Agreement and Disagreement between True and False-Positive Metrics in Recommender Systems Evaluation

Elisa Mena-Maldonado
RMIT University
elisa.mena.maldonado@rmit.edu.au

Rocío Cañamares
Universidad Autónoma de Madrid
rocio.cannamares@uam.es

Pablo Castells
Universidad Autónoma de Madrid
pablo.castells@uam.es

Yongli Ren
RMIT University
yongli.ren@rmit.edu.au

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

ABSTRACT

False-positive metrics can capture an important side of recommendation quality, focusing on the impact of suggestions that are disliked by users, as a complement of common metrics that only measure the amount of successful recommendations. In this paper we research the extent to which false-positive metrics agree or disagree with true-positive metrics in the offline evaluation of recommender systems. We discover a surprising degree of systematic disagreement that was occasionally noted but not explained in the literature by previous authors. We find an explanation for the discrepancy between the metrics in the effect of popularity biases, which impact false and true-positive metrics in very different ways: instead of rewarding the recommendation of popular items, as with true-positive, false-positive metrics penalize the popular. We determine precise conditions and cases in the general trends, with a formal explanation for our findings, which we confirm and illustrate empirically in experiments with different datasets.

CCS CONCEPTS

• Information systems → Recommender systems • Information systems → Evaluation of retrieval results.

KEYWORDS

Recommender systems; evaluation; metrics; false positives; popularity bias; non-random missing data.

ACM Reference format:
Elisa Mena-Maldonado, Rocío Cañamares, Pablo Castells, Yongly Ren, Mark Sanderson. 2020. Agreement and Disagreement between True and False-Positive Metrics in Recommender Systems Evaluation. In *Proc. of the 43rd ACM International Conference on Research and Development in Information Retrieval (SIGIR 2020). ACM, New York, NY, USA, 10 pages.* https://doi.org/10.1145/3397271.3401096.

1 Introduction

Matching what users like is a primary prerequisite for recommendation to be successful: it is commonly referred to as the accuracy of recommendation. Accuracy is generally measured as a function of the number of recommended items that users liked: the true positives, which can be counted, cut off, weighted, averaged, etc. Example true-positive metrics include precision, recall, mean reciprocal rank (MRR), or normalized discounted cumulative gain (nDCG).

Our focus is the flip side of accuracy: recommended items that users disliked –the false positives. This is not a common perspective, yet it has been occasionally considered and/or studied in the field [12,13,17,41], and is common in evaluation practice in specific business domains. We consider whether false-positive metrics capture anything different from true-positive. We find not just differences but blatant disagreements, the causes of which we then investigate. We further consider whether the disagreements are a matter of a complementarity in perspectives, or perhaps one type of metric is just delivering more correct and reliable measurements than the other.

We find that the agreement between true and false-positive metrics is tightly related to missing relevance information, and the fact that this information is missing not at random (MNAR) [30, 45,46]. The effect of such biases in recommendation algorithms and offline evaluation has become the object of growing research in the field [3,7,9,27,44,49]. In particular, algorithms and metrics have been found to be biased to favor the recommendation of popular items, beyond their objective quality, and progress has been made in managing these effects. But how such biases may affect false-positive metrics has not been studied or addressed, as far as we are aware.

We address these questions through a theoretical analysis based on the formalization of expected biased and unbiased metric values, and the rankings that optimize them. We find 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. We also find an explanation for the disagreements between false and true-positive metrics in recommender system comparisons, and we identify key elements to discern whether one of the two types of metrics may be more reliable than the other, in terms of capturing the underlying truth beyond the biases. We confirm and illustrate our analytical findings with empirical observations over different publicly available datasets.[1]

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 dec-

SIGIR'20, July 25–30, 2020, Virtual Event, China.

ACM ISBN 978-1-4503-8016-4/20/07...$15.00.
DOI: https://doi.org/10.1145/3397271.3401096

[1] The source code implementing all the experiments described in this paper is available at https://github.com/elikary/sigir2020.

ades [10,15,32]. 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 [23]. 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 [42]. Some attention has been paid to the flip side: the (evidence of) false positives [12,17,41]. 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 effect on the user experience. False positives are a clear concern in specific domains such as automatic music playlist generation [15]. Here, users may commonly tolerate background music that is just nice, but they may be annoyed by an occasional unpleasant track. As a consequence, the skip rate is a common metric in music and video streaming [36]. Skip behavior is also a common signal in recent challenges as a target for prediction, and/or as part of released data for evaluation (e.g. [5]). Some authors have likewise applied metrics of least non-relevant music to evaluate playlists [19].

The skip rate has been similarly used in Web search to assess the cost in reading and skipping effort of non-relevant results [48]. Dating online is also often mentioned as a domain where false positives involve a significant cost [37]. Information Retrieval (IR) metric frameworks have been likewise developed that consider the cost and benefit involved in delivering relevant and non-relevant documents [53]. 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 [14,38]. From a wider perspective, scholars have studied how a bad recommendation can hurt user trust [11]. 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 [2,51].

