Collaborative filtering algorithms are prone to mainstream-taste bias

Pantelis P. Analytis*
pantelis@sam.sdu.dk
University of Southern Denmark
Odense, Denmark

Philipp Hager*
p.k.hager@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

Collaborative filtering has been a dominant approach in the recommender systems community since the early 1990s. Collaborative filtering (and other) algorithms, however, have been predominantly evaluated by aggregating results across users or user groups. These performance averages hide large disparities: an algorithm may perform very well for some users (or groups) and poorly for others. We show that performance variation is large and systematic. In experiments on three large-scale datasets and using an array of collaborative filtering algorithms, we demonstrate large performance disparities across algorithms, datasets and metrics for different users. We then show that two key features that characterize users, their mean taste similarity and dispersion in taste similarity with other users, can systematically explain performance variation better than previously identified features. We use these two features to visualize algorithm performance for different users and we point out that this mapping can capture different categories of users that have been proposed before. Our results demonstrate an extensive mainstream-taste bias in collaborative filtering algorithms, which implies a fundamental fairness limitation that needs to be mitigated.

KEYWORDS

user features, performance variation, algorithmic fairness

ACM Reference Format:
Pantelis P. Analytis and Philipp Hager. 2023. Collaborative filtering algorithms are prone to mainstream-taste bias. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3604915.3608825

1 INTRODUCTION

Collaborative filtering is one of the most widely deployed, well-studied approaches that have emerged within the recommender systems community [30, 42, 43]. Traditional but also more recent and sophisticated collaborative filtering algorithms are often effective when performance is averaged across users. However, performance averages hide drastic performance variation between users [51]. Although performance variation is well documented [39], we still have a limited understanding of the reasons for it. Further, performance variation surfaces in many important contexts, such as when developing user categorization schemes [53], in algorithm selection [19], when deriving performance estimates for users [4, 19, 40], and more recently concerning algorithmic bias and fairness [3, 20, 51]. A better understanding of how performance varies across users could enable progress along several fronts and inspire the development of techniques to improve algorithm performance for disadvantaged groups of users.

In this work, we seek to understand the extent of performance variation in collaborative filtering by evaluating the performance of five staple collaborative filtering algorithms (k-nn user-user, NMF, k-nn item-item, FunkSVD, and EASE, see Table 1) for different users on the Jester [25], Faces [17], and MovieLens 1M [27] datasets. We demonstrate that performance for users varies drastically across collaborative filtering algorithms and datasets. We further show that all algorithms perform poorly (even below chance level) for a substantial group of users. We then explore the extent to which several previously proposed user variables [4, 19], such as variation in user ratings and mean user rating, can explain the performance discrepancies, and show that they explain only a small proportion of variance. By contrast, we demonstrate that two key user features, the mean taste similarity, and the dispersion in taste similarity between the target user and other individuals [6, 37], can explain a substantial proportion of performance variance across datasets, algorithms, and metrics. The performance variation is so structured and predictable that we can visualize it effectively across users. Collaborative filtering algorithms perform much better for users with high mean taste similarity, a phenomenon that we call *mainstream-taste bias*. They also perform better for users with high dispersion in taste similarity, which shows that this user feature can add further nuance to the notion of *mainstreamness* and can be used to identify different groups of users. Our work is among the first to document the extent of mainstream-taste bias across collaborative filtering algorithms, datasets, and metrics—a necessary first step toward developing new strategies to mitigate this crucial fairness issue.

2 RELATED WORK

Our work falls in a long tradition of contributions that stress the differential abilities of recommender systems to deal with certain items, users, and datasets [9, 45]. A well-studied phenomenon related to mainstream-taste bias is popularity bias, which refers to recommender systems overly recommending items that are already widely popular [1–3, 8, 32]. Most related to our approach, Abdollahpouri et al. [2] show the disparate impact of several recommendation algorithms for users who prefer niche, diverse, or popular

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM ISBN 979-8-4007-0241-9/23/09…$15.00
https://doi.org/10.1145/3604915.3608825

Model	First appearance	Parameters
KNN User-User	Resnick et al., 1994	weight skewness, number of neighbors
KNN Item-Item	Sarwar et al., 2001	weight skewness, number of neighbors
NMF	Zhang et al., 2006	number of components, l1 regularization, l2 regularization
FunkSVD	Funk, 2006	learning rate, regularization rate, epochs, embedding size
EASE	Steck, 2019	regularization rate

Table 1: The collaborative filtering models we compare and their parameters, ordered by the year of their first appearance.

content. Kowald et al. [36] replicate the findings from [2] for music recommendation and use the term *mainstream users* to refer to people who prefer popular items. In contrast to these contributions, we define mainstream taste through correlation patterns between user ratings (taste profiles) rather than item popularity.

