Popularity Debiasing from Exposure to Interaction in Collaborative Filtering

Yuanhao Liu
Data Intelligence System Research Center, Institute of Computing Technology, CAS
University of Chinese Academy of Sciences, Beijing, China
liuyuanhao20z@ict.ac.cn

Qi Cao*
Data Intelligence System Research Center, Institute of Computing Technology, CAS
caoqi@ict.ac.cn

Huawei Shen*
Data Intelligence System Research Center, Institute of Computing Technology, CAS
University of Chinese Academy of Sciences, Beijing, China
shenhuawei@ict.ac.cn

Yunfan Wu
Data Intelligence System Research Center, Institute of Computing Technology, CAS
University of Chinese Academy of Sciences, Beijing, China
wuyunfan19b@ict.ac.cn

Shuchang Tao
Data Intelligence System Research Center, Institute of Computing Technology, CAS
University of Chinese Academy of Sciences, Beijing, China
taoshuchang18z@ict.ac.cn

Xueqi Cheng
CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, CAS
University of Chinese Academy of Sciences, Beijing, China
cxq@ict.ac.cn

ABSTRACT

Recommender systems often suffer from popularity bias, where popular items are overly recommended while sacrificing unpopular items. Existing researches generally focus on ensuring the number of recommendations (*exposure*) of each item is equal or proportional, using inverse propensity weighting, causal intervention, or adversarial training. However, increasing the exposure of unpopular items may not bring more clicks or interactions, resulting in skewed benefits and failing in achieving real reasonable popularity debiasing. In this paper, we propose a new criterion for popularity debiasing, i.e., *in an unbiased recommender system, both popular and unpopular items should receive Interactions Proportional to the number of users who Like it, namely IPL criterion*. Under the guidance of the criterion, we then propose a debiasing framework with IPL regularization term which is theoretically shown to achieve a win-win situation of both popularity debiasing and recommendation performance. Experiments conducted on four public datasets demonstrate that when equipping two representative collaborative filtering models with our framework, the popularity bias is effectively alleviated while maintaining the recommendation performance.

CCS CONCEPTS

• **Human-centered computing → Collaborative filtering**.

KEYWORDS

collaborative filtering, popularity debiasing, new criterion

*Corresponding authors.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591947

ACM Reference Format:
Yuanhao Liu, Qi Cao, Huawei Shen, Yunfan Wu, Shuchang Tao, and Xueqi Cheng. 2023. Popularity Debiasing from Exposure to Interaction in Collaborative Filtering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591947

1 INTRODUCTION

Recommender systems aim to help users find interesting items and are used in various online services, e.g., online shopping [15, 16, 32], social networks [7, 28, 30], and video sites [6, 8, 23]. Collaborative filtering, a widely used recommendation approach [11, 12, 17, 34], learns models for recommendation from user-item interactions. Unfortunately, since the user-item interactions usually exhibit long-tail distribution in terms of item popularity [19, 36], the models trained via the skewed interactions are often observed to further amplify the popularity bias by over-recommending popular items and under-recommending unpopular items [2, 22]. Such popularity bias may lead to terrible results such as Matthew Effect [24], hindering the long-term and sustainable development of recommender systems.

Researchers propose different methods for popularity debiasing. Early works generally follow a paradigm of inverse propensity weighting [3, 14, 18, 20, 26], i.e., increase the weight of unpopular items during training. Later, researchers introduce causal methods to remove the harmful impact of popularity [31, 37–39]. Adversarial learning is also used to enhance the distribution similarity of prediction scores between items to alleviate popularity bias [21, 41].