2.2 False-Positive Metrics in Recommendation

The recommender systems literature is rich in accuracy evaluation methodologies [4,7,23,42,46]. 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 [42], which employ a false-positive metric –fallout– in the x axis. Frolov and Oseledets [17] and Sánchez and Bellogín [41] 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 [38] is the anti-metric for recall, and the ratio of returned non-relevant items (so-called anti-precision [17,29, 41]) is the anti-metric for precision. Fallout his also occasionally re-ported among other evaluation metrics in some work [12,13]. In IR, Lipani et al. [29] 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 [30,45], where heavy popularity biases pervade the data, impacting the algorithms and metrics that assess their accuracy. Important progress has been made in the last decade in confirming, measuring, explaining and coping with such popularity biases [7,9,27,44, 46,49]. 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.

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 [38] or anti-precision [29], can suitably meet this purpose; they are defined as:

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

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 a song in a music streaming session is taken as a sign of non-relevance, then antiP is the skip rate, i.e. the ratio of played songs that were skipped [5,15,35].

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:

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \Longrightarrow \text{antiP} = 1 - \text{P}$$

Figure 1 top-left illustrates this relationship in the metric values of antiP@10 vs. P@10 for a set of collaborative filtering algorithms (detailed later in Section 5.2) on MovieLens 1M [31], a widely used dataset example. The metrics in these graphs are measured taking so-called condensed rankings [6,40,50], where unrated (unjudged) items are excluded from the evaluated rankings before computing the metrics. 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 a different denominator in the two metrics. We can see in Figure 1 top-right

that they are still strongly negatively correlated. Since true-positive and false-positive metrics are inversely oriented (the lower antiP and fallout the better), this negative correlation means agreement on which of almost every pair of systems is best.

These example observations are obtained however with an experimental option (condensed rankings) that makes relevance knowledge artificially complete in the ranking top. Offline recommender system evaluation is generally conducted with highly incomplete relevance knowledge, and even though condensed rankings have been occasionally used in the literature [4,9,23,48], they are not generally considered the best option: they result in massive losses in effect size and statistical significance, and a deviation from the actual task that the evaluated systems are meant to solve –the systems need to rank all the items in the dataset, and not just the judged ones. In fact, condensed rankings just do not work in the common case when only positive ungraded user feedback is available for evaluation: all systems would get the same metric score in that case. Full rankings is therefore by far the most common option in recommender system evaluation nowadays [3,4,42].

Figure 1 bottom-left illustrates what happens with this more common experimental configuration. 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.[2] 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 5 on further datasets.

The contradiction is 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 TP + FP + U, where U denotes the number of unjudged returned items –and the metrics no longer add to 1. For instance, for the systems evaluated in Figure 1, the percentage of unrated items U at a top 10 cutoff is 83.16% on average over all the systems in the figure: most of the relevance knowledge is missing.

The disagreement is however not fully explained by this knowledge incompleteness: if judgments were simply missing at random, 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 [7,30,44,45] 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 popularity biases in recommender system evaluation [7]. We will then confirm and illustrate our theoretical findings with experiments and empirical observations under different angles.

4 Formal Analysis

4.1 Notation and Preliminaries

Given a set of users $\mathcal{U}$ and a set of items $\mathcal{I}$, we can formalize key elements in evaluation experiments by defining binary random

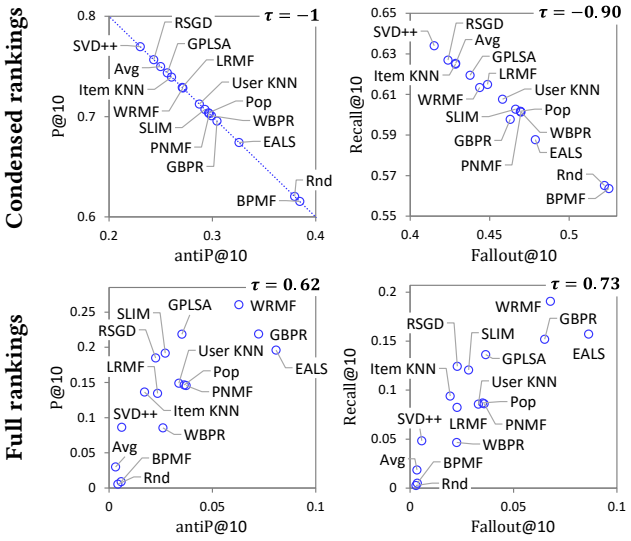

Figure 1: Anti-precision vs. precision (left) and fallout vs. recall (right), for condensed (top) vs. full (bottom) rankings in MovieLens 1M. Kendall τ correlation is shown for each plot.

variables in $\mathcal{U} \times \mathcal{I}$ that describe relationships between users and items [7]: 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.

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

Based on these variables, using rel as abbreviation for $rel = 1$, (and same for $train$ and $test$), we can express meaningful probabilities, such as $p(rel|i)$, denoting the ratio of users who like item $i \in \mathcal{I}$; $p(train|i)$, the ratio of users that the system has observed interacting with i –that is, the "popularity" of the item [7,44]; and $p(rel|train, i)$, the ratio of observed interactions involving item i that evidence a positive preference –the average binarized rating.

4.2 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 [17,29,41] 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 5.

[2] 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.

