How Do Recommendation Models Amplify Popularity Bias? An Analysis from the Spectral Perspective

Siyi Lin[†‡]
Zhejiang University
Hangzhou, China
lin452_lsy@zju.edu.cn

Chongming Gao
University of Science and Technology of China
Hefei, China
chongming.gao@gmail.com

Jiawei Chen[*†‡§]
Zhejiang University
Hangzhou, China
sleepyhunt@zju.edu.cn

Sheng Zhou
Zhejiang University
Hangzhou, China
zhousheng_zju@zju.edu.cn

Binbin Hu
Ant Group
Hangzhou, China
bin.hbb@antfin.com

Yan Feng[†‡]
Zhejiang University
Hangzhou, China
fengyan@zju.edu.cn

Chun Chen[†‡]
Zhejiang University
Hangzhou, China
chenc@zju.edu.cn

Can Wang[†§]
Zhejiang University
Hangzhou, China
wcan@zju.edu.cn

Abstract

Recommendation Systems (RS) are often plagued by popularity bias. When training a recommendation model on a typically long-tailed dataset, the model tends to not only inherit this bias but often exacerbate it, resulting in over-representation of popular items in the recommendation lists. This study conducts comprehensive empirical and theoretical analyses to expose the root causes of this phenomenon, yielding two core insights: 1) Item popularity is memorized in the principal spectrum of the score matrix predicted by the recommendation model; 2) The *dimension reduction* phenomenon amplifies the relative prominence of the principal spectrum, thereby intensifying the popularity bias.

Building on these insights, we propose a novel debiasing strategy that leverages a *spectral norm regularizer* to penalize the magnitude of the principal singular value. We have developed an efficient algorithm to expedite the calculation of the spectral norm by exploiting the spectral property of the score matrix. Extensive experiments across seven real-world datasets and three testing paradigms have been conducted to validate the superiority of the proposed method.

CCS Concepts

• **Information systems → Recommender systems**.

Keywords

Recommender System; Popularity Bias

ACM Reference Format:
Siyi Lin, Chongming Gao, Jiawei Chen, Sheng Zhou, Binbin Hu, Yan Feng, Chun Chen, and Can Wang. 2025. How Do Recommendation Models Amplify Popularity Bias? An Analysis from the Spectral Perspective. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25), March 10–14, 2025, Hannover, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3701551.3703579

1 Introduction

Recommender Systems (RS), with their capability to offer personalized suggestions, have found applications across various domains [18, 40, 70]. Collaborative filtering (CF), a widely-used technique within RS, learns user preference from historical interactions. However, their effectiveness in personalization is significantly compromised by popularity bias [10]. This bias emerges when user interaction data showcases a long-tailed distribution of item interaction frequencies. Subsequently, recommendation models trained on such data tend to inherit and even amplify this bias, leading to an overwhelming presence of popular items in recommendation results [53, 66, 73].

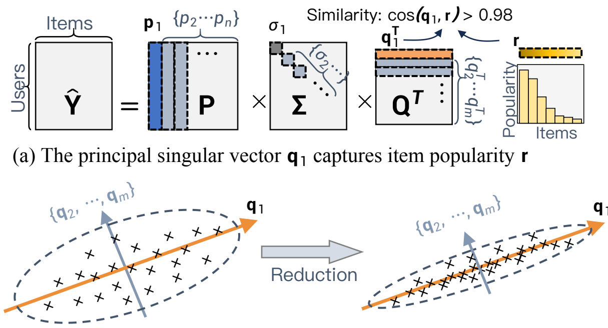
This notorious effect not only undermines the accuracy and fairness of recommendation [2, 3], but also exacerbates the Matthew Effect and the filter bubble through the user-system feedback loop [21, 22, 34].

Given the detrimental impact of popularity bias amplification, a thorough understanding of its root causes is crucial. Although some recent studies have endeavored to elucidate this, their investigations exhibit significant limitations: 1) Some researchers [52, 53, 66] have investigated popularity bias amplification through causal graphs. However, they merely postulate causal relations between item popularity and model predictions without deeply exploring the underlying mechanisms behind the relations. Moreover,

[*]Corresponding author.
[†]State Key Laboratory of Blockchain and Data Security, Zhejiang University.
[‡]College of Computer Science and Technology, Zhejiang University.
[§]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security.

WSDM '25, March 10–14, 2025, Hannover, Germany.

ACM ISBN 979-8-4007-1329-3/25/03
https://doi.org/10.1145/3701551.3703579



(a) The principal singular vector $\mathbf{q}_1$ captures item popularity $\mathbf{r}$

(b) Dimension reduction augments the relative prominence of the principal singular vector $\mathbf{q}_1$

Figure 1: Illustration of two core insights.

their analyses depend on hypothesized causal graphs, which may be flawed due to the widespread presence of unmeasured confounders [20, 65]. 2) Other studies [7, 12, 62, 67, 71] have revealed graph neural network (GNNs) can exacerbate popularity bias. However, these analyses primarily focus to GNNs rather than the mechanisms of generic recommendation models.

To bridge this research gap, we undertake extensive theoretical and empirical studies on popularity bias amplification. By investigating the spectrum of the ranking score matrix over all users and items predicted by recommendation models, we present the following insights:

1) Memorization Effect. When training a recommendation model on long-tailed data, *the information of item popularity is memorized in the principal spectrum (Figure 1(a)).* Empirically, we observe that the principal singular vector of the score matrix closely aligns with item popularity, with a cosine similarity consistently exceeding 0.98 across multiple representative recommendation models and datasets. Theoretically, we derive the lower bound of this cosine similarity, demonstrating that the similarity converges to one for highly long-tailed training datasets.

2) Amplification Effect. *The phenomenon known as dimension reduction augments the relatively prominence of the principal spectrum that captures item popularity, leading to bias amplification (Figure 1(b)).* We reveal that dimension reduction is pervasive in RS due to two primary reasons: i) The deliberate low-rank setting of user/item embeddings, employed either to conserve memory or to counteract overfitting, amplifies the impact of the principal spectrum; ii) The inherent training dynamics of gradient-based optimization prioritize the learning of the principal dimension, while the singular values of other dimensions are easily underestimated. Our further theoretical and empirical analyses establish the relationship between dimension reduction and popularity bias — larger principal singular values compared to other singular values lead to more popular items on the recommendations.

Our analysis not only explains the underlying mechanisms of bias amplification but also paves the way for the development of an innovative strategy to counteract this effect. Recognizing that the essence of this amplification lies in the undue contribution of the principal spectrum, we introduce a spectral norm regularizer [58] aimed at directly restraining the magnitude of the principal singular value. However, the direct computation of the spectral norm necessitates exhaustive processing of a large score matrix and numerous iterative procedures [48, 58], inducing significant computational costs. To address this challenge, we further develop an accelerated strategy by leveraging the intrinsic spectrum properties of the score matrix and matrix transformation techniques. Consequently, our method effectively mitigates popularity bias while imposing limited computational overhead.

In summary, our contributions are:

- Conducting comprehensive analyses to unravel the mechanisms behind popularity bias amplification in recommendation — item popularity is encoded within the principal singular vector, and its impact is exaggerated due to the dimension reduction phenomenon.
- Proposing an efficient method for mitigating the bias amplification through the regulation of the principal singular value.
- Performing extensive experiments across seven real-world datasets under three different testing scenarios, demonstrating the superiority of our method in reducing bias and enhancing recommendation quality.

2 Preliminaries

In this section, we present the background of the recommendation system and popularity bias amplification.