The term *mainstream bias* has been used before by Li et al. [38] and Zhu and Caverlee [54]. Li et al. [38] define mainstream users as people who prefer items liked by many people and react negatively to items widely disliked by others (whereas non-mainstream users show interest in rarely visited items or oppose widely accepted or rejected items). The authors identify non-mainstream users by selecting people for whom a collaborative filtering baseline performs worst. Then they show that their autoencoder architecture with an adversarial training objective improves performance for non-mainstream users. More closely related is the work by Zhu and Caverlee [54], who try to identify mainstream versus niche users and test different definitions of *mainstreamness*. One proposed definition, similar to ours, defines mainstream users in terms of their average similarity to other users (e.g., Jaccard similarity). We also show that taste dispersion, the standard deviation of the similarity of a user to other individuals, can add further nuance to the concept of taste mainstreamness.

Our work also builds on a line of work that tries to predict recommender system performance from user and dataset characteristics [4, 10, 13, 19], with applications to algorithm selection [11, 13, 19]. Lastly, our work also relates to work on identifying user categories (e.g., grey sheep users) using statistical properties of the data [26] and work on assessing algorithm performance for different user groups [21]. We show that mean taste similarity and dispersion in taste similarity are highly predictive of the performance of multiple collaborative filtering algorithms for different users and can explain a larger proportion of the performance variance than the user and dataset characteristics identified by Ekstrand and Riedl [19] and Adomavicius and Zhang [4]. Thus, these two features can be effectively used for mapping different user categories.

3 METHODS

3.1 Recommendation algorithms

We analyse four classic collaborative filtering algorithms and one state-of-the-art algorithm (see Table 1). First, we implement two neighborhood algorithms: the weighted user-user k-nearest neighbors algorithm [30, 42] and the weighted item-item k-nearest neighbors algorithm [44]. In addition, we investigate two staple matrix factorization approaches: non-negative matrix factorization (NMF) [52] and FunkSVD, a method popularized during the Netflix competition [22]. Lastly, we implement the EASE algorithm [49], a recent item-item approach inspired by shallow autoencoder models that achieves competitive performance comparable to strong linear approaches and deep neural networks [14]. We evaluate all models using stratified cross-validation, using 60% of each user's rating data for training, 20% for validation and hyperparameter selection, and the remaining 20% for testing. We repeat this process five times with different splits and report the mean test performance per user. Our code is available at: https://github.com/philipphager/recsys-mainstream-taste-bias.

3.2 Evaluation metrics

We evaluate algorithm performance using the normalized discounted cumulative gain (nDCG) [33], the random mean square error (RMSE), and the fraction of concordant pairs (FCP) [35]. FCP measures the percentage of item pairs that an algorithm ranked in the same way as a user. An FCP of 1.0 indicates perfect agreement between model and user, while an FCP of 0.5 means that a model's predictions are not better than chance. We adopt FCP due to its similarity to metrics in decision theory and its intuitive interpretation. Due to space constraints, we omit the results for RMSE and refer to our code repository for results across all metrics.

To evaluate the extent of performance variation across users, we also compute two notions of fairness for each ranking metric. First, we report each metric for the bottom 1% of users for whom the recommender system performs worst and the gap between them and the top 1% of users who benefit the most from their recommendations. This notion of fairness is a generalized version of a Rawlsian maximin principle [23, 41, 55], according to which a society should seek to improve the position for the most disadvantaged users. Our findings hold when varying the share of disadvantaged users from 1% to 5% or even 10%. Second, we quantify disparities across users by computing the Gini-coefficient for each ranking metric [12], which is commonly used to quantify inequality and dispersion in distributions.