The same as prior work [7,45] distinguished between the true and observed values of true-positive IR metrics (such as precision and nDCG), we can make the same distinction for false-positive metrics: the true 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" [7]. We shall use $\widehat{\text{P}}$ and anti$\widehat{\text{P}}$, with a "hat", to refer to the observed value of the respective metrics.

4.2.1 True-Positive Optimals. Cañamares and Castells [7] proved that the optimal recommendation that maximizes true P@1 ranks items by non-increasing value of the following ranking function:

$$\varphi_{\text{P}}(i) = p(rel|\neg train, i) \tag{1}$$

That is, ranking the items $i \in \mathcal{I}$ by decreasing order of $\varphi_{\text{P}}(i)$ produces a recommendation R that maximizes P@1 of R in expectation. The reader is referred to [7] 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* [39]), 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 $\widehat{\text{P}}$@1 is defined by:

$$\begin{aligned}\varphi_{\widehat{\text{P}}}(i) &= p(rel, test|\neg train, i) \\ &= \frac{p(rel, test, \neg train|i)}{p(\neg train|i)} \propto p(rel|train, i)\ \frac{p(train|i)}{1 - p(train|i)}\end{aligned} \tag{2}$$

where '$\propto$' denotes rank-equivalence, and in the second step we have applied Bayesian inversions plus the fact that $test \wedge \neg train = test$ (since a test rating is by definition not in training), and $p(test|rel, i) \propto p(train|rel, i)$ 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 [7].

Note that in [7] 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 2 above), because this represents the input that recommender systems can see, and it is more convenient for our line of analysis.

4.2.2 False-Positive Optimals. Given that anti-precision can be defined as precision on flipped relevance [17,29,41], we can directly infer that the optimal rankings that minimize anti-precision are defined by: 1) replacing rel for $\neg rel$ in equations 1 and 2; 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 lower the better). Thus, the optimal ranking functions for anti-precision are as follows:

$$\varphi_{\text{antiP}}(i) = -p(\neg rel|\neg train, i) \propto p(rel|\neg train, i) \tag{3}$$

$$\begin{aligned}\varphi_{\text{anti}\widehat{\text{P}}}(i) &= -p(\neg rel, test|\neg train, i) \\ &\propto -p(\neg rel|train, i)\ \frac{p(train|i)}{1 - p(train|i)}\end{aligned} \tag{4}$$

where we apply similar steps as in equation 2. From equations 1 and 3 a first conclusion follows right away:

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

The optimal rankings for true P and antiP being identical means that the two metrics agree on the comparison of any recommendation to the optimal and worst rankings (the latter being the inverse of the former). This agreement in comparisons to the extremes may make us expect that perhaps the metrics would tend to agree, at least as a general trend, in the comparisons between systems in between the two extremes. We will check this empirically in Section 5.

In contrast, we see in equations 2 and 4 that the optimal rankings in observed metric values are not quite the same. The optimal ranking function for observed anti$\widehat{\text{P}}$ (equation 4) is the product of:

i. The opposite of the "negated" rating $-p(\neg rel|train, i)$. This is a "double negation" of the corresponding term in equation 2.
ii. The popularity odds $p(train|i)/\big(1 - p(train|i)\big)$, a monotonically increasing function of popularity $p(train|i)$. The exact same component is present in equation 4.

Having popularity (in the term ii.) as a component of the ranking function, multiplied by a negative number (as is the term i.) 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.*

Interestingly, this is the exact opposite of the behavior of true-positive metrics [7], in which offline evaluation tends to reward the recommendation of popular items. In the next section, we analyze further consequences of such opposing trends.

4.3 Popularity Bias in False Positives

As noted, the difference in the optimal rankings for observed precision and anti-precision is in the $-p(\neg rel|train, i)$ term for observed anti$\widehat{\text{P}}$ in equation 4, in place of $p(rel|train, i)$ for observed $\widehat{\text{P}}$ in equation 2. 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 very 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 analyze 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 condition for the agreement or disagreement of optimal rankings over a given pair of items can be formalized by the

