

Received 23 May 2022, accepted 13 June 2022, date of publication 27 June 2022, date of current version 1 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3186719

# Bias and Unfairness of Collaborative Filtering Based Recommender Systems in MovieLens Dataset

ÁLVARO GONZÁLEZ<sup>1,3</sup>, FERNANDO ORTEGA<sup>2,3</sup>, DIEGO PÉREZ-LÓPEZ<sup>3</sup>,  
AND SANTIAGO ALONSO<sup>2,3</sup>

<sup>1</sup>Ingenio Labs, 28001 Madrid, Spain

<sup>2</sup>Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid, 28040 Madrid, Spain

<sup>3</sup>KNOWledge Discovery and Information Systems (KNODIS) Research Group, Universidad Politécnica de Madrid, 28040 Madrid, Spain

Corresponding author: Fernando Ortega (fernando.ortega@upm.es)

This work was supported in part by the Ministerio de Ciencia e Innovación of Spain and European Regional Development Fund (FEDER) under Grant PID2019-106493RB-I00 (DL-CEMG), and in part by the Comunidad de Madrid under Convenio Plurianual with the Universidad Politécnica de Madrid in the Actuation Line of Programa de Excelencia para el Profesorado Universitario.

**ABSTRACT** Recommender Systems have become one of the most important tools for streaming and marketplace systems in recent years. Their increased use has revealed clear bias and unfairness against minorities and underrepresented groups. This paper seeks the origin of these biases and unfairness. To this end, it analyzes the demographic characteristics of a gold standard dataset and its prediction performance when used in a multitude of Recommender Systems. In addition, this paper proposes Soft Matrix Factorization (SoftMF), which tries to balance the predictions of different types of users to reduce the present inequality. The experimental results show that those biases and unfairness are not introduced by the different recommendation models and that they come from the socio-psychological and demographic characteristics of the used dataset.

**INDEX TERMS** Recommender systems, collaborative filtering, fairness, MovieLens.

## I. INTRODUCTION


Recommender Systems (RSs) [1], [2] are tools that allow businesses to offer their customers products and/or services of their interest. They do this by understanding what product or service the customer is more inclined to like and recommending it. RSs started as a nice complement to search engines, but the increased popularity of marketplaces and streaming services, such as Spotify, Amazon or Netflix, has led to a situation in which RSs have become a necessity and, in some cases, the primary method of browsing. RSs can be classified into content-based RSs, knowledge-based RSs and collaborative filtering RSs.

Content-based RSs use keywords or characteristics associated with each item [3]–[5]. They work by recommending items similar to those the user has liked in the past. However,

these systems may become obsolete if users do not interact with new and different items [6].

Knowledge-based RSs work similarly to content-based, but they also add keywords that map the preferences and needs of each user [7], [8]. The system needs to know how an item meets the preferences of an user in order to recommend it. These kinds of RSs are very dependent on properly mapped keywords for each item and on users describing their preferences, something not always doable.

Finally, Collaborative Filtering (CF) based RSs [9]–[11], which are the focus of this manuscript, compute their recommendations by extracting knowledge from a sparse matrix where a set of users rates a set of items. There are two types of CF: memory-based and model-based. Memory-based models work directly on the rating matrix, and their most popular implementation is k-Nearest Neighbours (KNN) [12], [13]. Two flavors of KNN based CF can be found: user-user KNN, which works by finding the most similar users to an active one and recommending to him or her the items that these similar

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han .

users have liked; and item-item KNN, which recommends new items to the active user by identifying the items most similar to those rated positively by the active user. On the other hand, RSs based on model-based CF build a computational model of the rating matrix that is used to compute new recommendations. The most popular models are Matrix Factorization (MF) and Neural Networks (NN). MF [14]–[18] models transform the rating matrix in two new matrices that contain a low-dimensional representation of users and items and can be used to recreate the original matrix as closely as possible. NNs [19] models use deep networks to extract and join latent information from users and items in a non-linear fashion, modeling users' preferences.

Evaluation of RSs is a hot research topic and in recent years new ways to measure models have appeared. Beyond accuracy metrics, which try to identify the value of predictions with more than just the plain error, have become very popular in the last few years. Some example could be novelty and diversity, that evaluate more recommendations of larger number of items; or coverage, that measures the percent of items a model can recommend on a test set. In addition to this, the literature on different biases and demographic diversity in models has exploded. All the CF models described previously have been designed and optimized to better perform in certain backgrounds; however, without regard of the used model, there is always bias and unfairness in the recommendations. We can find examples in different papers that expose multiple bias in RS [20] and unfairness in the recommendations [20]–[22].

Bias is the tendency to better predict certain types of data. This results in better prediction for a specific type of item or user. An example of this could be how a user with a non-mainstream taste is going to be recommended mainstream items due to how many mainstream items are present in the dataset. Bias can come from many different sources, as exposed by [20]. There is a lot of literature on popularity bias [23], position bias [24] and many others [25], [26]. It is worth saying that bias usually creates a feedback loop that further improves the bias towards those items, as the higher number of recommendations increase the number of users' interactions, thus improving the likelihood of being recommended.

On the other hand, unfairness states the inequality of the recommendations for different groups of people. This means that certain users or groups of users have worse recommendations. These users could be from a known minority (f.ex., ethnic people) or not (f.ex., kids that watch documentaries). This inequality may come from different factors: age, gender, race, wealth, or even taste. Lambrecht and Tucker [21] shows how women are recommended jobs with lower pay, and [27] that male authors' books are more likely recommended than female authors' books.

The aforementioned examples show that to build a society more reliant on RS we should strive for equality of opportunity. With that in mind, it is imperative to develop recommendation models that provide not only good recommendations

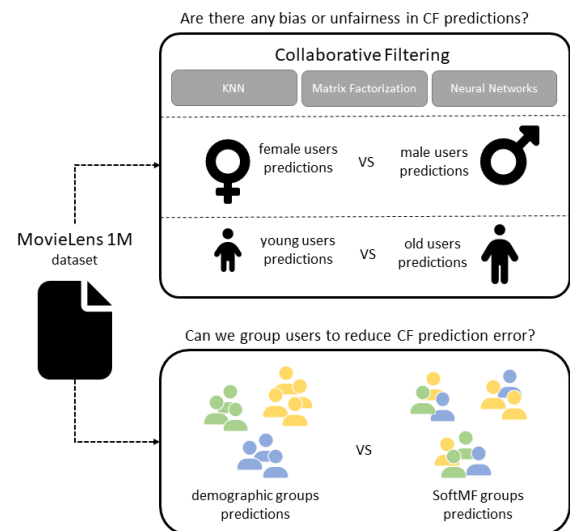


FIGURE 1. Main contributions of this paper.

but also fair and responsible ones. It is not only important that the models are unbiased, but also datasets should allow models to train without providing more types of bias.

