Revisiting Popularity and Demographic Biases in Recommender Evaluation and Effectiveness

Nicola Neophytou
neophytounicola@gmail.com
The University of Manchester
London, UK

Bhaskar Mitra
bmitra@microsoft.com
Microsoft
Montréal, Canada

Catherine Stinson
c.stinson@queensu.ca
Queen's University
Kingston, Canada

ABSTRACT

Recommender systems are susceptible to popularity bias and can disproportionately recommend popular items. Groups that are underrepresented in the training data may also receive less relevant recommendations from these algorithms compared to others. Ekstrand et al. [14] investigate how recommender performance varies according to popularity and demographics, and find statistically significant differences in recommendation utility between binary genders on two datasets, and significant effects based on age on one. Here we reproduce those results and extend them with additional analyses. We find statistically significant differences in recommender performance by both age and gender. We observe that recommendation utility steadily degrades for older users, and is lower for women than men. We also find that the utility is higher for users from countries with more representation in the dataset. Total usage and the popularity of consumed content are strong predictors of recommender performance and also vary significantly across demographic groups.

ACM Reference Format:
Nicola Neophytou, Bhaskar Mitra, and Catherine Stinson. 2021. Revisiting Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recommendation and search increasingly mediate our access to information, including news, entertainment, academic resources, and social connections. When evaluating result utility, it is common to report the mean performance over all users. Majority groups tend to dominate overall statistics, but utility may vary across individuals and demographics. Smaller demographic groups may not be well served by these algorithms that are optimized for mean performance. If these systems are unfair, in that the utility of search results and recommendations are systematically lower for some groups, members of those groups may be hindered in their decision-making abilities, access to relevant information, and access to opportunities.

While typical methods of evaluating the effectiveness of search and recommendation do not consider the disparate impact across demographics, several recent papers support the concern that differences in utility do exist. Mehrotra et al. [30] investigate how the needs of subgroups of the population are satisfied in the context of search. They study the impact on search quality by gender and age and find both query distribution and result quality vary across these groups. Ekstrand et al. [14] perform a similar study in the context of recommender systems, which they investigate through offline top-n evaluation. In our work, we reproduce the findings by Ekstrand et al., and extend the analysis to incorporate additional user attributes, such as the user's country, usage, and the popularity of the content they consume. Like them, we find statistically significant differences in recommender utility by age and gender. We observe recommendation utility on average is higher for men, and steadily degrades for older users. We also find the utility is higher for users from countries with more representation in the dataset. Our results indicate usage and popularity of consumed content are strong predictors of recommender performance. Both usage and content popularity vary significantly across groups and may provide a partial explanation for the observed differences in recommender utility, though low utility could also partially explain low usage. In summary, this work studies the following research questions in context of recommender systems:

RQ1 Does utility vary by demographic group?
RQ2 Does utility vary by usage and content popularity?
RQ3 Can usage and popularity explain demographic differences?

2 RELATED WORK

Recommender systems predict future user-item interactions based on past user-item interactions [33]. Past interactions are often subject to biases—such as selection bias [29], conformity bias [24, 28], exposure bias [26], and position bias [9, 20, 22]—and the collected data may reflect societal biases [23, 36]. Recommendation algorithms may further amplify these biases [35, 39] resulting in homogeneity of recommendations and reduced utility [7, 18]. Recommender systems often demonstrate popularity bias [2, 3] where popular items are recommended more than warranted by their popularity, and give lower quality recommendations to users with atypical tastes [5, 15, 16]. These biases raise fairness concerns [1, 6, 32]. For content producers, unfairness may involve disparate exposure over items of comparable relevance [11, 34]. For consumers, unfairness may involve different recommendation quality across demographics [14]. Our focus is on consumer-side fairness, building on prior work by Ekstrand et al. [14].

