Disparate Impact in Item Recommendation: A Case of Geographic Imbalance

Elizabeth Gómez[1], Ludovico Boratto[2(✉)], and Maria Salamó[1]

[1] Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
egomezye13@alumnes.ub.edu, maria.salamo@ub.edu
[2] Data Science and Big Data Analytics, Eurecat - Centre Tecnológic de Catalunya, Barcelona, Spain
ludovico.boratto@acm.org

Abstract. Recommender systems are key tools to push items' consumption. Imbalances in the data distribution can affect the exposure given to providers, thus affecting their experience in online platforms. To study this phenomenon, we enrich two datasets and characterize data imbalance w.r.t. the country of production of an item (*geographic imbalance*). We focus on movie and book recommendation, and divide items into two classes based on their country of production, in a majority-versus-rest setting. To assess if recommender systems generate a disparate impact and (dis)advantage a group, we introduce metrics to characterize the visibility and exposure a group receives in the recommendations. Then, we run state-of-the-art recommender systems and measure the visibility and exposure given to each group. Results show the presence of a disparate impact that mostly favors the majority; however, factorization approaches are still capable of capturing the preferences for the minority items, thus creating a positive impact for the group. To mitigate disparities, we propose an approach to reach the target visibility and exposure for the disadvantaged group, with a negligible loss in effectiveness.

Keywords: Recommender systems · Bias · Disparate impact

1 Introduction

Recommender systems learn patterns from users' behavior, to understand what might be of interest to them [37]. Natural imbalances in the data (e.g., in the amount of observations for popular items) might be embedded in the patterns. The produced recommendations can amplify these imbalances and create biases [9]. When a bias is associated to sensitive attributes of the users (e.g., gender or race), negative societal consequences can emerge, such as unfairness [22,23,30,33]. Unfairness can affect all the stakeholders of a system [1,5].

Data imbalances might be inherently connected to the way an industry is composed, e.g., with certain items mainly produced in certain parts of the world,

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12656, pp. 190–206, 2021.
https://doi.org/10.1007/978-3-030-72113-8_13

and with consumption patterns that differ based on the country of the users [4]. In this paper, we focus on geographic imbalance and study the problem of how the country of production of an item can create a disparate impact to providers in the recommendations. We assess disparate impact by considering both the *visibility* received by the providers of a group (i.e., the percentage of recommendations having them as providers) and their *exposure*, which accounts for the position in which items are recommended [41]. Hence, with these two metrics we measure respectively, (*i*) the share of recommendations of a group and (*ii*) the relevance that is given to that group. Both metrics are important to assess disparate impact in this context. Visibility alone might lead a group of providers not being reached by users in case they appear only at the bottom of the list, and exposure alone might not guarantee providers enough sales (a single item at the top of the list would mean these providers are recommended only once).

We assess disparate impact by comparing the visibility and exposure given to a group of providers with the representation of the group in the data. We study two forms of representation, based on (*i*) the amount of items a group offers, or (*ii*) the amount of ratings given to the items of a group.

We consider two of the main domains in which recommender systems operate, namely movies and books. We show, by extending two real-world datasets with the country of production of the items, that both movie and book data is imbalanced towards the United States. To understand the impact of this imbalance, we divide items into two groups, in a majority-versus-rest setting, and study how this imbalance is reflected in the visibility and exposure given to providers of the two groups when producing recommendations.

We consider state-of-the-art recommender systems, covering both model- and memory-based approaches, and point- and pair-wise algorithms. While commonly studied sensitive attributes, such as gender, show a disparate impact effect at the expense of the minority group, our use-case presents several peculiarities. Indeed, user preferences do not reflect these imbalances and users equally like items coming from the majority (the United States) and the minority (the rest of the countries) groups. This leads to disparity scenarios that affect either the majority or the minority group, according to patterns we present in this study.

To mitigate disparities, we propose a re-ranking that optimizes both the visibility and exposure given to providers, based on their representation in the data. Hence, we consider a distributive norm based on *equity* [43]. Our approach introduces in the recommendations items that increase the visibility and exposure of a group, causing the minimum possible loss in user relevance.

Our contributions can be summarized as follows:

- We study, for the first time, the impact of geographic imbalance in the data on the visibility and exposure given to different provider groups;
- We extend two real-world datasets with the country of production of each item and characterize the link between geographic imbalance and disparate impact, uncovering the factors that lead a group to be under-/over-exposed;
- We propose a re-ranking mitigation strategy that can lead to the target visibility and exposure with the minimum possible losses in effectiveness;

– We evaluate our approach, showing we can mitigate disparities with a negligible loss in effectiveness.

The rest of the paper details in Sect. 2 related work, while in Sect. 3 the scenario, metrics, recommenders, and datasets. Section 4 assesses disparate impact phenomena. Section 5 contains our mitigation algorithm and results. Section 6 concludes the paper.

2 Related Work

This section covers related studies, starting from the concepts of visibility and exposure in ranking, and continuing with the impact of recommendation for providers. We conclude by contextualizing our work with the existing studies.

