Popularity Bias in False-Positive Metrics for Recommender Systems Evaluation

ELISA MENA-MALDONADO, RMIT University, Australia
ROCÍO CAÑAMARES and PABLO CASTELLS, Universidad Autónoma de Madrid, Spain
YONGLI REN and MARK SANDERSON, RMIT University, Australia

We investigate the impact of popularity bias in false-positive metrics in the offline evaluation of recommender systems. Unlike their true-positive complements, false-positive metrics reward systems that minimize recommendations disliked by users. Our analysis is, to the best of our knowledge, the first to show that false-positive metrics tend to penalise popular items; the opposite behavior of true-positive metrics – causing a disagreement trend between both types of metrics in the presence of popularity biases. We present a theoretical analysis of the metrics, which identifies the reason that the metrics disagree and determines rare situations where the metrics might agree – the key to the situation lies in the relationship between popularity and relevance distributions, in terms of their *agreement* and *steepness* – two fundamental concepts we formalize. We then examine three well known datasets using multiple popular true and false-positive metrics on sixteen recommendation algorithms. Specific datasets are chosen to allow us to estimate both biased and unbiased metric values. The results of the empirical study confirm and illustrate our analytical findings. With the conditions of the disagreement of the two types of metrics established, we then determine under which circumstances true-positive or false-positive metrics should be used by researchers of offline evaluation in recommender systems.[1]

CCS Concepts: • **Information systems → Recommender systems**; **Evaluation of retrieval results**.

Additional Key Words and Phrases: recommender systems, evaluation, metric, false positives, popularity bias, non-random missing data

ACM Reference Format:
Elisa Mena-Maldonado, Rocío Cañamares, Pablo Castells, Yongli Ren, and Mark Sanderson. 2021. Popularity Bias in False-Positive Metrics for Recommender Systems Evaluation. *ACM Transactions on Information Systems*, 43 pages.

1 INTRODUCTION

While recommender systems are becoming popular, how to evaluate the quality of recommended items has gained increased attention from both practitioners and researchers. As the primary goal of recommender systems is to suggest content that users like, most existing research focuses on measuring the number of recommended items that users are positive about. So-called true-positive

[1]This paper significantly extends previous work presented in [51].

Authors' addresses: Elisa Mena-Maldonado, elisa.mena.maldonado@rmit.edu.au, RMIT University, 124 La Trobe Street, Melbourne, Australia, 3000; Rocío Cañamares, rocio.canamares@uam.es; Pablo Castells, pablo.castells@uam.es, Universidad Autónoma de Madrid, Escuela Politécnica Superior, C/ Francisco Tomás y Valiente 11, Madrid, Spain, 28049; Yongli Ren, yongli.ren@rmit.edu.au; Mark Sanderson, mark.sanderson@rmit.edu.au, RMIT University, 124 La Trobe Street, Melbourne, Australia, 3000.

1046-8188/2021/1-ART1 $15.00

metrics such as precision, recall, mean reciprocal rank (MRR), or normalized discounted cumulative gain (nDCG) are employed.

However, one can also consider avoiding what users dislike. While this important topic has been occasionally studied in the field [19, 20, 24, 65], it is usual in evaluation practice in specific business domains. For instance, the ratio of music items that a user skipped (and possibly did not like) is a common metric in the evaluation of music recommendation [8, 27, 58]. To the best of our knowledge, the only work seeking a deeper understanding of false-positive metrics in recommendation and their relation to true-positive metrics is the conference version of this paper [51].

The presence of various biases puts an extra challenge in this research. For example, it is known that users tend to give a small proportion of ratings to the large number of items managed by a recommender system, which results in a data sparsity problem. More importantly, a small number of popular items commonly take the majority of existing ratings and this (rating) information is missing not at random (MNAR) [47, 68, 69]. The effect of such biases on recommendation algorithms and offline evaluation has become the object of growing research in the field [5, 11, 14, 40, 66, 75]. In particular, algorithms and metrics have been found to be biased to favor the recommendation of popular items, beyond their objective quality. Progress has been made in managing these effects, but how such biases may affect false-positive metrics has not been studied, as far as we are aware.

In this paper, we consider whether false-positive metrics capture anything different from true-positives. In particular, we extend the conference version of our prior work [51], where we found systematic disagreements between false- and true-positive metrics, and sought a broad line of explanation for them. In the present paper, we extend our previous findings by developing a substantially wider and consolidated formal analytic support, upon which we develop more complete and conclusive answers to our research questions. In particular, we now examine if one type of metric is more likely to deliver more reliable measurements than the other, and we investigate when and why each metric may be better. We further examine the generality of our findings in empirical observations with additional metrics and evaluation protocols that our theoretical analysis does not cover.

We investigate the following research questions:

- *What is the relationship between false and true-positive metrics, and in particular to what extent do they agree with each other while evaluating the quality of recommendations?*
- *When true and false-positive metrics disagree, may one be more often right than the other?*
- *How are the false-positive metrics affected by popularity bias in offline evaluation of recommender systems?*

To answer these questions, we develop a formal theoretical analysis on how popularity bias, the unjudged recommended items, and the MNAR conditions affect what false-positive can capture in offline recommender system evaluation. We formalize the expected *biased* and *unbiased* metric values and the *rankings that optimize them*. Furthermore, we conduct the analysis by comparing false-positive with true-positive metrics to investigate the relationship between them. Finally, we perform a set of comprehensive empirical analyses on these questions over different publicly available datasets while considering multiple popular true and false-positive metrics (including precision and anti-precision, recall and fallout, MRR and anti-MRR, nDCG and normalised discounted cumulative loss – nDCL [24]), with sixteen classical and state-of-the-art recommendation algorithms.[2]

We find that: 1) there are fundamental differences in the manifestation of the biases with respect to prior work: somewhat paradoxically, false-positive metrics unfairly penalize the recommendation of popular items, just as true-positive metrics unfairly reward them; 2) the agreement between true and false-positive metrics is tightly related to missing relevance information, and the fact that this

[2]The source code implementing all the experiments described in this paper is available at https://github.com/elikary/tois2021.

information is MNAR; 3) the effect of popularity biases on both true and false-positive metrics lies upon relations between three main distributions in the data: the *popularity* of items, their *average rating*, and their *global relevance* (beyond observed ratings). Two fundamental conditions between these distributions determine how the metrics behave: their degree of *agreement* and their relative *steepness*. By agreement we mean whether the most popular items are also the ones with highest average rating and/or global relevance; and steepness is a measure of the "strength" of a bias – how uneven a distribution is, a property we explicitly formalize.

The contributions of the paper are:

- A theoretical analysis on how the popularity bias affects false-positive metrics in offline recommender system evaluation.
- The discovery and explanation of the agreement and disagreement between true and false-positive metrics for offline evaluation of recommender systems.
- The identification of key elements to discern whether true or false-positive metrics may be more reliable than the other, in terms of capturing the underlying truth beyond their biases.
- A comprehensive examination of a set of popular true and false-positive metrics on a set of state-of-the-art recommendation algorithms over three public datasets.

The rest of the paper is structured as follows. We start by motivating the importance of false positives in recommendation, and briefly review prior work that has used or studied such metrics in evaluating recommendations. Next, we show in Section 3 an initial example where contradictory results are observed in the offline evaluation of common recommender systems with false and true-positive metrics. In Section 4 we introduce the basic formalisms for our theoretical analysis, and the datasets with which we confirm and illustrate our findings along the analysis. The core of the analysis spans Sections 5 to 7: in Section 5 we find concise probabilistic formulations for the ranking functions that respectively optimize specific true and false-positive metrics, namely precision and "anti-precision", which we take as exemplars for our study. In Section 6, we analyze the agreements and disagreements between the two metrics, and in Section 7 we formally study to what extent each type of metric can be distorted, from an unbiased estimate of its actual value, by the popularity biases in the data. After the theory, we contrast our formal conclusions with empirical observations: we report in Sections 8 and 9 experiments where a set of state of the art recommendation algorithms are run on different datasets and evaluated with the two types of metrics. We end the paper with some closing conclusions. Additionally, detailed proofs for our most formal findings are provided in two appendices.

2 BACKGROUND AND RELATED WORK

The practical goal of a recommender system is defined by the particular purpose of recommendation within a specific application. The understanding of what a useful recommendation is has evolved and grown significantly beyond producing accurate rating estimates [1], towards considerably wider perspectives over the last two decades [15, 22, 50]. Amid many different (sometimes conflicting) objectives, matching the end-user's tastes can be understood as a primary requirement for recommendation to make sense. This dimension is broadly referred to as the accuracy of recommendation.

When assessing accuracy, different angles can be considered. The commonest approach focuses on the ability of a system to deliver as many good recommendations to as many people as possible [35]. Perfect accuracy is usually viewed as an impossible goal and users are expected to be tolerant of some error. If there are useful choices in the mix, uninteresting recommendations will hopefully be ignored. For these reasons, recommender system evaluation practice and research has largely focused on counting (evidence of) true positives [32]. Some attention has been paid to the flip side:

the (evidence of) false positives [19, 24, 65]. We too find it worthwhile pondering the potential negative effects that disliked recommendations can have on the user experience.

2.1 The Cost of a Bad Recommendation

Disliked recommended items have a negative impact on the user experience, that can have lasting effects [46]. False positives are thereby a main concern in many application domains. For instance, in automatic music playlist generation [22] users may commonly tolerate background music that is just nice, but they may be annoyed by an occasional unpleasant track [26]. As a consequence, the skip rate is a common metric in music and video streaming [58]. Skip behavior is also a common target for prediction in recent public challenges, where it is part of released data for evaluation (e.g. [8]). Some authors have likewise applied metrics of least non-relevant music to evaluate playlists [27].

False positives are also an important target to minimize in online advertisement – the prevalent revenue model for a substantial share of major services on the Internet – particularly so when users are not just indifferent to displayed ads, but explicitly annoyed by them. Broder et al., for instance, [6] addressed the prediction of bad advertisements, aiming to explicitly avoid false positives, considering that displaying no ad is preferable to inappropriate advertisement. Goldstein et al. [29] further studied the negative impact that annoying ads have on the user experience and thereby on the business revenue, as well as other long-term consequences (disengagement, reputation harm, etc.) for the involved stakeholders (the users, the advertisers, the platform). Later, Bron et al. [7] consider the action of closing ads by users as an explicit negative feedback signal. They define a false-positive metric based on this signal, the *hide rate*, as a distinct complement of the (positive) click-through rate.

The skip rate has been similarly used in Web search to assess the cost in reading and skipping effort of non-relevant results [73]. Dating online is also often mentioned as a domain where false positives involve a significant cost [59]. Information Retrieval (IR) metric frameworks have been likewise developed that consider the cost and benefit involved in delivering relevant and non-relevant documents [80]. Beyond IR and recommender systems, false positives in classification are important in particular domains, and cost-aware machine learning theory and methods have been long developed considering this [21, 60]. From a wider perspective, scholars have studied how a bad recommendation can hurt user trust [17]. Psychological studies have described a negativity bias in human perception, whereby bad impressions may sometimes outweigh good ones in our overall assessment of an experience [3, 78].

2.2 False-Positive Metrics in Recommendation

The recommender systems literature is rich in accuracy evaluation methodologies [4, 11, 32, 35, 69]. Most focus on true positives as the recommendation objective to be assessed. One notable exception to this is the use of ROC (Receiver Operating Characteristic) curves and the area underneath [32], which employ a false-positive metric – fallout – in the x axis. Frolov and Oseledets [24] and Sánchez and Bellogín [65] explicitly considered false positives in their definition and observation of "anti-metrics", which compute any common true-positive metric on flipped relevance judgments. For instance, fallout [60] is the anti-metric for recall, and the ratio of returned non-relevant items (so-called anti-precision [24, 44, 65]) is the anti-metric for precision. Fallout is also occasionally reported among other evaluation metrics in some work [19, 20]. In IR, Lipani et al. [44] related the disagreement between true-positive and false-positive metrics to the incompleteness of relevance knowledge, and used them to propose a correction method for the potential biases in pooling-based evaluation of search results.

While true-positive metrics consistently display a high correlation between each other throughout reported research, the aforementioned work systematically reports frequent contradictions between true-positive and false-positive metrics. While such disagreements were discussed in the corresponding work, a conclusive or systematic explanation has not been described, and is sought here. As we shall see, the discrepancies are caused by relevance knowledge incompleteness, a particularly acute condition in offline recommender system evaluation. More specifically, the cause lies in the fact that ratings are typically MNAR in common datasets [47, 68], where heavy popularity biases pervade the data, impacting the algorithms and metrics that assess their accuracy.

2.3 Popularity Bias in Recommendation

Important progress has been made in the last decade in confirming, measuring, explaining and coping with such popularity biases. Marlin et al. [47, 48] pointed out the MNAR condition [30] of recommender systems' input, and proposed algorithms that better coped with it. Following up on this, Steck [69] proposed metrics and algorithms to compensate for the popularity biases in the data. Jannach et al. [40] examined the popularity bias in different recommendation algorithms and suggested means to mitigate it. Bellogín et al. [5] showed that evaluation is biased towards rewarding popular items when information retrieval evaluation methodologies are applied to recommender systems.

More recently, Cañamares and Castells provided a formal explanation for the popularity bias in kNN algorithms [10] and developed a formal theory for whether the popularity bias is desirable or not, and to what extent offline evaluation can be deceitful to this respect [11]. In doing so, Cañamares and Castells formalized the distinction between biased and unbiased evaluation, and identified conditions for the two measurements to agree or disagree.

On a closely related line, drawing from related work in machine learning and statistics, a recent strand of research has addressed the bias in offline evaluation as an issue of mismatch between the data gathering policy (e.g. free user interaction with a deployed system) and item selection by the recommendation algorithms to be evaluated. Building on this perspective, techniques (such as inverse propensity scoring) have been explored to reduce the biases in the evaluation [28, 31, 71, 75] and the evaluated algorithms [36, 45, 66].

While all this prior research has focused on true-positive metrics, the question remains whether the results reached so far would similarly apply to false-positive metrics. We address the question here and, as we will see, the answer differs considerably from the corresponding prior findings. Our present research builds in many respects upon previous results by Cañamares and Castells [11]. While metric agreement analysis was a means to an end in that work, it is now in the main focus of our present research, and the popularity bias is, as we shall see, a component in fitting the pieces of our analysis together.

3 FALSE-POSITIVE VS. TRUE-POSITIVE METRICS

We thus find motivation for the use of evaluation metrics that assess bad recommendations, along with (or complementarily to) metrics that assess good. Simple metrics involving false positives, such as fallout [60] or anti-precision [44], can suitably meet this purpose; they are defined as:

$$\text{Fallout} = \frac{\text{FP}}{\text{FP} + \text{TN}} \qquad \text{antiP} = \frac{\text{FP}}{\text{TP} + \text{FP}} \tag{1}$$

where TP and FP denote the number of relevant and non-relevant retrieved (recommended) items respectively, and TN is the number of non-relevant items that are not returned. The measures can be respectively defined as recall and precision using non-relevance in place of relevance. If skipping

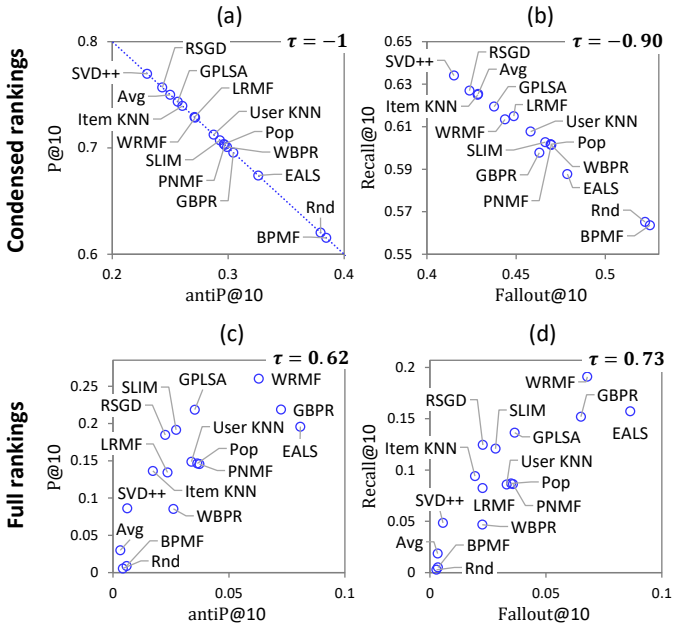


Fig. 1. True-positive vs. false-positive metrics, for condensed (top) vs. full (bottom) rankings in MovieLens 1M. Kendall τ correlation is shown for each plot.

a song in a music streaming session is taken as a sign of non-relevance, antiP is the skip rate (i.e. the ratio of played songs that were skipped) [8, 22, 57].

3.1 True and False Positives as Complementary Metrics

One may expect that false-positive metrics measure quite a similar thing to true-positive metrics, just "from the other end". False-positive can be expected to strongly (negatively) correlate with true-positive metrics. For instance, for anti-precision this relationship is direct and linear, as antiP is the exact arithmetic complement of precision:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \Rightarrow \mathrm{antiP} = 1 - \mathrm{P} \tag{2}$$

Figure 1 (a) illustrates this relationship in the metric values of antiP@10 vs. P@10 for a set of collaborative filtering algorithms (detailed later in Section 8) on MovieLens 1M [49], a widely used dataset example. The metrics in these graphs are measured taking so-called condensed rankings [9, 63, 64, 76, 77], where unrated (unjudged) items are excluded from the evaluated rankings before computing the metrics – an experimental setting that is not generally advised as we discuss shortly, but which we use here for illustrative purposes. We can see the algorithms stand on a straight $y = 1 - x$ line, confirming that antiP@10 = 1 − P@10. In essence, they are the same metric. The relationship between fallout and recall is not exactly linear due to the different denominators in the two metrics. We can see in Figure 1 (b) that they are still strongly negatively correlated. Since true-positive and false-positive metrics are inversely oriented (the lower antiP and fallout, the better the recommendation is considered), this negative correlation means agreement on which of almost every pair of systems is best.

These example observations are obtained however with an experimental setting (condensed rankings) that makes relevance knowledge artificially complete in the ranking. Offline recommender

system evaluation is generally conducted with highly incomplete relevance knowledge. Even though condensed rankings have been occasionally used or studied in the literature [2, 4, 14, 35, 54, 70], they are not generally considered the best option. Recent work [12] has found in fact that condensed rankings result in massive losses in effect size and statistical significance; and in deviation from the actual task that the evaluated systems are meant to solve: ranking all the items in the dataset, not just those judged [43]. For all these reasons, our present work assumes full rankings as the default and advised option for offline evaluation in recommender systems research [4, 5, 32] – we have showed condensed rankings initially just for contrast, to emphasize now the effect of missing judgments. For an in depth discussion of full vs. condensed rankings in recommendation we refer the reader to [12].

Figure 1 bottom (c and d) illustrates what happens with the common full rankings setting. As can be seen, the complementarity of true and false-positive metrics is not just lost, it is reversed, with high positive Kendall τ correlations reflecting disagreement in system comparisons.[3] Such a level of disagreement between true and false-positive metrics is rather intriguing, all the more so when it seems a quite systematic trend, as we will see in Section 8 on further datasets.

