UNMASKING POPULARITY BIAS: A Study on Fairness Evaluation and Mitigation in Recommendation Systems

SANMITHA SHETTY, University of Illinois at Chicago, USA
SHREYASH KADAM, University of Illinois at Chicago, USA

Recommendation systems are integral to navigating vast online content but often suffer from popularity bias: the tendency to over-recommend already popular items, potentially harming user experience and content diversity. This study investigates popularity bias propagation and its fairness implications [1], building upon prior work that revealed divergent findings across domains [5, 6] and highlighted the crucial role of evaluation strategies [3]. We conduct a focused reproducibility study in the Music and Movie domains, analyzing the impact of two key evaluation strategies (`UserTest`, `TrainItems`) and comparing three user grouping methods, including a novel `NicheConsumptionRate` approach focused directly on low-popularity item consumption. We evaluate bias using %ΔGAP [1] and accuracy using NDCG@10. Subsequently, we implement and assess a post-processing mitigation technique (multiplicative damping, $\alpha = 0.5$) to evaluate its effectiveness in reducing bias and its impact on accuracy. Our findings confirm that the evaluation strategy profoundly influences measured bias, with `TrainItems` revealing starker effects [3]. The `NicheConsumptionRate` grouping effectively identifies niche users who face similar biases to those highlighted by existing methods. The mitigation strategy successfully reduces both the magnitude and disparity of bias across groups but incurs a noticeable trade-off, generally decreasing NDCG@10 accuracy.

Additional Key Words and Phrases: Recommendation Systems, Popularity Bias, Fairness, Evaluation Strategies, Mitigation, Reproducibility

ACM Reference Format:
Sanmitha Shetty and Shreyash Kadam. 2025. UNMASKING POPULARITY BIAS: A Study on Fairness Evaluation and Mitigation in Recommendation Systems. 1, 1 (May 2025), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Background: The Role and Challenge of Recommendation Systems

In an era characterized by an overwhelming abundance of online information and choices, recommendation systems have become indispensable tools, helping users discover relevant content in domains ranging from e-commerce and entertainment to news and social media. By filtering and ranking items, these systems aim to personalize the user experience and increase engagement. Common techniques, particularly collaborative filtering, leverage user-item interaction patterns (ratings, clicks, views) to generate recommendations.

Authors' addresses: Sanmitha Shetty, University of Illinois at Chicago, Chicago, IL, USA; Shreyash Kadam, University of Illinois at Chicago, Chicago, IL, USA.

ACM XXXX-XXXX/2025/5-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.2 Problem Definition: The Pervasiveness of Popularity Bias

One of the most pervasive issues is **popularity bias**. This refers to the algorithmic tendency wherein items that are already popular (i.e., have significantly more interaction data) are disproportionately recommended over less popular, or "niche," items [2]. This leads to increased recommendations for these items, which in turn generates even more interaction data, further solidifying their dominance in the system's recommendations, a cycle often termed the "rich-get-richer" effect or a popularity feedback loop [1].

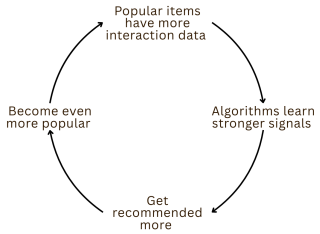


Fig. 1. The "rich-get-richer" effect in popularity bias

1.3 Motivation: Why Popularity Bias Demands Attention

Unchecked popularity bias has significant implications beyond statistical skew, affecting both users and content providers. It harms user experience by creating "filter bubbles" that limit discovery and fairness for niche users, who receive less relevant recommendations. Additionally, it reduces content diversity by disadvantaging creators of niche content, limiting their exposure and hindering market entry. Popularity bias also conflicts with the core goals of recommendation systems, which should aim to deliver accurate, diverse, and fair recommendations for all users.

