Information Processing and Management 59 (2022) 102766

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Recommender systems effect on the evolution of users' choices distribution

Naieme Hazrati *, Francesco Ricci

Free University of Bozen-Bolzano, Bolzano, Italy

ARTICLE INFO

Keywords:
Recommender systems
Simulation
Decision making
Choice models

ABSTRACT

Recommender systems' (RSs) research has mostly focused on algorithms aimed at improving platform owners' revenues and user's satisfaction. However, RSs have additional effects, which are related to their impact on users' choices. In order to avoid an undesired system behaviour and anticipate the effects of an RS, the literature suggests employing simulations.

In this article we present a novel, well grounded and flexible simulation framework. We adopt a stochastic user's choice model and simulate users' repeated choices for items in the presence of alternative RSs. Properties of the simulated choices, such as their diversity and their quality, are analysed. We state four research questions, also motivated by identified research gaps, which are addressed by conducting an experimental study where three different data sets and five alternative RSs are used. We identify some important effects of RSs. We find that non-personalised RSs result in choices for items that have a larger predicted rating compared to personalised RSs. Moreover, when a user's awareness set, which is the set containing the items that she can choose from, increases, then choices are more diverse, but the average quality (rating) of the choices decreases. Additionally, in order to achieve a higher choice diversity, increasing the awareness of the users is shown to be a more effective remedy than increasing the number of recommendations offered to the users.

1. Introduction

Recommender systems (RSs) have become indispensable tools for supporting online users in their decision making activity (Ricci et al., 2015). The research on RSs has been primarily dedicated to optimising the precision of the recommendations, which has been considered an important indicator of the goodness of the RS (Alhijawi & Kilani, 2020; Li et al., 2007). However, standard RSs evaluation methods, which have been used to measure *recommendation* precision, do not assess important properties of the actual users' *choices* for items, which are informed and stimulated by the recommendations. Choices are the items chosen by the users when exposed to the system's recommendations. It is important to stress this distinction between recommendations and choices, because not all the recommendations become users' choices, and users, in real scenarios, often choose items that have not been recommended.

Fortunately, nowadays, there is a growing attention to better understand the effects of RSs on users' decision making, also because it has been shown that RSs suffer from various types of biases, e.g., the popularity bias (Abdollahpouri et al., 2020; Fleder & Hosanagar, 2009; Mansoury et al., 2020; Matt et al., 2013; Yalcin & Bilge, 2021; Yao et al., 2021). Hence, the effect of RSs on users' choice behaviour has become a topic, studied by either conducting online user studies, i.e., by observing the effect of RSs on the choices of real users (Lee & Hosanagar, 2014, 2019, 2021; Matt et al., 2013; Zhu et al., 2018), or by designing algorithmic offline

* Corresponding author.
E-mail addresses: nhazrati@unibz.it (N. Hazrati), fricci@unibz.it (F. Ricci).

https://doi.org/10.1016/j.ipm.2021.102766
Received 14 May 2021; Received in revised form 2 September 2021; Accepted 14 September 2021
Available online 4 October 2021

simulations of the users' choices in the presence of an RS (Fleder & Hosanagar, 2007, 2009; Hazrati et al., 2020; Nadolski et al., 2009; Szlávik et al., 2011; Yao et al., 2021). While observing the choices of real subjects is probably the most reliable method, it also brings the complexity and cost of building an operational RS and running (many of) such evaluations. Conversely, the simulation of an RS's effect on users' choices is simpler to implement and enables to test alternative system configurations. In simulation studies, repeated simulated choices of the users, even for a long (simulated) temporal interval, are generated, while the users are influenced by a target RS. In the analysis of the simulated users' choices, various metrics, which capture the distribution and quality of the choices, have been considered: the Gini index (Fleder & Hosanagar, 2009), the catalogue coverage of the choices and the average users' ratings of the choices (Szlávik et al., 2011).

Simulation of users' choices has produced some interesting results, hence showing its validity and importance for understanding the RSs' effect on users. However, a simulation always provides a partial reconstruction of the reality and specific assumptions must be made. By analysing the state of the art, we found that some of these assumptions are too restricted and could be relaxed, hence a more reliable simulation framework can be produced. In fact, previous simulation studies, in order to define the possible choices and the users' choice model, have used synthetic data (items and users' profiles) instead of real users and item descriptions, which can be derived from operational system log data (Chaney et al., 2018; Fleder & Hosanagar, 2009). Others did not properly model the knowledge of the users for the catalogue of items and the user preferences for items, which influence their choices (Szlávik et al., 2011).

In our study, we address these limitations by deriving items and users information from three systems' log data sets of users' ratings for purchased (chosen) items (Amazon data sets). Moreover, we provide a more realistic definition of the users' knowledge of the items by defining a user specific awareness set. We also model the users' preferences for items with a novel debiased Matrix Factorisation rating prediction approach (Schnabel et al., 2016). These solutions were never used in previous simulations and make it possible to obtain a more reliable and realistic assessment of RSs effect. Moreover, previously conducted simulations did not evaluate the role of some important contextual settings: how many items are recommended and how wide is the users' knowledge of the items catalogue (awareness set). We fully analyse the impact of these important settings.

In the rest of this article, we first sketch the proposed approach and list the research questions that we address (Section 2). Then, after having discussed the state of the art in Section 3, we detail the proposed simulation framework (Section 4) and the design of the simulation experiments (Section 5). The experimental results are presented in Section 6. In Section 7 we compare our results with those obtained by previous studies. Finally, in Section 8 we summarise the obtained results, we discuss some limitations of our approach and we indicate important future lines of research.

2. Simulation framework and research questions

2.1. Proposed simulation framework

Fig. 1 shows the general schema of the proposed simulation framework. The full details of this model will be described later on, but we give here a description of the most important components: the simulated user, the user's awareness set and the recommender system. The *simulated user* is described by a choice model that, based on the estimated utility of items, makes it possible to simulate repeated choices for items belonging to the awareness set. The user specific *awareness set* contains items that the user is supposed to know before the recommendations are provided, plus the recommended items. In fact, the user is supposed to have some previous knowledge of popular items and items that are estimated to have a large utility, i.e., items that the user tends to appreciate. The utility of the items is estimated by using an existing rating data set, hence, we try to generate a precise model of the user's preferences by using all the available information about the user. Finally, the *recommender system* computes recommendations on the base of the simulated users' choices (this user and all the other simulated users). Hence, the RS, as new choices are simulated, adapts the recommendations to the observed preferences (choices) of the users. This creates a dynamic scenario where the RS adapts to the observed users' choices and the user choices are influenced by the RS. We study the variation of important metrics that describe the diversity and quality of the users' choices in this scenario.

2.2. Research questions

After the analysis of the literature, discussed in Section 3, we have decided to focus on the following important and yet not clearly answered research questions.

1. RQ1: How personalised and non-personalised RSs affect the evolution of choice diversity? What features of the RSs determine their specific impact? Personalised RSs produces a specific recommendation set for each user, hence, they should also produce more diverse choices, but how specific contextual parameters such as the number of recommendations and the user's awareness affect the diversity of the choices is worth to be analysed.
2. RQ2: Do personalised RSs suggest items that users rate higher than non-personalised RSs? Personalised RSs are supposed to perform better than non-personalised RSs, i.e., they should have better precision. But, it is important to verify under which conditions personalised RSs outperform non-personalised ones when measuring the quality of the users' choices, i.e., the users' ratings given to their choices.

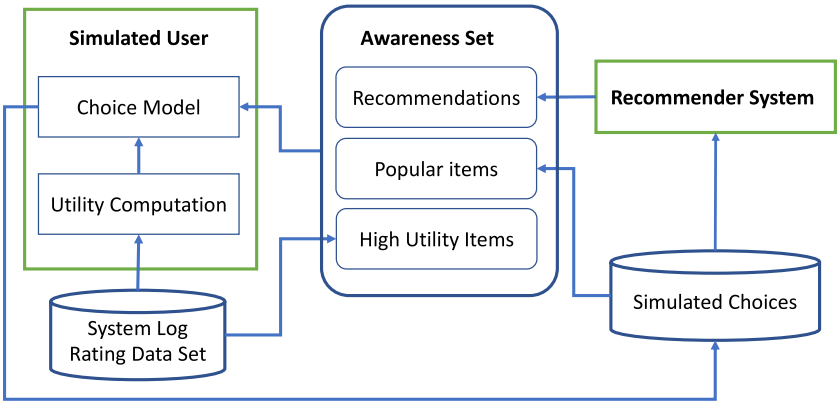

Fig. 1. General schema of the proposed simulation framework.

3. RQ3: Does a better users' awareness of the catalogue of the items, i.e., being aware of a larger number of items, lead to better choices, that is, higher users' rating for the choices? We expect that when the awareness of the catalogue increases, the diversity will increase as well, but the choice's rating could decrease because fewer recommendations are chosen, and recommendations are supposed to produce better choices (i.e., with higher ratings).
4. RQ4: Is a larger recommendation set producing an increase of choice diversity? If yes, has the awareness set size a larger or smaller effect on choice diversity than the recommendation set size? While recommending more items should result in a higher choice diversity, increasing the awareness set size may even be more effective in diversifying the choices.

3. Related work

The studies dedicated to analysing the effect of RSs on users' choice behaviour have followed two main approaches: online experiments and simulation of user behaviour. In online experiments, a web platform offering products to users, while alternative RSs make recommendations to users, is implemented. Then, the measurable effect of the used RSs on the actual users' choices is compared (Matt et al., 2013; Senecal et al., 2005; Zhu et al., 2018). Matt et al. (2013) designed a website offering music tracks to users. They randomly assigned users to five distinct groups. In each group a specific RS suggests items to the users. Then, they analyse the effect of the implemented RSs on users' choices by measuring diversity metrics such as the Gini index (Dorfman, 1979). Even though their results are interesting, they have tested their hypotheses only on a small set of users (32 users) and one single domain (music domain).

Online experiments have achieved interesting results that show some effects of RSs on users' choice behaviour. However, they offer a limited perspective, because only a few scenarios can be analysed and it is difficult to generalise the obtained results. As a consequence, few researchers have followed this approach.