Despite the various existing methods, they generally focus on making the number of recommendations (the *exposure*) of items equal to each other or proportional to their popularity [20, 31]. Fig. 1(a) shows the ratio of items that get unbiased exposure amount in existing representative debiasing method [31], which helps 23% unpopular items avoid being under-recommended. However, blindly increasing the exposure of unpopular items may not significantly

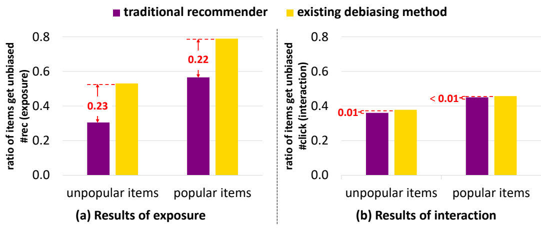

Figure 1: The proportion of items that get the unbiased (a) *exposure* (b) *interaction* amount before and after debiasing. The unbiased number of exposures/interactions of the item is proportional to the item's popularity. Popular items are the ones whose popularity exceeds 80% items.

increase their interactions [1], since they may be recommended to users who dislike them. Taking two similar books as an example, one is recommended to users who like it and the other is recommended more to those who don't like it. Although they are treated similarly in exposure, the latter is likely to earn less because fewer users will buy it, and thus still be plagued by popularity bias. As shown in Fig. 1(b), there are hardly any extra unpopular items (less than 0.01) that can get an unbiased amount of interactions through existing debiasing methods. In other words, simply increasing the exposure of unpopular items in existing methods may not necessarily lead to more clicks or interactions, resulting in skewed benefits and failing in achieving real reasonable popularity debiasing.

In this paper, we propose a new criterion for popularity debiasing, i.e., in a popularity unbiased recommender system, both popular and unpopular items should receive *interactions proportional* to the number of users who *like* it, namely ***IPL*** criterion. We propose a regularization-based debiasing framework following this criterion. However, in offline scenarios, we cannot observe the number of interactions that items can get in recommendation. In addition, due to the missing-not-at-random phenomenon of data, users who like the item are also invisible. In order to estimate both the number of interactions and users who like the items using observational data, we propose an inverse-propensity-score-based estimation method and use it to formalize the IPL regularization term for debiasing.

We conduct experiments on four public datasets and equip our framework with the representative matrix factorization [17] and the state-of-the-art LightGCN [11] for collaborative filtering. Experimental results show that our framework achieves the best effectiveness of debiasing while even further improving the recommendation performance. Further analysis is also offered to deepen the understanding of the success of our framework.

In summary, the contribution of this work is threefold.

- We propose a new criterion for popularity debiasing, i.e., an item should receive interactions proportional to the number of users who like it.
- We propose an estimation method to measure IPL in offline data, and further propose a debiasing framework, which is theoretically shown to achieve a win-win situation of both popularity debiasing and recommendation performance.
- Experiments conducted on four public datasets demonstrate the effectiveness of our framework when equipped with two representative collaborative filtering models.

[1]User *interactions*, such as clicking, purchasing, or expressing liking for the recommended item, can directly represent that the item benefits from the recommendation.

2 METHOD

This section describes the problem formulation and the proposed new criterion for popularity debiasing, as well as a general debiasing framework.

2.1 Collaborative Filtering

Suppose we have N users $\mathcal{U} = \{u_1, u_2, ..., u_N\}$ and M items $\mathcal{I} = \{i_1, i_2, ..., i_M\}$, collaborative filtering aims to predict the preference score $\widehat{y}_{u,i}$ of item i by user u based on observed interactions. For training a collaborative filtering model, Bayesian Personalized Ranking (BPR) [25] is one of the most widely adopted training loss,

$$L_{\text{BPR}} = -\sum_{u\in\mathcal{U}}\sum_{i^+\in\mathcal{I}_u^+}\sum_{i^-\in\mathcal{I}/\mathcal{I}_u^+} \ln(\sigma(\widehat{y}_{u,i^+} - \widehat{y}_{u,i^-}) + \lambda_\Theta\|\Theta\|_F^2,$$

where $\mathcal{I}_u^+$ is the set of items interacted with u in training data.

2.2 New Criterion for Popularity Debiasing

The new criterion is defined as a popular or unpopular item should receive interactions proportional to the number of users who like it, namely IPL criterion, which is formally defined as:

$$\forall i, j \in \mathcal{I}^{\text{pop}} or \mathcal{I}^{\text{unpop}}, \frac{C_i}{Q_i} = \frac{C_j}{Q_j}, \quad (1)$$

where $\mathcal{I}^{pop}$ and $\mathcal{I}^{unpop}$ are popular and unpopular items respectively, C_i is the number of interactions i can get from recommendation, and Q_i is the number of interactions that i can get when exposed to all users, reflecting the number of users who like item i. Such a criterion takes into account two aspects:

- The benefit of an item in recommendation is more about the interactions it receives, instead of massive useless recommendations. Hence, we take the interactions as the indicator.
- The more users like an item, the more interactions it deserves. Hence, an unbiased recommendation should ensure interactions is proportional to the users who like the item.