The fairness concerns in recommendation are not just theoretical questions, they often result in real-world harms. For example, women may see fewer recommendations for high-paying jobs than men [10, 25]. Ekstrand and Kluver [13] find that book recommenders favor male authors. Work on social networks [23, 36] finds that friend recommenders under-recommend minorities. On microlending platforms groups systemically receive smaller loans, or higher interest rates [27]. In ride-hailing platforms, bias can lead to producer-side starvation and loss of income for drivers [37, 38]. For an overview of fairness and bias in recommender systems, see Chen et al. [8], Ekstrand et al. [12].

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

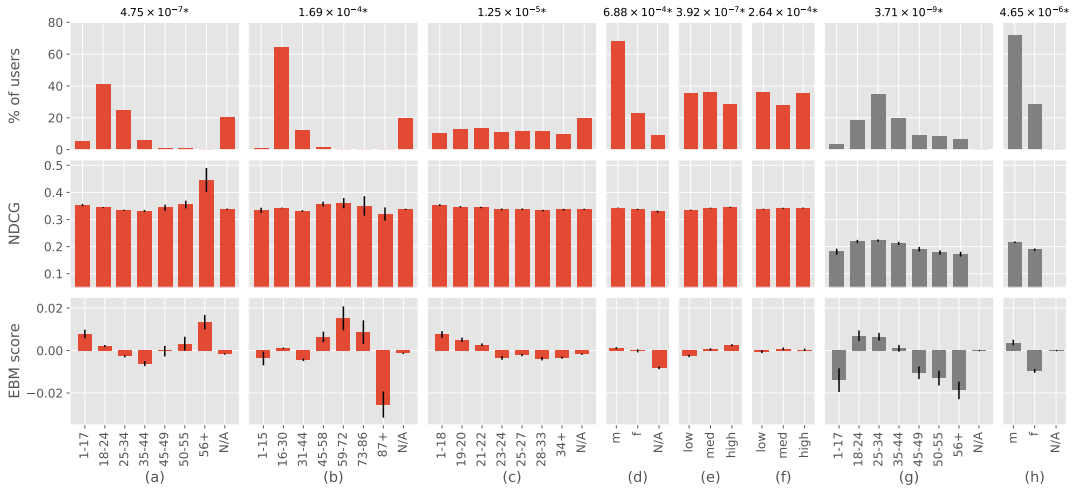

Figure 1: Comparison of binning strategies, metrics, and datasets on recommender utility by demographic variables. Red plots represent the LFM360K dataset and grey represent ML1M. For age, we consider the original bucketing scheme from Ekstrand et al. (a and g), and buckets by equal range (b) and equal number of users (c). (d) and (h) represent gender for LFM360K and ML1M, respectively. (e) and (f) represent country ordered by number of users and by GDP for LFM360K. P-values from Kruskal-Wallis significance tests on NDCG are reported above each column.

3 DEMOGRAPHICS AND POPULARITY

Like Ekstrand et al., we begin our analysis with age and binary gender. For age, in addition to their bucketing scheme, which had unequal age ranges and numbers of users per bucket, we use two additional schemes, such that each age bucket: (i) is equal in age range, and (ii) includes a roughly equal number of users. We also look at how performance varies by country. We bucket countries by the number of users in the dataset, and by the country's gross domestic product (GDP)[1], a proxy for socioeconomic status and cultural hegemony.

Users who have interacted more with the recommender system are likely to receive more relevant recommendations. To analyze how usage influences recommender utility, we bucket users by their number of interactions with items in the collection. We are also interested in the impact of popularity bias. The system may do a better job of recommending items to users who typically interact with items that are popular, compared to users with more niche interests. To investigate how item popularity affects utility, we introduce a novel *pop-index* attribute, defined as the largest value of p such that p% of items the user has interacted with have also received interactions from p% of other users. We take inspiration from the h-index [19], used to measure scholarly impact. We compare recommender utility for groups of users bucketed by pop-index.