3.3 User features

In a second step, we examine how well we can predict the user-level performance of each algorithm using different user characteristics. First, we employ user features, such as the mean user rating or the variance in user ratings, that were used to predict algorithmic performance for different users by Ekstrand and Riedl [19]. Second, we examine features used by Adomavicius and Zhang [4] to predict the performance of different methods across datasets. Table 3 lists all user features adopted from previous work that we could easily compute on our datasets.

In addition, we adopt two key user features, the mean taste similarity between a user and all other users and the dispersion in the

Dataset	Model	FCP				nDCG			
		mean	p1	diff	Gini	mean	p1	diff	Gini
MovieLens 1M	KNN User-User	0.6457	0.4327	0.3843	0.0574	0.9364	0.5501	0.4407	0.0202
	KNN Item-Item	0.6034	0.4224	0.3950	0.0651	0.9209	0.6062	0.3850	0.0225
	NMF	0.5999	0.4315	0.3961	0.0582	0.9240	0.6219	0.3696	0.0223
	FunkSVD	0.6496	0.4345	0.3794	0.0555	0.9393	0.5369	0.4539	0.0189
	EASE	0.6135	0.4302	0.3924	0.0543	0.9302	0.6074	0.3836	0.0206
Faces	KNN User-User	0.7096	0.4605	0.3406	0.0479	0.9014	0.5794	0.4113	0.0402
	KNN Item-Item	0.7046	0.4573	0.3435	0.0483	0.8958	0.5406	0.4542	0.0426
	NMF	0.6920	0.4144	0.3841	0.0499	0.8862	0.5457	0.4489	0.0477
	FunkSVD	0.7069	0.4491	0.3529	0.0489	0.8979	0.5551	0.4384	0.0433
	EASE	0.7056	0.4261	0.3736	0.0480	0.8987	0.5507	0.4436	0.0418
Jester	KNN User-User	0.6609	0.4501	0.3483	0.0491	0.9224	0.6948	0.2950	0.0251
	KNN Item-Item	0.6554	0.4467	0.3500	0.0499	0.9212	0.6947	0.2950	0.0256
	NMF	0.6121	0.4338	0.3572	0.0588	0.9026	0.6796	0.3103	0.0320
	FunkSVD	0.6527	0.4472	0.3507	0.0505	0.9204	0.6853	0.3042	0.0253
	EASE	0.6499	0.4420	0.3544	0.0501	0.9200	0.6909	0.2991	0.0243

Table 2: The performance of different collaborative filtering algorithms as measured in FCP and nDCG. We report the average performance across users (mean) and the performance for the bottom 1% of users for whom the algorithms perform worst (p1). We also report the performance difference between them and the performance for the top 1% of users for whom the algorithms perform best (diff). Lastly, we quantify the overall performance disparity across users in terms of Gini index.

observed taste similarities between the target user and other users. These two features correspond to statistical properties, such as the mean cue-criterion correlation (or mean correlation with other experts) that have been leveraged in psychology [15, 16, 18, 24], management [31], decision science [34, 37], and machine learning [47, 48] to predict the performance of heuristic decision-strategies and improper linear models in different decision environments. These cues have been shown to strongly predict the performance of simple variations of the k-nearest-neighbor algorithm [5].

For both features, we first compute the Pearson correlation between the rating vectors of all user pairs in a dataset. If two users did not rate at least two common items, their Pearson correlation coefficient is set to zero. The mean taste similarity is then defined as the mean Pearson correlation coefficient between a user and all other users [5]. The higher a user's mean taste similarity, the more their preference follows the opinion of others, i.e., is more mainstream. Consequently, a low mean taste similarity implies that the user's ratings diverge from the mainstream, and below 0, the user opposes commonplace preferences. The second dimension is a user's taste dispersion, the standard deviation of the user's Pearson correlation coefficient with all other users [5]. Intuitively, taste dispersion measures how consistently a user agrees or disagrees with other individuals. As a feature it adds further nuance to the concept of mainstreamness and it helps identify specific user groups. For example, users with taste similarity below zero and low dispersion consistently oppose the popular opinion and tend to relate to few other users. As we will show, this category of users consistently receives poor recommendations.