Conversely, the simulation of users' choice behaviour is simpler to implement and many alternative conditions can be studied. In a simulation, a collection of simulated users (agents) make (repeated) choices for items, while an RS is also simulated to affect the users' choices. Simulations of user behaviour in RSs have been primarily used to evaluate reinforcement-learning approaches (Huang et al., 2020; Ie et al., 2019; Zhao et al., 2019). However, some studies have adopted the simulation approach to understand users' choice behaviour in the presence of more general RSs (Bountouridis et al., 2019; Fleder & Hosanagar, 2007, 2009; Hazrati et al., 2019; Nadolski et al., 2009; Sie et al., 2010; Szlávik et al., 2011; Umeda et al., 2014).

Fleder and Hosanagar (2009) introduced a simulation procedure in which the users iteratively select items among a small set of candidate fictitious products based on a probabilistic multinomial-logit choice model (Brock & Durlauf, 2002). The model is based on *randomly* generated utility functions, one for each user: the higher is the computed item utility, the more likely the item is chosen. They were the first to assume and simulate that users cannot select an arbitrary catalogue item, but only items in a smaller, user specific set, called awareness set, that represents the users' knowledge of the items in the catalogue. Furthermore, a recommended item is assumed to receive an increased utility for the target user. This simulates the recommendation's effect to increase the visibility and salience of an item and consequently the user's estimation of its utility. Finally, they observed the effect of RSs on the users' choices in terms of choices' diversity. While, this study warned about the lack of diversity that could be produced by an RS, it was conducted on a small set of (50) fictitious users and items. Conversely, our simulation approach is designed to use more realistic and large data sets of logged user/item interactions (Amazon eCommerce platform). Moreover, while Fleder and Hosanagar (2009) used a random utility function for each user, in our simulation (see Fig. 1), the utility functions are derived from the above mentioned data sets of real users' choices and ratings. Furthermore, they assume that in each time interval of the simulation, each user makes a single choice, while we extract from the log data set the number of simulated choices that each user makes.

In another work, Bountouridis et al. (2019) adopt a simulation framework similar to that proposed in Fleder and Hosanagar (2009) and they apply it to the news domain. Their simulation allows content providers to select different RSs and analyse their

Table 1
Summary list of the most important symbols and notations.

U	Full set of users
I	Full set of items
R	Matrix of observed ratings in the log data
$\hat{R}$	Matrix of observed and predicted ratings
L	Number of the simulated time intervals
t_0	Starting time point of the simulation
t_l	End time point of the l'th time interval
$]t_{l-1}, t_l]$	Time interval between t_{l-1} (excluded) , t_l (included)
P	Binary $\|U\| \times \|I\|$ matrix of the observed choices in the log data
P^0	Matrix of the observed choices in the log data until t_0
P^l	Matrix of the observed choices in the log data in $]t_{l-1}, t_l]$
$\hat{P}^l$	Matrix of the simulated choices in $]t_{l-1}, t_l]$
c_l	List of the users who make simulated choices in $]t_{l-1}, t_l]$
M_l	Size of c_l
$\hat{Q}^l$	Matrix of the observed choices in $]t = 0, t_0]$ and simulated choices in $]t_0, t_l]$
t_{uj}	The time point of the observed user u choice of the item j
A^l_u	Awareness set of u in $]t_{l-1}, t_l]$
A	Size of all the users' awareness set
v_{ui}	Utility of the user u for choosing item i
δ	Multiplicative factor increase of utility of a recommended item
c	Scaling factor for transforming a rating into a utility value
k	Number of recommendations

effects with respect to two diversity metrics: long-tail diversity and unexpectedness (Vargas, 2015). Chaney et al. (2018) have also performed a simulation study of users' choices, where, similarly to our study, they aim at understanding the influence of alternative RSs on users. They simulate users' choices within consecutive time intervals with six alternative RSs that are retrained at the end of each time interval based on simulated choices. By measuring users' choice diversity, they show that RSs' feedback loop causes homogeneity of users in their preferences. However, unlike our simulations, their utility function is based on synthetic data and they assume that all the users make the same number of choices in the simulation.

Another important simulation study is described in Szlávik et al. (2011). The authors simulate six choice models and one RS (Funk, 2006). The choice models simulate users with different acceptance probability of the recommendations, namely, the simulated users may only select a recommended item or could also select other items. Choices are simulated for consecutive time intervals, and, similarly to our approach, they retrain the RS at the end of each time interval. They measured multiple metrics representing users' choice diversity. Comparing this study to our work, while Szlávik et al. (2011) aims at comparing the effect of "alternative choice models" on the choice behaviour, we aim at understanding the effect of "alternative RSs" on the choice. Moreover, Szlávik et al. (2011) do not consider the users' awareness of the catalogue, i.e., their simulated users can choose any item in the catalogue, which is not realistic. However, similarly to our study, they use a data set of logged users' choices in their simulation.

In a more recent work Yao et al. (2021) simulated users' choices, with alternative choice models. Similarly to our framework, they consider the users' awareness and their utility function is based on a real choice data set. But, their goal is to analyse RSs effect on items' popularity distribution, and therefore they considered different degrees of users' preferences for choosing popular items.

As we have noted above, the cited simulation studies make alternative assumptions on the users' preferences, choice models, items and frequency of choices. Hence, comparing the obtained results is by far easy and straightforward. We further discuss some of the differences between the finding of the previous studies and our findings in Section 7.

4. Simulation framework

Our target scenario is an information system where there are users and items. Users repeatedly select items over time. Here, an RS is supposed to support users' decision making. We assume that it is possible to access the log data of the system, i.e., ratings given to items, and in this way it is possible to estimate the utility of the items for the users. We want to build a simulation process that, starting from an initial set of system log data, the data present in the log up to a certain point in time, simulates the subsequent choices made by the users after that point in time, also simulating the effect of the RS, providing recommendations to the simulated users. Moreover, we would like to compare the simulated choices with the actual choices that are present in the log data set, after the time point when the simulation started. This can be used to study and understand the effect of an RS when is deployed in the target information system.

We denote with U the full set of users and I the full set of items that appear in the system data log. The most important symbols and notations used in this paper are listed in Table 1. The users' ratings for the items, which are in system log, are assumed to be stored in a $|U| \times |I|$ matrix R and r_{uj} indicates the logged rating of the user u for the item j. By using a rating prediction method (IPS-MF) we also predict missing ratings in R and we compute $\hat{R}$, where $\hat{r}_{uj}$ indicates the predicted rating of the user u for the item j. The predicted ratings provide the estimated utility of the items (see Fig. 1) that is used in the simulation of the users' choices. More details on rating prediction are given in Section 4.4.

Let P be the $|U| \times |I|$ matrix of system logged choices, where an element of this matrix p_{uj} is 1 if the user u, in the system log data, has chosen the item j and $p_{uj} = 0$ otherwise. We assume that ratings in the system log data are given to items that the users have chosen. Hence, we derive the matrix P of logged choices from the matrix R. Moreover, the logged users' choices are time-stamped: t_{uj} is the time when the user u chose item j. Assume that t_0 is a selected time point that is the time point of the start of the simulation; we denote with P^0, the initial choice matrix, formed by all the logged true choices p_{uj}, s.t. $t_{uj} \leq t_0$. The simulation procedure starts from this initial knowledge and aims at simulating users' choices made after this time point under alternative simulation conditions. We consider successive time intervals, after one or more months, from the time point t_0. For instance, $]t_0, t_1]$ denotes the time interval spanning from t_0 (excluded) to t_1 (included), t_1 is a time point one month after t_0 and P^1 is the matrix containing the observed, in the log data, choices in that interval. The simulation iterates on these time intervals to identify $\hat{P}^l$, which is the matrix of the simulated choices in $]t_{l-1}, t_l]$, and the aggregated simulated choices $Q^l = P^0 + \hat{P}^1 + \cdots \hat{P}^l$. We assume that in each time interval $]t_{l-1}, t_l]$, the users choose items one after another. A simulated user's choice involves three computational steps (cf. Fig. 1):

1. ***Awareness set***: The user's awareness set is built and contains the alternative options that the simulated user may actually choose (Section 4.1).
2. ***Recommendation***: An RS suggests some items to the user ad extends the awareness set. Moreover, the RS influences the user and this is implemented by increasing the estimated utility of an item. This also increases the probability that a recommended item is chosen, compared to the case when the item is present in the awareness set but is not recommended (Section 4.2).
3. ***Choice***: The user makes a choice among the items in the awareness set by using the choice model: a (stochastic) function of the estimated utilities of the items in the awareness set. The estimated utilities are proportional to the estimated ratings $\hat{r}_{ui}$ (Section 4.3).

4.1. Awareness set

As we mentioned earlier, we make the reasonable assumption that users are not aware of the entire catalogue of the items and can only choose items from a personalised subset of the catalogue called *awareness set*. The awareness set A_u^l contains the items that the user u can choose at the lth time interval, $]t_{l-1}, t_l]$. In the simulation described in this paper, A_u^l is assumed to be of a target size $|A_u^l| = A$, which is the same for all the users $u \in U$, e.g., it may contain 500 items of the catalogue. A_u^l includes the top A items in two ranked lists, Pop_u and Hut_u: the most popular items and the items with the largest predicted utility (rating), plus a small percentage of random items.

- Pop_u contains the items, which have not been chosen by u before, sorted with respect to their popularity. The popularity of an item is equal to the number of times the item was chosen by the other users in the previous time intervals. Hence, we make the assumption that users are more likely to know the popular items in the system (Sampaio et al., 2006; Teixeira et al., 2002).
- Hut_u contains the items, not chosen by u before the current time interval, sorted with respect to their predicted rating (or equivalently, utility). The predicted ratings in $\hat{R}$ are computed by using the IPS-MF model (see Section 4.4). In order to correctly simulate the users' preferences, the full knowledge of the system logged data is used in this prediction.

These two ranked lists are then combined by using the Borda count aggregation method (Saari, 1985). In Borda count, the ranks of the voters (Pop_u and Hut_u in our case) are converted into scores: the highest rank gets the largest score. Then the items are ranked based on the sum of their Pop_u and Hut_u scores. The top $\alpha * A$ items from this aggregated list are included in the awareness set ($\alpha \in [0, 1]$). Then, the remaining $(1 - \alpha) * A$ elements of the awareness set are random items. Random items are added to the awareness set because real users are actually unpredictable and do not only choose among popular and with high estimated rating items. We set α to 90%, so, most of the awareness set is composed of top items in the lists Pop_u and Hut_u.

