The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Investigating Potential Factors Associated with Gender Discrimination in Collaborative Recommender Systems

Masoud Mansoury *
Eindhoven University of Technology
m.mansoury@tue.nl

Himan Abdollahpouri
University of Colorado Boulde
himan.abdollahpouri@colorado.edu and

Jessie Smith
University of Colorado Boulder
jessie.smith-1@colorado.edu

Arman Dehpanah
DePaul University
adehpanah@depaul.edu

Mykola Pechenizkiy
Eindhoven University of Technology
m.pechenizkiy@tue.nl

Bamshad Mobasher
DePaul University
mobasher@cs.depaul.edu

Abstract

The proliferation of personalized recommendation technologies has raised concerns about discrepancies in their recommendation performance across different genders, age groups, and racial or ethnic populations. This varying degree of performance could impact users' trust in the system and may pose legal and ethical issues in domains where fairness and equity are critical concerns, like job recommendation. In this paper, we investigate several potential factors that could be associated with discriminatory performance of a recommendation algorithm for women versus men. We specifically study several characteristics of user profiles and analyze their possible associations with disparate behavior of the system towards different genders. These characteristics include the anomaly in rating behavior, the entropy of users' profiles, and the users' profile size. Our experimental results on a public dataset using four recommendation algorithms show that, based on all the three mentioned factors, women get less accurate recommendations than men indicating an unfair nature of recommendation algorithms across genders.

Introduction

Recommender systems are powerful tools for predicting users' preferences and generating personalized recommendations. It has been shown that these systems, while effective, can suffer from lack of fairness in their recommendation output. The generated recommendations by these systems are, in some cases, biased against certain groups (Ekstrand et al. (2018)). This discrimination among users could negatively affect users' satisfaction, and at worst, can lead to or perpetuate undesirable social dynamics.

Unfair recommendation is often defined as inconsistent performance of a recommendation algorithm for different groups of users. For example, suppose we have two user groups A and B. The recommender system would be considered unfair if it delivered significantly and consistently better recommendations for group A than B or vise versa. This unfairness can be even more problematic if it causes unintentional discrimination against individuals in protected classes such as gender, race, or age [1]. In this paper, we focus on one protected class: gender.

Mansoury et al. (2019) argue that the anomaly of users' rating behavior could be one of the reasons for receiving unfair recommendations. In this work, the authors show that groups with higher anomaly in their rating behavior receive less calibrated and, hence, less fair recommendations. However, they do not test their claim on protected groups, such as men versus women. In this paper, we explore this idea further.

We first show that the degree of anomalous rating behavior in male and female's profile does not explain why these two groups receive different levels of accuracy in their recommendations. As we show in the experimental results section, female groups still receive significantly less accurate recommendations than males with the same level of anomalous rating behavior. We then explore two other factors that could potentially be associated with poor recommender system performance for women. We will describe each of these factors in detail in the following sections.

Related work

Previous research has raised concerns about discrepancies in recommendation accuracy across different genders (Kamishima, Akaho, and Sakuma (2011); Yao and Huang (2017); Zhu, Hu, and Caverlee (2018)). For instance, Ekstrand et al. (2018) show that women on average receive less accurate, and consequently, less fair recommendations than men using a movie dataset.

It is crucial to figure out which factors might be leading to unfairness in recommendations to inform potential solutions to this problem. Abdollahpouri et al. (2019a,b) show that popularity bias can negatively affect the performance of the recommendations. In that work, authors show that algorithms with higher popularity bias give less accurate recommendations to the users.

In a more recent work, Mansoury et al. (2019) argued that anomalous rating behavior could be one of the reasons for unfair recommendations. In their experiments, there is indeed a correlation between the rating anomaly and the per-

*This author also has affiliation in School of Computing, DePaul University, Chicago, USA, mmansou4@depaul.edu.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] https://www.eeoc.gov/laws/types/

formance of the recommendations for different user groups. However, the authors did not specify user groups based on sensitive attributes (e.g., men versus women), but rather a general grouping based on the degree of anomaly of user profiles. In our work, given gender as a sensitive attribute, we show that anomalous rating behavior does correlate with recommendation performance for men. However, as we see in the experimental results section, it does not explain why women, on average, receive less accurate recommendations.