We use user features from previous work, mean-taste similarity, and dispersion in taste similarity separately and jointly to predict how well a collaborative filtering algorithm will perform for unseen users. As a predictor, we train a simple linear model and report the out-of-sample adjusted R^2 as a measure of the quality of the estimates. All reported R^2 estimates are obtained using 5-fold cross-validation. Lastly, we visualize the structure in user-level performance variability by projecting individuals onto a two-dimensional plane that consists of their mean taste similarity and their taste dispersion with other individuals.

3.4 Datasets

Our evaluation employs the MovieLens 1M [27], Jester [25], and Faces [17] datasets. MovieLens and Jester are prominent datasets used repeatedly by the recommender systems community to evaluate algorithm performance in collaborative filtering. Recently, DeBruine and Jones [17] published the Faces of London dataset that reports the ratings of 2,513 people of face portraits of a diverse group of London inhabitants. Although it was developed to study research questions in psychology, the dataset closely corresponds to the dating/matching apps domain, which leverages recommender systems. We use the Faces dataset because it is a complete dataset in which all users have evaluated all recommended "items". Thus, we can exclude the varying number of ratings per user as a cause for potentially observed performance differences. For similar reasons, we focus on the subset of 14,116 users in Jester who evaluated all 100 jokes in the dataset. Last, we use MovieLens 1M without additional filtering and demonstrate that the same concepts apply to common sparse datasets used in the community. We also chose these three datasets because they explore different taste domains (humor, people, and movies) and vary in the average degree of shared taste between users, which is reflected by their varying ranges of mean taste similarity. Because all three datasets report ratings on different scales, we apply min-max scaling and normalize

Dataset	Features	Model				
		User-User	Item-Item	NMF	FunkSVD	EASE
MovieLens	mean taste similarity	0.5204	0.2039	0.1214	0.4843	0.1345
	taste dispersion	0.0154	0.0545	0.0137	0.0213	0.0147
	mean taste similarity, taste dispersion	0.5443	0.2079	0.1212	0.4999	0.1344
	mean rating, rating variance, log rating count, log item popularity, user Gini, mean item Gini	0.0835	0.1816	0.03	0.118	0.043
	all	0.5601	0.2659	0.1349	0.5202	0.1505
Faces	mean taste similarity	0.6864	0.5665	0.8588	0.6593	0.7428
	taste dispersion	0.4536	0.5332	0.2039	0.4711	0.4317
	mean taste similarity, taste dispersion	0.7882	0.7227	0.8576	0.7541	0.8015
	mean rating, rating variance, log rating count, log item popularity, user Gini, mean item Gini	0.2274	0.2249	0.1665	0.2371	0.199
	all	0.8786	0.8326	0.9098	0.8591	0.8755
Jester	mean taste similarity	0.3699	0.2227	0.8623	0.3103	0.4372
	taste dispersion	0.5652	0.6768	0.118	0.5799	0.4499
	mean taste similarity, taste dispersion	0.6887	0.7077	0.8622	0.6654	0.6443
	mean rating, rating variance, log rating count, log item popularity, user Gini, mean item Gini	0.0015	0.0024	-0.0031	0.0146	0.0374
	all	0.7057	0.7213	0.8681	0.691	0.7135

Table 3: Predicting algorithm performance measured in FCP from user features using linear regression. We measure predictive performance in adjusted R^2 score averaged over 5-fold cross validation.

the ratings into a common 0-1 scale. We verified that this rating normalization did not negatively impact the performance of the evaluated models.

4 RESULTS

4.1 Performance (disparities)

Table 2 displays the performance of the five collaborative filtering algorithms as measured by FCP and nDCG. All algorithms show substantial performance variation for different users across all three datasets and measures. To give a sense of the degree of variation, for all strategies and datasets, the absolute difference in terms of FCP between the top 1% and the bottom 1% of users is larger than 34%. In some cases, it is as large as 39%. For example, EASE achieves an average FCP of 61.3% on MovieLens, but a performance of 82.3% for the top 1% and 43% (worse than chance) for the bottom 1%. Similarly, NMF performs on average at 69.2% on Faces but at 79.9% for the top 1% and at 41.4% for the bottom 1%. Similar patterns hold for all algorithms and datasets tested. Note that all algorithms perform at lower-than-chance levels for a non-negligible proportion of users on all datasets. Thus, a small category of users would be better off with random recommendations. Overall, we find major performance disparities between users across all algorithms, datasets, and metrics. However, there is no clear trend between models, which is also reflected in the fact that the top-bottom differences and Gini indexes achieved are almost identical for all models on the same dataset.