We also assume that a user does not choose an item more than one time. In fact, the data sets that we consider are for products (books, games and apps) that are typically purchased only once. Hence, an item j is removed from u's awareness set after it is selected once, so that it cannot be selected twice.

4.2. Recommender systems

An RS generates for each user u in a time interval l recommendations that are added to the user awareness set A_u^l. The number of recommendations k is a parameter that is varied in the simulations (Section 5.3). The RS, to simulate a realistic scenario, computes the recommendations only on the base of the initial set of ratings, before time t_0 and the successive simulated choices of the users. Conversely, the ratings, $\hat{R}$, which are predicted by the IPS-MF model (see Section 4.4) are only used to simulate the users' choices.

- PCF — Popularity-based Collaborative Filtering: is a neighbourhood-based CF RS that computes the cosine similarity between the 0/1 choices' vector of a target user u, q_u^{l-1} and the choice vector of the other users to find the nearest neighbours. The top-k popular items among the choices of the nearest neighbour users are recommended to the target user.
- LPCF — Low Popularity-based Collaborative Filtering: is similar to PCF, but it penalises the score computed by PCF by multiplying it with the inverse of their popularity. The top-k scored items are recommended.
- FM — Factor Model: is a Factor Model (FM) RS which generates recommendations following the approach proposed in Hu et al. (2008).

- POP — Popularity-based: The top-k popular items in terms of the number of times that they were selected by the users in the past are recommended.
- AR — Average Rating: The items are scored with a variation of the average predicted rating. A weighted average is calculated for each item as follows:

$$wr = \frac{v}{v+m} \times R + \frac{m}{v+m} \times B \tag{1}$$

Where R is the average rating for the item, v is the number of times that this item is rated, m is the minimum number of ratings required to be considered by the RS, B is the average of all of the ratings in the data set. The top-k scored items are recommended.

Additionally, we simulate a scenario, called NO-RS, where users do not receive any recommendations; they only choose items in their awareness sets.

Finally, it is worth noting that when we simulate the users' choices in each target time interval, the personalised RSs (PCF, LPCF and FM) do not compute recommendations for the *new users*, i.e., those who enter the system with their first choice on that month. New users are then simulated to make choices only among the items in their awareness sets. Obviously, new users, instead, always can receive recommendations generated by POP and AR since these two RSs are not personalised.

4.3. Choice model

During a time interval, a simulated user is given the chance to make some choices for items according to a multinomial logit model (Anas, 1983; Fleder & Hosanagar, 2007, 2009; Hazrati et al., 2020). We decided to apply the same choice model that was used in past experiments to make our results comparable to these ones; alternative choice models can be considered in the future.

When a user u makes a simulated choice she estimates the utility v_{ui} of the items $i \in A_u^l$ and then makes a choice with the following probability:

$$p(u \ chooses \ i) = \frac{e^{v_{ui}}}{\sum_{j \in A_u^l} e^{v_{uj}}} \tag{2}$$

Hence, items with larger utility are more likely to be chosen, but the user does not necessarily select the item with the largest utility. This assumption tries to take into account the potential errors introduced by the estimation of the utility and the fact that no decision maker is perfectly rational, i.e., follows the standard utility maximisation model (von Neumann & Morgenstern, 1953). The utility of an item i for the user u is proportional to the estimated rating of user u for item i:

$$v_{ui} = (\hat{r}_{ui} - 1) * c \tag{3}$$

In fact, ratings in the considered log data sets are in the $[1, 5]$ scale, hence utility is converted into a $[0, 4 * c]$ scale. The parameter c adjusts the influence of the rating in the choice model: the larger is c the more likely the user will be to choose an item with larger predicted rating. The value of c will be discussed later in this paper when we will describe the three log data sets that we have used in our experiments (Section 5.3).

As we have already mentioned, if an item i is recommended to the user u by an RS, it is added to the awareness set of u. But, in addition to that, the utility v_{ui} is increased (boosted) by a multiplicative factor $\delta > 1$, before the choice is made. Hence, the utility v_{ui} of a recommended item i is actually as follows:

$$v_{ui} = \delta * (\hat{r}_{ui} - 1) * c \tag{4}$$

Hence, the recommended item becomes marginally more likely to be chosen by the user, compared to an item with the same utility, but not recommended.

4.4. IPS-MF model for rating prediction

In order to predict an item's rating $\hat{r}_{ui}$ and then its utility v_{ui}, we use the full set of ratings actually present in the considered system log data set. However, in general, the observed ratings are subject to selection bias, and therefore any rating estimation based on observed ratings is also biased. For example, in a movie website, users typically watch and rate those movies that they like and rarely rate movies that they do not like (Pradel et al., 2012), unless they made a bad choice and watched a movie that they did not like. This produces a situation where data are said to be "Missing Not At Random" (MNAR) (Marlin & Zemel, 2009; Schnabel et al., 2016). MNAR data is typically subject to positivity bias, which happens because higher ratings are over-represented; thus, the average rating is skewed upwards (Boratto et al., 2021; Huang et al., 2020; Pradel et al., 2012; Schnabel et al., 2016; Yalcin & Bilge, 2021). Therefore, in order to debias rating predictions computed with the available data, we use *Inverse Propensity Score* Matrix Factorisation (IPS-MF) (Schnabel et al., 2016). IPS-MF modifies the loss function of a matrix factorisation model by taking into account the inverse probability of a user rating an item.

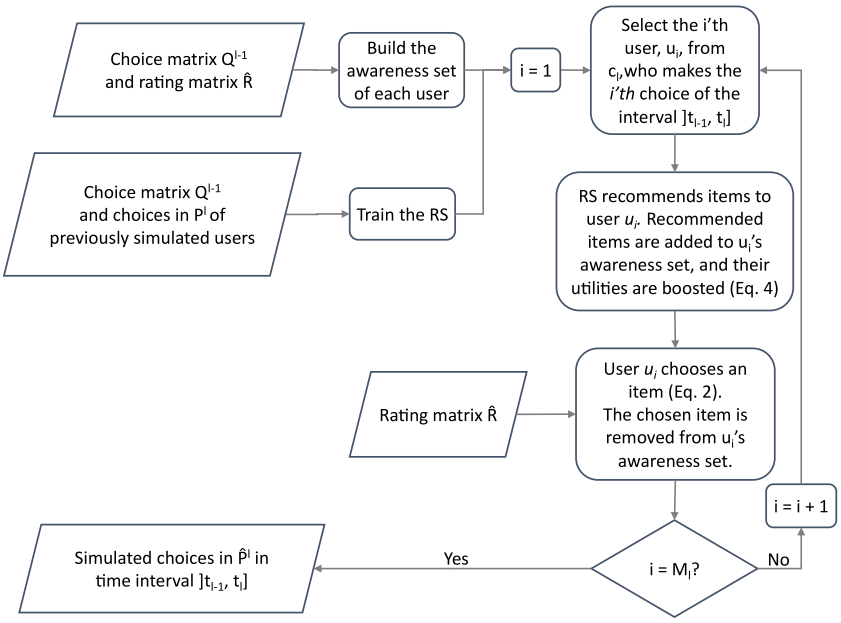

Fig. 2. Simulation procedure in time interval $]t_{l-1}, t_l]$ (M_l choices).

4.5. Simulation procedure

We summarise here the complete simulation procedure, which makes use of the modelling assumptions described before. In each time interval $]t_{l-1}, t_l]$, $l = 1, \ldots, L$, M_l choices are made by the users, where M_l is equal to number of choices observed in the lth time interval in the system log data. Moreover, the simulation considers five alternative RSs (PCF, LPCF, FM, POP, and AR) and generates recommendations before a user makes a choice.

Fig. 2 shows the procedure of simulating one time interval's choices (e.g., in the interval $]t_{l-1}, t_l]$). At the beginning, one of the RSs is trained and the awareness sets of all the users are built. Then, we insert, in a random way, the users (with repetitions) in the list $c_l = (u_1, \ldots, u_{M_l})$. Each user can appear in c_l several times: this is equal to the number of times she made a choice, as it is recorded in the log data set, in the considered time interval $]t_{l-1}, t_l]$. The list c_l gives the (inessential) order in which the simulated users are going to make their choices in this interval. Before a choice for user u_i is simulated, the recommended items for u_i are added to the awareness set $A^l_{u_i}$ and their utilities are boosted according to Eq. (4). At this point, a choice for u_i is simulated. Then, if i is not already equal to M_l, a new choice for user u_{i+1} is simulated. After all the M_l choices in $]t_{l-1}, t_l]$ are simulated, these choices are inserted in $\hat{P}^l$ and the full simulation procedure continues to the next time interval and fills the matrix $\hat{P}^{l+1}$, until all the L time intervals are considered (10 in our simulations).

5. Simulation experiment design

5.1. Evaluation metrics

We are interested in measuring the diversity and the quality of the simulated choices. Both are essential properties of the users' choices. Knowing the effect of an RS on these metrics can inform the decision to utilise a specific one in a real application. We introduce here metrics that capture these properties.

5.1.1. Gini index

The diversity of the users' choices is measured with the *Gini index*, which is often used to quantify inequality and has been previously adopted in related studies to measure sales diversity (Adamopoulos et al., 2015; Fleder & Hosanagar, 2007, 2009; Lee & Hosanagar, 2019; Matt et al., 2013; Szlávik et al., 2011). Dorfman (1979) discussed the Gini index in detail. A higher Gini index indicates lower choice diversity. For instance, a Gini index equal to 0 is obtained when the choices are perfectly uniformly distributed, i.e., all the items have been chosen the same number of times. Conversely, a Gini index close to 1 signals that the majority of the choices are concentrated on a very small set of items. We observe that diversity is considered a positive feature and, actually, the lack of diversity is perceived as an adverse effect of RSs (Fleder & Hosanagar, 2007; Matt et al., 2013; Szlávik et al., 2011).

5.1.2. Choice coverage and recommendation coverage

Choice Coverage captures another critical aspect of the users' choice diversity; it is the percentage of the items in the catalogue that have been *chosen* at least once by some user. Choice Coverage can show the ability of an RS in recommending the full potential set of available items. We note that Choice Coverage reveals a different facet of choice diversity, compared to the Gini index. While Choice Coverage captures the users' spread of choices over the catalogue, the Gini index measures how uniformly the choices are distributed over all the chosen items. Besides Choice Coverage, Recommendation Coverage measures the fraction of items present in the system that were *recommended* at least once during the simulation. This metric shows the ability of the RS to recommend diverse items.

