penz, D., Rekabsaz, N., Lesota, O., Schedl, M.: Mitigating consumer biases in recommendations with adversarial training. In: Proc. SIGIR (2022)
18. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: PMLR (2015)
19. Ge, Y., Tan, J., Zhu, Y., Xia, Y., Luo, J., Liu, S., Fu, Z., Geng, S., Li, Z., Zhang, Y.: Explainable fairness in recommendation. In: Proc. SIGIR (2022)
20. Geyik, S.C., Ambler, S., Kenthapadi, K.: Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In: Proc. KDD (2019)
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Proc. NIPS **27** (2014)
22. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal Machine Learning Research **13** (2012)
23. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM CSUR **51**(5) (2018)
24. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. ACM TIIS **5**(4) (2015)

25. Hauzenberger, L., Masoudian, S., Kumar, D., Schedl, M., Rekabsaz, N.: Modular and on-demand bias mitigation with attribute-removal subnetworks. In: Proc. ACL (2023)
26. Kumar, D., Lesota, O., Zerveas, G., Cohen, D., Eickhoff, C., Schedl, M., Rekabsaz, N.: Parameter-efficient modularised bias mitigation via AdapterFusion. In: Proc. ACL. (2023)
27. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: Proc. AAAI. vol. 32 (2018)
28. Li, R.Z., Urbano, J., Hanjalic, A.: Leave no user behind: Towards improving the utility of recommender systems for non-mainstream users. In: Proc. WSDM. pp. 103–111 (2021)
29. Li, Y., Chen, H., Xu, S., Ge, Y., Zhang, Y.: Towards personalized fairness based on causal notion. In: Proc. SIGIR (2021)
30. Liang, D., Krishnan, R.G., Hoffman, M.D., Jebara, T.: Variational autoencoders for collaborative filtering. In: Proc. WebConf (2018)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. ICLR (2019)
32. Masoudian, S., Volaucnik, C., Schedl, M., Rekabsaz, N.: Effective controllable bias mitigation for classification and retrieval using gate adapters. In: Proc. EACL (2024)
33. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM CSUR **54**(6) (2021)
34. Melchiorre, A.B., Rekabsaz, N., Ganhör, C., Schedl, M.: Protomf: Prototype-based matrix factorization for effective and explainable recommendations. In: Proc. RecSys (2022)
35. Melchiorre, A.B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., Schedl, M.: Investigating gender fairness of recommendation algorithms in the music domain. IP&M **58**(5) (2021)
36. Müllner, P., Lex, E., Schedl, M., Kowald, D.: Differential privacy in collaborative filtering recommender systems: a review. Frontiers in Big Data **6** (2023)
37. Pan, D., Li, X., Li, X., Zhu, D.: Explainable recommendation via interpretable feature mapping and evaluation of explainability. In: Proc. IJCAI (2020)
38. Pan, W., Cui, S., Bian, J., Zhang, C., Wang, F.: Explaining algorithmic fairness through fairness-aware causal path decomposition. In: Proc. KDD (2021)
39. Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: Adapterfusion: Non-destructive task composition for transfer learning. In: Proc. EACL (2021)
40. Shen, A., Han, X., Cohn, T., Baldwin, T., Frermann, L.: Does representational fairness imply empirical fairness? In: Proc. ACL (2022)
41. Spearman, C.: The proof and measurement of association between two things. The American Journal of Psychology **15** (1904)
42. Wu, L., Chen, L., Shao, P., Hong, R., Wang, X., Wang, M.: Learning fair representations for recommendation: A graph-based perspective. In: Proc. WebConf (2021)
43. Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G.: Controllable invariance through adversarial feature learning. Proc. NIPS **30** (2017)
44. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: Fa* ir: A fair top-k ranking algorithm. In: Proc. CIKM (2017)
45. Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., Ma, S.: Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: Proc. SIGIR (2014)

46. Zhao, C., Wu, L., Shao, P., Zhang, K., Hong, R., Wang, M.: Fair representation learning for recommendation: A mutual information perspective. In: Proc. AAAI (2023)
47. Zhu, Z., Wang, J., Caverlee, J.: Measuring and mitigating item under-recommendation bias in personalized ranking systems. In: Proc. SIGIR. pp. 449–458 (2020)