Factors associated with unfair recommendations

In this section, we discuss three different factors that might lead to a poor recommendation performance.

- **Profile anomaly ($\mathcal{A}$):** As discussed in Mansoury et al. (2019), one factor that could impact recommendation performance is the degree of anomalous rating behavior relative to other users. The authors in this paper showed that users whose rating behavior is more consistent with other users in the system as a whole receive better recommendations than those who have more anomalous ratings. This happens because users who rate more in line with typical users are likely to find more matching items or users. We measure the degree of profile anomaly based on how similarly a user rates items compared to the majority of other users who have rated that item. Since collaborative filtering approaches use opinions of other users (e.g. similar users) for generating recommendations for a target user, it is highly possible that users with anomalous ratings receive less accurate recommendations. Given a target user, u, and I_u as all items rated by u, profile anomaly of u can be calculated as:

$$\mathcal{A}_u = \frac{\sum_{i \in I_u} |r_{u,i} - \overline{r_i}|}{N_u} \quad (1)$$

  where $r_{u,i}$ is the rating given by u to item i, $\overline{r_i}$ is the average rating assigned to item i, and N_u is the number of items rated by u (i.e. the profile size of u).

- **Profile entropy ($\mathcal{E}$):** Another possible factor that could impact recommendation performance is how informative a user's profile is. The more diverse a user's ratings are, the higher their entropy is. For example, has the user only given high (or low) ratings to different items? Or are there a wide range of different ratings given by the user? We measure the entropy of user u's profile as follows:

$$\mathcal{E}_u = -\sum_{v \in V} D_u(v) \log D_u(v) \quad (2)$$

  where V is the set of discrete rating values (for example, 1,2,3,4,5) and D_u is the observed probability distribution over those values in u's profile.

- **Profile size ($\mathcal{S}$):** The last factor we investigate in this paper is the profile size of each user. We believe users who are more active in the system (and have rated a larger number of items) receive better recommendations compared to those with shorter profiles.

Table 1: Specification of ML1M for male and female users

	$\#users$	$\overline{\mathcal{A}}$	$\overline{\mathcal{E}}$	$\overline{\mathcal{S}}$
Male	4,331	0.781	4.174	139.2
Female	1,709	0.808	3.995	115.4

Table 2: Precision and miscalibration of recommendation algorithms for male and female users

algorithm	Precision		Miscalibration	
	Male	Female	Male	Female
`UserkNN`	0.235	0.162	0.915	0.971
`ItemkNN`	0.242	0.175	0.874	0.973
`SVD++`	0.133	0.095	1.156	1.130
`ListRankMF`	0.160	0.118	0.970	1.032

Methodology

For our experiments, we use the well-known MovieLens 1M (ML1M) dataset. In this dataset, 6,040 users provided 1,000,209 ratings (602,881 given by males and 197,286 given by females) on 3,706 movies. Table 1 shows the specification of ML1M dataset for male and female users. As shown in this table, there are more male users in the dataset than female users. Moreover, on average, male users have larger profiles, and their profile entropy is also higher than female users. In addition, the average anomaly of male users' profiles is slightly lower than female users.

We divide the dataset into training and test sets in an 80% - 20% ratio, respectively. The training set is then used to build the model. After training different recommendation algorithms, we generate recommendation lists of size 10 for each user in the test set.

We create 20 user groups separately for males and females by measuring different factors: degree of anomaly, entropy, and profile size, discussed further in the previous section. Specifically, we sort users based on each factor and then split them into 20 buckets in an ascending order. Users that fall within each bucket represent one group. In order to calculate the anomaly, entropy, profile size, precision, and miscalibration for each group, we average the corresponding measure over all the users in the group.