5.1.3. Choice's rating: Average predicted rating of the chosen items

The average predicted rating of the chosen items, which is also called in a more compact way, the Choice's Rating, is the mean of the (IPS-MF predicted) rating of the users' *chosen* items. We compute the average of the predicted rating of the chosen items for each user and then we average over all the users' mean values. We are interested in this metric to understand if an RS helps the users to identify the more valuable items; in fact, the predicted rating of a user for an item is the only measure that we have at our disposal to assess the quality of the choices.

5.1.4. Popularity

This metric measures the popularity of the chosen items. For every time interval (month) of simulated choices, we calculate the average popularity of the chosen items at the time that they were chosen. The Popularity of an item at a certain time-point is equal to the number of times the users have chosen this item from the beginning of the log data set (t=0) until that time point, divided by the overall number of choices from$t = 0$ until that time point.

5.1.5. Average predicted rating of the recommended items

This metric measures the mean of the (IPS-MF predicted) rating of the recommended items. This average predicted rating measures the quality of the recommendations. It is expected, given the adopted choice model, that the users choose the recommendations more often if they have high predicted ratings.

5.1.6. Recommendation acceptance

Recommendation Acceptance measures how often the simulated users make choices among the recommended items. This metric is computed by dividing the number of times the users have chosen a recommendation, by the number of times they have received some recommendations before making their choices. We recall that the users are choosing items in the awareness set and only some of these items are actually recommended. A higher acceptance ratio indicates a kind of *effectiveness* of the RS. Conversely, a low Recommendation Acceptance signals that the recommended items have low utilities (predicted ratings); otherwise, they would have been selected. This descends directly from the adopted choice model.

5.2. System log data sets

We searched for time-stamped choice and rating data sets that could be appropriate for our study.[1] A data set contains the log data of an information system and our simulation, starting from an initial subset of this data, simulates the next choices of the users. The actual choices of the users, contained in the log data set, are used to compare the simulated recommendation sessions with the observed users' behaviour. Moreover, the log data set is used to predict the simulated users' utilities for items, actually, their ratings and the users' awareness sets.

We have then selected three log data sets in the *Amazon* collection: Amazon $Kindle$, Amazon $Games$ and Amazon $Apps$ (He & McAuley, 2016). The reason for choosing Amazon data sets is that in the Amazon eCommerce platform, books, games and apps are typically rated after they are bought. Hence, the logged ratings by the users correctly signal their actual choices. Moreover, the chosen data sets are recording ratings for rather similar product types (mobile applications, Kindle electronic books and video games). This helps to make the results more easily comparable and enables us to better generalise the obtained results. Additionally, users in these three domains do not make repeated choices (purchases) for a single item; this is compliant with our simulation model.

However, the Amazon data sets contain many users that made very few purchases. These users are hard to be modelled and their choices hard to be correctly simulated. Hence, we decided to consider only users who made at least 20 choices and gave the corresponding ratings. Amazon $Apps$ data set contains users' ratings for Android applications over 52.5 months. We consider the choices collected in the first 42.5 months as starting observed data for building the RSs (time t_0) and then we simulated the last 10 months' choices. $Games$ data set contains users ratings given to Video games over 179 months. Similarly to the $Apps$ data set, we simulate the choice of the last 10 months.

The third data set, $Kindle$, which contains time-stamped ratings of users for books, is much more sparse than the previous twos. Hence, the analysis conducted on $Kindle$ is essential to understand the impact of data sparsity on recommendation and choice simulation. This data set is more sparse in the first 12 years (99.93% sparsity), hence, we decided to consider only these first 12 years to simulate choices in a sparse data set.

[1] We have evaluated: http://jmcauley.ucsd.edu/data/amazon/, MovieLens data sets and CiaoDVD from ciao.co.uk

Table 2
Important properties of the considered log data sets.

Property	*Apps*	*Games*	*Kindle*
# of ratings	154,033	80,305	28,346
# of users	4,579	2,061	2,575
# of items	23,921	20,060	16,017
# of time intervals (months)	52.5	179	144
Average number of ratings per user	33.63.	38.96	11.00
Sparsity	99.8%	99.8%	99.93%
# time intervals of simulated choices (n)	10	10	10
Average rating in the data	3.77	3.99	4.20
Average predicted rating	2.91	3.10	2.67
Gini index in the last n months	0.78	0.87	0.53
Coverage of the last n months choices	0.60	0.25	0.70
Average rating in the last n months	3.79	4.01	4.20
Average number of choices per month	3796.9	792	1572.5
New users' choices	9%	7%	17%

Table 3
Baseline and variation of the simulation parameters for each log data set. The values in the parenthesis are lower and higher values for each baseline value that we used in order investigate the impact of each parameter.

		Parameter			
		Utility scaling: c	Multiplicative factor: δ	Awareness set size: A (Variation)	Recommendation set size: k (Variation)
Data set	Apps	0.75	2	3000 (1000, 5000)	50 (10, 100)
	Games	0.7	2	2000 (1000, 3000)	50 (10, 100)
	Kindle	0.8	2	2000 (1000, 3000)	50 (10, 100)

Table 2 summarises the main characteristics of the considered data sets (after the above mentioned filters are applied). In this table, n shows the number of simulated months, which is 10 in this paper for all the data sets. *# of time intervals (months)* is the time difference (in months) between the last and the first rating recorded in the data set. *Average number of ratings per user* is the average number of ratings/choices of each user over the entire log data set. *Average rating in the data* is the average of the observed ratings in the whole log data set. *Average predicted ratings* is the average of all the predicted ratings using the unbiased IPS-MF model. *Average rating in the last n months* is the average of the observed ratings in the last n (10) months of the data set. *Average number of choices per month* is the number of choices a user has made on average in one single month during the last n months of the data set. *New users' choices* is the percentage of the choices made by the new users in a single month. This value is an average over all the last 10 months of the data set.

5.3. Simulation parameters

The proposed simulation depends on four important parameters (see Table 3). We first identified a baseline configuration of the parameters for each data set. Then we ran simulations varying the values of two of these parameters. To identify the baseline configuration, we run the simulation procedure when the FM recommender system was used. This RS is supposed to give the best recommendations among the five considered RSs. Using FM, we searched for parameter values that make the simulated choices distributed as close as possible, with respect to Gini index and Choice Coverage, to the observed choices in the data set.

The parameters δ and k take the same values in the three data sets. While the coefficient c, used to scale the utility, and the awareness set size A have been tuned in each data set. The utility coefficient must be tuned because of the differences in the distribution of the ratings in the data sets (see the Average predicted rating in Table 2). We set the utility coefficient in such a way that, after that normalisation, (1) the average utility of all the user–item pairs are similar in the data sets and (2) the Gini index and Choice Coverage are similar to the observed Gini index and Choice Coverage. Moreover, we tuned also the awareness set size in each data set to achieve the baseline Gini index and Choice Coverage values. The $Kindle$ data set has a very high sparsity and diversity. Hence, even with a small utility scaling coefficient and a large awareness set size, we could not come close to the observed Choice Coverage. Hence, we decided to keep parameters values similar to those used in the other data sets in order to better analyse the impact of data sparsity.

In order to understand the role of the *awareness set size* and *recommendation set size* on the users' choice behaviour, we have run the simulations with lower and higher values of these two parameters, as shown in Table 3.

6. Experimental results

Before discussing the experimental results, it is worth explaining how the performance metrics are presented. In order to address the first and the second research questions, which require the analysis of the *evolution* of the users' choices, we calculate each metric over the set of simulated users' choices from the beginning of the simulation period until the end of each month of simulated choices.

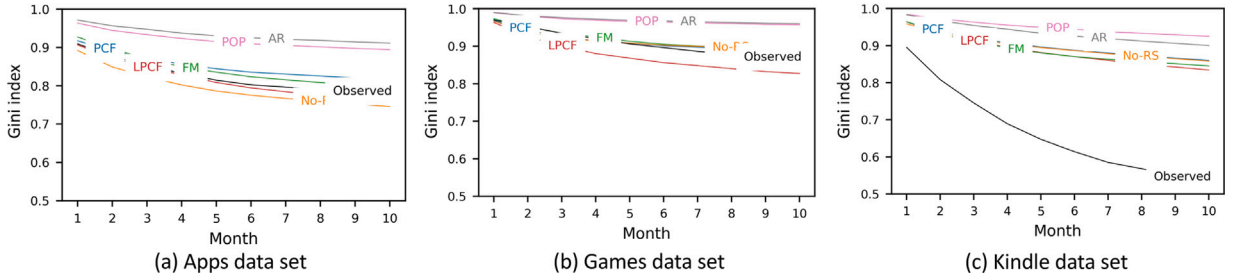

Fig. 3. Evolution of the Gini index of the observed and simulated choices under the effect of the five considered RSs.

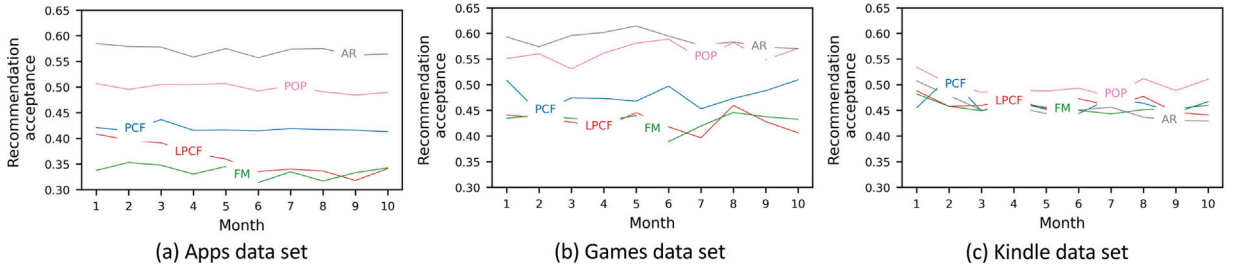

Fig. 4. Evolution of Recommendation Acceptance over the simulated choices under the effect of the five considered RSs.

For instance, in Fig. 3(a), the x axis shows the month of simulation and the y axis shows the values of a metric calculated over all the simulated choices up to the end of the month x. Hence, for instance, at month $x = 4$ it is shown the Gini index calculated over the accumulated choices simulated in months 1, 2, 3 and 4.