4.2 Using different user features to explain performance variation

In the previous section, we showed that the five algorithms perform very differently across users. This section investigates how predictive a user's mean taste similarity and dispersion are for these user-level performance differences. As a baseline, we employ user features that were proposed by Ekstrand and Riedl [19] (log of the number of ratings, the mean user rating, and the variance in the user's rating), as well as dataset-level features proposed to predict collaborative filtering performance by Adomavicius and Zhang [4] that we could easily convert to user-level features to explain performance variance (log item popularity, user Gini, mean item Gini, also see Methods section). A linear model trained with these features has an out-of-sample adjusted R^2 score ranging from 0.03 (NMF) to 0.18 (KNN Item-Item) on MovieLens, 0.17 (NMF) to 0.24 (FunkSVD) for the Faces dataset, and -0.01 (NMF) to 0.04 (EASE) for Jester. In contrast, a simple linear model that uses mean taste similarity and dispersion in taste similarity has an adjusted R^2 between 0.13 (EASE) and 0.52 (KNN User-User) on MovieLens, 0.72 (KNN Item-Item) and 0.86 (NMF) on the Faces dataset, and 0.65 (EASE) to 0.86 (NMF) for Jester. Table 3 gives the complete overview of how well we can predict the performance of each collaborative filtering algorithm for individual users. A model using both sets of features marginally improves in adjusted R^2. Still, when compared to previously considered features, the relative strength of mean taste similarity and, to a lesser extent, dispersion in taste similarity is remarkable.

4.3 Structure in performance variation

The analyses presented in Table 3 suggest that performance variation is systematic and can be explained mainly by the same two key variables for all algorithm/dataset combinations. Figure 1 visualizes the collaborative filtering performance for each user in terms of FCP and positions users according to their taste similarity and taste dispersion. We can see visually that all algorithms perform well for individuals with high mean taste similarity to others (mainstream tastes) and poorly for individuals with low mean taste similarity.