Task Formulation. This work mainly focus on the collaborative filtering (CF) [56], a widely-used recommendation scenario. Consider a RS with a user set $\mathcal{U}$ and an item set $\mathcal{I}$. Let n and m denote the total number of users and items. Historical interactions can be expressed by a matrix $\mathbf{Y} \in \{0,1\}^{n\times m}$, where the element y_{ui} indicates if user u has interacted with item i (*e.g.,* click). For convenience, we define the number of interactions of an item as $r_i = \sum_{u\in\mathcal{U}} y_{ui}$, and collect r_i over all items as a popularity vector $\mathbf{r}$. RS targets to suggest items to users based on their potential interests.

Recommendation Models. Embedding-based models are widely utilized in RS [56]. Such models convert user/item attributes (*e.g.,* IDs) into d-dimensional representations ($\mathbf{u}_u, \mathbf{v}_i$), and make predictions using the embedding similarity [56]. Given that the inner product is a conventional similarity metric due to its efficiency in retrieval and superior performance [32, 54, 60], this work also focuses on the inner product for analysis. Specifically, the model's predicted scores can be formulated as $\hat{y}_{ui} = \mu(\mathbf{u}_u^\top \mathbf{v}_i)$, where $\mu(.)$ denotes an activation function like Sigmoid. $\hat{y}_{ui}$ represents a user's preference for an item, which is then used for ranking to generate recommendations. For clarity of presentation, we also employ matrix notation. Let matrices $\hat{\mathbf{Y}}, \mathbf{U}, \mathbf{V}$ represent scores over all user-item combinations, embeddings over all users and items, respectively. Model predictions can be succinctly expressed as $\hat{\mathbf{Y}} = \mu(\mathbf{U}\mathbf{V}^\top)$.

Objective Functions. Common choices of loss functions for training a recommendation model include point-wise loss such as BCE and MSE [41], and pair-wise loss like BPR [42]. It is worth noting that BPR can be reconceptualized as a specialized pointwise loss.Concretely, BPR loss is expressed as:

$$\mathcal{L}_{BPR} = -\sum_{u\in\mathcal{U}} \sum_{i\in\mathcal{I}, y_{ui}=1} \sum_{j\in\mathcal{I}, y_{uj}=0} \log(\mu(\mathbf{u}_u^\top \mathbf{v}_i - \mathbf{u}_u^\top \mathbf{v}_j))$$

If construct a hyper-item space denoted as $\mathcal{I}' = \mathcal{I} \times \mathcal{I}$ derived from item pairs, and define the embeddings of hyper-items as $\mathbf{v}'_{ij} = \mathbf{v}_i - \mathbf{v}_j$

and assign new observed interactions to the combinations of users and hyper-items, *i.e.,* $y'_{u,ij} = 1$ for $y_{ui} = 1 \& y_{uj} = 0$ and $y'_{u,ij} = 0$ for $y_{ui} = 0 \& y_{uj} = 1$. BPR can be re-written as:

$$\mathcal{L}_{BPR} = -\frac{1}{2}\sum_{u\in\mathcal{U}}\Big(\sum_{\substack{(i,j)\in\mathcal{I}'\\ y'_{u,ij}=1}}\log(\mu(\mathbf{u}_u^\top\mathbf{v}'_{ij})) + \sum_{\substack{(i,j)\in\mathcal{I}'\\ y'_{u,ij}=0}}\log(\mu(-\mathbf{u}_u^\top\mathbf{v}'_{ij}))\Big)$$

where BPR can be reframed as a specific point-wise loss under the hyper-items space $\mathcal{I}'$. Therefore, for convenience, our analyses mainly focus on point-wise loss. But we will also discuss why our proposed debiased method is suitable for BPR (*cf.* Section 4.2) and validate its effectiveness in experiments (*cf.* Section 5).

Popularity Bias Amplification. Items' interaction frequency in recommendation data often follows a long-tailed distribution [6, 17, 47]. For instance, in a typical Douban dataset, a mere 20% of the most popular items account for 86.3% of all interactions. When models are trained on such skewed data, they tend to absorb and amplify this bias, frequently over-prioritizing popular items in their recommendations. For example, in the Douban dataset using the MF model, 20% of the most popular items occupy over 99.7% of the recommendation slots, while a mere 0.6% of the most popular items occupy more than 63% (*cf.* Appendix[1] B.1 more examples). This notorious effect significantly impacts the recommendation accuracy and fairness, even potentially posing detrimental effects on the entire ecosystem of RS[10]. Thus, understanding the underlying mechanisms behind this effect is crucial.

3 Understanding Popularity Bias Amplification

In this section, we conduct thorough analyses to answer:
1) How do recommendation models memorize the item popularity?
2) Why do recommendation models amplify popularity bias?

3.1 Popularity Bias Memorization Effect

3.1.1 Empirical Study. To discern how recommendation models memorize item popularity, we designed the following experiment: 1) We well trained three representative recommendation models, MF [35], LightGCN [24] and XSimGCL [59], on three real-world datasets (*cf.* Section 5 for experimental details); 2) We then performed SVD decomposition on the predicted score matrix, $\hat{\mathbf{Y}} = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top = \sum_{1\le k\le L}\sigma_k\mathbf{p}_k\mathbf{q}_k^\top$ where $L = min(n, m)$ and $\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_L$. We further computed cosine similarity between the right principal singular vector $\mathbf{q}_1$ and the item popularity $\mathbf{r}$. The outcomes are showcased in Table 1. From these experiments, we draw an impressive observation:

Observation 1. *The principal right singular vector $\mathbf{q}_1$ of the matrix $\hat{\mathbf{Y}}$ aligns significantly with the item popularity $\mathbf{r}$. The cosine similarity consistently surpasses 0.98 over multiple recommendation models and datasets.*

Given the orthogonal nature of different singular vectors, we can deduce that item popularity is almost entirely captured in the principal spectrum. This intriguing phenomenon elucidates how the recommendation model assimilates item popularity from the data and how this popularity influences recommendation outcomes.

[1]The complete, detailed **appendix** is available in the arXiv version of this paper at https://arxiv.org/abs/2404.12008.

Table 1: The cosine similarity between the principal singular vector ($\mathbf{q}_1$) and the item popularity (r) under different backbones and loss functions.

Backbone	Movielens			Douban			Globo		
	MSE	BCE	BPR	MSE	BCE	BPR	MSE	BCE	BPR
MF	0.993	0.988	0.991	0.992	0.991	0.993	0.993	0.989	0.992
LightGCN	0.992	0.991	0.992	0.990	0.988	0.990	0.992	0.990	0.991
XSimGCL	0.998	0.994	0.995	0.991	0.990	0.992	0.992	0.985	0.989

3.1.2 Theoretical Analyses. Prior to the theoretical validation of observation 1, we posit a power-law hypothesis pertaining to recommendation data:

Hypothesis 1. *The interaction frequency of items in recommendation data follows a power-law distribution (*a.k.a. *Zipf law) described by $r_g \propto g^{-\alpha}$.*

Here r_g signifies the popularity of the g-th most popular item, and α is a shape parameter indicating the distribution's slope. Power-law, as a typical long-tailed distribution, is prevalent across various natural and man-made phenomena [16]. Recent studies assert that item popularity in RS also aligns with this ubiquitous principle [6, 17, 47]. Then we have the following important theorem:

Theorem 1 (Popularity Memorization Effect). *Given an embedding-based recommendation model with sufficient capacity, when training the model on the data with power-law item popularity, the cosine similarity between item popularity $\mathbf{r}$ and the principal singular vector $\mathbf{q}_1$ of the predicted score matrix is bounded with:*

$$\cos(\mathbf{r}, \mathbf{q}_1) \ge \frac{\sigma_1^2}{r_{\max}\sqrt{\zeta(2\alpha)}}\sqrt{1 - \frac{r_{\max}(\zeta(\alpha)-1)}{\sigma_1^2}} \tag{1}$$