Visibility and Exposure in Rankings. Given a ranking, visibility and exposure metrics respectively assess the amount of times an item is present in the rankings [21,45] and *where* an item is ranked [8,46]. They were introduced in the context of non-personalized rankings, where the objects being ranked are individual users (e.g., job candidates). These metrics can operate at the *individual* level, thus guaranteeing that similar individuals are treated similarly [8,19], or at *group* level, by making sure that users belonging to different groups are given adequate visibility or exposure [19,45,46]. Under the group setting, the visibility/exposure of a group is proportional to its representation in the data [32,35,38,44].

Impact of Recommendations for Providers. The impact of the generated recommendations on the item providers is a concept known as *provider fairness (P-fairness)*. It guarantees that the providers of the recommended objects that belong to different groups or are similar at the individual level, will get recommended according to their representation in the data. In this domain, Ekstrand et al. [20] assessed that collaborative filtering methods recommend books of authors of a given gender with a distribution that differs from that of the original user profiles. Liu and Burke [29] propose a re-ranking function, which balances recommendation accuracy and fairness, by dynamically adding a bonus to the items of the uncovered providers. Sonboli and Burke [42] define the concept of local fairness, to equalize access to capital across all types of businesses. Mehrotra et al. [31] assess unfairness based on the popularity of the providers. Several policies are defined to study the trade-offs between user-relevance and fairness. Kamishima et al. [26] introduce recommendation independence, which leads to recommendations that are statistically independent of sensitive features.

Contextualizing Our Work. While our study draws from metrics derived from fairness, *this work does not directly mitigate fairness for the individual providers*. We study a broader phenomenon, i.e., *if an industry of a country is affected by how recommendations are produced in presence of data imbalance.*

Considering our use-cases, both cinema and literature are powerful vehicles for culture, education, leisure, and propaganda, as highlighted by the UNESCO[1]. Moreover, both domains have an impact on the economy of a country, with (sometimes public) investments for the production of movies/books that are expected to generate a return. Hence, considering how recommender systems can push the consumption of items of a country is a related but different problem w.r.t. provider fairness.

3 Preliminaries

Here, we present the preliminaries, to provide foundations to our work.

3.1 Recommendation Scenario

Let $U = \{u_1, u_2, ..., u_n\}$ be a set of users, $I = \{i_1, i_2, ..., i_j\}$ be a set of items, and V be a totally ordered set of values that can be used to express a preference. The set of ratings is a ternary relation $R \subseteq U \times I \times V$; each rating is denoted by r_{ui}. These ratings can directly feed an algorithm in the form of triplets (point-wise approaches) or shape user-item observations (pair-wise approaches).

To assess the real impact of the recommendations, we consider a temporal split of the data, where a fixed percentage of the ratings of the users (ordered by timestamp) goes to the training and the rest goes to the test set [6].

The recommendation goal is to learn a function f that estimates the relevance ($\hat{r}_{ui}$) of the user-item pairs that do not appear in the training data. We denote as $\hat{R}$ the set of recommendations, and as $\hat{R}_G$ those involving items of a group G.

Let C_i be the set of production countries of an item i. We use it to shape two groups, a majority $M = \{i \in I : 1 \in C_i\}$, and a minority $m = \{i \in I : 1 \notin C_i\}$. Note that 1 identifies the country associated to the majority group.

3.2 Metrics

Representation. The representation of a group is the amount of times that group appears in the data. We consider two forms of representation, based on (i) the amount of items offered by a group and (ii) the amount of ratings collected for that group. We define with $\mathcal{R}$ the *representation* of a group G ($G \in \{M, m\}$) ($\mathcal{R}_I$ denotes an item-based representation, while $\mathcal{R}_R$ a rating-based representation):

$$\mathcal{R}_I(G) = |G|/|I| \tag{1}$$

$$\mathcal{R}_R(G) = |\{r_{ui} : i \in G\}|/|R| \tag{2}$$

Equation (1) accounts for the proportion of items of a group, while Eq. (2) for the proportion of ratings associated to a group. Both metrics are between 0 and 1.

[1] https://publications.parliament.uk/pa/cm200203/cmselect/cmcumeds/667/667.pdf.

The representation of a group is measured by considering only the training set. It is trivial to notice that, given a group G, the representation of the other, $\overline{G}$, can be computed as $\mathcal{R}_*(\overline{G}) = 1 - \mathcal{R}_*(G)$ (where '*' refers to I or R).

Disparate Impact. We assess disparate impact with two metrics.

Definition 1 (Disparate visibility). *The* disparate visibility *of a group is computed as the difference between the share of recommendations for items of that group and the representation of that group:*

$$\Delta\mathcal{V}(G) = \frac{1}{|U|} \sum_{u \in U} \frac{|\{\hat{r}_{ui} : i \in \hat{R}_G\}|}{|\hat{R}|} - \mathcal{R}_*(G) \tag{3}$$

Its range is in $[-\mathcal{R}_*(G), 1 - \mathcal{R}_*(G)]$; it is 0 when there is no disparate visibility, while negative/positive values indicate that the group received a share of recommendations lower/higher than its representation. This metric is based on that considered by Fabbri et al. [21].