4 METHOD

Like Ekstrand et al., we conduct our experiments on Last.FM (LFM360K) [4] and MovieLens (ML1M) data [17]. LFM360K[2] represents a music recommendation task, and contains 358,868 users and 292,385 artists. For each user-artist pair, the dataset contains the number of plays. There are 17,535,605 user-artist pairs with at least one play, so the user-artist matrix is 99.98% sparse. Entries in the user-artist matrix were collected using "user.getTopArtists()" in the Lastfm API, so include only the top artists for each user. The number of artists varies across users, with values between one and 166, with a mean of 50. The dataset also contains user attributes, such as binary[3] gender (67% male, 24% female, 9% missing), age (20% missing), and country (none missing).

ML1M[4] represents a movie recommendation task. ML1M contains 3,952 movies and 6,040 users. Each user-movie pair has an associated 5-point rating assigned by the user. The dataset contains 1,000,209 ratings, corresponding to a 95.81% sparse user-movie matrix. Each user has rated at least 20 movies. The dataset also includes a binary gender, age, and occupation for each user.

We use an alternating Least Squares model [21], as implemented in the Implicit[5] code repository. We use the hyperparameters provided by setting factors to 50 and the regularization constant to 0.01. We train the model for 30 iterations in all experiments. The Implicit code performs some data cleanup[6]. All statistics reported in Section 5 are computed after this cleanup.

We use a five-fold cross-validation setting. For LFM360K, each test partition contains 5,000 randomly sampled users. For ML1M we partition the 6,040 users into five splits containing 1,208 users, for each iteration of cross-validation. We hold out 20% of the items each user has interacted with as test data. All other users and items are used for model training. To avoid the cold-start problem, we remove users who listened to 40 or fewer artists in the LFM360K dataset–roughly 10% of users. The ML1M dataset only includes users who have rated over 20 or more movies, so none are removed. For evaluation, we generate 1,000 recommendations per user, and measure the results using NDCG, MRR, and RBP metrics. To verify if differences in utility are significant across demographics, we perform Kruskall-Wallis significance tests on mean NDCG values between the demographic groups. We also run Bonferroni correction for multiple testing.

To understand the relative impact of user attributes on system performance, we train an Explainable Boosting Machine (EBM) model, as implemented in the InterpretML framework [31], to predict the mean NDCG for each user as a dependent variable. We represent

[1] https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
[2] http://ocelma.net/MusicRecommendationDataset/lastfm-360K.html

[3] We treat gender as a binary class due to the available attributes in the dataset. We do not intend to suggest that gender identities are binary.
[4] https://grouplens.org/datasets/movielens/1m/
[5] https://github.com/benfred/implicit
[6] https://github.com/benfred/bens-blog-code/blob/master/distance-metrics/musicdata.py#L39

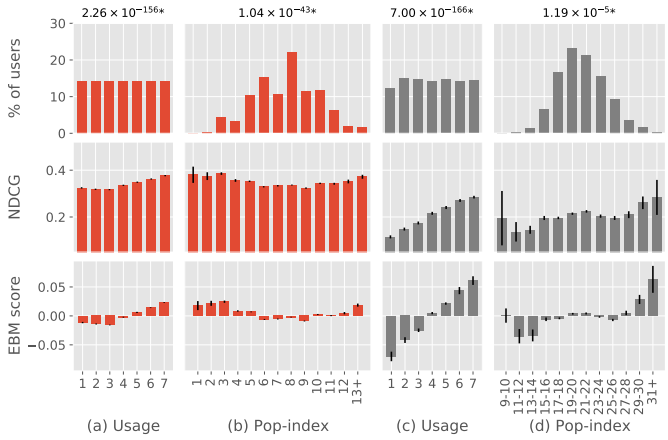

Figure 2: Recommendation utility by usage and content popularity. Red plots represent the LFM360K dataset, grey plots represent ML1M. p-values from Kruskal-Wallis significance tests on NDCG are reported above each column.