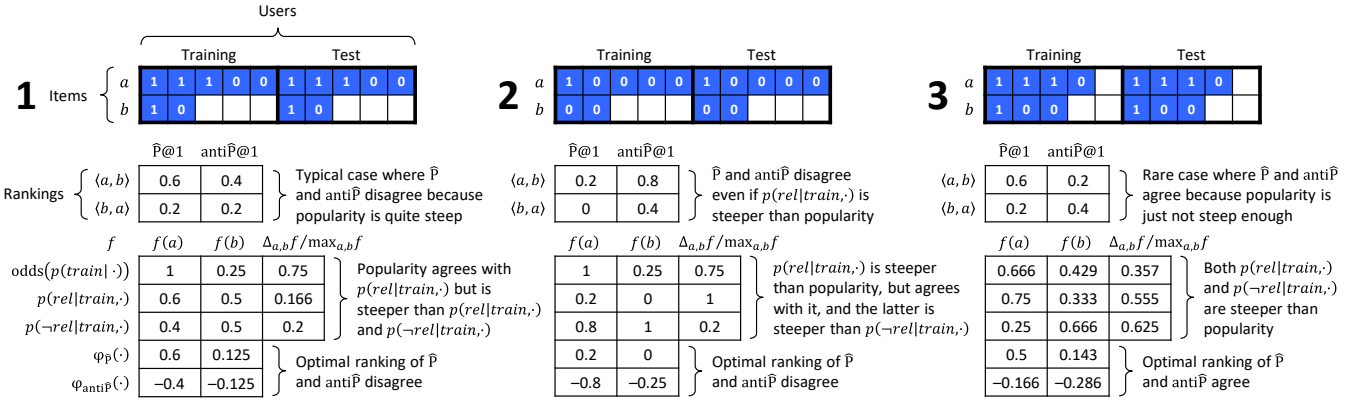

Figure 2: Toy examples illustrating the formal analysis for observed metric values. Blue cells represent ratings, and white cells unrated items. Calculations are straightforward with equations 2, 4, 5, and the definition of $p(train|i)$ and $p(rel|train,i)$ in Section 4.1. For instance, in example #1,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

following definition and lemma.

Definition – Given two functions $f, g: \mathcal{I} \to \mathbb{R}^+$, 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 ratio is higher:

$$\frac{\Delta_{a,b}f}{\max_{a,b}f} = \frac{|f(a) - f(b)|}{\max\big(f(a), f(b)\big)} > \frac{|g(a) - g(b)|}{\max\big(g(a), g(b)\big)} = \frac{\Delta_{a,b}g}{\max_{a,b}g} \quad (5)$$

Lemma – The optimal rankings for precision and anti-precision of any two given items disagree if and only if either of the two conditions hold:

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

Proof – By "disagreeing" we mean here that one item is more popular but has a lower average rating than the other. The lemma is proved by considering all possible cases of agreement or disagreement between popularity and the relevance density, and simple algebraic manipulation of inequalities, taking into account that $p(rel|train,i)$ and $p(\neg rel|train,i)$ always disagree with each other since $p(\neg rel|train,i) = 1 - p(rel|train,i)$. □

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 also bad (in 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 –and will hence contribute higher scores in both true and false-positives, thus producing disagreements between the two: while true-positive metrics would suggest ranking a before b, false-positive metrics would advise the opposite. Only if b had such an extremely higher (or lower) average rating than a might it add up more positive (or negative) ratings than a, in such a way that both metrics would agree to rank b before a (or a before b). We illustrate next such cases and intuitions along with the lemma through toy examples, followed by measurements on real data.

4.3.1 Toy Examples. Figure 2 shows three examples that illustrate the patterns characterized by the lemma. We consider we only have two items a and b, and ten users. Let us assume 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 exact 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 $\widehat{\text{P}}$@1 and anti$\widehat{\text{P}}$@1 and, as is not uncommon [3], in the example we only compute the metrics over target users who have at least some test rating. In this setting, $\widehat{\text{P}}$@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$\widehat{\text{P}}$@1 is the ratio of target users who have a negative one. The three examples in Figure 2 illustrate different cases in how the relation between popularity and the average rating determine 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: $p(rel|train,b)$ is 100% smaller than $p(rel|train,a)$, 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: $p(\neg rel|train,a)$ is just 20% smaller than $p(\neg rel|train,b)$. As a consequence, precision and anti-precision still disagree, confirming the lemma.

Finally, **example #3** represents an atypical case where popularity is rather flat, more than both the average rating and its complement, in such a way that now 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.

4.3.2 Observations on Real Data. For our illustrative purpose we take MovieLens 1M [31] as a common example (equivalent trends

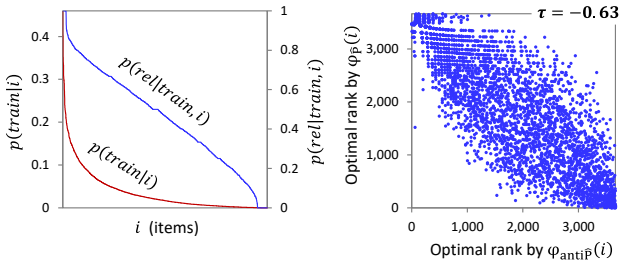

Figure 3: Comparison of the popularity and average rating distributions over items in MovieLens 1M (left), and how this results in an opposing trend between the optimal rankings for observed precision and anti-precision (right). The curves in the left graph are sorted in decreasing order of $p(train|i)$ and $p(rel|train,i)$ 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.

Table 1: Frequency of agreement / disagreement and steepness in MovieLens 1M.

$p(train\|\cdot)$ agrees with...	The odds of $p(train\|\cdot)$ is...	% item pairs	$\varphi_{anti\bar{P}}$ vs. $\varphi_{\bar{P}}$
$p(rel\|train,\cdot)$	steeper than $p(\neg rel\|train,\cdot)$	52.6%	Disagree
	less steep than $p(\neg rel\|train,\cdot)$	8.6%	Agree
$p(\neg rel\|train,\cdot)$	steeper than $p(rel\|train,\cdot)$	29.8%	Disagree
	less steep than $p(rel\|train,\cdot)$	9.0%	Agree

are observed in other similar datasets), taking rating values ≥ 4 as indicative of relevance, and < 4 as reflecting non-relevance. Table 1 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 $\varphi_{\bar{P}}$ and $\varphi_{anti\bar{P}}$ lies in the steepness of the popularity distribution, which is higher than the steepness of both $p(rel|train,i)$ and $p(\neg rel|train,i)$ 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 ($p(train|i)$ over i) are commonly exacerbated by a variety of exogenous factors in how users discover choices [7], some of which are further subject to self-reinforcement [16]. These factors do not affect intrinsic user tastes ($p(rel|train,i)$ over i) to any comparable extent. The popularity steepness is further amplified by the odds function $p/(1-p)$ in equations 2 and 4, 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, the lemma states that $\varphi_{\bar{P}}$ and $\varphi_{anti\bar{P}}$ still have a chance to disagree, if popularity agrees with the steepest 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 were unusually flat the optimals might tend to agree.*