1.4 Research Questions and Objectives

This project aims to contribute to the understanding and mitigation of popularity bias :

(1) **Impact of Evaluation Strategy:** How significantly do the `UserTest` and `TrainItems` evaluation strategies differ in their assessment of popularity bias (%ΔGAP) and recommendation accuracy (NDCG@10) for a common set of algorithms across the Music and Movie domains? Does this reproduction confirm the critical role of candidate item generation highlighted by Daniil et al. [3]?
(2) **Effectiveness of User Grouping Methods:** How do the fairness patterns (%ΔGAP and NDCG@10 disparities) compare when users are grouped based on `PopularPercentage`, `Average Popularity`, versus the novel `NicheConsumptionRate`?
(3) **Mitigation Efficacy (Bias):** How effective is a specific post-processing mitigation strategy (multiplicative damping) at reducing measured popularity bias (%ΔGAP) and narrowing the %ΔGAP disparities ?
(4) **Mitigation Impact (Accuracy):** What is the corresponding effect of this mitigation strategy on NDCG@10? Does it improve accuracy for niche groups, decrease it overall, or maintain existing levels?

1.5 Contributions and Novelty

While building on established research, this project offers specific novel contributions:

- **Direct Niche Consumption Grouping:** We introduce and systematically evaluate the NicheConsumptionRate method. This approach directly quantifies user affinity for the *least* popular items (bottom 30%), providing a potentially more direct and interpretable way to segment users based on niche taste compared to methods focusing on popular item interactions [1].
- **Targeted Reproducibility Study and Extension:** We perform a targeted reproduction of the Daniil et al. [3] framework, specifically examining the interplay of evaluation strategy and user grouping (including our novel method) in the Music and Movie domains. This provides focused validation and extension of their findings.
- **Integrated Mitigation Assessment:** We bridge the gap between bias analysis and practical intervention by implementing a specific mitigation strategy (post-processing damping with $\alpha = 0.5$) and evaluating its impact using the *same* rigorous framework (evaluation strategies, user groupings, metrics) employed in the baseline analysis. This allows for a direct assessment of the technique's effectiveness and trade-offs within the investigated contexts.

2 LITERATURE REVIEW

The study of popularity bias in recommendation systems has evolved from recognizing its existence to addressing fairness concerns and exploring mitigation strategies. Early research, like Celma and Cano [2], highlighted how popular items, such as artists, bias music recommendations and hinder niche discoveries.

2.1 User-Centric Fairness Perspective

Abdollahpouri et al. [1] shifted the focus from system-level item coverage to user-level fairness. They argued that popularity bias unfairly affects niche users, who prefer items not dominated by popularity. Their proposed %ΔGAP metric quantifies this bias by comparing the popularity of recommended items with users' historical preferences, demonstrating that niche users are often recommended more popular items than their profiles suggest.

2.2 Cross-Domain Reproductions and Divergence

Studies replicating Abdollahpouri et al.'s methodology in other domains revealed differing results:

- **Kowald et al. [5]** found bias in music recommendations using the LastFM dataset, though their results varied from the movie domain, suggesting dataset differences.
- **Naghiaei et al. [6]** reproduced the study in the book domain using the Book-Crossing dataset, highlighting the context-dependence of popularity bias and the potential for methodological inconsistencies.

These findings question the generalizability of popularity bias and the sensitivity of bias measurements to the experimental setup.

2.3 Methodological Investigation

Daniil et al. [3] explored the discrepancies in prior studies [5, 6], focusing on data characteristics, algorithms, user grouping, and evaluation strategy. They found that the evaluation strategy, especially how candidate items for ranking are generated—strongly influences bias and fairness metrics. Changing the item generation method led to different conclusions about popularity bias and fairness.

2.4 Positioning Our Work

This project directly engages with the findings and methodology presented by Daniil et al. [3]:

- We adopt their emphasis on the critical role of **evaluation strategy** (`UserTest` vs. `TrainItems`) and the importance of analyzing **user grouping** methods.
- We conduct a **targeted reproduction** of their framework in two domains (Music and Movies) using two key evaluation strategies (`UserTest`, `TrainItems`), as identified by Daniil et al. [3], with a representative subset of algorithms.
- We **innovate** by introducing the `NicheConsumptionRate` grouping as a direct measure of niche preference, offering a different lens compared to popular-item-focused methods used in prior work [1], thus enabling a richer comparison of how niche users are defined and impacted.
- We **extend** the analysis by evaluating a practical **mitigation strategy** (multiplicative damping) within this validated framework. This enables assessment not only of bias but also of potential remediation, including the associated costs (e.g., accuracy trade-offs).

3 METHODOLOGY

This research employed a two-phase methodology: Study 1 focused on reproducing and analyzing baseline popularity bias, while Study 2 evaluated a mitigation strategy.