3.2 The Uncertainty in Incomplete Relevance Judgments

The contradiction between true and false-positive metrics is in fact made possible by the implicit assumption that relevance knowledge is complete in order for the metrics to be strictly complementary. When it is not, the denominator of precision and anti-precision is no longer TP + FP, but $\widehat{\mathrm{TP}} + \widehat{\mathrm{FP}} + \mathrm{U}$, where U denotes the number of unjudged returned items, and the "hat" over variables and metrics ($\widehat{\mathrm{TP}}$, $\widehat{\mathrm{P}}$, etc.) denotes the corresponding counts and measurements with incomplete relevance judgments:

$$\widehat{\mathrm{P}} = \frac{\widehat{\mathrm{TP}}}{\widehat{\mathrm{TP}} + \widehat{\mathrm{FP}} + \mathrm{U}} \qquad \widehat{\mathrm{anti\mathrm{P}}} = \frac{\widehat{\mathrm{FP}}}{\widehat{\mathrm{TP}} + \widehat{\mathrm{FP}} + \mathrm{U}} \tag{3}$$

Now the metrics (precision and anti-precision) no longer add to 1:

$$\widehat{\mathrm{P}} + \widehat{\mathrm{anti\mathrm{P}}} = 1 - \delta \qquad \text{with } \delta = \frac{\mathrm{U}}{\widehat{\mathrm{TP}} + \widehat{\mathrm{FP}} + \mathrm{U}} \tag{4}$$

The difference δ from unity is the fraction of recommended unjudged items. Moffat et al. [52] call this ratio the *residual* of the metric (here, precision), and take it as a measure of the uncertainty involved in the metric values computed with the available, incomplete relevance knowledge. Lipani et al. [44] consider this ratio as well in their analysis of the bias in IR evaluation with pooled judgments. For instance, for the systems evaluated in Figure 1, $\delta@10 = 0.83$ on average over all the systems in the figure: most of the relevance knowledge is missing (the figure is actually displaying "hat" metrics; though we omitted the "hats" at this point in the figure labels for simplicity – our notation will be fully rigorous to this respect in the sequel).

The disagreement is, however, not fully explained by this knowledge incompleteness alone: if judgments were simply missing at random (MAR), we should expect the correlation between metrics to just decrease rather than becoming consistently negative. We can therefore anticipate that the disagreement should relate to the strongly MNAR effects [11, 47, 66, 68] in the user-item observations that are commonly available for offline recommender system evaluation. Our analysis will seek to clarify how these issues relate to each other. We start by seeking answers through a formal analysis of the metrics and their optimization, following up on related prior work on

[3] All the correlations are statistically significant at $p < 0.05$, and so are all the Kendall τ and Pearson correlation values reported everywhere in the rest of the paper (except in Figure 18).

popularity biases in recommender system evaluation [11]. We will then confirm and illustrate our theoretical findings with experiments and empirical observations under different angles.

4 PRELIMINARIES FOR ANALYSIS: NOTATION AND DATA

Before starting our formal analysis we briefly introduce the notation and basic formalities that are used in the rest of the paper, which we summarize in Table 1. When we develop the theoretical analysis in the following sections we will use toy examples and instance observations with real data to provide intuitions and illustrate the effects of the analytical findings. We therefore introduce the datasets to be used for this purpose already in this preliminary section.

4.1 Notation

Given a set of users $\mathcal{U}$ and a set of items $\mathcal{I}$, we formalize key elements in evaluation experiments by defining binary random variables in $\mathcal{U} \times \mathcal{I}$ that describe relationships between users and items, as in [11, 51]: we define the variable rel as taking value 1 iff the user likes the item. We define rated = 1 iff the user has been observed interacting with the item, in such a way that evidence of positive or negative preference (i.e. an observation of rel for the user-item pair at hand) is obtained – we will say, for short, that a "rating" is present in the available data records. Note that for all our purposes, rel can be understood as a (binary) rating value. We will hence work with binary (or binarized) ratings all along the theoretical and empirical developments of our study.

The available observations (ratings) for an experiment are commonly divided (either by natural design, or by an artificial split) into a training set and a test set. We shall therefore define the variables train and test as being 1 iff rated = 1 and the rating was assigned to the training or test subset, respectively. Since the two subsets are disjoint, we always have train · test = 0. The training input for a recommender system is therefore $\{(u, i, \mathsf{rel}(u, i)) \in \mathcal{U} \times \mathcal{I} \times \{0, 1\} \mid \mathsf{train}(u, i) = 1\}$, and the set $\{(u, i, \mathsf{rel}(u, i)) \in \mathcal{U} \times \mathcal{I} \times \{0, 1\} \mid \mathsf{test}(u, i) = 1\}$ is used as the equivalent to relevance judgments for the computation of evaluation metrics.

Following common convention in IR, we shall use rel, train, test as abbreviation for rel = 1, train = 1, test = 1, relying on context for disambiguation. Based on these variables we can express meaningful probabilities, such as $p(\mathsf{rel}|i)$, denoting the ratio of users who like item $i \in \mathcal{I}$; $p(\mathsf{train}|i)$, the ratio of users that the system has observed interacting with item i – that is, the "popularity" of the item [11, 66]; and $p(\mathsf{rel}|\mathsf{train}, i)$, the ratio of observed interactions involving item i that evidence a positive preference – the average binarized training rating:[4]

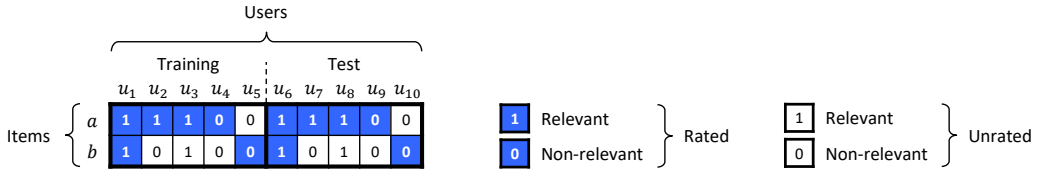
$$
\begin{aligned}
p(\mathsf{rel}|i) &= \frac{|\{u \in \mathcal{U} \mid \mathsf{rel}(u, i) = 1\}|}{|\mathcal{U}|} \qquad & p(\mathsf{rel}|\mathsf{train}, i) &= \frac{|\{u \in \mathcal{U} \mid \mathsf{rel}(u, i)\ \mathsf{train}(u, i) = 1\}|}{|\{u \in \mathcal{U} \mid \mathsf{train}(u, i) = 1\}|} \\
p(\mathsf{train}|i) &= \frac{|\{u \in \mathcal{U} \mid \mathsf{train}(u, i) = 1\}|}{|\mathcal{U}|} \qquad & p(\mathsf{train}|\mathsf{rel}, i) &= \frac{|\{u \in \mathcal{U} \mid \mathsf{rel}(u, i)\ \mathsf{train}(u, i) = 1\}|}{|\{u \in \mathcal{U} \mid \mathsf{rel}(u, i) = 1\}|}
\end{aligned}
\tag{5}
$$

And so forth. Figure 2 illustrates the computation of a few such probabilities on a toy example involving ten users and two items.

4.2 Data

In the following sections, we will complement and illustrate our theoretical analysis with example observations on real data. We shall use for this purpose three public datasets: MovieLens 1M [49], Yahoo! R3 [47] and CM100k [11]. The same datasets are used later in section 8, where we run and

[4]The probability $p(\mathsf{rel}|\mathsf{train}, i)$ amounts to the average (binarized) rating of item i, since binarized ratings (having value 1 when the item is relevant to the user, and zero otherwise) are but relevance judgments, and this probability simply computes the average of such 0/1 rating values over training ratings – the average training rating.



	$p(\text{rel}\|i)$	$p(\text{train}\|i)$	$p(\text{test}\|i)$	$p(\text{rel}\|\text{train}, i)$	$p(\text{rel}\|\text{test}, i)$	$p(\text{rel}\|\neg\text{train}, i)$	$p(\text{rel}, \text{test}\|i)$	$p(\text{train}\|\text{rel}, i)$
$i = a$	6/10	4/10	4/10	3/4	3/4	3/6	3/10	3/6
$i = b$	4/10	2/10	2/10	1/2	1/2	3/8	1/10	1/4

$p(\text{rel})$	$p(\text{train})$	$p(\text{test})$	$p(\text{rel}\|\text{train})$	$p(\text{rel}\|\text{test})$	$p(\text{rel}\|\neg\text{train})$	$p(\text{rel}, \text{test})$	$p(\text{train}\|\text{rel})$
10/20	6/20	6/20	4/6	4/6	6/14	4/20	4/10

Fig. 2. Toy rating and relevance matrix example with just two items a and b, ten users and a fictitious 50-50% training/test rating split (where, for simplicity, users u_1 to u_5 have only training ratings, and u_6 to u_{10} only test ratings). Blue cells represent ratings, white cells unrated items, and the binary value in the cells represents relevance (value of the rel variable). The computation of some illustrative probabilities involving relevance, ratings and the two sides of the split is exemplified in the figure.

evaluate a set of recommendation algorithms and verify the extent to which the results of our theoretical analysis are empirically confirmed. The details of the datasets are shown in Table 2.

MovieLens is possibly the most popular dataset in the recommender systems literature, and is hence a good representative of typical evaluation practice material. The 1M subset contains about one million ratings for movies by users in the MovieLens application.[5] The ratings were collected from spontaneous user activity in the system [49], and hence this is a classic instance where the available observations (as input for the evaluated systems and as test data for evaluation) are subject to strong bias and display a marked MNAR pattern in the data distributions.

A common dataset such as MovieLens enables a good deal of analysis already, as we shall see. We can compute distributions such as the popularity $p(\text{train}|i)$ and the average rating $p(\text{rel}|\text{train}, i)$ of an item i, as defined in equation 5, based only on the supplied rating data: we interpret rating values greater than or equal to a threshold (4 in this dataset) as indicative of rel = 1 and values below as meaning rel = 0, and we simply need to apply a rating split (typically random) into a training subset and a test subset in order to work with the train variable – we shall apply a 80-20% training-test split for this purpose.

It is not possible however to properly estimate distributions such as $p(\text{rel}|i)$ or $p(\text{train}|\text{rel}, i)$ – or even $p(\text{rel})$ – with a common MNAR dataset as MovieLens, as the observations of relevance are highly incomplete and biased: we lack any relevance knowledge outside the (highly biased) rating sample. We use Yahoo! R3 and CM100k to complement this limitation: these datasets provide relevance judgments sampled uniformly at random, in different ways.

The CM100k data[6] includes about 100 ratings per user for music selected uniformly at random from the Deezer[7] catalog [11]. This sample of user tastes enables natural unbiased estimates of distributions involving the probability of relevance. For instance, $p(\text{rel}|i) \sim |\{u \in S(i) \mid \text{rel}(u, i) = 1\}| \,/\, |S(i)|$ is estimated as the ratio of users who liked i among the set of people – denoted as

[5]https://grouplens.org/datasets/movielens/1m
[6]http://ir.ii.uam.es/cm100k
[7]https://www.deezer.com

Table 1. Notation summary.

$\mathcal{U}, \mathcal{I}$	Set of all users and all items, respectively.
$\mathsf{rel} : \mathcal{U} \times \mathcal{I} \to \{0, 1\}$	Random variable indicating whether a user likes an item.
$\mathsf{rated} : \mathcal{U} \times \mathcal{I} \to \{0, 1\}$	Random variable indicating whether a user-item interaction record (a rating) is available (is logged) in a dataset for offline evaluation. We have $\mathsf{rated} = \mathsf{train} \vee \mathsf{test}$.
$\mathsf{train} : \mathcal{U} \times \mathcal{I} \to \{0, 1\}$	Random variable indicating whether a user has been observed interacting with an item by the evaluated system.
$\mathsf{test} : \mathcal{U} \times \mathcal{I} \to \{0, 1\}$	Random variable indicating whether the relevance of an item for a user is known for evaluation but not to the system.
$p(\mathsf{train}\|i)$	Popularity of item i: ratio of users who have been observed interacting with item i by the evaluated system.
$p(\mathsf{rel}\|\mathsf{train}, i)$	Average binarized rating of item i: ratio of interactions with i observed by the system that indicate a positive preference.
$M : \mathrm{Aut}(\mathcal{I}) \times \mathcal{U} \to \mathbb{R}$	Oracle "real" value – or unbiased estimate – of a metric M, assuming full relevance knowledge – or an unbiased sample thereof. $\mathrm{Aut}(\mathcal{I})$ denotes the set of all permutations (i.e. rankings) of $\mathcal{I}$.
$\widehat{M} : \mathrm{Aut}(\mathcal{I}) \times \mathcal{U} \to \mathbb{R}$	"Observed" estimate of a metric M computed in an offline experiment drawing (incomplete) relevance knowledge from a split-based sample of logged user interaction with items.
$\varphi_M : \mathcal{I} \to \mathbb{R}$	Optimal ranking function for a metric M. If the ranking is personalized (i.e. different for each user), a user variable is implicit.
$\mathrm{dr}_{a,b} f$	Decrement rate of a real function f over two items a, b. See equation 11 for mathematical definition.

Table 2. Details of the datasets.

Dataset	#Users	#Items	#Ratings
MovieLens 1M	6,040	3,706	1,000,209
Yahoo! R3	5,400	1,000	183,179 = 129,179 training + 54,000 test
CM100k	1,054	1,084	103,584 = 11,594 on familiar music + 91,990 on unfamiliar music

$S(i) \subset \mathcal{U}$ – who were asked to rate i in the survey. And other probabilities are similarly estimated taking $S(i)$ in place of $\mathcal{U}$ as appropriate.

In order to work with a MNAR train random variable in this dataset, we use the extra information supplied in the data about whether users were familiar or not with the music before being surveyed, as in [11]: we consider as "MNAR ratings" only the ratings for music that users had already heard before, as a proxy for spontaneous user-item interaction. In accordance with this, in this

dataset we estimate $p(\text{train}|i)$ over users in $S(i)$ rather than $\mathcal{U}$ (i.e. with $|S(i)|$ instead of $|\mathcal{U}|$ in the denominator), just as we do for $p(\text{rel}|i)$.

Yahoo! R3 contains ratings for music entered by users in the Yahoo! LaunchCast streaming service [47]. The dataset includes two subsets:

- A "training" set that was collected from spontaneous user interaction with music in the system – this set of training data is therefore MNAR.
- A "test" set containing ratings that each user was asked to enter for ten music tracks sampled uniformly at random – hence test ratings are MAR.

The MNAR training set enables the computation of the same probabilities as MovieLens. And it is straightforward to estimate, for instance, $p(\text{rel}|i)$ using the test set just in the same way as we described for CM100k. It is trickier to estimate other probabilities such as $p(\text{train}|\text{rel})$, because the MNAR (biased training) and MAR (unbiased relevance) samples are disjoint in this dataset. Since CM100k does not have this limitation we will use it primarily, along with MovieLens 1M, in the formal analysis sections that follow.

5 OPTIMAL RANKING FOR FALSE POSITIVES

Our analysis of the agreement or disagreement between true and false-positive metrics is developed in terms of a comparison of the optimal rankings that – respectively – maximize and minimize each metric. We select for this purpose precision and anti-precision [24, 44, 65] as our primary metrics, because of their exact arithmetic relation, and as a simple and most tractable case we shall take P@1 and antiP@1. We have observed in our experiments that our analytical findings generalize well to other false-positive metrics and cutoffs, as we will show with examples in Section 8.

The same as prior work [11, 68] distinguished between the "observed" and "true" values of true-positive IR metrics (such as precision and nDCG), we can make the same distinction for false-positive metrics. The true values – to which we shall refer as "*real*"[8] – and the observed values are computed with complete and incomplete relevance knowledge, respectively. We shall often refer to them as biased (for observed) and unbiased (for real) values as well, since an incomplete relevance sample introduces bias in the metric estimates – unless the relevance knowledge is sampled uniformly at random over user-items pairs.

Thus, the real value of antiP@1 of a ranked recommendation R is 1 if the first item in R is disliked by the target user, and 0 otherwise. And the observed value of antiP@1 is 1 if the target user dislikes the first recommended item and a rating (hence denoting a negative preference) is present in the test set for this user-item pair. Analogous definitions apply to precision, with "like" in place of "dislike" [11]. We shall use $\widehat{\text{P}}$ and $\widehat{\text{antiP}}$, with a "hat", to refer to the observed value of the respective metrics, while metric names without the hat shall designate their oracle values.

For instance, with the toy example in Figure 2, the recommendation $\langle a, b\rangle$ has real precision $\text{P} = 2/2$ for user u_8 because this user likes both recommended items a and b. But the observed precision is $\widehat{\text{P}}@1 = 1/2$, because only a has a test rating by u_8, and b is unrated by this user. Likewise, the real anti-precision for u_7 is $\text{antiP} = 1/2$ because this user only dislikes item b. But $\widehat{\text{antiP}} = 0$, because this negative preference is unobserved (the test rating by u_7 for b is missing). For u_6, the observed and real metric values are the same, $\widehat{\text{P}} = \text{P} = 1$ and $\widehat{\text{antiP}} = \text{antiP} = 0$, because the relevance knowledge is complete for this user.

[8] This is to avoid overloading the word "true" here, which we are already using in the phrase "*true* positives", and thus avoid confusion.

5.1 True-Positive Optimals

Using the distinction between true and observed metric values, Cañamares and Castells [11] proved that the optimal recommendation that maximizes the *real* value of P@1 ranks items by non-increasing values of the following ranking function:

$$\varphi_{\mathrm{P}}(i) = p(\mathsf{rel}|\neg\mathsf{train}, i) \tag{6}$$

That is, ranking the items $i \in \mathcal{I}$ by decreasing order of $\varphi_{\mathrm{P}}(i)$ produces a recommendation R that maximizes P@1 of R in expectation. The reader is referred to [11] for the detailed proof, but the intuition is that the optimal recommendation is obtained by ranking items by decreasing probability of relevance (as in Robertson's probability ranking principle [62]), with the additional condition that the target user has not been observed interacting in the system with the recommended items before (i.e. no training rating is present for the user-item pair), a requirement for discovery that is usual in most recommendation scenarios.

In the same lemma, they proved that when observed precision is computed by using a random split of available ratings, the optimal ranking for observed $\widehat{\mathrm{P}}$@1 is defined by $\varphi_{\widehat{\mathrm{P}}}(i) = p(\mathsf{rel}, \mathsf{test}|\neg\mathsf{train}, i)$. The intuition is clear: the observed precision only computes an item as relevant if a positive test rating attests such relevance (i.e. we have $\mathsf{rel} \wedge \mathsf{test}$). By applying rules and definitions of probability, we can rewrite this objective function as follows:

$$\begin{aligned}
\varphi_{\widehat{\mathrm{P}}}(i) = p(\mathsf{rel}, \mathsf{test}|\neg\mathsf{train}, i) &= \frac{p(\mathsf{rel}, \mathsf{test}, \neg\mathsf{train}|i)}{p(\neg\mathsf{train}|i)} \\
&= \frac{p(\mathsf{rel}, \mathsf{test}|i)}{1 - p(\mathsf{train}|i)} \sim \frac{p(\mathsf{rel}, \mathsf{train}|i)}{1 - p(\mathsf{train}|i)} \qquad (7) \\
&\propto p(\mathsf{rel}|\mathsf{train}, i)\, \frac{p(\mathsf{train}|i)}{1 - p(\mathsf{train}|i)} \qquad (8)
\end{aligned}$$

where '$\propto$' denotes rank-equivalence, and we apply the definition of conditional probability. In step 7 we use the fact that $\mathsf{test} \wedge \neg\mathsf{train} = \mathsf{test}$ (since a test rating is by definition not in training), and $p(\mathsf{test}|\mathsf{rel}, i) \propto p(\mathsf{train}|\mathsf{rel}, i)$ because:

(1) When ratings are partitioned into training and test subsets uniformly at random (by a given split ratio), the probability of test and training are proportional to the probability of rating (multiplied by the corresponding ratio).
(2) The probability that a rating goes to either side of the split is the same for all items, and is therefore independent from any item characteristic such as its relevance [11].