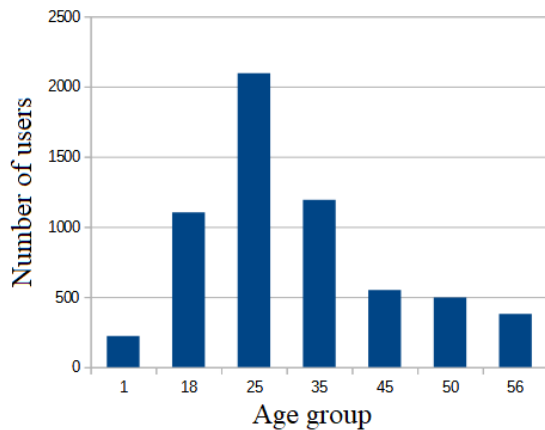
In this paper, we will study the biases and unfairness of CF based RSs on the MovieLens dataset. Figure 1 summarizes the main contributions of the paper. On the one hand, we will analyze the average prediction error for the different demographic groups existing in MovieLens (i.e. male users/female users and young users/old users). On the other hand, we propose a new MF model, SoftMF, which searches for the optimal user grouping to minimize prediction error regardless of the demographic groups to which users belong.

The rest of the paper is structured as follows: Section II analyzes the demographic information of the MovieLens dataset, looking for an imbalance between different user groupings; Section III surveys CF models and their prediction accuracy when training with MovieLens; Section IV proposes and tests a new MF model that finds the best user grouping while training. Finally, section V exposes the conclusions of our research.

## II. THE MovieLens DATASET

MovieLens is one of the most used datasets in RSs research [28]. It started with 100,000 ratings (MovieLens 100K) but has been expanded to 25 million ratings (MovieLens 25M). In this paper, we will use MovieLens1M, the largest MovieLens dataset (it contains 1 million ratings) that provides demographic information about its users.

MovieLens 1M stores the following demographic information related to the users: gender, age, occupation, and zip-code. In our research, we have focused on the users' gender and age, since we consider that occupation and zip-code are less relevant for the bias and unfairness of the recommendations. Regarding users' age, in order to protect users' data, it has been masked by dividing it into seven different ranges:



**FIGURE 2.** Number of users by age in MovieLens 1M dataset.

1 (age under 18), 18 (age between 18 and 24), 25 (age between 25 and 34), 35 (age between 35 and 44), 45 (age between 45 and 49), 50 (age between 50 and 55) and 56 (age of 56 or more).

As we can see in fig. 2, the age histogram of MovieLens follows a normal distribution, slightly skewed towards the left. The group with the most users is 25 (age between 25 and 34), while the group with the smallest number of users is 1 (age under 18). This can be explained easily, as this dataset was collected by the University of Minnesota and the large majority of students and professors would be in the 18-44 age range.

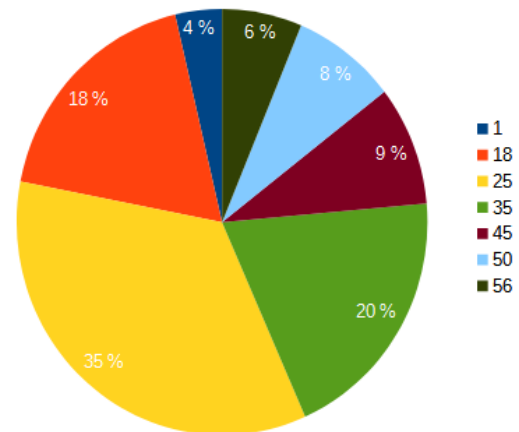
As we mentioned above, the other relevant information is the user's gender. The gender in MovieLens1M is divided into M, for male users, and F, for female users. There are 4331 male users and 1709 female users. This difference in the number of users of each gender, along with the difference in age, makes for a clearly unbalanced dataset.

As we observe, there is a high disparity between gender and/or age. This fact may be reflected in a worse prediction for minority groups, such as female users or users belonging to age group 1 (age under 18). In section III we address this phenomenon. Furthermore, beyond the number of users belonging to each demographic group, the behavior of the users in each group can exacerbate the bias and unfairness of the recommendations performed to the minority groups. For example, if users from the major groups vote more items than users from the minor groups, this imbalance will be magnified. For that reason, we performed a deeper analysis of the rating behavior of each group by measuring its number of votes (tables 1 and 2), the rating distribution (tables 3 and 4), and the group cohesion (tables 5 and 6).

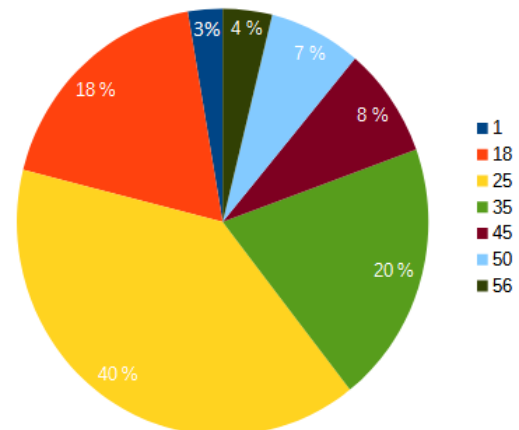
Table 1 shows the average number of ratings by users of each age group. As we can see, users from majority groups (Age 18, 25 and 35) vote for more items than users from minority groups (Age 1, 45, 50 and 56). This situation further reduces the relevance of minority groups in the system. This is shown in figs. 3 and 4, where we compare the percentage of users by age with respect to the percentage of ratings by

**TABLE 1.** Average number of ratings by users of each age group.

Age	Mean Number of votes
1	122.6
18	166.4
25	188.7
35	166.8
45	152.1
50	146.1
56	102.1



**FIGURE 3.** Proportion of data in MovieLens provided by each age group. Percentage of total users by group.



**FIGURE 4.** Proportion of data in MovieLens provided by each age group. Percentage of total ratings by group.

users' age. It is observed that the minority groups represents the 27% of the users of MovieLens1M (fig. 3) and 22% of the total ratings of MovieLens1M (fig. 4).

Table 2 repeats this experiment by dividing users by gender instead of by age. We can observe the same trend with age division; the majority group (M) has more ratings than the minority group (F). Analyzing the relative information provided by each group to the dataset, we find that the M group represents the 71% of users and provides the 75% of ratings, and the group F represents the 29% of users and provides the 25% of the ratings. This imbalance in the dataset could stress bias and unfairness problems.

**TABLE 2.** Mean number of votes by users' gender.

Gender	Mean Number of votes
F	144.2
M	174.0

Another relevant information regarded to the users' behavior is the way in which users vote. If a group of users focuses all their votes on one of the possible MovieLens scores (from 1 to 5 stars), predictions for that group will be easier than if their votes are heterogeneous. Table 3 contains the mean vote of each age group and its standard deviation, and table 4 contains the mean vote of each gender group and its standard deviation. We can see that the mean vote and standard deviation barely change between groups. This means that there is no statistical evidence of differences in the way in which users vote.

**TABLE 3.** Users' rating behaviour of demographic groups. Age groups.

Age	Mean vote	Standard deviation
1	3.54	1.2
18	3.5	1.16
25	3.54	1.12
35	3.61	1.07
45	3.63	1.06
50	3.71	1.06
56	3.76	1.06

**TABLE 4.** Users' rating behaviour of demographic groups. Gender groups.

Gender	Mean vote	Standard deviation
F	3.62	1.11
M	3.56	1.11

Finally, we wanted to check the cohesion of the different groups. If a group is less cohesive, it means that its users are less similar. A group with more different users will be harder to predict as they will have less in common. We define the cohesion of a group  $G$  as the average similarity of the users who belong to the group based on their ratings:

$$cohesion_G = \frac{\sum_{u \in G} \sum_{v \in G \| u \neq v} sim(u, v)}{\#G^2 - \#G} \quad (1)$$

where  $sim(u, v)$  represents the similarity of users  $u$  and  $v$  using a similarity metric such as JMSD, cosine, correlation, and singularities.