We also calculate the performance metrics on the observed choices, i.e., those stored in the log data sets in the considered months of the simulation. This enables us to compare our simulation results with logged users' choices. Even though we do not know the exact conditions that have determined the choices logged in the data sets and if, for instance, an RS was used, the metrics computed on logged data can be used as a qualitative reference point for better assessing the impact of the considered RSs. The label "Observed", which is used in the graphs, refers to a measured metric on the logged choices observed in the data set, while the other curves are relative to the simulated choices.

Additionally, to address the research questions 3 and 4, we analyse the calculated metrics over all of the choices, i.e., from the first to the last simulated month of choices. These information is presented in Tables 5 and 6.

6.1. Recommender systems' effect on choice diversity: Gini index

We start by addressing the first research question: "RQ1 - How personalised and non-personalised RSs affect the evolution of choice diversity? What features of the RSs determine their specific impact?"

The evolution of the Gini index is shown in Fig. 3. We observe a clear tendency of the choices to grow in diversity with time: the Gini index is always monotonically decreasing. However, comparing the Gini index in the presence and absence of RSs, we note that it is very difficult that an RS can produce a higher diversity than when no RS is used. In fact, only in Fig. 3(c) (*Kindle* data set), LPCF and FM produce slightly smaller Gini index values than the baseline case, called NO-RS, while in Fig. 3(b) (*Games* data set), only LPCF has a lower Gini index. Nevertheless, considering the comparison of either LPCF or FM with the NO-RS case in the three data sets, we can conclude that the evolution of the Gini index in the presence or the absence of an RS strictly depends on the log data.

Another evident observation is that in all of the considered data sets, the personalised RSs (PCF, LPCF and FM) produce lower Gini index values compared to the non-personalised ones (AR and POP). This means a higher choice diversity for the personalised RSs. This is more evident in *Apps* and *Games*. This effect is clearly motivated by the fact that the non-personalised RSs make the same recommendations to all the users, while the personalised RSs adapt to each individual user's profile. However, there is another reason why the Gini index is larger for non-personalised RSs: the Recommendation Acceptance of the non-personalised RSs is higher than for the personalised RSs, especially in the *Apps* and *Games* data sets (see Fig. 4). Hence, the choices influenced by the non-personalised RSs are more often made among the narrower set of recommended items and not among the other items in the awareness set.

When discussing choice diversity, it is worth considering Choice Coverage (see Fig. 5), which gives another perspective on the broad dimension of diversity. Personalised RSs produce choices that cover a larger part of the catalogue compared to the non-personalised RSs. Hence, the choices determined by the personalised RSs are covering a larger spectrum of items (Choice Coverage) and they are distributed more uniformly on those items (the Gini index).

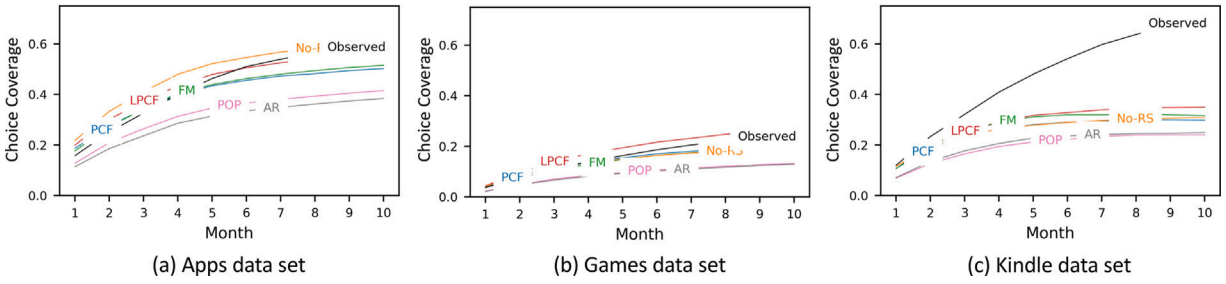

Fig. 5. Evolution of Choice Coverage over the observed and simulated choices under five different RSs.

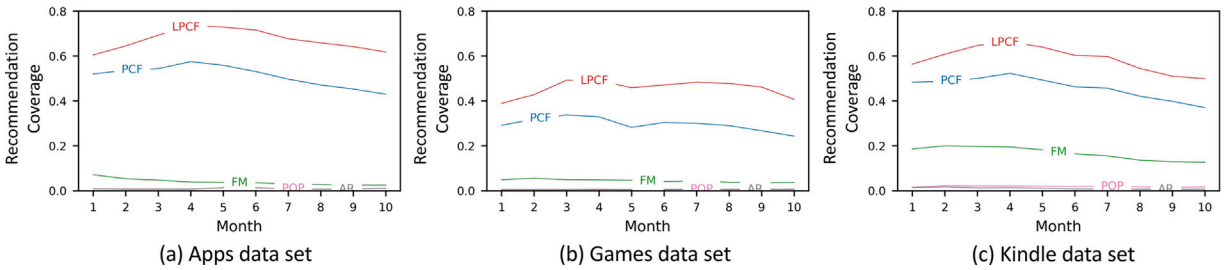

Fig. 6. Evolution of Recommendation Coverage over the simulated choices under the effect of the five considered RSs.

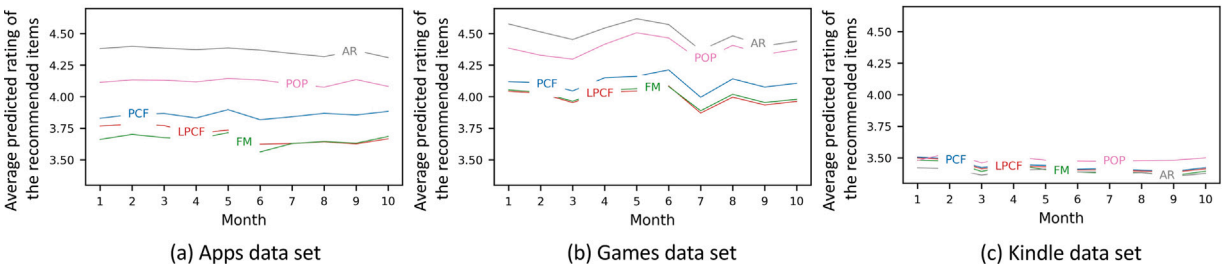

Fig. 7. Evolution of the average predicted rating of the recommended items over the simulated choices under the influence of the five considered RSs.

Focusing now on the performance differences between the personalised RSs (PCF, LPCF and FM), we can observe that LPCF results in the lowest Gini index (highest diversity). We note that LPCF, by definition, recommends less popular items than the other two personalised RSs. Hence, LPCF is somewhat expected to lead to more diverse choices. This is confirmed by the Recommendation Coverage results (see Fig. 6): LPCF has the highest Recommendation Coverage, i.e., the recommended items (but also the choices, as shown in Fig. 5) cover a larger part of the catalogue of recommendable items than the other RSs.

Among the non-personalised RSs, POP and AR have slightly different effects on the Gini index. While in the *Apps* data set, POP has a lower Gini than AR, both POP and AR have similar Gini index values in *Games*. However, in *Kindle*, AR produces a lower Gini index. The high Gini index of the non-personalised RSs can also be determined by a high (predicted) rating of the recommendations: when the recommendations have a large *utility*, the users are more likely to choose them, i.e., there is a high Recommendation Acceptance, which can result in a high frequency of choices over these items, which gives a high Gini index. In fact, we observe in Figs. 3, 4 and 7 that this relationship exists. For instance, we observe in Fig. 7(a) that AR's recommendations have higher average predicted ratings and consequently, a high Recommendation Acceptance (see Fig. 4(a)). This results in a higher Gini index compared to POP (Fig. 3(a)). In the other two data sets, we observe similar relations between the predicted ratings of the recommendations and the Gini index.

The analysis of the Popularity of the chosen items, shown in Fig. 8, confirms the observations made while discussing the Gini index and the Choice Coverage. Even though this metric shows a somewhat different evolution in the three considered data sets, it clearly indicates that the non-personalised RSs tend to recommend more popular items and this bias is increased month by month.

6.2. Personalised vs Non-personalised recommender systems effect on users' choice's rating

We now focus on the second research question: "RQ2 - Do personalised RSs suggest items that users rate higher than non-personalised RSs?". In fact, personalised RSs are considered more accurate and preferable compared with non-personalised RSs, hence, they are expected to produce better choices.

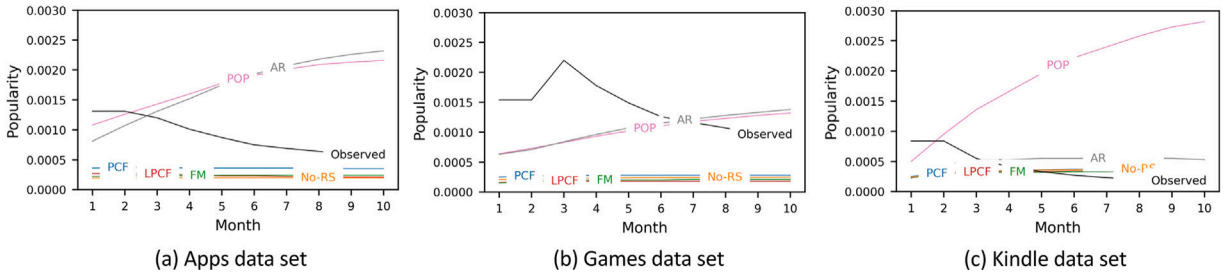

Fig. 8. Evolution of Popularity of the chosen items for the observed and simulated choices under the influence of the five considered RSs.

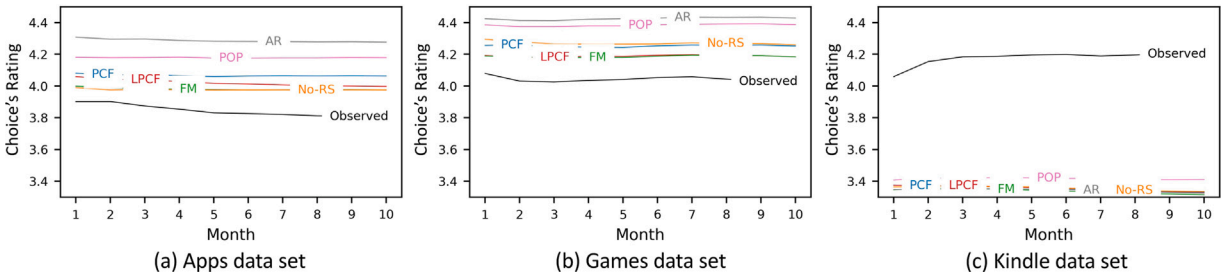

Fig. 9. Evolution of the Choice's Rating (average predicted rating of the choices) over the observed and simulated choices under the influence of the five considered RSs.

