Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems

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

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

ABSTRACT

The use of IR methodology in the evaluation of recommender systems has become common practice in recent years. IR metrics have been found however to be strongly biased towards rewarding algorithms that recommend popular items –the same bias that state of the art recommendation algorithms display. Recent research has confirmed and measured such biases, and proposed methods to avoid them. The fundamental question remains open though whether popularity is really a bias we should avoid or not; whether it could be a useful and reliable signal in recommendation, or it may be unfairly rewarded by the experimental biases. We address this question at a formal level by identifying and modeling the conditions that can determine the answer, in terms of dependencies between key random variables, involving item rating, discovery and relevance. We find conditions that guarantee popularity to be effective or quite the opposite, and for the measured metric values to reflect a true effectiveness, or qualitatively deviate from it. We exemplify and confirm the theoretical findings with empirical results. We build a crowdsourced dataset devoid of the usual biases displayed by common publicly available data, in which we illustrate contradictions between the accuracy that would be measured in a common biased offline experimental setting, and the actual accuracy that can be measured with unbiased observations.

KEYWORDS

recommender systems; popularity; evaluation; bias; accuracy; non-random missing data; collaborative filtering

1 INTRODUCTION

The use of IR methodologies and metrics for the evaluation of recommender systems has spread in recent years and is becoming common practice in the area, under the understanding of recommendation as a ranking task [14]. Yet IR metrics have been found to be strongly biased towards rewarding algorithms that recommend popular items, that is, items that many people know, like, rate or interact with [4,21,35]. At the same time, state of the art recommendation algorithms have similarly been found to display a marked bias towards recommending items most people like [21]. This may naturally cast doubt on the reliability of common experiments and the outcome on which the best algorithms really are.

This problem has been of no particular concern to IR methodology, as popularity biases do not occur, or not in such a dramatic way, in traditional search and IR tasks. The popularity bias is so strong in common datasets for recommender system evaluation that even a pure and simple popularity ranking appears to achieve suboptimal but non-negligible recommendation accuracy compared to the best state of the art personalized algorithms [14]. And it is in fact not necessarily trivial to outperform, for instance, in high rating sparsity conditions. Research has therefore been recently undertaken addressing the issue, so far mainly focusing on confirming and measuring the popularity biases, and removing them [4,21,34,35]. But a basic question remains yet unanswered: is the popularity bias actually something we should get rid of at all? If recommending popular items happened to be the right thing to do, then should not both the evaluation metrics and the recommendation algorithms rightfully favor them?

The majority opinion is indeed useful information for people –it is a simple yet fair and useful default criterion we keep in sight most of the time through our human decisions, even when we do not follow it. And we in fact often do adopt it, for instance, in the absence of enough evidence to form one's own personal choice, or as guidance to reduce the cost of building a decision from scratch, or as a social learning mechanism [3]. From an application point of view, a recommendation based on the choices of many can be acceptable in many circumstances [16] –and requires minimum development skills and maintenance costs. It is actually a widespread approach that many applications display in the form of top charts, best-selling lists, average people's ratings, etc. Even in the presence of a full-fledged personalized recommender system, majority listings are still a good resort for new or cold users.

The effectiveness of majority taste makes indeed statistical sense: the items that many people like (according to the records of observed user activity) are liked by many people (in test data for evaluation) [19]. Yet from an experimental perspective, if the observations are somehow biased, and the bias is consistent across training to test data, the majority bias in recommendation might be accurately guessing where the observations have been placed by the experimenter, rather than where true user tastes are being actually most satisfied. Moreover, the majority signal might be contaminated by trends that deviate from actual user appreciation [5,29]. Recent studies show that majority formation involves a degree of chance, by which different outcomes are possible as to what choices make it to the top of popularity [31]. Crowd dynamics are moreover known to be exposed to external and internal influence and bias factors [26,27,29], such as mass media [7], marketing, opinion management [6], algorithmic bias [28], or social conformity [13].

The issue is therefore open whether or not popularity is a truly effective ingredient to achieve accurate recommendations, to

ACM Reference Format:
Rocío Cañamares and Pablo Castells. 2018. Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems. In Proceedings of ACM SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA. ACM, NY, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210014

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA.

ACM 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210014

what extent and in what cases, and whether we are measuring it properly. We address the question by considering, analyzing and comparing two views on IR metrics: biased and unbiased. The former represents what is measured in common offline experiments in the literature, where relevance information is *missing not at random* (MNAR) [23,24,25,34,35], and the latter represents the true metric value that would be obtained if the missing information became available.

We do this at both a theoretical and an empirical level. At the analytical level, we formulate a probabilistic expression of the problem. Starting by a revised probability ranking principle [30] for recommender systems, we analyze popularity-based recommendation by comparison to the optimal ranking. We find that the effectiveness or ineffectiveness of popularity depends on the interplay of three main variables: item relevance, item discovery by users, and the decision by users to interact with discovered items. We identify the key probabilistic dependencies among these factors that determine the outcome for popularity, and we characterize a set of trends defined by different independence assumptions, each resulting in a particular pattern of behavior for popularity. We back our theoretical findings with empirical observations with a dataset we build on a crowdsourcing platform, in which we remove several of the common biases of publicly available datasets.

Among other findings, we prove and illustrate qualitative contradictions between the accuracy that is measured in a common offline experimental setting, and the actual accuracy that can be estimated with unbiased observations. We identify conditions that guarantee popularity to be a safe element in recommendation, and we characterize and exemplify situations where, on the contrary, popularity can be a totally misleading direction to follow, to the point of leading to worse effectiveness than random recommendation. We furthermore find that the average rating can be more effective than the number of ratings as a trend to follow in recommendation in many cases, contrarily to what the biased metric values suggest –which represent what the literature commonly reports [14,21]. Finally, we look at the signification our findings can have in personalized collaborative filtering algorithms.

2 POPULARITY IN RECOMMENDATION

Several authors have recently paid attention to the role and effects of popularity in recommender systems. Fleder and Hosanagar [15] observed the concentration effect of the recommendation feedback loop. Cremonesi et al. [14] were among the first to point at and analyze the fair results of popularity in the top-k recommendation task. Fig. 1 illustrates a typical observation in line with those findings, on two popular public datasets [17,20] (algorithm configuration and details are the same as in [9]), where popularity performs about half as well as the personalized algorithms (kNN [14] and matrix factorization [20]). Cremonesi et al. furthermore observed that even though recommendation by average rating value achieves worse accuracy than ranking items by their number of ratings (as is the case in Fig. 1), the comparison gets reversed when the top few most popular items are removed from the data. An explanation for such different outcomes was yet to be given, and will hopefully be found in the present paper.

Steck [34,35] raised awareness on the fact that ratings are missing not at random [23,24] and subject to biases affecting both the input for algorithms and the data for evaluation. He proposed metrics and algorithmic corrections to better cope with popularity. Bellogín et al. [4] studied the strong popularity biases that surface in IR evaluation methodologies when applied to recommendation, and proposed further experimental methods to neutralize

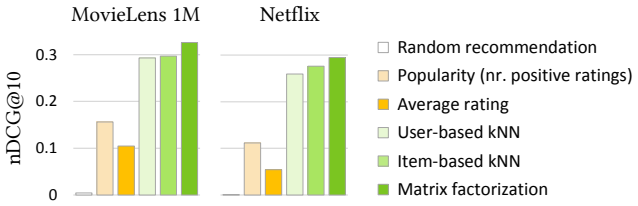

Figure 1: Typical offline experimental results for non-personalized popularity-based recommendation compared to personalized algorithms on two public datasets.

the biases. Jannach et al. [21] verified and measured the correlation with popularity observed in common state of the art algorithms, and proposed approaches to counter it. In our own prior work [9] we formalized the popularity bias as an intrinsic trend in memory-based collaborative filtering. Earlier on we studied the effect of social mouth-to-mouth on the raise of popularity distributions and the positive or negative effect on the accuracy of popularity-based recommendation [8].