Table 5 shows the cohesion of each age group and table 6 shows the cohesion of each gender group. We can observe that the groups with fewer users (i.e., groups of young users, old users, and female users) are less cohesive than the groups with higher number of users. This tendency is observed for every similarity metrics analyzed.

### III. MovieLens BIAS AND UNFAIRNESS

The previous section evidences that there exists a large imbalance between demographic groups in the MovieLens dataset.

**TABLE 5.** Demographic groups cohesion. Age groups.

Age	JMSD	Cosine	Correlation	Singularities
1	0.0508	0.8852	0.5321	0.4634
18	0.0582	0.9166	0.5798	0.4847
25	0.0604	0.9274	0.5774	0.4935
35	0.051	0.9108	0.544	0.4843
45	0.0448	0.896	0.537	0.4684
50	0.0461	0.8985	0.5354	0.4686
56	0.0368	0.8488	0.499	0.4101

**TABLE 6.** Demographic groups cohesion. Gender groups.

Gender	JMSD	Cosine	Correlation	Singularities
F	0.0453	0.9029	0.5372	0.4840
M	0.057	0.9221	0.5674	0.4914

In this section, we analyze whether this imbalance causes bias and unfair recommendations in the MovieLens dataset. In this section, we have evaluated the quality predictions performed by the most popular CF based RS: KNN based CF has been evaluated using correlation [2], cosine [2], JMSD [29], PIP [30], and singularities [31] similarity metrics; MF based CF has been tested using Probabistic Matrix Factorization (PMF) [14], Biased Matrix Factorization (BiasedMF) [15], Bernoulli Matrix Factorization (BeMF) [32], Binomial Non-negative Matrix Factorization (BNMF) [33], and Non-negative Matrix Factorization (NMF) [34]; and NN based CF has been verified using Neural Collaborative Filtering (NCF) [19]. These recommendation models have been chosen as a heterogeneous sample of the different types of CF that exist. On the one hand, models for the three most common CF implementations have been selected: KNN, MF, and NN. On the other hand, within each type of implementation, a wide sample of algorithms has been chosen: different similarity metrics for KNN based CF and varied factorization techniques for MF based CF. For NN based CF, because it is an emerging topic, only the reference model in the state-of-the-art of NN based CF has been selected.

All these CF based RS contain hyper-parameters that must be tuned to achieve a fine tuning of the algorithms in MovieLens datasets. These hyper-parameters have been selected using a Grid Search that minimizes the Mean Absolute Error (MAE) of the predictions. Table 7 contains the results of this search and the selected hyper-parameters.

**TABLE 7.** Hyper-parameters of the evaluated recommendation models.

Model	Hyper-parameters
PMF	factors = 8, $\gamma = 0.01$ , $\lambda = 0.045$
BiasedMF	factors = 6, $\gamma = 0.01$ , $\lambda = 0.055$
BeMF	k = 2, $\gamma = 0.006$ , $\lambda = 0.16$
BNMF	factors = 10, $\alpha = 0.6$ , $\beta = 5$
NMF	factors = 2
kNN	k = 75
NCF	optimizer = adam, epochs = 10

All experiments included in this section have been carried out using CF4J [35] and the train-test split included in the

library that contains 911,031 training ratings and 89,178 test ratings.

### A. AGE PREDICTION ANALYSIS

The first experiment aims to determine the performance of the recommendation models for each age group. To this end, we have trained the selected CF based RS with the training split of MovieLens 1M and computed the prediction error using MAE for each age group. The results of these experiments can be shown in table 8. As we can observe, the MF and NN models perform more accurate predictions than the KNN models. Predictions for group 1 (age under 18) and group 18 (age between 18 and 24) are usually worse than predictions for the rest of the groups, while predictions for group 25 (age between 25 and 34) and group 45 (age between 45 and 49) are better.

This experiment also notes that the youngest users receive worse predictions than the older ones. However, as shown in fig. 2, the number of users in age group 1 is much lower than the number of users of the other groups. Consequently, this unfairness to the youngest users may be due to the amount of data used by each group, rather than the age of the users. For this reason, we repeated the experiment but deleting random users from each group until all groups had the same number of users. The results of this new experiment are shown in table 9. We can see that the different prediction errors follow the same logic as before: The youngest users are usually recommended to be worse than the oldest users. Therefore, we conclude that the unfair predictions of the youngest users do not have a close relationship with the imbalance in the number of users and must be caused by another reason.

Previous experiments show that there are two age groups (age group 1 and age group 18) with a worse prediction even when the dataset is properly balanced. At this point, we consider that this unfairness can be caused by one of the following reasons: (a) the tastes of users who belong to these groups are completely different; or (b) the CF models are biased towards the tastes of older users. To distinguish between these two possibilities, we have performed a new experiment. We have trained the different recommendation models for each age group separately. That is, we have divided the MovieLens dataset into as many datasets as there are age groups, and we have trained each of those datasets separately. If age groups have fewer prediction errors when trained separately, this means that the rest of the dataset is having a negative impact on their predictions, so there exists a bias in the CF model. On the contrary, if the age groups have more prediction error when trained separately, we can confirm that the age of the users has no influence on the tastes of the users, so we need users of all ages to perform accurate predictions.

The results of this experiment are shown in table 10. Also, we can observe the same trends as in the previous experiment (table 8) since the MF and NN models perform better predictions than the KNN models. However, we can also observe new interesting facts. While groups with moderate or high number

of users report the same error, groups with low number of users, such as age group 1 and age group 56, have quite higher errors in their predictions.

To highlight the improvement of training age groups together or separately, we made a table showing the gap between table 8 and table 10. Results are shown in table 11. We find that training each group alone only benefits KNN models, which are the worst performing of the three types. An exception is the age group 18, which finds training alone a benefit in all cases, although it is a slight one.

Finally, to summarize all these experiments, table 12 shows the overall results of the previous experiments grouped by CF based RS type.

These experiments let us conclude that there is a bias towards older users, but this bias is intrinsic to the dataset and it has not been introduced by the recommendation models since there is no any evidence that the users of different age groups have a negative influence on the users of other age groups. The higher prediction error for young users seems to be caused by sociological reasons, such as a wide variety of tastes in young users and a better defined taste of older people.

### B. GENDER PREDICTION ANALYSIS

Once the influence of user age on CF based RS has been studied, we have repeated the same experiments to analyze whether there is bias and / or unfairness in the predictions computed for male and female users. First, we have computed the prediction error of the models analyzed for each gender. Table 13 contains these errors. We can observe that the female group has a worse prediction error than the male group.

