Causal Intervention for Popularity Bias in the Recommender System

Ziling Gong
zgong@ucsd.edu

Yihan Xue
y6xue@ucsd.edu

Jiawei Wang
jiw076@ucsd.edu

Biwei Huang
bih007@ucsd.edu

Babak Salimi
bsalimi@ucsd.edu

Abstract

The prevalence of popularity bias in recommender systems contributes to the promotion of globally popular items, diminishing the diversity of recommended products and suppressing personalized selections. This bias acts as a confounding factor in the intricate relationship between items and user interactions. As the recommender system continues to prioritize popular items, it reinforces their promotion to users, amplifying the bias in a self-reinforcing loop. This project places users as a central stakeholder in recommender system applications, aiming to diversify the presented products for a more individualized experience. We introduce two distinct approaches to address popularity bias. The Popularity-bias De-confounding and Adjusting (PDA) model enables control over the strength of popularity bias. The Disentangling Interest and Conformity with Causal Embedding (DICE) model captures causes separately to be integrated into RecSys. This paper concludes that leveraging the contribution of popularity bias in the Recommender System leads to better recommendation performance, achieving higher recall rates.

Code: https://github.com/bettygong/DSC180B-popularity-bias-in-Netflix-dataset.git

1 Introduction

In recent years, the pervasive integration of Artificial Intelligence has revolutionized personal convenience, particularly through the widespread adoption of recommender systems across diverse platforms. These systems, designed to analyze individual user preferences and offer diverse content recommendations, have become integral to enhancing user experiences. However, conventional recommender systems incorporate various biases into their algorithms, impacting the recommended content for each user.

Despite the advancements these systems bring, a significant challenge undermines their potential: the inherent biases within their algorithms, most notably popularity bias. Popularity bias occurs when a recommender system disproportionately favors items that are already popular, overshadowing personalized user preferences. This bias leads to a uniformity in recommendations, whereby widely appreciated items gain further visibility at the expense of lesser-known products. It reduces the diversity of content suggested to users, which not only narrows the scope of discovery but also concerns about the fairness and inclusivity of the recommender systems.

Addressing this bias is crucial for the evolution of recommender systems. Existing research has taken varied approaches to mitigate the effects of popularity bias, from strategies that seek to eliminate the influence of item popularity to more complex methods that incorporate different measures of popularity into the recommendation process. The goal is to ensure the algorithms offer a balanced mix of content that accurately reflects the diverse preferences of their user base. By promoting a broader spectrum of suggestions, recommender systems can move towards a more equitable and inclusive model that values diversity as much as relevance. Moreover, tackling popularity bias opens the door to enhanced discoverability for emerging content creators and niche products. It refines algorithms and shapes our digital experiences.

In response to the challenges of popularity bias, our work introduces a causal inference methodology to intricately de-bias user history data used by recommender systems. The two models, the Popularity-bias De-confounding and Adjusting model (PDA) and the Disentangling Interest and Conformity with causal Embeddings model (DICE), aim to enhance personalization in content recommendations. Our findings suggest that integrating popularity bias within our algorithms, instead of removing it, significantly improves the performance, leading to higher recall rates. This approach aligns recommendations more closely with user preferences.

1.1 Literature Review

Popularity bias plays a crucial role in recommendation systems. Zhang et al. (2021) suggests that blindly pursuing unbiased learning may remove beneficial patterns from the data. Projects exhibit uneven interaction frequencies, and over-recommending popular items can amplify biases. Thus, they propose the PDA model to enhance recommendation accuracy by utilizing popularity bias. PDA eliminates confounding popularity bias during model training and adjusts recommendation scores and expected popularity bias through causal intervention, enabling the system to offer fairer and more just recommendations.

Another approach to reducing popularity bias is employed by Abdollahpouri, Burke and Mobasher (2019), who utilize a probabilistic framework called xQuAD. The xQuAD framework initially ranks recommendation lists based on user information and then reshuffles them to generate new ranked lists for users. Both PDA and xQuAD show improved performance, demonstrating the influence of popularity bias on recommendation systems, but this approach needs a more logical process. However, compared to xQuAD, PDA's method is more rational and accurate.