However, whether popularity is actually a good or bad feature to have –and whether its measured accuracy is reliable or not– is an implicit question that has not been directly explained yet. One obvious, negative answer has been given considering that popularity is the antithesis of novelty, a key ingredient in most cases for recommendations to be useful [1,11,12]. But from a broad perspective, this answer is partial, and does not refute the usefulness of *some* degree of popularity. While lack of novelty is an obvious drawback of popularity, the effect of popularity on pure accuracy should be properly understood. Even avoiding the head of the popularity distribution, even anywhere in the long tail, some items are more popular than others, and we need to understand the difference when we settle for recommendation at one precise point or the other on the popularity curve. Furthermore, since top-performing recommendation algorithms are strongly biased towards popular items, the question concerns state of the art methods as well, and any findings on the issue would help better understand and properly compare the effectiveness of personalized algorithms.

3 THEORETICAL FORMULATION

We start our study by setting out a mathematical formalization of the relevant involved elements. We first settle some definitions, a general framing for the recommendation task, and some formal notation to be used in the rest of the paper.

3.1 Preliminaries

The recommendation task [3] considers a set of users $\mathcal{U}$, a set of items $\mathcal{I}$, and a set of observed rating values for a subset of $\mathcal{U} \times \mathcal{I}$. We need not make any specific assumption about what ratings exactly consist of: explicit scores, implicit interaction records, etc.; it is sufficient for our purposes to consider they reflect some evidence of a positive or non-positive preference by the user for the rated item. The ratings are supplied as input (training data) to recommendation algorithms, which return a ranking of items for each user. In offline experiments, a subset of the available ratings is held out as test data for evaluation, and the rest of ratings are fed as input to the algorithms under evaluation [32]. In online evaluation, all available ratings are potentially used as input, and user feedback in response to recommendations in a live system is taken as test data. In both settings, test ratings are used as relevance judgments to compute the evaluation metrics of interest.

3.1.1 Popularity-Based Recommendation. "Popular" generally means what many people like or know [21]. In the recommender systems literature, popularity is commonly defined as the number of users who have rated (who have been observed interacting with) an item, regardless of whether the interaction reflected positive or negative preference [4,14,21]. We find it more meaningful to consider an alternative refinement that only counts the interactions evidencing a positive preference (as in [9,34]). In usual datasets this makes no noticeable difference, but it can be proved that the total number of votes is never better than the number of positive votes as a signal for recommendation, whereby we shall focus on the latter definition.

Another sensible and common notion in the scope of popularity is the average rating value [14,21]. We shall use a simplified –and empirically equivalent in all our experiments– binarized definition of average rating, as the ratio of users who liked an item, which is better suited to a probabilistic analysis. The average rating tends to display lower empirical effectiveness than popularity in common datasets, though it has been show to outperform popularity when, for instance, the few top most popular items are removed from the data, as reported by Cremonesi et al. [14].

Popularity notions can be used for recommendation by just delivering the popularity ranking to all users. Formally, we shall denote by $pop(i)$ and $avg(i)$ the ranking functions of popularity and average rating, respectively, for $i \in \mathcal{I}$. Given a data split, we define $pop(i) = |i^+_{train}|$ as the number of positive training ratings the item has, and $avg(i) = |i^+_{train}|/|i_{train}|$ as the ratio of positive training ratings.

3.1.2 Observed vs. True Accuracy. Ranking-based recommendation accuracy metrics such as precision, recall, nDCG, MAP, MRR, etc., measure how well a recommendation ranks the relevant items above non-relevant ones. In the recommendation context, an item is considered relevant for a user if a positive rating by the user for the item is available in the test data. Such relevance knowledge is however generally incomplete –this is particularly true in recommender system experiments, where most of the user preferences are unknown, by definition of the recommendation task [2]. The difference between the metric value we can measure in common experiments, and the true metric value we would compute if we had full relevance knowledge, is a key distinction in our study.

3.1.3 Data Split Protocol. Our analysis shall assume an offline experimental design based on a random rating data split with a given ratio $\rho \in (0,1)$ of training data. We consider a common data partition procedure which consists of iterating over each of the available ratings in the dataset at hand, assigning it to the training or test subset with probability ρ and $1-\rho$ respectively. As in the most usual settings, we consider a recommendation task definition where the system should not recommend items to users who have already rated them (in the input training data) [2]. The data split is not necessary when we consider true metric values, which assume full (or at least unbiased) relevance knowledge: all ratings can be supplied as input to the algorithm (as if $\rho = 1$), and the metrics use extra (unrated) relevance information obtained somehow. In our theoretical analysis we will abstract ourselves from the problem of obtaining this relevance knowledge, and we will later describe how we manage to get it in our experiments.

3.1.4 User-Item Random Variables. We shall formalize key elements involved in the problem as random variables, in order to reason in terms of probabilities and expected values. First, we define the random variable $rated: \mathcal{U} \times \mathcal{I} \to \{0,1\}$ over the set of user-item pairs as $rated = 1$ if a rating by the user for the item is available in the dataset and 0 otherwise. Given a rating split, we similarly define the variables $train$ and $test$ on user-item pairs as taking value 1 if $rated = 1$ and the rating was assigned to the training or test partition respectively, and 0 otherwise. Similarly, we define $rel: \mathcal{U} \times \mathcal{I} \to \{0,1\}$ as $rel = 1$ if the user likes the item (regardless of the presence or absence of a rating), and 0 otherwise. Throughout the paper we will use the abbreviation $p(rated)$, $p(rel)$, etc., for $p(rated = 1)$, $p(rel = 1)$, and so forth. Now considering a random variable $I: \mathcal{U} \times \mathcal{I} \to \mathcal{I}$ defined as the item in a user-item pair, we can handle conditional probabilities such as $p(rated|i)$, $p(rel|i)$, and so on, for $i \in \mathcal{I}$ –where i shall stand as an abbreviation of $I = i$. The probability $p(rel|i)$, for instance, represents the ratio of users who like item i.

Using these random variables and the definitions of section 3.1.1 above, we have $pop(i) \propto |i^+_{train}|/|\mathcal{U}| = p(train, rel|i)$ and $avg(i) = p(rel|train, i)$. Since $p(train|\neg rated, i) = 0$ and, in the random split procedure described earlier in section 3.1.3, the probability for a rating to be sampled for training or test is independent from both the item and the rating value, and is equal to the split ratio ρ, we have:

$$pop(i) \propto p(rated, rel|i) \qquad avg(i) \sim p(rel|rated, i)$$

We shall use the popularity and average rating ranking functions in this form in the rest of the paper.