As was the case with younger users, the group of female users is in the minority compared to the group of male users. To avoid this disparity, we have balanced both groups by randomly deleting users from the male group until both groups had the same number of users. Then we have tested again the prediction accuracy of all recommendation models. The results can be seen in table 14. We can appreciate that the balancing did not help in equating prediction errors, and the female group is still worse predicted. However, in this case, balancing the dataset resulted in a better accuracy for almost all recommendation models tested.

Despite the improvement in the predictions for both groups after balancing them, there are still unfair predictions for the minority group (that is, female users). To find the cause of this unfairness, we have separately trained the recommendation models for male users and female users. Tables 15 and 16 contain the results of this experiment when the dataset is unbalanced and balanced, respectively. We can observe that female users still obtain worse predictions than male users, so there is a bias in the predictions.

To better compare the results of these experiments, tables 17 and 18 show an improvement gap that occurs when we train the recommendation models with male users and female users separately instead of together in the balanced and imbalanced dataset, respectively. We can observe that



**TABLE 8.** MAE of the predictions by age groups.

Age	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Cor.	Sing.	NCF
1	0.7851	0.7666	0.8422	0.7825	0.8157	0.8983	0.9408	0.8196	0.9596	0.8552	0.7678
18	0.766	0.7532	0.7985	0.7437	0.7872	0.8308	1.0593	0.7976	0.9862	0.8368	0.7339
25	0.7041	0.6921	0.7218	0.6676	0.7191	0.7697	0.9991	0.7432	0.9786	0.7773	0.7007
35	0.7173	0.705	0.7399	0.6847	0.719	0.7777	0.9784	0.7582	0.9836	0.7784	0.6932
45	0.6789	0.6603	0.688	0.6402	0.695	0.7351	0.9173	0.7134	0.9398	0.7314	0.6901
50	0.7046	0.6952	0.6933	0.6879	0.727	0.7718	0.957	0.7485	0.9333	0.7684	0.6904
56	0.7109	0.6885	0.7027	0.6462	0.7176	0.7532	0.9441	0.7486	0.9496	0.7478	0.6768

**TABLE 9.** MAE of the predictions by balanced age groups.

Age	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Cor.	Sing.	NCF
1	0.8296	0.7941	0.8305	0.7867	0.8586	0.8557	1.0612	0.8496	1.0052	0.8582	0.6983
18	0.7713	0.7669	0.8398	0.7446	0.8085	0.8473	0.9418	0.8063	0.9913	0.8113	0.7605
25	0.79	0.7866	0.8401	0.7804	0.787	0.808	0.9445	0.8232	0.9997	0.8388	0.7243
35	0.7631	0.7215	0.7806	0.6908	0.7433	0.758	0.9698	0.7682	0.9327	0.7664	0.7005
45	0.6983	0.6687	0.6814	0.6828	0.7043	0.7406	0.9281	0.706	0.9107	0.7123	0.7101
50	0.7391	0.7081	0.7608	0.6854	0.7424	0.7726	0.9732	0.744	0.9147	0.7416	0.7175
56	0.6569	0.6466	0.6548	0.6182	0.6711	0.7347	1.0171	0.7059	0.9142	0.6924	0.7016

**TABLE 10.** MAE of predictions by separately trained age groups.

Age	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Cor.	Sing.	NCF
1	0.996	0.9456	0.9718	0.8935	0.9594	0.9199	0.9479	0.9145	0.9192	0.9283	0.9159
18	0.7521	0.7386	0.7602	0.7266	0.7586	0.7772	0.9473	0.7685	0.8986	0.7755	0.7442
25	0.7373	0.7259	0.7652	0.7069	0.7496	0.8027	0.9632	0.7717	0.9445	0.7716	0.6968
35	0.7252	0.7034	0.7335	0.6821	0.7237	0.7363	0.9297	0.7255	0.8556	0.7331	0.7033
45	0.7766	0.7213	0.7195	0.6842	0.7422	0.7173	0.8666	0.7163	0.8505	0.7241	0.7153
50	0.8203	0.7635	0.7533	0.7133	0.7831	0.7454	0.8887	0.7365	0.8402	0.751	0.7272
56	0.9137	0.7711	0.7548	0.7521	0.8608	0.7282	0.8406	0.7354	0.8095	0.7273	0.7618

**TABLE 11.** Improvement gap in age groups in predictions by training recommendation models separately rather than together.

Age	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Correlation	Singularities	NCF
1	-0.2109	-0.1790	-0.1296	-0.1110	-0.1437	-0.0216	-0.0071	-0.0949	+0.0404	-0.0731	-0.1481
18	+0.0139	+0.0146	+0.0383	+0.0171	+0.0286	+0.0536	+0.1120	+0.0291	+0.0876	+0.0613	-0.0103
25	-0.0332	-0.0338	-0.0434	-0.0393	-0.0305	-0.0330	+0.0359	-0.0285	+0.0341	+0.0057	+0.0039
35	-0.0079	+0.0016	+0.0064	+0.0026	-0.0047	+0.0414	+0.0487	+0.0327	+0.1280	+0.0453	-0.0101
45	-0.0977	-0.0610	-0.0315	-0.0440	-0.0472	+0.0178	+0.0507	-0.0029	+0.0893	+0.0073	-0.0252
50	-0.1157	-0.0683	-0.0600	-0.0254	-0.0561	+0.0264	+0.0683	+0.0120	+0.0931	+0.0174	-0.0368
56	-0.2028	-0.0826	-0.0521	-0.1059	-0.1432	+0.0250	+0.1035	+0.0132	+0.1401	+0.0205	-0.0850

**TABLE 12.** MAE of the predictions by age groups grouped by CF based RS type.

Age Group	Standard					Balanced					Separated				
	kNN		MF		NN	kNN		MF		NN	kNN		MF		NN
	Mean	StD	Mean	StD		Mean	StD	Mean	StD		Mean	StD	Mean	StD	
1	0.7984	0.0302	0.8947	0.0582	0.7666	0.8199	0.0295	0.9260	0.0999	0.6983	0.9533	0.0382	0.9260	0.0132	0.9159
18	0.7697	0.0229	0.9021	0.1141	0.7339	0.7862	0.0377	0.8796	0.0828	0.7446	0.7472	0.0143	0.8334	0.0836	0.7442
25	0.7009	0.0222	0.8536	0.1243	0.6676	0.7968	0.0244	0.8828	0.0845	0.7243	0.7370	0.0223	0.8507	0.0952	0.6968
35	0.7132	0.0203	0.8553	0.1151	0.6847	0.7399	0.0352	0.8390	0.1034	0.6908	0.7136	0.0208	0.7960	0.0921	0.7033
45	0.6725	0.0222	0.8074	0.1112	0.6402	0.6871	0.0142	0.7995	0.1104	0.6687	0.7288	0.0339	0.7750	0.0766	0.7153
50	0.7016	0.0154	0.8358	0.1006	0.6879	0.7272	0.0301	0.8292	0.1075	0.6854	0.7667	0.0393	0.7924	0.0682	0.7272
56	0.6932	0.0284	0.8287	0.1079	0.6462	0.6495	0.0196	0.8129	0.1450	0.6182	0.8105	0.0729	0.7682	0.0531	0.7618