Another method for controlling popularity bias is the DICE model proposed by Zheng et al. (2021). The DICE model distinguishes between user interest and conformity, embedding them into different models. In this case, popularity bias is considered a cause of user behavior, and this model exhibits promising results in experiments.

Tobias Schnabel and his team propose another method for controlling popularity bias called The Inverse-Propensity-Scoring (IPS) method. This approach handles selection biases by adapting models and adjusting the data distribution to be even through interaction reweighting. However, unlike DICE, this method does not perform well in practice due to its high model variance.(Schnabel et al. 2016)

1.2 Dataset Description

Two datasets, Douban and Netflix, are utilized for model comparison.

Douban dataset, encompassing movie and book ratings collected from a Chinese social networking platform. These datasets encompass essential fields such as user_id, item_id, timestamp, and users' ratings spanning the period from 2005 to 2017. Ratings within this dataset range from 1 to 5, with an empty rating indicating a user's negative interaction with the item, ie. click. To optimize our model's efficiency, we considered all rating records as positive samples and applied a filter to include data only up to 2010. Following the data cleaning phase, our dataset retained 7,174,218 interactions involving 47,890 users and 26,047 items.

Netflix dataset is employed from Kaggle Netflix Prize data, specifically, we used the `combined data_1_subset`. This subset was chosen due to its considerable size and relevance to the research objectives. It includes necessary features, such as user ID, item ID, user-assigned ratings, and timestamp details. To accommodate PDA-model's requirements, the dataset undergoes temporal splitting, addressing the influence of popularity drift over time. On the other hand, for the DICE model, rating-5 data is extracted to be positive samples, and negative samples are drawn artificially. The dataset comprises 122,958 unique users and 1,343 distinct items. The split training, test, and validation dataset consists of 12,442,388 entries, 1,942,172 entries, and 833,866 entries.

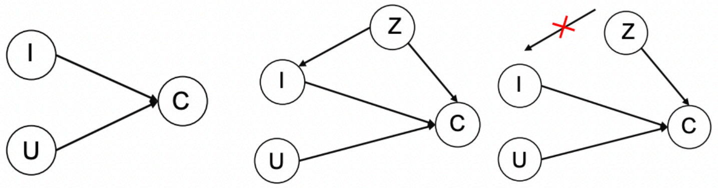
2 Methods

2.1 PDA

2.1.1 Popularity Bias Causal Graph

Illustrated in 1(a), conventional recommendation methods consider user information (U node) and item information (I node) to explain the interaction, ie. the click (C node). The causal relationship is known as the collider effect. However, in real-world scenarios, a third factor, popularity (Z node), influences both the interaction (C node) and the item information (I node), introducing a confounding element depicted in 1(b).

On one hand, individuals often exhibit a conformity mentality, leaning towards popular items in their purchasing decisions (Z -> C). Consequently, the more popular items become, the higher the likelihood of user interaction. On the other hand, recommender models tend to perpetuate biases present in the data, disproportionately highlighting popular items (Z -> I -> C), thus exacerbating the popularity bias (Zhang et al. 2021). The popularity factor in the second path serves as a bias amplification, becoming the target of the removal process.



(a) Causal graph of traditional methods. (b) Causal graph that considers item popularity. (c) We cut off $Z \rightarrow I$ for model training.

Figure 1: PDA causal graph

To eliminate the Z -> I path, we employ do-calculus, as advocated by (Zhang et al. 2021), intervening in the recommended item I to render it impervious to the effects of popularity Z. This approach, rooted in causal inference, serves as a remedy, effectively debiasing the data and mitigating the impact of popularity.