Definition 2 (Disparate exposure). *The* disparate exposure *of a group is the difference between the exposure obtained by the group in the recommendation lists [41] and the representation of that group:*

$$\Delta\mathcal{E}(G) = \frac{1}{|U|} \sum_{u \in U} \frac{\sum_{pos=1}^{k} \frac{1}{log_2(pos+1)}, \forall i \in \hat{R}_G}{\sum_{pos=1}^{k} \frac{1}{log_2(pos+1)}} - \mathcal{R}_*(G) \tag{4}$$

where *pos* is the position of an item in the top-k recommendations.

This metric also ranges in $[-\mathcal{R}_*(G), 1 - \mathcal{R}_*(G)]$; it is 0 when there is no disparate exposure, while negative/positive values indicate that the exposure given to the group in the recommendations is lower/higher than its representation.

Notice that the disparate visibility/exposure of one group can be computed as the opposite of the value obtained for the other group.

Remark. *We do not define a unique "disparate impact" metric, to control both visibility and exposure, so that* providers are recommended enough times and with enough exposure. *A unique metric would not allow us to balance both, by compressing everything in a unique number.*

3.3 Recommendation Algorithms

We consider five state-of-the-art Collaborative Filtering algorithms. As memory-based approaches, we consider the UserKNN [24] and ItemKNN [39] algorithms. For the class of matrix factorization based approaches, we consider the BPR [36], BiasedMF [28], and SVD++ [27] algorithms. To contextualize our results, we also consider two non-personalized algorithms (MostPopular and RandomGuess).

3.4 Datasets

MovieLens-1M (Movies). The dataset provides 1M ratings (range 1–5), provided by 6,040 users, to 3,600 movies. It contains the IMDb ID of each movie, which allowed us to associate it to its country of production thanks to the OMDB APIs[2] (note that *each movie may have more than one country of production*).

Book Crossing (Books). The dataset contains 356k ratings (in the range 1–10), given by 10,409 users, to 14,137 books. The dataset contained the ISBN code of each book, which was used to add information about its countries of production thanks to the APIs offered by the Global Register of Publishers[3].

For both datasets, we encoded the country of production with an integer, with the United States (which represents the majority group in both datasets) having ID 1, and the rest of the countries having subsequent IDs.

4 Disparate Impact Assessment

In this section, we run the algorithms presented in Sect. 3.3 to assess their effectiveness and the disparate impact they generate.

4.1 Experimental Setting

For both datasets presented in Sect. 3.4, the test set was composed by the most recent 20% of the ratings of each user. To run the recommendation algorithms presented in Sect. 3.3, we considered the LibRec library (version 2). For each user, we generate 150 recommendations (denoted in the paper as the top-n) so that we can mitigate disparate impact through a re-ranking algorithm. The final recommendation list for each user is composed by 20 items (denoted as top-k).

Each algorithm was run with the following hyper-parameters:

- **UserKNN.** similarity: Pearson; neighbors: 50; similarity shrinkage: 10;
- **ItemKNN.** similarity: Cosine for Movies and Pearson for Books; neighbors: 200 (Movies), 50 (Books); similarity shrinkage: 10;
- **BPR.** iterator learnrate: 0.1; iterator learnrate maximum: 0.01; iterator maximum: 150; user regularization: 0.01; item regularization: 0.01; factor number: 10; learnrate bolddriver: false; learnrate decay = 1.0;
- **BiasedMF.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 20 (Movies), 1 (Books); user regularization: 0.01; item regularization: 0.01; bias regularization: 0.01; number of factors: 10; learnrate bolddriver: false; learnrate decay: 1.0;
- **SVD++.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 10 (Movies), 1 (Books); user regularization: 0.01; item regularization: 0.01; impItem regularization: 0.001; number of factors: 10; learnrate bolddriver: false; learnrate decay: 1.0.

[2] http://www.omdbapi.com/.
[3] https://grp.isbn-international.org/search/piid_cineca_solr.

To evaluate recommendation effectiveness, we measure the ranking quality of the lists by measuring the *Normalized Discounted Cumulative Gain* (NDCG) [25].

$$DCG@k = \sum_{u \in U} \hat{r}_{ui}^{pos} + \sum_{pos=2}^{k} \frac{\hat{r}_{ui}^{pos}}{log_2(pos)} \qquad NDCG@k = \frac{DCG@k}{IDCG@k} \tag{5}$$

where $\hat{r}_{ui}^{pos}$ is relevance of item i recommended to user u at position pos. The ideal DCG is calculated by sorting items based on decreasing true relevance (true relevance is 1 if the user interacted with the item in the test set, 0 otherwise).

4.2 Characterizing User Behavior

This section characterizes the group representation and users' rating behavior.

Group Representation. In the Movies dataset, $\mathcal{R}_I(m) = 0.3$ and $\mathcal{R}_R(m) = 0.23$. In the Books dataset, instead, $\mathcal{R}_I(m) = 0.12$ and $\mathcal{R}_R(m) = 0.08$. Both datasets show a strong geographic imbalance, with the majority group covering 70% of the items in the first dataset and 88% in the second. This imbalance is worsened when we consider the ratings, since in the movie context the ratings associated to the majority are 77%, while in the book content the rating representation for the majority is 92%. It becomes natural to ask ourselves if the majority group also attracts better ratings, to assess if this exacerbated imbalance is because majority items are perceived as of higher quality.

