A Unifying and General Account of Fairness Measurement in Recommender Systems

ENRIQUE AMIGÓ, Universidad Nacional de Educación a Distancia (UNED), Spain
YASHAR DELDJOO, Polytechnic University of Bari, Italy
STEFANO MIZZARO, University of Udine, Italy
ALEJANDRO BELLOGÍN, Universidad Autónoma de Madrid, Spain

Fairness is fundamental to all information access systems, including recommender systems. However, the landscape of fairness definition and measurement is quite scattered with many competing definitions that are partial and often incompatible. There is much work focusing on specific – and different – notions of fairness and there exist dozens of metrics of fairness in the literature, many of them redundant and most of them incompatible. In contrast, to our knowledge, there is no formal framework that covers all possible variants of fairness and allows developers to choose the most appropriate variant depending on the particular scenario. In this paper, we aim to define a general, flexible, and parameterizable framework that covers a whole range of fairness evaluation possibilities. Instead of modeling the metrics based on an abstract definition of fairness, the distinctive feature of this study compared to the current state of the art is that we start from the metrics applied in the literature to obtain a unified model by generalization. The framework is grounded on a general work hypothesis: interpreting the space of users and items as a probabilistic sample space, two fundamental measures in information theory (Kullback-Leibler Divergence and Mutual Information) can capture the majority of possible scenarios for measuring fairness on recommender system outputs. In addition, earlier research on fairness in recommender systems could be viewed as single-sided, trying to optimize some form of equity across either user groups or provider/procurer groups, without considering the user/item space in conjunction, thereby overlooking/disregarding the interplay between user and item groups. Instead, our framework includes the notion of statistical independence between user and item groups. We finally validate our approach experimentally on both synthetic and real data according to a wide range of state-of-the-art recommendation algorithms and real-world data sets, showing that with our framework we can measure fairness in a general, uniform, and meaningful way.

1 INTRODUCTION

The notion of fairness has recently attracted considerable attention. Fairness is studied in general in artificial intelligence and machine learning, typically focusing on classification problems [53, 74], and also in Information Retrieval (IR), with a focus on fair rankings [14, 21, 24]. However, in the field of Recommender Systems (RSs) the notion of fairness becomes multi-faceted and arguably presents a richer scenario [1, 11, 24, 48]. To evaluate if a RS is fair, one must take into account a variety of factors, including the stakeholders (consumers, producers, side-stakeholders), the kind of benefit impacting

Authors' addresses: Enrique Amigó, Universidad Nacional de Educación a Distancia (UNED) , Madrid, Spain, enrique@lsi.uned.es; Yashar Deldjoo, Polytechnic University of Bari, Bari, Italy, yashar.deldjoo@poliba.it; Stefano Mizzaro, University of Udine, Udine, Italy, mizzaro@uniud.it; Alejandro Bellogín, Universidad Autónoma de Madrid, Madrid, Spain, alejandro.bellogin@uam.es.

2022. XXXX-XXXX/2022/10-ART $15.00
https://doi.org/10.1145/1122445.1122456

the consumers and businesses/producers (perceived utility, item exposure), the context, morality, time among other variables [74]. For instance, almost every online platform we interact with, like Spotify and Amazon, functions as a marketplace connecting consumers with product producers or service providers. From the consumers' perspective, fairness mostly concerns an even distribution of effectiveness among users, avoiding the penalization of protected groups like, for example, female or black candidates in job applications.[1] Conversely, producers and item providers, who seek increased visibility, are primarily concerned with exposure fairness that should not be penalized, for example, on the basis of producers' popularity or country. Let us also remark that a fair system might provide unequal distribution of resources, as receiving a privilege can be based on merits and needs [19] or fitness [57]. Given the complexity of such a scenario, it is not surprising that the notion of fairness in RSs lacks a unified understanding. There are many definitions, which are different if not even incompatible [24, 74].