A recommender system is ideally unbiased if it satisfies the IPL criterion above, which not only provides a more reasonable goal for debiasing but also eliminates the trade-off between debiasing and recommendation performance (see Section 2.5).

2.3 Estimation in offline data

The IPL criterion provides a great goal for popularity debiasing. However, due to the incomplete observation in offline data, it is hard to obtain the real number of interactions received from recommendations (C_i) and from all users (Q_i). Although some methods tried to consider interaction amount in debiasing [40, 41], they generally obtain the C_i directly in the online scenario or ignore the incompleteness of the data, making it difficult to apply offline.

To solve the above problem, we propose an estimator for C_i and Q_i based on the inverse-propensity-score (IPS) [13] technique.

For estimating C_i, note that $C_i = \sum_{u\in\mathcal{U}_i} R_{u,i}$, where $\mathcal{U}_i$ is users interact with i when i is exposed to all users, and $R_{u,i} = 1$ if i is recommended to u otherwise 0. Given $C_i^{\text{IPS}} = \sum_{u\in\mathcal{U}_i} \frac{O_{u,i}}{P_{u,i}} R_{u,i}$ where $P_{u,i} = \Pr(O_{u,i} = 1)$ and $O_{u,i}$ is a binary variable of which 1 means that an interaction between u and i is observed, we can easily have $\mathbb{E}_O\left[C_i^{\text{IPS}}\right] = C_i$. In other words, C_i^{IPS} is an unbiased estimation of C_i. For estimating $P_{u,i}$ in C_i^{IPS}, we follow the broadly adopted approach in [35], which assume that $P_{u,i} = P_{*,i} = p_{*,i}^{\text{expose}} * p_{*,i}^{\text{like}}$, where $p_{*,i}^{\text{expose}}$ and $p_{*,i}^{\text{like}}$ are the probability that item i is exposed and liked, and $p_{*,i}^{\text{like}}$ is estimated to be proportional to $Q_i = |\mathcal{U}_i|$. Then we have $P_{*,i} \propto p_{*,i}^{\text{expose}} * Q_i$. Denoting the observed popularity of item i as $Q_i^* = |\mathcal{U}_i^*|$, we have a binomial distribution of $Q_i^* \sim \mathcal{B}(Q_i, p_{*,i}^{\text{expose}})$. Then we have $Q_i^* = p_{*,i}^{\text{expose}} * Q_i \propto P_{*,i}$, and

$$C_i^{\text{IPS}} = \sum_{u\in\mathcal{U}_i} \frac{O_{u,i}}{P_{u,i}} R_{u,i} = \sum_{u\in\mathcal{U}_i^*} \frac{R_{u,i}}{P_{u,i}} = \frac{\sum_{u\in\mathcal{U}_i^*} R_{u,i}}{P_{*,i}} \propto \frac{C_i^*}{Q_i^*}, \tag{2}$$

where $C_i^* = \sum_{u\in\mathcal{U}_i^*} R_{u,i}$ is the number of observed interactions in recommendation.

To estimate Q_i, recall that $Q_i = Q_i^*/p_{*,i}^{\text{expose}}$. We assume that $p_{*,i}^{\text{expose}} \propto \left(Q_i^*\right)^\gamma$ [35], where γ characterizes the popularity bias during data generation [27], and can be estimated by fitting the data. Then we have

$$Q_i = \frac{Q_i^*}{p_{*,i}^{\text{expose}}} \propto (Q_i^*)^{1-\gamma}. \tag{3}$$

With observed C_i^* and Q_i^* in offline data, we can measure the IPL criterion by $\frac{C_i^{\text{IPS}}}{Q_i} \propto \frac{C_i^*}{(Q_i^*)^{2-\gamma}} = r_i$. Then we can rewrite Eq. (1) as:

$$\forall i, j \in \mathcal{I}^{\text{pop}} or \mathcal{I}^{\text{unpop}}, r_i = r_j. \tag{4}$$

2.4 Debiasing Framework

We further propose a debiasing framework with a regularization term considering the IPL criterion during model training.