3.1 Datasets and Preprocessing

Two publicly available datasets representing different media domains were used:

- **Music (LastFM):**
  The first dataset represented the **Music** domain, utilizing a pre-processed subset derived from the comprehensive LastFM-1b collection [8]. This subset was curated to ensure a reasonable level of data density by filtering to include only users and artists with at least 20 interactions. The resulting dataset comprised 3,000 users, 12,690 unique artists (items), and 1,008,479 user-artist interactions (listening events). Given the nature of the original data (listening counts), this dataset primarily represents implicit user feedback. Its sparsity level was calculated at 97.35%. A crucial preprocessing step, consistent with the methodology applied by Kowald et al. [5] and Daniil et al. [3], involved scaling the raw listening counts per user.
- **Movies (MovieLens1M):**
  The second dataset represented the **Movie** domain, using the widely adopted MovieLens1M dataset [4]. This dataset is a standard benchmark in Recommendation systems research and contains explicit user feedback. It includes 1,000,209 ratings provided by 6,040 users for approximately 3,900 distinct movies. The ratings are expressed on an integer scale from 1 (lowest preference) to 5 (highest preference). The sparsity of this dataset is approximately 95.75%. In contrast to the Music dataset, no additional scaling or transformation was applied to the MovieLens1M ratings; the original 1–5 explicit rating values were used directly in all experiments.

Item Popularity Calculation: For both domains, item popularity was defined as the interaction frequency (i.e., the number of unique users rating or interacting with an item), calculated from the *entire processed dataset* prior to the train-test split.

$$\texttt{normalized\_popularity} = \frac{\texttt{interaction\_count}}{\texttt{total\_number\_of\_users}}$$

Normalized popularity was derived and used for Δ%GAP calculations and the mitigation step. Sets of the top 20% most popular items and bottom 30% least popular (niche) items were identified based on these interaction counts.

3.2 Algorithms

We evaluated a selection of six algorithms, implemented using the Cornac library [7], chosen to represent different recommendation paradigms relevant to the reproduced studies:

- **MostPop:** A non-personalized baseline that recommends the most popular items to all users.
- **UserKNN:** User-based collaborative filtering (k = 40, cosine similarity).
- **ItemKNN:** Item-based collaborative filtering (k = 40, cosine similarity).
- **PMF:** Probabilistic Matrix Factorization with 10 latent factors.
- **NMF:** Non-negative Matrix Factorization using the Cornac implementation, with 15 latent factors.
- **HPF:** Hierarchical Poisson Factorization, applied to explicit or scaled implicit data, with 50 latent factors.

Default hyperparameters from the Cornac library were used for all models, consistent with the setup in the reproduced studies [3, 5, 6].

3.3 Evaluation Framework

Our evaluation framework is designed for consistency and reproducibility across domains and studies. For each dataset, a fixed 80% training / 20% test split was created using Cornac's `RatioSplit` method (`my_seed=0`, `exclude_unknowns=True`).
A key component of the framework is the comparison of two **evaluation strategies** for candidate item generation, following Daniil et al. [3]:

- **UserTest (`eva_two`):** Ranks only the items that the user interacted with in the test set (evaluates known interactions).
- **TrainItems (`eva_three`):** Ranks all items not interacted with by the user in the training set (simulates recommending novel items, which is crucial for assessing bias generalization).

For **recommendation generation**, scores were predicted for the candidate items defined by each strategy. In Study 1, items were ranked directly by their predicted scores. In Study 2, scores were adjusted using multiplicative damping:

$$\texttt{new\_score} = \frac{\texttt{original\_score}}{\texttt{item\_popularity}^{0.5}}$$

before ranking. The top-10 items were selected as the final recommendation list in both studies.

3.4 User Grouping Methods

Fairness was analyzed by segmenting test set users into Low/Niche (bottom 20%), Medium/Diverse (middle 60%), and High/Blockbuster (top 20%) groups based on their training profile characteristics:

- **`PopularPercentage`:** Users were sorted by the percentage of Top-20% popular items in their training profile.
- **`AveragePopularity`:** Users were sorted by the average normalized popularity of items in their training profile.
- **`NicheConsumptionRate(Novel)`:** Users were sorted by the percentage of Bottom-30% niche items in their training profile. The High group represents the most niche-focused users.

3.5 Evaluation Metrics

- **%ΔGAP (Percent Delta Group Average Popularity):** Measures the popularity shift from a user's training profile to their recommendations, computed per user group [1].

