Overcoming Undesired Effects in Personalized Recommender Systems by Leveraging Uncertainty

Paula Gómez Duran[1*], Pere Gilabert[1], Santi Seguí[1], Jordi Vitrià[1]

[1*]Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, Barcelona, 08007, Spain.

*Corresponding author(s). E-mail(s): paulagomezduran@gmail.com;

Abstract

In today's digital landscape, Recommender Systems (RSs) have become an omnipresent tool for guiding users towards products, services, and content tailored to their interests. However, despite their widespread adoption and a long track of research, these systems are not immune to shortcomings. A significant challenge faced by RSs is the perpetuation of existing biases, resulting in a multitude of undesirable effects, most notably popularity bias. This bias tends to restrict the variety of recommended items, limiting users' exposure to blockbuster or popular content. Consequently, it can exacerbate societal concerns, including the erosion of trust in media organizations that struggle to deliver truly personalized recommendations. This deficiency in serving users who deviate from mainstream trends further impedes user diversity and engagement, underscoring the need for improved RSs algorithms.

In order to tackle the undesired effects of RSs, we propose a stochastic ranking method for serving truly personalized recommendations. By harnessing the uncertainty inherent in RSs predictions, our approach facilitates the delivery of more responsible and diverse recommendation lists. Our approach places a premium on fairness and personalization, advocating a paradigm shift in RSs optimization objectives that differentiates them in a valuable way from merely predicting a user's next action. Through extensive experimentation, we substantiate the efficacy of our approach in mitigating the adverse impact of popularity bias from both user and item perspectives, ultimately enhancing various beyond accuracy metrics. This study underscores the significance of responsible and equitable recommendations in cultivating a healthier online environment and represents a meaningful stride toward more accountable RSs.

Keywords: Recommender Systems, Uncertainty, Beyond Accuracy Metrics, Fairness, Diversity.

1 Introduction

In today's digital era, Recommender Systems (RSs) have become an indispensable tool to aid users in exploring fresh and varied content that resonates with their interests and values. However, despite their immense utility, these systems are susceptible to the obstacles posed by particularities in data collection as well as by undesired effects, which are unintentionally caused or intensified by recommendation algorithms.

Recent research has emphasized the importance of increasing our understanding of how algorithms behave and advocating for unbiased data collection, which has led to a new focus on ensuring fair, equitable, and diverse recommendations. This has given rise to the concept of Responsible Recommendations (RRs) [1]. However, the goal of RRs is in conflict with certain unintended effects of RSs, which in turn give rise to issues of unfairness. While it is commonly argued that these undesirable effects stem from the functioning and optimization objectives of the underlying algorithms, failing to address their causes can lead to self-reinforcing feedback loops that perpetuate detrimental consequences, such as the propagation of popularity bias and the erosion of user diversity, among others. Therefore, it is crucial to consider the implications of these unintended effects and prioritize fair and diverse recommendations in order to foster responsible and equitable online environments.

Recommendation algorithms have traditionally focused on optimizing user behavior over time and then presenting content that only matches past behavior. However, optimizing metrics for this task can create a feedback loop and contribute to addiction to social media platforms, which has become a growing concern in recent years [2, 3]. In order to mitigate these effects, some organizations have been motivated to follow a specific agenda in order to move toward more responsible recommenders, and thus every time more companies are becoming interested in avoiding negative consequences that may arise from existing recommendation algorithms [1]. For example, in the field of Public Service Media (PSM) in Europe, many TV broadcasters have explicitly stated their mission to provide unbiased information and deliver diverse content [4–7].

An important undesired effect is 'popularity bias', which occurs because the algorithm pushes just for accuracy by reinforcing existing biases and presenting items that strongly align with observed users' behaviors, often coinciding with the *mainstream*. This can limit the discovery of fresh or niche items, hence reducing diversity in recommendations and also harming those users who do not follow the trend. Real-world datasets frequently exhibit a significant portion of interactions that involve a small number of popular ('blockbuster') items, representing what is commonly known as the 'short head' users. The remaining data is typically associated with the 'medium tail' users, who consume average ('diverse') items, and the 'long tail' users, who consume more unique ('niche') items [8–10]. Despite, in theory, RSs are designed to help users explore new and relevant content which probably lies on the *medium* or *long tail*, these algorithms often suffer from inherent biases that prioritize the accuracy metrics and end up promoting the 'blockbuster' items or mainstream content, leading to an increase in the viewership of the *short head* [11, 12]. Those undesired effects can worsen societal issues and damage trust in media platforms. Therefore, prioritizing fair and diverse recommendations is necessary for responsible and equitable online

environments and further research needs to be conducted on the implications of RSs and the role of RRs for mitigating these negative effects.

In this article, we delve into the impact of biases on RSs, particularly focusing on popularity bias, and present a novel approach that steers RSs toward RRs by incorporating prediction uncertainty in the ranking (serving) stage. We critically examine previous efforts to mitigate undesired effects in RSs and highlight their advantages. Moreover, we contend that a paradigm shift in the RSs task is imperative to advocate for genuinely personalized and RRs, aligning with the proposition put forth by Pellegrini et al. [13]. Additionally, we aim to conduct an interpretability assessment to gain insights into the ranking generation process for both traditional and new responsible tasks. Overall, our methodology seeks to facilitate the discovery of relevant content for users while also delivering value to media providers in real-world production settings [14].

2 Related work

Recognized as a powerful tool for alleviating information overload, RSs have revolutionized numerous applications by providing personalized suggestions to individual users. However, the prevalence of bias within RSs poses significant challenges that can undermine the effectiveness and fairness of recommendations. Recent years have witnessed a resurgence of research interest in addressing the issue of unfairness in RSs, particularly within the domains of machine learning and artificial intelligence [1]. Various types of bias commonly afflict RSs, originating from several key factors which can arise within three different stages: Collection, Learning and Serving [15]. Each stage introduces its own set of biases that can impact the overall recommendation process.

In the **Collection** stage, biases can arise due to the reliance on observational data to capture user behaviors. These data biases mainly arise from two perspectives. On the one hand, from users' behaviors being influenced by the items they are exposed to. This can lead to *User Exposure Bias* [16, 17] and also several confounding factors which stem from how the RS exposes items. One of these factors is *Position Bias* [18], where users tend to interact with items in higher positions of the list. Another factor is *User Selection Bias* [19, 20], which occurs when users are free to choose which items to rate, resulting in observed ratings that may not be representative of all ratings. Finally, there is the *Conformity Bias* [21], which reflects users' tendency to behave similarly to others in a group. On the other hand, due to the uneven distribution of item presentations in the data. Certain items, owing to their popularity, receive more user interactions, which disproportionately impact model training and subsequently bias the recommendations toward these popular items, commonly known as *item popularity bias* [22].

Moving on to the **Learning** stage, which involves the training of recommendation models based on the collected data, the feedback loop inherent in RS perpetuates and intensifies biases over time. The exposure mechanism of RS shapes user behaviors, which are then used as training data for the models. This feedback loop creates a self-reinforcing cycle known as the Matthew effect, 'the rich get richer' phenomenon. The

resultant *popularity bias algorithm* leads to a situation where highly popular items are excessively recommended, while less popular items struggle to gain visibility [23, 24].

Finally, in the **Serving** phase, where the recommendation results are returned to users, several undesired effects can arise as a consequence of the biases present in the previous stages. The biased training process often leads to recommendations that lack diversity, making it challenging for less popular items to receive sufficient exposure. This can result in *unfairness*, meaning that the system systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others. This is often reflected in *item novelty bias*, where recommendations tend to favor new or recently released items over older ones [25, 26] and other undesired effects, such as *low catalog coverage*, *low serendipity*, and a *lack of diversity* within the user's recommended lists can also occur.

Addressing these biases and their detrimental effects is paramount. By developing advanced techniques that account for bias and promote fairness, RSs can deliver more diverse and unbiased recommendations, enhancing user experience and fostering equitable access to a wide range of content, finally aiming toward RRs. In order to address this problem there have been multiple approaches. Although many of them focus on debiasing the *Collection stage* [27–29], implementing these approaches often requires making certain assumptions about data generation, leading to high variance or challenging training processes. Consequently, many research works have directed their efforts toward addressing popularity bias algorithm and the issue of unfairness in RSs [8, 30–33]. However, un-biasing a RS is not a straightforward task, and these proposed methods have often resulted in a decrease in accuracy, which has discouraged providers from actively advocating for more fair and RRs. Furthermore, while various research endeavors have targeted biases in the *Serving stage* [17, 34, 35], these methods often face the challenge of compromising the user experience, making it difficult to implement them effectively. Overcoming biases in the serving stage remains an ongoing and complex challenge in the field of RSs.