3.2 Expected Precision

We choose precision for an accuracy metric, as a representative yet tractable option for theoretical analysis. Moreover and for the same reason we shall take $P@1$. In the experiments of section 6 we will see that our analysis and results generalize well empirically to other accuracy metrics and common deeper cutoffs. Given a recommendation $R = \langle R_1, R_2, \ldots, R_n \rangle$, $P@1$ is a binary value that is equal to 1 if the target user likes the top ranked item R_1, and 0 if she does not. This makes it easier to reason about the expected value of this metric. As a binary function, the expectation of $P@1$ for a given recommendation R is $\mathbb{E}[P@1|R] = p(P@1 = 1|R) = p(rel|R_1)$. As we have stressed in section 3.1.2, we shall distinguish between *observed precision*, which we shall denote by the symbol $\hat{P}$, and *true precision*, which we denote as P. We have $P@1 = 1$ iff the target user likes the top-ranked item, whereas $\hat{P}@1 = 1$ iff the user likes the top item *and* a rating by the user for the item is present in the test set. Therefore, $\mathbb{E}\left[\hat{P}@1\middle|R\right] = p(rel, test|R_1)$ for observed precision.

Now we need to be more precise with the computation of the metrics: in fact $P@1 = 1$ iff the first ranked *recommendable* item in R is relevant (and analogously for $\hat{P}$). Let this item be R_k, ranked at the k-th position of R. As mentioned in section 3.1.3, recommendable means that R_k does not have a training rating by the target user, and being the first means that all the items $R_1, R_2, \ldots, R_{k-1}$ above R_k are not recommendable because they do have a training rating. If we marginalize $p(P@1 = 1|R)$ and $p\left(\hat{P}@1 = 1\middle|R\right)$ by the possibility that the k-th item is the first recommendable, and we make the mild assumption that whether two items are rated or not by some user are mutually independent events (whereby so is the $train$ event, since ratings are independently sampled in the random split as described in section 3.1.3), we have:

Table 1: Summary of cases. "Optimal" is meant in the scope of non-personalized recommendations. We indicate the conclusion numbering corresponding to each case, and the figure(s) where it is shown or tested.

Rating independence assumptions/cases	Corresponding assumptions on rating decision + discovery		Subcases	# conclusion	Fig.	Popularity $\mathbb{E}[\hat{P}@1]$	Popularity $\mathbb{E}[P@1]$	Average rating $\mathbb{E}[\hat{P}@1]$	Average rating $\mathbb{E}[P@1]$
Item-independent $rated \perp i \mid rel$	$rated \perp i \mid seen, rel$	$seen \perp i \mid rel$	–	1	5d	Optimal			
Relevance-independent $rated \perp rel \mid i$	$rated \perp rel \mid seen, i$	$seen \perp rel \mid i$	a. $p(rel\|i)$ steeper enough than $p(rated\|i)$	2+3	5a				
			b. $p(rated\|i)$ steeper enough than $p(rel\|i)$		–		Random		
			c. Neither dominates		5c		Better than random	Almost random	Better than pop
No assumption	–		–	4	3, 5b				

$$\mathbb{E}[P@1|R] \sim \sum_{k=1}^{n} p(rel, \neg train|R_k) \prod_{j=1}^{k-1} p(train|R_j) \quad (1)$$

$$\mathbb{E}[\hat{P}@1|R] \sim \sum_{k=1}^{n} p(rel, test|R_k) \prod_{j=1}^{k-1} p(train|R_j) \quad (2)$$

where in equation 2 we can remove the condition $\neg train$ in $p(rel, test, \neg train|R_k)$ as it is redundant: if a rating is present in the test set it cannot be present in the training set.

3.3 Optimal Recommendation

We can now set forth a first result on the optimal rankings for expected observed and true precisions.

Lemma 1 – Assuming pairwise item rating independence, the optimal recommendation that maximizes the (true) $P@1$ expectation under a random rating split ranks items $i \in \mathcal{I}$ by non-increasing value of:

$$f(i) = p(rel|\neg train, i) = p(rel|i) \frac{1 - \rho\, p(rated|rel, i)}{1 - \rho\, p(rated|i)} \quad (3)$$

Under the same assumptions, the optimal recommendation that maximizes the expected (observed) $\hat{P}@1$ ranks items by non-increasing value of:

$$\hat{f}(i) = \frac{p(rel, test|i)}{p(\neg train|i)} \propto p(rel|i) \frac{p(rated|rel, i)}{1 - \rho\, p(rated|i)} \quad (4)$$

Proof In order to show that the above rankings maximize the corresponding precision, it suffices to show that a consecutive swap against f or $\hat{f}$ in a ranking produces a smaller value for $\mathbb{E}[P@1|R]$ or $\mathbb{E}[\hat{P}@1|R]$ respectively [10]. Given that any ranking can be generated by a sequence of pairwise counter-order swaps on any other ranking (as per e.g. the proof of correction of bubble sort), we would have proven our point. For true precision, let R be some ranking so that $f(R_k) \geq f(R_{k+1})$ for some k, and let us consider a ranking R' consisting of swapping R_k and R_{k+1} in R. Using equation 2 it is easy to see that, by trivial algebraic cancellation and rearrangement of terms, we have:

$$\mathbb{E}[P@1|R] \geq \mathbb{E}[P@1|R']$$

$$\Leftrightarrow \frac{p(rel, \neg train|R_k)}{1 - p(train|R_k)} \geq \frac{p(rel, \neg train|R_{k+1})}{1 - p(train|R_{k+1})} \Leftrightarrow f(R_k) \geq f(R_{k+1})$$

Which is true by description of R. That is, swapping R_k and R_{k+1} decreases $\mathbb{E}[P@1|R]$. And an analogous reasoning proves the corresponding statement for observed precision.

The right-side form of f and $\hat{f}$ in equations 3 and 4 is trivially obtained by applying $p(train|i) = \rho\, p(rated|i)$ and $p(test|i) = (1 - \rho)\, p(rated|i)$ as explained in section 3.1.4. □

4 RELEVANCE BIAS IN RATING DISTRIBUTION

We now analyze how the relation between relevance and rating can determine the effectiveness of popularity. We do so by examining how the popularity and average rating rankings relate to the optimal ranking, and random recommendation.

We start by considering two extreme cases in the relation between rating and relevance: a) the probability that a user rates an item depends only on relevance; and b) the probability that a user rates an item is independent from relevance. These two conditions can be expressed as conditional independence assumptions between rating, relevance and items: a) $rated \perp i \mid rel$ and $rated \perp i \mid \neg rel$, and b) $rated \perp rel \mid i$, respectively. We analyze the consequences of each of these conditions in the next subsections. The analytic findings we will reach therein are summarized in Table 1.

4.1 Conditional Item Independence

The independence assumption means $p(rated|rel, i) \sim p(rated|rel)$ and $p(rated|\neg rel, i) \sim p(rated|\neg rel)$. Applying this to equation 3, we get that the optimal ranking for true precision is given by:

$$f(i) \sim \frac{(1 - \rho\, a)\, p(rel|i)}{1 - \rho\, b + \rho\, (b - a)\, p(rel|i)} \propto p(rel|i)$$

with constants $a = p(rated|rel)$, $b = p(rated|\neg rel)$. The rank equivalence holds because $x/(c_1 + c_2\, x)$ is a monotonically increasing function of x as long as $c_1 > 0$, whatever the value of c_2.

For observed precision, we similarly get:

$$\hat{f}(i) \propto \frac{a\, p(rel|i)}{1 - \rho\, b + \rho\, (b - a)\, p(rel|i)} \propto p(rel|i)$$

where a and b are defined as before. We thus find, in particular, that if the rating probability depends only on relevance, then observed and true precision are consistent as to what the optimal recommendation is.