Table 1. Results of state-of-the-art recommender systems. Normalized Discounted Cumulative Gain (NDCG); Disparate Visibility for the minority group when considering the item representation as a reference ($\Delta\mathcal{V}_I$); Disparate Exposure for the minority group when considering the item representation as a reference ($\Delta\mathcal{E}_I$); Disparate Visibility for the minority group when considering the rating- representation as a reference ($\Delta\mathcal{V}_R$); Disparate Exposure for the minority group when considering the rating representation as a reference ($\Delta\mathcal{E}_R$). The values in bold indicate the best result.

	Movies					Books				
Algorithm	NDCG	$\Delta\mathcal{V}_I$	$\Delta\mathcal{E}_I$	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$	NDCG	$\Delta\mathcal{V}_I$	$\Delta\mathcal{E}_I$	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$
MostPop	0.1109	−0.1802	−0.2016	−0.1089	−0.1302	0.0089	−0.1239	−0.1239	−0.0839	−0.0840
RandomG	0.0105	**0.0020**	**0.0027**	0.0733	0.0740	8.91E+11	**0.0013**	**0.0015**	0.0412	0.0415
UserKNN	0.1247	−0.1544	−0.1668	−0.0831	−0.0955	0.0053	−0.0438	−0.0360	**−0.0039**	**0.0039**
ItemKNN	0.1199	−0.1744	−0.1926	−0.1031	−0.1212	0.0075	−0.0799	−0.0790	−0.0400	−0.0390
BPR	**0.1395**	−0.1054	−0.1087	**−0.0340**	**−0.0373**	0.0054	−0.0257	−0.0259	0.0142	0.0141
BiasedMF	0.0588	0.0901	0.0954	0.1614	0.1668	**0.0103**	−0.1239	−0.1239	−0.0840	−0.0840
SVD++	0.0684	0.0742	0.0762	0.1455	0.1475	**0.0103**	−0.1239	−0.1239	−0.0840	−0.0840

Rating Behavior. We considered the average rating associated to the items of each group. In the Movies dataset, the average rating for the majority group is 3.56, while that of the minority group is 3.61. In the Books dataset, we observed an average rating of 4.38 for the majority, and of 4.43 for the minority. This shows that the preference of the users for the two groups does not differ.

Observation 1. *Both datasets expose a big geographic imbalance in the representation of each group, in terms of offered items. The majority group usually attracts more ratings, thus increasing the existing imbalance. However, the minority items are not considered as of lower quality for the users, since the average rating for both groups is the same in both datasets.*

4.3 Assessing Effectiveness and Disparate Impact

We assess disparate impact in terms of visibility and exposure. Table 1 presents the results obtained when generating a top-20 ranking for each user, considering as a reference the minority group. The first phenomenon that emerges is that both groups can be affected by disparate impact and that, when one group receives more visibility, it also receives more exposure; hence, when a group is favored in the amount of recommendations, it is also ranked higher.

Considering the Movies dataset, MostPop, UserKNN, ItemKNN, and BPR present a disparate visibility and exposure that disadvantage the minority, for both forms of representation. The point-wise Matrix Factorization algorithms (BiasedMF and SVD++) and RandomGuess, instead, advantage the minority. This goes in contrast with the literature on algorithmic bias and fairness, where the minority is usually disadvantaged. We conjecture that, since recommender systems do not receive any information about the geographic groups and since users equally prefer the items of the two groups, the point-wise Matrix Factorization approaches create factors that capture user preferences as a whole. Our results align with those of Cremonesi et al. [14], who showed the capability of factorization approaches to recommend long-tail items. Interestingly, when considering disparate visibility and exposure, the best results for the item-based representation are those of RandomGuess; nevertheless, the algorithm is also the least effective in terms of NDCG. No algorithm can offer both effectiveness and adapt to the offer of a country. When considering the rating-based representation, BPR is the most effective and has the lowest disparate visibility and exposure. Hence, the combination between factorization approaches and a pair-wise training can connect effectiveness and equity of visibility and exposure.

In the Books dataset, besides MostPop, all the approaches advantage the majority. This opposite trend in terms of disparate impact of the point-wise Matrix Factorization algorithms (BiasedMF and SVD++) w.r.t. the Movies dataset, can be explained by considering that the items having more ratings will lead to factors that have more weight at prediction stage; here, the majority is much larger than in the Movies dataset, so this leads to the group being advantaged in terms of visibility and exposure. This dataset is much also more

sparse, so effectiveness is strongly reduced, and the point-wise Matrix Factorization approaches are the most effective. There is no connection between effectiveness and equity of exposure and visibility. Indeed, RandomGuess and UserKNN are, respectively, the best algorithms when considering the item-/rating-based representation of the groups. This good visibility and exposure provided by UserKNN in the rating-based setting can be connected to phenomena observed by Cañamares and Castells [11] since, under sparsity, the algorithm adapts to item popularity.

Observation 2. *Geographic imbalance almost always affects the minority group, since we feed algorithms with much more instances than their counterpart. Matrix Factorization based approaches can help the minority receive more visibility and exposure, with latent factors that capture preferences also of the minority. However, if the imbalance is too severe, the minority is always affected by disparate impact.*

5 Mitigating Disparate Impact

The previous section allowed us to observe a new phenomenon that departs from the existing algorithmic fairness studies, since *the minority group is not always the disadvantaged one when considering geographic imbalance*. Still, our results show that we can always observe a group receiving a disproportional visibility and exposure with respect to its representation in the data.