All in all, despite the existence of many methods that aim to address biases in RSs, a comprehensive and holistic approach that considers the entire life-cycle of recommendation is still lacking. Current methods often focus on addressing specific biases individually, but fail to account for the interconnected nature of biases throughout the recommendation process. Moreover, these methods may rely on assumptions about data generation or present implementation challenges. To effectively tackle biases, it is crucial to adopt a joint perspective that considers biases that may arise at each stage of the recommendation process. This holistic approach will enable researchers to overcome limitations of individual bias mitigation methods and provide more effective and comprehensive solutions to the biases present in RSs.

3 Proposed approach

In this work, we make three significant contributions. Firstly, our work introduces a novel approach to the recommendation task by undertaking a comprehensive analysis from the viewpoint of the full-ranking. This perspective allows us to simulate a real-world product stage scenario, enabling a more accurate assessment of RSs. By

optimizing the new task proposed in [13], we aim to uncover the profound impact of this innovative concept by utilizing metrics that go beyond accuracy [15, 36], providing a more holistic evaluation of the system's performance and its ability to achieve true personalization. Secondly, we propose an enrichment of the recommendation concept to promote RRs. This is accomplished by harnessing the uncertainty in the predictions, allowing us to develop a principled method that incorporates considerations of fairness, equity, and diversity. Lastly, we delve into the analysis of various RSs behaviors, aiming to enhance the interpretability of ranking predictions and provide users with a greater sense of trust. Additionally, we investigate whether RRs can yield more accurate recommendations for different user types.

First contribution: Holistic Evaluation of new RS task

In the realm of RSs, evaluating model candidates plays a pivotal role in the development process. While online evaluation is the preferred method for assessing recommender models [37], it is not always practical or efficient during the initial stages, particularly when searching for optimal hyper-parameter settings. Consequently, offline evaluation remains the primary approach for the recommender community to assess new models during development.

As the number of items to be recommended by RSs has increased, evaluating the ranking per user across the entire catalog has become increasingly time-consuming. To tackle this challenge, an alternative and more efficient evaluation method was introduced by Koren [38] in 2008. This method involved computing metrics on a small subset of items, which consisted of all relevant items and a specific number C of randomly sampled non-relevant items from the complete set. The objective was to create a representative test subset that accurately reflected the distribution of the entire test set, enabling model evaluation as if it were performed on the complete item set. This approach gained wide adoption in subsequent works [39–41] due to its advantage of reducing evaluation time by avoiding the computation of metrics on the full item set, commonly referred to as *full-ranking* evaluation. While full-ranking strategies closely resembled real-world production scenarios for RSs, evaluating within a subset of items provided an encouraging approach to significantly reduce evaluation time while still allowing for model comparison and selection. However, recent studies over time have demonstrated that evaluating on subsets does not consistently align with full-ranking evaluations [42, 43]. While subsequent research aimed to create more reliable and representative test subsets that accurately captured the distribution of the full set (e.g., selecting C candidate items based on popularity distribution) [44], the final conclusions indicated that neither sampling strategy produced satisfactory approximations of the full-ranking that could be consistently applied across multiple datasets.

Pellegrini et al. [13] introduced a fresh perspective on the RSs task. They argued that traditional RSs have excelled at predicting the most probable item a user will interact with next, but this standard approach falls short in achieving robust personalization, which is a crucial component of effective RSs. Instead, they advocated for the advancement toward a truly personalized RSs that offers users highly tailored recommendations aligned with their specific interests. To achieve this, they proposed

optimizing RSs for a task that involves identifying the preferred items within a sampled set of items based on popularity distribution. In order to evaluate this task, they conducted evaluations on a test set that was sampled also according to popularity distribution to assess the performance of this system. However, there remains a crucial aspect to consider: how an RS trained for this task will perform in a real-world production scenario, which necessitates a more comprehensive offline evaluation approach that resembles the full-ranking process.

In their seminal work, Abdollahpouri et al. [45] emphasizes the importance of considering multiple metrics simultaneously to draw meaningful conclusions about the suitability of different RSs approaches. Similarly, Pellegrini et al. [13] advocates for quantitative evaluations of the new recommendation task from various perspectives and encourages the exploration of multiple evaluation approaches. It is widely recognized that RSs algorithms can be susceptible to biases and may even reinforce them. However, understanding the precise mechanisms behind these effects and finding effective mitigation strategies is not always straightforward.

To accomplish our objective, we propose evaluating the standard Factorization Machine [46] model using the full-ranking process on four different datasets. We assess the model's performance in both the traditional RS task of next-item prediction and the new RS task focused on achieving true personalization. In order to gain a comprehensive understanding of how each model behaves in a production stage setting, we introduce a set of metrics which go beyond accuracy. By assessing a RS with all those metrics we aim to provide an analysis that indicates its best performance within a full-ranking scenario from the following perspectives (see section 4.3 for details about the metrics):

1. **Accuracy:** We assess the RS accuracy using two commonly used metrics, namely Hit Ratio (HR) and normalized Discounted Cumulative Gain (nDCG). Accuracy metrics provide insights into the value attributed to media platforms based on user preferences. However, it is important to acknowledge that accuracy metrics may not always provide a completely accurate assessment, as biases in the offline evaluation data can introduce potential distortions. Therefore, the reliability of accuracy reports is enhanced when the RS employed for data collection during the production stage is aware of these biases and takes measures to prevent the reinforcement of biased loops.
2. **Fairness:** We address fairness from the item perspective by introducing a novel metric called 'Error of Exposure'. This metric measures the reduction in item exposure error, which is defined as the difference between an item's exposure during the training stage and its exposure during the inference stage. By evaluating fairness, we aim to ensure equitable item exposure across the recommendations.
3. **Serendipity:** This measure refers to the phenomenon of discovering valuable content that was not actively sought. To measure serendipity, we assess the performance of the inverse metric called 'Average Recommendation Popularity' (ARP). With the new concept of RS that strives for true personalization, we expect the recommended content to exhibit serendipitous qualities such as unexpected discoveries, content variety or serendipitous connections.

4. **Novelty:** We aim to measure the novelty of the content that users receive. Novelty is defined as the degree to which the recommended items differ from the user's previous interactions. By assessing novelty, we can determine whether the RS is capable of providing fresh and diverse content to users.
5. **Coverage:** Popularity biases in common RSs often lead to the reinforcement of popular content, resulting in a lack of personalized recommendations specially for those users who do not follow the trend. To address this, we use the 'Aggregated Diversity' (Agg-Div) metric, which measures the number of items from the entire media catalog that are being recommended. Higher values indicate a broader range of content being shown to users.
6. **Diversity:** We provide insights into the diversity of recommended items. However, it is important to note that high diversity may not always be desirable and depends on the context. In this study, when measuring diversity per user, we consider that the expected value of diversity should be high but without compromising other metrics. Striking a balance between diversity and personalized recommendations ensures that diversity is maximized within the user's preferred item set.

Second contribution: Stochastic Ranker

In addition to striving for truly personalized recommendations, another crucial aspect that adds significant value to the recommendation process is the ability to provide RRs. Achieving this goal requires striking a balance between accuracy and beyond accuracy metrics, taking into account the objectives of both users and producers. However, manually defining this trade-off can be complex and contentious. To address this challenge, we propose a principled method called *Stochastic Ranker* (SR). By leveraging the uncertainty in the predictions, our approach avoids relying on heuristic definitions of the trade-off between accuracy and beyond accuracy metrics. Instead, it consists on a stochastic ranking mechanism that is based on an objective measure: uncertainty prediction. This perspective allows us to improve exposure and coverage metrics by avoiding overconfidence in our predictions. In essence, we acknowledge that a more cautious approach over predictions can lead to better performance in terms of reaching a wider audience and providing broader catalog coverage.