Table 4
Precision of Recommender Systems in the classic evaluation scenario.

		Data set		
		Apps	Games	Kindle
Recommender System	PCF	66.5	73.7	80.1
	LPCF	66.5	73.2	80.1
	FM	**67.2**	**75.4**	**80.3**
	POP	67.0	74.7	80.1
	AR	67.0	72.9	80.1

The Choice's Rating metric is the average (IPS-MF) predicted rating of the chosen items. As we previously mentioned, it tells us how good are the users' choices. The evolution of these metrics is depicted in Fig. 9.

We observe that the non-personalised RSs produce choices for items that are better evaluated by the users: the Choice's Rating metric is higher compared to the more sophisticated personalised RSs, such as, FM. This result is matching the previously discussed result on the average predicted ratings of the recommendations (Fig. 7): not only the user choices influenced by non-personalised RSs, but even the recommendations of these RSs have a higher average predicted rating, compared to the personalised ones. This clearly contrasts with common knowledge, i.e., that personalised RSs are more accurate and therefore, should recommend items that the users rate higher. In order to clarify this aspect, we have measured the precision of the considered RSs.

Precision of the recommender systems. Table 4 shows the precision of the considered RSs on our three data sets. We include in this evaluation only users with 20 or more recorded ratings. We split the available ratings into 75% training and 25% testing data. The split is performed based on the time-stamps of the ratings, that is, 25% of the last recorded ratings of each user is considered as her test ratings. We calculate the precision over the top-5 recommendations for each user and we show here the average of these values.

Overall, non-personalised RSs perform slightly better than PCF and LPCF and slightly worse than FM. Hence, in these data sets, non-personalised RSs are actually also rather precise, with a marginally small advantage for FM. It is therefore clear that these differences in system's precision are providing only a partial indication of RSs' quality and impact: other factors, which are studied in this paper, influence the system's performance, such as the variety and the predicted ratings of recommendations.

Moving back to the simulation results, an additional interesting observation is that in the *Apps* and *Games* data sets, the Choice's Rating metric computed on the observed choices (those in the log data) takes slightly smaller values than those obtained by the personalised RSs, especially FM. While in *Kindle*, the Choice's Rating metric is much higher on the observed choices than on all the simulated choices, whatever is the RS that influenced these choices. Moreover, according to Fig. 7, also the recommendations in *Kindle* have a smaller average predicted rating (roughly 3.40) compared to the average of the observed ratings of the last 10 months of simulated choices (4.20). Instead, in the *Apps* data set, the predicted rating of the recommendations of LPCF and FM

Table 5
Impact of the users' awareness set size on the considered users' choices related metrics.

	RS	Apps			Games			Kindle		
		Awareness set size			Awareness set size			Awareness set size		
		1000	3000	5000	1000	2000	3000	1000	2000	3000
Gini	PCF	0.90	0.82	0.73	0.92	0.89	0.86	0.90	0.86	0.82
	LPCF	0.81	0.76	0.69	0.83	0.83	0.81	0.87	0.83	0.80
	FM	0.90	0.80	0.71	0.92	0.89	0.86	0.88	0.85	0.81
	POP	0.97	0.89	0.82	0.98	0.96	0.93	0.97	0.93	0.88
	AR	0.98	0.91	0.84	0.98	0.96	0.93	0.94	0.90	0.86
Coverage	PCF	0.39	0.50	0.60	0.17	0.20	0.23	0.26	0.30	0.34
	LPCF	0.50	0.56	0.63	0.27	0.27	0.29	0.32	0.35	0.38
	FM	0.35	0.52	0.61	0.15	0.20	0.24	0.28	0.32	0.35
	POP	0.21	0.42	0.53	0.08	0.13	0.18	0.15	0.24	0.31
	AR	0.18	0.38	0.51	0.08	0.13	0.17	0.18	0.25	0.30
Popularity	PCF	0.00063	0.00035	0.00025	0.00037	0.00028	0.00024	0.00054	0.00037	0.00029
	LPCF	0.00033	0.00021	0.00015	0.00020	0.00018	0.00016	0.00047	0.00034	0.00027
	FM	0.00041	0.00024	0.00017	0.00027	0.00021	0.00018	0.00046	0.00034	0.00028
	POP	0.00382	0.00216	0.00152	0.00178	0.00132	0.00103	0.0048	0.00282	0.00197
	AR	0.00382	0.00232	0.00162	0.00188	0.00138	0.00107	0.00081	0.00053	0.00039
Choice's Rating	PCF	4.14	4.06	4.00	4.27	4.25	4.23	3.34	3.33	3.33
	LPCF	4.03	4.00	3.95	4.17	4.18	4.18	3.34	3.33	3.32
	FM	4.00	3.98	3.94	4.17	4.18	4.18	3.32	3.32	3.31
	POP	4.28	4.18	4.11	4.44	4.39	4.35	3.44	3.41	3.39
	AR	4.40	4.28	4.19	4.48	4.43	4.39	3.32	3.33	3.32
Recommendation Coverage	PCF	0.38	0.43	0.50	0.24	0.24	0.26	0.33	0.37	0.42
	LPCF	0.59	0.62	0.68	0.40	0.41	0.42	0.42	0.50	0.54
	FM	0.03	0.03	0.03	0.03	0.04	0.04	0.12	0.13	0.13
	POP	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02
	AR	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Recommendation acceptance	PCF	0.67	0.41	0.31	0.64	0.51	0.37	0.63	0.47	0.37
	LPCF	0.57	0.34	0.24	0.58	0.41	0.35	0.61	0.44	0.36
	FM	0.58	0.34	0.24	0.55	0.43	0.33	0.61	0.46	0.35
	POP	0.72	0.49	0.40	0.72	0.57	0.46	0.66	0.51	0.42
	AR	0.77	0.56	0.45	0.74	0.57	0.50	0.62	0.43	0.37

(around 3.70) is similar to the average rating of the observed choices (3.79). This clearly shows that all the considered RSs struggle to find good recommendations in the *Kindle* data set and consequently poor choices (i.e., low Choice's Rating) are influenced by these RSs. This is another analysis that shows how important is a simulation experiment in order to understand the impact of an RS in a specific data set.

6.3. Impact of the awareness set size on choice diversity and choice's rating

We now address the third research question: "RQ3 - Does a better users' awareness of the catalogue of the items, i.e., being aware of a larger number of items, lead to better choices, that is, higher users' rating for the choices?" To address this research question, we run the simulation procedure for different awareness set sizes, see Table 5. In order to simplify the analysis of the results, we show here the metrics calculated on all the simulated choices (from the first month of the simulation to the last one, i.e., month 10).

By increasing the awareness set size, as expected, the users' choices become more diverse: the Gini index and the Popularity metrics both decrease, while choices cover a larger part of the catalogue, as Choice Coverage increases. However, the Choice's Rating, i.e., the users' satisfaction with the choices, tends to decrease. Moreover, Recommendation Acceptance decreases as the awareness set size increases. Accordingly, we can state that when users make fewer choices among the recommended items, because the range of options among which they can choose expands (awareness set size), then the satisfaction for the choices decreases. Hence, RSs can actually help users to find good (high rated) items, especially when users have a smaller knowledge of the items' catalogue. According to our simulation, increasing the users' awareness about the catalogue can be effective for diversifying the choices and mitigating the concentration bias introduced by the RS. But, higher awareness sets leads to worse choices.

As it was mentioned earlier, with a higher awareness set size, Recommendation Acceptance decreases. One can then expect that a decrease of the Recommendation Acceptance is associated to a decrease of the Recommendation Coverage, because when users do not accept recommendations, the RS keeps suggesting the same set of items. However, we observe the opposite: when the awareness set size increases, Recommendation Coverage increases as well. This is an interesting result that can be explained by noting that with a larger awareness set size, users choose more diverse items (according to the Gini index and Choice Coverage). Consequently, at the beginning of each month, the RSs models, which are re-trained on the base of these diverse choices, produce more diverse recommendations as well. Interestingly, this could contribute to the observed increase in the Gini index and Choice Coverage. In

Table 6
Impact of the recommendation set size on the considered users' choices related metrics.

	RS	Apps			Games			Kindle		
		Recommendation set size			Recommendation set size			Recommendation set size		
		10	50	100	10	50	100	10	50	100
Gini	PCF	0.80	0.82	0.82	0.90	0.89	0.88	0.86	0.86	0.86
	LPCF	0.75	0.76	0.80	0.87	0.83	0.85	0.84	0.83	0.85
	FM	0.78	0.80	0.81	0.89	0.89	0.88	0.85	0.85	0.83
	POP	0.81	0.89	0.93	0.91	0.96	0.97	0.87	0.93	0.95
	AR	0.83	0.91	0.94	0.92	0.96	0.96	0.87	0.90	0.90
Coverage	PCF	0.53	0.50	0.36	0.18	0.20	0.22	0.31	0.30	0.29
	LPCF	0.59	0.56	0.50	0.23	0.27	0.25	0.34	0.35	0.33
	FM	0.55	0.52	0.32	0.19	0.20	0.21	0.32	0.32	0.33
	POP	0.52	0.42	0.11	0.19	0.13	0.10	0.31	0.24	0.20
	AR	0.51	0.38	0.09	0.18	0.13	0.12	0.29	0.25	0.22
Popularity	PCF	0.00025	0.00035	0.00077	0.00027	0.00028	0.00029	0.00038	0.00037	0.00037
	LPCF	0.00020	0.00021	0.00024	0.00022	0.00018	0.00021	0.00034	0.00034	0.00035
	FM	0.00023	0.00024	0.00054	0.00024	0.00021	0.00019	0.00037	0.00034	0.00031
	POP	0.00135	0.00216	0.00569	0.00084	0.00132	0.00125	0.00160	0.00282	0.00274
	AR	0.00150	0.00232	0.00523	0.00099	0.00138	0.00114	0.00047	0.00053	0.00048
Choice's Rating	PCF	4.05	4.06	4.20	4.27	4.25	4.23	3.34	3.33	3.32
	LPCF	4.01	4.00	4.03	4.23	4.18	4.19	3.33	3.33	3.32
	FM	4.02	3.98	3.96	4.23	4.18	4.14	3.33	3.32	3.31
	POP	4.11	4.18	4.36	4.33	4.39	4.40	3.40	3.41	3.41
	AR	4.14	4.28	4.47	4.37	4.43	4.35	3.34	3.33	3.30
Recommendation Coverage	PCF	0.20	0.43	0.38	0.07	0.24	0.40	0.21	0.37	0.42
	LPCF	0.29	0.62	0.61	0.10	0.41	0.51	0.31	0.50	0.50
	FM	0.01	0.03	0.03	0.01	0.04	0.06	0.04	0.13	0.18
	POP	0.00	0.01	0.02	0.00	0.01	0.01	0.01	0.02	0.03
	AR	0.00	0.01	0.02	0.00	0.01	0.01	0.00	0.01	0.01
Recommendation acceptance	PCF	0.15	0.41	0.92	0.16	0.51	0.62	0.16	0.47	0.62
	LPCF	0.09	0.34	0.87	0.13	0.41	0.61	0.15	0.44	0.63
	FM	0.11	0.34	0.87	0.14	0.43	0.59	0.14	0.46	0.62
	POP	0.18	0.49	0.94	0.21	0.57	0.70	0.16	0.51	0.66
	AR	0.21	0.56	0.95	0.25	0.57	0.67	0.16	0.43	0.61

conclusion, we can state that an increase in the awareness set size results in an increase of the choice and recommendation diversity, while, paradoxically, it also produces a decrease in the quality of the choices.