Considering a recommender model with parameter Θ, we have

$$\widehat{r}_i = \mathbb{E}[r_i|\Theta] = \frac{\mathbb{E}\left[C_i^*|\Theta\right]}{(Q_i^*)^{2-\gamma}} = \frac{\sum_{u\in\mathcal{U}_i^*} \Pr(u, i|\Theta)}{\left|\mathcal{U}_i^*\right|^{2-\gamma}} = \frac{\sum_{u\in\mathcal{U}_i^*} \sigma(\widehat{y}_{u,i})}{\left|\mathcal{U}_i^*\right|^{2-\gamma}},$$

where $\sigma(\widehat{y}_{u,i})$ indicates the probability that i will be recommended to u, $\mathcal{U}_i^*$ is all the users observed interact with i in training data.

Then the regularization term ensures IPL criterion is written as:

$$L_{\text{IPL}} = std_{i\in\mathcal{I}}(\widehat{r}_i) = \sqrt{\frac{1}{M}\sum_{i\in\mathcal{I}}(\widehat{r}_i - \frac{1}{M}\sum_{i\in\mathcal{I}}\widehat{r}_i)^2}, \tag{5}$$

which aims to reduce the r_i gap between all the M items according to Eq. (4) to achieve unbiased recommendation under the IPL criterion.

The final loss function for debiasing model training is:

$$\min_{\Theta} L_{\text{debias}} = L_{\text{BPR}} + \lambda_f * L_{\text{IPL}}, \tag{6}$$

where λ_f controls the strength of popularity debiasing.

2.5 Theoretical Analysis

Generally, popularity debiasing may bring a decrease in recommendation performance. In this section, we give a theoretical analysis to show that our proposed criterion can achieve a win-win situation for both recommendation performance and popularity debiasing.

Proposition 2.1. *Given an acceptable recommendation performance with recall = c, there always exists a recommendation result that satisfies the requirements of both the recall c and the IPL criterion.*

Proof. For recommendations with recall $c = \frac{\sum_{i\in\mathcal{I}} C_i}{\sum_{i\in\mathcal{I}} Q_i}$, to meet the IPL criterion, we should have $\forall i \in \mathcal{I}, r_i = \frac{C_i}{Q_i} = \frac{\sum_{i\in\mathcal{I}} C_i}{\sum_{i\in\mathcal{I}} Q_i} = c$.

Given $k \in \{1, 2, ..., M\}$ denoting the number of items recommended to each user, if a recommendation with recall c cannot meet the IPL criterion, then there always exists a discriminated item i that has $r_i = \frac{C_i}{Q_i} = \frac{C_i}{|\mathcal{U}_i|} < c$. In other words, there are at least $(1-c)|\mathcal{U}_i|$ users like item i while cannot further have i in their recommendation list, which means that such a user u will necessarily lose the recommendation of another item $j \in \mathcal{I}_u$ (u likes j) and make $r_j < c$ if adding item i into the recommendation list of user u. For better understanding, we denote the set of items at risk of discrimination (like j) as $\mathcal{I}^- \triangleq \left\{j\left|\frac{|\mathcal{U}^-\cap\mathcal{U}_j|}{|U_j|} > 1-c\right.\right\}$, where $\mathcal{U}^- \triangleq \left\{u\middle||\mathcal{I}_u| > k\right\}$. The above-mentioned user u likes at least k items in $\mathcal{I}^-$. Now we consider the existence of u, which is the necessary condition for the existence of the discriminated item i:

$$\textbf{condition-1}: \exists u \in \mathcal{U}, |\mathcal{I}_u \cap \mathcal{I}^-| > k. \tag{7}$$