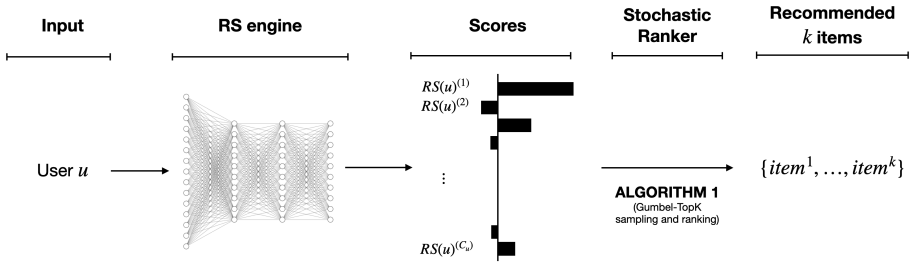

Fig. 1 The Gumbel-Topk algorithm can be employed to sample items directly from the implicit categorical distribution of a Recommender System (RS). This can be done by applying Algorithm 1 to the list of predicted scores from a user, thus generating its final list of recommended items.

Without loss of generality, we can say that the goal of any recommender is to predict an expected probability vector $\hat{\mathbf{y}}_u = \{y_u^{(1)}, \ldots, y_u^{(C_u)}\}$ for each user u, where C_u represents the number of candidate items for user u, and $y_u^{(c)}$ represents the predicted probability of user u consuming item c. This vector should represent our expectations regarding the utility of the recommendation by taking into account all stakeholders of the RS. The most suitable probability distribution to model this vector is a categorical distribution.

In practice, the probabilistic view of a recommender is relaxed and the output of a RS is usually a score vector $RS(u) = \{RS(u)^{(1)}, \ldots, RS(u)^{(C_u)}\}$ that represents the predicted relevance of items for a particular user. In most of the RSs, these scores are used to rank the items and determine the order in which a small subset of them is presented to the user.

Relying on the probabilistic view of a classifier can indeed be useful when considering the utility of recommendations in a RS: the confidence level associated with recommendations can be assessed in a sound way by leveraging this information during the ranking process. More specifically, we can measure the aleatoric uncertainty [47] of recommendations to build a system that takes decisions in a way that is coherent with the uncertainty level of the prediction.

So, with the aim of leveraging uncertainty and being able to build a principled method that aims to be trusted by all stakeholders, we propose to implement a stochastic ranking strategy which we call SR. By using this method, the RS is able to generate a list of k items by sampling k times, without replacement, from the probabilities $\hat{\mathbf{y}}_u$. The Gumbel-TopK trick, introduced by [48], provides a method to directly sample from the candidate item scores outputted by any RS model, including those that output non-normalized scores, in a way that is consistent with the distribution of $\hat{\mathbf{y}}_u$.

Let's consider a parametrization a categorical distribution in terms of an unconstrained vector of numbers that correspond to $RS(u) = \{RS(u)^{(1)}, \ldots, RS(u)^{(C_u)}\}$. In the case of a generic RS, $RS(u)$ are the scores of each item given a user u. The Gumbel-Topk trick works by perturbing the scores of all possible items, and then selecting the top-k of these perturbed probabilities as it can be shown in the Algorithm 1.

Algorithm 1 Stochastic Ranker using the Gumbel-TopK trick

1: Given a user u, obtain the set of scores $\{\text{RS}(u)^{(1)}, \ldots, \text{RS}(u)^{(C_u)}\}$
2: Compute $r_c = -\log(-\log(\varepsilon_c)) + \text{RS}(u)^{(c)}$, $\varepsilon_c \sim U(0,1)$, for each item candidate $c \in \{1, ..., C_u\}$
3: Return the k largest keys from $\{r_1, \ldots, r_{C_u}\}$.

Formally, we are given a set of C_u elements with weights $\{\text{RS}(u)^{(1)}, \ldots, \text{RS}(u)^{(C_u)}\}$ and we want to sample k elements, $K_u = \{i_1, \ldots, i_k\}$, without replacement. Given the total weight $W = \sum_{i=1}^{C_u} \text{RS}(u)^{(i)}$, the distribution for k-element subsets is given by:

$$P(K_u) = \frac{i_1}{W} \frac{i_2}{(W - \text{RS}(u)^{(1)})} \cdots \frac{i_k}{(W - \sum_{j=1}^{k-1} \text{RS}(u)^{(i_j)})}$$

It can be shown that by choosing the k largest Gumbel random variables r_c, we can sample subsets according to the sampling without replacement probability given by $P(K_u)$. Figure 1 illustrates how to apply this strategy in a RS pipeline that outputs non-normalized scores for a given user over a set of candidate items.

Third contribution: Ranking Interpretability

While assessing the performance of a RS using various metrics provides valuable insights into how recommendations are generated, directly interpreting how the model constructs ranking lists for different users can enhance trust and comprehension. Visualizing the ranking process enables a deeper understanding of how RSs generate recommendations and whether they prioritize popular items for all users or exhibit personalized behaviors based on individual consumption patterns.

To address this, we propose visualizing the ranking construction for different user clusters, determined by their consumption behavior. By examining how the models recommend items based on popularity, we can gain insights into how they are trained and whether they exhibit diversity in their recommendations. By combining comprehensive metric assessments with visualizations of the ranking process, we can enhance our understanding of RSs and their recommendation generation mechanisms. This approach fosters trust and enables us to make more informed judgments about the performance and behavior of RS models.

3.1 Datasets

Here, we briefly describe the three datasets we have analyzed in the experiments. We transformed all datasets to work with implicit feedback and, further, applied some pre-processing to them, as discussed below. In Table 1, we present a comprehensive summary of the statistical information for the final versions of the datasets, including a new column called 'weight Top_{10} items', which indicates the percentage of interactions accounted by the ten most popular items in each dataset. Furthermore, Figure 2 offers a visual depiction of the distribution of users and items according to their respective number of interactions.

	#users	#items	#interactions	sparsity	weight Top_{10}items
ML-1M	6,040	3,062	999,611	95.16%	4.084%
Netflix	18,503	8,898	78,071	99.95%	6.58%
Pinterest	55,187	9,916	1,445,622	99.73%	0.82%
PSM	14,658	552	83,082	97.38%	47.41%

Table 1 Dataset statistics in terms of number of users, number of items, number of interactions among them and sparsity of the rating matrix. The last column indicates the percentage of interactions accounted by the ten most popular items. Statistics are reported after applying all data transformations needed.

Movielens 1M

To facilitate the reproducibility of our results, we have applied our RSs models to the publicly available MovieLens 1M dataset[1]. It is composed of approximately 1M ratings from about 6,000 users on 3,000 movies. The ratings are given to us in the form of *<userID, itemID, rating, timestamp>* tuples and each user has a minimum of twenty ratings. In order to treat all datasets the same way, we cut the item frequency to be at least five, as we did with the data from the public broadcaster.

Netflix

The Netflix Prize dataset [49] was made publicly available on October 2006, the same day that the Netflix Prize competition was launched. The dataset consists of over 100 million movie ratings from more than 480,000 Netflix subscribers and, when released, was one of the largest and most complex datasets of its kind. As part of this release, the organizers also released a subset of data, the Netflix Probe subset, that comprises 6 years of data (2000 - 2005). The dataset contains 100,000 ratings from approximately 19,000 users for 9,000 movies. The challenge lies in capturing users' interests, given that there are, at most, nine interactions per user. This is what makes the dataset interesting, as it does not suffer from a significant imbalance in terms of user representation.

Pinterest

The Pinterest dataset is one of the largest social curation networks and it was released in 2015 by Geng et al. [50]. Its source data is very particular for being content-centric network and it is composed of approximately 1.5M ratings from about 55,000 users on 10,000 images. The dataset already provides a train-test split which has been achieved by following *leave-one-out* strategy.