In this section, we mitigate these phenomena by presenting a re-ranking algorithm that introduces items of the disadvantaged group in the recommendation list, to reach a visibility and an exposure proportional to its representation.

A re-ranking algorithm is the only option when optimizing ranking-based metrics, like visibility and exposure. An in-processing regularization, such as those presented in [7,26], would not be possible, since at prediction stage the algorithm does not predict *if and where* an item will be ranked in a list. Re-rankings have been introduced to reduce disparities, both for non-personalized rankings [8,13,32,41,45,46] and for recommender systems [10,31], with approaches such as Maximal Marginal Relevance [12]. These algorithms optimize only one property (visibility or exposure), so no direct comparison is possible.

5.1 Algorithm

The foundation behind our mitigation algorithm is to *move up in the recommendation list the item that causes the minimum loss in prediction for all the users*. We start by targeting the desired visibility, to make sure the items of the disadvantaged group are recommended enough times. Then we move items up inside the recommendation list to reach the target exposure.

The mitigation is described in Algorithm 1. The inputs are the recommendations (top-n items), the current visibility and exposure of the disadvantaged

Input: *recList*: ranked list (records contain *user*, *item*, *prediction*, *exposure*, *group*, *position*), *vis*: visibility of disadvantaged group, *exp*: exposure of disadvantaged group, *rep*: representation of disadvantaged group, *advG*: ID of advantaged group, *disadvG*: ID of disadvantaged group
Output: *reRankedList*: ranked list adjusted by visibility and exposure

```
define optimizeVisibilityExposure (recList, vis, exp, rep)
begin
    reRankedList ← mitigation(recList, vis, rep, advG, disadvG,
    "visibility")
    reRankedList ← mitigation(reRankedList, exp, rep, advG, disadvG,
    "exposure")
    return reRankedList
end

define mitigation (list, VE, rep, advG, disadvG, rankingType)
begin
    for user ∈ list.users do
        losses.add(calculateLoss(list, user, rankingType, advG, disadvG)
    end
    while VE < rep do
        minLoss ← losses.sortByLoss(0)
        list ← swap(list, minlLoss.itemAdvG, minLoss.itemDisadvG)
        if reRankingType == "visibility" then
            VE ← VE + 1
        else
            VE ← (VE − minLoss.itemDisadvG.exposure) +
            minLoss.itemAdvG.exposure
        end
        losses.add(calculateLoss(list, user, rankingType, advG, disadvG))
    end
    return list
end

define calculateLoss (list, user, rankingType, advG, disadvG)
begin
    itemAdvGroup ← getlastItem(list, user,top-k,advGroup)
    if reRankingType == "visibility" then
        itemDisadvGroup ← getfirstItem(list, user,last-n,disadvGroup)
    else
        while itemAdvGroup.position > itemDisadvGroup.position do
            itemDisadvGroup ← getnextItem(list, user,top-k,disadvGroup)
        end
    end
    loss ← itemAdvGroup.prediction - itemDisadvGroup.prediction
    lossUser ← [user, itemAdvGroup, itemDisadvGroup, loss]
    return lossUser
end
```

Algorithm 1: Visibility and exposure mitigation algorithm

group and its representation in the data (our target), and the IDs of the advantaged and disadvantaged groups. The output is the re-ranked list of items.

The *optimizeVisibilityExposure* method (lines 1–6), executes the mitigation, firstly to regulate the visibility of the disadvantaged group (by adding their items to the top-k) and secondly to regulate the exposure (by moving their items up in the top-k). The *mitigation* method (lines 7–23) regulates the visibility and exposure of the recommendation list. First, we loop over the users (lines 9–11) and call the *calculateLoss* method, to calculate the loss (in terms of items' predicted relevance) we would have in each user's list when swapping the items of the two groups. The while loop (lines 12–21) swaps the items until the target visibility/exposure is reached; line 13 returns the user that causes the minimum loss and line 14 swaps their items. If the goal is to reach a target visibility, lines 15–16 increase the visibility of the group by 1; if the swap is done to reach a target exposure, lines 17–19 subtract the exposure of the old item and add that of the new one. Finally, the *calculateLoss* method recalculates the loss for the user object of the swap and returns the re-ranked list.

The *calculateLoss* method (lines 24–37) identifies the user causing the minimal loss of predicted relevance. We select two items in the list of each user. The first is the last item of the advantaged group in the top-k (line 26). If we are

Table 2. Impact of mitigation on recommended lists with item-based representation. Normalized Discounted Cumulative Gain (NDCG); Disparate Visibility ($\Delta\mathcal{V}_I$) for the minority; Disparate Exposure ($\Delta\mathcal{E}_I$) for the minority. We report below gain/loss of each setting w.r.t. the original one (left side of Table 1).