Given $|\mathcal{I}_u|$ follows Pareto distribution $\Pr(|\mathcal{I}_u|=x) = (\beta-1)x^{-\beta}$ with $\beta > 1$, the probability that user $u \in \mathcal{U}^-$ equals to $p = \Pr(|\mathcal{I}_u| > k) = k^{1-\beta}$. Therefore we have $\Pr(|\mathcal{U}^-\cap\mathcal{U}_i| = x) \sim B(x; |\mathcal{U}_i|, p)$, where $B(\cdot; n, p)$ is the Binomial distribution with the number of experiments n and the probability of success p. Then we have $\Pr(i \in \mathcal{I}^-) = \Pr(|\mathcal{U}^- \cap \mathcal{U}_i| > (1-c)|\mathcal{U}_i|) < F(c|\mathcal{U}_i|; |\mathcal{U}_i|, 1-p)$, where $F(\cdot; n, p)$ is the cumulative distribution function of $B(\cdot; n, p)$. With the Chernoff bound $F(x; n, p) \le \exp\left(-nD(\frac{x}{n}||p)\right)$, where $D(a||p) = a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}$, we further have $\Pr(i \in \mathcal{I}^-) \le \exp\left(-\left((1-c)\log\frac{(1-c)}{p} + c\log\frac{c}{1-p}\right)\right)$.

Then the probability that satisfying *condition-1* is:

$$\begin{aligned}
&1-\prod_{\substack{u\in\mathcal{U}\\|\mathcal{I}_u|>k}}\left(1-\sum_{j=k+1}^{|\mathcal{I}_u|}\binom{|\mathcal{I}_u|}{j}\Pr(i\in\mathcal{I}^-)^j(1-\Pr(i\in\mathcal{I}^-))^{|\mathcal{I}_u|-j}\right)\\
\le&1-\prod_{\substack{u\in\mathcal{U}\\|\mathcal{I}_u|>k}}\left(1-\sum_{j=k+1}^{|\mathcal{I}_u|}\binom{|\mathcal{I}_u|}{j}\Pr(i\in\mathcal{I}^-)^j\right)\\
\le&1-\prod_{\substack{u\in\mathcal{U}\\|\mathcal{I}_u|>k}}\left(1-\sum_{j=k+1}^{|\mathcal{I}_u|}\binom{|\mathcal{I}_u|}{j}\exp\left(-\left((1-c)\log\frac{(1-c)}{p}+c\log\frac{c}{1-p}\right)\right)^j\right)
\end{aligned} \tag{8}$$

Taking real-world datasets as an illustration, when specifying interaction recall $c = 0.99$, the probability of *condition-1* $< 10^{-10}$ on MovieLens-1M. Similar results were shown on other datasets. This show that there is almost always a recommendation result that satisfies IPL criterion while ensuring recommendation performance. □

3 EXPERIMENTAL RESULTS

3.1 Experiment Setup

Datasets. We conduct experiments on four public benchmark datasets: **MovieLens-1M** [9], **Gowalla** [4], **Yelp** [1], and **Amazon Book** [10]. We uniformly sample interaction records from each item, splitting them into 70%, 10%, and 20% as train, validation, and test sets. In addition, we follow [35] and use MF [17] to estimate γ of the datasets, which are 1.826, 1.285, 1.552, 1.446, respectively.

Baselines. We implement our method with MF [17] and LightGCN [11] [2], comparing with the following baselines: **IPW** [20] is a IPS-based re-weighting method. **PO** [40] eliminates popularity bias in

[2]The model implementations are accessible on Github: https://github.com/UnitDan/IPL

Table 1: Performance of recommendation and debiasing. Bold and underlined scores are the best and second-best performances.