Public Service Media

We obtained an anonymized dataset corresponding to historical data of user views of the online catalogue of a PSM, specifically a TV broadcaster named TV3, collected throughout a whole calendar year (2021)[2]. The raw data contained information on user interactions indicating *userID*, *itemID* and some contextual information of the

[1] https://grouplens.org/datasets/movielens
[2] PSM dataset is obtained from https://zenodo.org/record/7940658

interaction such as a timestamp. Items were identified at the single episode level (e.g., morning and evening news had separate IDs, as had each episode of TV series) but, as it would not be helpful for the RS to have to rank different episodes of the same TV show, we aggregated all episodes from the same program into the same *itemID*. This resulted in a large reduction in the number of items, composed of approximately 80,000 ratings from about 14,000 users on 552 different contents. Besides, we applied filters that required a minimum of five interactions per user and also a minimum frequency of five visualizations per item, with the aim of removing outliers and work with more stable data. After filtering, we adapted the data to build a dataset which consists of <*userID, itemID, rating, timestamp*> tuple interactions.

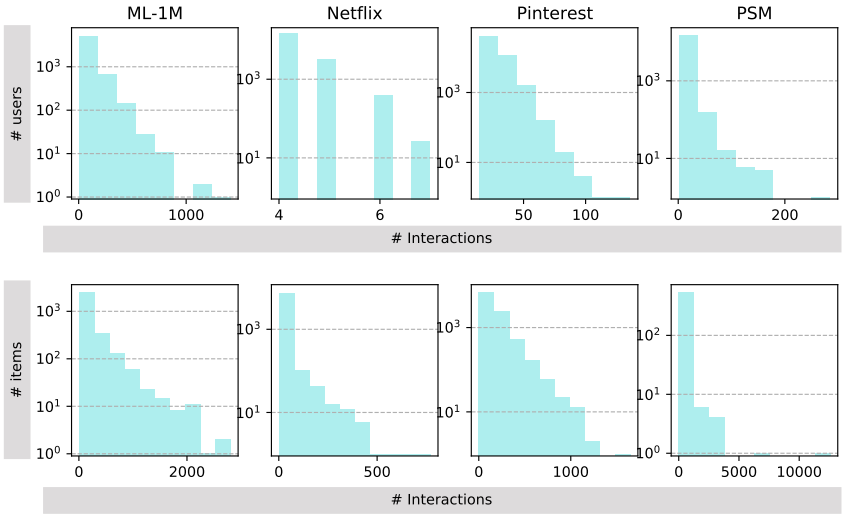

Fig. 2 Distribution of number of users (upper row) or items (lower row) given number of interactions, in log scale, for the four datasets under consideration: ML-1M, Netflix, Pinterest and PSM.

4 Experiments

In this section, our aim is to showcase the effectiveness of our proposed approach. We begin by presenting a preliminary analysis to gain insights into the influence of popularity bias on users and items respectively. Subsequently, we discuss the experimental setup and the metrics employed to evaluate our system. Finally, we present the results in both quantitative and qualitative formats, illustrating the successful attainment of RRs. Throughout this study, we seek answers to the following research questions (RQs):

RQ1. How does the quantitative assessment of a RS change, including accuracy and beyond accuracy metrics, when it is trained to optimize a new policy which is strongly focused on achieving genuine personalization?
RQ2. Can we construct RRs in a principled manner by incorporating uncertainty into the recommendation process?
RQ3. Do the outcomes of our proposed methodology maintain consistency across different datasets that exhibit varying degrees of popularity bias, and if not, can we establish the circumstances under which our method is appropriate for application?

4.1 Preliminary analysis

Prior to commencing the experiments, it is essential to conduct a comprehensive analysis of the popularity biases across the various datasets. Thus, in this section, we present an extensive examination of different clusters, considering both the user and item perspectives. For this purpose, we have classified users and items into three distinct clusters, utilizing the methodology proposed by Borges and Stefanidis [51] for user clustering, and the one proposed by Abdollahpouri et al. [45] for item clustering.

Item clusters

In order to analyze popularity bias from items perspective, they are grouped based on the percentage of interactions received across the entire consumption. Items with over 40% of the interactions fall into the **Blockbuster** cluster, those with between 20% and 40% are categorized into the **Diverse** cluster, while those with less than 20% are placed in the **Niche** cluster. This categorization allows to examine how different models distribute items from each cluster on the final ranking list presented to the user.

User clusters

In order to analyze popularity bias from users perspective, they are characterized based on the percentage of interactions made with items from the previously defined clusters (Blockbuster, Diverse, and Niche items). We then sort users according to their popularity distribution, with those consuming more popular items in the **Short Head** cluster, those with between 20% and 60% of interactions from popular items in the **Medium Tail** cluster, and those with less than 20% of interactions from popular items in the **Long Tail** cluster.

In Table 2, we present the number of users and items belonging to each cluster, along with their percentage in the entire target and catalog, respectively. We want to highlight that the statistics for the Public Service Media (PSM) dataset are particularly interesting, as popularity bias poses a major challenge for PSM, and specifically TV broadcasters. This popularity bias observed can be attributed to multiple factors. It is important to highlight that a crucial factor contributing to this bias is the utilization of a popularity-based model during the production stage. Furthermore, the ground-truth data used for evaluating the model might result in high accuracy during offline evaluation, but it may not necessarily correspond to similar outcomes during online evaluation.

	User Bias Distribution			Item Bias Distribution		
	Short Head	Medium Tail	Long tail	Blockbuster	Diverse	Niche
ML-1M	199 (3%)	3,472 (58%)	2,356 (39%)	217 (7%)	681 (22%)	2164 (71%)
Netflix	520 (3%)	6,071 (32%)	11,912 (65%)	165 (2%)	1321 (15%)	7412 (83%)
Pinterest	159 (1%)	12,511 (22%)	42,517 (77%)	1476 (15%)	3350 (34%)	5090 (51%)
PSM	974 (7%)	3,852 (26%)	9,832 (67%)	6 (1%)	42 (8%)	504 (91%)

Table 2 We provide statistical information about the number of users and items in each cluster, along with the percentage they represent in relation to the total number of users and items, respectively. The left side of the table shows the number of users in each cluster along with the percentage they represent in brackets. On the right side, we present the same information for items.

4.2 Experimental setup

In this section, we provide an overview of different RSs used for our comparative analysis of the impact of RRs on four distinct datasets. Additionally, we describe the training and evaluation procedures to ensure the reproducibility of the results.

Baselines

To comprehensively evaluate the performance of RSs, we include two non-personalized models, namely the `Random model` and `MostPop model`, as reference points for comparison and easing the assessment of various metrics. These models serve as baselines for evaluating the effectiveness of different tasks, including the conventional 'next-item prediction' task, and the novel task proposed by Pellegrini et al. [13] that emphasizes true personalization and RRs.

Despite the recent focus on sequential recommendations, Matrix Factorization (MF) models have consistently demonstrated their efficacy and versatility in the field of RSs [52]. Factorization Machines (FM) [46], an extension of the MF algorithm, have gained recognition as a valuable option for building accurate and personalized recommendations. Their capability to incorporate contextual information and handle diverse data structures [40, 53, 54] makes them a viable choice in the current landscape [13, 23]. Therefore, FM has been selected as the baseline model due to its alignment with the constraints discussed in section 3. It is important to note that changing the model will not substantially affect the observed behavior because the significance of our contributions lies not in the specific model choice, but in the training methodology and the presentation of items to users. Additionally, in order to provide concise references in Table 3 showcasing the quantitative results, we introduce the following acronyms:

- *Random*: this model does not take into consideration the distribution of the data. It makes recommendations for items in a uniform manner.
- *MostPop*: it refers to the Most Popular items recommender, which consistently recommends the k most frequently consumed items.
- *FM*: it refers to the FM model architecture trained for the typical recommendation task, 'next-item prediction'. In our scenario, this means that $y_u^{(c)}$ has been trained to predict the next item given the observed preferences of a user.

- *FM-PR*: it refers to the FM model architecture trained for Prob-Ratio (PR) optimization defined in [13], which aims to predict which item a user would choose among popularity sampled items. This approach can be understood as a way to estimate the genuine interest of the users.
- *FM-SR*: it improves the abovementioned FM-PR model by leveraging the uncertainty in the predictions using the SR introduced in this work.

Training and Evaluation

To ensure a fair comparison of the models' performance, we train all of them by using Adam optimizer and the Binary Cross Entropy (BCE) loss, as for FM family models it has been shown the advantage of using point-wise losses over pair-wise ones [40]. In fact, a crucial reason to use point-wise loss function is that we can pair one positive example with many negative ones, and thus we can flexibly control the sampling ratio of negative examples during training. The optimal sampling ratio being found to be between 3 and 6, we have used a value of 4 for our experiments.