On the other hand, with the independence assumption at hand, the ranking functions for popularity and average rating become:

$$pop(i) \sim a\, p(rel|i) \propto f(i) \propto \hat{f}(i)$$

$$avg(i) \sim \frac{a\, p(rel|i)}{b + (a - b)\, p(rel|i)} \propto p(rel|i) \propto f(i) \propto \hat{f}(i)$$

We thus come to:

Conclusion 1 – If the probability of rating depends just on whether the item is liked, then 1) *the expected observed and true precision agree*, and 2) *both popularity and average rating produce the optimal non-personalized recommendation.*

Note that the scope of this finding, and all the ones that shall follow, is non-personalized: popularity, for instance, ranks items

by $p(rated|rel,i)$, but not specifically by $p(rated|rel,i,u)$ for each user –and analogously for the average rating. Hence optimality is in those precise terms: without having the ranking depend on the user, thus applying lemma 1 in a non-personalized version. Note also that by *optimal* we shall always be meaning in expectation (of precision) with respect to the random data split and the detailed placement of ratings in the user-item matrix.

4.2 Conditional Relevance Independence

The relevance independence assumption means $p(rated|rel,i) \sim p(rated|i)$. Under this assumption, the optimal rankings obtained in equations 3 and 4 become:

$$f(i) \sim p(rel|i)\frac{1-\rho\, p(rated|i)}{1-\rho\, p(rated|i)} = p(rel|i)$$

$$\hat{f}(i) \sim p(rel|i)\frac{p(rated|i)}{1-\rho\, p(rated|i)} \propto p(rel|i)\, p(rated|i)$$

where the final rank equivalence for $\hat{f}$ holds because $x/(1-\rho\, x)$ is monotonically increasing in x and almost equal to the identity function for small values of x. Observed and true precision are thus not necessarily consistent when the rating probability depends not just on relevance.

The popularity rankings, on their side, become:

$$pop(i) \propto p(rel,rated|i) = p(rel|i)\, p(rated|i) \propto \hat{f}(i)$$

$$avg(i) \propto p(rel|rated,i) = p(rel|i) \propto f(i)$$

whereby popularity and average rating would match the optimal ranking for observed and true precision, respectively. We therefore find:

Conclusion 2 – If the probability of rating does not depend on relevance, then *the average rating is optimal in true precision, whereas popularity is optimal in observed precision.*

We can draw further conclusions depending on which distribution, relevance or rating, is steeper. If the relevance distribution is steeper enough than the rating distribution, then $p(rel|i)$ would dominate over $p(rated|i)$ when multiplying them, and we would have approximately $\hat{f}(i) \propto p(rel|i)\, p(rated|i) \propto p(rel|i) \propto f(i)$. If on the contrary the rating distribution is steeper enough than relevance, we would have $\hat{f}(i) \propto p(rel|i)\, p(rated|i) \propto p(rated|i)$, whereby $pop(i)$ would be totally unrelated to $f(i)$ –hence equivalent to random recommendation–, and same for $avg(i)$ with respect to $\hat{f}(i)$. We thus conclude:

Conclusion 3 – If the probability of rating does not depend on relevance, we have: a) if relevance is steeper enough than rating, *true and observed precision come close to agree, whereby popularity and the average rating approximate the optimal in both metrics*; b) if rating *dominates* over relevance, *popularity and the average rating tend to become equivalent to random recommendation in true and observed precision, respectively*; and c) we can *further* conclude that in the average case where neither rating nor relevance are clearly steeper than the other, *popularity and the average rating can be expected to be not optimal but still better than random recommendation in true and observed precision, respectively.*

We hence find a contradiction between observed and true precision when rating and relevance are independent, unless the relevance distribution is very much steeper than the rating distribution. If the latter is very skewed, the contradiction can become extreme: observed and true precision report opposite outcomes.

4.3 General Case

The simplifying assumptions considered above are not meant to reflect situations that would strictly occur in real scenarios: they just serve the purpose of identifying and understanding fundamental factors that make part, as mixed trends, of real situations. Moreover, we will show later that it is actually possible to enforce them in a controlled experiment.

However, in the general case, with no particular independence assumptions, any outcome is actually possible. It is easy to build simple toy examples where popularity and the average rating are better or worse than each other and/or random recommendation, either in terms of true or observed precision. We may nonetheless expect, based on the findings of the previous subsections, that to the extent that rating dependence on either relevance or items (while coexisting) dominate one over the other, the results will be closer to the corresponding trends characterized so far.

We may also realize that since $\hat{f}(i) = pop(i)/(1-\rho\, p(rated|i))$, popularity will tend to get a favorable assessment in observed precision –unless the items with the highest number of relevant ratings have a low total number of ratings, which is quite unlikely– whereas the average rating does not have such a direct relation to observed precision. We may hence expect to see the average rating lagging behind popularity in evaluations such as the one shown in Fig. 1, which need not necessarily reflect the actual situation if true precision could be measured.

We seek further insights in the next section, by analyzing where the probabilistic dependences between rating, relevance, and items may arise from.

5 THE INTERPLAY OF RATING, DISCOVERY AND RELEVANCE

In order to better understand how rating may come to depend on relevance and individual items, we can consider the basic question: how does a rating come to existence? For a rating to be generated, the user must first of all discover the item at hand somehow. Then, she needs to examine, consume, buy, use (whatever applies in the application domain) the item, in order to form an opinion about it; and finally, she needs to decide to enter a rating. For simplicity, we shall collapse consumption and rating as a single event (as is in fact the case for systems working with implicit user preference feedback), which is sufficient for our analysis.

The characterization of popularity distributions can be thus decomposed into (and explained by) the discovery and rating decision distributions that give rise to the rating distribution. To reflect this view, let $seen: \mathcal{U} \times \mathcal{I} \to \{0,1\}$ be a binary random variable that takes value 1 if the user knows the item exists, and 0 otherwise. We can marginalize the probability that an item has been rated by the event that it has been discovered or not. Given that$(rated|\neg seen,i) = 0$ (a user cannot rate items she has not yet discovered), and further marginalizing over relevance, we have:

$$\begin{aligned} p(rated|i) &= p(rated|seen,i)p(seen|i) \\ &= p(rated|seen,rel,i)p(seen|rel,i)p(rel|i) \\ &\quad + p(rated|seen,\neg rel,i)p(seen|\neg rel,i)\big(1-p(rel|i)\big) \end{aligned}$$

If we rewrite all the equations in section 4 using this decomposition, we realize that the ideal rankings –and hence the precision of popularity– depend on, and can be fully expressed in terms of, the following factors appearing above:

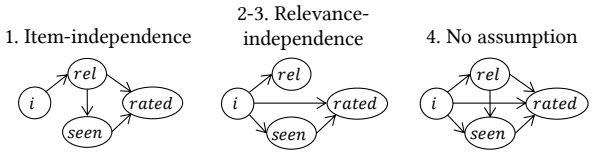

Figure 2: Graphical models summarizing the conditional independence assumptions in conclusions 1 through 4.

- The bias in the *user decision* towards rating discovered items depending on whether they like them or not, reflected in $p(rated|seen, rel, i)$ and $p(rated|seen, \neg rel, i)$.
- The potential bias in *item discovery*, towards finding liked or non-liked items more often, represented in $p(seen|rel, i)$, and $p(seen|\neg rel, i)$.
- The *relevance* distribution over items $p(rel|i)$, reflecting that some items may be liked by more people than others.