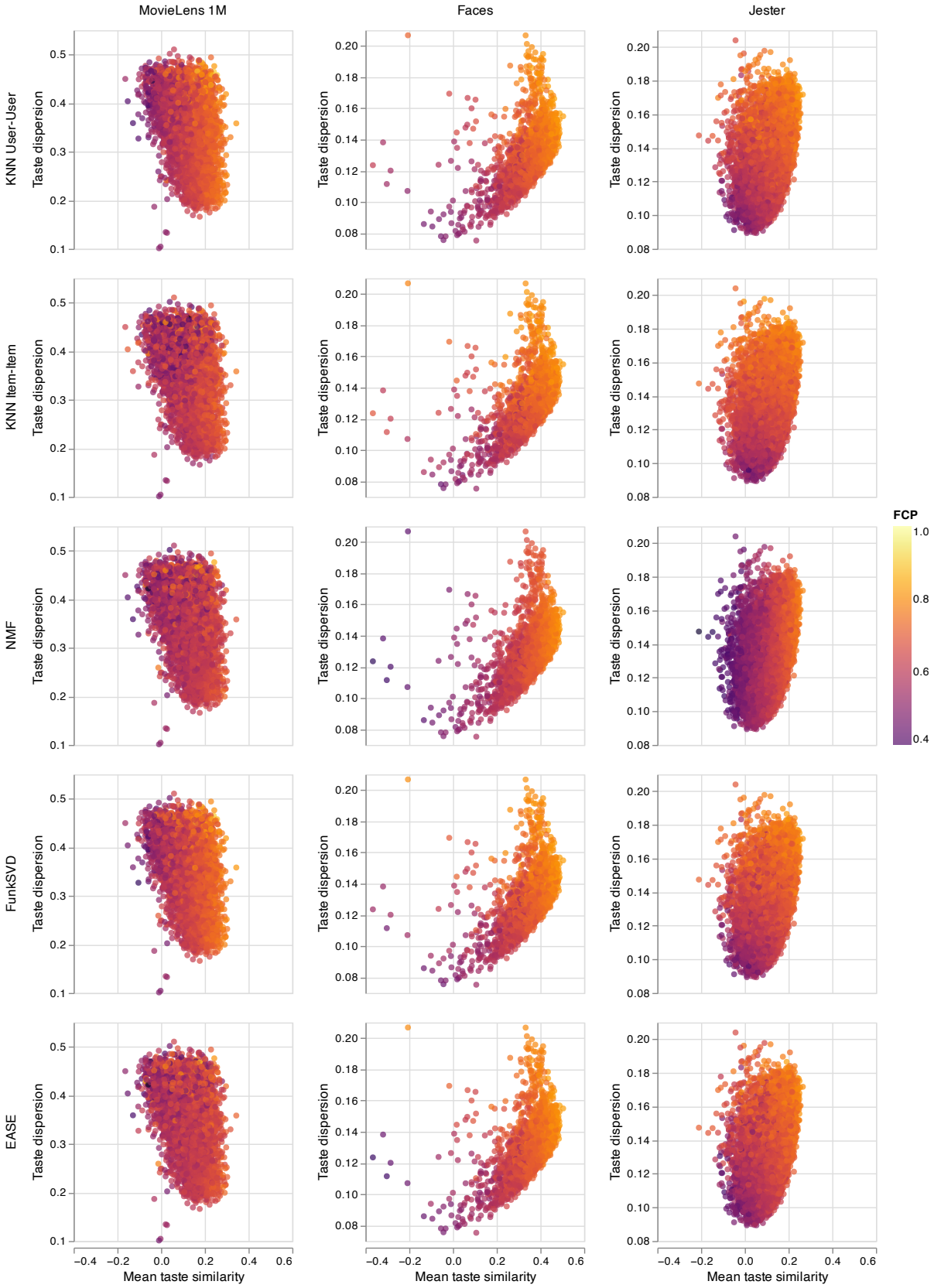

Figure 1: User-level recommendation performance measured in FCP for five collaborative filtering algorithms across three datasets. Users are placed on a 2-dimensional plane depending on their mean taste similarity and dispersion in taste similarity with other users in the dataset. Each point is an individual user.

In essence, collaborative filtering has a strong built-in mainstream-taste bias. It performs much better for people with mainstream tastes than people with alternative or unusual tastes (low or negative mean-taste similarity). In addition, the algorithms tend to perform better for individuals with higher dispersion in taste similarity. This observation is especially noticeable in the complete datasets Faces and Jester, in which every user rated every item. For most models and datasets, performance variation is structured in such a way that we can see a clear performance gradient from left to right on the graphs (see Figure 1, larger mean taste similarity) and from the bottom to the top (see Figure 1, larger dispersion in taste similarity).

5 GENERAL DISCUSSION

We assessed and exposed the large performance variability of five representative collaborative filtering algorithms and showed that they are prone to a mainstream-taste bias—they perform much better for mainstream users than those with non-mainstream or unusual tastes. Our results shed light on a previously little-known issue and open the way for future research.

5.1 Implications for evaluation and fairness

The term *mainstream bias* has been used to account for the large performance disparities between users with mainstream and non-mainstream tastes [38, 54]. Yet we still lack a precise definition of mainstream users. In this work, we proposed an approach based on the statistical properties of user profiles, and we showed its potential to expose the sizeable mainstream bias that appears to be inherent in many collaborative filtering algorithms across settings and metrics. Such performance disparities have direct implications for algorithm evaluation and fairness. So far, algorithms have been evaluated in terms of mean performance, but our work stresses the importance of going further and looking at performance at the individual level. Further, our findings emphasize the need for fairness evaluation when developing new recommender systems to reduce user performance disparities or to focus on specific user categories. For example, one could conceive performance metrics that penalize for larger disparities in the population, drawing inspiration from the social sciences [7, 46]. Additionally, metrics such as the Rawlsian maximin principle metric [23, 41, 55], might be valuable when designers want to improve performance for specific groups of users, such as the most disadvantaged users.