6.4. Impact of changing the number of recommendations on choice diversity

We finally focus on the fourth research question: "RQ4 - Is a larger recommendation set producing an increase of choice diversity? If yes, has the awareness set size a larger or smaller effect on choice diversity than the recommendation set size?"

The experimental results are shown in Table 6. One could expect that increasing the number of recommendations should result in an increase in the diversity of the choices; if a user chooses items from a larger set of options, then the choices should be more diverse. But, in our simulation, the user can also make choices from the awareness set, for items that were not recommended. Hence, the actual effect of increasing the number of recommendations is not easy to anticipate. In fact, we observe that with the personalised RSs, when the recommendation set size is increased, the Gini index is not substantially affected: it can either grow a little (in the *Apps* data set) or remains substantially stable (in the *Games* and *Kindle* data sets). While the effect on Choice Coverage is more clear: it decreases. The observed decrease in Coverage is explained by the fact that when a higher number of items are recommended, the choices tend to become more focused on these recommendations rather than the awareness set. How the choices are distributed among the chosen items is not affected too much, i.e., the Gini index is not changing a lot.

On the other hand, with non-personalised RSs, by increasing the recommendation set size, both the Gini index and Choice Coverage clearly signal a decrease in the diversity (a higher Gini index and a much lower Choice Coverage). It is very interesting to note that while with 10 recommendations, the Choice Coverage of personalised and non-personalised RSs are similar, with 100 recommendations, the situation changes significantly: non-personalised RSs produce very low Coverage. In this situation, users choose more among the recommendations and less among their awareness sets and consequently Choice Coverage decreases. Similar arguments explain the larger Gini index for these RSs.

Lastly, in order to better understand the situation, we look at the detailed evolution of the Gini index, month by month, in the *Apps* data set and we compare a baseline condition (Fig. 10(a)) with two alternative conditions: when the awareness set size is increased (Fig. 10(b)) and when the recommendation set size is increased (Fig. 10(c)). It can be easily noted that the Gini index is both lower and faster-decreasing with a larger awareness set size (5000 vs 3000). In contrast, the Gini index is both larger and less quickly decreasing when the number of recommendations is increased. This shows that increasing the awareness set size is more effective in increasing choice diversity than recommending more items.

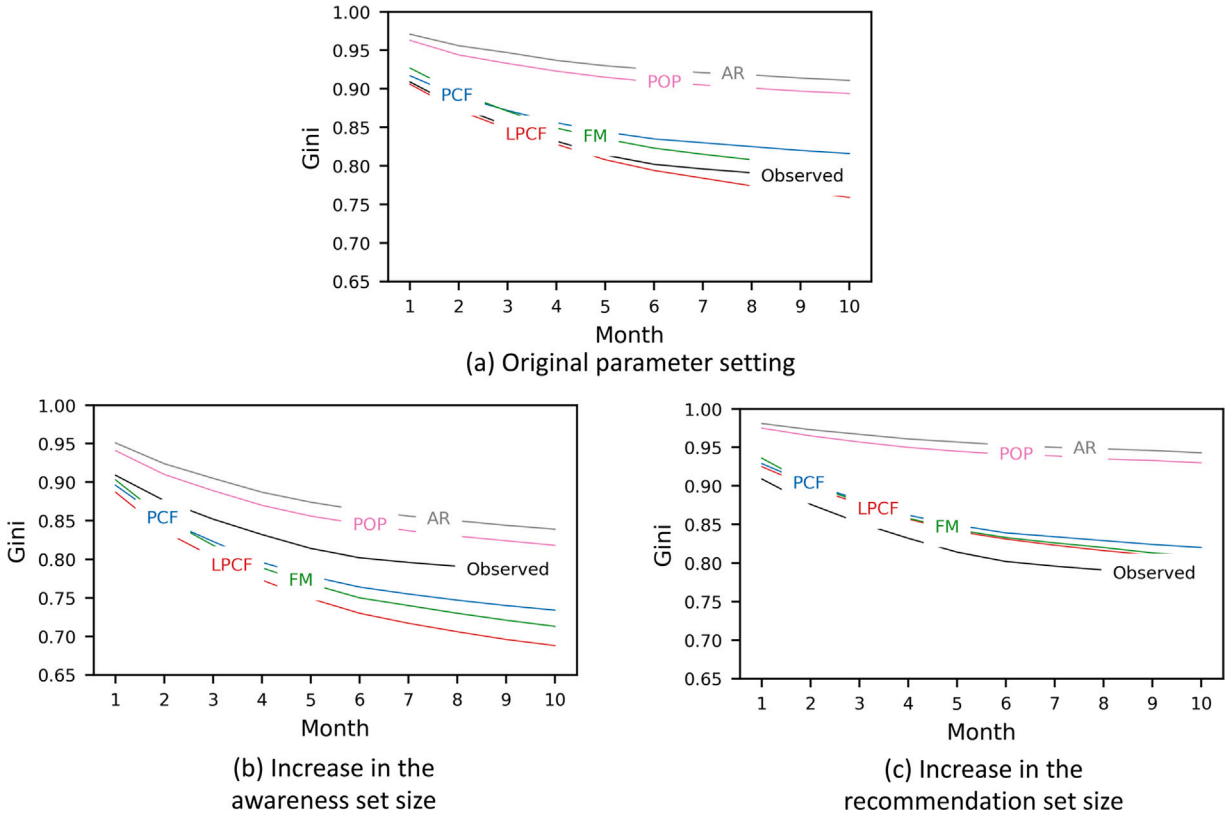

Fig. 10. Apps data set. Evolution of the Gini index over the observed and simulated choices under five different RSs when (a) original simulation parameters are used, (b) awareness set size is increased from original (3000) to 5000 and (c) recommendation set size is increased from original (50) to 100.

7. Comparison with previous simulation studies

As it is discussed in Section 3, previous simulation studies make very diverse design assumptions. Hence, it is not possible here to make precise comparisons, e.g., by running these state of the art procedures and make one-to-one comparisons with our outcomes. Our simulation framework is quite different from those previously cited. For instance, Fleder and Hosanagar (2009) use a small synthetic data set of users and items and also allows repeated choices. Conversely, we use large system log data to define users and items and to estimate user utility and simulate choices. However, in this section, we make some comparisons, trying to relate the findings of our study with those of previous works. We focus on the two studies that more closely match our simulation framework.

Fleder and Hosanagar (2009) have shown that adopting RSs always results in a lower choice diversity compared to the case where the is no RS. However, in our results we show that the extent of this effect largely depends on the RS's algorithm: some RSs decrease the diversity much more than others. Moreover, we found that although in the *Apps* data set, all the considered RSs produce a lower diversity (higher Gini index) than the NO-RS case, in the other two data sets LPCF produces higher diversity. This shows that the choice diversity in the presence or the absence of an RS strictly depends on the application domain, and one can design RSs that can tame this problem. Additionally, Fleder and Hosanagar (2009) have found that when δ is increased, hence, the salience of recommendations and the probability of recommendation acceptance becomes larger, also a higher Gini index is observed. We have not analysed the specific impact of δ, but we have found that when the awareness set size is decreased, that results in a higher recommendation acceptance (see Section 6.3), as if δ is increased and the Gini index increases as well. Hence, we have shown that a decrease in choice diversity can be due to either a smaller awareness of the users or to an increased salience of the recommendations.

Szlávik et al. (2011) have discovered that different choice models can lead to different users' choice behaviours, hence they stressed the importance of properly modelling the user, in addition to the RS. For instance, by increasing the probability of choosing the recommended items, they have showed that choice diversity can decrease, but, at the same time, users make better choices (higher ratings). In our analysis, we focus on the awareness of the users, which has an indirect effect on the acceptance of the recommendations, which was analysed by Szlávik et al. (2011). The results obtained by manipulating the users' awareness of the catalogue (Section 6.3) show that by decreasing the awareness set size, the users are more likely to choose among the recommendations only and choice diversity decreases. Hence, our results, confirm the results found by Szlávik et al. (2011), i.e., a higher acceptance of recommendations leads to lower choice diversity (higher Gini index). Moreover, in our simulation,

by considering alternative RSs, we have discovered that when an RS is producing recommendations that are accepted more often, then users also choose better items. This was also found by Szlávik et al. (2011).

8. Conclusion and future work

8.1. Summary of the results and implications

We have presented a simulation framework where the iterative choice making procedure of users, influenced by an RS, is simulated.

We have discovered that on top of a general tendency in successive months to *increase* diversity, only LPCF and FM can produce a larger choice diversity with respect to when no RS is used and this is true only in two data sets. Hence, in practice, only relying on RSs that explicitly penalises the most popular items, such as LPCF, one can increase diversity with an RS, compared to the situation where no RS is used. However, personalised RSs can increase the diversity of the choices much more than non-personalised RSs, which conversely influence users to make choices for more popular items and covering a smaller part of the items' catalogue.