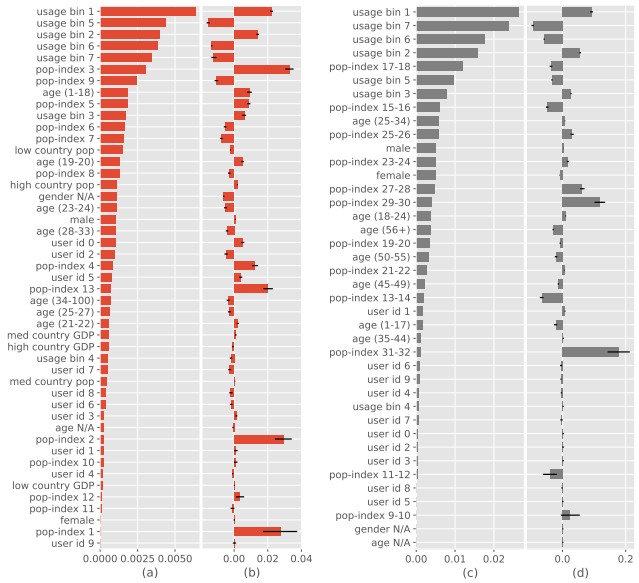

Figure 3: Ranked features and their scores from the EBM analysis. In (a) and (c) equal numbers of users are sampled for each factor. In (b) and (d) the full database is used.

each user by the following features: (i) Age, (ii) Gender, (iii) Country, ordered by prevalence in the dataset and bucketed (LFM360K only), (iv) Country, ordered by GDP and bucketed (LFM360K only), (v) Usage (*i.e.*, total number of listens for LFM360K and total number of movies rated for ML1M), (vi) Pop-index, and finally (vii) The last digit of the user ID. The last digit of the user ID serves as a control feature which should have no effect on performance. We run the EBM model individually for each feature group, and with all features included for cross feature-group comparison.

5 RESULTS

We reproduce the main results from Ekstrand et al., and inquire in more detail into how recommender utility varies by age, gender, and country. We also study the impact of usage and pop-index on utility, and how they interplay with demographics.

RQ1 Does utility vary by demographic group?

Figure 1 shows the distribution of users, recommender utility (mean NDCG), and the EBM scores corresponding to different demographic variables. For each column, we run the Kruskall-Wallis significance test and on all metrics. P-value for NDCG is reported above each column.

Impact on age. Ekstrand et al. find significant differences in recommender utility across different user age brackets according to the Kruskal-Wallis significance test. Our analysis confirms these findings on both datasets, as we also report significant differences based on Kruskal-Wallis significance test ($p < 0.01$) across the same age brackets (Figure 1a and 1g). We also find significant differences when we try alternative binning strategies on LFM360K, corresponding to bins with equal age range (Figure 1b) and bins with equal number of users (Figure 1c). While we only report p-values corresponding to NDCG, we have verified the differences are also statistically significant for MRR and RBP, except for MRR for ML1M.

The first row shows on both datasets that the age distribution is skewed towards young adults, more so for LFM360K than ML1M. Because the age buckets were irregular, we show the results with buckets of uniform range (Figure 1b). We also posit that a skewed distribution of users across age buckets may make it difficult to detect differences in utility across ages, because some age buckets contain very few users. Therefore, we additionally try buckets containing approximately equal numbers of users (Figure 1c). When the number of users in each bucket are comparable, we find a gradual downward trend in recommender utility, as age increases. This effect was not visible in Ekstrand et al. We also observe a similar downward trend on ML1M as seen in Figure 1g. This trend is further confirmed by the EBM scores in Figures 1c and 1g where younger ages correspond to higher EBM scores when the number of users in each bucket are approximately equal.