For $\alpha > 2$, this can be further bounded with:

$$\cos(\mathbf{r}, \mathbf{q}_1) \ge \sqrt{\frac{2-\zeta(\alpha)}{\zeta(2\alpha)}} \tag{2}$$

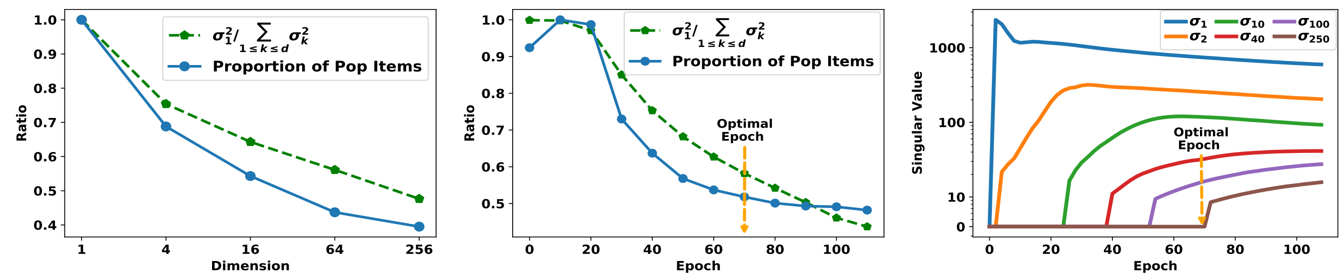
where $r_{\max}$ is the popularity of the most popular item, and $\zeta(\alpha)$ is Riemann zeta function with $\zeta(\alpha) = \sum_{j=1}^{\infty}\frac{1}{j^\alpha}$.

Proof can be found in Appendix A.1. Notably, as the long-tailed nature of item popularity intensifies (*i.e.,* $\alpha \to \infty$ suggesting $\zeta(\alpha) \to 1$), the right side of Eq. (2) converges to one, implying a near-perfect alignment between $\mathbf{r}$ and $\mathbf{q}_1$. Even when the data isn't markedly skewed and has a considerable $\zeta(\alpha)$, we typically observe σ_1^2 to vastly exceed $r_{\max}$, *e.g.,* 5.6×10^5 vs. 4.6×10^3 in the dataset Movielens (with more examples presented in Appendix B.2). Thus, from Eq. (1), a high similarity between $\mathbf{r}$ and $\mathbf{q}_1$ emerges. This theorem provides theoretical validation for our observation 1.

3.2 Popularity Bias Amplification Effect

Earlier discussions illuminate that the principal spectrum memorizes item popularity. In this subsection, we reveal the phenomenon of dimension reduction in RS, which amplifies the effect of the principal spectrum, leading to popularity bias amplification.

3.2.1 Empirical Study. The occurrence of dimension reduction in RS is largely attributable to two factors: 1) explicit low-rank configuration of user/item embeddings [24, 36], and 2) intrinsic training dynamics associated with gradient-based optimization [4, 15, 44]. Here, we present experiments to validate these points and examine their impacts on popularity bias.



(a) Exploring different dimensions. (b) Exploring different epochs. (c) Exploring singular values.

Figure 2: Illustration of how dimension reduction impacts popularity bias in Movielens: (a)-(b) the proportion of popular items in recommendations and the ratio of the largest singular value ($\sigma_1^2/\sum_{1\le k\le L}\sigma_k^2$) with varying embedding dimensions and training epochs, respectively; (c) how singular values evolves during training.

Impact of Low-Rank Configuration. Figure 2(a) displays the proportion of popular items in recommendations from well-trained MF models with varying embedding dimensions d. We also present the magnitude of the largest singular value σ_1 compared with other singular values. We report $\sigma_1^2/\sum_{1\le k\le L}\sigma_k^2$ as it is easily calculable, where the denominator equals the sum of the diagonal elements of $\hat{\mathbf{Y}}$. We observe:

Observation 2. *As the model embedding dimension d is reduced, the relative prominence of the principal singular value increases ($\sigma_1^2/\sum_{1\le k\le L}\sigma_k^2 \uparrow$), and the recommendation increasingly favors popular items.*

This observation reveals the impact of low-rank embeddings. A smaller d squeezes the dimensions (causing singular values of more dimensions to become zero), thereby relatively amplifying the effect of the principal spectrum. Consequently, item popularity contributes more significantly to ranking, resulting in more severe popularity bias.

Dimension Collapse from Gradient Optimization. Figure 2(c) illustrates the evolution of singular values as training progresses using a gradient-based optimizer; and Figure 2(b) offers a dynamic view of popularity bias and the ratio $\sigma_1^2/\sum_{1\le k\le L}\sigma_k^2$ over the training procedure. We observe:

Observation 3. *The principal singular value grows preferentially and swiftly, while others exhibit a more gradual increment. Notably, many singular values appear to be far from convergence even at the end of the training process. Accordingly, popularity bias is severe at the beginning but exhibits a relative decline as training advances. But even at the end of training, unless an extensive number of epochs are employed (which could result in computational overhead and potential over-fitting), the bias remains pronounced.*

This phenomenon reveals the dynamic of singular values during gradient optimization — the principal dimension is prioritized, while singular values of other dimensions are easily under-estimated. This inherent mechanism could readily lead to dimension collapse, relatively enhancing the impact of the principal spectrum, and thereby inducing popularity bias.

3.2.2 Theoretical Analyses. In this subsection, we focus on establishing a theoretical relationship between singular values and the ratio of popular items in recommendations. For readers interested in the theoretical support of the impact of gradient optimization, we refer them to the Appendix A.3, which are relatively straightforward by invoking recent gradient theory [15, 44]. For convenience, our analysis here concentrates on the ratio of the most popular item in top-1 recommendations. We have:

Theorem 2 (Popularity bias amplification). *Given hypothesis 1 and nearly perfect alignment between $\mathbf{q}_1$ and $\mathbf{r}$, the ratio of the most popular item in top-1 recommendations over all users is bounded by:*

$$\eta \ge \frac{1}{n}\phi\left(\frac{\sqrt{2\zeta(2\alpha)}}{1-2^{-\alpha}}\left(\frac{\sum_{1\le k\le L}\sigma_k}{\sigma_1}-1\right)\right) \tag{3}$$

where $\phi(a) = \sum_{u\in\mathcal{U}} \mathbb{I}[p_{1u} > a]$ is an inverse cumulative function calculating the number of elements p_{1u} in the left principal singular vector $\mathbf{p_1}$ exceeding a given value a, and the function $\mathbb{I}[.]$ signifies an indicator function.

The detailed proof is available in Appendix A.2. This theorem vividly showcases the influence of dimension reduction on popularity bias. Essentially, as dimension reduction intensifies the relative prominence of the principle singular value ($\frac{\sigma_1}{\sum_{1\le k\le L}\sigma_k}\uparrow$), the input of the function $\phi(.)$ decreases ($\frac{\sqrt{2\zeta(2\alpha)}}{1-2^{-\alpha}}(\frac{\sum_{1\le k\le L}\sigma_k}{\sigma_1}-1)\downarrow$). Given the monotonically decreasing nature of $\phi(.)$, dimension reduction thus escalates the ratio of most popular items in recommendations. Interestingly, the theorem illustrates the impact of a long-tailed distribution on popularity bias. A larger α (indicating a more skewed item popularity distribution) decreases the value of $\frac{\sqrt{2\zeta(2\alpha)}}{1-2^{-\alpha}}$, further elevating the lower bound of the ratio, intensifying bias.

4 Proposed Method

In this section, we first introduce our proposed debiasing method, followed by a discussion of its properties and a comparison with other debiasing approaches.