Thus train and test have different probabilities, but these produce the same rankings. Getting rid of the test variable in the optimal ranking function for observed precision will enable us to implement it as a legitimate recommendation algorithm taking just the same input (training data) as any other recommender system, to which it can be fairly compared in experiments (as opposed to oracle implementations that are given "unfair" access to test data). The optimal for real precision in contrast does require oracle relevance knowledge to estimate $p(\mathsf{rel}|\neg\mathsf{train}, i)$, and is therefore useful mainly as a theoretical construct. Note also that in past work [11] the probabilities are expressed in terms of the rated variable, while for our aims in this paper we rewrite them, equivalently, in terms of train (e.g. in equation 3 above) because, again, this represents the input that recommender systems can see, and it is more convenient for our line of analysis.

5.2 False-Positive Optimals

Given that anti-precision can be defined as precision on flipped relevance [24, 44, 65], we can directly infer that the optimal rankings that minimize anti-precision are defined by: 1) replacing

rel for ¬rel in equations 6 and 8; and 2) reversing the ranking, e.g. by a negative sign on the ranking function. We reverse the ranking function because while the optimal ranking for precision should maximize the metric, the opposite is the case for false-positive metrics: the optimal ranking for anti-precision should minimize anti-precision (the fewer false positives the better). Thus, the optimal ranking functions for anti-precision are as follows:

$$\varphi_{\text{antiP}}(i) = -p(\neg\text{rel}|\neg\text{train}, i) \propto p(\text{rel}|\neg\text{train}, i) \tag{9}$$

$$\varphi_{\text{anti}\widehat{\text{P}}}(i) = -p(\neg\text{rel}, \text{test}|\neg\text{train}, i) \propto -p(\neg\text{rel}|\text{train}, i)\ \frac{p(\text{train}|i)}{1 - p(\text{train}|i)} \tag{10}$$

where we apply similar steps as in equation 8. From equations 6 and 9 a first conclusion follows right away:

Conclusion 1. *The optimal ranking for real precision and real anti-precision are identical, as their ranking functions are equivalent:* $\varphi_{\text{antiP}}(i) \propto \varphi_{\text{P}}(i)$.

The optimal rankings for real precision and anti-precision being identical means that the two metrics agree on the comparison of any recommendation algorithm to the optimal ranking and its inverse, the worst. This agreement in comparisons to such extremes may make us expect that perhaps the metrics would tend to agree, at least as a general trend. We will check this empirically in Section 8.

In contrast, we see in equations 8 and 10 that the optimal rankings in observed metric values are not quite the same. The optimal ranking function for observed anti$\widehat{P}$ (equation 10) is the product of two terms:

(1) The opposite of the "negated" rating $-p(\neg\text{rel}|\text{train}, i)$. This is a "double negation" of the corresponding term in equation 8.
(2) The popularity odds $p(\text{train}|i)/(1 - p(\text{train}|i))$, a monotonically increasing function of popularity $p(\text{train}|i)$. The same component is present in equation 8.

Having popularity – in the term (2) – as a component of the ranking function, multiplied by a negative number – as is the term (1) – means that the more popular an item is, the lower it is placed in the optimal ranking. We can therefore conclude, to begin with, that:

Conclusion 2. *The popularity bias tends to work against observed anti-precision: the metric is biased to favor the recommendation of unpopular items in offline evaluation.*

This is the exact opposite of the behavior of true-positive metrics [11], in which offline evaluation tends to reward the recommendation of popular items. In the next sections, we analyze further consequences of such opposing trends. We begin by examining the relation between precision and anti-precision in observed values. After that in Section 7 we analyze the relation between the observed and real value of each metric.

As a final side-comment, note that if relevance knowledge were complete (i.e. all user-item pairs had either a training or a test rating), we would have $\neg\text{train} \Rightarrow \text{test}$, therefore $p(\neg\text{rel}, \text{test}|\neg\text{train}, i) = p(\neg\text{rel}|\neg\text{train}, i)$, and by equations 6-10, $\varphi_{\text{anti}\widehat{\text{P}}}(i) = \varphi_{\text{antiP}}(i) \propto \varphi_{\text{P}}(i) = \varphi_{\widehat{\text{P}}}(i)$: all the optimal rankings would be just the same – and in fact anti$\widehat{\text{P}} = 1 - \widehat{\text{P}}$, as argued in Section 3.1 (equation 2). More generally, is not difficult to see that we get the same situation – in expectation – if the test data were randomly sampled (i.e. if test is independent from rel and i anywhere). The possibility to have different behaviors for different metrics is therefore, once again, a consequence of relevance knowledge being incomplete (i.e. $\text{U} \neq 0$ and $\delta \neq 0$ in equations 3 and 4 in Section 3.2) and missing not at random.

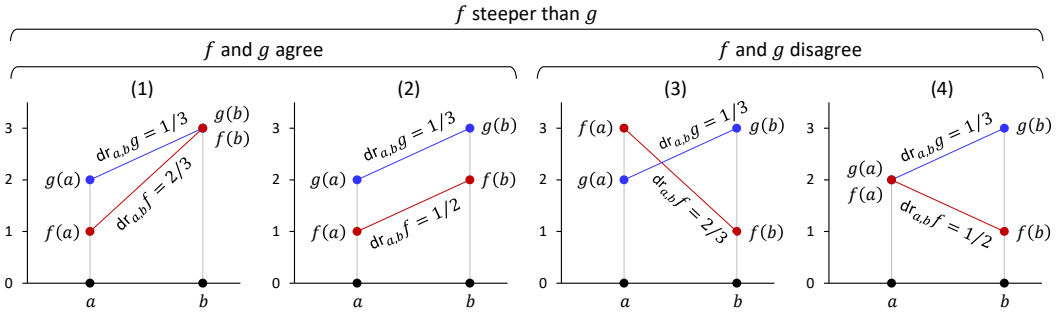


Fig. 3. Illustration of the concepts of steepness and agreement. In all the examples, f is steeper than g: in cases (1) and (3) f has a larger difference than g, $|1-3| > |2-3|$; and both have the same maximum value of 3. In examples (2) and (4) f and g have the same difference $|2-3| = |1-2|$, but g has a higher maximum value $3 > 2$.

6 POPULARITY BIAS IN FALSE POSITIVES: DISAGREEMENTS IN OBSERVED VALUES

We start by analyzing the observed values of the true and false-positive metrics, and how they compare to each other. As noted, the difference in the optimal rankings for observed precision and anti-precision is in the $-p(\neg\text{rel}|\text{train}, i)$ term for observed anti-precision in equation 10, in place of $p(\text{rel}|\text{train}, i)$ for observed precision in equation 8. This term is responsible for the opposite effect of popularity in the two metrics, and also explains why the true and false-positive metrics may come to disagree with each other, as we shall see. But we shall identify specific conditions for this – or the opposite – to be the case, which essentially relate to the strength of the popularity bias, and whether it goes along with or against the relevance of items. We do so, first, in terms of a generic pair of items and the order in which the two optimal recommendations would rank them. After that, we will inspect the global trends that arise in specific datasets as a result of, simply put, how many pairs of items the two rankings agree or disagree upon.

The precise conditions for the agreement or disagreement of optimal rankings over a given pair of items are a matter of *agreement* and *steepness* of the components of the ranking functions. We formalize and explain this in the following definitions and a subsequent lemma.

Definition 1. *Given two functions $f : \mathcal{I} \to \mathbb{R}^+, g : \mathcal{I} \to \mathbb{R}^+$, over the set of items $\mathcal{I}$, we say that f and g* agree *over two items $a, b \in \mathcal{I}$ if they rank them in the same order: either $f(a) < f(b) \wedge g(a) < g(b)$, or $f(a) > f(b) \wedge g(a) > g(b)$.*

We introduce this definition just as a convenient short name for a quite simple condition that we use in our analysis all along. Note that in the above definition we omit the case where $f(a) = f(b) \wedge g(a) = g(b)$.[9] Implicitly, this means that when we say two functions agree or disagree, neither should have the same value for a and b. This is intentional, and aims to simplify our statements and explanations – we will henceforth disregard everywhere the case where functions are equal for two items. For completeness, we provide in Annex B the conclusions that result when equality occurs – which are a smooth completion of our main analysis, as we shall see.[10]

[9]Likewise, when we say that f and g *disagree* over two items we shall mean that either $f(a) < f(b) \wedge g(a) > g(b)$, or $f(a) > f(b) \wedge g(a) < g(b)$.

[10]Note also that the condition that f and g be positive functions (i.e. ranging in $\mathbb{R}^+$) is not strictly needed in the definition of agreement, but we take it here for homogeneity with the definition that follows.

Table 3. Summary of cases established in Lemma 1.

$p(\text{train}\|\cdot)$ and $p(\text{rel}\|\text{train},\cdot)$ agree	$\wedge$	odds of $p(\text{train}\|\cdot)$ steeper than $p(\neg\text{rel}\|\text{train},\cdot)$	$\Rightarrow$	$\varphi_{\text{anti}\widehat{\text{P}}}$ and $\varphi_{\widehat{\text{P}}}$ disagree
$p(\text{train}\|\cdot)$ and $p(\text{rel}\|\text{train},\cdot)$ agree	$\wedge$	odds of $p(\text{train}\|\cdot)$ less steep than $p(\neg\text{rel}\|\text{train},\cdot)$	$\Rightarrow$	$\varphi_{\text{anti}\widehat{\text{P}}}$ and $\varphi_{\widehat{\text{P}}}$ agree
$p(\text{train}\|\cdot)$ and $p(\text{rel}\|\text{train},\cdot)$ disagree	$\wedge$	odds of $p(\text{train}\|\cdot)$ steeper than $p(\text{rel}\|\text{train},\cdot)$	$\Rightarrow$	$\varphi_{\text{anti}\widehat{\text{P}}}$ and $\varphi_{\widehat{\text{P}}}$ disagree
$p(\text{train}\|\cdot)$ and $p(\text{rel}\|\text{train},\cdot)$ disagree	$\wedge$	odds of $p(\text{train}\|\cdot)$ less steep than $p(\text{rel}\|\text{train},\cdot)$	$\Rightarrow$	$\varphi_{\text{anti}\widehat{\text{P}}}$ and $\varphi_{\widehat{\text{P}}}$ agree

The notion of steepness was hinted at – but not formalized – as part of an analytical argument by Cañamares and Castells [11], related to our own analysis here, and we now provide a precise formal definition.

Definition 2. *Given two positive functions $f : \mathcal{I} \to \mathbb{R}^+, g : \mathcal{I} \to \mathbb{R}^+$, over the set of items $\mathcal{I}$, we say that f is* steeper *than g over two items $a, b \in \mathcal{I}$ if its decrement rate is higher:*

$$\text{dr}_{a,b} f = \frac{|f(a) - f(b)|}{\max\left(f(a), f(b)\right)} > \frac{|g(a) - g(b)|}{\max\left(g(a), g(b)\right)} = \text{dr}_{a,b} g \tag{11}$$

Steepness is the decrement rate [72] of a function over two points – the decrement is divided by the maximum value – and measures the decrease over two points. We could alternatively define steepness as the *increment* rate [72], by replacing min for max in the denominators of equation 11, and the definition would be strictly equivalent: it is not immediately obvious but it is easy to check that a function has a higher decrement rate than another with respect to the maximums $\Leftrightarrow$ it has a higher increment rate with respect to the minimums. Figure 3 illustrates the notion of steepness with simple examples.

Based on these definitions, the following statement establishes the agreement between precision and anti-precision in terms of their observed values.

Lemma 1. *The optimal rankings for (the observed value of) precision and anti-precision of any two given items disagree if and only if either of the two conditions hold:*

(a) *The odds of $p(\text{train}|\cdot)$ is steeper than $p(\text{rel}|\text{train},\cdot)$, and disagrees with $p(\text{rel}|\text{train},\cdot)$ in comparing the two items.*
(b) *The odds of $p(\text{train}|\cdot)$ is steeper than $p(\neg\text{rel}|\text{train},\cdot)$, and disagrees with $p(\neg\text{rel}|\text{train},\cdot)$ in comparing the two items.*

Proof. See Appendix A.

Table 3 summarizes the cases that Lemma 1 establishes. By "disagreeing" in the lemma we mean that one item is more popular but has a lower average rating than the other (as per Definition 1). The intuition behind this lemma is that popular items with many ratings tend to get a higher chance of accumulating more positive ratings, which can turn into true positives in evaluation. But by the same reason, popular items also carry a high statistical potential for producing false positives. Thus, recommending these items is found to be good in terms of true positives but is found to be bad in terms of false positives. If an item a is much more popular than another item b (popularity is very "steep" over such two items), then chances are a will have both more positive and negative ratings than b. Thus a will yield higher scores in both true and false-positives, producing disagreements between optimal rankings: while true-positive metrics suggest ranking a before b, false-positive metrics suggest the opposite. Only if b had such a substantially higher (or lower) average rating than a both metrics might agree to rank b before a (respectively, a before b). We illustrate next such cases and intuitions along with the lemma through toy examples, followed by measurements on real data.

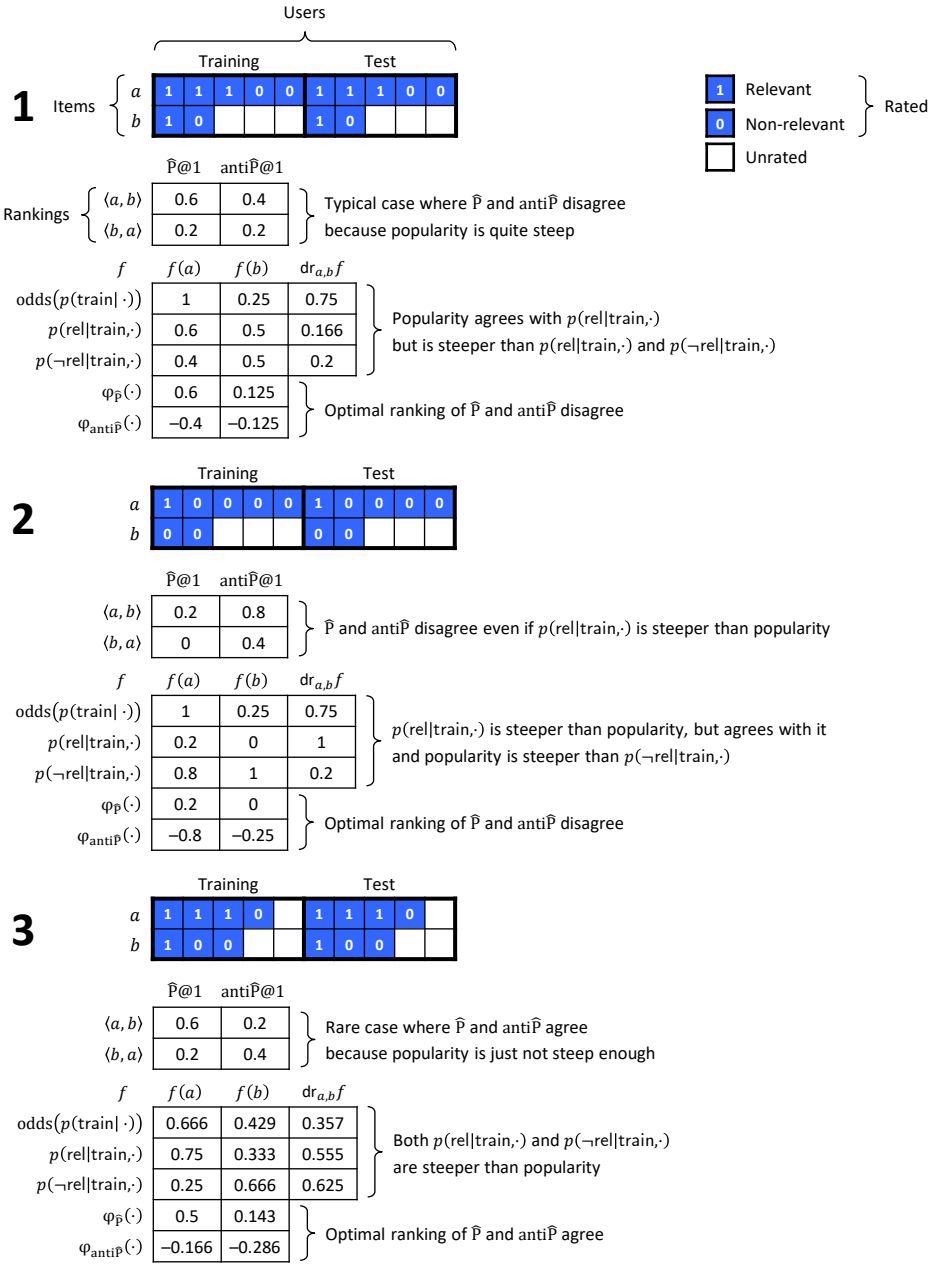


Fig. 4. Toy examples illustrating the formal analysis for observed metric values. Blue cells represent ratings, and white cells unrated items. Calculations are straightforward with equations 8, 10, 11, and the definition of $p(\text{train}|\cdot)$ and $p(\text{rel}|\text{train},\cdot)$ in Section 4.1. For instance, in example #1, $p(\text{rel}|\text{train},a) = 3/5 = 0.6$, $p(\text{rel}|\text{train},b) = 1/2 = 0.5$, and $\text{dr}_{a,b}f = |0.6 - 0.5|/0.6 = 0.1\overline{6}$ for $f(\cdot) = p(\text{rel}|\text{train},\cdot)$.

6.1 Toy Examples

Figure 4 shows three examples that illustrate the patterns characterized by Lemma 1. Similarly to the toy example in Figure 2, we consider we have just two items a and b, and ten users. We apply a 50-50% random split into training and test for whatever ratings are available, in such a way that, as a simplification, the same number of ratings (and rating values) fall on each side of the split. To further simplify the presentation of the example, we assume users fall entirely on only one side of the split – it is easy to see that this does not involve any loss of generality in the points we aim to illustrate.

In this toy setting, we compare two recommendations, that are delivered to all users: one that ranks a before b, and one that does the opposite. We compute $\widehat{\mathrm{P}}@1$ and anti$\widehat{\mathrm{P}}@1$ and, as is not uncommon [5], in the example we only compute the metrics over target users who have at least some test rating. In this setting, $\widehat{\mathrm{P}}@1$ of the ranking $\langle a, b\rangle$, for instance, is equal to the ratio of target users who have a positive test rating for a, and anti$\widehat{\mathrm{P}}@1$ is the ratio of target users who have a negative one. The three examples in Figure 4 illustrate different cases in how the relation between popularity and the average rating determines the agreement or disagreement between the two metrics.