Impact on gender. Both LFM360K (Figure 1d) and ML1M (Figure 1h) datasets contain many more male than female users. As in Ekstrand et al., we observe statistically significant differences in utility by gender based on Kruskal-Wallis significance test ($p < 0.01$), with better recommendation utility for male than female users. This is observed in both datasets, except for MRR and RBP for LFM360K, and MRR for ML1M. Given the unbalanced user distribution across genders in these datasets, this can either be the result of a popularity bias, or a demographic bias. We revisit this question later in this section in the context of RQ3.

Impact on country. An additional demographic variable available in the LFM360K dataset, but not in ML1M, is users' country of residence. Ekstrand et al. did not analyze whether there is evidence of recommender utility differences by country, but we perform this analysis here. We group the countries in two ways: into low, medium and high buckets based on the number of users from that country, and also based on the country GDP. Figures 1e and 1f show the results corresponding to the two analyses.

We find statistically significant differences by country on both measures, except for MRR and RBP for GDP. The model has higher recommender utility for users from countries with more representation in the dataset. The same trend is not observed, however, when countries are ordered by GDP.

RQ2 Does utility vary by usage and content popularity?

It is not obvious when to attribute utility differences across groups of users to popularity bias, rather than bias specifically affecting demographic groups, because marginalized groups are often also less represented in training datasets. To explore this issue, we first investigate how recommender utility is affected by two measures of popularity: usage and pop-index. For a given user, high usage implies more representation in the data, while a higher pop-index corresponds to affinity towards items that are popular with other users in the dataset. In Figure 2 we compare both these measures on the LFM360K and ML1M datasets. For both datasets there is a trend toward greater NDCG as usage increases. The EBM analysis shows the same trend, where low usage corresponds to a negative effect on the EBM score, and high usage corresponds to a positive effect. We also investigate popularity in the sense of how popular items preferred by a user are among the user population as a whole. Our hypothesis is that users whose playlists contain more popular items will likely have greater recommendation utility. On ML1M (Figure 2d), we observe a trend which supports our hypothesis. However, on LFM360K (Figure 2b), we observe a U-shaped trend, with higher utility associated with both groups of users with maintstream and unique tastes. We suspect differences in observations on the two datasets may be partially explained by the semantics of user interactions in the two cases. In LFM360K, the user interacts with an artist by listening to them, and they can listen to the same artist multiple times. So, for users with more distinctive tastes, the recommender algorithm may still achieve reasonable performance by recommending items the user interacted with before. In contrast, in ML1M the user interacts with the item by providing a rating and therefore the recommender must suggest new items the user has not interacted with before, which is a more difficult challenge, specifically when the user has a distinctive taste.

RQ3 Can usage and popularity explain demographic differences?

One of our goals is to better understand the relative importance of different demographic and popularity features to explain the differences in mean recommender utility amongst users. Towards that goal, we train an EBM model to predict mean recommender utility based on these user attributes. Figure 3 shows that on both datasets (LFM360K and ML1M) the usage features emerge as the most predictive, followed by pop-index. Among the demographic attributes, some of the age-related features are ranked highest on both datasets. On LFM360K, age is followed by country (ordered by number of users) and gender as the next most predictive user attributes. In the absence of country information, on the ML1M dataset we observe gender to be high in the feature ranking after age. The high feature importance for usage and pop-index provides evidence than some of the demographic differences may be explained by representation in the data. This is not to argue that the recommender system under study is fair to different demographics of users. Disparity of utility across demographics may directly influence user retention [14] and usage. This creates a vicious cycle where a small difference in utility across user groups may be further amplified by subsequent disparity in system adoption and usage across demographics, leading to even bigger disparities in utility. Table 1 shows how usage and pop-index are distributed across demographic groups, further demonstrating how they may correlate with historical marginalization.

Table 1: Percentage of users in different usage and pop-index buckets corresponding to each demographic groups for LFM360K. For younger users and men a higher proportion of the population correspond to higher usage buckets. The trend for pop-index is less clear.