4.1 ReSN: Regulation with Spectral Norm

The above analyses elucidate the essence of the popularity bias amplification — the undue influence of the principal spectrum. Therefore, the core of an effective debiasing strategy naturally lies on mitigating this effect. To address this, we propose ReSN which leverages *Spectral Norm* Regularizer to penalize the magnitude of principal singular value:

$$\mathcal{L}_{\mathrm{Re}SN} = \mathcal{L}_R(\mathbf{Y}, \hat{\mathbf{Y}}) + \beta||\hat{\mathbf{Y}}||_2^2 \tag{4}$$

where $\mathcal{L}_R(\mathbf{Y}, \hat{\mathbf{Y}})$ is original recommendation loss, and $||.||_2$ denote the spectral norm of a matrix measuring its principle singular value; β controls the contribution from the regularizer.

However, there are practical challenges: 1) the $n \times m$ dimensional matrix $\hat{\mathbf{Y}}$ can become exceptionally large, often comprising billions of entries, making direct calculations computationally untenable; 2) Existing methods to determine the gradient of the spectral norm are iterative [48, 58], which further adds computational overhead.

To circumvent these challenges, we make two refinements:

Firstly, given the alignment of the principal singular vector $\mathbf{q_1}$ with item popularity $\mathbf{r}$, the calculating of the spectral norm can be simplified into: $||\hat{\mathbf{Y}}||_2^2 = ||\hat{\mathbf{Y}}\mathbf{q_1}||^2 \approx ||\hat{\mathbf{Y}}\mathbf{r}||^2/||\mathbf{r}||^2$, where $||.||$ denotes the L2-norm of a vector. It transforms the calculation of the complex spectral norm of a matrix to a simple L2-norm of a vector, avoiding iterative algorithms by leveraging the singular vector property. Further, the item popularity $\mathbf{r}$ can be quickly computed via $\mathbf{r} = \mathbf{Y}^\top \mathbf{e}$, where $\mathbf{e}$ represents a n-dimension vector filled with ones.

Secondly, we exploit the low-rank nature of the matrix $\hat{\mathbf{Y}}$. For models based on embeddings, $\hat{\mathbf{Y}}$ can be expressed as $\hat{\mathbf{Y}} = \mu(\mathbf{U}\mathbf{V}^\top)$, where $\mathbf{U}$ and $\mathbf{V}$ represent the embeddings associated with users and items, respectively, and $\mu(.)$ designates an activation function. Our approach turns to penalize the spectral norm of the matrix before the introduction of the activation function. This is motivated by the ease of computation: $||\mathbf{U}\mathbf{V}^\top||_2^2 = ||\mathbf{U}(\mathbf{V}^\top\tilde{\mathbf{q}}_1)||^2$, where $\tilde{\mathbf{q}}_1$ denotes the right principal vector of the matrix $\mathbf{U}\mathbf{V}^\top$. By adopting this method, we circumvent the computationally-intensive task of processing the entire matrix $\hat{\mathbf{Y}}$. Nonetheless, this method introduces a challenge: accurately computing $\tilde{\mathbf{q}}_1$, since it doesn't inherently align with item popularity. To rectify this, we may simply mirror the calculation of $\mathbf{q_1} \leftarrow \frac{\mathbf{Y}^\top\mathbf{e}}{||\mathbf{Y}^\top\mathbf{e}||}$ to $\tilde{\mathbf{q}}_1 \leftarrow \frac{\mathbf{V}\mathbf{U}^\top\mathbf{e}}{||\mathbf{V}\mathbf{U}^\top\mathbf{e}||}$. This approach is clued by our Observation 1 and Theorem 1: a matrix's principal singular vector tends to align with the column sum vector, especially when the vector showcases a long-tailed distribution.

To empirically validate the accuracy and rationality of the proposed method, we computed the ideal value of $||\mathbf{U}\mathbf{V}^\top||_2^2$, as well as the estimated $\frac{||\mathbf{U}\mathbf{V}^\top\mathbf{V}\mathbf{U}^\top\mathbf{e}||^2}{||\mathbf{U}\mathbf{V}^\top\mathbf{e}||^2}$ from ReSN, training the MF model with two losses on three datasets. The results are shown in the Table 2. According to the table, we found that the actual spectral norms and our approximate estimates are very close across diverse losses and datasets.This indicates that the singular vector $\tilde{\mathbf{q}}_1$ obtained through $\frac{\mathbf{U}\mathbf{V}^\top\mathbf{e}}{||\mathbf{U}\mathbf{V}^\top\mathbf{e}||}$, serves as an accurate surrogate for the true value of $\mathbf{q_1}$. Therefore, the estimated regularization term is a accurate surrogate for the spectral norm $||\mathbf{U}\mathbf{V}^\top||_2^2$ which validates the precision of this strategy.

In essence, our ReSN optimizes the following loss function:

$$\tilde{\mathcal{L}}_{ReSN} = \mathcal{L}_R(\mathbf{Y}, \hat{\mathbf{Y}}) + \frac{\beta}{||\mathbf{V}\mathbf{U}^\top\mathbf{e}||^2}||\mathbf{U}\mathbf{V}^\top\mathbf{V}\mathbf{U}^\top\mathbf{e}||^2 \tag{5}$$

4.2 Discussions

The proposed ReSN have the following aspects:

Model-Agnostic: The proposed ReSN is model-agnostic and easy to implement. Given that ReSN introduces merely a regularization term, it can be easily plugged into existing embedding-based methods with minimal code augmentation.

Efficiency: The regularizer can be fast computed from right to left — it predominantly requires the multiplication of a $n \times d$ (or $m \times d$) matrix with a vector. With a time complexity of $O((n+m)d)$, ReSN is highly efficient. Section 5.5 also provides empirical evidence. The additional time for calculating the regularizer is negligible.

Table 2: Comparison between the actual spectral norm and the estimated approximation.

Datasets	MSE		BCE	
	$\|\|\mathbf{U}\mathbf{V}^\top\|\|_2^2$	$\|\|\mathbf{U}(\mathbf{V}^\top\tilde{\mathbf{q}}_1)\|\|^2$	$\|\|\mathbf{U}\mathbf{V}^\top\|\|_2^2$	$\|\|\mathbf{U}(\mathbf{V}^\top\tilde{\mathbf{q}}_1)\|\|^2$
Movielens-1M	5.627×10^5	5.613×10^5	5.629×10^5	5.620×10^5
Douban	1.160×10^7	1.155×10^7	1.161×10^7	1.157×10^7
Globo	8.321×10^6	8.309×10^6	8.327×10^6	8.316×10^6

Suitable for BPR Loss: As delineated in Section 2, while BPR can be regarded as a specialized point-wise loss, it involves the concept of hyper-items. It means that the regularizer should be conducted on the embedding matrix of hyper-items $V' \in \mathbb{R}^{m^2 \times d}$, *i.e.,* $\frac{||\mathbf{U}\mathbf{V}'^\top\mathbf{V}'\mathbf{U}^\top\mathbf{e}||^2}{||\mathbf{V}'\mathbf{U}^\top\mathbf{e}||^2}$, rather than $\frac{||\mathbf{U}\mathbf{V}^\top\mathbf{V}\mathbf{U}^\top\mathbf{e}||^2}{||\mathbf{V}\mathbf{U}^\top\mathbf{e}||^2}$. In the following, we will build their approximations. For the numerator part, we have:

$$\mathbf{V}'^\top\mathbf{V}' = \sum_{i,j \in I}(\mathbf{v}_i - \mathbf{v}_j)^\top(\mathbf{v}_i - \mathbf{v}_j) = 2m\mathbf{V}^\top\mathbf{V} - 2m^2\bar{\mathbf{v}}^\top\bar{\mathbf{v}}$$

where $\bar{\mathbf{v}} = \sum_{i \in I}\mathbf{v}_i/m$ denote the mean vector of the item embeddings. Furthermore, current literature posits that an ideal item representation should emulate a uniform distribution over the unit ball [50]. This implies that $\bar{\mathbf{v}}$ tends to gravitate towards the origin. Thus, $\mathbf{V}'^\top\mathbf{V}'$ can be approximated by $\mathbf{V}^\top\mathbf{V}$ and $||\mathbf{U}\mathbf{V}'^\top\mathbf{V}'\mathbf{U}^\top\mathbf{e}||^2$ can be approximated by $||\mathbf{U}\mathbf{V}^\top\mathbf{V}\mathbf{U}^\top\mathbf{e}||^2$.