2.1.2 Derive Predictive Model P(C|do(U, I))

Denote G as Figure 1 (b) and G' be figure 1(c).
P(C|do(U, I)) means how likely the user clicks (C) the item by fixing variables in nodes U and I. The reason for do(U, I) instead of do(I) is that the recommender system takes both U and I as input, and without losing generality, the backdoor path I <- Z ->C in G is blocked by do(U, I).

$$\begin{aligned} P(C \mid do(U, I)) &\overset{(1)}{=} P_{G'}(C \mid U, I) \\ &\overset{(2)}{=} \sum_{z} P_{G'}(C \mid U, I, z) P_{G'}(z \mid U, I) \\ &\overset{(3)}{=} \sum_{z} P_{G'}(C \mid U, I, z) P_{G'}(z) \\ &\overset{(4)}{=} \sum_{\sim} P(C \mid U, I, z) P(z), \end{aligned}$$

2.1.3 Step 1: Estimating P(C|do(U, I))

Let the parameters of the conditional probability function be θ, and we parameterize U=u, I=i, and popularity $Z = m_i^t$. We use L2 to optimize the pairwise BPR objective function on historical data D.

$$\max_{\Theta} \sum_{(u,i,j)\in\mathscr{D}} \log \sigma \left(P_{\Theta}\left(c = 1 \mid u, i, m_i^t\right) - P_{\Theta}\left(c = 1 \mid u, j, m_j^t\right)\right) \tag{1}$$

Here m_i^t represents the local popularity of item i on the stage t, $m_i^t = D_t^i / \sum D_j^t$, where we divided data D into T stages and D_i^t is the number of observed interactions for item i at stage t. According to (Zhang et al. 2021), the usage of local popularity is because the most recent data has a larger impact on the system's exposure mechanism.
Subsequently, we parametrize $P_\theta(c = 1|u, i, m_i^t) = \mathrm{ELU}'(f_\theta(u, i)) \cdot (m_i^t)^\gamma$.
The function $f_\theta(u, i)$ represents any user-item matching model, with Machine Factorization (MF) being our choice in this context. The parameter γ governs the intensity of the conformity effect; for instance, $\gamma = 0$ characterizes our PD model, which is devoid of popularity bias. In contrast, in the PDA model, a higher γ corresponds to a more pronounced impact.

In addition, the activation function ELU' is employed to guarantee positivity.

$$ELU'(x) = \begin{cases} e^x, & \text{if } x \leq 0 \\ x+1, & \text{else} \end{cases} \tag{2}$$

2.1.4 Step 2: Estimating $\sum P(C|U,I,z)P(z)$

$$\begin{aligned} P(C \mid do(U,I)) &= \sum_z P(C \mid U,I,z)P(z) \\ &= \sum_z ELU'\left(f_\Theta(u,i)\right) \times z^\gamma P(z) \\ &= ELU'\left(f_\Theta(u,i)\right) \sum_z z^\gamma P(z) \\ &= ELU'\left(f_\Theta(u,i)\right) E\left(Z^\gamma\right) \end{aligned}$$

We replace P(C|U, I, z) by the equation in step 1, and take ELU' out of summation. $\sum_z z^\gamma * P(z)$ is expectation equation of Z^γ. Since the expectation value is constant, we neglect the term and use $ELU(f_\theta(u,i))$ to estimate $P(C|do(U,I))$.

2.2 DICE

2.2.1 Causal implication

Causality implication: interest -> click <- conformity. The click is the collider of interest and conformity. By collider effect, interest and conformity are independent, but dependent conditioned on click.

2.2.2 Causal relationship for click, interest and conformity

SCM explains how click is generated from interest and conformity causes. The big picture is that:

$$\begin{aligned} X_{ui}^{int} &:= f_1\left(u, i, N^{int}\right), \\ X_{ui}^{con} &:= f_2\left(u, i, N^{con}\right), \\ Y_{ui}^{click} &:= f_3\left(X_{ui}^{int}, X_{ui}^{con}, N^{click}\right), \end{aligned} \tag{3}$$

2.2.3 Separate Embedding Overview

Separate embeddings for interest and conformity are trained for user (u_{int} and u_{con}) and item (i_{int} and i_{con}). The recommendation score is calculated by

$$\begin{aligned} s_{ui}^{\text{int}} &= \left\langle u^{(\text{int})}, i^{(\text{int})}\right\rangle, \quad s_{ui}^{con} = \left\langle u^{(\text{con})}, i^{(\text{con})}\right\rangle, \\ s_{ui}^{\text{click}} &= s_{ui}^{\text{int}} + s_{ui}^{\text{con}} \end{aligned} \tag{4}$$

Particularly, $s_{\text{click}} = \langle u_{\text{int}}, i_{\text{int}}\rangle + \langle u_{\text{con}}, i_{\text{con}}\rangle$, where $\langle ., .\rangle$ means dot product. From training and optimization, the embeddings are captured.