  Formula: $\%\Delta\text{GAP} = \frac{\text{AvgPop}_{\text{Rec}} - \text{AvgPop}_{\text{Profile}}}{\text{AvgPop}_{\text{Profile}}} \times 100.$

- **NDCG@10 (Normalized Discounted Cumulative Gain at 10):** Evaluates the ranking quality of the top-10 recommended items by rewarding relevant items appearing earlier in the list. Values range from 0 (worst) to 1 (best).
- **T-tests:** Independent two-sample Welch's t-tests were conducted to compare metric distributions between user groups; a p-value below 0.05 was taken as statistically significant.

3.6 Study 1: Baseline Reproducibility Study

The primary objective of Study 1 was to establish a baseline understanding of how popularity bias manifests and affects different user groups within our chosen experimental setup, closely following and reproducing key aspects of the methodology investigated by Daniil et al. [3]. This involved several sequential steps.

First, **data preparation** required loading and preprocessing the selected datasets. For both domains, overall item popularity (based on interaction frequency) was calculated, and items were categorized into Top-20% popular and Bottom-30% niche sets. Second, standard **model training** procedures were followed. Third, **baseline recommendations** were generated for users present in the test set. For each trained algorithm, recommendations were generated using two distinct evaluation strategies as defined by Daniil et al. [3] The Top-10 ranked items for each user under each condition were stored. Fourth, a comprehensive **baseline evaluation** was conducted. For every combination of domain, evaluation strategy, and algorithm, users were segmented into Low/Niche, Medium/Diverse, and High/Blockbuster groups based on each of the three user grouping methods Finally, to assess the statistical significance of observed differences between fairness groups, independent two-sample **t-tests** were performed on the distributions of individual user %ΔGAP and NDCG@10 scores for each relevant group comparison.

3.7 Study 2: Bias Mitigation Study

Building upon the baseline results established in Study 1, the second phase of our research focused on implementing and evaluating a specific intervention aimed at mitigating popularity bias. The goal was to assess the effectiveness of this strategy in reducing bias metrics) and to understand its consequent impact on recommendation accuracy.

First, this study directly **leveraged the outputs of Study 1**, specifically the trained recommendation models and the calculated base item popularity and user metrics. Second, the chosen **mitigation technique** was a post-processing approach known as *Multiplicative Damping*. This technique operates by adjusting the recommendation scores predicted by the baseline algorithms before final ranking. The core idea is to penalize items based on their popularity, making highly popular items less likely to appear at the top of the list compared to less popular but potentially relevant ones. Third, the specific implementation involved applying the following formula to each candidate item's original score generated by a baseline algorithm:

$$\text{new\_score} = \frac{\text{original\_score}}{\text{item\_popularity}^{\alpha}}$$

For all experiments in this study, we fixed $\alpha = 0.5$, which corresponds to a moderate damping effect (i.e., a square root penalty on popularity). Fourth, the process of **generating mitigated recommendations** was executed. For each baseline algorithm and evaluation strategy (`UserTest` and `TrainItems`), the candidate item scores were recalculated using the damping formula. The candidate items were then re-ranked based on these *mitigated scores*, and the Top-10 items were selected to form the final mitigated recommendation lists.

Finally, a **comparative analysis** was conducted by directly comparing the %ΔGAP, NDCG@10, and t-test results from Study 2 (Mitigated) with those from Study 1 (Baseline) under corresponding

conditions. This comparison allowed us to quantify the effectiveness of the mitigation strategy in reducing bias magnitude and disparity, and to characterize the associated accuracy trade-offs.

4 RESULTS AND ANALYSIS

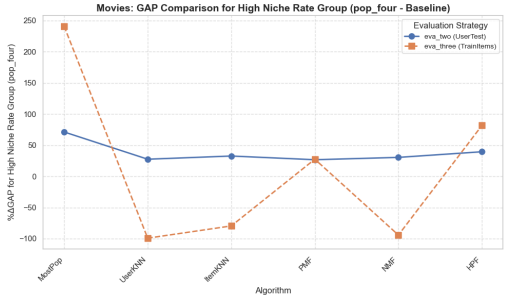
This section presents the empirical findings from both studies, referencing the summary results in Table 1 and the visualizations in Figures 2 through 5.