The evaluation of all models follows the standard offline top-k evaluation, where the target is to generate a ranking list (according to the predicted scores) of k items that a user is most likely to interact with. We use the *leave-one-out* strategy, which has been widely adopted in literature [40, 54–56], for splitting the dataset to train, validation and test sets. Indeed, we want to highlight the importance of tuning on a validation set and then report the metrics on the test set, as conclusions on validation results cannot be directly trusted because they do not guarantee that the model is generalizing well. This is not always stated in several papers when it comes to *leave-one-out* split, which can lead to confusion when reproducing the results. We run all the experiments for a maximum of 100 epochs and perform early stopping when the HR accuracy metric stops improving for more than 10 consecutive epochs.

4.3 Metrics

Recent research has made a strong turn toward RRs, and several new metrics have been proposed which go beyond accuracy [6, 45, 57, 58]. Although several discussions regarding the best way of performing sampling on evaluation have arisen, there have been claims regarding the importance and robustness of performing full-ranking assessment [42–44]. Moreover, full-ranking strategies more closely resemble production scenarios for recommendation tasks, and there are some RSs behaviors that cannot be assessed without following this methodology, as it is the case of measuring the percentage of items recommended across the entire catalog, commonly known as coverage or aggregate-diversity [45, 59].

Even though fairness measures are becoming more and more important nowadays, there is still a need for RSs to achieve certain level of accuracy. In their study [60], the authors compare different assessment methods to measure the similarity between users' predicted preferences and true preferences. They find that top-k based recommendation tasks, using a ranking assessment, outperform error loss methods in terms of realism. Therefore, we adopt ranking performance evaluation metrics such as HR and nDCG to assess the performance of deep learning recommendation models. This

aligns with the importance emphasized by [56] of simulating an evaluation scenario that resembles a real production stage in the field of RSs.

In this work, our objective is to thoroughly analyze RSs from a responsible perspective. To achieve this, we introduce five fairness metrics that serve as assessment tools for evaluating recommendations in a comprehensive and responsible manner.

Metrics description

We define the training dataset $\mathcal{D}$ as a set of user-item interactions $d_j,\ j \in \{1, \ldots, N\}$. Within, let L_u be the recommended set of items for user $u \in U$ (note that L_u is a list truncated at any desired k, where k is the number of items that can be displayed to a given user). Let L be the combined list of all recommendation lists or rankings given to all users, being $L = |\cup_{u \in U} L_u|$. Let I be the set of all items in the catalog and U be the set of all users. We define **the merit of an item** $M_{\mathcal{S}}(i)$ as the number of times that an item i was selected by a user divided by the total number of interactions in a set $\mathcal{S}$, and we claim that **merit** is a good way to standardize what in many cases is differently defined as *item popularity*.

- Average Recommendation Popularity (ARP): This measure defined by Yin et al. [61] and further used in many works [45, 62], aims to calculate the average popularity of the recommended items in each list, also averaged across all users. The opposite of this metric is Serendipity, as lower scores of ARP denote high serendipity [63].

$$\text{ARP} = \frac{1}{|U|} \sum_{u \in U} \frac{\sum_{i \in L_u} M_{\mathcal{D}}(i)}{|L_u|} \tag{1}$$

- Aggregate-Diversity (Agg-Div): This measure indicates the ratio of unique recommended items along the catalogue averaged for all users.

$$\text{Agg-Div} = \frac{|\{L\}|}{|I|} \tag{2}$$

- Error of Exposure (EoE): It is the difference between the expected and actual number of times an item is recommended to a user. Measuring EoE requires taking into account whether the original merit distribution of items is proportionally reflected in the recommendations or not. Over-recommendation of popular items should be prevented in order to ensure fairness, and maintaining the original merit distribution from a dataset $\mathcal{D}$, in the recommendation set of items $\mathcal{R}$, would actually be the ideal outcome. The EoE equation can be defined as following:

$$\text{EoE} = \sum_{i \in I} |M_{\mathcal{D}}(i) - M_{\mathcal{R}}(i)| \tag{3}$$

- Novelty (Nov): This measure defined in [57] reflects how likely the user has been exposed to the item based on the population historical engagement and it is defined as:

$$\text{Nov} = 1 - \sum_{i \in I} \frac{|M_{\mathcal{R}}(i)|}{C_i}, \tag{4}$$

where C_i is the number of users who did not interact with the item i during training stage.

- Diversity (Div): It measures the breadth of recommended items, ranging from narrow to wide. From the perspective of popularity bias, it is not clear whether high diversity is always desirable. While it is desirable to recommend a diverse range of music across multiple artists, high diversity could also result in recommending both Blockbuster and Niche items for a user in the Long Tail, which may not reflect responsible and fair recommendations. To calculate diversity, we use the *Cosine Similarity* measure among the top-k recommended items, then sum all distances and take the average for all users.

4.4 Results

To offer a thorough evaluation of the various methodologies and recommendation tasks across all four datasets, we present both quantitative and qualitative analyses of the outcomes.

Quantitative evaluation

Quantitative evaluation of the different methodologies and recommendation tasks can be observed in Table 3, where the best results have been highlighted in bold. It is important to note that the Random and MostPop models have not been highlighted as they are included only for comparison purposes. As evaluating whether a RS is responsible and fair requires consideration of multiple metrics simultaneously [1, 45, 62], we will provide an overview of each model's performance in the following sections, emphasizing the strengths and weaknesses of each approach.

When we examine the three benchmark datasets (ML-1M, Netflix, and Pinterest), it can be observed that training a RS for the new task that drives toward true personalization, consistently outperforms other models across all the metrics. More specifically, observing the results from FM-PR and FM-SR models, which are actually optimized for this new task, it can be seen that they consistently outperform in terms of accuracy metrics, hence indicating a better fit with user preferences and resulting in better value for media companies. Notably, improvements are observed in the ARP and Agg-Div metrics, indicating that a larger number of unique items are being presented to users in their recommendations. At the same time, a low result on the EoE metric demonstrates a better fit between the popularity item distribution from the training set and the recommendation set. This guarantees that the tastes of each user - particularly in terms of Blockbuster, Diverse, and Niche items - are respected, hence increasing novelty and pushing for new, original, and unusual recommendations.

Conversely, when examining the collected PSM dataset, there is a significant drop in accuracy metrics, even notable improvements are observed in beyond accuracy metrics. The statistics from Table 2 reveal a high popularity bias, as the data interactions were originally gathered by a MostPop model at the production stage. In fact, the reason why the MostPop model performs very well for this dataset, is because the ground-truth items from the test set belong exclusively to the popular bins (Blockbuster and Diverse items).

		Accuracy			Beyond accuracy			
		HR ↑	nDCG ↑	ARP ↓	Agg-Div ↑	EoE ↓	Novelty ↑	Diversity
ML - 1M	**Random**	0.0032	0.0017	0.0003	1.0000	1.0559	0.9966	-
	MostPop	0.0405	0.0192	0.0038	0.0376	1.7227	0.8645	-
	FM	0.0639	0.0318	0.0031	0.1750	1.4482	0.9732	0.4721
	FM-PR	**0.0692**	**0.0349**	0.0015	0.3217	0.8811	0.9879	0.2024
	FM-SR	0.0553	0.0259	**0.0013**	**0.5013**	**0.6802**	**0.9925**	0.2409
Netflix	**Random**	0.0017	0.0008	0.0001	1.0000	1.2727	0.9989	-
	MostPop	0.0223	0.0110	0.0065	0.0018	1.8466	0.3573	-
	FM	0.0652	0.0328	0.0045	**0.0652**	1.4461	**0.9824**	0.4710
	FM-PR	0.0699	0.0341	**0.0030**	0.0373	1.1171	0.9695	0.1842
	FM-SR	**0.0755**	**0.0369**	**0.0030**	0.0561	**1.0939**	0.9792	0.1513
Pinterest	**Random**	0.0020	0.0009	0.0001	1.0000	0.6588	0.9990	-
	MostPop	0.0080	0.0039	0.0008	0.0016	1.9792	0.3618	-
	FM	0.0105	0.0050	0.0005	0.0080	1.9447	0.8717	0.1756
	FM-PR	0.0102	0.0051	0.0006	0.0141	1.9383	0.9275	0.2082
	FM-SR	**0.0117**	**0.0055**	**0.0004**	**0.0604**	**1.7542**	**0.9832**	0.0924
PSM	**Random**	0.0187	0.0084	0.0017	1.0000	1.4970	0.9817	-
	MostPop	0.5514	0.2863	0.0407	0.0725	1.0520	0.6886	-
	FM	**0.5826**	**0.3229**	0.0376	0.1246	1.0117	0.8999	0.8498
	FM-PR	0.2481	0.1754	0.0103	0.6014	0.8169	0.9679	0.1834
	FM-SR	0.2164	0.1569	**0.0088**	**0.9384**	**0.8150**	**0.9796**	0.2219

Table 3 We present the quantitative results for five different models across four datasets, assessed using seven metrics. The evaluation is performed at $k = 10$ for all metrics, and the Stochastic Ranker (SR) is applied to the initial 100 item predictions. Best results are highlighted in the table; however, the *Random* and *MostPop* models, included for comparison purposes, are never emphasized in the results.