	MovieLens-1M					Gowalla					Yelp					Amazon book				
	Pre↑	Recall↑	NDCG↑	MI↓	DI↓	Pre↑	Recall↑	NDCG↑	MI↓	DI↓	Pre↑	Recall↑	NDCG↑	MI↓	DI↓	Pre↑	Recall↑	NDCG↑	MI↓	DI↓
MF	19.977	27.685	30.777	2.138	3.723	4.738	17.763	13.149	0.601	3.067	2.557	10.871	6.848	0.782	3.989	2.618	12.066	7.979	0.757	3.061
PO	17.329	23.097	25.877	1.997	3.461	4.442	16.862	12.374	0.593	2.918	1.951	8.075	5.078	0.742	4.594	2.606	12.021	7.965	0.742	2.928
DPR	16.630	23.010	25.240	1.985	3.467	3.983	14.404	10.862	0.611	4.042	2.104	8.485	5.407	0.766	5.757	2.238	10.219	7.596	0.728	5.194
IPW	16.350	22.110	23.540	2.073	3.500	4.021	15.151	10.360	0.588	2.974	1.950	8.081	4.813	0.718	3.882	2.220	8.751	7.011	0.735	3.004
MACR	17.248	23.253	25.782	2.211	3.458	3.977	14.694	10.879	0.584	2.905	1.685	6.732	4.285	0.747	4.891	2.140	11.875	7.889	0.726	3.173
IPL	**20.397**	**28.642**	**31.397**	**1.942**	**3.254**	**5.136**	**19.220**	**14.450**	**0.559**	**2.882**	**2.715**	**11.578**	**7.278**	**0.625**	**3.808**	**2.891**	**13.274**	**8.914**	**0.725**	**2.889**
LGCN	19.808	27.217	30.599	2.010	4.032	5.381	19.894	15.355	0.719	2.623	2.934	12.410	7.948	0.723	4.403	2.905	12.978	8.805	0.777	3.016
PO	19.830	27.206	30.475	1.998	3.356	5.221	18.373	15.176	0.692	3.131	2.861	12.220	7.767	0.738	4.487	2.880	11.675	7.992	0.720	3.031
DPR	17.682	22.861	28.199	1.922	3.417	4.533	16.654	12.458	0.638	2.760	2.106	11.255	6.152	0.728	6.809	2.422	9.427	6.854	0.736	4.334
IPW	13.689	16.700	19.590	1.959	3.777	2.950	10.960	8.340	0.685	3.357	1.913	9.159	5.822	0.722	6.026	2.075	8.427	6.999	0.716	4.922
MACR	18.724	26.251	27.945	1.985	3.180	5.183	19.022	14.831	0.630	3.058	2.335	11.684	7.702	0.725	4.651	2.108	10.133	7.255	0.757	3.799
IPL	**19.979**	**27.545**	**30.949**	**1.921**	**3.005**	**5.385**	**19.897**	**15.365**	**0.621**	**2.011**	**2.954**	**12.440**	**8.013**	**0.720**	**3.848**	**2.913**	**12.991**	**8.844**	**0.704**	**2.886**

item ranking by regularization. **MACR** [31] removes the causal effect of popularity on preference scores. **DPR** [41] uses adversarial training for item group fairness. In addition, we also optimize **MF** and **LightGCN** using BPR loss as the basis for comparison.

Evaluation metrics. We evaluate recommendation performance by *Precision@k*, *Recall@k* and *NDCG@k*. We set $k = 20$ following [29], and show all these metrics after multiplying by 100 [33]. To evaluate the effectiveness of popularity debiasing, we use the mutual information (*MI*) [5] of r_i and item popularity Q_i^* to measure the correlation between them. We also propose a metric that is consistent with the IPL criterion, called *Deviation of Interaction distribution* (*DI*). $DI = \frac{std_{i \in \mathcal{I}}(r_i)}{mean_{i \in \mathcal{I}}(r_i)}$. Lower values of MI and DI indicate less popularity bias in the recommendation result.

3.2 Performance Evaluation

To demonstrate the effectiveness of our method, we evaluate recommendation performance and debiasing results on four benchmark datasets. As shown in Table 1, MF and LightGCN achieve good recommendation performance, but with a serious popularity bias, reflected in a high value of MI and DI. For debiasing baselines, they mitigate the popularity bias to some extent, however, at the cost of severely sacrificing recommendation performance. For our method, IPL shows superiority in both recommendation accuracy and debiasing. Especially, IPL achieves the best recommendation performance on all datasets, even outperforming MF and LightGCN. As for debiasing, our IPL performs best on MI and DI. These results demonstrate the effectiveness of our IPL method in terms of both recommendation performance and popularity debiasing.