4.1 Study 1: Baseline Popularity Bias Analysis

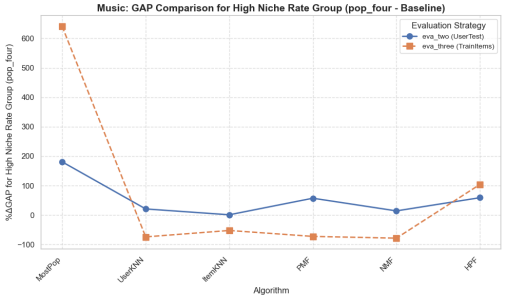
4.1.1 Overall Bias & Impact of Evaluation Strategy. Consistent with the findings of Daniil et al. [3], the choice of evaluation strategy profoundly influenced the observed bias. As shown in Figure 2, the `TrainItems` (`eva_three`, dashed orange lines) strategy consistently revealed larger magnitudes and variations in baseline %ΔGAP compared to the `UserTest` (`eva_two`, solid blue lines) strategy across both domains for the high-niche user group.

4.1.2 Unfairness Towards User Groups (%ΔGAP Analysis). Under the `TrainItems` strategy, significant fairness disparities emerged impacting the high-niche user group (`pop_four`-High):

- **Grouping Method Comparison:** While Table 1 focuses on `pop_four`, broader analysis indicated that regardless of the method, users identified as niche consistently experienced the most extreme %ΔGAP shifts under `TrainItems` compared to diverse or blockbuster groups, with t-tests frequently confirming significance. The `pop_four` method effectively captured this vulnerable group.
- **Domain Differences:** The magnitude of bias varied significantly. In Music, the baseline `eva_three` %ΔGAP for the high-niche group ranged dramatically from +641.1% (MostPop) to −78.3% (NMF). In Movies, the range was also wide but generally less extreme, from +240.7% (MostPop) to −99.2% (UserKNN).
- **Algorithm Differences:** Under `TrainItems`, algorithms displayed distinct baseline bias profiles for niche users: MostPop, HPF, and sometimes PMF (in Movies, +27.5%) showed strong positive %ΔGAP, pushing popular items onto niche users.



(a) Baseline %ΔGAP Comparison for High Niche Rate Group in Movie Domain



(b) Baseline %ΔGAP Comparison for High Niche Rate Group in Music Domain

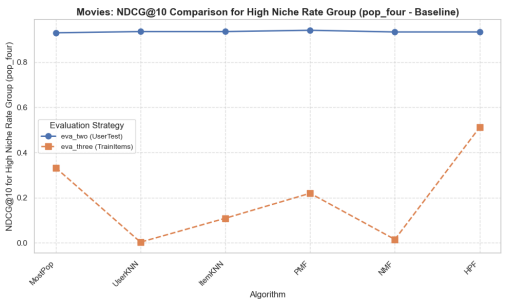
Fig. 2. Baseline %ΔGAP for high-niche groups across both domains

4.1.3 Accuracy Analysis (NDCG@10 Analysis). Baseline accuracy for the high-niche user group also showed critical variations:

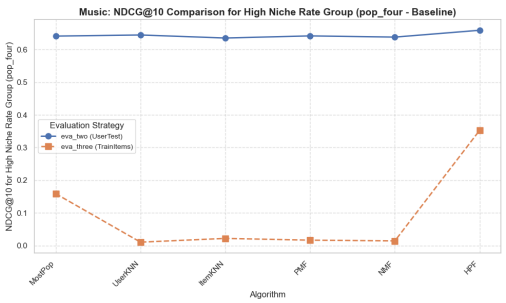
- **Evaluation Strategy Impact:** As seen in Figure 3, accuracy (NDCG@10) was dramatically higher under `UserTest` (solid blue) compared to `TrainItems` , dashed orange). In Movies

(Fig. 3a), `eva_two` NDCG consistently exceeded 0.9, while `eva_three` (Table 1). A similar pattern held for Music (Fig. 3b), with `eva_two` NDCG generally above 0.6, but `eva_three` NDCG falling near zero for KNNs, PMF (0.017), and NMF (0.014).

- **Accuracy Disparities (Fairness):** The extremely low NDCG@10 scores under `TrainItems` for the high-niche group signify a major fairness issue. These users receive recommendations of very poor ranking quality when the system suggests novel items, especially from algorithms like UserKNN, ItemKNN, PMF (in Music), and NMF.