- In example #1 the difference in average rating between the two items is small, whereas b is 75% less popular (in odds) than a. This makes precision and anti-precision disagree in observed value.
- Example #2 shows a case where the average rating is extremely steep: $p(\mathsf{rel}|\mathsf{train}, b)$ is 100% smaller than $p(\mathsf{rel}|\mathsf{train}, a)$, while the popularity difference is the same as in example #1. However, the average rating agrees with popularity, and the latter is still steeper than the complement of the former: $p(\neg\mathsf{rel}|\mathsf{train}, a)$ is just 20% smaller than $p(\neg\mathsf{rel}|\mathsf{train}, b)$. As a consequence, precision and anti-precision still disagree, confirming Lemma 1.
- Finally, example #3 represents an atypical case where popularity is much less steep than the average rating and its complement. Consequently, the two metrics agree.

Having analyzed the agreement or disagreement of optimal rankings in terms of a generic individual item pair, we now examine the trends that can be observed in real data that are commonly used for offline recommender system evaluation.

6.2 Observations on Real Data

For our illustrative purpose we take MovieLens 1M [49] as a common example (equivalent trends are observed in other similar datasets). Figure 5 confirms the correspondence of distribution steepness and alignment with metric agreement, and shows how frequent each case is in the relation between popularity and relevance density. We find that the optimal rankings for precision and anti-precision disagree on the vast majority (82.4%) of item pairs. The main cause for the disagreement between $\varphi_{\widehat{\mathrm{P}}}$ and $\varphi_{\mathrm{anti}\widehat{\mathrm{P}}}$ lies in the steepness of the popularity distribution, which is higher than the steepness of both $p(\mathsf{rel}|\mathsf{train}, i)$ and $p(\neg\mathsf{rel}|\mathsf{train}, i)$ in 63.3% of all item pairs.

The popularity distribution can be expected to be steeper than the relevance density (and its complement) in common recommendation environments: popularity biases ($p(\mathsf{train}|i)$ over i) are commonly exacerbated by a variety of exogenous factors in how users discover choices [11], some of which are further subject to self-reinforcement [23]. These factors do not affect intrinsic user tastes ($p(\mathsf{rel}|\mathsf{train}, i)$ over i) to any comparable extent. The popularity steepness is further amplified by the odds function $p/(1-p)$ in equations 8 and 10, that has a slope $\gg 1$. If the odds of popularity happens to be not steeper than the average rating or its complement for specific item pairs, Lemma 1 states that $\varphi_{\widehat{\mathrm{P}}}$ and $\varphi_{\mathrm{anti}\widehat{\mathrm{P}}}$ still have a chance to disagree, if popularity agrees with the steepest

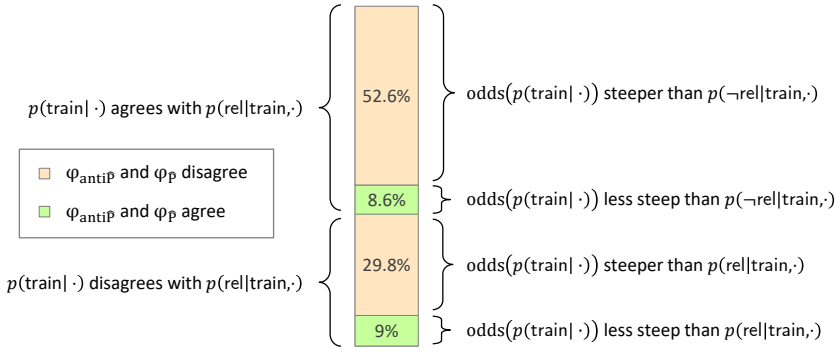


Fig. 5. Frequency – in terms of number of item pairs – of agreement / disagreement and steepness between popularity and the average rating in MovieLens 1M, and the resulting agreement or disagreement in the optimal rankings, as established by Lemma 1. The percentages shown in this figure provide a quantitative perspective of how frequent the cases described in Lemma 1 and Table 3 can be in practice, in a typical dataset.

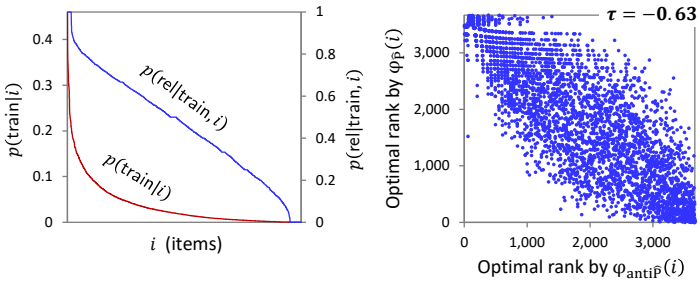


Fig. 6. Comparison of the popularity and average rating distributions over items in MovieLens 1M (left), and how this results in an opposing trend between the optimal rankings for observed precision and anti-precision (right). The curves in the left graph are sorted in decreasing order of $p(\text{train}|i)$ and $p(\text{rel}|\text{train}, i)$ respectively (hence the x axis is ordered differently for each curve). Each dot in the right plot is an item, and its coordinates reflect its rank in the corresponding optimal ranking. The Kendall τ correlation between the rankings is shown.

average rating (as in toy example #2). We can thereby state the following conclusion, which we will further contrast empirically in Section 5:

Conclusion 3. *The optimal rankings of observed precision and anti-precision can be expected to oppose each other as a general trend in common datasets for offline evaluation. Only if the popularity distribution was unusually flat would the optimals possibly agree.*

Figures 6 and 7 further illustrate our line of analysis. Figure 6 right confirms that the optimal rankings for observed precision and anti-precision tend to oppose each other in MovieLens 1M. Figure 6 left shows how the average rating (and therefore its complement) has a gentle linear decrease, while the popularity distribution $p(\text{train}|i)$ has a steeper decrease in comparison. Figure 7 bottom shows how the ranking function $\varphi_{\text{anti}\widehat{\text{P}}}$ has a strong (negative) correlation with popularity $p\,(\text{train}|i)$ (left), and no correlation with the average rating $p\,(\text{rel}|\text{train}, i)$ (right).[11] Since the

[11]Note that for correlation purposes, the average rating and its double negation are equivalent, as they are a linear function of each other: $-p\,(\neg\text{rel}|\text{train}, i) = p\,(\text{rel}|\text{train}, i) - 1$.

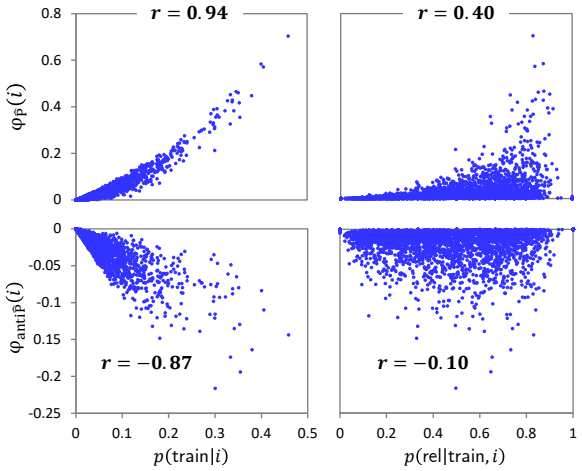


Fig. 7. Comparison of the optimal ranking functions for observed precision (top) and anti-precision (bottom) against their two main components: popularity (left) and the average rating (right). Each dot in the graphs is an item, and the values are computed on the MovieLens 1M rating data. Pearson correlation is shown for each graph. The ranking functions of precision and anti-precision tend to grow and decrease, respectively, with popularity (left). The average rating component (right) has a rather negligible effect on the optimal ranking in comparison – the slight correlations are in fact a transitive effect of the positive correlation between the average rating and popularity in MovieLens 1M.

popularity odds is multiplied by a negative value $-p(\neg\text{rel}|\text{train}, i)$, $\varphi_{\text{anti}\widehat{\text{P}}}(i)$ decreases with the popularity of i (Figure 7 bottom left), and popular items should therefore be ranked low for an optimal ranking. The opposite is the case for $\varphi_{\widehat{\text{P}}}$, which strongly correlates with popularity (left) because it is multiplied by a positive number $p(\text{rel}|\text{train}, i)$ (Figure 7 top).

6.3 Summary

The essential outcome of the analysis in this section is explaining why true and false-positive metrics tend to disagree, even though the definition of the metrics would suggest they should be complementary. The key reason for this is the presence of a popularity bias in the metrics, and the strength (steepness) of this bias. The precise answer is condensed in Lemma 1, which indicates that the observed values of precision and anti-precision disagree when popularity is sufficiently unevenly distributed (i.e. steep) over items. If it is more uneven than both relevance and non-relevance, the disagreement is guaranteed. If popularity happened to be somewhere less steep than both relevance and non-relevance, then the true and false-positive metrics would agree. But this is a rather rare situation when popularity is the product of spontaneous interaction, as the statistics show in Section 6.2 for MovieLens 1M (Figure 5).

Our formal analysis is developed in terms of optimal rankings, and relations between probabilities, between rankings and between metrics at a micro-level, that is, over pairs of items. From the aggregation of such patterns, one macro-level trend or another emerges, depending on which pattern is more frequent among the set of all item pairs. Later in Section 8 we will examine what all this ultimately amounts to at the global level by running actual recommendation algorithms over different datasets.

But first we continue our micro-level analysis to address a fundamental question: which metric – true or false-positive – can be more trustworthy when they disagree which each other. For this

Table 4. Summary of cases established in Lemmas 2 and 3.

				$\varphi_{\widehat{P}}$ and φ_P		$\varphi_{\widehat{\text{anti}P}}$ and $\varphi_{\text{anti}P}$
$p(\text{train}\|\text{rel},\cdot) \sim p(\text{train}\|\text{rel})$	–	–	$\Rightarrow$	Agree	$\wedge$	Agree
$p(\text{train}\|\text{rel},\cdot) \sim p(\text{train}\|\cdot)$ $\wedge$ $p(\text{train}\|\cdot)$ and $p(\text{rel}\|\cdot)$ agree	$\wedge$ odds of $p(\text{train}\|\cdot)$ steeper than $p(\neg\text{rel}\|\cdot)$	$\Rightarrow$	Agree	$\wedge$	Disagree
$p(\text{train}\|\text{rel},\cdot) \sim p(\text{train}\|\cdot)$ $\wedge$ $p(\text{train}\|\cdot)$ and $p(\text{rel}\|\cdot)$ agree	$\wedge$ odds of $p(\text{train}\|\cdot)$ less steep than $p(\neg\text{rel}\|\cdot)$	$\Rightarrow$	Agree	$\wedge$	Agree
$p(\text{train}\|\text{rel},\cdot) \sim p(\text{train}\|\cdot)$ $\wedge$ $p(\text{train}\|\cdot)$ and $p(\text{rel}\|\cdot)$ disagree	$\wedge$ odds of $p(\text{train}\|\cdot)$ steeper than $p(\text{rel}\|\cdot)$	$\Rightarrow$	Disagree	$\wedge$	Agree
$p(\text{train}\|\text{rel},\cdot) \sim p(\text{train}\|\cdot)$ $\wedge$ $p(\text{train}\|\cdot)$ and $p(\text{rel}\|\cdot)$ disagree	$\wedge$ odds of $p(\text{train}\|\cdot)$ less steep than $p(\text{rel}\|\cdot)$	$\Rightarrow$	Agree	$\wedge$	Agree

purpose we now seek to relate, in the next section, the optimal ranking functions for observed metric values to their real value counterparts. This is more challenging than our analysis so far; we will deal with the formal difficulty by considering simplified cases, and seeking to generalize from there.

7 AGREEMENT WITH REAL METRIC VALUES

Our analysis so far thus finds that the observed values of true and false-positive metrics will tend to stand in contradiction of each other in offline recommender system experiments. We should naturally wonder if, given this situation, either of the measurements should be misleading, or both could be providing a correct observation in their own way. By "correct" we mean agreeing with real metric values in the comparison between systems. Since we have seen that false and true-positive metrics tend to disagree in their observed comparisons, but they fully agree in their real values, then only one can be correct in its observed value – the question is, which one?

There is not a single answer for all cases, but it is possible to analyze extreme cases in the dependence between ratings, relevance and items, which we summarize in Table 4. Specifically, the two extremes we consider are the same as Cañamares and Castells [11] analyzed: the probability of rating being independent from the item given relevance, and ratings being independent from relevance given an item.

(a) The probability that a user rates an item (in the training set) is totally determined by whether or not they like the item. Formally, this means the train variable is conditionally independent on the item variable given rel: $p(\text{train}|\text{rel}, i) \sim p(\text{train}|\text{rel})$ and $p(\text{train}|\neg\text{rel}, i) \sim p(\text{train}|\neg\text{rel})$.
(b) The probability that a user rates an item is totally determined by what the item is, regardless of whether or not the user likes it. Formally, this means $p(\text{train}|\text{rel}, i) \sim p(\text{train}|i)$.

We will consider these two extreme situations and see that it is possible to identify conditions that determine the agreement and disagreement between observed and real metric values.

7.1 Relevance-Dependent Rating

If rating is conditionally independent from the item given relevance, Cañamares and Castells [11] proved that the observed and real value of precision agree. Considering that anti-precision can be defined as precision on flipped relevance, this principle also applies to the anti-metric. We can express this as a lemma:

Lemma 2. *If the probability of rating depends mainly on relevance above any other property of individual items, then both precision and anti-precision are consistent with their real value, in terms of the optimal rankings.*

Proof. The proof is given in Appendix A.

The event that a user rates an item can be understood as involving two events: the discovery of the item by the user, and the decision of the user to engage with it (to rate it) once discovered [11]. Conditional item independence reflects situations where the probability of these two events (discovery and decision to rate when discovered) is mainly determined by whether the user likes the item or not. This can be the case for instance when items are found by effective search and recommendation or good advice from friends [11], and when users prefer to rate things they like.

7.2 Item-Dependent Rating

The item-dependence case results in different situations depending on the relationship between the rating and relevance distributions. We can formally establish the cases and conditions by the following lemma.

Lemma 3. *If the probability of rating in the training set is strongly dependent on items beyond their relevance, then:*

(a) *If the probability of rating and relevance agree, then* (i) *the optimal for observed precision is consistent with the real optimal ranking, and* (ii) *anti-precision is consistent in observed and real optimal rankings if and only if non-relevance is steeper than the odds of popularity.*
(b) *If the probability of rating and relevance disagree, then* (i) *the optimal for observed anti-precision is consistent with the real optimal ranking, and* (ii) *precision is consistent in observed and real optimal rankings if and only if relevance is steeper than the odds of popularity.*

Proof. The proof is given in Appendix A.

A strong item dependence occurs when individual items are subject to very different distribution processes, for circumstances unrelated to what people may actually like to find (and then rate). This is the case for instance when items are subject to intense advertisement, fad, viral phenomena, etc. Even the time of release of an item can determine the amount of feedback it can get: older items have an advantage over new ones, as they have been exposed to user feedback in the system for a longer time.

Under the relevance-independence premise, the agreement or disagreement between popularity and relevance is a key primary condition in the lemma that determines the consistency of the true and false-positive metrics. As we shall discuss and analyze in the sequel, these two probabilities can be expected to agree more often than not, at least in a "positive world". But more negative situations are not unthinkable and we may want to be aware of their implications – we show a simulated example later in the experiments section.

As to the steepness conditions, note that relevance (or non-relevance) being steeper than popularity is not impossible, but is a rare condition, as argued in Section 6.2, as popularity biases are usually extremely strong. By the logic established in the lemma 3, this means that precision and anti-precision will not tend to be consistent at the same time in typical recommendation environments. We will check this however through inspection and experiments on real data.

7.3 Toy Examples

The toy examples in Figure 8 illustrate the cases established in this lemma, by extending the first example in Figure 4 with unobserved relevance.

- In example #1a, the item a is more popular than b (it has more ratings), and it also has a higher probability of relevance (it is liked by more users – counting those who rated the items and those who did not). This is sufficient to make observed and real precision consistent. Since the odds of popularity are steeper than non-relevance, observed anti-precision is not consistent with its real value. This corresponds to Lemma 3 case (a).

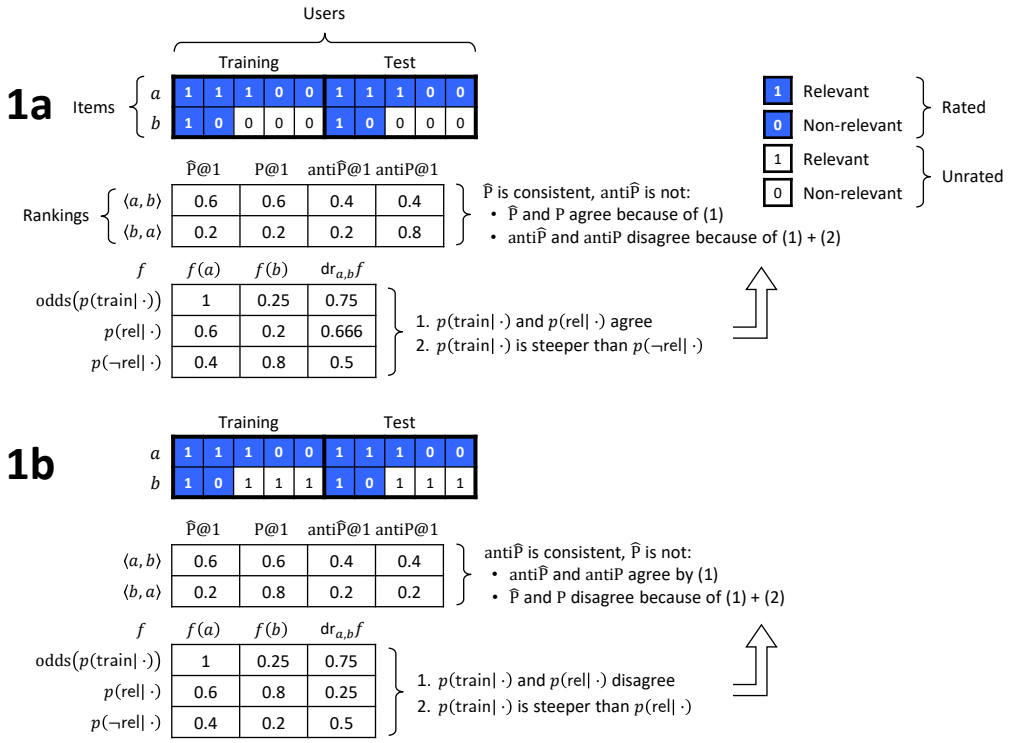


Fig. 8. Toy examples illustrating the formal analysis for observed vs. real metric values. As in Figure 4, blue cells represent ratings, and white cells unrated items. The two examples here extend example #1 in Figure 4 with the relevance value of unrated items. Calculations are directly derived from the basic definition of anti-precision in equation 1, steepness in equation 11, and the definition of $p(\text{train}|i)$ and $p(\text{rel}|i)$ in Section 4.1.

- Example #1b shows the opposite case: popularity and relevance go in opposite ways, whereby observed anti-precision agrees with its real value. Since popularity is steeper than relevance, observed precision gets distorted from its real value. This corresponds to Lemma 3 case (b).

Even if the relevance independence premise of Lemma 3 is not properly satisfied (it is easy to check that $p(\text{rel}|\text{train}, a) = p(\text{rel}|a)$, but $p(\text{rel}|\text{train}, b) \neq p(\text{rel}|b)$ in both examples), the results align with the theoretical conclusion.