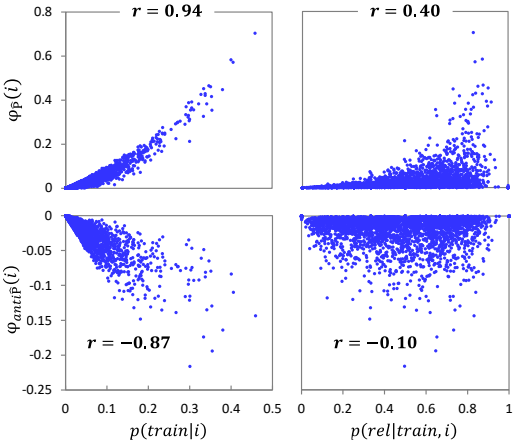

Figure 4: Comparison of the optimal ranking functions for observed precision (top) and anti-precision (bottom) against their two main components: popularity (left) and the average rating (right). Each dot in the graphs is an item, and the values are computed on the MovieLens 1M rating data. Pearson correlation 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.

Figures 3 and 4 further illustrate our line of analysis. Figure 3 right confirms that indeed the optimal rankings for observed precision and anti-precision are quite in opposition to each other in MovieLens 1M. Figure 3 left shows how the average rating (and therefore its complement) has a rather smooth linear decrease, while the popularity distribution $p(train|i)$ has a much more aggressive decrease in comparison. Figure 4 bottom shows how popularity is stronger than the relevance ratio when multiplied in $\varphi_{anti\bar{P}}$: the ranking function has a strong (negative) correlation with popularity ($p(train|i)$, left), and a very weak correlation with $p(rel|train,i)$ (right). Since the popularity odds is multiplied by a negative value $-p(\neg rel|train,i)$, $\varphi_{anti\bar{P}}(i)$ decreases with popularity (Figure 4 bottom left), and popular items should therefore be ranked low for an optimal ranking. The opposite is the case for $\varphi_{\bar{P}}$ (Figure 4 top): it very strongly correlates with popularity (left) because it is multiplied by a positive number $p(rel|train,i)$.

4.3.3 Which Metric is Right? Our analysis 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 true 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 true values, then only one can be correct in its observed value –the question is, which one? Answering this question on a formal basis is a challenge that we envision as future work. As a step in that direction, we seek empirical insights on the issue in the next section.

5 Empirical Observations

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 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 true metric values.

5.1 Data

In addition to MovieLens 1M [31], which we used in previous sections, we shall use the Yahoo! R3 [30] and CM100k [7] datasets, which provide relevance judgments sampled uniformly at random, thus enabling unbiased estimates of true metric values. The details of the datasets are shown in Table 2.

The Yahoo! R3 data includes a training set that was collected from spontaneous user interaction with music in the Yahoo! LaunchCast streaming service (hence training data are MNAR); and a test set containing ratings that each user was asked to enter for ten items sampled uniformly at random [30] (hence test ratings are MAR –missing at random). Metrics computed on this data configuration are thus unbiased estimates of the values that would be computed with a full rating matrix (i.e. the "true" metric values). We can also reproduce typical biased measurements ("observed" metric values such as $\widehat{\mathrm{P}}$ and anti$\widehat{\mathrm{P}}$) in this dataset by splitting the training set uniformly at random into a training and a test subset –in such a way that test ratings are MNAR.

The CM100k data includes about 100 ratings per user for music selected uniformly at random. As a test set, we uniformly sample 20% of this data, thus obtaining MAR relevance judgments for unbiased estimation of true metric values. However, the remaining subset for training is also MAR. In order to reproduce MNAR input data as in real recommendation settings, we use information available in this dataset about whether users were familiar with the music or not before being surveyed, as in [7]: we take as training data only the non-test ratings for music that users had already heard before, as a proxy for spontaneous user-item interaction. On the other hand, in order to reproduce the computation of biased (observed) metric values, we simply take the set of all ratings for familiar music as a MNAR dataset, which we randomly split into training and (MNAR) test subsets.

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. We binarize rating values based on a minimum relevance threshold value, which is 4 on MovieLens and Yahoo! R3, and 3 on CM100k.

5.2 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 [21]: user-based and item-based kNN with cosine similarity [33], BPoissMF [20] (BPMF in the figures), EALS [22], GBPR [36], ListRankMF [43] (LRMF), PNMF [52], GPLSA [24], RankSGD [26] (RSGD), SLIM [34], SVD++ [28], WBPR [18], and WRMF [25]. Since the optimality of algorithms is not the object 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 $\mu = 1$ [7], 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 1 to 4. For true metric values, the optimal ranking function (equation 1) is given "oracle" access to test data in order to estimate $p(rel|\neg train, i)$.