We run our experiments using four recommendation algorithms: user-based collaborative filtering (`UserKNN`), item-based collaborative filtering (`ItemKNN`), singular value decomposition (`SVD++`), and list-wise matrix factorization (`ListRankMF`). All recommendation models are optimized using Grid Search over hyperparameters and the configuration with the highest precision is selected. The precision values for `UserKNN`, `ItemKNN`, `SVD++`, and `ListRankMF` are 0.214, 0.223, 0.122, and 0.148, respectively. We used *librec-auto* and LibRec 2.0 for all experiments (Guo et al. (2015); Mansoury et al. (2018)).

Experimental Results

Table 2 shows the performance of recommendation algorithms for male and female users. In terms of precision,

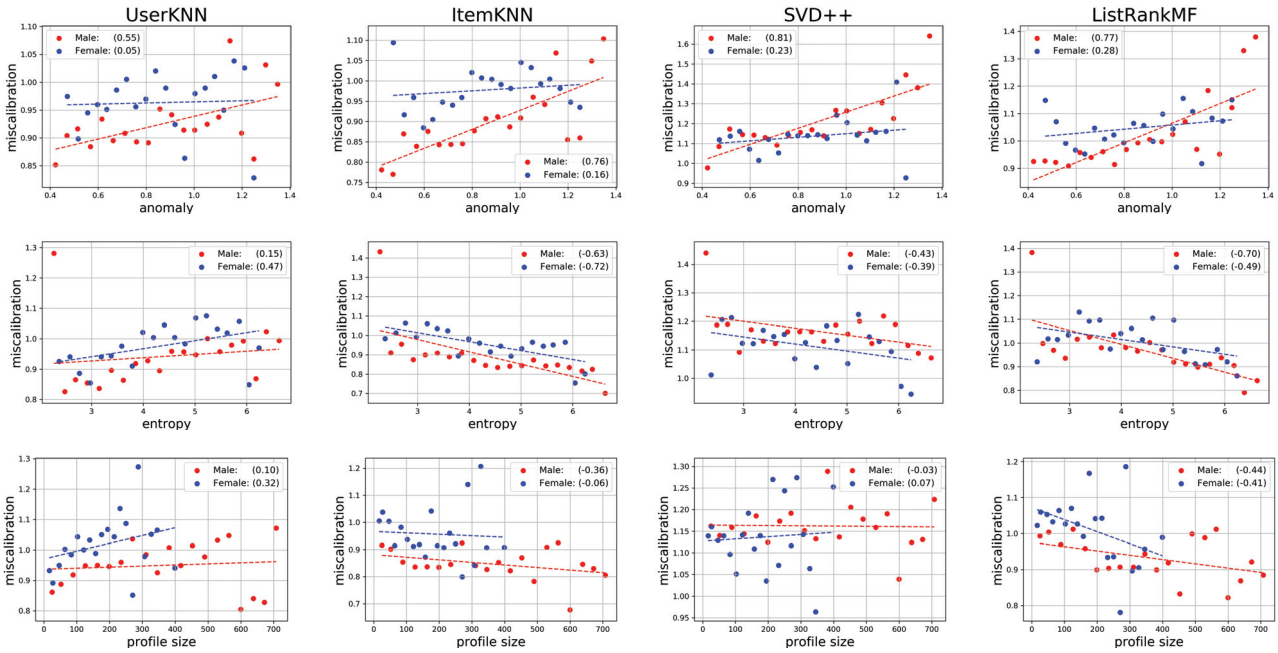


Figure 1: The Correlation between anomaly, entropy, and size of the users' profiles and miscalibration of the recommendations generated for them. Numbers next to the legends in the plots show the correlation coefficient for each user group.

male users consistently receive more accurate recommendations than females and in terms of miscalibration, except for SVD++, male users receive less miscalibrated (i.e. more calibrated) recommendations than females. Lower miscalibration for females than males on SVD++ shows an interesting result in our experiments that needs further investigation.