2.2.4 Mining Cause-specific Data

As we do not know the cause of the click, but the effect. We can only split observed training data O into two sets of cause-specific data, denoted O_1 and O_2. Specifically, O_1 represents conformity-dominated data, the instances where negative samples are less popular than positive samples. O_2 represents the interest-dominated data, where negative samples are more popular than positive ones.

To be elaborated, we result in these two data sets from two cause-specific cases.

- Case 1 (M_I): The negative item is less popular than the positive one.
  - The user clicks a popular item a, while not clicking an unpopular item b, which does not indicate where the user's interest lies. The clicks may come from conformity cause.

$$\begin{aligned} &M^c_{ua} > M^c_{ub}, \\ &M^I_{ua} + M^C_{ua} > M^I_{ub} + M^c_{ub}. \end{aligned} \tag{5}$$

- Case 2 (M_C): The negative item is more popular than the positive one.
  - The user clicks an unpopular item c, while not clicking the popular item d. The click is largely due to interest. Plus, the conformity implication is abased.

$$\begin{aligned} &M^I_{uc} > M^I_{ud}, M^C_{uc} < M^C_{ud} \\ &M^I_{uc} + M^C_{uc} > M^I_{ud} + M^C_{ud} \end{aligned} \tag{6}$$

From this learning of relative relations, we successfully separate user-item interaction into two cause-specific data sets O_1 and O_2.

2.2.5 Loss functions

To train the model, we decompose the loss function into four parts.

- The conformity modeling was applied on O_1 and O_2 in case 1 (W_I). (attach eq6)

$$L^{O_2}_{\text{interest}} = \sum_{(u,i,j)\in O_2} \text{BPR}\left(\left\langle u^{(\text{int})}, i^{(\text{int})}\right\rangle, \left\langle u^{(\text{int})}, j^{(\text{int})}\right\rangle\right). \tag{7}$$

- The interest modeling applied on O2 in case 2 (W_C). (attach eq7)

$$\begin{aligned}
L_{\text{conformity}}^{O_1} &= \sum_{(u,i,j)\in O_1} \mathrm{BPR}\left(\left\langle u^{(\text{con })}, i^{(\text{con })}\right\rangle, \left\langle u^{(\text{con })}, j^{(\text{con })}\right\rangle\right), \\
L_{\text{conformity}}^{O_2} &= \sum_{(u,i,j)\in O_2} -\mathrm{BPR}\left(\left\langle u^{(\text{con })}, i^{(\text{con })}\right\rangle, \left\langle u^{(\text{con })}, j^{(\text{con })}\right\rangle\right), \\
L_{\text{conformity}}^{O_1+O_2} &= L_{\text{conformity}}^{O_1} + L_{\text{conformity}}^{O_2}
\end{aligned} \tag{8}$$

- Estimating Clicks is the main target of RecSys applied on the entire data O, which is to maximize the margin between scores of positive items and negative items. (attach eq8)

$$L_{\text{click}}^{O_1+O_2} = \sum_{(u,i,j)\in O} \mathrm{BPR}\left(\left\langle u^t, i^t\right\rangle, \left\langle u^t, j^t\right\rangle\right) \tag{9}$$

  where u^t, i^t and j^t are concatenation of interest embedding and conformity embedding for user and item: $u^t = u^{(\text{int })} \left\|u^{(\text{con })}, i^t = i^{(\text{int })}\right\| i^{(\text{con })}, j^t = j^{(\text{int })} \| j^{(\text{con })}$

- The discrepancy task works as a constraint on interest embedding and conformity embedding.

Our final loss function is (attach eq10)

$$L = L_{\text{click}}^{O_1+O_2} + \alpha(L_{\text{interest}}^{O_2} + L_{\text{conformity}}^{O_1+O_2}) + \beta(L_{\text{discrepancy}}) \tag{10}$$

3 Results

The evaluation of the Popularity-bias De-confounding and Adjusting (PDA) and Disentangling Interest and Conformity with Causal Embedding (DICE) models on the Netflix and Douban datasets on Table1 reveals insightful variations in performance, which are indicative of the underlying user behaviors specific to each platform.