Similarly, for the denominator:

$$\begin{aligned}||\mathbf{V}'\mathbf{U}^\top\mathbf{e}||^2 &= \sum_{i,j \in I}(\mathbf{v}_i\mathbf{U}^\top\mathbf{e} - \mathbf{v}_j\mathbf{U}^\top\mathbf{e})^\top(\mathbf{v}_i\mathbf{U}^\top\mathbf{e} - \mathbf{v}_j\mathbf{U}^\top\mathbf{e}) \\ &= 2m\sum_{i \in I}(\mathbf{v}_i\mathbf{U}^\top\mathbf{e})^\top(\mathbf{v}_i\mathbf{U}^\top\mathbf{e}) - 2(\sum_{i \in I}\mathbf{v}_i\mathbf{U}^\top\mathbf{e})^\top(\sum_{i \in I}\mathbf{v}_i\mathbf{U}^\top\mathbf{e}) \\ &= 2m(\mathbf{V}\mathbf{U}^\top\mathbf{e})^\top(\mathbf{V}\mathbf{U}^\top\mathbf{e}) - 2m^2(\bar{\mathbf{v}}\mathbf{U}^\top\mathbf{e})^\top(\bar{\mathbf{v}}\mathbf{U}^\top\mathbf{e})\end{aligned}$$

We can deduce $||\mathbf{V}'\mathbf{U}^\top\mathbf{e}||^2$ can be approximated by $||\mathbf{V}\mathbf{U}^\top\mathbf{e}||^2$. Consequently, ReSN emerges as a logical regularizer even for the BPR loss. This assertion is also validated by our experiments.

Differences from Methods on Dimensional Collapse: Recent studies [8, 50, 67] has also employed regularizers to alleviate the dimensional collapse of user/item embeddings. Our ReSN diverges from these methods in two key aspects: 1) ReSN imposes constraints directly onto the prediction matrix, unlike the embedding matrix constraints utilized in these methods. This distinction is of significance due to the inherent spectral gap between the embeddings and the prediction matrix. 2) ReSN explicitly modulates the influence of the principal spectrum that captures popularity information, while these methods mainly focuses on promoting embedding uniformity. ReSN directly and solely mitigates the impact of the memorized popularity signal, thus demonstrating high efficacy in mitigating popularity bias; while others may disrupt the spectral structure of the prediction, potentially compromising model accuracy.

Differences from Regularization-based Debiasing methods: Various regularizers are introduced to combat popularity bias [33, 43, 67, 73]. However, except [67] as discussed before, existing approaches are typically heuristic, applying strong constraints to model predictions that may break the model's original spectrum. While it could mitigate popularity bias, this approach may also impair the model's ability to capture other useful signals, significantly compromising recommendation accuracy. Contrasting this, our ReSN is a light and theoretic-grounding approach — it motivated by the core reason of bias amplification and only modulates the influence of the principle spectrum.

5 Experiments

We conduct experiments to address the following questions:
RQ1: How does ReSN perform compared with other methods?
RQ2: Is ReSN suitable for diversified loss functions and backbones?
RQ3: What is the impact of regularizer coefficient β?
RQ4: How is the efficiency of ReSN ?

5.1 Experiment Settings

Datasets and Metrics. We adopt seven real-world datasets, Yelp2018 [24], Douban [46], Movielens [61], Gowalla [25], Globo [19], Yahoo!R3 [35] and Coat [45] for evaluating our model performance. Details about these datasets refer to Appendix C.1.

We adopt three representative testing paradigms for comprehensive evaluations: 1) **Common**: We employ the conventional testing paradigm in RS, wherein the datasets are randomly partitioned into training (70%), validation (10%), and testing (20%). We also report the accuracy-fairness trade-off in this setting. 2) **Debiased**: Closely referring to [5, 53, 69], we sample an debiased test set where items are uniformly distributed, aiming to evaluate the model's efficacy in mitigating popularity bias. 3) **Uniform-exposure**: We also adopt the uniform exposure paradigm for model testing as the recent work [31, 49, 63]. Notably, the datasets Yahoo!R3 and Coat contain a small dataset collected through a random recommendation policy. Such data isolate the popularity bias from uneven exposure, offering a more precise estimation of user preferences. Consequently, we train our recommendation model on conventionally biased data and then test it on these uniformly-exposed data.

For evaluation metrics, we adopt the widely-used **NDCG@K** for evaluating accuracy [29]. We simply adopt $K = 5$ for Yahoo and Coat datasets and $K = 20$ for the other datasets as recent work [24, 59, 63]. We observe similar results with other metrics. We also employ the **ratio of pop/unpopular items** for illustrating the severity of popularity bias in recommendations. Here we closely refer to recent work [66] to define popular and unpopular items. We sort the items according to their popularity in descending order, and divide items into five groups ensuring the aggregated popularity of items within each group is the same. We define the items in the most popular groups as popular items, while the others as unpopular.

Baselines. The following methods are compared: 1) **MACR** (KDD'21 [53]), **PDA** (SIGIR'21 [66]): the representative causality-based debiasing methods, which posit a causal graph [39] for the recommendation procedure and leverage causal inference to mitigate popularity bias accordingly; 2) **InvCF** (WWW'23 [63]): the SOTA method that addresses popularity bias by disentangling the popularity from user preference. 3) **Zerosum** (Recsys'22 [43]) , **IPL** (SIGIR'23 [33]): the representative methods based on regularizers, which penalize the score differences or constrain the ratio of the predicted preference with the exposure.

For fair comparisons, we implement all compared methods with uniform MF backbone and MSE loss. We also explore the performance with other backbones and losses in subsection 5.3. Besides

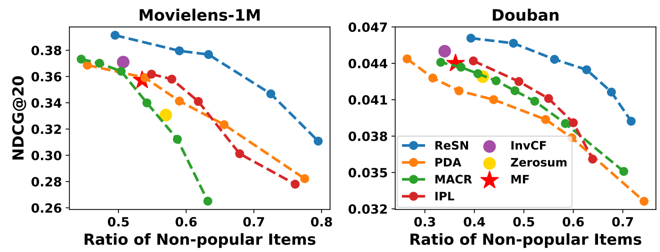


Figure 3: Pareto curves of compared methods illustrating the trade-off between accuracy and fairness under the common testing paradigm.

above baselines, we also compare our method with the methods on mitigating dimension collapse, including nCL [8] and DirectAU [50]; and the debiasing methods tailored for GNN-based methods including APDA [71] and GCF_{logdet} [67] when using GNN-based backbones.

Parameter Settings. The embedding dimension d is 256 while other dimensions are explored in 5.4 . Grid search is utilized find the optimal hyperparameters. More details refer to Appendix C.2

5.2 Performance Comparison (RQ1)

Comparison under three testing paradigms. Table 3 showcases the NDCG@20 comparison across seven datasets over three testing paradigms. Under the Common testing paradigm, our ReSN, with few exceptions, consistently outperforms compared methods. This superior performance can be attributed to the rigorous theoretical foundations of ReSN, which pinpoint and address the root cause of bias amplification. By curbing this bias amplification, ReSN achieves significant improvements in recommendation accuracy. Transitioning to the Debiased and Uniform-exposure testing paradigms, the improvements by ReSN become even more impressive, demonstrating its effectiveness in mitigating popularity bias.