Secondly, we have found that non-personalised RSs result in choices for items that have a larger predicted rating compared to personalised RSs. This result, together with the conducted analysis of the precision of the considered RSs, indicates that non-personalised RS can actually be strong baselines if the goal is to recommend items that the user will like, i.e., if there is no need to diversify the choices of the users and supporting them to discover novel items. In this second case, non-personalised RS must be avoided.

Thirdly, we have found that when the awareness of the users about the catalogue of items is increased, choices are more diverse, but there is a clear decrease in the acceptance of the recommendations, which leads to choices with smaller Choice's Rating. So, in conclusion, paradoxically, being aware of more items in the catalogue does not help users to make better choices.

Fourthly, our results show that increasing the recommendation set size has a marginal effect on the Gini index of personalised RSs, but decreases the Coverage of these RSs substantially. This is an apparently strange result, but it is due to the fact that when more recommendations are offered, the choices are more frequent among this restricted set of options, instead of ranging over the larger awareness set and hence the coverage diminishes.

We believe that the proposed simulation framework can become a powerful tool for RSs researchers and developers to anticipate the effects of a novel RS before actually deploying it to a real system. In addition to predicting the effect of an RS in the future, this simulation framework can also be employed for exploring counterfactual scenarios, i.e., understanding the effects that a target RS could have had on the choice distribution if it was used in a system. In fact, the proposed simulation framework is general and it is straightforward to be adjusted for specific settings. For instance, one can easily conduct simulations with other data sets, modify the simulation parameters, such as the number of recommendations, the time intervals, the size of the catalogue awareness of the users and the simulated effect of the RS on the choices.

8.2. Limitations and future work

There are still some limitations to the simulations conducted in this study that should be addressed in future studies. First of all, it is clearly important to compare these offline experiments with an online study, where users are observed while making real choices, informed by the considered RSs.

Moreover, since we use log data sets in domains such as books, video games, which are applications where a user typically consumes an item only once, it would be interesting to discover the effect of relaxing this condition and allow simulated users to make repeated choices. This analysis can be very instructive in application domains, such as music or movies, where this behaviour is observed. So far, we do not have any result that can indicate how the same parameters that we have manipulated (the awareness set size and the number of recommendations) can have on the users' choice distribution.

Another important limitation of the conducted simulation relates to the fact that we predict the full user–item pairs' ratings at the beginning of the simulation and we use them to estimate the utility for each user–item pair. In fact, estimating all the utilities at the beginning of the simulation is equivalent to make the assumption that users' preferences are independent from their previous choices. But, this is actually not true; a user's choice is typically dependent on her most recent choices: e.g., users listen to music tracks similar to their recently played tracks. It is possible to address this limitation by actively updating the rating prediction model with the simulated choices over time. But, this is not a simple task, as it is not clear how the choice model must actually incorporate the knowledge of the previously made choices. This is a challenging question.

We also note that in our simulation, we use some information about the future user's behaviour that is derived from the log data set. We do observe the users' preferences and this is necessary to conduct a reliable simulation. However, we have used a debiased model (IPS-MF) to correctly estimate the users' preferences. But we also observe the number of choices made by a user in a time interval. Hence, in some sense, we look at some properties of the future choices when simulating them. It is surely possible to address this limitation by estimating the number of choices of a user during each time interval based on her previous choices.

However, notwithstanding the above mentioned limitations, the results that we have described clearly show a number of interesting effects caused by the manipulation of parameters that an RS designer can actually perform. Hence, we give concrete and directly exploitable knowledge about the effects of RS technologies on the users' choice behaviour.

CRediT authorship contribution statement

Naieme Hazrati: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – reviewing and editing. **Francesco Ricci:** Supervision, Formal analysis, Conceptualization, Resources, Methodology, Writing – original draft, Writing – review & editing.

References

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2020). The connection between popularity bias, calibration, and fairness in recommendation. In *Fourteenth ACM conference on recommender systems* (pp. 726–731).

Adamopoulos, P., Tuzhilin, A., & Mountanos, P. (2015). Measuring the concentration reinforcement bias of recommender systems. *RN (I)*, *1*, 2.

Alhijawi, B., & Kilani, Y. (2020). A collaborative filtering recommender system using genetic algorithm. *Information Processing & Management*, *57*(6), Article 102310.

Anas, A. (1983). Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research, Part B (Methodological)*, *17*(1), 13–23.

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

Bountouridis, D., Harambam, J., Makhortykh, M., Marrero, M., Tintarev, N., & Hauff, C. (2019). SIREN: A simulation framework for understanding the effects of recommender systems in online news environments. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 150–159). ACM.

Brock, W. A., & Durlauf, S. N. (2002). A multinomial-choice model of neighborhood effects. *American Economic Review*, *92*(2), 298–303.

Chaney, A. J., Stewart, B. M., & Engelhardt, B. E. (2018). How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 224–232).

Dorfman, R. (1979). A formula for the gini coefficient. *The Review of Economics and Statistics*, 146–149.

Fleder, D. M., & Hosanagar, K. (2007). Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM conference on electronic commerce* (pp. 192–199). ACM.

Fleder, D., & Hosanagar, K. (2009). Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science*, *55*(5), 697–712.

Funk, S. (2006). Netflix update: Try this at home.

Hazrati, N., Elahi, M., & Ricci, F. (2019). Analysing recommender systems impact on users' choices. In *1st international workshop on the impact of recommender systems at RecSys*.

Hazrati, N., Elahi, M., & Ricci, F. (2020). Simulating the impact of recommender systems on the evolution of collective users' choices. In *Proceedings of the 31st ACM conference on hypertext and social media* (pp. 207–212).

He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th international conference on world wide web* (pp. 507–517).

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 eighth IEEE international conference on data mining* (pp. 263–272). IEEE.

Huang, J., Oosterhuis, H., de Rijke, M., & van Hoof, H. (2020). Keeping dataset biases out of the simulation: A debiased simulator for reinforcement learning based recommender systems. In *Fourteenth ACM conference on recommender systems* (pp. 190–199).

Ie, E., Hsu, C.-w., Mladenov, M., Jain, V., Narvekar, S., Wang, J., Wu, R., & Boutilier, C. (2019). Recsim: A configurable simulation platform for recommender systems. arXiv preprint arXiv:1909.04847.

Lee, D., & Hosanagar, K. (2014). Impact of recommender systems on sales volume and diversity. In *2 thirty fifth international conference on information systems*. Citeseer.

Lee, D., & Hosanagar, K. (2019). How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment. *Information Systems Research*, *30*(1), 239–259.

Lee, D., & Hosanagar, K. (2021). How do product attributes and reviews moderate the impact of recommender systems through purchase stages? *Management Science*, *67*(1), 524–546.

Li, Q., Myaeng, S. H., & Kim, B. M. (2007). A probabilistic music recommender considering user opinions and audio features. *Information Processing & Management*, *43*(2), 473–487.

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020). Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 2145–2148).

Marlin, B. M., & Zemel, R. S. (2009). Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on recommender systems* (pp. 5–12).

Matt, C., Hess, T., & Weiß, C. (2013). *The differences between recommender technologies in their impact on sales diversity*. ICIS.

Nadolski, R. J., Van den Berg, B., Berlanga, A. J., Drachsler, H., Hummel, H. G., Koper, R., & Sloep, P. B. (2009). Simulating light-weight personalised recommender systems in learning networks: A case for pedagogy-oriented and rating-based hybrid recommendation strategies. *Journal of Artificial Societies and Social Simulation*, *12*(1), 4.

von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton University Press.

Pradel, B., Usunier, N., & Gallinari, P. (2012). Ranking with non-random missing ratings: influence of popularity and positivity on evaluation metrics. In *Proceedings of the sixth ACM conference on recommender systems* (pp. 147–154).

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In *Recommender systems handbook* (pp. 1–34). Springer.

Saari, D. G. (1985). The optimal ranking method in the borda count. pure.iiasa.ac.at.

Sampaio, I., Ramalho, G., Corruble, V., & Prudêncio, R. (2006). Acquiring the preferences of new users in recommender systems-the role of item controversy. In *Proceedings of the ECAI 2006 workshop on recommender systems* (pp. 107–110).

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., & Joachims, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. In *International conference on machine learning* (pp. 1670–1679). PMLR.

Senecal, S., Kalczynski, P. J., & Nantel, J. (2005). Consumers' decision-making process and their online shopping behavior: A clickstream analysis. *Journal of Business Research*, *58*(11), 1599–1608.

Sie, R. L., Bitter-Rijpkema, M., & Sloep, P. B. (2010). A simulation for content-based and utility-based recommendation of candidate coalitions in virtual creativity teams. *Procedia Computer Science*, *1*(2), 2883–2888.

Szlávik, Z., Kowalczyk, W., & Schut, M. (2011). Diversity measurement of recommender systems under different user choice models. In *Fifth international AAAI conference on weblogs and social media*.

Teixeira, I. R., Carvalho, F. d. A. T. d., Ramalho, G., & Corruble, V. (2002). ActiveCP: A method for speeding up user preferences acquisition in collaborative filtering systems. In *Proceedings of the 16th brazilian symposium on artificial intelligence: Advances in artificial intelligence* (pp. 237–247). London, UK, UK: Springer-Verlag, URL: http://dl.acm.org/citation.cfm?id=645853.669613.

Umeda, T., Ichikawa, M., Koyama, Y., & Deguchi, H. (2014). Evaluation of collaborative filtering by agent-based simulation considering market environment. In *Developments in business simulation and experiential learning: Proceedings of the annual ABSEL conference (vol. 36)*.

Vargas, S. (2015). *Novelty and diversity evaluation and enhancement in recommender systems* (Ph.D. thesis, Ph.D. dissertation), Universidad Autónoma de Madrid.

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

Yao, S., Halpern, Y., Thain, N., Wang, X., Lee, K., Prost, F., Chi, E. H., Chen, J., & Beutel, A. (2021). Measuring recommender system effects with simulated users. arXiv preprint arXiv:2101.04526.

Zhao, X., Xia, L., Zou, L., Yin, D., & Tang, J. (2019). Toward simulating environments in reinforcement learning based recommendations. arXiv preprint arXiv:1906.11462.

Zhu, D. H., Wang, Y. W., & Chang, Y. P. (2018). The influence of online cross-recommendation on consumers' instant cross-buying intention: The moderating role of decision-making difficulty. *Internet Research*, *28*(3), 604–622.