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 train + 54,000 test
CM100k	1,054	1,084	103,584	= 11,594 on familiar music
				+ 91,990 on unfamiliar music

5.3 Observed Metric Disagreements

Figure 5a 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, 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 2: though we omit further graphs in the interest of space, it is easy to check that the popularity is steeper than the average rating in all three datasets, and correlates positively with the average rating for most item pairs (two reasons for the observed value of metrics to disagree).

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 the metrics would come to an agreeing trend.*

We also find validation for our formulation of the optimal rankings, which are confirmed as bounds of non-personalized algorithms for the respective metric (the optimal ranking for anti$\widehat{\mathrm{P}}$ in the x axis and the optimal for $\widehat{\mathrm{P}}$ in the y axis). 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

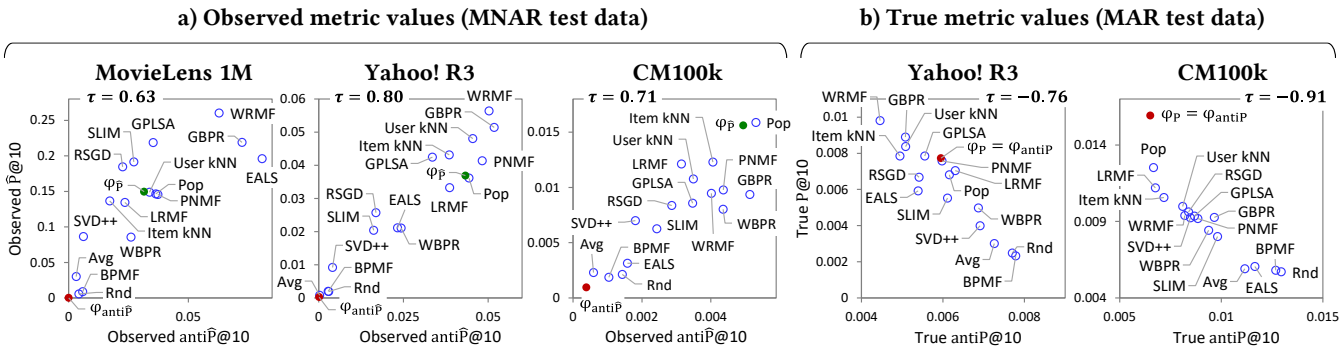

Figure 5: Regular evaluation of true and false-positive metrics on MNAR data (left) and true metric values bases on unbiased MAR test data (right). The optimal recommendations for precision and anti-precision are shown as a green and red dot, respectively. The optimals for true metric values are the same on the right graphs (as per Conclusion #1), displayed as a unique red dot. Kendall τ correlation of the system rankings is shown in each graph. The respective relevance judgment coverage ratios at cutoff 10 for observed metric values (left) are 15.97%, 5.43% and 1.10% (average across systems) for MovieLens 1M, Yahoo! R3 and CM100k, respectively. For true metric values (right), it is 1.23% and 1.83% for Yahoo! R3 and CM100k respectively.

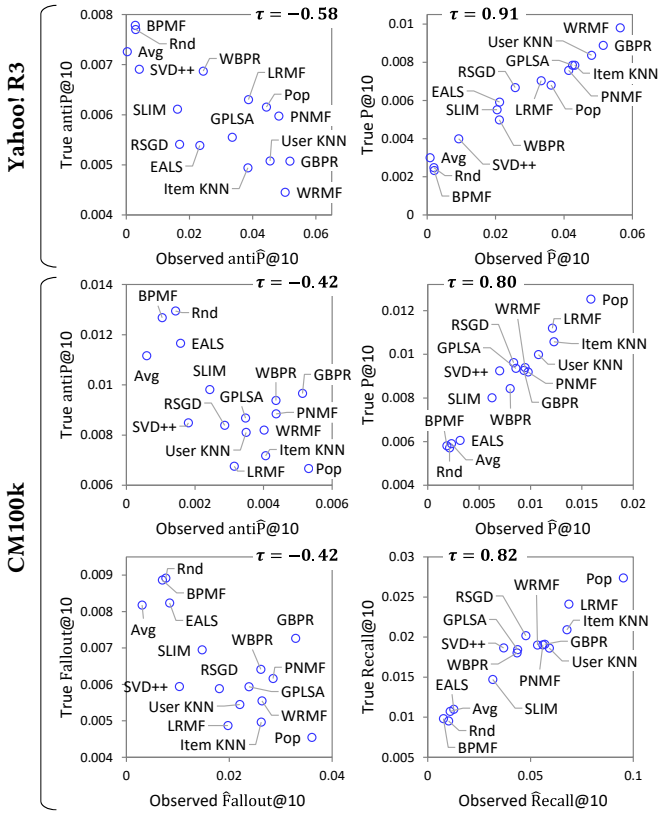

Figure 6: Agreement and disagreement between observed and true values of true and false-positive metrics on Yahoo! R3 and CM100k. Kendall τ correlation of the system rankings is shown in the graphs.

the optimal ranking for $\widehat{\mathrm{P}}$, and is therefore among the worst recommendations in anti$\widehat{\mathrm{P}}$, as predicted in Section 4.3 (Conclusion #2).