The fact is that measuring fairness has different facets, such as the consumer or producer perspective, modeling benefit in terms of exposure [22, 32, 81, 82] or utility [3, 45, 77, 90] between groups, target benefit distribution in terms of equity or merit based, etc. To our knowledge, there is no formal framework that covers all possible variants of fairness and allows developers to choose the most appropriate variant depending on the particular scenario. The result is that in many cases the authors do not identify the most appropriate metric and in many other cases different authors apply different metrics for the same purpose. This situation is of course an obstacle to progress in the field. In perspective, providing a uniform, general, standard, and unified account of fairness in RSs would be instrumental to remove such an obstacle, and that is precisely the aim of this paper.

The main contribution of this paper is the definition of a general, flexible, and parameterizable framework that covers most possibilities in fairness measurement. There is prior research that analyzes existing fairness metrics comprehensively based on a set of dimensions [78]. Theoretical frameworks of metrics that are adaptable to various circumstances have also been outlined in [19, 40, 64, 79]. Compared to previous approaches, our proposal makes the following specific contributions:

- We define such a novel framework on the basis of a comprehensive analysis of existing metrics via categorization dimensions. Thus, we start from the metrics applied in the literature to obtain a unified model by generalization, rather than starting from a unique abstract fairness definition.
- We show that by modeling exposure, utility, and effectiveness as probability distributions over the user/item space, it is possible to capture most existing fairness metrics by means of two information theory measures, namely Kullback-Leibler Divergence and Mutual Information.
- The framework allows to model some features that are absent in previously proposed metrics, such as the independence between user/item groups or individuals regardless of any ideal target benefit distribution.
- Besides, on the basis of its coverage of existing metrics, this flexible framework is validated over a synthetic data set and recommender system outputs artificially defined to cover different fairness strengths and weaknesses. The framework behavior is also checked on real data sets.

More in detail, we adopt the following methodology. To be able to comprehensively analyze the existing metrics, we first establish five dimensions along which the metrics can be classified (Section 2). As a second step, we then perform a classification of the existing fairness metrics, still focusing on

[1] In this work, we will frequently use the phrases user or customer fairness, item or producer or supplier fairness, and protected or sensitive features interchangeably.

RSs but also relating to more classical fairness definitions in classification (Section 3). Although the goal of this paper is not to serve as a systematic or semi-systematic literature review, such a thorough analysis of the literature to analyze a vast majority of the existing metrics allows us to show that the above proposed five dimensions can be used to describe in an organized and coherent way the fairness metrics landscape. As a third step, we identify the most flexible metric models in the literature to propose a general and formal framework that is based on information theory and allows us to measure fairness in a unified way. We obtain a framework that, based on a series of questions, allows us to identify the most appropriate metric for each specific scenario (Section 4). As a fourth and final step, the theoretical work is complemented with an experimental analysis on both synthetic and real data where we check the behavior of all the metrics derived from the proposed framework, and evaluate the fairness of state-of-the-art recommendation algorithms, including classical and neural algorithms, tested on real-world data sets (Section 5).

2 FAIRNESS DIMENSIONS

By analyzing the vast literature on fairness, one can note a variety of different approaches. Some research works focus on specific notions of fairness. Some others attempt to include different fairness notions. Others appear to have given up hope for a singular definition of fairness, conceding that «the English word "fairness" will need a multitude of definitions»[42, p. 11]. To make a historical comparison, the situation is not different from that about the notion of relevance in IR in the 1970s. At that time, the seminal paper by Saracevic [67] was a breakthrough that helped to clarify (i) the existence of different kinds of relevance and (ii) the possibility of classifying all of them under a common framework, by identifying some classification dimensions. This study is an effort to accomplish the same for the concept of fairness, which we feel is possible or at least worth exploring.