5.2 Categories (or a continuum) of users

Our visualization approach in Figure 1 can account for and help organize previously proposed categories of users and special user profiles (e.g., mainstream or non-mainstream users, grey sheep users, and power users in shilling attacks). Although several user categories have been proposed, there has been no attempt to integrate them into a common map or schema. For example, the term grey sheep users [11] refers to those whose opinions do not correlate with those of other users and who rarely receive good recommendations. In essence, there is no signal to be captured for these users, and correlations with other users are mostly haphazard and spurious. Thus, these user profiles would be similar to randomly generated ratings; their mean taste similarity will be zero and recommendation performance will be close to chance for all collaborative filtering algorithms. Further, our work suggests the existence of many nuanced categories of mainstream or non-mainstream users. For example, non-mainstream users with low dispersion in taste similarity seem to be hard to accommodate for any collaborative filtering algorithm. This group of users seems to be suppressed by collaborative filtering algorithms (i.e., worse than chance performance), and they might even be better off receiving randomly generated recommendations. In contrast, non-mainstream users with higher dispersion in taste similarity seem to have some community they can relate to, and collaborative filtering algorithms can achieve decent prediction rates for them.

5.3 Limitations and future work

Although we demonstrated mainstream-taste bias for five representative collaborative filtering algorithms, our list is far from complete. New collaborative filtering algorithms are added to the recommender systems arsenal at a fast pace [28, 29, 50]. Thus, future work should aim to demonstrate the potential generality of mainstream-taste bias for new algorithms, but also in further taste domains and contexts (i.e. implicit feedback). Algorithms or approaches that can effectively mitigate mainstream-taste bias by improving the performance for disadvantaged users could generate a substantial breakthrough in the field, and lead to an improvement both in terms of average performance and in terms of different measures of fairness. We believe that this is a promising avenue for future research.

ACKNOWLEDGMENTS

We would like to thank Joseph Konstan and Thorsten Joachims for their insightful remarks and K. Rhett Nichols for editing the manuscript. The presented research was supported by a Sapere Aude starting grant to Pantelis P. Analytis bestowed by the Independent Research Fund Denmark.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing popularity bias in recommender systems with personalized re-ranking. *arXiv preprint arXiv:1901.07555* (2019).
[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* (2019).
[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 726–731.
[4] Gediminas Adomavicius and Jingjing Zhang. 2012. Impact of data characteristics on recommender systems performance. *ACM Transactions on Management Information Systems* 3, 1 (2012), 1–17.
[5] Pantelis P. Analytis, Daniel Barkoczi, and Stefan M Herzog. 2018. Social learning strategies for matters of taste. *Nature Human Behaviour* 2, 6 (2018), 415–424.
[6] Pantelis P. Analytis, Daniel Barkoczi, Philipp Lorenz-Spreen, and Stefan Herzog. 2020. The structure of social influence in recommender networks. In *Proceedings of The Web Conference 2020*. 2655–2661.
[7] Anthony B Atkinson et al. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 3 (1970), 244–263.
[8] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2017. Statistical biases in information retrieval metrics for recommender systems. *Information Retrieval Journal* 20 (2017), 606–634.
[9] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Transactions on Information Systems* 41, 3, Article 67 (2023), 39 pages.

[10] Richard Chow, Hongxia Jin, Bart Knijnenburg, and Gokay Saldamli. 2013. Differential Data Analysis for Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 323–326.

[11] Mark Claypool, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes, and Matthew Sartin. 1999. Combining content-based and collaborative filters in an online newspaper. In *Proceedings of ACM SIGIR Workshop on Recommender Systems*, Vol. 60. 1853–1870.

[12] Frank Alan Cowell. 2000. Measurement of inequality. *Handbook of income distribution* 1 (2000), 87–166.

[13] Tiago Cunha, Carlos Soares, and André CPLF de Carvalho. 2018. Metalearning and Recommender Systems: A literature review and empirical study on the algorithm selection problem for Collaborative Filtering. *Information Sciences* 423 (2018), 128–144.

[14] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems* 39, 2 (2021), 1–49.

[15] Robyn M Dawes. 1979. The robust beauty of improper linear models in decision making. *American Psychologist* 34, 7 (1979), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

[16] Robyn M Dawes and Bernard Corrigan. 1974. Linear models in decision making. *Psychological bulletin* 81, 2 (1974), 95.

[17] Lisa DeBruine and Benedict Jones. 2021. Face Research Lab London Set. (4 2021). https://doi.org/10.6084/m9.figshare.5047666.v5

[18] Hillel J Einhorn and Robin M Hogarth. 1975. Unit weighting schemes for decision making. *Organizational Behavior and Human Performance* 13, 2 (1975), 171–192.