	Age (bucketed by equal number of users)								**Gender**		
	1-18	**19-20**	**21-22**	**23-24**	**25-27**	**28-33**	**34+**	**N/A**	**m**	**f**	**N/A**
Usage											
1	11%	8%	9%	11%	12%	15%	24%	21%	13%	16%	20%
2	13%	11%	12%	13%	13%	17%	17%	17%	14%	15%	15%
3	14%	15%	13%	13%	15%	15%	15%	14%	14%	16%	15%
4	16%	15%	15%	14%	14%	14%	13%	14%	14%	15%	13%
5	15%	16%	15%	16%	16%	14%	11%	13%	14%	14%	13%
6	15%	17%	18%	17%	14%	13%	11%	11%	15%	13%	12%
7	17%	18%	18%	16%	15%	12%	9%	10%	16%	11%	11%
Pop-index											
1	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2	0%	0%	0%	0%	0%	0%	1%	1%	0%	0%	0%
3	3%	3%	3%	4%	4%	5%	8%	6%	5%	4%	5%
4	2%	2%	3%	3%	3%	3%	5%	4%	4%	2%	4%
5	7%	9%	9%	10%	11%	12%	15%	11%	11%	8%	11%
6	15%	15%	14%	16%	14%	15%	15%	17%	15%	14%	17%
7	10%	10%	10%	10%	11%	10%	12%	11%	11%	10%	11%
8	23%	23%	23%	23%	24%	22%	19%	20%	21%	25%	21%
9	14%	13%	12%	12%	11%	10%	9%	11%	11%	13%	12%
10	14%	12%	13%	12%	12%	12%	10%	10%	12%	13%	8%
11	7%	8%	7%	6%	6%	6%	4%	5%	6%	7%	6%
12	2%	3%	2%	2%	2%	2%	1%	2%	2%	2%	2%
13+	2%	2%	2%	2%	1%	2%	1%	1%	2%	2%	1%

6 DISCUSSION AND CONCLUSION

We confirmed that recommender systems are prone to unfairness across the demographic attributes available in the datasets used here. To explore this question more thoroughly, one would need access to more detailed demographic data, and the ability to observe temporal dynamics of how recommendations affect usage and usage affects recommendations. In order to answer questions like what caused the U-shaped pattern we found in recommender utility by usage, we would need the ability to intervene on recommendations in real time.

Mehrotra et al. [30] point out that users for whom a search engine is least satisfactory can paradoxically end up having the highest measured utility. They found when utility is bad enough to make a user stop using the service for everyday needs, they still use the search engine for very easy queries that they assume even a poor search engine could get right. Such searches end up being successful, resulting in artificially high utility scores. User attrition is an issue we cannot track given the datasets used here. It may be that users who have the highest usage are a self-selecting group for whom recommenders happen to work well.

For both datasets there is a trend toward greater utility as usage increases. This is unsurprising, given that users with higher usage will provide more labels, with which the recommender can build a more accurate model of user preferences. One anomalous effect we observed is in the LastFM dataset; users with least usage have higher utility recommendations than users with slightly more usage. This could be evidence of the same effect as observed by Mehrotra et al. [30]. If LastFM gives poor recommendations for a given user, that user might stop using it for everyday music streaming, but still use it when they are looking for something very mainstream. Another possibility is since LastFM users input a few artists they like when setting up their accounts, early listens will be dominated by artists which the user identified as being among their favourites, rather than recommendations provided by the model. Utility may therefore be artificially high during early use.

The social harms that can result from unfair recommendation go well beyond some people choosing not to use a tool that others find fun and convenient. Recommendation algorithms are used to

make major life decisions, like mortgage lending, job searching, and for basic access to information. The body of work we are adding to here demonstrates that fair recommendation is a problem requiring serious attention.

REFERENCES

[1] Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv preprint arXiv:1907.13158* (2019).

[2] Himan Abdollahpouri and Masoud Mansoury. 2020. Multi-sided exposure bias in recommendation. *arXiv preprint arXiv:2006.15772* (2020).