Exploring Accuracy-fairness Trade-off. Given the conventional accuracy-fairness trade-off observed in RS, we delve deeper into examining this effect across various methods. After well-training various methods with differing hyper-parameters (details of hyper-parameters tuning refer to Appendix C.2), we depict the Pareto frontier in Figure 3. It highlights the relationship between accuracy (NDCG@20) and fairness (ratio of unpopular items) under the Common testing paradigm. Here, positions in the top-right corner indicate superior performance. We observe that ReSN exhibits a more favorable Pareto curve in comparison to other baselines. When fairness is held constant, ReSN showcases superior accuracy. Conversely, when accuracy is fixed, ReSN delivers enhanced fairness. This suggests that ReSN effectively navigates the fairness-accuracy trade-off, primarily through its capability to counteract popularity bias amplification — it only mitigates the effect of the principle spectrum without disturbing other spectrum.

Compared with the Methods on Tackling Dimension Collapse. Table 4 shows the results of our ReSN compared with existing methods on tackling dimension collapse on debiased testing paradigm. nCL and DirectAU can indeed mitigate the popularity bias. However, their performance is inferior to ReSN. The reason is that our ReSN is designed for debiasing, directly modulating the effect of the item popularity on predictions, and thus yielding better performance.

Table 3: Performance comparison in terms of NDCG between ReSN and other baselines across seven datasets and three testing paradigms. The "Com"(refers to "Common") represents the paradigm where the training and test datasets are partitioned randomly; "Deb"(refers to "Debiased") represents the paradigm where a debiased test dataset is formulated based on item popularity; "Uni"(refers to "Uniform-exposure") represents the paradigm where the test data is uniformly-exposed. The best result is bolded and the runner-up is underlined. The mark '*' denotes the improvement achieved by ReSN over best baseline is significant with $p < 0.05$.

	Movielens		Douban		Yelp2018		Gowalla		Globo		Yahoo	Coat
	Com	Deb	Com	Deb	Com	Deb	Com	Deb	Com	Deb	Uni	Uni
MF	0.3572	0.1490	0.0440	0.0116	0.0416	0.0164	0.1182	0.0438	0.1709	0.0028	0.6672	0.5551
Zerosum	0.3309	0.1411	0.0434	0.0110	0.0415	0.0137	0.1063	0.0421	0.1630	0.0036	0.6665	0.5633
MACR	0.3732	0.1647	0.0441	0.0145	0.0404	0.0208	0.1107	0.0545	**0.1782**	0.0253	0.6714	0.5661
PDA	0.3688	0.1662	0.0446	0.0171	0.0437	0.0229	0.1283	0.0675	0.1725	0.0243	0.6756	0.5676
InvCF	0.3723	0.1567	0.0450	0.0152	0.0433	0.0183	0.1302	0.0592	0.1671	0.0194	0.6519	0.5715
IPL	0.3618	0.1621	0.0442	0.0173	0.0419	0.0219	0.1318	0.0623	0.1715	0.0203	0.6691	0.5602
ReSN	**0.3857***	**0.1745***	**0.0456***	**0.0186***	**0.0445***	**0.0254***	**0.1343***	**0.0703***	0.1682	**0.0256***	**0.6792***	**0.5871***

Table 4: NDCG@20 comparison with methods for addressing Dimension Collapse under the debiased testing paradigm.

	Movielens	Douban	Gowalla
MF	0.1529	0.0116	0.0438
nCL	0.1572	0.0112	0.0451
DirectAU	0.1691	0.0131	0.0622
ReSN	**0.1788**	**0.0188**	**0.0712**

Table 5: NDCG@20 comparison with GNN-based backbones (LightGCN, XSimGCL) under the debiased testing paradigm.

	Movielens		Douban		Gowalla	
	LGCN	XSGCL	LGCN	XSGCL	LGCN	XSGCL
Backbone	0.1531	0.1686	0.0117	0.0132	0.0446	0.0563
Zerosum	0.1363	0.1438	0.0112	0.0129	0.0437	0.0498
MACR	0.1682	0.1692	0.0157	0.0164	0.0543	0.0623
PDA	0.1684	0.1732	0.0182	0.0190	0.0689	0.0732
InvCF	0.1602	0.1672	0.0153	0.0169	0.0599	0.0687
IPL	0.1653	0.1701	0.0166	0.0193	0.0642	0.0699
APDA	0.1657	0.1713	0.0156	0.0189	0.0468	0.0522
$\mathrm{GCF}_{\mathrm{logdet}}$	0.1672	0.1724	0.0124	0.0141	0.0403	0.0492
ReSN	**0.1758**	**0.1810**	**0.0194**	**0.0202**	**0.0717**	**0.0763**

Table 6: NDCG@20 comparison with different Loss functions under the debiased testing paradigm.

	Movielens		Douban		Gowalla	
	+BCE	+BPR	+BCE	+BPR	+BCE	+BPR
MF	0.1529	0.1540	0.0117	0.0120	0.0432	0.0431
Zerosum	0.1472	0.1498	0.0109	0.0106	0.0423	0.0425
MACR	0.1682	0.1629	0.0155	0.0149	0.0574	0.0546
PDA	0.1635	0.1633	0.0176	0.0173	0.0661	0.0675
InvCF	0.1574	0.1582	0.0153	0.0154	0.0553	0.0583
IPL	0.1612	0.1628	0.0173	0.0177	0.0612	0.0626
ReSN	**0.1788**	**0.1693**	**0.0188**	**0.0180**	**0.0712**	**0.0702**

5.3 Adaptability Exploration (RQ2)

To investigate the adaptability of ReSN, we evaluate it with various backbones and loss functions under debiased testing paradigm. Table 5 showcases the performance of ReSN under LightGCN [24] and XSimGCL [59], where APDA [71] and $\mathrm{GCF}_{\mathrm{logdet}}$ [67] that tailored for GNN-based backbones are included. Besides, Table 6 depicts the results with BCE and BPR losses. Notably, ReSN consistently outperforms compared methods, irrespective of the chosen backbone or loss function. These findings affirm the great adaptability of ReSN, underscoring its ability to seamlessly integrate with diverse recommendation models. Our ReSN also outperforms those debiasing methods tailored for GNN-based methods. They reason is that they only consider to mitigate bias amplification raised by GNNs while ignoring the bias from the generic recommendation mechanism.

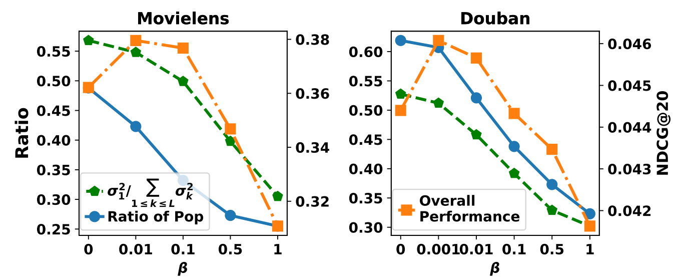


Figure 4: The proportion of popular items in recommendations and the ratio of the largest singular value ($\sigma_1^2/\sum_{1\le k\le L}\sigma_k^2$) and NDCG@20 with varying β.

Table 7: Running time comparison (s/Epoch), where ReSN-Direct directly calculates spectral norm without acceleration.

	Movielens	Douban	Gowalla
MF	0.177	2.098	0.634
ReSN	0.181	2.124	0.675
ReSN-Direct	649	59239	15280
Speedup Ratio	3585	27890	22637

5.4 Hyperparameter Study (RQ3)

Figures 4 presents the recommendation accuracy, the ratio of popular items, and the ratio of principle singular value ($\sigma_1^2/\sum_{1\le k\le L}\sigma_k^2$) in ReSN as the hyperparameter β varies. Notably, as β increases, the ratio of principle singular value and the severity of popularity bias reduces. This trend affirms the efficacy of our regularizer. Regarding recommendation accuracy, it initially rises and then declines with an increase in β. This can be attributed to the fact that popularity bias isn't intrinsically detrimental [66, 68]. Indeed, item popularity can also convey beneficial information about the item's appeal or quality, which ought to be retained. Hence, the strategic approach for popularity bias lies in mitigating its bias amplification, rather than eliminating it entirely. That is also our target.