[19] Michael Ekstrand and John Riedl. 2012. When recommenders fail: predicting recommender failure for algorithm selection and combination. In *Proceedings of the 6th ACM Conference on Recommender Systems*. 233–236.

[20] Michael D Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and discrimination in recommendation and retrieval. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 576–577.

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

[22] Simon Funk. 2006. Netflix update: Try this at home.

[23] David García-Soriano and Francesco Bonchi. 2021. Maxmin-fair ranking: individual fairness under group-fairness constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 436–446.

[24] Gerd Gigerenzer and Wolfgang Gaissmaier. 2011. Heuristic decision making. *Annual Review of Psychology* 62 (2011), 451–482.

[25] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. 2001. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval* 4, 2 (2001), 133–151.

[26] Benjamin Gras, Armelle Brun, and Anne Boyer. 2016. Identifying grey sheep users in collaborative filtering: a distribution-based technique. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 17–26.

[27] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (2016), 19.

[28] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 639–648.

[29] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[30] Jonathan L Herlocker, Joseph A Konstan, Al Borchers, and John Riedl. 1999. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 230–237.

[31] Robin M Hogarth. 1978. A note on aggregating opinions. *Organizational Behavior and Human Performance* 21, 1 (1978), 40–46.

[32] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What Recommenders Recommend: An Analysis of Recommendation Biases and Possible Countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (2015), 427–491.

[33] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[34] Konstantinos V Katsikopoulos, Lael J Schooler, and Ralph Hertwig. 2010. The robust beauty of ordinary information. *Psychological Review* 117, 4 (2010), 1259.

[35] Yehuda Koren and Joe Sill. 2011. Ordrec: an ordinal model for predicting personalized item rating distributions. In *Proceedings of the 5th ACM Conference on Recommender Systems*. 117–124.

[36] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Proceedings of the 42nd European Conference on IR Research*. 35–42.

[37] Ralf HJM Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Mehdi Moussaid, Giuseppe Argenziano, Iris Zalaudek, Patty A Carney, and Max Wolf. 2019. How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances* 5, 11 (2019).

[38] Roger Zhe Li, Julián Urbano, and Alan Hanjalic. 2021. Leave No User Behind: Towards Improving the Utility of Recommender Systems for Non-mainstream Users. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 103–111.

[39] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021*. 624–632.

[40] Gustavo Penha and Rodrygo LT Santos. 2020. Exploiting Performance Estimates for Augmenting Recommendation Ensembles. In *Fourteenth ACM Conference on Recommender Systems*. 111–119.

[41] John Rawls. 2020. A theory of justice. In *A Theory of Justice*. Harvard University Press.

[42] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*. ACM, 175–186. https://doi.org/10.1145/192844.192905

[43] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–58.

[44] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*. ACM, 285–295. https://doi.org/10.1145/371920.372071

[45] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*. 253–260.

[46] Amartya Sen. 1976. Real national income. *The Review of Economic Studies* 43, 1 (1976), 19–39.

[47] Özgür Şimşek. 2013. Linear decision rule as aspiration for simple decision heuristics. In *Advances in Neural Information Processing Systems*. 2904–2912.

[48] Özgür Şimşek and Marcus Buckmann. 2015. Learning From Small Samples: An Analysis of Simple Decision Heuristics. In *Advances in Neural Information Processing Systems*. 3141–3149.

[49] Harald Steck. 2019. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference*. 3251–3257.

[50] Longqi Yang, Tobias Schnabel, Paul N Bennett, and Susan Dumais. 2021. Local factor models for large-scale inductive recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 252–262.

[51] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems*. 2921–2930.

[52] Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon. 2006. Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the 2006 SIAM International Conference on Data Mining*. 549–553.

[53] Yong Zheng, Mayur Agnani, and Mili Singh. 2017. Identifying grey sheep users by the distribution of user similarities in collaborative filtering. In *Proceedings of the 6th Annual Conference on Research in Information Technology*. 1–6.

[54] Ziwei Zhu and James Caverlee. 2022. Fighting mainstream bias in recommender systems via local fine tuning. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. 1497–1506.

[55] Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. 2021. Fairness among new items in cold start recommender systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 767–776.