(a) Baseline NDCG@10 Comparison for High Niche Rate Group in Movie Domain



(b) Baseline NDCG@10 Comparison for High Niche Rate Group in Music Domain

Fig. 3. Baseline NDCG@10 for high-niche groups across both domains

4.2 Study 2: Mitigation Evaluation

4.2.1 Impact on Bias (%ΔGAP Reduction). The multiplicative damping mitigation ($\alpha = 0.5$) proved highly effective in reducing popularity bias magnitude and disparity:

- **Magnitude Reduction:** Mitigation substantially reduced the absolute %ΔGAP values. Comparing baseline vs. mitigated results in Table 1 for `eva_three`, the extreme positive baseline value for Music/MostPop (+641.1%) was reduced to −36.5%. The large negative baseline value for Movies/UserKNN (−99.2%) remained highly negative (−99.9%), suggesting the mitigation might struggle when baseline scores for popular items are already extremely low for certain users or items, or that $\alpha = 0.5$ is insufficient in such cases. An exception was Music/PMF, where the mitigated GAP became highly positive (+512.2%). Figure 4 visually confirms the general trend of reduction across all conditions.
- **Disparity Reduction:** By reducing the extreme values (with the noted exception of Music/PMF), the mitigation generally narrowed the gap in %ΔGAP experienced by niche users compared to other groups (inferred from broader analysis and the flattening effect in Fig. 4). T-tests comparing groups often became non-significant after mitigation, suggesting more equitable popularity treatment overall.

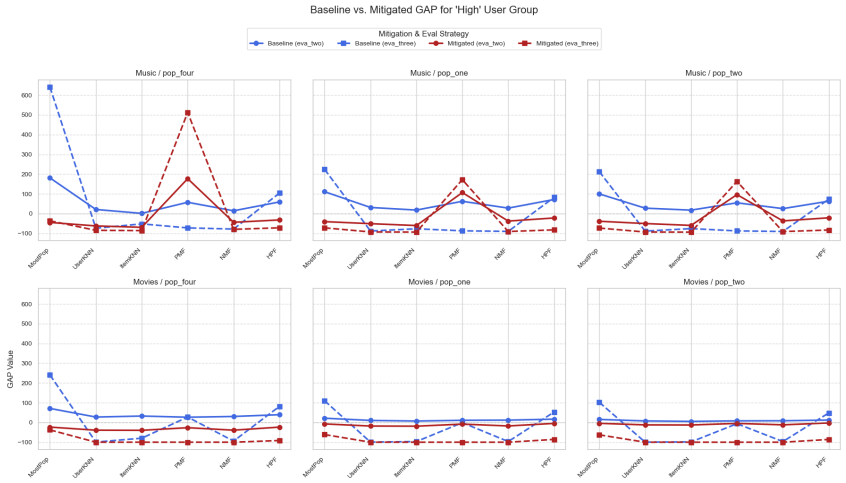


Fig. 4. Baseline vs Mitigated %ΔGAP Comparison for High Niche Rate Group in Movie Domain

4.2.2 Impact on Accuracy (NDCG@10 Trade-off). The success in bias reduction came with a distinct trade-off regarding ranking accuracy:

- **General Decrease:** As shown in Table 1 and Figure 5, mitigated NDCG@10 values were generally lower than or, at best, similar to baseline values. This effect was most pronounced under the `TrainItems` strategy. For example, Movies/HPF baseline NDCG@10 under `eva_three` was 0.512, which dropped to 0.039 after mitigation. Music/HPF dropped from 0.353 to 0.033.
- **Niche User Accuracy:** Mitigation did *not* improve—and often worsened—the already low accuracy for the high-niche group under `TrainItems`. For instance, Movies/UserKNN baseline NDCG was 0.003 and became 0.000 post-mitigation.
- **Persistent Gaps:** Accuracy disparities between user groups often persisted—or were even exacerbated—in relative terms when overall accuracy dropped. The mitigation strategy, focused on popularity.