The conditional dependence between rating, relevance, and items can be therefore reformulated as conditional dependencies of user behavior and item discovery on relevance and items, as we summarize in Fig. 2: if (and only if) rating decision and item discovery only depend on relevance, then rating only depends on relevance: $p(rated|rel, i) = p(rated|seen, rel, i)\ p(seen|rel, i) \sim p(rated|seen, rel)\ p(seen|rel) = p(rated|rel)$, and conclusion 1 holds. Analogously, if the former are conditionally independent from relevance, so is the later, and conclusions 2 and 3 hold. And when discovery and user behavior are not conditionally independent from the same thing (relevance or the item), rating depends on both relevance and the item. We briefly reflect on the meaning of dependencies at this level the next subsections.

5.1 User Rating Behavior

Rating decision, in face of a discovered item, is essentially determined by human behavior [18]. The user bias towards rating relevant items more often than non-relevant ones has been mentioned as a likely possibility (a MNAR case) in prior work [25], and some studies have confirmed this trend in particular environments [24]. Considering that relevance is the main intrinsic property of consumed items that may bias the user's rating decisions (conditional independence from the item given its relevance) may be a reasonable simplification for many purposes. It is moreover possible to make the decision independent from both the item and user tastes e.g. in a controlled experiment, where users are prompted for explicit feedback on items they did not freely choose, as we shall describe in section 6.

5.2 Item Discovery

Item discovery results from a more complex combination of actions by the user (e.g. searching and browsing) and external agents (advertisement, mouth-to-mouth [8], recommender systems [14,21], random chance, etc.). Discovery thus results from the interplay of a variety of processes, some of which are typically not the same for all items, and thus certainly do depend on the specific item. For instance, the item producer and/or marketer is one active, item-specific agent in the dissemination of the item towards potential consumers. At the same time, discovery may depend on relevance, as is generally the case when items are found by users through a search engine, a recommender system, or a suggestion by a friend. If such discovery means are more accurate than random, discovery will be biased towards items that users will like.

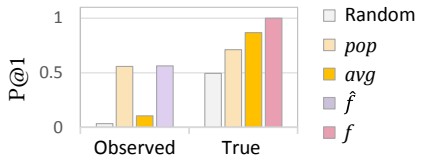

Figure 3: Expected precision with no assumption. The expectation is estimated by Monte Carlo random sampling over the set of all possible valid values of $p(rated|seen, rel)$, $p(rated|seen, \neg rel)$, $p(seen|rel, i)$, $p(seen|\neg reli)$, $p(rel|i)$, for $i \in \mathcal{I}$ with $|\mathcal{I}| = 3,700$, $\rho = 1$ for true precision and $\rho = 0.8$ for observed precision, taking 10^4 random samples.

Discovery independence from the item given its relevance represents a fair situation in which all items have equal opportunity to be discovered, except for favoring positive matches in the interest of users. On the other end, conditional discovery independence from relevance represents a scenario where each item has its own, somewhat arbitrary degree of promotion, which fails to properly consider what users may or may not like –users are at the mercy of marketing or fashion [5,7], and dispense with search or recommendation aids, or these are ineffective.

5.3 Multiple Dependence

We noted at the end of section 4 that in the general case with no assumption any outcome can occur for the effectiveness of popularity, agreements or disagreements between observed and true precision. We may however seek further insights beyond that, by estimating the expected situation considering all the possible values that the five fundamental distributions may take: $p(rated|seen, rel, i)$, $p(rated|seen, \neg rel, i)$, $p(seen|rel, i)$, $p(seen|\neg rel, i)$, and $p(rel|i)$. This means assessing the expected precisions for popularity, average rating, and the optimal rankings over the space of such values. Formally, we should compute:

$$\mathbb{E}[P@1|\theta] = \int_{\Omega^n} \mathbb{E}[P@1|\theta, \omega] d\omega$$

and similarly for $\mathbb{E}\left[\hat{P}@1\middle|\theta\right]$, with θ denoting the different rankers, and Ω^n representing the set of all possible valid values of the five conditional probabilities for all $i \in \mathcal{I}$ –namely $\Omega = [0,1]^5$ and $n = |\mathcal{I}|$. This expectation can be estimated in a Monte Carlo approach, by sampling points $\omega \in \Omega^n$ uniformly at random, computing the ranking R that each recommender θ returns given ω (which is straightforward since pop, avg, f, and $\hat{f}$ are direct functions of the five probabilities in Ω), and computing $\mathbb{E}[P@1|R]$ and $\mathbb{E}\left[\hat{P}@1\middle|R\right]$, which again are functions of the same probabilities.

Fig. 3 shows the result for $|\mathcal{I}| = 3{,}700$ (using the MovieLens 1M size [17] as an example) –it is easy to check that the results do not qualitatively depend on $|\mathcal{I}|$. We see that we may expect a substantial and qualitative contradiction between the observed and true precision:

Conclusion 4 – In the absence of any independence assumption, *while according to observed precision* (as we would measure in a standard experiment) *popularity would appear to be optimal and the average rating would seem barely better than random, popularity can be expected to be in truth just better than random, and the average rating to be better than popularity.*

Note the difference in extent of this finding compared to conclusions 1-3 in section 4. Whereas the results here are *in expectation*

over all possible values of the conditional discovery, conditional rating decision and relevance probabilities (which means that the results may differ for specific probability values), the conclusions in section 4 hold strictly *for any value* of such probabilities, as long as the stated assumptions hold. Moreover, in the Monte Carlo estimation we have assumed a uniform "meta-distribution" of the probability values, while in practice some configurations can be expected to be more likely than others. Be that as it may, this neutral expectation provides an additional reference point in our analysis.

6 EMPIRICAL OBSERVATIONS

We now run some experiments to see if the analytical results match empirical observations and to what extent. We build for this purpose a new dataset simultaneously supporting measurements under MAR (data *missing at random* [23,24]) and MNAR conditions.

6.1 A Crowdsourced MAR Dataset

Common publicly available datasets [12,17] usually provide user ratings for items in some domain, which have been collected through some natural process where users freely interact with items, under the influence of a myriad of discovery sources, some internal to the system where ratings are collected, and others exogenous. When using such collections there is generally no way to know precisely the discovery and behavior distributions from which the rating distribution resulted. We can only compute observed metric values, aware that they are measured upon MNAR data [25,34], and we have no means to contrast this to true unbiased values. Fig. 1 showed earlier an example of this usual situation on two well-known datasets. We see that popularity performs fairly well, far above random recommendation, and the average rating stands behind popularity by a clear difference. In our experiments we use a smoothed rating average, to avoid an extremely poor accuracy due to the high variance in the items with lowest number of ratings. We use Dirichlet smoothing with $\mu = \text{avg}_{i \in \mathcal{I}} |i_{train}|$ –the average sample size– as a fair default setting [36]. We may suspect the presence of biases and some effect on the observed results, but we cannot explain or verify much further with the available data.

Seeking further insights, we build a new dataset where data is (essentially) MAR, by a crowdsourcing approach as follows.[1] We sample around 1,000 music tracks from the Deezer database[2] uniformly at random using the public API. Then, we work with around 1,000 users (after discarding unreliable input through several checks and filters) in the CrowdFlower[3] service. We randomly assign tracks to users in such a way that each track is assigned to about 100 users, and each user gets about 100 tracks assigned, adding to a total of about 100,000 assignments. For each assignment, we ask the user to play the music and tell whether or not a) she likes it, and b) she had heard it before this survey.