7.4 Discussion and Observations on Real Data

Our analysis thus considers two ideal cases where rating probability depends either only on relevance, or on anything but relevance ("item dependence"). A pure conditional rating independence from relevance or items is not likely to occur in the real world, and both dependencies can be expected to coexist [11]. Yet as trends in a complex mix, the separate dependencies may help explain and expect different degrees of consistency of true and false-positive metrics in different situations. In the CM100k dataset, for instance, the conclusion of Lemma 2 – that both precision and anti-precision are consistent – holds only for 18.3% of item pairs. This means that the item-independence assumption of the lemma is far from being observed in this dataset. However, we can see in the statistics displayed in Figure 9 that a very high percentage of item pairs satisfy Lemma 3,

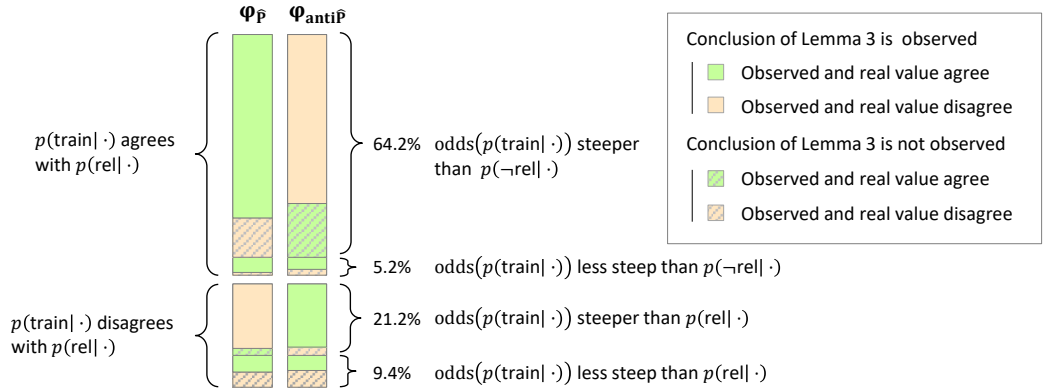


Fig. 9. Frequency – in terms of number of item pairs – of agreement / disagreement and steepness between popularity and the relevance probability in CM100k, and the resulting agreement or disagreement in the optimal rankings. The percentages shown in this figure provide a quantitative perspective of how frequent the cases described in Table 4 can be in practice, in a specific dataset. The non-striped areas indicate the proportion of item pairs of each segment that satisfy Lemma 3 – they add up to 81.2% for precision and 75.3% for anti-precision. The situation described by Lemma 3 is therefore largely observed even if the independence assumption does not strictly hold. The optimal rankings of observed and real precision agree (green segments) on 64.1% of item pairs, whereas anti-precision is consistent for 42.1% of item pairs.

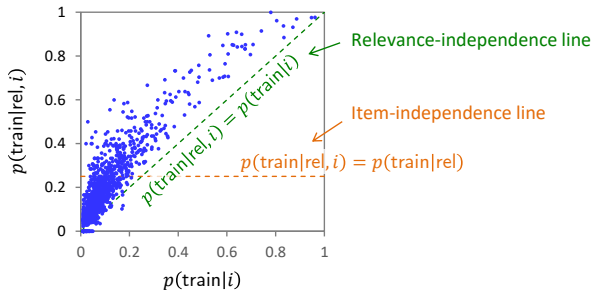


Fig. 10. Prior vs. relevance-conditional rating probability in CM100k. Each dot in the scatterplot is an item, and the plot shows how rating and relevance are not that far away from being conditionally independent given an item. The straight line $p(\text{train}|\text{rel}, i) = p(\text{train}|i)$ represents the point of strict item-conditional rating-relevance independence, and the line $p(\text{train}|\text{rel}, i) = p(\text{train}|\text{rel})$ represents the point of strict relevance-conditional rating-item independence. Based on our model representation with the CM100k data described in Section 4.2, $p(\text{train}|\text{rel}) = p(\text{train}, \text{rel})/p(\text{rel})$ is computed here as the number of user-item pairs in the dataset where the item is liked and familiar to the user, divided by the number of pairs where the user likes the item. Analogously, $p(\text{train}|\text{rel}, i)$ is computed as the ratio of users who are familiar with item i among those who like i.

even if the independence assumption does not hold – this may suggest that the item dependence (the premise of Lemma 3) is dominant over the relevance dependence (the premise of Lemma 2).

Figure 10 provides a complementary perspective of how close ratings are to being conditionally independent from relevance given an item, by plotting $p(\text{train}|\text{rel}, i)$ against $p(\text{train}|i)$. We see that most items stand above the conditional item independence line, that is $p(\text{train}|\text{rel}, i) > p(\text{train}|i)$ for most items. This means that rating indeed depends on relevance: users are more prone to rate items

they like than items they do not. However the points are not far from the relevance independence line – they are certainly closer to that than to the item independence line $p(\text{train}|\text{rel}, i) = p(\text{train}|\text{rel})$. The strength of this item-dependence seems to account for Lemma 3 – and not Lemma 2 – being observed.

7.5 Summary

We have examined in this section the correspondence between observed and real (i.e. biased and unbiased) metric values. We find that their relation is determined, first of all, by what the main factor is that attracts user interaction (i.e. popularity): relevance, or something else. If relevance is the dominant factor, then both true and false-positive metrics agree in observed and real values – and both metric types therefore agree with each other. Even if observed measurements are subject to biases, the bias does not deviate them, in qualitative comparisons, from the unbiased measurements. This seems not to be the case in the data we have examined, where relevance certainly fosters user interaction, but does not seem to be the strongest factor, as Figure 10 suggests.

If there is some other interaction-attracting force different from and stronger than relevance, then one metric type (either true-positive or false-positive) will tend to be inconsistent with its real value, and the other will be consistent. Which of either metric type is consistent is determined by the relative monotonicity of the popularity and relevance distributions: if items liked by many people tend to attract much interaction, then false-positive metrics are inconsistent and true-positive metrics are consistent. The opposite is the case if the items most people dislike are the ones getting most interaction. The first situation can be expected to be more frequent than the second, as is found in our CM100k data inspection (Figure 9), but the second is not impossible. When relevance is not the main factor determining popularity, there is still a theoretically possible exception where all metrics can be consistent: when user interaction is distributed evenly enough over items (more so than relevance). But we have discussed already that this is rarely a case found "in the wild".

Our formal analysis essentially concludes here. In the next two sections, we empirically examine how our theoretical study – the micro-level item pair perspective and the macro-level observations on raw data – transcends to the comparative evaluation of actual recommender systems using datasets with MNAR and MAR data.

8 EMPIRICAL OBSERVATIONS: COMPARISON OF RECOMMENDATION ALGORITHMS

We run experiments in order to check and illustrate to what extent and how faithfully our theoretical results are observed in experiments with real data, exposed to empirical variance and the potential violation of our theoretical simplifications. Also, while our analysis so far was in terms of optimal rankings, we aim to examine how this generalizes to the comparison of rankings other than the optimal, as can be returned by common recommendation algorithms. On the other hand, we seek to observe whether either false or true-positive metrics are more robust than the other to evaluation biases, by checking their degree of agreement with unbiased estimates of the corresponding real metric values.

8.1 Algorithms

The particular choice of recommender systems for our experiments is not a critical point: we just need a representative set of well-behaved state-of-the-art algorithms. For this purpose we select several collaborative filtering algorithms implemented in the LibRec library [33]: user-based and item-based kNN with cosine similarity [55], BPoissMF [30] (BPMF in the figures), EALS [34], GBPR [58], ListRankMF [67] (LRMF), PNMF [79], GPLSA [37], RankSGD [39] (RSGD), SLIM [56], SVD++ [41], WBPR [25], and WRMF [38]. Since the optimality of algorithms is not the object

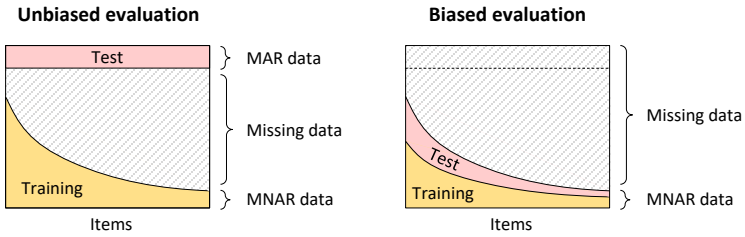


Fig. 11. Data splitting approach into training and test subsets to produce biased (right) and unbiased (left) evaluation with datasets that include MAR data. In Yahoo! R3, the MAR and MNAR subsets are directly supplied in the data release. In CM100k, the MAR data "band" is a 20% random subsample of the whole dataset (which is MAR), and the MNAR subset is formed by including the ratings, in the remaining 80%, where users declared to be familiar with the rated item.

of our analysis here, we simply take the default configuration of these algorithms in the LibRec library, which achieves a reasonable performance in most cases. Some of them do nonetheless underperform, and they are useful in our experiments as well: it is as important for an evaluation methodology to properly identify poorly performing systems as it is to single out the most effective ones. The potential metric distortions we are studying here concern, for instance, the early stages of parameter tuning from a suboptimal starting point as much as the comparison of highly optimized algorithms.

In addition, we include three non-personalized recommendations: ranking by decreasing popularity (Pop), by decreasing average rating (Avg) with Dirichlet smoothing $\mu = 1$ [11], and random (Rnd). Finally, whenever useful, we will include non-personalized versions of the optimal rankings for true or observed metric values as defined by equations 2 to 5. For true metric values, the optimal ranking function (equations 2, 4) is given "oracle" access to test data in order to estimate $p(\text{rel}|\neg\text{train}, i)$.

8.2 Computing Metrics

As mentioned in Section 4.2, we will use the MovieLens 1M [49], Yahoo! R3 [47] and CM100k [11] datasets in our experiments here. Since all the data in MovieLens is naturally biased (MNAR), evaluation metrics computed with this dataset can only be equally biased: we get "observed" metric values in our terminology [11], such as $\widehat{\text{P}}$ and anti$\widehat{\text{P}}$. In contrast, as described earlier in Section 4.2, the Yahoo! R3 and CM100k datasets provide relevance judgments sampled uniformly at random, thus enabling unbiased estimates of real metric values.

We handle the training and test subsets released in Yahoo! R3 as follows. We supply the MNAR training set as input to the evaluated algorithms, and we use the MAR test set as relevance judgments for evaluation; as the test judgments are an unbiased sample of full relevance knowledge, we obtain unbiased estimates of the metric values that would be computed with a full rating matrix – i.e. the "real" metric values. Leaving out the test set supplied in the release, we can also reproduce typical biased measurements – observed metric values – in this dataset by simply splitting the MNAR training set uniformly at random into a training and a test subset, in such a way that both the training and test ratings are MNAR – thus handling the Yahoo! R3 training set in the same way as e.g. MovieLens 1M. Figure 11 illustrates this.

With CM100k, we use the information about familiar music as a reasonable proxy for a MNAR training data, as described in Section 4.2. However, we need to be careful to ensure that we have disjoint training and test sets now. To achieve this, we first uniformly sample 20% of all the data as a test set, thus obtaining MAR relevance judgments for unbiased estimation of real metric values.

Then, to feed the algorithms with MNAR input data, we take the subset of ratings for music that users had already heard before in the remaining 80% as the training set. On the other hand, to reproduce the computation of biased (observed) metric values, we simply take the set of all ratings for familiar music as a MNAR dataset (just as MovieLens or the training set of Yahoo! R3), and we randomly split it into training and (MNAR) test subsets (just as we do for MovieLens and for the Yahoo! training set).

In all three datasets, for observed metric value computation, the split ratio of the MNAR data is 80% ratings for training and 20% for testing, under 5-fold cross-validation. As mentioned in Sections 4.1 and 4.2, we binarize rating values by a minimum relevance threshold value: 4 on MovieLens and Yahoo! R3, and 3 on CM100k, based on the description of what the respective rating scales stand for.

8.3 Observed Metric Disagreements

The top row of Figure 12 quite clearly confirms the contradiction between the observed values of precision and anti-precision. This goes beyond our analysis in terms of optimal rankings: a distinct positive correlation is displayed between both metrics in all three datasets over the 16 different systems, meaning that observed precision and anti-precision disagree in more pairwise system comparisons than those upon which they agree. The observations we find here would correspond to our toy example #1 in Figure 4, and is a consequence of the fact that, for most item pairs, the popularity is steeper than the average rating, and correlates positively with the average rating (two reasons for the observed value of metrics to disagree based on Lemma 1). We checked this for MovieLens 1M earlier in Figure 5, and it is easy to check that similar trends are found in the two other datasets. As an illustration of how our analysis generalizes to other metrics, the bottom row of Figure 12 shows almost identical trends for recall and its anti-metric fallout.

The behavior of the metrics in terms of optimality can thus provide an explanation for their observed overall contradicting trend in system comparisons. In particular, we thus find that:

Conclusion 4. *Observed precision and anti-precision tend to disagree with each other in the comparison between systems (not just in optimal rankings) in offline experiments with common datasets. Only if the popularity distribution were unusually flat would the metrics come to an agreeing trend.*

We also find validation in these results for our formulation of the optimal rankings, which are confirmed as bounds of non-personalized algorithms (popularity, average rating and random) for the respective metric (the optimal ranking for anti$\widehat{\text{P}}$ in the x axis and the optimal for $\widehat{\text{P}}$ in the y axis in Figure 12). We confirm that the optimal recommendation in each metric is worst – or nearly – for the other metric. We further see that popularity is very close to the optimal ranking for $\widehat{\text{P}}$, and is therefore among the worst recommendations in anti$\widehat{\text{P}}$, as predicted in Section 6 (Conclusion #2). Moreover, we see that the findings for precision and anti-precision generalize to another true/false-positive metric pair, namely recall and fallout, for which the observed trends in Figure 12 are equivalent.

We see in the figure that personalized algorithms can do better than the non-personalized optimal, also as one might expect. This is the case in Yahoo! R3, though not in CM100k where collaborative filtering fails to improve over popularity due to data sparsity, as reported in [10]). We nonetheless see that the disagreement between the two metrics generalizes to non-personalized algorithms: for instance, while WRMF is the best system in observed precision in Yahoo! R3, it appears to be the second worst in observed anti-precision.

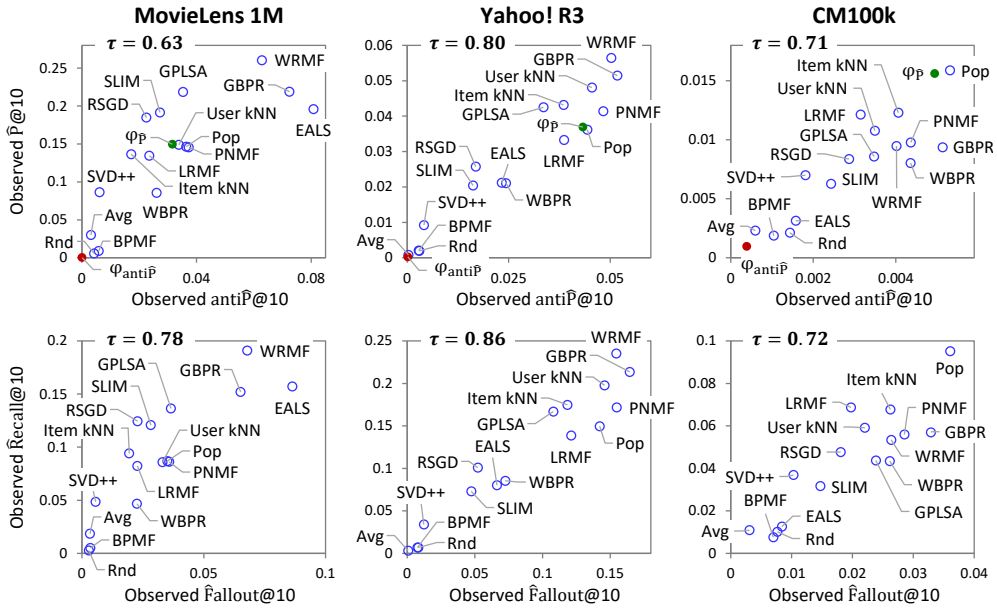


Fig. 12. Evaluation in "observed" true and false-positive metrics on common biased MNAR data. The theoretical optimal recommendations for precision and anti-precision (equations 9 and 10) are shown as a green and red dot, respectively – we omit them in the recall/fallout graphs to avoid confusion because these recommendations optimize different metrics than are shown in these graphs. Kendall τ correlation of the system rankings is shown in each graph. The respective relevance judgment coverage ratios at cutoff 10 are 15.97%, 5.43% and 1.10% (average across systems) for MovieLens 1M, Yahoo! R3 and CM100k, respectively.

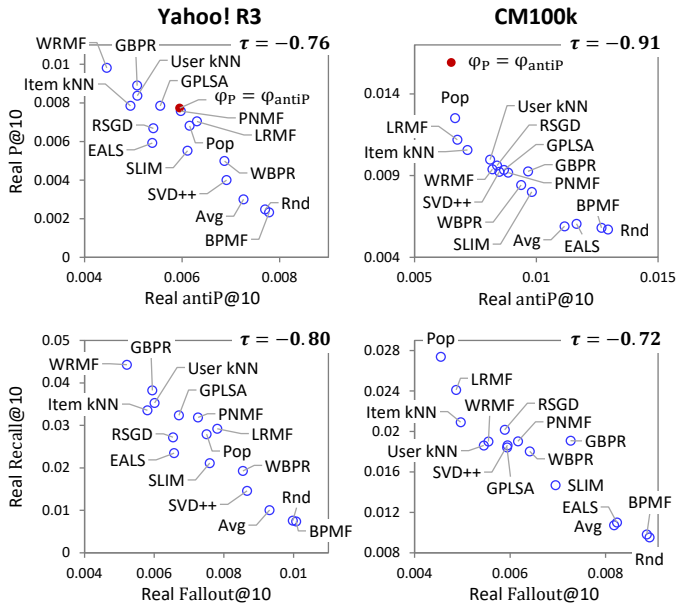


Fig. 13. Evaluation in estimated "real" values of true and false-positive metrics with unbiased MAR test data. The optimal recommendations for real precision and anti-precision are the same (as per Conclusion #1), displayed as a unique red dot – as in Figure 12, we omit them in the recall/fallout graphs to avoid confusion because these recommendations optimize different metrics than are shown in these graphs. Kendall τ correlation of the system rankings is shown in each graph. The relevance judgment coverage ratios at cutoff 10 for these real metric value estimates are 1.23% for Yahoo! R3 and 1.83% for CM100k.

8.4 Real Metric Agreement

We now look at the real metric values, using the two datasets that support unbiased metric estimates: Yahoo! R3 and CM100k, as explained in Section 4.2. Figure 13 shows the results. The same as for $\widehat{\mathrm{P}}$ and anti$\widehat{\mathrm{P}}$ in Figure 12, some personalized algorithms do better than the non-personalized optimal ranking, but the latter is confirmed as an optimal bound of non-personalized recommendations (popularity, average rating and random). We see in the top row of the figure that the negative correlation between real precision and anti-precision is quite strong – i.e. they highly agree in ranking systems. Again, the observation generalizes to recall/fallout as we can see in the bottom row of Figure 13. This is a manifestation of our analytical result in Conclusion #1, where now we see that the exact coincidence of the optimal rankings for precision and anti-precision generalizes to a – not exact but – strong agreement between the two metrics in comparing any two systems other than the optimals:

Conclusion 5. *Real precision and anti-precision tend to agree with each other in the comparison between systems (not just in optimal rankings) in offline experiments with unbiased test data. Based on the theoretical analysis, we may expect this to be independent from the shape of the popularity distribution.*

We thus empirically confirm that we can expect agreement between real precision and anti-precision (Figure 13), and disagreement between their observed values (Figure 12). This means that only one of the metrics can agree with itself in terms of observed and real value, and we would want to know which one does – both situations (where either metric gets the right system comparisons) are theoretically possible according to Lemmas 2 and 3. We check what is observed empirically, by plotting the real and observed values of each metric against each other in Figure 14. We see that while the MNAR measurements of precision are quite consistent with the unbiased MAR estimates (bottom row in the figure), anti-precision seems to suffer from a severe distortion by the MNAR bias (top row), to the point that the system comparisons are almost reversed. Once again, the figure shows equivalent patterns for fallout vs. recall as for precision and anti-precision. We can see that the patterns are quite equivalent.