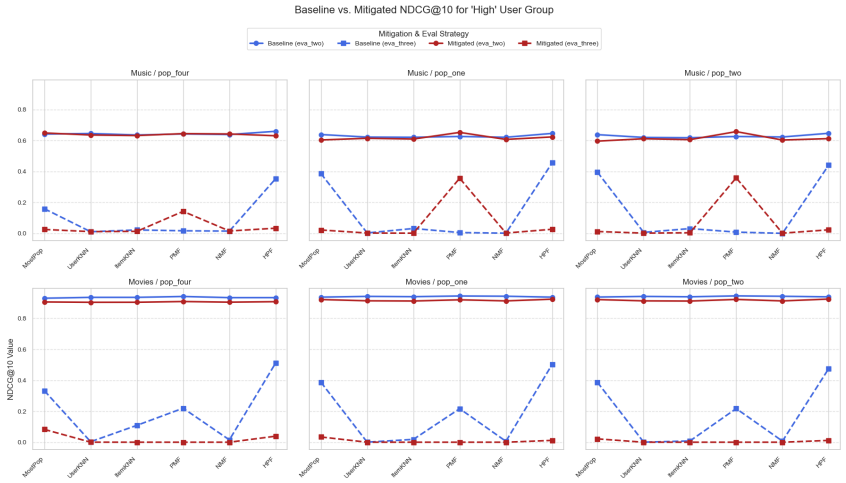


Fig. 5. Baseline vs Mitigated NDCG@10 Comparison for High Niche Rate Group in Movie Domain

Domain	Algorithm	%ΔGAP (pop_four-High)		NDCG@10 (pop_four-High)	
		Baseline	Mitigated ($\alpha = 0.5$)	Baseline	Mitigated ($\alpha = 0.5$)
music	MostPop	641.1	-36.5	0.159	0.025
music	UserKNN	-74.0	-84.7	0.010	0.011
music	ItemKNN	-52.1	-86.0	0.022	0.013
music	PMF	-72.3	512.2	0.017	0.142
music	NMF	-78.3	-79.8	0.014	0.015
music	HPF	104.2	-71.8	0.353	0.033
movies	MostPop	240.7	-37.5	0.333	0.083
movies	UserKNN	-99.2	-99.9	0.003	0.000
movies	ItemKNN	-79.6	-99.9	0.109	0.000
movies	PMF	27.5	-99.8	0.220	0.000
movies	NMF	-94.2	-99.9	0.015	0.000
movies	HPF	81.7	-91.3	0.512	0.039

Table 1. Summary Comparison of Baseline vs. Mitigated Results for High Niche Group under `TrainItems`.

5 DISCUSSION

Our findings provide several important insights into measuring and mitigating popularity bias in Recommendation systems.

- **Reconfirming the Primacy of Evaluation Strategy:** This study strongly reinforces the central message from Daniil et al. [3]: the measurement of popularity bias is highly sensitive to methodological choices, particularly the evaluation strategy. The stark contrast between the `UserTest` (`eva_two`) and `TrainItems` (`eva_three`) outcomes across both domains and all algorithms underscores this point. `UserTest`, by focusing on known user-item interactions within the test set, provides a limited view that often masks the true extent to which algorithmic biases generalize to novel recommendation scenarios.

- **Validating `NicheConsumptionRate`:** The novel `NicheConsumptionRate` (`pop_four`) grouping method proved to be a valid and useful tool. It effectively identified users with strong preferences for low-popularity items and demonstrated that these users face fairness challenges (high bias shifts, low accuracy from certain algorithms) comparable to users identified via low popular item consumption (`pop_one`, `pop_two`). This suggests that defining niche users based on positive affinity for unpopular content can be as informative.
- **Mitigation Effectiveness & Trade-offs:** The implementation of multiplicative damping ($\alpha = 0.5$) demonstrated clear success in achieving its direct goal: reducing the magnitude of popularity bias (%ΔGAP) and making its impact more equitable across user groups. However, the study clearly highlights the common bias-accuracy trade-off. Improving fairness in terms of recommendation popularity came at the cost of overall ranking accuracy (NDCG@10), and crucially, it did not solve the accuracy disparity for niche users. TThe choice of $\alpha = 0.5$ is just one point on this trade-off curve; different values would yield different balances.
- **Context Matters:** The variations observed between the Music and Movie domains, even with efforts to align methodologies (like scaling music ratings), underscore that popularity bias effects and the impact of mitigation are context-dependent. Findings from one domain may not directly translate to another without careful consideration and adaptation.