**TABLE 13.** MAE of the predictions by gender groups.

Gender	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Cor.	Sing.	NCF
F	0.7864	0.7566	0.786	0.7656	0.7849	0.7973	1.0214	0.7849	0.9339	0.7928	0.7443
M	0.7394	0.7198	0.7689	0.6954	0.7538	0.7878	0.9578	0.7706	0.9637	0.7716	0.6983

training each group separately in the imbalanced dataset provides a slight improvement in prediction accuracy for most recommendation models. In contrast, training each

group separately in the balanced dataset provides a slight degradation in prediction accuracy for most recommendation models. However, these improvements/degradations are

**TABLE 14.** MAE of the predictions by balanced gender groups.

Gender	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Cor.	Sing.	NCF
F	0.7458	0.7302	0.7554	0.7101	0.7707	0.7883	0.9834	0.768	0.9625	0.7774	0.7492
M	0.7127	0.6983	0.7132	0.6753	0.7364	0.7488	0.9742	0.7543	0.971	0.7614	0.7155

**TABLE 15.** MAE of predictions from separately trained gender groups. Imbalanced groups.

Gender	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Cor.	Sing.	NCF
F	0.7594	0.755	0.7904	0.7294	0.7676	0.7989	0.9987	0.7653	0.9597	0.7791	0.735
M	0.7126	0.6996	0.7317	0.6884	0.7347	0.77	0.9788	0.7625	0.9576	0.7732	0.6902

**TABLE 16.** MAE of predictions from separately trained gender groups. Balanced groups.

Gender	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Cor.	Sing.	NCF
F	0.7873	0.7781	0.7986	0.7526	0.7829	0.8185	1.0171	0.8024	0.9741	0.8118	0.7412
M	0.7246	0.7073	0.7201	0.6843	0.7251	0.7439	0.9472	0.7412	0.9138	0.7436	0.7067

not significant, since the error varies in the second decimal place.

Finally, to summarize all these experiments, table 19 shows the overall results of the previous experiments grouped by CF based RS type.

These experiments let us conclude that recommendation models do not introduce gender bias despite the unfairness of predictions performed for female users. The cause of this unfairness is not the MovieLens dataset or the recommendation algorithm used, the cause should be based on other sociological reasons.

#### IV. SOFT MATRIX FACTORIZATION (SoftMF) MODEL

Section III evidences that there is a disparity in the quality of the predictions for different users with respect to their demographic information. However, the mechanisms tested to reduce the unfairness in the predictions of female users and young users (i.e., balancing the groups in the dataset and training the groups separately) have failed to avoid this problem. Nevertheless, the bias in the dataset exists, so we hypothesize that it is possible to improve the quality of CF based RS predictions by grouping users by nondemographic characteristics.

In this section, we propose a novel MF based CF model that learns the most optimal grouping for users that reduces the prediction error. We name this model **SoftMF**, since it includes a softmax classification to properly group the users of the MovieLens dataset. That is, users with similar tastes are grouped together during the training phase to avoid negative influence from non-similar users.

We define the model as follows. Being  $R$  the sparse rating matrix that contains the ratings of a set  $U$  of users to a set  $I$  of items, our model factorizes this matrix as

$$R_{U \times I} \approx \text{softmax}(W_{U \times G}) \cdot P_{U \times K \times G} \cdot Q_{I \times K \times G} \quad (2)$$

where  $W$  is a dense matrix that defines the membership of each user to the groups  $G$ ,  $P$  is a dense matrix that contains the latent representation of the users in a latent space of  $K$  dimensions, and  $Q$  is a dense matrix that contains the latent

representation of the items in the same latent space. Note that in this factorization model, both  $P$  and  $Q$  have 3 dimensions, as opposed to the standard model (PMF [14]) where they only have 2 dimensions. That is because each group has its own latent factors defined for each user. Users have a probability of being part of a group and the higher the probability, the higher the importance of the factorization of each group. The softmax function is used by  $W$  to transform the learned parameters into a probability distribution. Let  $\vec{w}_u$  be a row of the matrix  $W$ , we define

$$sm_g(\vec{w}_u) = \frac{e^{\beta \cdot w_{u,g}}}{e^{\beta \cdot w_{u,1}} + \dots + e^{\beta \cdot w_{u,G}}} \quad (3)$$

where  $g$  is the group for which we want to calculate the user's membership probability  $u$ , and  $\beta$  is an hyper-parameter of our model which controls the probability that a user belongs to more than one group. The higher the value of  $\beta$ , the lower the probability of belonging to more than one group. Fixing a high value for  $\beta$ , this model would be comparable to training each group independently.

The loss function of our model is defined as

$$\begin{aligned} loss = \sum_{(u,i) \in R} \sum_{g=1}^G sm_g(\vec{w}_u) (r_{u,i} - p_u^g \cdot q_i^g)^2 \\ + \lambda_p ||p^g||^2 + \lambda_q ||q^g||^2 + \lambda_w w_{u,g}^2 \end{aligned} \quad (4)$$

where  $\lambda_p$ ,  $\lambda_q$  and  $\lambda_w$  are the regularization hyper-parameters.

Using gradient descent, we can optimize this function to minimize the previous loss function. The update rules for each parameter are defined as follows

$$p_u^g \leftarrow p_u^g + \gamma_p (sm_g(\vec{w}_u) \cdot (r_{u,i} - p_u^g \cdot q_i^g) \cdot q_i^g - \lambda_p \cdot p_u^g) \quad (5)$$

$$q_i^g \leftarrow q_i^g + \gamma_q (sm_g(\vec{w}_u) \cdot (r_{u,i} - p_u^g \cdot q_i^g) \cdot p_u^g - \lambda_q \cdot q_i^g) \quad (6)$$

$$w_{u,g} \leftarrow w_{u,g} - \gamma_w \left( \sum_{k=1}^G sm'_{g,k}(\vec{w}_u) \cdot (r_{u,i} - p_u^g \cdot q_i^g)^2 + \lambda_w w_{u,g} \right) \quad (7)$$

**TABLE 17.** Improvement gap in age groups in predictions by training recommendation models separately rather than together. Imbalanced groups.

Gender	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Cor.	Sing.	NCF
F	+0.0270	+0.0016	-0.0044	+0.0362	+0.0173	-0.0016	+0.0227	+0.0196	-0.0258	+0.0137	+0.0093
M	+0.0268	+0.0202	+0.0372	+0.0070	+0.0191	+0.0178	-0.0210	+0.0081	+0.0061	-0.0016	+0.0081

**TABLE 18.** Improvement gap in age groups in predictions by training recommendation models separately rather than together. Balanced groups.