Several offline experiments were conducted to analyze the effect of removing the first $t \in [10, 15, 20]$ popular items from the dataset. The results demonstrated a more stable behavior, similar to benchmark datasets, providing promising indications for the FM-PR and FM-SR models. However, due to the preprocessing explained in section 3.1, removing these popular items resulted in a significantly reduced number of interactions, as approximately 50% rely on the top ten items as shown in Table 1. Consequently, conducting trustworthy experiments with such limited data points becomes challenging. Nonetheless, considering all metrics together, the results of the FM-PR and FM-SR models show promise, suggesting that deploying these models in production could facilitate less biased data collection and ultimately improve accuracy metrics in subsequent training iterations.

Lastly, it is important to examine the Diversity metric on its own. As mentioned earlier in the description of metrics (see section 4.3), there is a lack of consensus regarding whether diversity should be high or low in RSs. However, it is important to note that achieving accurate item recommendations should be the primary focus before

aiming for high diversity when evaluating all metrics simultaneously. In the case of FM models, where diversity is often very high, it may not necessarily be a good thing, as this diversity is achieved by recommending Blockbuster items to Long Tail users. This leads to higher diversity, but not for the right reasons. What we ultimately want is to achieve the highest possible diversity within the recommendations that a user expects to have. Therefore, there is no point in recommending Blockbuster items to all users, as it would increase diversity but not necessarily accuracy or user satisfaction. Hence, we claim that diversity should always be analyzed within its context, taking into account other metrics that can help us evaluate diversity results. It is important to note that diversity should not be measured solely from a popularity bias perspective, but also from a thematic perspective, among others.

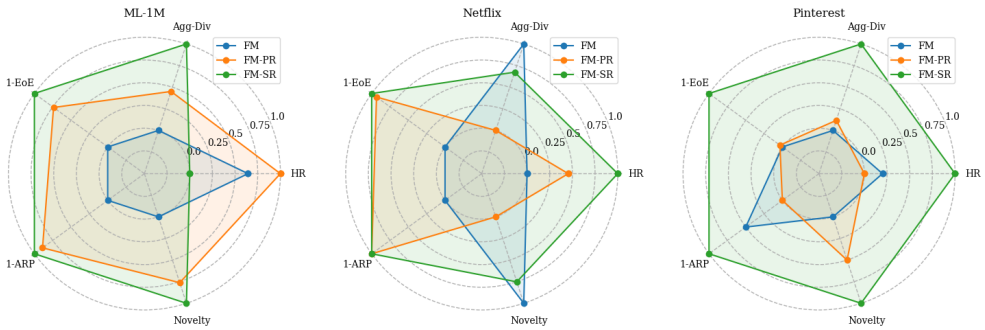

Fig. 3 Radar charts are employed to showcase the evaluation metrics across different RS models on three distinct datasets. A larger shaded area in the chart indicates a superior trade-off between the metrics achieved by a specific method. The results consistently indicate that the FM-SR model outperforms the other models.

In order to provide a visual comparison of the performance of FM, FM-PR, and FM-SR across three datasets (ML-1M, Netflix, and Pinterest), we have created a radar chart presented in Figure 3. To improve clarity, we have omitted the Random and MostPop models. We have also discussed the PSM dataset separately, which has an extreme popularity bias that is commonly seen in TV broadcasters. Each chart presents five axes that exhibit the normalized values of five metrics: Agg-Div, HR, Novelty, ARP, and EoE. The overall performance of a specific method across all these metrics is determined by the area of the shape, with a larger area indicating better performance. Before plotting the metrics, they all have been normalized using a min-max scaler to bring them within the range of 0-1. It is worth noting that EoE and ARP are represented as 1-EoE and 1-ARP on the chart to facilitate comparison with the other metrics, ensuring that higher values on all the metrics correspond to better performance. The results show that FM-SR outperforms the other methods on all three datasets, as demonstrated by its larger area on the radar chart. However, there is a drop in the HR metric for the ML-1M dataset, which we believe is due to the nature of the data, particularly on the test set. Unlike Netflix and Pinterest datasets, which collect genuine user preferences (Netflix has between 4 and 15 interactions per

user, and Pinterest is a very item-focused dataset where users just pin very similar items), ML-1M dataset has a less focused user profiling, and thus avoiding popular items may not lead to good HR metrics, even if it would provide more interesting recommendations. Furthermore, it should be noted that the HR numerical drop for ML-1M (see Table 3) is not as significant as it seems in the visualization.

Qualitative evaluation

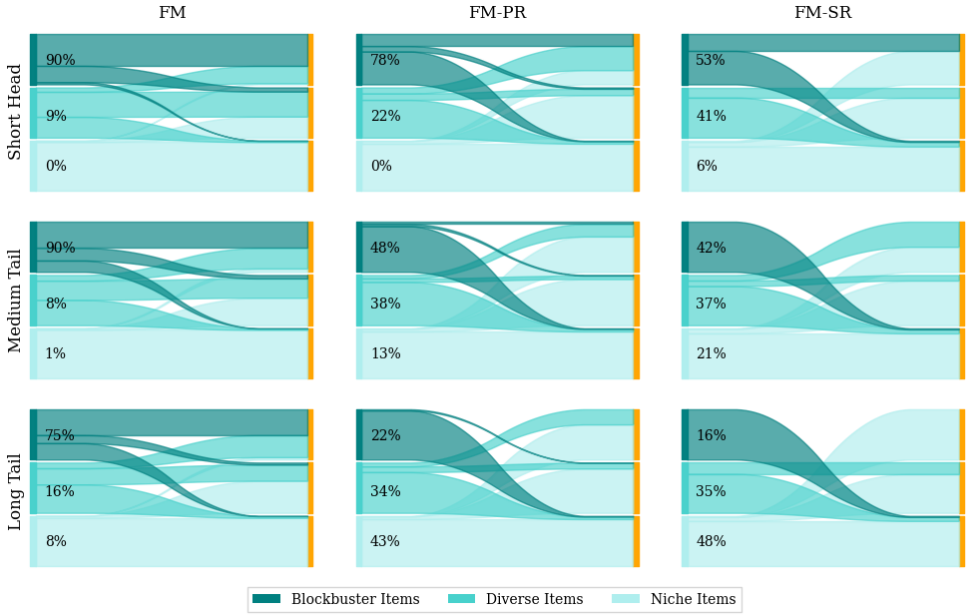

Fig. 4 This visual representation showcases, for the ML-1M dataset, the distribution of items across the ranking produced for different user clusters (Short Head, Medium Tail, and Long Tail). The aim is to provide deeper insights on how each method positions items across the ranking list showed to a user. Each row corresponds to a randomly selected user within a cluster, and each column represents a RS model (FM, FM-PR, FM-SR). Within each chart, items posed by a particular model are displayed in different colors to better visualize their respective clusters. The left axis illustrates the item distribution of the MostPop model, which serves as the ground-truth, with Blockbuster items at the top, Diverse items in the middle, and Niche items at the tail. The right axis show how each RS model impacts this item distribution, providing insights into how the model positions items in the final ranking list. Numerical values in each chart indicate the percentage of items from each cluster that contribute to the 100% HR metric achieved by each model in Table 3.