According to our analysis in Section 7, the strength of the item-dependence in the rating probability, along with the agreement between rating and relevance, and the steepness of the training rating distribution, may play a role in determining which of precision or anti-precision is more consistent with its real value. The inconsistency of anti-precision and the consistency of precision is the situation characterized in Lemma 3 (a) when popularity agrees with but is steeper than relevance. We have seen earlier in Section 7 that this is the case much more often than not in the studied datasets: popularity and relevance agree in 69% of item pairs in CM100k, and the popularity odds is steeper in 90% of item pairs, which may explain what we are observing.

In consonance with our analysis around Lemmas 2 and 3 in Section 7.2, we may also infer from these results that the rating dependence on items may play a stronger role than the dependence on relevance, at least concerning the observed outcomes: if the rating dependence on relevance were strong enough, anti-precision should be consistent with its real value, which is not the case. Even if the lemma premise of conditional relevance independence is not strictly true, the data seems relatively "close to" satisfying the premise, as Figure 10 would suggest earlier in Section 7 when looking at the data, and is again hinted now, judging by the observed trend when comparing actual recommender systems. The observations are therefore in line with Lemma 3 case (a), and example #1a in Figure 8: popularity and relevance tend to agree, and popularity is steeper. Lemma 3 can therefore explain why observed precision agrees with its true value, and anti-precision does not.

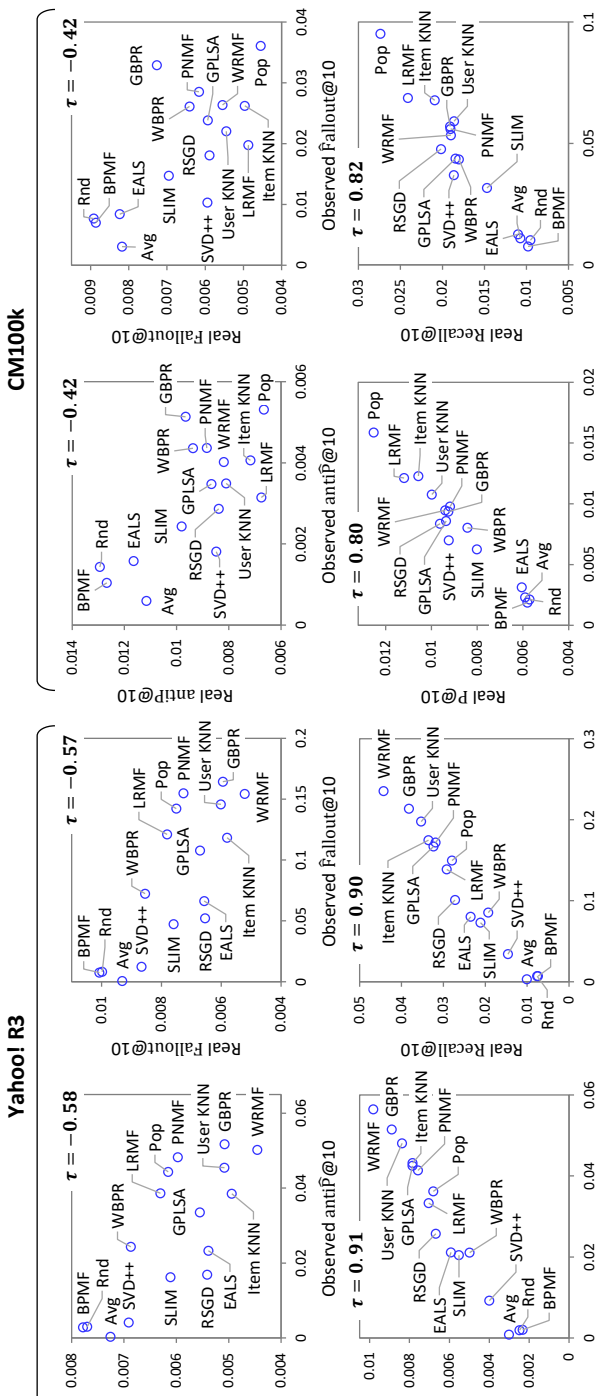


Fig. 14. Agreement and disagreement between observed and true values of true-positive (top) and false-positive (bottom) metrics on Yahoo! R3 (left) and CM100k (right). Kendall τ correlation of the system rankings is shown in the graphs.

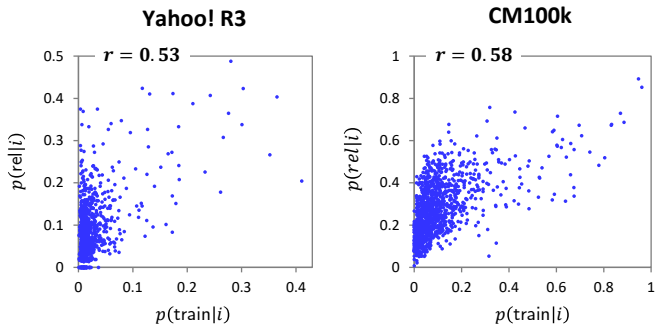


Fig. 15. Rating vs. relevance probability in Yahoo! R3 and CM100k. $p(\text{rel}|i)$ is estimated for each item i as the ratio of positive ratings of i in the MAR test sample.

8.5 Rating against Relevance

We have seen so far that a steep popularity and a positive relation between rating and relevance is plausibly a key factor that can make true-positive metrics be more resilient to the popularity bias than false-positive metrics. Different configurations where popularity is evenly distributed over items are rather unlikely in real settings, for reasons that we discussed in earlier sections. An agreeing trend between rating and relevance is also to be expected in general, as long as people, the media, etc., find, talk about, and engage with things people like more often that things people dislike. Figure 15 further illustrates this type of pattern by the observed correlation between the rating and relevance distributions in Yahoo! R3 and CM100k: we see that the two probabilities agree more often than not (as we had already observed earlier in terms of agreement ratios in Figure 9).

Even though this situation can be quite common, other scenarios are also possible. The relation between rating and relevance might become negative in certain domains or particular situations. This can happen, for instance, if the lowest quality choices are the most advertised (e.g. as a compensation mechanism). Or even a simpler and neutral situation such as an increase in item quality over time: at a steady rate of user engagement the newest and better items would have less user feedback (as they have been less time in the system) than older ones. Without some relevance bias in item exposition and user engagement – or some correction for freshness in the models – our theory would predict that false-positive metrics might better capture the truth than true-positive ones in such cases: rather than recommending the choices that attracted the most praise, we may want to recommend the options that got the least complaints.

It is possible to simulate this type of pattern in datasets such as CM100k by shuffling the rating distribution over items in such a way that the items that more people dislike get a higher number of ratings. Figure 16 shows the result and indeed we see that the false-positive metrics find a clear agreement between observed and real values, just the opposite to the true-positive metrics, which now become misleading: the best systems in observed $\widehat{\text{P}}$@10 and $\widehat{\text{R}}$ecall@10 (GBPR, WBPR, SLIM) are actually the worst (along with PNMF) in real P@10 and Recall@10. These results now correspond to Lemma 3 case (b), and example #1b in Figure 8: popularity is still steeper, but disagrees with relevance.

We may draw a final, simple and practical consequence from these observations and our findings formalized in Section 7 and summarized in Figure 9. Since anti-precision is always consistent if popularity and the relevance distribution disagree, while precision is not always consistent in that case, we may conclude that:

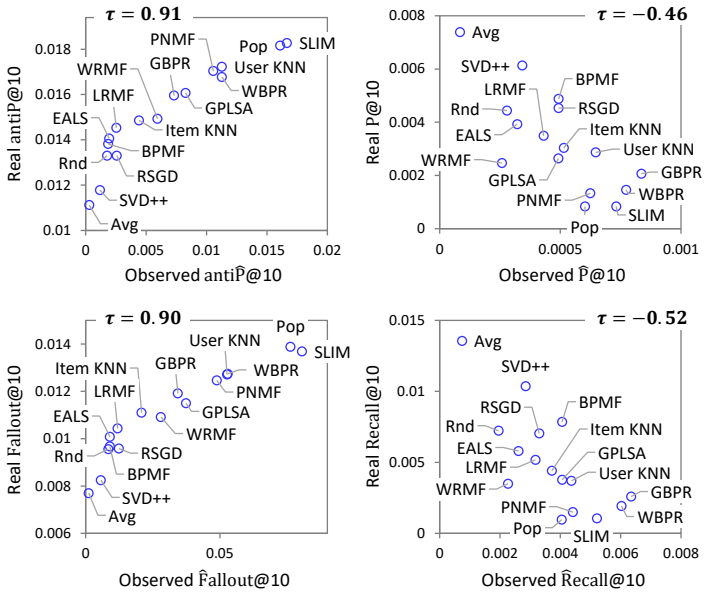


Fig. 16. Agreement between observed and real values of true and false-positive metrics when a negative relation between the distribution of ratings and relevance is simulated in CM100k. Kendall τ correlation shown in the graphs.

Conclusion 6. *If the popularity and the relevance distributions tend to disagree in the top recommended items by two systems, a false-positive metric is likely to be more reliable than a true-positive metric in the comparison of such two systems.*

A collateral finding is worth noting here. We can notice in Figure 16 that all personalized algorithms except SVD++ (and to a negligible extent RSGD and BPMF) perform substantially worse – in terms of real metric values – than random recommendation in the simulated conditions. This means that state of the art algorithms seem to be implicitly relying on a positive correlation between the rating and relevance distributions, to work properly. Even if this correlation seems natural and we can observe it in the datasets we have examined, a word of caution is worth being raised that if this correlation ever breaks, common algorithms could go badly wrong as in Figure 16 – and worse yet, this would go unnoticed in offline evaluation with true-positive metrics. In such cases, if we are able to detect the negative correlation between rating and relevance, we could anticipate the problem and resort to false-positive metrics as being more reliable.

9 GENERALITY

Our initial formal analysis in Sections 5, 6 and 7 involves three simplifications: a specific metric pair (precision and anti-precision), a specific cutoff ($k = 1$), and a specific rating split approach (random). We have then seen in Section 8 that the formal findings generalize empirically to deeper cutoffs (we took $k = 10$ as an example) and an additional true/false-positive metric pair: recall and fallout. Exploring how far our findings can be further generalized is a good subject for future work, that exceeds the current scope and desirable length of the present paper. We take a glimpse in that direction here nonetheless. In particular, we check how our observations generalize to additional false-positive metrics, and to a different rating splitting procedure.

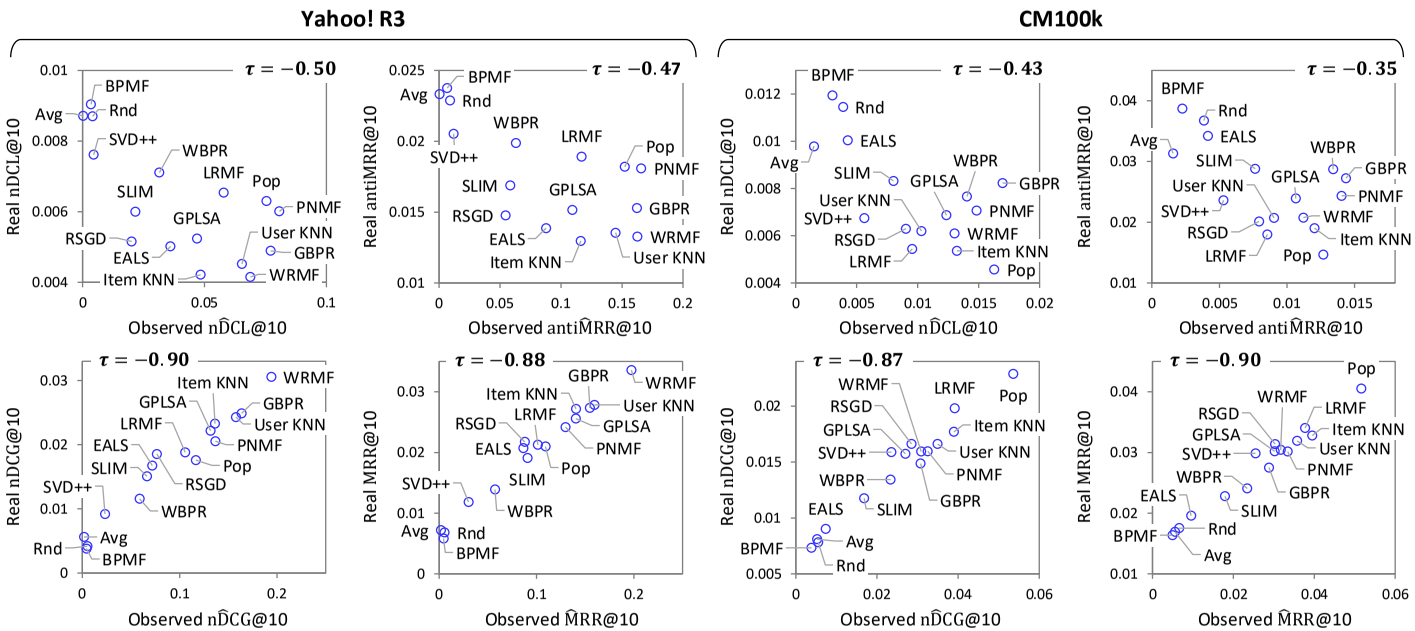


Fig. 17. Agreement and disagreement between observed and real values of nDCG and MRR, and their respective anti-metrics nDCL and anti-MRR on Yahoo! R3 (left) and CM100k (right), extending Figure 14 with these additional metric pairs.

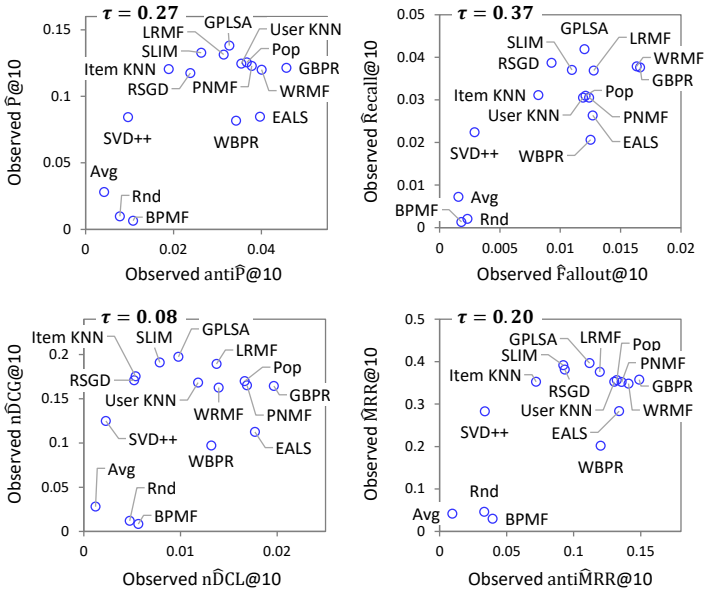


Fig. 18. Evaluation in observed true vs. false-positive metrics on a temporal split of the MovieLens 1M dataset. The measurements are the same as in Figure 12 with a temporal instead of random split.

Figure 17 extends Figure 14 with two additional metric/anti-metric pairs: nDCG and nDCL [24] and MRR and anti-MRR where, the same as other anti-metrics (anti-precision vs. precision and fallout vs. recall), nDCL and anti-MRR are false-positive metrics respectively defined as nDCG and MRR on flipped relevance judgments. We see that the trends are quite equivalent to what we observed earlier for precision and recall. This is just an example, and similar equivalence is observed for all the other measurements shown in the paper this far (Figures 1, 12, 13 and 16).

Figure 18 shows the same measurements as in Figure 12, but applying a temporal – rather than random – split [42] of MovieLens 1M ratings, where the 80% earlier-entered ratings are included in the training set and the remaining 20% are used as test data. We see that the strength of correlation between metrics and anti-metrics decreases considerably compared to the corresponding results displayed in Figure 12. This illustrates the importance of the random split assumption: with a more realistic temporal split, the relevant distributions (popularity, etc.) are not preserved now to the same extent across the data split – an assumption our theoretical derivations in Section 5 relied upon. Yet, we still observe some degree of positive correlation, that is, disagreement between the true and false-positive metrics. The correlation is weakest with nDCG vs. nDCL, based on which these two metrics would seem to have no particular mutual relation.

In spite of the absence or weakness of correlation, we see that the metric pairs do strongly and systematically disagree on specific recommendations such as popularity, the average rating and random recommendation, among others. The average rating and random are among the top 3 best recommendations according to all false-positive metrics, which are at the same time the top 3 worst in true-positives. Likewise, popular recommendations are in the top-right quadrant of the graphs, meaning they are considered good recommendations by true-positive metrics, while they are found poor recommendations by false-positive metrics. Such disagreements thus seem to stand the difference between a random and a temporal rating split. Hence, even if true and false-positive metrics do not disagree in the comparison of *most* system pairs, they do strongly

disagree in many, and our analysis provides an explanation for these. In particular, the opposite effect of the popularity bias on true and false-positive metrics appears to be noticeable in the observed disagreements, even if the bias is now mixed with additional distortions when the time variable (or metric complexities such as the rank discount of nDCG) are not removed from the experiments.

As a final note on the generality of our analysis, making the false positive notion operational (e.g. running experiments involving false positives) requires the availability of negative labels – negative user feedback – for the computation of false-positive metrics. From this perspective, the empirical side of our research cannot be applied in cases when only positive user feedback is available, as all false-positive metrics would yield a value of zero for any system output, providing zero information to the experiment. False-positive metrics would however be a non-problem in these cases.

The availability of negative feedback is on the other hand widespread, in different forms, in many common applications, as discussed in Section 2 [6–8, 17, 26, 27, 29, 46, 58, 59, 73, 80]. In addition to explicit negative ratings and "dislikes", implicit negative signals can be derived from user behavior with delivered recommendations. Dislike can be inferred, for instance, from the absence of engagement with presented recommendations, combined with an understanding of user browsing behavior – e.g. clicks after scrolling can be taken as a clear sign that the user is uninterested in the items they scrolled beyond without clicking [61].

10 WRAP-UP

We have carried out an extensive in-depth study of the situations in which true and false-positive metrics agree or disagree in comparing recommender systems. A system's performance can be assessed from different perspectives. Studies have reported that a bad experience can outperform a good one in an overall assessment; in many business domains, displeasing users can be as much a priority concern (or more) as is pleasing them. Counting the number of false positives instead of true positives can therefore provide a relevant complementary view of recommendation performance.

We discover in our research that missing relevance judgements are a cause of a loss of complementarity between true and false-positive metrics. That the judgements are missing not at random further results in a potential discrepancy between the metrics, which we observe systematically in our experiments and inspection of data at different levels. Our research to this respect addresses two essential high-level questions: when and why do true and false-positive metrics agree with each other, and when and why do the observed measurements agree with the real metric values, for each of these two metric types.