5.5 Efficiency Study (RQ4)

To further validate the effectiveness of our acceleration strategy, we test the running time per epoch of ReSN and the original brute-force strategy employed to compute the gradient of the spectral norm. Also, we present the baseline MF for comparison. The results are presented in Table 7. As can be seen, our acceleration strategy achieves over 3600, 27000 and 22000 times impressive speed-up in both datasets, respectively. Moreover, compared with MF, our ReSN does not incur much computational overhead.

6 Related Work

Analyses on Popularity Bias. In RS, items frequently exhibit a long-tailed distribution in terms of interaction frequency. Models trained on skewed data are susceptible to inheriting and exacerbating such bias [2, 3, 26, 55, 72, 73]. The crux of tackling popularity bias lies in understanding why and how recommendation models intensify popularity bias. Several recent efforts aim to elucidate this. Among these, causality-based investigations stand out. For instance, Wang et al. [52], Zhang et al. [66] developed a causal graph of the data generative process, attributing the amplification of popularity bias to a confounding effect; Wei et al. [53] presented an alternate causal graph, exploring the direct and indirect causal influence of popularity bias on predictions. A common limitation among these causality-based methods is their surface-level engagement with the causal relationships among variables, rather than delving deeper into the underlying mechanisms. For example, these studies usually operate on the assumption that item popularity directly affects predictions. However, the specifics of how and why predictions memorize and are influenced by item popularity remain largely unexplored. Worse still, their effectiveness hinges on the accuracy of their respective causal graphs, which might not always hold due to the unmeasured confounders [20, 30].

There were other investigations into popularity bias. For instance, Zhu et al. [73] demonstrate that model predictions inherit item popularity, yet they failed to elucidate the amplification. Also, their conclusions rely on a strong assumption that the preference scores maintain same distribution across different user-item pairs. The study by [38] shed light on the limited expressiveness of low-rank embeddings, giving clues of popularity bias in recommendations. Yet they did not factor in the impact of long-tailed training data. In fact, popularity bias origins from long-tailed data [66, 73], amplified during training, which would be more serious than the theoretically analyses presented in [38]. Some efforts [13, 28] examined popularity bias through embedding magnitude, their theoretical analysis can only applied in the early stages of training. Other researchers delved into how graph neural networks amplify popularity bias through influence functions [11], the hub effect [71] or dimensional collapse [67]. However, their conclusions can not be extended to general recommendation models.

Methods on Tackling Popularity Bias. Recent efforts on addressing popularity bias are mainly four types: 1) Causality-driven methods assume a causal graph to identify popularity bias and employ causal inference techniques for rectification. While they have demonstrated efficacy, their success is closely tied to the accuracy of the causal graph. This poses challenges due to the prevalence of unmeasured confounders [20, 30, 37]. 2) Propensity-based methods [9, 23, 45, 51, 64] adjust the data distribution by reweighting the training data instances. While this approach directly negates popularity bias in the data, it may inadvertently obscure other valuable signals, such as item quality. Consequently, these methods often underperform compared to causality-driven ones. 3) Regularizer-based methods [1, 27, 33, 43, 73] constrain predictions by introducing regularization terms. For example, Zhu et al. [73] employs a Pearson coefficient regularizer to diminish the correlation between item popularity and model predictions; Zhang et al. [67] adopts a regularizer for mitigating embeddings collapse; Rhee et al. [43] proposes to regularize the score differences; [33] constrains the predictions with IPL criterion. As discussed in section 4.2, their constraints are too strong, may significantly compromising accuracy. 4) Disentanglement-based methods [14, 57, 63] target at learning disentangled embeddings that segregate the influence of popularity from genuine user preferences. While promising, achieving a perfect disentanglement of popularity bias from true preferences remains a formidable challenge in RS.

Among the related work, the one most closely related to ours is [67], but we emphasize that our work differs in two key aspects: 1) Their theoretical justification of bias amplification focuses solely on GNNs, whereas our analysis applies to generic recommendation mechanisms. 2) Their regularizer aims to mitigate collapse of user/item embeddings, while our ReSN specifically targets the mitigation of the principal spectrum's influence. Section 4.2 provides a detailed discussion of these differences, demonstrating that ReSN is more effective in debiasing. Table 4 also offers empirical evidence supporting our claims.

7 Conclusion

In this study, we delve into the root cause of popularity bias amplification. Our analyses offer two core insights: 1) Item popularity is encoded in the principal spectrum of model predictions; 2) The phenomenon of dimension reduction accentuates the influence of the principal spectrum. Based on these insights, we introduce ReSN, an efficient technique aimed at mitigating popularity bias by penalizing the principle singular value. A potential limitation of our study pertains to the static perspective on popularity bias, neglecting its dynamic nature as it evolves temporally. It could be more insightful to investigate the mechanism of bias amplification in the context of temporal sequential recommendations, and to examine its evolution during the feedback loop.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62372399, 62476244) and the advanced computing resources provided by the Supercomputing Center of Hangzhou City University.

References

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*. 42–46.

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The impact of popularity bias on fairness and calibration in recommendation. *arXiv preprint arXiv:1910.05755* (2019).

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 726–731.

[4] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. 2019. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems* 32 (2019).

[5] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *Proceedings of the 12th ACM conference on recommender systems*. 104–112.

[6] Rodrigo Borges and Kostas Stefanidis. 2020. On Measuring Popularity Bias in Collaborative Filtering Data. In *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020*, Vol. 2578.

[7] Hao Chen, Yuanchen Bei, Qijie Shen, Yue Xu, Sheng Zhou, Wenbing Huang, Feiran Huang, Senzhang Wang, and Xiao Huang. 2024. Macro graph neural networks for online billion-scale recommender systems. In *Proceedings of the ACM on Web Conference 2024*. 3598–3608.

[8] Huiyuan Chen, Vivian Lai, Hongye Jin, Zhimeng Jiang, Mahashweta Das, and Xia Hu. 2023. Towards Mitigating Dimensional Collapse of Representations in Collaborative Filtering. *arXiv preprint arXiv:2312.17468* (2023).

[9] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. 2021. Autodebias: Learning to debias for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 21–30.

[10] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[11] Jiajia Chen, Jiancan Wu, Jiawei Chen, Xin Xin, Yong Li, and Xiangnan He. 2023. How Graph Convolutions Amplify Popularity Bias for Recommendation? *arXiv preprint arXiv:2305.14886* (2023).

[12] Jiajia Chen, Jiancan Wu, Jiawei Chen, Xin Xin, Yong Li, and Xiangnan He. 2024. How graph convolutions amplify popularity bias for recommendation? *Frontiers of Computer Science* 18, 5 (2024), 185603.

[13] Jiawei Chen, Junkang Wu, Jiancan Wu, Xuezhi Cao, Sheng Zhou, and Xiangnan He. 2023. Adap-τ: Adaptively Modulating Embedding Magnitude for Recommendation. In *Proceedings of the ACM Web Conference 2023*. 1085–1096.

[14] Zhihong Chen, Jiawei Wu, Chenliang Li, Jingxu Chen, Rong Xiao, and Binqiang Zhao. 2022. Co-training disentangled domain adaptation network for leveraging popularity bias in recommenders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 60–69.

[15] Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. 2023. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *Applied and Computational Harmonic Analysis* (2023), 101595.