We can realize in the figure that personalized algorithms can do better than the non-personalized optimal, also as one might expect (except in CM100k where collaborative filtering fails to improve over popularity due to data sparsity, as reported in [8]). 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.

5.4 True Metric Agreement

We now take a look at the true metric values, using the two datasets that support unbiased metric estimates: Yahoo! R3 and CM100k, as explained in Section 5.1. Figure 5b shows the results. We see that the negative correlation between true precision and anti-precision is quite strong –i.e. they highly agree in ranking systems. 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 – *True 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 true precision and anti-precision (Figure 5b), and disagreement between their observed values (Figure 5a). However, the formal analysis does not establish which metric should agree or not with its true value –both situations are theoretically possible. We analyze this question empirically by plotting the true and observed values of each metric against each other in Figure 6. We see that while the MNAR measurements of precision are quite consistent with the unbiased MAR estimates, anti-precision seems to suffer from a severe distortion by the MNAR bias, to the point that the system comparisons are almost reversed. As an illustration of how our analysis generalizes to other false-positive metrics, the figure shows similar relations for fallout vs. recall in CM100k. We can see that the patterns are quite equivalent –this is just one example, and analogous observations are also obtained for any of the comparisons shown in all previous figures.

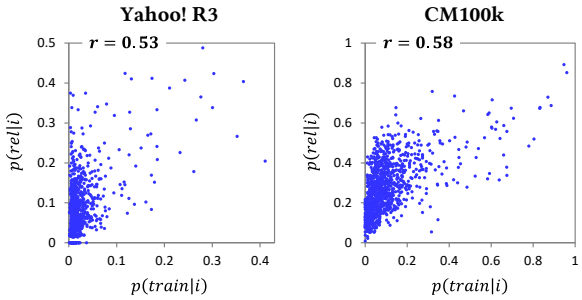

Figure 7: Rating vs. relevance probability in Yahoo! R3 and CM100k. $p(rel|i)$ is estimated for each item i as the ratio of positive ratings of i in the (MAR) test sample.

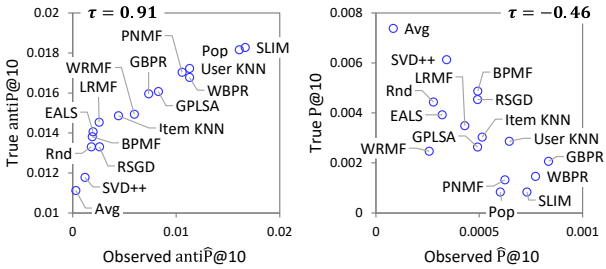

Figure 8: Agreement between observed and true values of true and false-positive metrics when a negative relation between the distribution of ratings and relevance is simulated in CM100k. Kendall τ correlation shown in the graphs.

5.5 Rating against Relevance

The consistency between observed and true precision can be related to the fact that the relevance and popularity distributions tend to agree in the two datasets, whereby $\varphi_{\hat{P}}$, φ_P and thereby φ_{antiP} tend to agree, whereas $\varphi_{anti\hat{P}}$ tends to disagree with φ_{antiP} as soon as it disagrees with $\varphi_{\hat{P}}$. Figure 7 shows the correlation between rating and relevance probabilities in Yahoo! R3 and CM100k. The high correlation hints a strong dependence between the two random variables, confirming that the rating and relevance probability agree more often than not, which may explain why precision measurements are reliable while observed anti-precision is not.

Even though this situation can be quite common, other scenarios are also possible. The relation between rating and relevance could become negative in certain domains, where people are more prone to express negative opinions than positive ones. This is often the case, for instance, in customer-support channels such as social media accounts of airlines, car brands, telecom providers, and the like, where the items that get the most feedback are typically the least satisfying. 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 (a car brand, an airline, etc.) that got the least complaints.

It is possible to simulate this type of pattern in a dataset 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 8 shows the result and indeed we see that the false-positive metric finds a clear agreement between observed and true values, just the opposite to the true-positive metric, which now becomes misleading: the best systems in observed $\hat{P}$@10 (GBPR, WBPR, SLIM) are actually the worst in true P@10.

A collateral finding is worth noting in this figure: all personalized algorithms except SVD++ (and to a negligible extent RSGD and BPMF) perform substantially worse –in terms of true metric values– than random recommendation in these 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 8 –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.

6 Conclusions

We find that false-positive metrics are heavily affected by popularity biases, just as true-positive metrics had been found to be [3,7,45]. Paradoxically, the popularity bias in false-positive metrics is negative: the recommendation of popular items is penalized, just as it was rewarded by true-positive metrics. Moreover, we observe systematic disagreements between true and false-positive metrics in recommender system evaluation, and we find an explanation for them: both types of metrics are influenced by the same popularity biases, but just in opposite ways.

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 distributions are not opposed to relevance distributions, 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 the natural next step as future research, just as debiasing evaluation techniques and algorithms are being researched for true-positive metrics [44,49].