Our theoretical research approach is based on a probabilistic representation of the data involved in recommendation experiments, upon which we express the metrics of interest, we formalize the distinction between complete and incomplete relevance knowledge in computing such metrics, and we represent the rankings that optimize each of them. Reaching this point, our analysis is based on just three basic conditions involving probabilities, random variables, and ranking functions: agreement, steepness, and conditional (in)dependence. Our formal analysis establishes a set of cases and configurations that determine different situations for agreement and consistency between metrics and variants, but does not state which situation is more likely to occur. Inspection of real data and experiments with recommendation algorithms provide insights in that respect, and are consistent with our theoretical findings.

As a final perspective, we explore a simulated atypical situation where the least liked items become the most popular – the results are again consistent with theory. We furthermore put the generality of our analysis to test beyond our simplifications in aspects such as the rating split, metric complexity, and metric depth. The expected trends persist in these additional experiments.

Among the tested assumptions, lifting the random split assumption makes the trends still manifest but weaker, illustrating the natural limits of our analysis.

11 CONCLUSIONS

Our study confirms that false-positive metrics are heavily affected by popularity biases, just as true-positive metrics had been found to be [5, 11, 68]. Paradoxically, false-positive metrics are negatively biased towards popularity: the recommendation of popular items is penalized, just as it was rewarded by true-positive metrics. We find that this results in systematic disagreements between true and false-positive metrics as soon as a popularity bias is present in the test data for recommender system evaluation, as is typically the case with offline data logged from spontaneous user activity.

We further find that true-positive metrics may tend to be more reliable than false-positive metrics as to their correspondence with unbiased evaluation: in common cases where item popularity is not distributed in opposition to relevance, false-positive measurements tend to contradict the metric values we would compute if we had full relevance knowledge. The opposite situation is also possible nonetheless, as we illustrate empirically.

This is the first time, as far as we are aware, that the described biases on false-positive metrics are identified, systematically characterized and explained. Beyond our current findings, avoiding or coping with these biases would be a natural next step in future research, just as debiasing evaluation techniques and algorithms are being researched for true-positive metrics [66, 75]. A straightforward starting point in this direction is to check the effect of these debiasing techniques on false-positive metrics, and on scenarios where the popularity might play in reverse to relevance. Testing our findings on further datasets of different characteristics, domains and scale might likewise complement our insights and provide new ones. A closer look at our analyzed effects on specific algorithm types, parameter configurations, and evaluation goals (e.g. parameter tuning) is also a natural direction worth being pursued.

On the other hand, false-positive metrics have not been analyzed, as far as we are aware, under a user model perspective, as extensively researched for true-positive metrics for over a decade in the IR field [13, 16, 18, 53, 74]. Identifying the user behaviors that false-positive metrics might represent would help better understand the meaning and usefulness of these metrics, their relation to false-positive metrics, and might even motivate the definition of new principled false-positive metrics.

ACKNOWLEDGMENTS

This work was partially supported by the Australian Research Council Discovery (DP190101485), the Australian Technology Network (ATN-LATAM Research Scholarship), and the Spanish Government (grant ref. PID2019-108965GB-I00).

REFERENCES

[1] G. Adomavicius and A. Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17, 6 (June 2005), 734–749.

[2] C. Basu, H. Hirsh, and W. W. Cohen. 1998. Recommendation as classification: using social and content-based information in recommendation. In *Proceedings of the* 15^{th} *National Conference on Artificial Intelligence (AAAI 1998).* AAAI Press, Menlo Park, CA, USA, 714–720.

[3] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs. 2001. Bad is Stronger than Good. *Review of General Psychology* 5, 4 (Dec. 2001), 323–370.

[4] A. Bellogín, P. Castells, and I. Cantador. 2011. Precision-Oriented Evaluation of Recommender Systems: An Algorithmic Comparison. In *Proceedings of the* 5^{th} *ACM Conference on Recommender Systems (RecSys 2011).* ACM, New York, NY,

USA, 333–336.

[5] A. Bellogín, P. Castells, and I. Cantador. 2017. Statistical biases in information retrieval metrics for recommender systems. *Information Retrieval* 20, 6 (July 2017), 606–634.

[6] A. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. 2008. To Swing or Not to Swing: Learning When (Not) to Advertise. In *Proceedings of the* 17^{th} *ACM Conference on Information and Knowledge Management (CIKM 2013)*. ACM, New York, NY, USA, 1003–1012.

[7] Marc Bron, Ke Zhou, Andy Haines, and Mounia Lalmas. 2019. Uncovering Bias in Ad Feedback Data Analyses & Applications. In 3^{rd} *International workshop on Augmenting Intelligence with Bias-Aware Humans in the Loop (HumBL @ WWW 2019)*. ACM, New York, NY, USA, 614–623.

[8] B. Brost, R. Mehrotra, and T. Jehan. 2019. The Music Streaming Sessions Dataset. In *Proceedings of The World Wide Web Conference (WWW 2019)*. ACM, New York, NY, USA, 2594–2600.

[9] C. Buckley and E. M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information. In *Proceedings of the* 27^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*. ACM, New York, NY, USA, 25–32.

[10] R. Cañamares and P. Castells. 2017. A probabilistic reformulation of memory-based collaborative filtering: implications on popularity biases. In *Proceedings of the* 40^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*. ACM, New York, NY, USA, 215–224.

[11] R. Cañamares and P. Castells. 2018. Should I follow the crowd? A probabilistic analysis of the effectiveness of popularity in recommender systems. In *Proceedings of the* 41^{st} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018)*. ACM, New York, NY, USA, 415–424.

[12] R. Cañamares and P. Castells. 2020. On Target Item Sampling in Offline Recommender System Evaluation. In 14^{th} *ACM Conference on Recommender Systems (RecSys 2020)*. ACM, New York, NY, USA, 259–268.

[13] B. Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proceedings of the* 34^{th} *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*. ACM, New York, NY, USA, 903–912.

[14] P. Castells and R. Cañamares. 2018. Characterization of fair experiments for recommender system evaluation: A formal analysis. In *Proceedings of the Workshop on Offline Evaluation for Recommender Systems (REVEAL 2018) at the* 12^{th} *ACM Conference on Recommender Systems (RecSys 2018)*.

[15] P. Castells, N. J. Hurley, and S. Vargas. 2015. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook,* 2^{nd} *edition*, F. Ricci, L. Rokach, and B. Shapira (Eds.). Springer, New York, NY, USA, 881–918.

[16] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the* 18^{th} *ACM Conference on Information and Knowledge Management (CIKM 2009)*. ACM, New York, NY, USA, 621–630.

[17] P. Y. K. Chau, S. Y. Ho, K. K. W. Ho, and Y. Yao. 2013. Examining the effects of malfunctioning personalized services on online users' distrust and behaviors. *Decision Support Systems* 56, C (Dec. 2013), 180–191.

[18] C.L.A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. 2011. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In *Proceedings of the* 4^{th} *ACM International Conference on Web Search and Data Mining (WSDM 2011)*. ACM, New York, NY, USA, 75–84.

[19] P. Cremonesi, F. Garzotto, S. Negro, A. V. Papadopoulos, and R. Turrin. 2011. Looking for "Good" Recommendations: A Comparative Evaluation of Recommender Systems. In *Proceedings of Human-Computer Interaction – INTERACT 2011 –* 13^{th} *International Conference (Interact 2011)*. Springer, New York, NY, USA, 152–168.

[20] P. Cremonesi, F. Garzotto, and R. Turrin. 2013. User-Centric vs. System-Centric Evaluation of Recommender Systems. In *Proceedings of Human-Computer Interaction – INTERACT 2013 –* 14^{th} *International Conference (Interact 2013)*. Springer, New York, NY, USA, 334–351.

[21] C. Elkan. 2001. The foundations of cost-sensitive learning. In *Proceedings of the* 17^{th} *International Joint Conference of Artificial Intelligence (IJCAI 2001)*. Morgan Kaufmann, Burlington, MA, USA, 973–978.

[22] B. Fields. 2011. *Contextualize Your Listening: The Playlist as Recommendation Engine.* Ph.D. Dissertation. Goldsmiths, University of London.

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

[24] E. Frolov and I. Oseledets. 2016. Fifty Shades of Ratings: How to Benefit from a Negative Feedback in Top-N Recommendations Tasks. In *Proceedings of the* 10^{th} *ACM Conference on Recommender Systems (RecSys 2016)*. ACM, New York, NY, USA, 91–98.

[25] Z. Gantner, L. Drumond, C. Freudenthaler, and L. Schmidt-Thieme. 2011. Personalized Ranking for Non-Uniformly Sampled Items. In *Proceedings of the International Conference on KDD Cup 2011 (KDDCUP 2011)*. JLMR.org, 231–247.

[26] J. Garcia-Gathright, B. St. Thomas, C. Hosey, Z. Nazari, and F. Diaz. 2018. Understanding and Evaluating User Satisfaction with Music Discovery. In *Proceedings of the* 41^{st} *Annual International ACM SIGIR Conference on Research*

and Development in Information Retrieval (SIGIR 2018). ACM, New York, NY, USA, 55–64.

[27] A. Germain and J. Chakareski. 2013. Spotify Me: Facebook-assisted automatic playlist generation. In *Proceedings of the IEEE* 15th *International Workshop on Multimedia Signal Processing (MMSP 2013)*. IEEE Press, Piscataway, NJ, USA, 25–28.

[28] A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. 2018. Offline A/B testing for recommender systems. In *Proceedings of the* 11th *ACM International Conference on Web Search and Data Mining (WSDM 2018)*. ACM, New York, NY, USA, 198–206.

[29] D. G. Goldstein, R. P. McAfee, and S. Suri. 2013. The Cost of Annoying Ads. In *Proceedings of the* 22nd *International Conference on World Wide Web (WWW 2013)*. ACM, New York, NY, USA, 459–470.

[30] P. Gopalan, J. M. Hofman, and D. M. Blei. 2015. Scalable Recommendation with Poisson Factorization. In *Proceedings of the* 31st *Conference on Uncertainty in Artificial Intelligence (UAI 2015)*. AUAI Press, Arlington, VA, USA, 326–335.

[31] A. Gruson, P. Chandar, C. Charbuillet, J. McInerney, S. Hansen, D. Tardieu, and B. Carterette. 2019. Offline evaluation to make decisions about playlist recommendation. In *Proceedings of the* 12th *ACM International Conference on Web Search and Data Mining (WSDM 2019)*. ACM, New York, NY, USA, 420–428.

[32] A. Gunawardana and G. Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook, 2nd edition*, F. Ricci, L. Rokach, and B. Shapira (Eds.). Springer, New York, NY, USA, 265–308.

[33] G. Guo, J. Zhang, Z. Sun, and N. Yorke-Smith. 2015. LibRec: A Java Library for Recommender Systems. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the* 23rd *Conference on User Modelling, Adaptation and Personalization (UMAP 2015)*.

[34] X. He, H. Zhang, M.Y. Kan, and T.S. Chua. 2016. Fast Matrix Factorization for Online Recommendation with Implicit Feedback. In *Proceedings of the* 39th *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*. ACM, New York, NY, USA, 549–558.

[35] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22, 1 (Jan. 2004), 5–53.

[36] J. M. Hernández-Lobato, N. Houlsby, and Z. Ghahramani. 2014. Probabilistic Matrix Factorization with Non-Random Missing Data. In *Proceedings of the* 31st *International Conference on Machine Learning (ICML 2014)*. Proc. of Machine Learning Research, Sheffield, UK, 1512–1520.

[37] T. Hofmann. 2004. Latent Semantic Models for Collaborative Filtering. *ACM Transactions on Information Systems* 22, 1 (Jan. 2004), 89–115.

[38] Y. Hu, Y. Koren, and C. Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the* 8th *IEEE International Conference on Data Mining (ICDM 2008)*. IEEE Computer Society, Washington, DC, USA, 15–19.

[39] M. Jahrer and A. Töscher. 2011. Collaborative filtering ensemble for ranking. In *Proceedings of the International Conference on KDD Cup 2011 – Volume 18 (KDDCUP 2011)*. JLMR.org, 153–167.

[40] D. Jannach, L. Lerche, I. Kamehkhosh, and M. Jugovac. 2015. What Recommenders Recommend: an Analysis of Recommendation Biases and Possible Countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (Dec. 2015), 427–491.

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

[42] Y. Koren. 2009. Collaborative Filtering with Temporal Dynamics. In *Proceedings of the* 15th *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009)*. ACM, New York, NY, USA, 447–456.

[43] W. Krichene and S. Rendle. 2020. On Sampled Metrics for Item Recommendation. In *Proceedings of the* 26th *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2020)*. ACM, New York, NY, USA, 1748–1757.

[44] A. Lipani, M. Lupu, and A. Hanbury. 2015. Splitting Water: Precision and Anti-Precision to Reduce Pool Bias. In *Proceedings of the* 38th *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*. ACM, New York, NY, USA, 103–112.

[45] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A General Knowledge Distillation Framework for Counterfactual Recommendation via Uniform Data. In *Proceedings of the* 43rd *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*. ACM, New York, NY, USA, 831–840.

[46] H. Lu, M. Zhang, W. Ma, C. Wang, F. xia, Y. Liu, L. Lin, and S. Ma. 2019. Effects of User Negative Experience in Mobile News Streaming. In *Proceedings of the* 42nd *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*. 705–714.

[47] B. M. Marlin and R. S. Zemel. 2009. Collaborative Prediction and Ranking with Non-random Missing Data. In *Proceedings of the* 3rd *ACM Conference on Recommender Systems (RecSys 2009)*. ACM, New York, NY, USA, 5–12.

[48] B. M. Marlin, R. S. Zemel, S. T. Roweis, and M. Slaney. 2007. Collaborative filtering and the missing at random assumption. In *Proceedings of the* 23rd *Conference on Uncertainty in Artificial Intelligence (UAI 2007)*. 267–275.

[49] F. M. Maxwell and J. A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (Dec. 2015).

[50] S. M. McNee, J. Riedl, and J. A. Konstan. 2006. Being Accurate is not enough: How Accuracy Metrics Have Hurt Recommender Systems. In *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems (CHI 2006)*. ACM, New York, NY, USA, 1097–1101.

[51] E. Mena-Maldonado, R. Cañamares, P. Castells, Y. Ren, and M. Sanderson. 2020. Agreement and Disagreement between True and False-Positive Metrics in Recommender Systems Evaluation. In *Proceedings of the* 43rd *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*. ACM, New York, NY, USA, 841–850.

[52] A. Moffat, F. Scholer, and Z. Yang. [n.d.]. Estimating Measurement Uncertainty for Information Retrieval Effectiveness Metrics. *Journal of Data and Information Quality* 10, 3 ([n. d.]).

[53] A. Moffat and J. Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems* 27, 1 (Dec. 2008).

[54] R. J. Mooney and L. Roy. 1999. Content-Based Book Recommending Using Learning for Text Categorization. In *Proceedings of the* 5th *ACM Conference on Digital Libraries*. ACM, New York, NY, USA, 195–204.

[55] X. Ning, C. Desrosiers, and G. Karypis. 2015. A Comprehensive Survey of Neighborhood-Based Recommender Systems. In *Recommender Systems Handbook,* 2nd *edition*, F. Ricci, L. Rokach, and B. Shapira (Eds.). Springer, New York, NY, USA, 37–76.

[56] X. Ning and G. Karypis. 2011. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In *Proceedings of the IEEE* 11th *International Conference on Data Mining (ICDM 2011)*. IEEE Computer Society, Washington, DC, USA, 497–506.

[57] E. Pampalk, T. Pohle, and G. Widmer. 2005. Dynamic playlist generation based on skip-ping behavior. In *Proceedings of the* 6th *International Conference on Music Information Retrieval (ISMIR 2005)*. 634–637.

[58] W. Pan and L. Chen. 2013. GBPR: Group Preference Based Bayesian Personalized Ranking for One-Class Collaborative Filtering. In *Proceedings of the* 23rd *International Joint Conference of Artificial Intelligence (IJCAI 2013)*. AAAI Press, 2691–2697.

[59] L. A. S. Pizzato, T. Rej, J. Akehurst, I. Koprinska, K. Yacef, and J. Kay. 2013. Recommending people to people: the nature of reciprocal recommenders with a case study in online dating. *User Modeling and User-Adapted Interaction* 23, 5 (Nov. 2013), 447–488.

[60] F. Provost and T. Fawcett. 2001. Robust classification for imprecise environments. *Machine Learning* 42, 3 (March 2001), 203–231.

[61] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How Does Clickthrough Data Reflect Retrieval Quality?. In *Proceedings of the* 17th *ACM Conference on Information and Knowledge Management (CIKM 2008)*. ACM, New York, NY, USA.

[62] S. E. Robertson. 1977. The Probability Ranking in IR. *Journal of Documentation* 33, 4 (Jan. 1977), 294–304.

[63] T. Sakai. 2007. Alternatives to Bpref. In *Proceedings of the* 30th *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*. ACM, New York, NY, USA, 71–78.

[64] T. Sakai and N. Kando. 2008. On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments. *Information Retrieval* 11, 5 (March 2008), 447–470.

[65] P. Sánchez and A. Bellogín. 2018. Measuring anti-relevance: a study on when recommendation algorithms produce bad suggestions. In *Proceedings of the* 12th *ACM Conference on Recommender Systems (RecSys 2018)*. ACM, New York, NY, USA, 367–371.

[66] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proceedings of the* 3rd *International Conference on Machine Learning (ICML 2016)*. Proceedings of Machine Learning Research, Sheffield, UK, 1670–1679.

[67] Y. Shi, M. Larson, and A. Hanjalic. 2010. List-wise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of the* 4th *ACM Conference on Recommender Systems (RecSys 2010)*. ACM, New York, NY, USA, 269–272.

[68] H. Steck. 2010. Training and Testing of Recommender Systems on Data Missing not at Random. In *Proceedings of the* 16th *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*. ACM, New York, NY, USA, 713–722.

[69] H. Steck. 2011. Item Popularity and Recommendation Accuracy. In *Proceedings of the* 5th *ACM Conference on Recommender Systems (RecSys 2011)*. ACM, New York, NY, USA, 125–132.

[70] H. Steck. 2013. Evaluation of recommendations: rating prediction and ranking. In *Proceedings of the* 7th *ACM Conference on Recommender Systems (RecSys 2013)*. ACM, New York, NY, USA, 213–220.