	Mitigation visibility & exposure					
	Movies			Books		
Algorithm	NDCG	$\Delta\mathcal{V}_I$	$\Delta\mathcal{E}_I$	NDCG	$\Delta\mathcal{V}_I$	$\Delta\mathcal{E}_I$
MostPop	0.1052	−0.0017	−0.0017	0.0087	−0.0039	−0.0039
(gain/loss)	−0.0057	0.1785	0.1999	−0.0002	0.1200	0.1200
RandomG	0.0106	−0.0017	−0.0017	8.91E+11	−0.0039	−0.0039
(gain/loss)	0.0001	−0.0036	−0.0043	3.24E+09	−0.0052	−0.0055
UserKNN	0.1205	−0.0017	−0.0017	0.0050	−0.0039	−0.0039
(gain/loss)	−0.0042	0.1528	0.1652	−0.0003	0.0399	0.0321
ItemKNN	0.1173	−0.0017	−0.0017	0.0075	−0.0039	−0.0039
(gain/loss)	−0.0027	0.1727	0.1909	0.0000	0.0760	0.0751
BPR	**0.1372**	**−0.0017**	**−0.0017**	0.0055	−0.0039	−0.0039
(gain/loss)	−0.0023	0.1037	0.1070	0.0001	0.0218	0.0220
BiasedMF	0.0623	−0.0017	−0.0017	**0.0119**	**−0.0039**	**−0.0039**
(gain/loss)	0.0035	−0.0918	−0.0971	0.0016	0.1200	0.1200
SVD++	0.0712	−0.0017	−0.0017	0.0113	−0.0039	−0.0039
(gain/loss)	0.0028	−0.0759	−0.0779	0.0011	0.1200	0.1200

regulating visibility, lines 27–28 select the first item of the disadvantaged group out of the top-k (denoted as last-n). Lines 29–33 mitigate for exposure; the while selects an item of the disadvantaged group that in the top-k is currently ranked lower than that of its counterpart. Once we obtain the pair of items for the user, we calculate the loss by considering the *prediction* attribute (line 34). Finally, line 35 collects the loss of the user, which is returned in line 36.

5.2 Impact of Mitigation

In this section, we assess the impact of our mitigation. Since we split data temporally, we cannot run statistical tests to assess the difference in the results, so we highlight the gain/loss obtained for each measure.

Results are reported in Tables 2 and 3 separating them between item- and rating-based representation of the groups. Trivially, given a target representation and a dataset, all algorithms achieve the same disparate visibility/exposure. Let us consider the trade-off between disparate visibility/exposure and effectiveness. Considering the Movies dataset, in both representations of the groups, BPR is the algorithm with the best trade-off between effectiveness and equity of visibility and exposure. It was already the most accurate algorithm, and thanks to our mitigation based on the minimum-loss principle, the loss in NDCG was

Table 3. Impact of mitigation on recommended lists with rating-based representation. Normalized Discounted Cumulative Gain (NDCG); Disparate Visibility ($\Delta\mathcal{V}_R$) for the minority; Disparate Exposure ($\Delta\mathcal{E}_R$) for the minority. We report below gain/loss of each setting w.r.t. the original one (left side of Table 1).

	Mitigation visibility & exposure					
	Movies			Books		
Algorithm	NDCG	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$	NDCG	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$
MostPop	0.1076	−0.0003	−0.0003	0.0089	−0.0040	−0.0040
(gain/loss)	−0.0032	0.1085	0.1299	−0.0006	0.0800	0.0800
RandomG	0.0112	−0.0003	−0.0003	8.54E+11	−0.0040	−0.0040
(gain/loss)	0.0006	−0.0736	−0.0743	−2.37E+10	−0.0452	−0.0455
UserKNN	0.1239	−0.0003	−0.0003	0.0050	−0.0040	−0.0040
(gain/loss)	−0.0008	0.0828	0.0952	−0.0003	−0.0001	−0.0079
ItemKNN	0.1185	−0.0003	−0.0003	0.0075	−0.0040	−0.0040
(gain/loss)	−0.0015	0.1027	0.1209	0.0001	0.0360	0.0351
BPR	**0.1390**	**−0.0003**	**−0.0003**	0.0053	−0.0040	−0.0040
(gain/loss)	−0.0005	0.0337	0.0370	−0.0001	−0.0182	−0.0180
BiasedMF	0.0648	−0.0003	−0.0003	**0.0122**	**−0.0040**	**−0.0040**
(gain/loss)	0.0060	−0.1618	−0.1671	0.0016	0.0800	0.0800
SVD++	0.0735	−0.0003	−0.0003	0.0113	−0.0040	−0.0040
(gain/loss)	0.0051	−0.1459	−0.1479	0.0011	0.0800	0.0800

negligible. In the Books dataset, BiasedMF confirms to be the best approach, in both effectiveness and equity of visibility and exposure. It is interesting to observe that, in both scenarios, MostPop is the second most effective algorithm and now provides the same visibility and exposure as the other algorithms; this is due to popularity bias phenomena [2], and their analysis is left as future work.

Observation 3. *When providing a re-ranking based on minimal predicted loss, the effectiveness remains stable, but disparate visibility and disparate exposure are mitigated.*

6 Conclusions and Future Work

In this paper, we considered data imbalance in the items' country of production of items (*geographic imbalance*). We considered a group setting based on a majority-versus-rest split of the items and defined measures to assess disparate visibility and disparate exposure for groups. The results of five collaborative filtering approaches show that the minority group is not always disadvantaged.