3.3 IPL Method Win-Win Evaluation

To verify the conclusion in Section 2.5 that our proposed method can achieve a win-win situation in recommendation and debiasing, we compared the performance of IPL with different strengths of popularity debiasing and basic models. We used $1/DI$ to represent the debiasing performance and an unbiased recall metric ($Recall_{\text{SNIPS}}$) [35] to represent the recommendation accuracy. We experimented with 20 equally spaced values of λ_f on a logarithmic scale from $1e-2$ to $1e-6$. Fig. 2 demonstrates that the experimental results of IPL are consistently located in the upper right space of the basic model results, indicating that IPL can simultaneously enhance recommendation and debiasing performance.

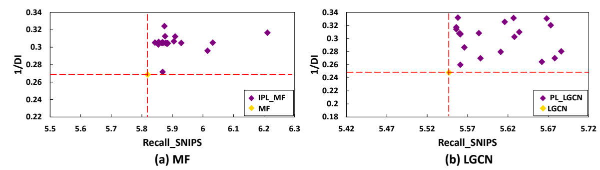

Figure 2: Performance of basic models and IPL with different λ_f on recommendation and debiasing.

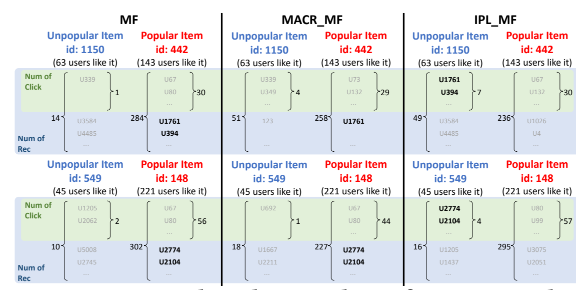

Figure 3: Case study. The number of recommendations (blue) is on the left, and that of effective recommendations (green) is on the right. Users who prefer unpopular items are bold.

3.4 Case Study

Fig. 3 demonstrates how IPL can achieve a win-win situation using Movie-Lens as an example. By comparing recommendation lists of two pairs of popular and unpopular items, we found that MF amplifies popularity bias by over-recommending popular items (442 and 148) and under-recommending unpopular ones (1150 and 549). The existing debiasing method (MACR_MF) blindly recommends more to unpopular items, sacrificing recommendations of popular items. In contrast, our IPL_MF recommends the unpopular item (1150) to users (U1761 and U394) who prefer unpopular items, thereby alleviating the popularity bias while achieving better recommendation accuracy.

4 CONCLUSION

In this work, we propose a new criterion to measure popularity bias, i.e., an item should get interactions proportional to the number of users who like it (IPL). We further propose a debiasing framework based on IPL. Both theoretical analysis and experimental results demonstrate that the proposed method can achieve a win-win situation for both recommendation boosting and popularity debiasing.

ACKNOWLEDGMENTS

This work is funded by the National Key R&D Program of China (2022YFB3103700, 2022YFB3103701), and the National Natural Science Foundation of China under Grant Nos. 62272125, U21B2046. Huawei Shen is also supported by Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

[1] 2022. Yelp Open Dataset. https://www.yelp.com/dataset.

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Unfairness of Popularity Bias in Recommendation. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (CEUR@RecSys '19)*.

[3] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D Smith. 2006. From niches to riches: Anatomy of the long tail. *Sloan management review* 47, 4 (2006), 67–71.

[4] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '11)*. 1082–1090.

[5] Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

[6] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the 4th ACM Conference on Recommender systems (RecSys '10)*. 293–296.

[7] Souvik Debnath, Niloy Ganguly, and Pabitra Mitra. 2008. Feature weighting in content based recommendation system using social network analysis. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. 1041–1042.

[8] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. 2016. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics* 5, 2 (2016), 99–113.

[9] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TIIS)* 5, 4 (2015), 1–19.

[10] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. 507–517.

[11] 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 (SIGIR '20)*. 639–648.

[12] Junjie Huang, Qi Cao, Ruobing Xie, Shaoliang Zhang, Feng Xia, Huawei Shen, and Xueqi Cheng. 2023. Adversarial Learning Data Augmentation for Graph Contrastive Learning in Recommendation. In *Database Systems for Advanced Applications: 28th International Conference, DASFAA 2023, Tianjin, China, April 17–20, 2023, Proceedings, Part II*. Springer, 373–388.