Specifically, on the Netflix dataset, the DICE model outperforms the others with a recall@20 of 0.1906, surpassing the PDA model's 0.181. The outperformance of DICE on Netflix can be attributed to its ability to disentangle the elements of individual preference and collective popularity. Netflix's user base may display more pronounced herd behavior, which is a tendency to follow popular choices, thereby making the platform's recommendation landscape more susceptible to popularity bias. The DICE model's effectiveness in separating the interest and conformity factors allows it to account for this herd behavior. It adjusts recommendations to ensure that the items' intrinsic appeal is not overshadowed by their popularity. By doing so, DICE can harness the popularity component as a valid input to the recommendation process rather than allowing it to skew the recommendations unfairly.

Conversely, the PDA model shines on the Douban platform, with a recall@20 of 0.0565, outperforming the baseline PD model's 0.0453. Douban, a platform with a strong emphasis on in-depth reviews and reflective user engagement, particularly including book reviews, is less influenced by mainstream herd behavior. The PDA model leverages this characteristic through its intervention method that curtails the reinforcement of popularity bias during model training. Additionally, it allows for a specific injection of popularity strength during the inference stage.

Table 1: Recommendation performance taking top-20 recommendations

Method	Douban	Netflix
PD	0.0453	0.133
PDA	**0.0565**	0.181
DICE	0.0273	**0.1906**

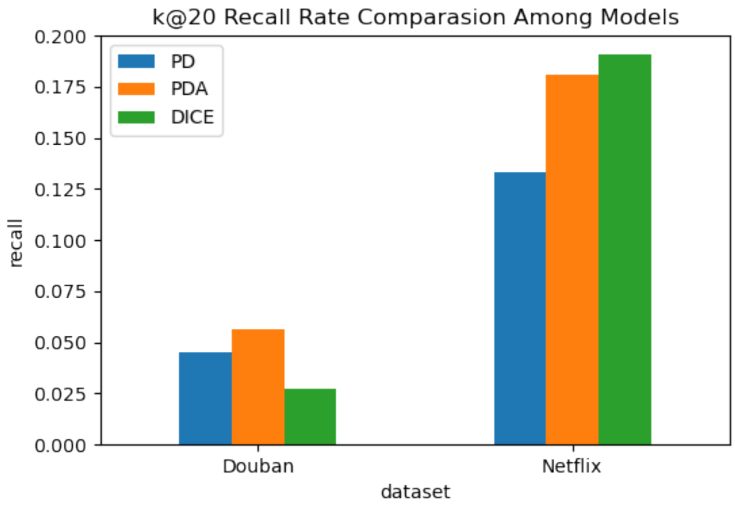


Figure 2: Recall@20 of PDA and DICE models

Table 2: Douban top-20 top-50 recommendations

Method	Top20	Top50
PD	0.0453	0.0843
$PDA-D_i$	0.0564	0.1066
$PDA-D_i^t$	**0.0565**	**0.1066**
DICE	0.273	0.0513

4 Applications, Implications, and Strategic Considerations

This section delves into the multifaceted implications of addressing popularity bias within recommender systems, emphasizing its real-world applications, accruing benefits, potential concerns, and inherent limitations. Our investigation aims to provide a holistic understanding of the impact of implementing bias-mitigation models, underscoring the necessity for achieving a harmonious balance between innovation and responsibility is paramount and necessitating ongoing efforts in innovation, transparency, and ethical practices to fully realize the potential of recommender systems in a diverse and dynamic digital landscape.