Gender	PMF	BiasedMF	BeMF	BNMF	NMF	JMSD	Cosine	PIP	Cor.	Sing.	NCF
F	-0.0415	-0.0479	-0.0432	-0.0425	-0.0122	-0.0302	-0.0337	-0.0344	-0.0116	-0.0344	+0.0080
M	-0.0119	-0.0090	-0.0069	-0.0090	+0.0113	+0.0049	+0.0270	+0.0131	+0.0572	+0.0178	+0.0088

**TABLE 19.** MAE of the predictions by gender groups grouped by CF based RS type.

Gender	Standard					Balanced					Separated				
	kNN		MF		NN	kNN		MF		NN	kNN		MF		NN
	Mean	StD	Mean	StD		Mean	StD	Mean	StD		Mean	StD	Mean	StD	
F	0.7759	0.0139	0.8661	0.1066	0.7443	0.7424	0.0233	0.8559	0.1073	0.7492	0.76045	0.02205	0.86035	0.11005	0.735
M	0.7355	0.0288	0.8503	0.1011	0.6983	0.7072	0.0224	0.8419	0.1194	0.7155	0.71345	0.02005	0.84845	0.10975	0.6902

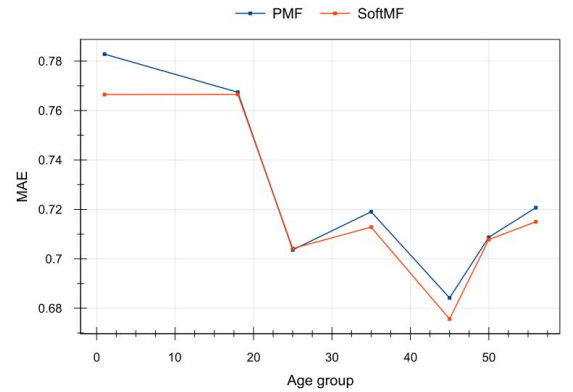
where

$$sm'_{g,k}(\vec{w}_u) = \begin{cases} -sm_g(\vec{w}_u) \cdot sm_k(\vec{w}_u) & \text{if } g \neq k \\ (1 - sm_g(\vec{w}_u)) \cdot sm_g(\vec{w}_u) & \text{if } g = k \end{cases} \quad (8)$$

and  $\gamma_p$ ,  $\gamma_q$  and  $\gamma_w$  are hyper-parameters of the model to control the learning rate.

The algorithm 1 contains the pseudo-code for the fitting method of our proposed model. The algorithm receives as input the rating matrix  $R$  and the hyper-parameters of the model: the number of latent factors  $k$ , the number of groups  $G$ , the learning rates ( $\gamma_p$ ,  $\gamma_q$  and  $\gamma_w$ ), the regularizations ( $\lambda_p$ ,  $\lambda_q$  and  $\lambda_w$ ), the shape of the softmax function  $\beta$  (see eq. 3) and the number of iterations  $m$ . The algorithm returns a matrix with the latent factors of users  $P$  and items  $Q$  as well as the groups of each user  $W$ . The algorithm contains three main loops: the loop from line 2 to line 28 that controls the number of iterations, the loop from line 3 to line 20 that updates the latent factors  $P$  and groups  $W$  of users; and the loop from line 21 to line 27 that updates the latent factors  $Q$  of items. The computational complexity of the algorithm is  $\mathcal{O}(m \cdot U \cdot I \cdot G^2 + m \cdot I \cdot U \cdot G)$ . However, since users only rate a small proportion of available items, we can replace  $U \cdot I$  with  $R$  (the number of ratings). Furthermore, since the number of groups is much smaller than the number of users  $U$  or items  $I$ , we can consider  $G$  highly negligible. In this way, the computational complexity of the algorithm can be simplified to  $\mathcal{O}(2 \cdot m \cdot R)$ . Finally, note that the proposed algorithm can compute user and item updates, for loops from lines 3 to 20 and from lines 21 to 27 respectively, in parallel to each user or item because the update of each user or item does not depend on other users or items.

Our hypothesis is that if there is a clustering of users that allows us to reduce the unfairness of the predictions, the proposed model will find it. To this end, the defined cost function seeks to reduce the prediction error by appropriate grouping of users.

**FIGURE 5.** Prediction error with  $\beta = 1$  by user age.

Before verifying whether our hypothesis is true, we must tune the hyper-parameters of the model to make it fit the MovieLens 1M dataset as well as possible. To do so, we have used Grid Search fixing  $\beta = 1$ . The results of this search are shown in table 20.

Once the hyper-parameters have been tuned, we have compared the prediction accuracy of the proposed factorization model with the standard PMF. We have performed this comparison for the 7 age groups that exist in MovieLens 1M. Figure 5 and table 21 show those results. As we can observe, SoftMF reaches roughly the same error as PMF.

In view of these results, we would like to answer the following research question: Does the grouping performed by SoftMF correspond to the demographic groups? To answer this question, we have assigned each user to his/her most probable group (that is, the group  $g$  with a higher value  $sm'_{g,k}(\vec{w}_u)$ ) and we have computed the percentage of users of each age/gender group that belongs to the 7 learned groups. Table 22 contains the results of this analysis. We can observe that there is no correlation between age/gender and the groups made by SoftMF.



**TABLE 20.** SoftMF hyper-parameters that reports best MAE values.

	Iters	Factors (k)	Groups (G)	$\gamma_p$	$\gamma_q$	$\gamma_w$	$\lambda_p$	$\lambda_q$	$\lambda_w$	$\beta$
SoftMF	50	5	7	0.08	0.08	0.5	0.01	0.01	0.02	1

**Algorithm 1:** SoftMF Model Fitting Algorithm

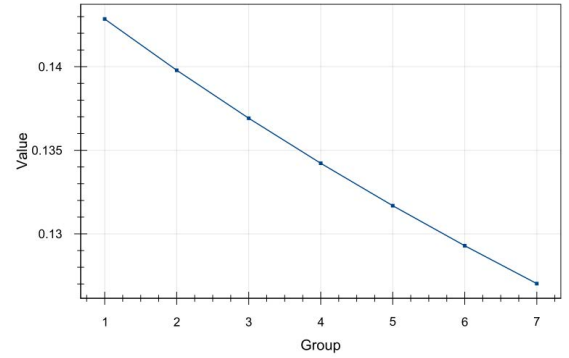
```

input :  $R, k, G, \gamma_p, \gamma_q, \gamma_w, \lambda_p, \lambda_q, \lambda_w, \beta, m$ 
output:  $\mathbf{P}, \mathbf{Q}, \mathbf{W}$ 
1 Initialize  $\mathbf{P} \leftarrow U(0, 1), \mathbf{Q} \leftarrow U(0, 1), \mathbf{W} \leftarrow U(0, 1)$ 
2 repeat
3   for each user  $u$  do
4     for each item  $i$  rated by user  $u$ :  $R_{u,i}$  do
5       Initialize  $\Delta$  to 0
6       for each  $g \in \{1, \dots, G\}$  do
7         Update  $p_u^g$  according to eq. 5
8         for each  $k \in \{1, \dots, G\}$  do
9           if  $g = k$  then
10             $\Delta_g += (1 - sm_g(\vec{w}_u)) \cdot sm_g(\vec{w}_u)$ 
11          else
12             $\Delta_g -= sm_g(\vec{w}_u) \cdot sm_k(\vec{w}_u)$ 
13          end
14        end
15      end
16      for each  $g \in \{1, \dots, G\}$  do
17         $w_{u,g} \leftarrow w_{u,g} - \gamma_w (\Delta \cdot (r_{u,i} - p_u^g \cdot q_i^g)^2 + \lambda_w \cdot w_{u,g})$ 
18      end
19    end
20  end
21  for each item  $i$  do
22    for each user  $u$  that rated item  $i$ :  $R_{u,i}$  do
23      for each  $g \in \{1, \dots, G\}$  do
24        Update  $q_i^g$  according to eq. 6
25      end
26    end
27  end
28 until  $m$  iterations

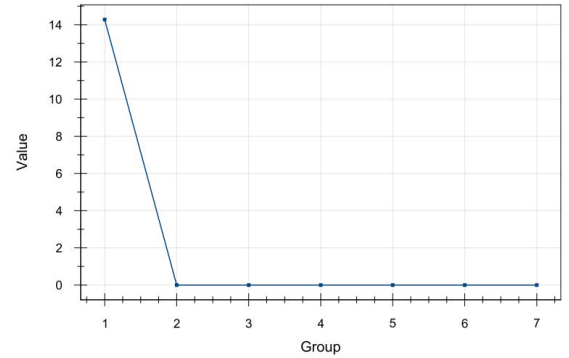
```