[13] Thorsten Joachims and Adith Swaminathan. 2016. Counterfactual evaluation and learning for search, recommendation and ad placement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '16)*. 1199–1201.

[14] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2014. Correcting Popularity Bias by Enhancing Recommendation Neutrality.. In *Proceedings of the Workshop on ACM Conference on Recommender Systems (CEUR@RecSys '14)*.

[15] Hyea Kyeong Kim, Jae Kyeong Kim, and Young U Ryu. 2009. Personalized recommendation over a customer network for ubiquitous shopping. *IEEE Transactions on Services Computing* 2, 2 (2009), 140–151.

[16] Jae Kyeong Kim, Yoon Ho Cho, Woo Ju Kim, Je Ran Kim, and Ji Hae Suh. 2002. A personalized recommendation procedure for Internet shopping support. *Electronic commerce research and applications* 1, 3-4 (2002), 301–313.

[17] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[18] Adit Krishnan, Ashish Sharma, Aravind Sankar, and Hari Sundaram. 2018. An adversarial approach to improve long-tail performance in neural collaborative filtering. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 1491–1494.

[19] Jingjing Li, Ke Lu, Zi Huang, and Heng Tao Shen. 2019. On both cold-start and long-tail recommendation with social data. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 33, 1 (2019), 194–208.

[20] Dawen Liang, Laurent Charlin, and David M Blei. 2016. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI. AUAI*.

[21] Zhongzhou Liu, Yuan Fang, and Min Wu. 2022. Mitigating Popularity Bias for Users and Items with Fairness-centric Adaptive Recommendation. *ACM Transactions on Information Systems (TOIS)* (2022).

[22] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2145–2148.

[23] Tao Mei, Bo Yang, Xian-Sheng Hua, and Shipeng Li. 2011. Contextual video recommendation by multimodal relevance and user feedback. *ACM Transactions on Information Systems (TOIS)* 29, 2 (2011), 1–24.

[24] Matjaž Perc. 2014. The Matthew effect in empirical data. *Journal of The Royal Society Interface* 11, 98 (2014), 20140378.

[25] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI '09)*. 452–461.

[26] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2017. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'17)*.

[27] Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM conference on Recommender systems*. 125–132.

[28] Frank Edward Walter, Stefano Battiston, and Frank Schweitzer. 2008. A model of a trust-based recommendation system on a social network. *Autonomous Agents and Multi-Agent Systems* 16, 1 (2008), 57–74.

[29] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. 165–174.

[30] Zhi Wang, Lifeng Sun, Wenwu Zhu, Shiqiang Yang, Hongzhi Li, and Dapeng Wu. 2012. Joint social and content recommendation for user-generated videos in online social network. *IEEE Transactions on Multimedia* 15, 3 (2012), 698–709.

[31] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1791–1800.

[32] Le Wu, Peijie Sun, Yanjie Fu, Richang Hong, Xiting Wang, and Meng Wang. 2019. A neural influence diffusion model for social recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. 235–244.

[33] Yunfan Wu, Qi Cao, Huawei Shen, Shuchang Tao, and Xueqi Cheng. 2022. INMO: A Model-Agnostic and Scalable Module for Inductive Collaborative Filtering. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. 91–101.

[34] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the 9th International Conference on Web Search and Data Mining (WSDM '16)*. 153–162.

[35] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM conference on recommender systems (RecSys '18)*. 279–287.

[36] Yin Zhang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Lichan Hong, and Ed H Chi. 2021. A model of two tales: Dual transfer learning framework for improved long-tail item recommendation. In *Proceedings of the Web Conference 2021 (WWW '21)*. 2220–2231.

[37] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. 11–20.

[38] Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, and Wei Wu. 2021. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *arXiv preprint arXiv:2109.07946* (2021).

[39] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021 (WWW '21)*. 2980–2991.

[40] Ziwei Zhu, Yun He, Xing Zhao, Yin Zhang, Jianling Wang, and James Caverlee. 2021. Popularity-opportunity bias in collaborative filtering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*. 85–93.

[41] Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. Measuring and mitigating item under-recommendation bias in personalized ranking systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 449–458.