4.1 Real-World Applications

The mitigation of popularity bias in recommender systems finds application across a spectrum of digital platforms, facilitating a more nuanced interaction between users and content. Notably:

- **E-Commerce Platforms:** By ensuring a more diversified product recommendation, these models can significantly enhance customer satisfaction and retention on e-commerce sites. Users are more likely to discover niche products that align with their specific interests, potentially increasing sales diversity and supporting smaller or new vendors.
- **Content Streaming Services:** For platforms like Netflix or Spotify, the models promise to improve content discovery, pushing beyond mainstream hits to offer personalized suggestions based on a user's unique tastes. This not only improves user engagement but can also elevate lesser-known artists, creators, and genres, contributing to a more vibrant and diverse cultural landscape.
- **Social Media Platforms:** By broadening the range of content presented to users, recommender systems can mitigate echo chamber effects and promote a plurality of perspectives, enhancing the quality of discourse.
- **Online Advertising:** Improved targeting precision based on refined understanding of user interests has the potential to increase the effectiveness of advertising cam-

paigns and user satisfaction with ad content.

4.2 Accruing Benefits

The strategic incorporation of models designed to counteract popularity bias presents several key benefits:

- **Enhanced Personalization:** Systems capable of discerning beyond mainstream popularity can offer recommendations that genuinely reflect individual user preferences, thereby elevating the user experience.
- **Promotion of Content Diversity:** The deliberate reduction of popularity bias supports an ecosystem where a wider array of content, creators, and products can thrive, enriching the available choices for users.
- **Empowerment of Emerging Talent:** Equitable exposure for emerging talents fosters innovation and diversity within various industries, promoting healthy competition and cultural enrichment.
- **Increased User Engagement:** Platforms that effectively mirror and expand upon user interests are likely to see heightened engagement levels, as users value discoveries that resonate with their unique tastes and preferences.

4.3 Potential Concerns

While the initiatives to mitigate popularity bias are commendable, they are not without potential pitfalls:

- **Algorithmic Complexity:** Addressing popularity bias introduces additional complexity to recommendation algorithms, potentially complicating their maintenance, understanding, and scalability.
- **Risk of Filter Bubbles:** There exists a danger that overly personalized recommendations might limit user exposure to diverse viewpoints, inadvertently reinforcing pre-existing biases.

4.4 Inherent Limitations

The endeavor to integrate advanced models for mitigating popularity bias into existing recommender systems confronts several limitations:

- **Technical and Financial Challenges:** The technical integration of sophisticated models poses significant challenges and requires substantial financial investment, with the benefits not always immediately quantifiable.
- **Cost-Benefit Balancing:** The economic implications of adopting advanced recommender systems necessitate careful consideration, particularly for platforms operating with limited financial leeway.

5 Conclusion

Integrating the findings from both datasets, it becomes clear that the effectiveness of bias-mitigation models like PDA and DICE is contingent upon the behavioral patterns of the users they serve. But both popularity-bias-involved models perform better than the popularity-eliminated model (PD). It indicates that completely removing popularity bias is not helpful for recommendation. It is reasonable because customers or users inevitably have a conformity mindset and want to access popular items. However, model trainers should control the contribution, or the involvement, of the popularity bias.

The DICE model, with its emphasis on separating popularity and individual interests, is better suited for environments where collective preferences significantly influence user choices, as seen on Netflix. In contrast, the PDA model is more effective in contexts where individualized and prolonged user engagement is emphasized, as demonstrated by its performance on Douban. This understanding of user behavior and platform-specific dynamics is crucial for developing and implementing recommender systems that can accurately reflect both the collective trending and the rich individuality of user preferences.

References

Abdollahpouri, Himan, Robin Burke, and Bamshad Mobasher. 2019. “Managing Popularity Bias in Recommender Systems with Personalized Re-ranking.” In *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference*.

Gao, Chen, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. 2022. “Causal inference in recommender systems: A survey and future directions.” *arXiv preprint arXiv:2208.12397*

Schnabel, Tobias, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. “Recommendations as Treatments: Debiasing Learning and Evaluation.” In *Proceedings of the 33rd International Conference on Machine Learning*.

Spirtes, Peter, and Kun Zhang. 2016. “Causal discovery and inference: concepts and recent methodological advances.” In *Applied informatics*. SpringerOpen

Zhang, Yang, 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*.

Zheng, Yu, Chen Gao, Xiang Li, Xiangnan He, Depeng Jin, and Yong Li. 2021. “Disentangling User Interest and Conformity for Recommendation with Causal Embedding.” *International World Wide Web Conference Committee*. [Link]