We proposed a mitigation algorithm that produces a re-ranking, by adding to the recommendation lists items that cause the minimum loss in predicted relevance. Results show that *thanks to our approach, any recommendation algorithm can bring equity of visibility and exposure to providers, without impacting the end-users in terms of effectiveness.*

Future work will study geographic imbalance in education, to explore country-based disparities for teachers [3,16–18]. Moreover, we will evaluate divergence-based disparity metrics [15]) and consider multi-class group settings. Other issues emerging from imbalanced groups, such as bribing [34,40], will be considered.

Acknowledgments. This research was partially funded by project 2017-SGR-341, MISMIS-LANGUAGE (grant No. PGC2018-096212-B-C33) from the Spanish Ministry of Science and Innovation, and NanoMoocs (grant No. COMRDI18-1-0010) from ACCIÓ. L. Boratto acknowledges Agència per a la Competivitat de l'Empresa, ACCIÓ, for their support under project "Fair and Explainable Artificial Intelligence (FX-AI)".

References

1. Abdollahpouri, H., et al.: Multistakeholder recommendation: survey and research directions. User Model. User-Adap. Interact. **30**(1), 127–158 (2020). https://doi.org/10.1007/s11257-019-09256-1
2. Abdollahpouri, H., Mansoury, M.: Multi-sided exposure bias in recommendation (2020)
3. Barra, S., Marras, M., Fenu, G.: Continuous authentication on smartphone by means of periocular and virtual keystroke. In: Au, M.H., et al. (eds.) NSS 2018. LNCS, vol. 11058, pp. 212–220. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02744-5_16

4. Bauer, C., Schedl, M.: Global and country-specific mainstreaminess measures: definitions, analysis, and usage for improving personalized music recommendation systems. PLOS ONE **14**(6), 1–36 (2019). https://doi.org/10.1371/journal.pone.0217389
5. Bauer, C., Zangerle, E.: Leveraging multi-method evaluation for multi-stakeholder settings. CoRR abs/2001.04348 (2020)
6. Bellogín, A., Castells, P., Cantador, I.: Statistical biases in information retrieval metrics for recommender systems. Inf. Retrieval J. **20**(6), 606–634 (2017). https://doi.org/10.1007/s10791-017-9312-z
7. Beutel, A., et al.: Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, pp. 2212–2220. ACM (2019). https://doi.org/10.1145/3292500.3330745
8. Biega, A.J., Gummadi, K.P., Weikum, G.: Equity of attention: amortizing individual fairness in rankings. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, pp. 405–414. ACM (2018). https://doi.org/10.1145/3209978.3210063
9. Boratto, L., Fenu, G., Marras, M.: The effect of algorithmic bias on recommender systems for massive open online courses. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11437, pp. 457–472. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15712-8_30
10. Burke, R., Sonboli, N., Ordonez-Gauger, A.: Balanced neighborhoods for multi-sided fairness in recommendation. In: Conference on Fairness, Accountability and Transparency, FAT 2018, Proceedings of Machine Learning Research, vol. 81, pp. 202–214. PMLR (2018)
11. Cañamares, R., Castells, P.: A probabilistic reformulation of memory-based collaborative filtering: implications on popularity biases. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 215–224. ACM (2017). https://doi.org/10.1145/3077136.3080836
12. Carbonell, J.G., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR 1998: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335–336. ACM (1998). https://doi.org/10.1145/290941.291025
13. Celis, L.E., Straszak, D., Vishnoi, N.K.: Ranking with fairness constraints. In: 45th International Colloquium on Automata, Languages, and Programming, ICALP 2018. LIPIcs, vol. 107, pp. 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2018). https://doi.org/10.4230/LIPIcs.ICALP.2018.28
14. Cremonesi, P., Koren, Y., Turrin, R.: Performance of recommender algorithms on top-n recommendation tasks. In: Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, pp. 39–46. ACM (2010). https://doi.org/10.1145/1864708.1864721
15. Deldjoo, Y., Anelli, V.W., Zamani, H., Kouki, A.B., Noia, T.D.: Recommender systems fairness evaluation via generalized cross entropy. In: Burke, R., Abdollahpouri, H., Malthouse, E.C., Thai, K.P., Zhang, Y. (eds.) Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments Co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, 20 September 2019, CEUR Workshop Proceedings, vol. 2440. CEUR-WS.org (2019)