In Table 3, we present a numerical comparison of the results to better understand the behavior of accuracy and beyond-accuracy metrics for each RS. It can be observed that the FM-PR model leads to a nice compromise between media service value and user engagement, considering there should be a trade-off between accuracy and beyond-accuracy metrics. In this section, we aim to complement these results by analyzing the differences between the 'popularity sampled' objective loss defined in [13] - which

corresponds to FM-PR model - and the one aimed at optimizing next-item predictions from the perspective of the item's ranking distribution - which corresponds to FM model. Additionally, we further investigate the results of applying the SR method defined in section 3 hence demonstrating how we can leverage the inherent uncertainty in model predictions to promote RRs.

Figure 4 provides valuable insights into how the distribution of items changes in a full ranking. By examining the recommendation lists of three randomly selected users from distinct clusters, we can enhance our comprehension of how each trained model (FM, FM-PR, and FM-SR) constructs rankings in comparison to the MostPop model, which serves as the benchmark on the left axis. Additionally, we have provided numerical values indicating the percentage of items from each cluster that have contributed to the HR metrics.

The FM model, which is trained to predict the next item, tends to prioritize Blockbuster items due to the popularity bias amplification effect that it suffers. Hence, its recommendations are similar for all types of users because it has been optimized to predict what is most likely to be clicked. We observe that the popularity distribution of the ranking is similar to the MostPop model distribution.

In contrast, the FM-PR model is optimized for a more responsible task: discovering the user's preferences among distinct popular items, which aims to guide toward true personalization. As a result, it selects items that are favored by the user, reducing the impact of popularity bias effect. In fact, this is particularly noticeable for Medium and Long Tail users, for whom the model prioritizes similar items to those originally consumed by the user, resulting in recommendations that emphasize items from Diverse and Niche clusters on the top positions of the ranking. Remarkably, the top ten recommended items for the Long Tail user do not include any Blockbuster items, reflecting the items actually shown to the user.

Lastly, FM-SR model - which is the FM-PR model with the SR applied - results in a significant impact on the ranking. By leveraging the uncertainty in the predictions the model is able to prioritize the most preferred items by each user type. For instance, it selects Blockbuster items for the Short Head user, Diverse items for the Medium Tail user, and Niche items for the Long Tail user within the top positions. By adopting a SR in a task that optimizes for true personalization, the method moves toward responsible and equitable recommendations without compromising the value for media companies.

Lastly, it is worth noting that the trends observed in the ML-1M dataset shown in Figure 4, are consistent with the results observed in the Netflix and Pinterest datasets, as evident from the analysis of Figure 3.

5 Conclusions

The field of AI is witnessing a growing significance of ethical concerns on a daily basis, posing new challenges. In particular, the research in recommendations is shifting toward building RR. However, this effort has given rise to numerous challenges when attempting to build RSs that are fair, equitable, and accurate. The primary concern lies in the ability to measure the effectiveness of a RS from multiple perspectives

simultaneously. This necessitates addressing two key requirements: 1) the clear definition of a set of metrics that go beyond accuracy and provide practical insights into the behavior of RS, and 2) achieving a suitable trade-off between accuracy and beyond accuracy metrics, ensuring that RSs are not only fair but also able to maintain high accuracy and offer a compromise that adds value to both consumers and providers. Moreover, in order to further advance toward RRs, it is crucial to gain better understanding of RSs and uncover how do they strike this balance. To this end, the use of explainability techniques and visually appealing representations become imperative and should be given greater emphasis when building new RSs approaches.

In this research, significant breakthroughs have been achieved in three crucial areas. Firstly, we have proposed a comprehensive set of metrics to evaluate RRs, which provide a thorough assessment of fairness, accuracy, diversity, and other important aspects of RSs performance. Additionally, a novel metric called *Error of Exposure* has been introduced to measure the disparity between the distribution of original user consumption and the distribution of the recommended content. This innovative approach offers a holistic perspective on its effectiveness and aligns with the evolving demands of the digital landscape. Secondly, our study embraces a recently proposed methodology that modifies the optimization function of a RS aiming to move toward true personalization. In addition, we introduce a novel approach called the *Stochastic Ranker*, which addresses the challenge of balancing accuracy and beyond accuracy metrics in a principled manner. Through the effective utilization of the inherent uncertainty in the model's predictions, our approach achieves significant improvements in fairness, diversity, and novelty while maintaining the overall effectiveness of the system. An important fact is that this is achieved without the need for heuristic trade-offs or compromising the system's performance. Lastly, we enhance the understanding of the RS optimization process and its influence on ranking construction by incorporating visual elements. Through visually engaging graphics, we offer clear insights into how these RSs generate rankings for diverse user clusters. These visualizations facilitate a deeper comprehension of the improvements made by the models and gain from each specific metrics that contribute to the overall enhancement of a RS, hence leading to RRs.

In terms of future work, we suggest replicating our analysis and experiments using different deep models, such as graph convolutional network models and sequential models, which are also prominent in recommendation research. However, we believe that our findings should hold true for other recommendation algorithms, as they are not specific to a particular RS model but rather to the task the model is optimized for. Additionally, it is important to acknowledge that all methods rely on the training data used. While efforts are being made to improve data collection processes for fairer and less biased datasets, we argue that implementing responsible RS in the online domain can lead to the accumulation of superior datasets and overall enhance the quality of RSs, thereby mitigating undesirable effects.

References

[1] Elahi, M., Jannach, D., Skjærven, L., Knudsen, E., Sjøvaag, H., Tolonen, K., Holmstad, Ø., Pipkin, I., Throndsen, E., Stenbom, A., et al.: Towards responsible

media recommendation. AI and Ethics, 1–12 (2022)

[2] Schwär, H.: How instagram and facebook are intentionally designed to mimic addictive painkillers. Business Insider. Retrieved October **25**, 2021 (2021)

[3] Zakon, A.: Optimized for addiction: Extending product liability concepts to defectively designed social media algorithms and overcoming the communications decency act. Wis. L. REv., 1107 (2020)

[4] Sjøvaag, H., Krumsvik, A.H.: In search of journalism funding: scenarios for future media policy in norway. Journalism practice **12**(9), 1201–1219 (2018)

[5] Wirtz, B.W., Weyerer, J.C., Geyer, C.: Artificial intelligence and the public sector—applications and challenges. International Journal of Public Administration **42**(7), 596–615 (2019)

[6] Macgregor, M.: Responsible ai at the bbc: Our machine learning engine principles. BBC R&D (2021)

[7] Liu, R., Gupta, S., Patel, P.: The application of the principles of responsible ai on social media marketing for digital health. Information Systems Frontiers, 1–25 (2021)

[8] Abdollahpouri, H., Burke, R., Mobasher, B.: Controlling popularity bias in learning-to-rank recommendation. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, pp. 42–46 (2017)

[9] Boratto, L., Fenu, G., Marras, M.: Connecting user and item perspectives in popularity debiasing for collaborative recommendation. Information Processing & Management **58**(1), 102387 (2021)

[10] Elahi, M., Kholgh, D.K., Kiarostami, M.S., Saghari, S., Rad, S.P., Tkalčič, M.: Investigating the impact of recommender systems on user-based and item-based popularity bias. Information Processing & Management **58**(5), 102655 (2021)

[11] Jannach, D., Lerche, L., Kamehkhosh, I., Jugovac, M.: What recommenders recommend: an analysis of recommendation biases and possible countermeasures. User Modeling and User-Adapted Interaction **25**, 427–491 (2015)

[12] Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., Burke, R.: Feedback loop and bias amplification in recommender systems. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 2145–2148 (2020)

[13] Pellegrini, R., Zhao, W., Murray, I.: Don't recommend the obvious: estimate probability ratios. In: Proceedings of the 16th ACM Conference on Recommender Systems, pp. 188–197 (2022)

[14] Jannach, D., Jugovac, M.: Measuring the business value of recommender systems. ACM Transactions on Management Information Systems (TMIS) **10**(4), 1–23 (2019)

[15] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., He, X.: Bias and debias in recommender system: A survey and future directions. ACM Transactions on Information Systems **41**(3), 1–39 (2023)