6 TECHNICAL CHALLENGES

Conducting this research involved navigating several technical and methodological challenges:

- **Computational Resources:** The combinatorial nature of the experiments (Domains × Algorithms × Evaluation Strategies × Grouping Methods × Baseline/Mitigation) required significant computation time for training, recommendation generation, and evaluation, particularly for the larger LastFM dataset and more complex algorithms.
- **Data Heterogeneity & Preprocessing:** The fundamental difference between the implicit LastFM data (listening counts) and explicit MovieLens data (1–5 ratings) posed a challenge for direct comparison. The interpretation of the 'rating' signal differs across domains, limiting the direct comparison of absolute metric magnitudes (e.g., NDCG, absolute %ΔGAP shift). This highlights a broader challenge in standardizing bias evaluation across datasets with different feedback types.
- **Mitigation Parameter Tuning:** Implementing the mitigation involved selecting a specific technique (damping) and strength ($\alpha = 0.5$). Finding the best balance between bias reduction and accuracy preservation would require extensive hyperparameter tuning, potentially involving cross-validation or defining a multi-objective optimization function.
- **Metric Interpretation Complexity:** Evaluating fairness using multiple metrics like %ΔGAP and NDCG@10 requires careful interpretation. Our results showed that mitigation could successfully reduce bias disparities (%ΔGAP) while simultaneously decreasing accuracy (NDCG@10). Deciding whether this trade-off is acceptable depends heavily on the specific application goals and ethical considerations.

7 CONCLUSION

This study highlights that addressing popularity bias in recommender systems demands a nuanced, multi-faceted approach. We found that the evaluation strategy particularly the use of `TrainItems` significantly influences how fairness and bias are perceived. Niche users, identified effectively through our proposed `NicheConsumptionRate` method, consistently experience both greater popularity shifts and lower ranking accuracy. While the post-processing mitigation using multiplicative

damping ($\alpha = 0.5$) effectively reduced bias, it did so at a cost to NDCG@10, underscoring a persistent trade-off between fairness and accuracy. Ultimately, our results affirm that no single solution suffices; effective mitigation requires a thoughtful balance of evaluation design, user segmentation, algorithmic intervention, and fairness goals.

8 FUTURE WORK

This study opens several promising avenues for future research to further understand and address popularity bias.

Firstly, **mitigation strategies** warrant deeper exploration. This includes systematically tuning the damping parameter α to fully characterize the bias-accuracy trade-off, comparing multiplicative damping against pre-processing and in-processing techniques, and investigating adaptive mitigation methods tailored to specific user or item characteristics.

Secondly, the **scope of the analysis** could be expanded by incorporating additional domains such as book recommendations [6], evaluating newer deep learning-based algorithms, and including alternative evaluation strategies like 'Modified TrainItems' [3] for a more comprehensive methodological comparison.

Finally, **deeper analysis** is needed to quantitatively link dataset characteristics (e.g., sparsity and popularity distribution) to bias levels, to formally evaluate the effects of data preprocessing choices such as scaling, and to develop more nuanced user grouping approaches that integrate multiple dimensions of user behavior.

—

REFERENCES

[1] 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 (RMSE'19), Co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019) (CEUR Workshop Proceedings, Vol. 2440)*, Alexander Tuzhilin, Bracha Shapira, Robin Burke, and Gonen Dror (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2440/paper4.pdf

[2] Òscar Celma and Pedro Cano. 2008. From hits to niches? Or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition, NETFLIX '08, Las Vegas, Nevada, USA, August 24, 2008*. 1–8. https://doi.org/10.1145/1722149.1722154

[3] Savvina Daniil, Mirjam Cuper, Cynthia C. S. Liem, Jacco van Ossenbruggen, and Laura Hollink. 2024. Reproducing Popularity Bias in Recommendation: The Effect of Evaluation Strategies. *ACM Trans. Recomm. Syst.* 2, 1, Article 5 (2024), 39 pages. https://doi.org/10.1145/3637066

[4] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (2015), 19 pages. https://doi.org/10.1145/2827872

[5] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12036)*. Springer, 35–42. https://doi.org/10.1007/978-3-030-45442-5_5

[6] Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Mahdi Dehghan. 2022. The Unfairness of Popularity Bias in Book Recommendation. arXiv:2202.13446 [cs.IR] https://arxiv.org/abs/2202.13446

[7] Aghiles Salah, Quoc-Tuan Truong, and Hady W. Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *J. Mach. Learn. Res.* 21, 95 (2020), 1–5. http://jmlr.org/papers/v21/19-805.html

[8] Markus Schedl. 2016. The LFM-1b Dataset for Music Retrieval and Recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (New York, New York, USA) *(ICMR '16)*. Association for Computing Machinery, New York, NY, USA, 103–110. https://doi.org/10.1145/2911996.2912004