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. TIN2016-80630-P).

REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, 17, 6 (Jun. 2005). IEEE, Piscataway, NJ, USA, 734–749.

[2] R. F. Baumeister, E. Bratslavsky, C. Finkenauer and K. D. Vohs. Bad is Stronger than Good. *Review of General Psychology*, 5, 4 (December 2001). American Psychological Association, Wasnington, D.C., USA, 323–370.

[3] A. Bellogín, P. Castells and I. Cantador. Statistical Biases in Information Retrieval Metrics for Recommender Systems. *Information Retrieval* 20, 6 (Jul. 2017). Springer, Dordrecht, Netherlands, 606–634.

[4] A. Bellogín, P. Castells and I. Cantador. Precision-Oriented Evaluation of Recommender Systems: An Algorithmic Comparison. In *Proc. of the 5th ACM Conf. on Recommender Systems* (RecSys 2011). ACM, New York, NY, USA, 333–336.

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

[6] C. Buckley and E. M. Voorhees. Retrieval Evaluation with Incomplete Information. In *Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval* (SIGIR 2004). ACM, New York, NY, USA, 25–32.

[7] R. Cañamares and P. Castells. Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems. *In Proc. of the 41st Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2018). ACM, New York, NY, USA, 415–424.

[8] R. Cañamares and P. Castells. A Probabilistic Reformulation of Memory-Based Collaborative Filtering – Implications on Popularity Biases. In *Proc. of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2017). ACM, New York, NY, USA, 215–224.

[9] P. Castells and R. Cañamares. Characterization of Fair Experiments for Recommender System Evaluation – A Formal Analysis. In *Proc. of the Workshop on Offline Evaluation for Recommender Systems* (REVEAL 2018) *at the 12th ACM Conference on Recommender Systems* (RecSys 2018).

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

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

[12] P. Cremonesi, F. Garzotto, S. Negro, A. V. Papadopoulos and R. Turrin. Looking for "Good" Recommendations: A Comparative Evaluation of Recommender Systems. In *Proc. of Human-Computer Interaction – INTERACT 2013 – 14th International Conference* (Interact 2013). Springer, New York, NY, USA, 152–168.

[13] P. Cremonesi, F. Garzotto and R. Turrin. User-Centric vs. System-Centric Evaluation of Recommender Systems. In *Proc. of Human-Computer Interaction – INTERACT 2013 – 14th IFIP TC 13 International Conference* (Interact 2013). Springer, New York, NY, USA, 334–351.

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

[15] B. Fields. *Contextualize Your Listening: The Playlist as Recommendation Engine.* Doctoral thesis, Goldsmiths, University of London, 2011.

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

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

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

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

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

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

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

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

[24] T. Hofmann. Latent Semantic Models for Collaborative Filtering. *ACM Transactions on Information Systems* 22, 1 (Jan. 2004). ACM, New York, NY, USA.

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

[26] M. Jahrer and A. Töscher. Collaborative filtering ensemble for ranking. In *Proc. of the International Conf. on KDD Cup 2011* (KDDCUP 2011). JMLR.org, 153–167.

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

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

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

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

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

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

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

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

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

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

[37] L. A. S. Pizzato, T. Rej, J. Akehurst, I. Koprinska, K. Yacef and J. Kay. 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). Springer, New York, NY, USA, 447–488.

[38] F. Provost and T. Fawcett. Robust classification for imprecise environments. Machine Learning 42, 3 (Mar. 2001). Springer, New York, NY, USA, 203–231.

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

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

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

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

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

[44] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak and T. Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proc. of the 33rd International Conference on Machine Learning* (ICML 2016). Proc. of Machine Learning Research, Sheffield, UK, 1670–1679.

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

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

[47] H. Steck. 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.

[48] K. Wang, T. Walker and Z. Zheng. PSkip: estimating relevance ranking quality from web search clickthrough data. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD 2019). ACM, New York, NY, USA, 1355–1364.

[49] L. Yang, Y. Cui, Y. Xuan, C. Wang, S. Belongie and D. Estrin. Unbiased Offline Recommender Evaluation for Missing-Not-At-Random Implicit Feedback. In *Proc. of the 12th ACM Conference on Recommender Systems* (RecSys 2018). ACM, New York, NY, USA, 279–287.

[50] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. of the 15th ACM International Conf. on Information and Knowledge Management* (CIKM 2006). ACM, New York, NY, USA, 102–111.

[51] D. Yin, S. D. Bond and H. Zhang. Are bad reviews always stronger than good? Asymmetric negativity bias in the formation of online consumer trust. In *Proc. of the 31st International Conference on Information Systems* (ICIS 2010). Association for Information Systems, pp. 1–18.

[52] Z. Yuan and E. Oja. Projective Nonnegative Matrix Factorization for Image Compression and Feature Extraction. In *Proc. of the 14th Scandinavian conference on Image Analysis* (SCIA 2005). Springer-Verlag, Berlin, Heidelberg, 333–342.

[53] C. Zhai and J. Lafferty. A risk minimization framework for information retrieval. *Information Processing and Management* 42, 1 (Jan. 2006). Pergamon Press, Inc. Elmsford, NY, USA, 31–55.