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The Connection Between Popularity Bias, Calibration, and Fairness in Recommendation. In *Proc. RecSys*. 726–731.

[4] áOscar Celma. 2010. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer.

[5] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12, 4 (2002), 331–370.

[6] Robin Burke. 2017. Multisided Fairness for Recommendation. *CoRR* abs/1707.00093 (2017).

[7] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *Proc. RecSys (RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 224–232.

[8] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and Debias in Recommender System: A Survey and Future Directions. *arXiv preprint arXiv:2010.03240* (2020).

[9] Andrew Collins, Dominika Tkaczyk, Akiko Aizawa, and Joeran Beel. 2018. A study of position bias in digital library recommender systems. *arXiv preprint arXiv:1802.06565* (2018).

[10] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[11] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proc. CIKM*. 275–284.

[12] Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2021. Fairness and Discrimination in Information Access Systems. *arXiv preprint arXiv:2105.05779* (2021).

[13] Michael D Ekstrand and Daniel Kluver. 2021. Exploring author gender in book rating and recommendation. *User Modeling and User-Adapted Interaction* (2021), 1–44.

[14] 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*. 172–186.

[15] Mustansar Ghazanfar and Adam Prugel-Bennett. 2011. Fulfilling the needs of gray-sheep users in recommender systems, a clustering solution. (2011).

[16] Benjamin Gras, Armelle Brun, and Anne Boyer. 2015. When Users with preferences different from others get inaccurate recommendations. In *11th International Conference on Web Information Systems and Technologies*. Springer, 191–210.

[17] F. Maxwell Harper and Joseph A. Konstan. 2016. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4 (2016), 19:1–19:19.

[18] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *Proc. ICML*. PMLR, 1929–1938.

[19] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* 102, 46 (2005), 16569–16572.

[20] Kajta Hofmann, Bhaskar Mitra, Filip Radlinski, and Milad Shokouhi. 2014. An Eye-tracking Study of User Interactions with Query Auto Completion. In *Proc. CIKM*. ACM, 549–558.

[21] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*. Ieee, 263–272.

[22] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. 2007. Evaluating the accuracy of Implicit feedback from Clicks and Query Reformulations in Web Search. *ACM TOIS* 25, 2 (2007).

[23] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2018. Homophily influences ranking of minorities in social networks. *Scientific reports* 8, 1 (2018), 1–12.

[24] Sanjay Krishnan, Jay Patel, Michael J Franklin, and Ken Goldberg. 2014. A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. In *Proc. RecSys*. 137–144.

[25] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science* 65, 7 (2019), 2966–2981.

[26] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proc. SIGIR*. 831–840.

[27] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *Proc. RecSys*. 467–471.

[28] Yiming Liu, Xuezhi Cao, and Yong Yu. 2016. Are You Influenced by Others When Rating? Improve Rating Prediction by Conformity Modeling. In *Proc. RecSys*. 269–272.

[29] Benjamin M Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. 2007. Collaborative filtering and the missing at random assumption. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*. 267–275.

[30] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing search engines for differential satisfaction across demographics. In *Proc. WWW*. 626–633.

[31] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[32] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. 2020. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proc. Web Conference*. 1194–1204.

[33] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.

[34] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proc. SIGKDD*. 2219–2228.

[35] Catherine Stinson. 2021. Algorithms are not neutral: Bias in collaborative filtering. *arXiv preprint arXiv:2105.01031* (2021). arXiv:2105.01031 [cs.CY]

[36] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*. 923–932.

[37] Tom Sühr, Asia J Biega, Meike Zehlike, Krishna P Gummadi, and Abhijnan Chakraborty. 2019. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proc. SIGKDD*. 3082–3092.

[38] Guang Wang, Yongfeng Zhang, Zhihan Fang, Shuai Wang, Fan Zhang, and Desheng Zhang. 2020. FairCharge: A data-driven fairness-aware charging recommendation system for large-scale electric taxi fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–25.

[39] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).