Figure 1 shows the relationship between the degree of anomaly, entropy, and profile size for 20 user groups for both male and female users and the miscalibration of the recommendations they received. As we can see in the first row (anomaly vs miscalibration), in all algorithms except for SVD++, the recommendations given to the female users have higher miscalibration (they are less calibrated) regardless of the anomaly of their ratings compared to the male user groups. Also, we can see that the positive correlation between profile anomaly and recommendation miscalibration discussed in Mansoury et al. (2019) can only be seen on male users. The second row of Figure 1 shows the relationship between the entropy of the ratings and the miscalibration of their recommendations. Again, it can be seen that except for SVD++, for all other algorithms, female user groups have higher miscalibration in their recommendations regardless of the amount of entropy of their ratings.

Finally, the last row of Figure 1 shows the correlation between the average profile size of different user groups and the miscalibration of their recommendations. Looking at this plot, we can see that there is no significant correlation between these two indicating the profile size of the users does not affect the miscalibration of their recommendations. However, except for SVD++, again all algorithms have higher miscalibration for female user groups regardless of their profile size. It seems that SVD++ is indeed the *fairest* algorithm among the four as it gives a comparable performance for both male and female users. It can also be seen that there is no data point for female groups when the value of the x axis is larger than 400, meaning the largest average profile size for female groups is 400 while there are some male user groups with an average profile size of around 700.

Figure 2 shows the correlation between the aforementioned factors for different user groups and the precision of their recommendations. Unlike miscalibration, it seems the correlations of these three factors with precision are much stronger. For example, the first row of this Figure shows that the higher the inconsistency of the ratings, the lower the precision is, which is what we expected. The second row shows a strong correlation (correlation coefficient ≈ 0.9) indicating that user groups with higher entropy (more information gain) in their ratings receive more accurate (higher precision) recommendations. Also, from the same Figure, we can see for the lower values of entropy, the algorithms behave more fairly, but, the larger the entropy gets, the discrimination between female and male user groups becomes more apparent (higher precision for male user groups). The relationship between average profile size and precision is also shown in the last row of Figure 2. As expected, user groups with larger profiles benefit from more accurate recommendations for both males and females. However, the discrimination can still be seen for some algorithms such as UserKNN where female users with the same profile size still receive recommendations with lower precision compared to the male users.

Conclusion and Future Work

Unfairness in recommendation is an important issue that needs to be diagnosed and treated properly. In this paper, we