[71] A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudík, J. Langford, D. Jose, and I. Zitouni. 2017. Off-policy evaluation for slate recommendation. In *Proceedings of the* 31^{st} *Conference on Neural Information Processing Systems (NIPS 2017)*. Curran Associates, Inc., Red Hook, NY, USA, 3635–3645.
[72] L. Törnqvist, P. Vartia, and Y. O. Vartia. 1985. How Should Relative Changes be Measured. *The American Statistician* 39, 1 (Feb. 1985), 43–46.
[73] K. Wang, T. Walker, and Z. Zheng. 2019. PSkip: estimating relevance ranking quality from web search clickthrough data. In *Proceedings of the* 25^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2019)*. ACM, New York, NY, USA, 1355–1364.
[74] A. F. Wicaksono and A. Moffat. 2020. Metrics, User Models, and Satisfaction. In *Proceedings of the* 13^{th} *ACM International Conference on Web Search and Data Mining (WSDM 2020)*. ACM, New York, NY, USA, 654–662.
[75] L. Yang, Y. Cui, Y. Xuan, C. Wang, S. Belongie, and D. Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the* 12^{th} *ACM Conference on Recommender Systems (RecSys 2018)*. ACM, New York, NY, USA, 279–287.
[76] E. Yilmaz and J. A. Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the* 15^{th} *ACM International Conference on Information and Knowledge Management (CIKM 2006)*. ACM, New York, NY, USA, 102–111.
[77] E. Yilmaz and J. A. Aslam. 2008. Estimating Average Precision when Judgments are Incomplete. *Knowledge and Information Systems* 16, 2 (Aug. 2008), 173–211.
[78] D. Yin, S. D. Bond, and H. Zhang. 2010. Are bad reviews always stronger than good? Asymmetric negativity bias in the formation of online consumer trust. In *Proceedings of the* 31^{st} *International Conference on Information Systems (ICIS 2010)*. Association for Information Systems, 1–18.
[79] Z. Yuan and E. Oja. 2005. Projective Nonnegative Matrix Factorization for Image Compression and Feature Extraction. In *Proceedings of the* 14^{th} *Scandinavian Conference on Image Analysis (SCIA 2005)*. Springer-Verlag, Berlin, Heidelberg, 333–342.
[80] C. Zhai and J. Lafferty. 2006. A risk minimization framework for information retrieval. *Information Processing and Management* 42, 1 (Jan. 2006), 31–55.

A PROOF OF LEMMAS

We provide here complete proofs for Lemmas 1, 2 and 3. For reading convenience, we shall recall the statement of the lemma before each proof.

A.1 Proof of Lemma 1

We first state and prove the following intermediate lemma, that simplifies the proof of Lemma 1:

Lemma 4. *For any two functions* $f : \mathcal{I} \to \mathbb{R}^+$, $g : \mathcal{I} \to \mathbb{R}^+$, *we have:*

$f \cdot g$ agrees with f *on two items* $a, b \in \mathcal{I} \Leftrightarrow f$ *agrees with* g *or is steeper than* g *for* a, b.

Proof. We need to prove a double implication.

$\Leftarrow$) Let us assume that f agrees with g or is steeper than g for a, b.

If f agrees with g then their product agrees with both. If f disagrees with g, let us assume $f(a) < f(b)$ and $g(a) > g(b)$ – the opposite case is proved symmetrically. If f is steeper than g we have:

$$\frac{|f(a) - f(b)|}{\max(f(a), f(b))} > \frac{|g(a) - g(b)|}{\max(g(a), g(b))} \Leftrightarrow \frac{f(b) - f(a)}{f(b)} > \frac{g(a) - g(b)}{g(a)}$$
$$\Leftrightarrow f(b)g(a) - f(a)g(a) > f(b)g(a) - f(b)g(b) \Leftrightarrow f(b)g(b) > f(a)g(a)$$

Whereby $f \cdot g$ agrees with f.

$\Rightarrow$) Let us assume that $f \cdot g$ agrees with f for a, b.

If f agrees with g then we have proved the lemma. Otherwise, let us assume $f(a) < f(b)$ and $g(a) > g(b)$ – the opposite case is proved symmetrically. Since $f \cdot g$ agrees with f we have

$f(b)g(b) > f(a)g(a)$. Subtracting $f(b)g(a)$ from both sides of the inequality, we get:

$$f(b)g(b) - f(b)g(a) > f(a)g(a) - f(b)g(a) \Leftrightarrow \frac{g(b) - g(a)}{g(a)} > \frac{f(a) - f(b)}{f(b)}$$
$$\Leftrightarrow \frac{|f(a) - f(b)|}{\max(f(a), f(b))} > \frac{|g(a) - g(b)|}{\max(g(a), g(b))}$$

Whereby f is steeper than g. □

We now recall and prove Lemma 1.

Lemma 1. *The optimal rankings for the observed value of precision and anti-precision of any two given items disagree if and only if either of the two conditions hold:*

(a) *The odds of* $p(\text{train}|\cdot)$ *is steeper than* $p(\text{rel}|\text{train}, \cdot)$ *and disagrees with* $p(\text{rel}|\text{train}, \cdot)$ *in comparing the two items.*
(b) *The odds of* $p(\text{train}|\cdot)$ *is steeper than* $p(\neg\text{rel}|\text{train}, \cdot)$*, and disagrees with* $p(\neg\text{rel}|\text{train}, \cdot)$ *in comparing the two items.*

Proof. The lemma is proved by chaining agreements and disagreements, and using Lemma 4 to bring the ranking optimal functions into play, as involving a product of agreeing or disagreeing components. To abbreviate the proof we will use the notation $f \propto g$ to indicate that two functions f and g agree, and $f \propto -g$ to indicate that f and g disagree (over two items at hand in the lemma).

We need to prove a double implication.

⇐) Let us assume that either (a) or (b) is true.

(a) If the odds of $p(\text{train}|\cdot)$ is steeper than $p(\text{rel}|\text{train}, \cdot)$, then $\varphi_{\widehat{\text{P}}}$, as the product of the former two, agrees with $p(\text{train}|\cdot)$ by Lemma 4 (⇐). That is, $\varphi_{\widehat{\text{P}}}(\cdot) \propto p(\text{train}|\cdot)$.

If $p(\text{train}|\cdot)$ disagrees with $p(\text{rel}|\text{train}, \cdot)$, then $\text{odds}(p(\text{train}|\cdot)) \propto p(\text{train}|\cdot) \propto -p(\text{rel}|\text{train}, \cdot) \propto 1 - p(\text{rel}|\text{train}, \cdot) = p(\neg\text{rel}|\text{train}, \cdot)$. That is, $\text{odds}(p(\text{train}|\cdot))$ agrees with $p(\neg\text{rel}|\text{train}, \cdot)$.

If these two functions agree, then by Lemma 4 (⇐) their product, which is equal to $-\varphi_{\text{anti}\widehat{\text{P}}}(\cdot)$, agrees with both of them: $-\varphi_{\text{anti}\widehat{\text{P}}}(\cdot) = \text{odds}(p(\text{train}|\cdot))\, p(\neg\text{rel}|\text{train}, \cdot) \propto p(\text{train}|\cdot) \propto \varphi_{\widehat{\text{P}}}(\cdot)$. That is, the observed precision and anti-precision disagree: $\varphi_{\text{anti}\widehat{\text{P}}} \propto -\varphi_{\widehat{\text{P}}}$.

(b) If the odds of $p(\text{train}|\cdot)$ disagrees with $p(\neg\text{rel}|\text{train}, \cdot)$, then it agrees with $p(\text{rel}|\text{train}, \cdot)$, and by Lemma 4 (⇐), the product $\varphi_{\widehat{\text{P}}}(\cdot) = p(\text{rel}|\text{train}, \cdot)\, \text{odds}(p(\text{train}|\cdot)) \propto p(\text{train}|\cdot)$.

If the odds of $p(\text{train}|\cdot)$ is steeper than $p(\neg\text{rel}|\text{train}, \cdot)$, then by Lemma 4 (⇐) their product $-\varphi_{\text{anti}\widehat{\text{P}}}(\cdot) \propto p(\text{train}|\cdot) \propto \varphi_{\widehat{\text{P}}}(\cdot)$. Hence the optimal rankings disagree: $\varphi_{\text{anti}\widehat{\text{P}}} \propto -\varphi_{\widehat{\text{P}}}$.

⇒) By *reductio ad absurdum*, let us assume the negation of (a) and (b) in the lemma, and let us prove that the optimal rankings must agree.

If both (a) and (b) are false, then either $p(\text{rel}|\text{train}, \cdot)$ or $p(\neg\text{rel}|\text{train}, \cdot)$ disagree with and is steeper than the odds of $p(\text{train}|\cdot)$.

(a) If $p(\text{rel}|\text{train}, \cdot)$ is steeper than the odds of $p(\text{train}|\cdot)$, then their product $\varphi_{\widehat{\text{P}}}(\cdot) \propto p(\text{rel}|\text{train}, \cdot)$ by Lemma 4 (⇐). If $p(\text{rel}|\text{train}, \cdot)$ disagrees with $p(\text{train}|\cdot)$, then we have $p(\neg\text{rel}|\text{train}, \cdot) \propto \text{odds}(p(\text{train}|\cdot))$, and by Lemma 4 (⇐) their product is $-\varphi_{\text{anti}\widehat{\text{P}}}(\cdot) \propto p(\neg\text{rel}|\text{train}, \cdot) \propto -p(\text{rel}|\text{train}, \cdot) \propto -\varphi_{\widehat{\text{P}}}(\cdot)$. Hence $\varphi_{\text{anti}\widehat{\text{P}}} \propto \varphi_{\widehat{\text{P}}}$.

(b) If $p(\neg\text{rel}|\text{train}, \cdot)$ disagrees with $p(\text{train}|\cdot)$ then $\text{odds}(p(\text{train}|\cdot)) \propto -p(\neg\text{rel}|\text{train}, \cdot) \propto p(\text{rel}|\text{train}, \cdot)$ and by Lemma 4 (⇐) their product $\varphi_{\widehat{\text{P}}}(\cdot) = \text{odds}(p(\text{train}|\cdot))\, p(\text{rel}|\text{train}, \cdot) \propto p(\text{train}|\cdot)$.

If $p(\neg\text{rel}|\text{train},\cdot)$ is steeper than the odds of $p(\text{train}|\cdot)$ then by Lemma 4 ($\Leftarrow$) their product $-\varphi_{\text{anti}\widehat{\text{P}}}(\cdot) \propto p(\neg\text{rel}|\text{train},\cdot) \propto -p(\text{train}|\cdot) \propto -\varphi_{\widehat{\text{P}}}(\cdot)$. Hence $\varphi_{\text{anti}\widehat{\text{P}}} \propto \varphi_{\widehat{\text{P}}}$.

□

A.2 Proof of Lemma 2

Lemma 2. *If the probability of rating depends mainly on relevance above any other property of individual items, then both precision and anti-precision are consistent with their true value, in terms of the optimal rankings.*

Proof. If rating is conditionally independent from items given relevance, that is, $p(\text{train}|\text{rel},i) \sim p(\text{train}|\text{rel})$ and $p(\text{train}|\neg\text{rel},i) \sim p(\text{train}|\neg\text{rel})$ for all $i \in \mathcal{I}$, let us first decompose the probability of rating by marginalizing with respect to relevance as follows:

$$p(\text{train}|i) \sim p(\text{train}|\text{rel})\,p(\text{rel}|i) + p(\text{train}|\neg\text{rel})\,(1-p(\text{rel}|i)) = b + (a-b)\,p(\text{rel}|i)$$

where $a = p(\text{train}|\neg\text{rel})$ and $b = p(\text{train}|\text{rel})$ are constants over items.

Now applying this to the ranking function for observed anti-precision (equation 9), and using the independence assumption, we get:

$$\begin{aligned}\varphi_{\text{anti}\widehat{\text{P}}}(i) &\sim -p(\text{train}|\neg\text{rel})\,\frac{p(\neg\text{rel}|i)}{p(\text{train}|i)}\cdot\frac{p(\text{train}|i)}{1-p(\text{train}|i)} \propto -\frac{1-p(\text{rel}|i)}{1-p(\text{train}|i)}\\ &\sim \frac{p(\text{rel}|i)-1}{1-b-(a-b)p(\text{rel}|i)} \propto p(\text{rel}|i)\end{aligned}$$

where in the first step we apply the Bayes rule to $p(\neg\text{rel}|\text{train},i)$, and the last step follows because the function $f(x) = (x-1)\,/\,(c_1+c_2x)$ is monotonically increasing in x if $c_1+c_2 = 1-p(\text{train}|\text{rel}) > 0$. Hence the optimal ranking function for observed anti-precision is equivalent – as a function on items – to the probability of relevance $p(\text{rel}|\cdot)$.

Now applying similar steps to the optimal ranking for true anti-precision, we have:

$$\begin{aligned}\varphi_{\text{antiP}}(i) = p(\text{rel}|\neg\text{train},i) &\sim p(\neg\text{train}|\text{rel})\,\frac{p(\text{rel}|i)}{p(\neg\text{train}|i)} \propto \frac{p(\text{rel}|i)}{1-p(\text{train}|i)}\\ &\sim \frac{p(\text{rel}|i)}{1-b-(a-b)p(\text{rel}|i)} \propto p(\text{rel}|i)\end{aligned}$$

where the last step follows because the function $f(x) = x\,/\,(c_1+c_2x)$ is monotonically increasing in x if $c_1 = 1-p(\text{train}|\neg\text{rel}) > 0$.

The optimal ranking function for true anti-precision is therefore also equivalent to the probability of relevance $p(\text{rel}|i)$, which means it is also equivalent to the optimal ranking for observed anti-precision.

The optimal rankings for true and observed precision are proved to be the same analogously, just flipping ¬rel to rel. This result for true and observed precision was in fact already proven by Cañamares and Castells [11]. □

A.3 Proof of Lemma 3

Lemma 3. *If the probability of rating in the training set is strongly dependent on items beyond their relevance, then:*

(a) *If the probability of rating and relevance agree, then* (i) *the optimal for observed precision is consistent with the true optimal ranking, and* (ii) *anti-precision is consistent in observed and true optimal rankings if and only if non-relevance is steeper than the odds of popularity.*

(b) *If the probability of rating and relevance disagree, then* (i) *the optimal for observed anti-precision is consistent with the true optimal ranking, and* (ii) *precision is consistent in observed and true optimal rankings if and only if relevance is steeper than the odds of popularity.*

Proof. As in the proof of Lemma 1, we shall use $f \propto g$ and $f \propto -g$ as short for agreement and disagreement between two functions f and g (over two items at hand), respectively.

If rating and relevance are conditionally independent given an item, that is, $p(\text{rel}|\text{train}, i) \sim p(\text{rel}|i)$ given any item $i \in \mathcal{I}$, then the ranking functions (equations 6, 8, 9, 10) become:

$$\varphi_{\text{P}}(i) \sim p(\text{rel}|i) \qquad \varphi_{\widehat{\text{P}}}(i) \sim p(\text{rel}|i)\,\frac{p(\text{train}|i)}{1-p(\text{train}|i)} \tag{12}$$

$$\varphi_{\text{antiP}}(i) \sim p(\text{rel}|i) \qquad \varphi_{\widehat{\text{antiP}}}(i) \sim -p(\neg\text{rel}|i)\,\frac{p(\text{train}|i)}{1-p(\text{train}|i)} \tag{13}$$

The consistency of the metrics depends on whether $p(\text{train}|i)$ agrees with $p(\text{rel}|i)$ or not.

(a) If $p(\text{train}|\cdot)$ and $p(\text{rel}|\cdot)$ agree, then the odds of $p(\text{train}|\cdot)$ also agrees with $p(\text{rel}|\cdot)$, and by Lemma 4 their product $\varphi_{\widehat{\text{P}}}$ agrees with both and we have $\varphi_{\widehat{\text{P}}}(\cdot) \propto p(\text{rel}|\cdot) \sim \varphi_{\text{P}}(\cdot)$, whereby precision is consistent, thus proving (i).

The consistency of anti-precision further depends on the popularity steepness: if its odds is steeper than $p(\neg\text{rel}|\cdot)$ then $\varphi_{\widehat{\text{antiP}}} \propto -\varphi_{\widehat{\text{P}}}$ by Lemma 1 (b), because popularity disagrees with non-relevance, since we are assuming $p(\text{train}|\cdot) \propto p(\text{rel}|) \propto p(\text{rel}|) - 1 \propto -p(\neg\text{rel}|\cdot)$. By transitivity, $\varphi_{\widehat{\text{antiP}}} \propto -\varphi_{\widehat{\text{P}}} \propto -\varphi_{\text{P}} \propto -\varphi_{\text{antiP}}$ (because we just proved (i) $\varphi_{\widehat{\text{P}}}(\cdot) \propto \varphi_{\text{P}}(\cdot)$).

By an analogous rationale, if the popularity odds is not steeper than $p(\neg\text{rel}|i)$, then anti-precision is consistent, and we have (ii).

(b) If $p(\text{train}|i)$ disagrees with $p(\text{rel}|i)$, then it agrees with $1 - p(\text{rel}|i) = p(\neg\text{rel}|i)$, and this case is just symmetric to the previous one, swapping relevance and non-relevance.

□

B FLAT DISTRIBUTIONS

We analyze here the equality cases that where excluded hitherto in all definitions and Lemmas 1 and 3: the cases where the probabilities we have studied in our analysis are equal for two items. These situations are rather infrequent, but we study them here for the sake of completeness. Again, in the following we use $f \propto g$ and $f \propto -g$ to denote agreement and disagreement between two functions f and g respectively.

B.1 Flat Popularity

The case where $p(\text{train}|a) = p(\text{train}|b)$ for two items $a, b \in \mathcal{I}$ represents an extreme where the popularity bias is absent in the data, at least between a and b. Considering this equality in equations 8 and 10 we get $\varphi_{\widehat{\text{P}}}(\cdot) \propto p(\text{rel}|\text{train}, \cdot)$ and $\varphi_{\widehat{\text{antiP}}}(\cdot) \propto -p(\neg\text{rel}|\text{train}, \cdot) \propto \varphi_{\widehat{\text{P}}}(\cdot)$. That is, observed precision and anti-precision agree on the optimal ranking. This is in line with Lemma 1, with an absent popularity bias as a particular case of a weak bias.

If rating is conditionally independent from relevance given items, as in the premise of Lemma 3, then with flat popularity Equation 12 becomes $\varphi_{\widehat{\text{P}}}(\cdot) \propto p(\text{rel}|\cdot) \sim \varphi_{\text{P}}(\cdot)$, and Equation 13 gives $\varphi_{\widehat{\text{antiP}}}(\cdot) \propto -p(\neg\text{rel}|\cdot) \propto p(\text{rel}|\cdot) \sim \varphi_{\text{P}}(\cdot)$. There is unanimous agreement among all four optimal rankings.

B.2 Flat Relevance

If $p(\text{rel}|\text{train}, a) = p(\text{rel}|\text{train}, b)$, then by equations 8 and 10 we get $\varphi_{\widehat{\text{P}}}(\cdot) \propto p(\text{train}|, \cdot)$ and $\varphi_{\widehat{\text{antiP}}}(\cdot) \propto -p(\text{train}|\cdot) \propto -\varphi_{\widehat{\text{P}}}(\cdot)$. That is, observed precision and anti-precision disagree on the optimal ranking. This is again in line with Lemma 1, where here popularity is "infinitely" steeper than the average rating and hence makes precision and anti-precision fully opposed.

For Lemma 3, with the relevance independence premise, flat relevance, and equations 12 and 13, we have that $\varphi_{\text{P}}(a) \sim \varphi_{\text{P}}(b)$ and $\varphi_{\text{antiP}}(a) \sim \varphi_{\text{antiP}}(b)$. That is, a and b are tied in the optimal rankings for the true metric values. However, $\varphi_{\widehat{\text{P}}}(\cdot) \propto p(\text{train}|\cdot)$ and $\varphi_{\widehat{\text{antiP}}}(\cdot) \propto -p(\text{train}|\cdot)$ are opposed to each other for a and b, which we may count as a soft discrepancy with the optimal true metric rankings: observed precision and anti-precision would each rank the opposite item on top of the other, while this is indifferent to the true metric values – unless popularity were also flat and the items would then be tied in all four optimal rankings.