**TABLE 21.** Prediction error with  $\beta = 1$  by users' gender.

Gender	PMF	SoftMF
F	0.7439	0.7385
M	0.7113	0.7093



**FIGURE 6.** Average probability of belongs to  $n$ -th most probable group using  $\beta = 1$ .



**FIGURE 7.** Average probability of belonging to  $n$ -th most probable group using  $\beta = 100$ .

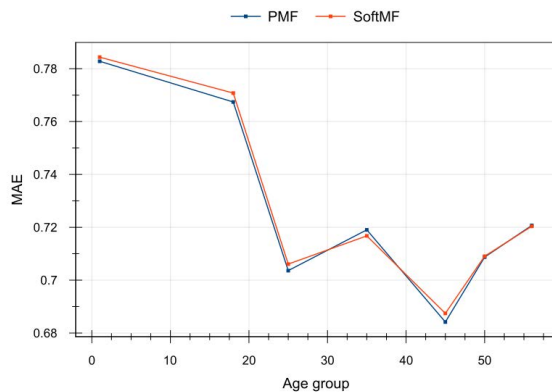
Next, we want to check if SoftMF learns that users must belong to one group or to many groups. To do so, we have sorted the probability of belonging to each group  $g$  for each user (that is, the output of  $sm'_{g,k}(\vec{w}_u)$ ) from highest to lowest value. We then calculated the mean and standard deviation of that probability for the most probable group, the second most probable group, etc. This has been done because we wanted to know how far the probability of belonging from the most probable group to the least probable one was. The results of this analysis are shown in fig. 6. We can observe that most of the users have been assigned to the seven groups with approximately the same probability. The value of a uniform probability distribution with seven groups is  $1/7 \approx 0.14$ , and the scale of the y-axis is from 0.13 to 0.14. This means that the grouping is very soft and barely divides the dataset.

After examining these results, we decided to raise the  $\beta$  parameter to force the model to extremely divide the users. We decided to fix  $\beta = 100$  with the same hyper-parameters and compared the results. Figure 7 shows the results for the new experiment. There is a clear differentiation between the group to which the user is mainly assigned and the rest. This division has no impact on the prediction error, as shown in fig. 8 and table 23. Furthermore, this division does not prevent the model from giving worse results for younger users or female users.

SoftMF has been designed to learn the best way to group users according to their rating behavior. However, the results show that the model either does not group the users and instead tries to train them together or, forcing the grouping, does not get better results compared to other MF models. With

**TABLE 22.** Percentage of users by age or gender in each learned group.

		Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
Age groups	Age 1	4.05	4.29	3.52	3.04	3.57	2.91	4.3
	Age 18	18.72	18.66	18.85	18.91	18.35	16.08	18.181
	Age 25	34.17	33.82	36.98	31.08	36.94	34.73	35.52
	Age 35	16.42	20.81	18.61	20.38	20.61	19.81	21.77
	Age 45	11.39	9.38	8.15	10.02	7.5	9.32	7.65
	Age 50	8.1	8.03	8.51	9.68	6.91	9.32	6.81
Gender groups	Age 56	7.11	4.97	5.35	6.86	6.07	7.8	5.74
	F	29.57	26.35	28.34	28.49	26.34	27.5	29.18
	M	69.22	66.62	71.65	69.59	73.65	72.49	70.81

**FIGURE 8.** Prediction error for  $\beta = 100$  according to the age of users.**TABLE 23.** Prediction error for  $\beta = 100$  by users' gender.

Gender	PMF	SoftMF
F	0.7439	0.7444
M	0.7113	0.7130

this in mind, we can conclude that there is no good way of grouping users in the MovieLens 1M dataset. We can also say that the recommendation models in this article do not introduce any bias or unfairness in the prediction, as all of them follow the same logic (younger users or female users are worse recommended), and there is no way to reduce the differences in predictions they have.

## V. CONCLUSION

In this paper, we have analyzed the bias and unfairness of the recommendation models in the MovieLens 1M dataset. We have evaluated bias and unfairness from two points of view: the age of users and the gender of users. This analysis has been performed using 11 recommendation models from 3 different approaches: KNN, MF, and NN. The main conclusion drawn is that there are unfair predictions towards younger users and toward women. This shows a clear bias in the predictions, which favors male users and old users. However, all the experimental results show that this bias has not been introduced by the recommendation models.

The first experiment concerned the analysis of the error in the predictions about the demographic groups. Once the existence of unfair predictions was verified, the cause was

sought. An attempt was made to balance the number of users in each group, but the unfairness did not disappear. We tried to train each demographic group separately to avoid the influence of the rest of the groups on the model, but were unable to improve the predictive capacity of any group. Finally, a new MF model was defined with the objective of determining the optimal grouping of users to avoid bad influence between users without common tastes. This model determined that no clustering was necessary.

Based on this evidence, we can state that the cause of bias and unfair predictions in the MovieLens 1M dataset is not found in the recommendation model learning process. We hypothesize that it is rooted in other sociopsychological factors. As future work, we propose to further investigate the cause of these problems. A multidisciplinary study that analyzes the problem not only from a machine learning point of view may be the key to solve this problem. It would also be interesting to analyze whether the absence of demographic bias in MovieLens 1M occurs in other data sets.

## REFERENCES

- [1] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender Systems Handbook*. Boston, MA, USA: Springer, 2011, pp. 1–35.
- [2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowl. Syst.*, vol. 46, pp. 109–132, Jul. 2013.
- [3] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in *Proc. 5th ACM Conf. Digit. Libraries (DL)*, 2000, pp. 195–204.
- [4] A. Van Den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proc. Neural Inf. Process. Syst. Conf. (NIPS)*, vol. 26, 2013, pp. 1–9.
- [5] C.-J. Lin, T.-T. Kuo, and S.-D. Lin, "A content-based matrix factorization model for recipe recommendation," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2014, pp. 560–571.
- [6] L. Iaquinta, M. de Gemmis, P. Lops, G. Semeraro, M. Filannino, and P. Molino, "Introducing serendipity in a content-based recommender system," in *Proc. 8th Int. Conf. Hybrid Intell. Syst.*, Sep. 2008, pp. 168–173.
- [7] W. Carrer-Neto, M. L. Hernández-Alcaraz, R. Valencia-García, and F. García-Sánchez, "Social knowledge-based recommender system. Application to the movies domain," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10990–11000, 2012.
- [8] Q. Lin, Y. Niu, Y. Zhu, H. Lu, K. Z. Mushonga, and Z. Niu, "Heterogeneous knowledge-based attentive neural networks for short-term music recommendations," *IEEE Access*, vol. 6, pp. 58990–59000, 2018.
- [9] Y. Koren and R. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook*. New York, NY, USA: Springer, 2015, pp. 77–118.
- [10] D. Sánchez-Moreno, A. B. G. González, M. D. M. Vicente, V. F. L. Batista, and M. N. M. García, "A collaborative filtering method for music recommendation using playing coefficients for artists and users," *Expert Syst. Appl.*, vol. 66, pp. 234–244, Dec. 2016.