16. Dessì, D., Dragoni, M., Fenu, G., Marras, M., Reforgiato Recupero, D.: Evaluating neural word embeddings created from online course reviews for sentiment analysis. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC 2019, pp. 2124–2127. ACM (2019). https://doi.org/10.1145/3297280.3297620
17. Dessì, D., Fenu, G., Marras, M., Reforgiato Recupero, D.: Leveraging cognitive computing for multi-class classification of E-learning videos. In: Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10577, pp. 21–25. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70407-4_5
18. Dessì, D., Fenu, G., Marras, M., Reforgiato Recupero, D.: COCO: semantic-enriched collection of online courses at scale with experimental use cases. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) WorldCIST'18 2018. AISC, vol. 746, pp. 1386–1396. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-77712-2_133
19. Diaz, F., Mitra, B., Ekstrand, M.D., Biega, A.J., Carterette, B.: Evaluating stochastic rankings with expected exposure. CoRR abs/2004.13157 (2020)
20. Ekstrand, M.D., Tian, M., Kazi, M.R.I., Mehrpouyan, H., Kluver, D.: Exploring author gender in book rating and recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, pp. 242–250. ACM (2018). https://doi.org/10.1145/3240323.3240373
21. Fabbri, F., Bonchi, F., Boratto, L., Castillo, C.: The effect of homophily on disparate visibility of minorities in people recommender systems. In: Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, pp. 165–175. AAAI Press (2020)
22. Fenu, G., Lafhouli, H., Marras, M.: Exploring algorithmic fairness in deep speaker verification. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12252, pp. 77–93. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58811-3_6
23. Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: from discrimination discovery to fairness-aware data mining. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2125–2126. ACM (2016). https://doi.org/10.1145/2939672.2945386
24. Herlocker, J.L., Konstan, J.A., Riedl, J.: An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. Inf. Retrieval **5**(4), 287–310 (2002). https://doi.org/10.1023/A:1020443909834
25. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002). https://doi.org/10.1145/582415.582418
26. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Recommendation independence. In: Conference on Fairness, Accountability and Transparency, FAT 2018, Proceedings of Machine Learning Research, vol. 81, pp. 187–201. PMLR (2018)
27. 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. ACM (2008). https://doi.org/10.1145/1401890.1401944
28. Koren, Y., Bell, R.M., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Comput. **42**(8), 30–37 (2009). https://doi.org/10.1109/MC.2009.263
29. Liu, W., Burke, R.: Personalizing fairness-aware re-ranking. CoRR abs/1809.02921 (2018)

30. Marras, M., Korus, P., Memon, N.D., Fenu, G.: Adversarial optimization for dictionary attacks on speaker verification. In: Kubin, G., Kacic, Z. (eds.) Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019, pp. 2913–2917. ISCA (2019). https://doi.org/10.21437/Interspeech.2019-2430
31. Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., Diaz, F.: Towards a fair marketplace: counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, pp. 2243–2251. ACM (2018). https://doi.org/10.1145/3269206.3272027
32. Patro, G.K., Biswas, A., Ganguly, N., Gummadi, K.P., Chakraborty, A.: FairRec: two-sided fairness for personalized recommendations in two-sided platforms. In: WWW 2020: The Web Conference 2020, pp. 1194–1204. ACM/IW3C2 (2020). https://doi.org/10.1145/3366423.3380196
33. Ramos, G., Boratto, L.: Reputation (in)dependence in ranking systems: demographics influence over output disparities. In: Huang, J., et al. (eds.) Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Virtual Event, China, 25–30 July 2020, pp. 2061–2064. ACM (2020). https://doi.org/10.1145/3397271.3401278
34. Ramos, G., Boratto, L., Caleiro, C.: On the negative impact of social influence in recommender systems: a study of bribery in collaborative hybrid algorithms. Inf. Process. Manag. **57**(2), 102058 (2020). https://doi.org/10.1016/j.ipm.2019.102058
35. Ramos, G., Caleiro, C.: A novel similarity measure for group recommender systems with optimal time complexity. In: Boratto, L., Faralli, S., Marras, M., Stilo, G. (eds.) BIAS 2020. CCIS, vol. 1245, pp. 95–109. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52485-2_10
36. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian personalized ranking from implicit feedback. In: UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 452–461. AUAI Press (2009)
37. Ricci, F., Rokach, L., Shapira, B.: Recommender systems: introduction and challenges. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 1–34. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_1
38. Sapiezynski, P., Zeng, W., Robertson, R.E., Mislove, A., Wilson, C.: Quantifying the impact of user attentionon fair group representation in ranked lists. In: Companion of The 2019 World Wide Web Conference, WWW 2019, pp. 553–562. ACM (2019). https://doi.org/10.1145/3308560.3317595
39. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the Tenth International World Wide Web Conference, WWW 10, pp. 285–295. ACM (2001). https://doi.org/10.1145/371920.372071
40. Saúde, J., Ramos, G., Caleiro, C., Kar, S.: Reputation-based ranking systems and their resistance to bribery. In: Raghavan, V., Aluru, S., Karypis, G., Miele, L., Wu, X. (eds.) 2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, 18–21 November 2017, pp. 1063–1068. IEEE Computer Society (2017). https://doi.org/10.1109/ICDM.2017.139
41. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, pp. 2219–2228. ACM (2018). https://doi.org/10.1145/3219819.3220088

42. Sonboli, N., Burke, R.: Localized fairness in recommender systems. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, UMAP 2019, pp. 295–300. ACM (2019). https://doi.org/10.1145/3314183.3323845
43. Walster, E., Berscheid, E., Walster, G.W.: New directions in equity research. J. Pers. Soc. Psychol. **25**(2), 151 (1973)
44. Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pp. 22:1–22:6. ACM (2017). https://doi.org/10.1145/3085504.3085526
45. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: FA*IR: a fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, pp. 1569–1578. ACM (2017). https://doi.org/10.1145/3132847.3132938
46. Zehlike, M., Castillo, C.: Reducing disparate exposure in ranking: a learning to rank approach. In: WWW 2020: The Web Conference 2020, pp. 2849–2855. ACM/IW3C2 (2020). https://doi.org/10.1145/3366424.3380048