The novelty of the resulting dataset with respect to others is that we completely remove the discovery bias by sampling and assigning items to users uniformly at random. Moreover, we completely remove the rating decision bias by requiring users to rate everything they are presented with, that is $p(rated|seen) = 1$ in the resulting dataset. Furthermore, the declared user music knowledge information enables recreating MNAR data conditions, as we shall see. Fig. 4 shows the rating and relevant rating distributions $p(rated|i)$ and $p(rated, rel|i)$ in our crowdsourced dataset and in MovieLens. For the former we show, additionally, the discovery distribution $p(seen|rated, i)$. We can see that the rating distribution in our dataset corresponds to a uniform probability (with a natural binomial sampling variance), whereas discovery and relevance are heavy-headed.

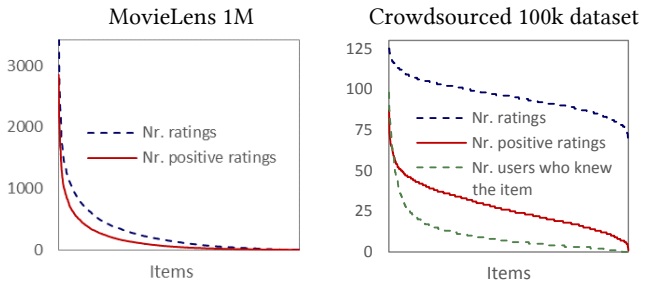

Figure 4: Data distribution in MovieLens 1M (left) and our crowdsourced dataset (right). Note that each curve has axis x (items) sorted by decreasing order of the corresponding distribution so as to better show its shape –the x values of the curves therefore do not match with each other.

6.2 Evaluation under Different Scenarios

The unbiased data thus collected enables reproducing different scenarios for experimentation, resulting in different outcomes for popularity, which are shown in Fig. 5. The dataset as is enables two different scenarios, and two additional ones are recreated by resampling the ratings in different ways. We describe each scenario and the corresponding experimental results in turn in the following paragraphs labeled a-d, matching the labels in Fig. 5. In all scenarios, we split the ratings into training and test with $\rho = 0.8$ (5-fold validation) and we interpret the absence of rating as non-relevance (alike to negative ratings). We average the results over 10 executions to reduce the variance in all experiments with the crowdsourced data. Along with the evaluated methods we show the metric values for the optimal non-personalized rankings defined by f and $\hat{f}$ in equation 5, as skyline oracles that are given access to relevance information that is hidden from the other recommenders. We also show results on MovieLens for quick reference.

a) Rating fully independent from relevance and items. Since rating decision is, by our data collection design, independent from items and relevance, the dataset as is fits in the case described in section 4.2, where conclusions 2 and 3 apply. Because the rating distribution is uniform over items, the relevance distribution is much steeper in comparison, and we have specifically conclusion 3a. The results can be seen in Fig. 5a: popularity and the average rating perform significantly better than random, and not far from the optimal non-personalized ranking, which confirms the analytical expectation. We can only measure observed precision here, since we do not have further relevance knowledge beyond the collected ratings.

We can also see the advantage of popularity over random is smaller than in MovieLens: because of the flat rating distribution, the relevant rating distribution is much less steep in our dataset

[1] The dataset is publicly available at http://ir.ii.uam.es/cm100k
[2] https://www.deezer.com
[3] https://www.crowdflower.com

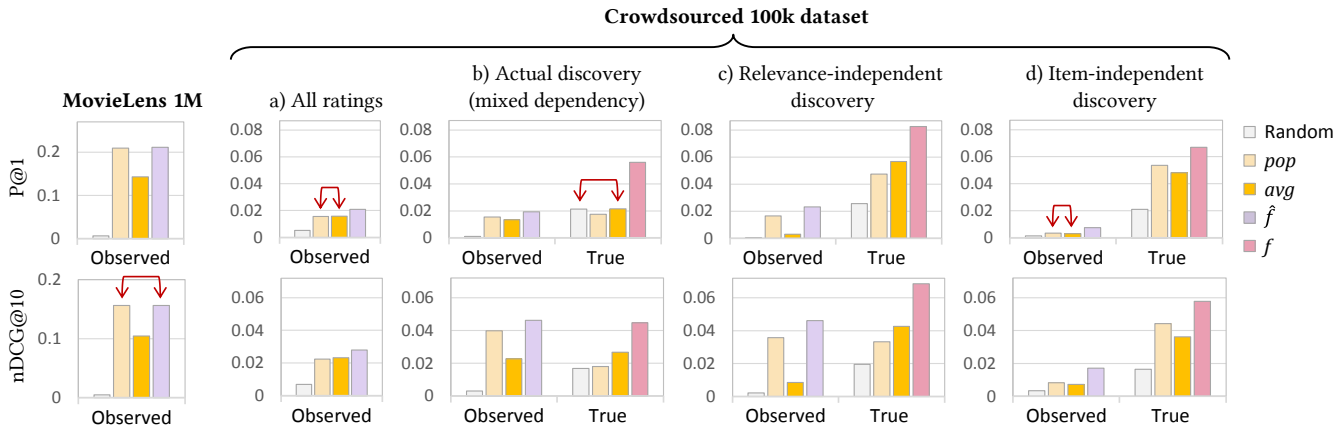

Figure 5: Empirical confirmation of the analytical results. Column a) reflects and confirms conclusions 2 and 3a; c) corresponds to conclusions 2 and 3c, d) matches conclusion 1, and b) exemplifies the general scenario addressed in conclusion 4. We confirm several inconsistencies between observed and true accuracy, and find below-random recommendation performance for popularity in scenario d. We also show the accuracy of the (oracle) optimal non-personalized rankings. Non-statistically significant differences (2-tailed Student's t-test at $p < 0.01$) are indicated in the graphs with a red double arrow.

(see Fig. 4), and popularity therefore gets a considerably lesser advantage. Moreover, we find that the accuracy of popularity and the average rating become statistically equivalent. This provides an explanation for the results reported by Cremonesi et al. [14], where removing the head of the rating distribution reversed the comparison between popularity and the average rating, as the rating distribution was made flatter –moreover, popularity was defined in [14] as the total number of ratings (rather than just positive ones), which naturally converges to random recommendation as the rating distribution tends towards uniformity.

b) Mixed discovery dependencies. Based on the user feedback on what music they had heard before, we can reproduce a natural MNAR discovery bias by providing the recommender systems as input only the ratings for music that users declared to know. We likewise compute observed precision by counting relevant items only when they were known to the user. But at the same time, we can estimate true precision by using the full available MAR relevance information. This knowledge only covers about 10% of items for each user but, as a uniform sample, it enables an unbiased estimate of true precision. We still have $p(rated|seen) = 1$ (where "$seen$" now means the user declared to know the item before the experiment) and rating decision of a discovered item is independent from items and relevance, therefore the resulting setting corresponds to the mixed situation described in sections 4.3 and 5.3. We have no particular reason to assume discovery (by whatever processes lead users to discover music before our experiment) could be independent from neither items, nor relevance. In fact, the results shown in Fig. 5b suggest, by their difference to cases c and d (to be described next), that both dependences must be present in the data.