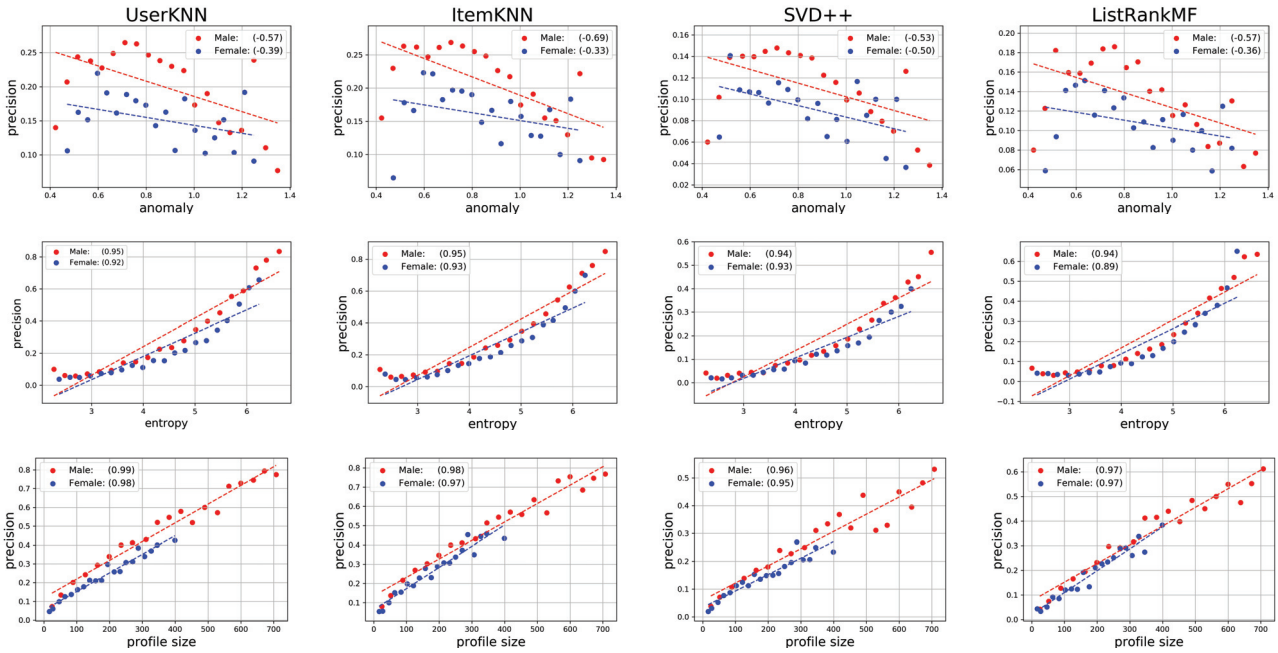


Figure 2: The Correlation between anomaly, entropy, and size of the users' profiles and precision of the recommendations generated for them. Numbers next to the legends in the plots show the correlation coefficient for each user group.

investigated the impact of three different factors on recommendation performance and fairness: the degree of anomaly, entropy, and profile size. We made several interesting observations that need further research. First, we observed that neighborhood-based algorithms such as `UserKNN` and `ItemKNN` discriminate more against women in this data set, in part because they are more affected by the specific characteristics of the profiles such as profiles size and entropy as these characteristics are present more in the female profiles than male ones. On the other hand, the Matrix Factorization based methods (`ListRankMF` and `SVD++`) that rely on lower dimensional latent space and focus more on user-item interactions seem to even out this effect and make the recommendations fairer across different genders. In particular, the `SVD++` algorithm was able to give a comparable performance for both male and female users, which provided a more *fair* treatment across genders than the other algorithms. For future work, we intend to use more datasets for our analysis. We will also investigate the contribution of each of the mentioned factors on recommendation performance. Finally, we will explore other potential factors that could play a role in the unfairness of recommendation algorithms.

References

Abdollahpouri, H.; Mansoury, M.; Burke, R.; and Mobasher, B. 2019a. The impact of popularity bias on fairness and calibration in recommendation. *arXiv preprint arXiv:1910.05755*.

Abdollahpouri, H.; Mansoury, M.; Burke, R.; and Mobasher, B. 2019b. The unfairness of popularity bias in recommendation. *In Workshop on Recommendation in Multistakeholder Environments (RMSE)*.

Ekstrand, M. D.; Tian, M.; Azpiazu, I. M.; Ekstrand, J. D.; Anuyah, O.; McNeill, D.; and Pera, M. S. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *In Conference on Fairness, Accountability and Transparency*, 172–186.

Guo, G.; Zhang, J.; Sun, Z.; and Yorke-Smith, N. 2015. Librec: A java library for recommender systems. In *UMAP Workshops*.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *In 11th International Conference on Data Mining Workshops*, 643–650.

Mansoury, M.; Burke, R.; Ordonez-Gauger, A.; and Sepulveda, X. 2018. Automating recommender systems experimentation with librec-auto. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 500–501. ACM.

Mansoury, M.; Abdollahpouri, H.; Rombouts, J.; and Pechenizkiy, M. 2019. The relationship between the consistency of users' ratings and recommendation calibration. *arXiv preprint arXiv:1911.00852*.

Yao, S., and Huang, B. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *In Advances in Neural Information Processing Systems*, 2921–2930.

Zhu, Z.; Hu, X.; and Caverlee, J. 2018. Fairness-aware tensor-based recommendation. In *In Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1153–1162.