[16] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[17] Ludovik Çoba, Panagiotis Symeonidis, and Markus Zanker. 2017. Visual analysis of recommendation performance. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 362–363.

[18] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.

[19] Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. 2018. News session-based recommendations using deep neural networks. In *Proceedings of the 3rd workshop on deep learning for recommender systems*. 15–23.

[20] Sihao Ding, Peng Wu, Fuli Feng, Yitong Wang, Xiangnan He, Yong Liao, and Yongdong Zhang. 2022. Addressing unmeasured confounder for recommendation with sensitivity analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 305–315.

[21] Chongming Gao, Kexin Huang, Jiawei Chen, Yuan Zhang, Biao Li, Peng Jiang, Shiqi Wang, Zhong Zhang, and Xiangnan He. 2023. Alleviating Matthew Effect of Offline Reinforcement Learning in Interactive Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 238–248.

[22] Chongming Gao, Shiqi Wang, Shijun Li, Jiawei Chen, Xiangnan He, Wenqiang Lei, Biao Li, Yuan Zhang, and Peng Jiang. 2023. CIRS: Bursting filter bubbles by counterfactual interactive recommender system. *ACM Transactions on Information Systems* 42, 1 (2023), 1–27.

[23] Alois Gruson, Praveen Chandar, Christophe Charbuillet, James McInerney, Samantha Hansen, Damien Tardieu, and Ben Carterette. 2019. Offline evaluation to make decisions about playlistrecommendation algorithms. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 420–428.

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

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

[26] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25 (2015), 427–491.

[27] Jiarui Jin, Zexue He, Mengyue Yang, Weinan Zhang, Yong Yu, Jun Wang, and Julian McAuley. 2024. InfoRank: Unbiased Learning-to-Rank via Conditional Mutual Information Minimization. In *Proceedings of the ACM on Web Conference 2024*. 1350–1361.

[28] Dain Kim, Jinhyeok Park, and Dongwoo Kim. 2023. Test Time Embedding Normalization for Popularity Bias Mitigation. *arXiv preprint arXiv:2308.11288* (2023).

[29] Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1748–1757.

[30] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, and Peng Wu. 2023. Balancing unobserved confounding with a few unbiased ratings in debiased recommendations. In *Proceedings of the ACM Web Conference 2023*. 1305–1313.

[31] Siyi Lin, Sheng Zhou, Jiawei Chen, Yan Feng, Qihao Shi, Chun Chen, Ying Li, and Can Wang. 2024. ReCRec: Reasoning the Causes of Implicit Feedback for Debiased Recommendation. *ACM Transactions on Information Systems* 42, 6 (2024), 26 pages.

[32] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. 2022. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the ACM Web Conference 2022*. 2320–2329.

[33] 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*. 1801–1805.

[34] 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.

[35] Benjamin M Marlin and Richard S Zemel. 2009. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*. 5–12.

[36] Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems* 20 (2007).

[37] Wentao Ning, Reynold Cheng, Xiao Yan, Ben Kao, Nan Huo, Nur Al Hasan Haldar, and Bo Tang. 2024. Debiasing Recommendation with Personal Popularity. In *Proceedings of the ACM on Web Conference 2024*. 3400–3409.

[38] Naoto Ohsaka and Riku Togashi. 2023. Curse of "Low" Dimensionality in Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 537–547.

[39] Judea Pearl. 2009. *Causality*. Cambridge university press.

[40] Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. 2018. Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2110–2119.

[41] Steffen Rendle. 2022. Item recommendation from implicit feedback. In *Recommender Systems Handbook*. Springer.

[42] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.

[43] Wondo Rhee, Sung Min Cho, and Bongwon Suh. 2022. Countering Popularity Bias by Regularizing Score Differences. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 145–155.

[44] Andrew M Saxe, James L McClelland, and Surya Ganguli. 2019. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences* 116, 23 (2019), 11537–11546.

[45] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*. PMLR, 1670–1679.

[46] Weiping Song, Zhiping Xiao, Yifan Wang, Laurent Charlin, Ming Zhang, and Jian Tang. 2019. Session-based social recommendation via dynamic graph attention networks. In *Proceedings of the Twelfth ACM international conference on web search and data mining*. 555–563.

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

[48] Shuhan Sun, Zhiyong Xu, and Jianlin Zhang. 2021. Spectral norm regularization for blind image deblurring. *Symmetry* 13, 10 (2021), 1856.

[49] Bohao Wang, Jiawei Chen, Changdong Li, Sheng Zhou, Qihao Shi, Yang Gao, Yan Feng, Chun Chen, and Can Wang. 2024. Distributionally Robust Graph-based Recommendation System. In *Proceedings of the ACM on Web Conference 2024*. 3777–3788.

[50] Chenyang Wang, Yuanqing Yu, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. 2022. Towards representation alignment and uniformity in collaborative filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1816–1825.

[51] Lei Wang, Chen Ma, Xian Wu, Zhaopeng Qiu, Yefeng Zheng, and Xu Chen. 2024. Causally Debiased Time-aware Recommendation. In *Proceedings of the ACM on Web Conference 2024*. 3331–3342.

[52] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. 2021. Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1717–1725.

[53] 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 Conference on Knowledge Discovery & Data Mining*. 1791–1800.

[54] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 726–735.

[55] Jiancan Wu, Xiang Wang, Xingyu Gao, Jiawei Chen, Hongcheng Fu, Tianyu Qiu, and Xiangnan He. 2024. On the Effectiveness of Sampled Softmax Loss for Item Recommendation. *ACM Transactions on Information Systems* 42, 4 (2024).

[56] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2022. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2022), 4425–4445.

[57] Guipeng Xv, Chen Lin, Hui Li, Jinsong Su, Weiyao Ye, and Yewang Chen. 2022. Neutralizing popularity bias in recommendation models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2623–2628.

[58] Yuichi Yoshida and Takeru Miyato. 2017. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941* (2017).

[59] Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Nguyen Quoc Viet Hung, and Hongzhi Yin. 2023. XSimGCL: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[60] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1294–1303.

[61] Wenhui Yu and Zheng Qin. 2020. Graph convolutional network for recommendation with low-pass collaborative filters. In *International Conference on Machine Learning*. PMLR, 10936–10945.

[62] An Zhang, Wenchang Ma, Pengbo Wei, Leheng Sheng, and Xiang Wang. 2024. General Debiasing for Graph-based Collaborative Filtering via Adversarial Graph Dropout. In *Proceedings of the ACM on Web Conference 2024*. 3864–3875.

[63] An Zhang, Jingnan Zheng, Xiang Wang, Yancheng Yuan, and Tat-Seng Chua. 2023. Invariant Collaborative Filtering to Popularity Distribution Shift. In *Proceedings of the ACM Web Conference 2023*. 1240–1251.

[64] Fan Zhang and Qijie Shen. 2023. A Model-Agnostic Popularity Debias Training Framework for Click-Through Rate Prediction in Recommender System. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1760–1764.

[65] Xiang Zhang, Douglas E Faries, Hu Li, James D Stamey, and Guido W Imbens. 2018. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiology and drug safety* 27, 4 (2018), 373–382.

[66] 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*. 11–20.

[67] Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, Irwin King, et al. 2023. Mitigating the Popularity Bias of Graph Collaborative Filtering: A Dimensional Collapse Perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[68] Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, and Wei Wu. 2022. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[69] 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*. 2980–2991.

[70] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.

[71] Huachi Zhou, Hao Chen, Junnan Dong, Daochen Zha, Chuang Zhou, and Xiao Huang. 2023. Adaptive Popularity Debiasing Aggregator for Graph Collaborative Filtering. (2023), 7–17.

[72] Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. 2022. Evolution of Popularity Bias: Empirical Study and Debiasing. *arXiv preprint arXiv:2207.03372* (2022).

[73] 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*. 85–93.