While observed precision depicts a comparable outcome to the results on MovieLens, true precision tells quite a different story, revealing a quite inadequate performance, even slightly (but statistically significantly in $P@1$) below random. On the other hand, the average rating performs somewhat poorly but better than popularity in true accuracy, most particularly in terms of nDCG@10. Overall, the results seem not far from conclusion 4.

c) Relevance-independent discovery. We can reproduce separate discovery biases by simple resampling procedures. By randomly shuffling the discovery distribution over items, i.e. reassigning each $p(seen|i)$ to a random item j, we can decouple discovery from relevance, that is we remove any dependence there might be between $seen$ and rel given an item. Discovery therefore only depends (by random arbitrary assignment) on the item. Discovery (to which the rating distribution is proportional in this case) seems to have a slightly steeper distribution than relevance in Fig. 4, but this does not seem to be enough and the setting tends to fit in conclusion 3c. The contradiction between observed and true accuracy is most striking in this scenario: popularity stands out in observed precision, where the average rating is just above random, while almost the opposite is the case in true precision – though popularity is still better than random, as expected.

d) Item-independent discovery. We can reproduce a relevance-only dependent discovery by randomly reassigning (fictional) discovery to user-item pairs with probability $p(seen|rel)$ if the user likes the item, and with probability $p(seen|\neg rel)$ otherwise, using the values of $p(seen|rel)$ and $p(seen|\neg rel)$ estimated from our initial data. By doing so, the resulting discovery distribution will just depend on the relevance of items, and we remove the potential direct dependence on the item. As in cases b and c above, we only use the ratings (as system input and for observed precision computation) on items that users have "$seen$". Once again, we see the results in Fig. 5d match the analytical prediction (conclusion 1). Popularity seems to take better advantage of the relevance dependence than the average rating. The advantage is nonetheless quantitatively small.

The empirical results are thus largely coherent with the analytically characterized behaviors. The non-personalized rankings approach but do not fully reach the oracle theoretical optima when the theory suggests it should. This can be attributed to the sampling variance involved in rating assignment and the data split. The comparisons between popularity, average rating and random rankings are however similarly affected by the variance, and do seem to match more closely what theory expects. Though we focused on

P@1 as a tractable metric, it seems to consistently generalize empirically to other metrics (we just show nDCG@10 for the sake of space, but other metrics follow similar patterns). In fact, deeper cutoffs align slightly more tightly with the analytical findings, possibly because they are more robust to the potential peculiarities of a single top 1 item. We have also found some sensitiveness to the split ratio at specific points, which we envisage analyzing along with the sampling protocol in further depth in future work.

The crowdsourced dataset provides information that allows supplying MAR (scenario a) or MNAR (scenarios b, c, d) input data for the algorithms; and MAR or MNAR relevance information for computing the observed and true value, respectively, of evaluation metrics in any scenario. In other words, *we evaluate with ratings that are usually missing*. Related to this, Marlin et al. [24,25] also used semi-randomly polled ratings with a different focus: learning and correcting for the MNAR biases in recommendation algorithms. They did not explain how the biases result in metric disagreements, but they empirically observe them. Nor do they analyze popularity or consider the role of item discovery, but the overall idea of randomly sampling ratings for unbiased metric estimates is a direct precedent of our experimental approach.

6.3 Popularity-Biased Personalized Algorithms

As we discussed in section 2, popularity is known to be a major trend in most state of the art collaborative filtering methods [9,14,21]. We may therefore wonder whether similar patterns may manifest in some way in the algorithms that are biased to this feature. Focusing on the most general, mixed dependence scenario (case c), we test, as just a sample and representative method, two variants of the user-based k nearest neighbors (kNN) algorithm: normalized and not normalized, as defined in e.g. [14]. The latter is known to be popularity-biased, whereas the former is biased towards the average rating [9]. We tune k on MovieLens 1M for $P@1$ by grid search by multiples of 10, then 100 then 1,000, using a validation subset of the training data. On the crowdsourced dataset we just take all users as neighbors to avoid the burden of tuning for true precision –we nevertheless checked that doing so only makes the resulting differences larger and clearer.

Fig. 6 shows the result: the popularity-biased algorithm appears to be clearly better than the one biased to the average rating in terms of observed metric values (in line with prior reported results [9,14]), while the true values reveal rather the opposite is the case. One may wonder if we would see a similar result in MovieLens had we had an unbiased glimpse of the unseen relevance for this dataset.

7 DISCUSSION

Our study confirms the general popularity effectiveness trend [14,21], formally proving and explaining where this comes from. We also find that the apparent accuracy can be misleading (i.e. does not match the true accuracy) in many cases: this mainly happens, in essence, when item discovery is broadly detached from user taste.

We show that common experiments (i.e. observed accuracy metric values) may be rather unfair to the average rating, and its personalized derivatives. Contrarily to what has been observed in the literature so far [14,21], the average rating may be in fact a better, safer, more robust signal than the number of positive ratings in terms of true achieved accuracy in most general situations –a quick glimpse at Table 1 or Fig. 3 and 5 evidences that while the observed accuracy of popularity would appear better than the

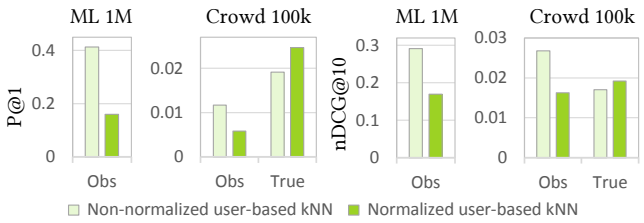

Figure 6: User-based kNN in MovieLens 1M and the crowdsourced 100k dataset, mixed dependency scenario. All pairwise differences are statistically significant (2-tailed Student's t test at $p < 0.001$).

average rating in many cases, the latter actually outperforms –or at worst is not far from– the former in true accuracy. In exchange, the rating average needs smoothing and hence parameter tuning –we see that a simple default configuration works quite well nonetheless. Furthermore, if unbiased item judgments are available for training, the average rating can definitely and systematically outperform popularity (we omit such experiments here for the sake of space).

We further find out that among the factors producing MNAR conditions in rating data [24,25,34], taste biases in users' decision to rate items may not have exactly the role that has been suggested in the literature. In particular, it does not matter whether liked items are rated more often or less, when it comes to the effectiveness of popularity or the average rating, and its measurement. The situation with regards to our analytical findings is the same regardless, for all purposes, since we did not need to assume $p(rated|seen, rel) > p(rated|seen, \neg rel)$ or the opposite anywhere in our analysis. What matters is just whether rating depends on relevance or not, and whether the dependence is full or partial; but not in what direction.

The ideal conditions for popularity and the average rating to be truly accurate are cases when discovery mainly depends on relevance, or barely depends on it. The average rating seems more robust than popularity to relevance-independence –so it would be preferable over popularity in highly biased (e.g. marketing-driven) scenarios. Popularity might take a slightly better advantage of relevance dependence, though further experiments would be needed to check this point, as the difference is small in our observations, and the theoretical conclusions would suggest a tie. Problems can arise in the mixed case where, in our particular experiment, we strikingly discover not only a contradiction between observed and true accuracy, but a below random performance for popularity.

Finally, we observe that the findings for popularity, worthy of research in themselves, can have further signification on top-performing recommendation algorithms as far as they are popularity-biased. For instance, Cremonesi et al. [14] had found that a normalized kNN algorithm outperformed the non-normalized version when the top head items were removed. We now find an explanation in our analysis, by the trends described in scenario a (section 6.2) for a flattened rating distribution, along with the respective bias of the kNN variants to popularity and the average rating.

Our findings may call for a second look at the algorithmic state of the art in light of new approximations to true accuracy. In this perspective, preference data on randomly sampled users and items may be costly to obtain, but can be a highly clarifying complement of common experiments with biased user interaction data.

8 CONCLUSIONS

We have developed a formal analysis of the effectiveness of popularity-based recommendation, upon the identification and formalization of key factors on a probabilistic basis. Our findings provide some principled explanation of the general trend observed in experiments reported in the recent literature in the field [14,21]. At the same time, insights from a deeper analysis suggest we may want to scratch beneath the surface of common experiments as we may discover unperceived and potentially different outcomes. To the best of our knowledge, these represent the first specific results to be reached on the question whether popularity is an effective or misleading signal in recommendation –and the first to suggest the average rating might be preferable to the number of favorable preferences as a non-personalized signal.