- [11] C. Panagiotakis, H. Papadakis, A. Papagrigoriou, and P. Fragopoulou, "Improving recommender systems via a dual training error based correction approach," *Expert Syst. Appl.*, vol. 183, Nov. 2021, Art. no. 115386.
- [12] L. T. Ponnamp, S. D. Punyasamudram, S. N. Nallagulla, and S. Yellamati, "Movie recommender system using item based collaborative filtering technique," in *Proc. Int. Conf. Emerg. Trends Eng., Technol. Sci. (ICETETS)*, Feb. 2016, pp. 1–5.
- [13] R. Ahuja, A. Solanki, and A. Nayyar, "Movie recommender system using K-means clustering and K-nearest neighbor," in *Proc. 9th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Jan. 2019, pp. 263–268.
- [14] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Proc. 20th Int. Conf. Neural Inf. Process. Syst.*, 2008, pp. 1257–1264.
- [15] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Comput.*, vol. 42, no. 8, pp. 30–37, Aug. 2009.
- [16] F. Tan, L. Li, Z. Zhang, and Y. Guo, "A multi-attribute probabilistic matrix factorization model for personalized recommendation," *Pattern Anal. Appl.*, vol. 19, no. 3, pp. 857–866, Aug. 2016.
- [17] B. Zeng, Q. Shang, X. Han, F. Zeng, and M. Zhang, "RACMF: Robust attention convolutional matrix factorization for rating prediction," *Pattern Anal. Appl.*, vol. 22, no. 4, pp. 1655–1666, Nov. 2019.
- [18] A. Pujahari and D. S. Sisodia, "Pair-wise preference relation based probabilistic matrix factorization for collaborative filtering in recommender system," *Knowl.-Based Syst.*, vol. 196, May 2020, Art. no. 105798.
- [19] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.
- [20] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," 2020, *arXiv:2010.03240*.
- [21] A. Lambrecht and C. Tucker, "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads," *Manage. Sci.*, vol. 65, no. 7, pp. 2966–2981, Jul. 2019.
- [22] A.-A. Stoica, C. Riederer, and A. Chaintreau, "Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity," in *Proc. World Wide Web Conf.*, Apr. 2018, pp. 923–932.
- [23] H. Abdollahpouri and M. Mansoury, "Multi-sided exposure bias in recommendation," 2020, *arXiv:2006.15772*.
- [24] M. O. Brien and M. T. Keane, "Modeling result-list searching in the world wide web: The role of relevance topologies and trust bias," in *Proc. 28th Annu. Conf. Cognit. Sci. Soc.*, vol. 28, Princeton, NJ, USA: Citeseer, 2006, pp. 1881–1886.
- [25] B. Marlin, R. S. Zemel, S. Roweis, and M. Slaney, "Collaborative filtering and the missing at random assumption," 2012, *arXiv:1206.5267*.
- [26] T. Wang and D. Wang, "Why Amazon's ratings might mislead you: The story of herding effects," *Big Data*, vol. 2, no. 4, pp. 196–204, Dec. 2014.
- [27] M. D. Ekstrand and D. Kluver, "Exploring author gender in book rating and recommendation," *User Model. User-Adapted Interact.*, vol. 31, no. 3, pp. 377–420, 2021.
- [28] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 1–19, 2016.
- [29] J. Bobadilla, F. Serradilla, and J. Bernal, "A new collaborative filtering metric that improves the behavior of recommender systems," *Knowl.-Based Syst.*, vol. 23, no. 6, pp. 520–528, 2010.
- [30] H. J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," *Inf. Sci.*, vol. 178, no. 1, pp. 37–51, Jan. 2008.
- [31] J. Bobadilla, F. Ortega, and A. Hernando, "A collaborative filtering similarity measure based on singularities," *Inf. Process. Manage.*, vol. 48, no. 2, pp. 204–217, 2012.
- [32] F. Ortega, R. Lara-Cabrera, Á. González-Prieto, and J. Bobadilla, "Providing reliability in recommender systems through Bernoulli matrix factorization," *Inf. Sci.*, vol. 553, pp. 110–128, Apr. 2021.
- [33] A. Hernando, J. Bobadilla, and F. Ortega, "A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model," *Knowl.-Based Syst.*, vol. 97, pp. 188–202, Apr. 2016.
- [34] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
- [35] F. Ortega, J. Mayor, D. López-Fernández, and R. Lara-Cabrera, "CF4J 2.0: Adapting collaborative filtering for Java to new challenges of collaborative filtering based recommender systems," *Knowl.-Based Syst.*, vol. 215, Mar. 2021, Art. no. 106629.



Álvaro González was born in Madrid, Spain, in 1997. He received the B.S. degree in software engineering from the Universidad Politécnica de Madrid, in 2020, where he is currently pursuing the M.S. degree in artificial intelligence. Since 2020, he has been working with private companies, such as Ingenio Laboratories or Aingura IoT in research related projects, focusing mainly in recommender systems. All the projects were in collaboration with the Universidad Politécnica de Madrid. His research interests include machine learning, data analysis, and artificial intelligence.



Fernando Ortega was born in Madrid, Spain, in 1988. He received the B.S. degree in software engineering, the M.S. degree in artificial intelligence, and the Ph.D. degree in computer sciences from the Universidad Politécnica de Madrid, in 2010, 2011, and 2015, respectively. He is currently an Associate Professor with the Universidad Politécnica de Madrid. He is the author of more than 40 research articles in most prestigious international journals. He leads several national projects to include machine learning algorithms into the society. His research interests include machine learning, data analysis, and artificial intelligence. He is the Head Researcher of the KNOWledge Discovery and Information Systems (KNODIS) Research Group.



Diego Pérez-López was born in Cuenca, Spain, in 1996. He received the Vocational Education and Training (VET) in microcomputer systems and networks and the Certificate of Higher Education (HNC) in web application development from the Pedro Mercedes High School, Cuenca, in 2015 and 2017, respectively. He is currently pursuing the double degrees in software engineering and technologies for the information society with the Universidad Politécnica of Madrid. From January 2021 to June 2021, he studied at the Blekinge Institute of Technology, Sweden, as part of the European Erasmus Program.



Santiago Alonso received the B.S. degree in software engineering and the Ph.D. degree in computer science and artificial intelligence from the Universidad Politécnica de Madrid, in 2015. He is currently an Associate Professor with the Universidad Politécnica de Madrid, participating in master and degree subjects and doing work related with advanced databases. His main research interests include natural computing (P-systems) and did some work on genetic algorithms. His current research interests include machine learning, data analysis, and artificial intelligence.

...