[16] Chen, J., Wang, C., Zhou, S., Shi, Q., Feng, Y., Chen, C.: Samwalker: Social recommendation with informative sampling strategy. In: The World Wide Web Conference, pp. 228–239 (2019)

[17] Liu, D., Cheng, P., Dong, Z., He, X., Pan, W., Ming, Z.: A general knowledge distillation framework for counterfactual recommendation via uniform data. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 831–840 (2020)

[18] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Acm Sigir Forum, vol. 51, pp. 4–11 (2017). Acm New York, NY, USA

[19] Steck, H.: Evaluation of recommendations: rating-prediction and ranking. In: Proceedings of the 7th ACM Conference on Recommender Systems, pp. 213–220 (2013)

[20] Hernández-Lobato, J.M., Houlsby, N., Ghahramani, Z.: Probabilistic matrix factorization with non-random missing data. In: International Conference on Machine Learning, pp. 1512–1520 (2014). PMLR

[21] Liu, Y., Cao, X., Yu, Y.: Are you influenced by others when rating? improve rating prediction by conformity modeling. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 269–272 (2016)

[22] Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.: The unfairness of popularity bias in recommendation. In: RecSys Workshop on Recommendation in Multistakeholder Environments (RMSE) (2019)

[23] Zhang, Y., Feng, F., He, X., Wei, T., Song, C., Ling, G., Zhang, Y.: Causal intervention for leveraging popularity bias in recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 11–20 (2021)

[24] Yalcin, E., Bilge, A.: Investigating and counteracting popularity bias in group recommendations. Information Processing & Management **58**(5), 102608 (2021)

[25] Bedi, P., Gautam, A., Sharma, C., *et al.*: Using novelty score of unseen items to

handle popularity bias in recommender systems. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I), pp. 934–939 (2014). IEEE

[26] Lika, B., Kolomvatsos, K., Hadjiefthymiades, S.: Facing the cold start problem in recommender systems. Expert systems with applications **41**(4), 2065–2073 (2014)

[27] Chen, J., Dong, H., Qiu, Y., He, X., Xin, X., Chen, L., Lin, G., Yang, K.: Autodebias: Learning to debias for recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21–30 (2021)

[28] Saito, Y., Yaginuma, S., Nishino, Y., Sakata, H., Nakata, K.: Unbiased recommender learning from missing-not-at-random implicit feedback. In: Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 501–509 (2020)

[29] Lin, Z., Liu, D., Pan, W., Ming, Z.: Transfer learning in collaborative recommendation for bias reduction. In: Proceedings of the 15th ACM Conference on Recommender Systems, pp. 736–740 (2021)

[30] Zhu, Z., He, Y., Zhao, X., Zhang, Y., Wang, J., Caverlee, J.: Popularity-opportunity bias in collaborative filtering. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 85–93 (2021)

[31] Asudeh, A., Jagadish, H., Stoyanovich, J., Das, G.: Designing fair ranking schemes. In: Proceedings of the 2019 International Conference on Management of Data, pp. 1259–1276 (2019)

[32] Buyl, M., De Bie, T.: Debayes: a bayesian method for debiasing network embeddings. In: International Conference on Machine Learning, pp. 1220–1229 (2020). PMLR

[33] Li, Y., Chen, H., Fu, Z., Ge, Y., Zhang, Y.: User-oriented fairness in recommendation. In: Proceedings of the Web Conference 2021, pp. 624–632 (2021)

[34] Bonner, S., Vasile, F.: Causal embeddings for recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 104–112 (2018)

[35] McInerney, J., Brost, B., Chandar, P., Mehrotra, R., Carterette, B.: Counterfactual evaluation of slate recommendations with sequential reward interactions. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1779–1788 (2020)

[36] Yalcin, E.: Exploring potential biases towards blockbuster items in ranking-based recommendations. Data Mining and Knowledge Discovery, 1–41 (2022)

[37] Liang, D., Krishnan, R.G., Hoffman, M.D., Jebara, T.: Variational autoencoders for collaborative filtering. In: Proceedings of the 2018 World Wide Web Conference, pp. 689–698 (2018)

[38] Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 426–434 (2008)

[39] Elkahky, A.M., Song, Y., He, X.: A multi-view deep learning approach for cross domain user modeling in recommendation systems. In: Proceedings of the 24th International Conference on World Wide Web, pp. 278–288 (2015)

[40] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.-S.: Neural collaborative filtering. In: Proceedings of the 26th International Conference on World Wide Web, pp. 173–182 (2017)

[41] Kang, W.-C., McAuley, J.: Self-attentive sequential recommendation. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 197–206 (2018). IEEE

[42] Bera, S.K., Seshadhri, C.: How to count triangles, without seeing the whole graph. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 306–316 (2020)

[43] Li, D., Jin, R., Gao, J., Liu, Z.: On sampling top-k recommendation evaluation. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2114–2124 (2020)

[44] Dallmann, A., Zoller, D., Hotho, A.: A case study on sampling strategies for evaluating neural sequential item recommendation models. In: Proceedings of the 15th ACM Conference on Recommender Systems, pp. 505–514 (2021)

[45] Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., Malthouse, E.: User-centered evaluation of popularity bias in recommender systems. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, pp. 119–129 (2021)

[46] Rendle, S.: Factorization machines. In: 2010 IEEE International Conference on Data Mining, pp. 995–1000 (2010). IEEE

[47] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)

[48] Kool, W., Van Hoof, H., Welling, M.: Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In: International Conference on Machine Learning, pp. 3499–3508 (2019). PMLR

[49] Bennett, J., Lanning, S., *et al.*: The netflix prize. In: Proceedings of KDD Cup and Workshop, vol. 2007, p. 35 (2007). New York

[50] Geng, X., Zhang, H., Bian, J., Chua, T.-S.: Learning image and user features for recommendation in social networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4274–4282 (2015)

[51] Borges, R., Stefanidis, K.: On measuring popularity bias in collaborative filtering data (2020)

[52] Bobadilla, J., Dueñas-Lerín, J., Ortega, F., Gutierrez, A.: Comprehensive evaluation of matrix factorization models for collaborative filtering recommender systems (2023)

[53] Koren, Y.: Collaborative filtering with temporal dynamics. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 447–456 (2009)

[54] Duran, P.G., Karatzoglou, A., Vitria, J., Xin, X., Arapakis, I.: Graph convolutional embeddings for recommender systems. IEEE Access **9**, 100173–100184 (2021)

[55] Sun, Z., Fang, H., Yang, J., Qu, X., Liu, H., Yu, D., Ong, Y.-S., Zhang, J.: Daisyrec 2.0: Benchmarking recommendation for rigorous evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

[56] Liu, H., Wang, W., Zhang, Y., Gu, R., Hao, Y.: Neural matrix factorization recommendation for user preference prediction based on explicit and implicit feedback. Comput Intell Neurosci **2022**, 9593957 (2022)

[57] Yan, Z.: Serendipity: Accuracy's unpopular best friend in recommenders. eugeneyan.com (2020)

[58] Ahanger, A.B., Aalam, S.W., Bhat, M.R., Assad, A.: Popularity bias in recommender systems-a review. In: Emerging Technologies in Computer Engineering: Cognitive Computing and Intelligent IoT: 5th International Conference, ICETCE 2022, Jaipur, India, February 4–5, 2022, Revised Selected Papers, pp. 431–444 (2022). Springer

[59] Karakaya, M.Ö., Aytekin, T.: Effective methods for increasing aggregate diversity in recommender systems. knowledge and Information Systems **56**, 355–372 (2018)

[60] Wang, W., Lu, Y.: Analysis of the mean absolute error (mae) and the root mean square error (rmse) in assessing rounding model. In: IOP Conference Series: Materials Science and Engineering, vol. 324, p. 012049 (2018). IOP Publishing

[61] Yin, H., Cui, B., Li, J., Yao, J., Chen, C.: Challenging the long tail recommendation. Proceedings of the VLDB Endowment **5**(9), 896–907 (2012)

[62] Abdollahpouri, H., Burke, R., Mobasher, B.: Managing popularity bias in recommender systems with personalized re-ranking. arXiv preprint arXiv:1901.07555 (2019)

[63] Ziegler, C.-N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web, pp. 22–32 (2005)