The presented findings can be useful in different ways. In a working application, we may wish to know if popularity is truly effective or not, in order to leave it or not as a trend in our algorithms. In an evaluation experiment, we may want to neutralize the interference of the popularity bias by an experimental design where popularity amounts to random recommendation [4,21,35]. Or we might want to rethink recommendation algorithms in light of what formal analysis or new experiments on true precision can reveal. Our reported experiments show that getting such estimates on unbiased samples is feasible.

Our research can be extended in many directions. To begin with, our findings may have implications on state of the art recommendation algorithms, inasmuch as they are strongly biased towards popularity. Re-examining their effectiveness in view of our findings may deserve further study. We also envision the construction of further and larger datasets as a worthy endeavor, perhaps by more coordinated efforts in the community. These should allow to further confirm, revise, or extend our findings. Different scenarios defined by different –or fewer– assumptions could be explored as well. For instance, even though random data splitting is very common practice in the recommender systems literature, we find it worthwhile exploring beyond this and consider, for instance, temporal data splits, which better represent a real setting.

ACKNOWLEDGMENTS

This work was supported by the national Spanish Government (grant nr. TIN2016-80630-P).

REFERENCES

[1] P. Adamopoulos and A. Tuzhilin. 2014. On unexpectedness in recommender systems: or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology* **5**, 4, (Jan. 2014). ACM, New York, NY, 1–32.

[2] 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 TKDE* 17, 6 (June 2005). IEEE, Piscataway, NJ, USA, 734–749.

[3] A. Bandura. 1971. *Social Learning Theory.* General Learning Press, New York.

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

[5] S. Bikhchandani, D. Hirshleifer, and I. Welch. 1992. A Theory of Fads, Custom, and Cultural Change as Informational Cascades. *The Journal of Political Economy* 100, 5 (Oct. 1992). University of Chicago Press, Chicago, IL, USA, 992–1026.

[6] R. Bredereck and E. Elkind. 2017. Manipulating Opinion Diffusion in Social Networks. In *Proc. of the 26th International Joint Conference on Artificial Intelligence* (IJCAI 2017). Morgan Kaufmann Publishers, San Francisco, CA, USA 894–900.

[7] J. Bryant and M. B. Oliver (Eds.). 2008. *Media Effects: Advances in Theory and Research, 3rd edition.* Routledge, Abingdon, UK.

[8] R. Cañamares and P. Castells. 2014. Exploring social network effects on popularity biases in recommender systems. In *6th ACM RecSys Workshop on Recommender Systems and the Social Web* (RSWeb 2014). Foster City, CA, Oct. 2014.

[9] R. Cañamares and P. Castells. 2017. 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, USA, 215–224.

[10] R. Cañamares and P. Castells. 2017. On the Optimal Non-Personalized Recommendation: From the PRP to the Discovery False Negative Principle. *ACM SIGIR Workshop on Axiomatic Thinking for Information Retrieval and Related Tasks* (ATIR 2017). Tokyo, Japan, Aug. 2017.

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

[12] O. Celma and P. Herrera. 2008. A new approach to evaluating novel recommendations. In *Proc. of the 2nd ACM Conference on Recommender Systems* (RecSys 2008). ACM, New York, NY, USA, 179–186.

[13] R. B. Cialdini and N. J. Goldstein. 2004. Social Influence: Compliance and Conformity. *Annual Review of Psychology* 55 (Feb. 2004). Palo Alto, CA, USA, 591–621.

[14] P. Cremonesi, Y. Koren, and R. Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proc. of the 4th ACM Conference on Recommender Systems* (RecSys 2010). ACM, New York, NY, USA, 39–46.

[15] 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). Informs, Catonsville, MD, USA, 697–712.

[16] S. Goel, A. Broder, E. Gabrilovich, and B. Pang. 2010. Anatomy of the long tail: ordinary people with extraordinary tastes. In *Proc. of the 3rd ACM Int. Conf. on Web Search and Data Mining* (WSDM 2010). ACM, New York, NY, USA, 201–210.

[17] F. M. Harper and J. A. Konstan. 2016. The MovieLens Datasets: History and Context. *ACM TOIS* 5, 4 (Jan. 2016). ACM, New York, NY, USA.

[18] F. M. Harper, X. Li, Y. Chen and J. A. Konstan. 2005. An Economic Model of User Rating in an Online Recommender System. In *Proc. of the 10th International Conference on User Modeling* (UM 2005). Springer, Berlin, Germany, 307–316.

[19] H. He and E. A. Garcia. 2009. Learning from Imbalanced Data. *IEEE TKDE* 21, 9 (Sept.2009). IEEE, Piscataway, NJ, USA, 1263–1284.

[20] Y. Hu, Y. Koren, and C. Volinsky. 2008. 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.

[21] 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). Kluwer Academic Publishers, Hingham, MA, USA, 427–491.

[22] G. Linden, B. Smith, and J. York. 2003. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing* 7, 1 (Jan. 2003). IEEE, Piscataway, NJ, USA, 76–80.

[23] R. J. A. Little and D. B. Rubin. 1987. *Statistical analysis with missing data.* John Wiley & Sons, Hoboken, NJ, USA.

[24] B. M. Marlin, R. S. Zemel, S. T. Roweis, and M. Slaney. 2007. Collaborative Filtering and the Missing at Random Assumption. In *Proc. of the 23rd Conf. on Uncertainty in Artificial Intelligence* (UAI 2007). AUAI Press, Arlington, VA, 267–275.

[25] B. Marlin and R. Zemel. 2009. 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.

[26] M. Moussaïd, J. E. Kämmer, P. P. Analytis, and H. Neth. 2013. Social Influence and the Collective Dynamics of Opinion Formation. *PLoS One* 8, 11 (Nov. 2013). Public Library of Science, San Francisco, CA, USA.

[27] S. A. Myers, C. Zhu, and J. Leskovec. 2012. Information diffusion and external influence in networks. In *Proc of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (KDD 2012). ACM, New York, NY, USA, 33–41.

[28] E. Pariser. 2012. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think.* Penguin Books, London, UK.

[29] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer and A. Vespignani. 2010. Characterizing and Modeling the Dynamics of Online Popularity. *Physical Review Letters* 105, 15 (Oct. 2010). APS, Ridge, NY, USA.

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

[31] M. J. Salganik, P. S. Dodds and D. J. Watts. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311, 5762 (Feb. 2006). AAAS, Washington, D.C., USA, 854–856.

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

[33] A. Sinha, D. F. Gleich, and K. Ramani. 2016. Deconvolving Feedback Loops in Recommender Systems. In *Proc. of the 30th Annual Conference on Neural Information Processing Systems* (NIPS 2016). Barcelona, Spain, Dec. 2016, 3243–3251.

[34] H. Steck. 2010. Training and testing of recommender systems on data missing not at random. In *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (KDD 2010). ACM, New York, NY, USA, 713–722.

[35] H. Steck. 2011. Item popularity and recommendation accuracy. In *Proc. of the 5th ACM Conference on Recommender Systems* (RecSys 2011). ACM, New York, NY, USA, 125–132.

[36] C. Zhai and, J. D. Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems* 22, 2 (April 2004). ACM, New York, NY, USA, 179–14.