After a careful exploratory process for the compilation of fairness metrics, we propose to define five orthogonal dimensions (D1–D5) as independent categorization criteria on which to categorize the existing metrics on recommender systems' fairness evaluation. That is, each dimension has a number of variants associated with it. Ideally, a fairness assessment framework should be flexible enough to cover all combinations of variants of the different dimensions. The dimensions defined in this paper have been compiled from different works [24, 46, 78]. The five dimensions are described below; for each of them we provide a name, a set of possible values, and a brief description.

- *D1 – Benefit (Exposure, Effectiveness).* The first dimension concerns the type of benefit that needs to be distributed in a fair way. It can essentially take two alternative forms. The first one is to what extent items are exposed to users. Note that some previous research has named the exposure binary case as "visibility" and the ranking case as "exposure" [10, 33]; herein with "exposure" we mean both of them. The second one, i.e., the effectiveness benefit criterion, is to what extent this exposure is useful to the user. Wang et al. [78] named this dimension as *treatment* (exposure) versus *impact* (effectiveness) *optimization object.*
- *D2 – Stakeholder (Users, Items/Providers).* A core characteristic of RSs is the duality of user- and vendor-centered utility [19], also known as user/consumers and provider/producers fairness

in the literature, or for short C-Fairness and P-Fairness [12].[2] They aim at a fair treatment of the users and of the producers, respectively. Both can be considered simultaneously, which is called *two-sided fairness* [2, 24, 47, 80] and some authors extend this idea to multi-stakeholders, including the own system interest [18]. This dimension has been referred as *subject* [46, 78] and also as *Consumer* vs. *provider fairness* [24].

- *D3 – Partition Granularity (Two Groups, Many Groups, Individuals).* Fairness usually entails comparing, on average, the benefit received by the members of different groups. The granularity of the partition into groups can vary along a spectrum: at one extreme, only two groups are defined, i.e., privileged and unprivileged groups as defined by their protected attributes; in more general cases, there are several groups over which maintain equity; and at the other extreme, fairness is studied between individuals, e.g., equal recommendation effectiveness for each user. In general, fairness covers everything from the division of user or item spaces into two groups, to many groups, to consider individuals (which subsumes the previous two cases). We will see that not all metrics capture all possibilities. Wang et al. [78] named this dimension as *target*.
- *D4 – Exposure Scheme (Rating, Set, Ranking).* The existing fairness metrics differ depending on how the items are displayed to the user. Capturing different information access user interfaces is crucial for the generality of fairness measurement. In some cases, the RS exposes the items to the users according to an estimated rating (e.g., 1 to 5 scale). In other cases, the user interface consists of a set of recommended items without any priority order. In most cases, items are organized in a ranking, or into a ranking of categories, or even a ranking of rankings [34]. The fairness measurement framework should be able to weight the exposure of items in all these situations. We will see that most current metrics are oriented to specific exposure schemes, while others encapsulate this dimension in an exposure weight parameter. This dimension has been referred to as provider *representation measure* [24, 40] or *attention* [31].
- *D5 – Fairness Criterion (Parity, Size Proportionality, Utility Proportionality, Independence).* This last dimension concerns the overall criterion used to state that distributing in a certain way the benefits across individual users/items or groups of users/items is fair. According to Kirnap et al. [40], there are three main ways to define such a target distribution, i.e., on the basis of parity, proportionality to the corpus presence, and proportionality to utility [40]. Parity implies that all groups receive the same exposure mass, proportionality implies that the exposure is proportional to the group size, and utility implies that the exposure is proportional to the relevance mass of the group. Deciding the target benefit distribution that enables a reasonable allocation of resources can be task-specific [19] and extremely complex for the system designer, since it can follow norms of short-term user satisfaction, long-term business growth, morality, among others.

Besides these fairness criteria, fairness can be measured on the basis of statistical *Independence* of user's or item's protected attributes. For example, in a job recommendation setting, the exposure to executive vs. low qualified jobs could be required to be independent of protected characteristics of users

[2]Note that in the literature, consumer and user fairness are frequently used as synonyms. The same applies for item fairness, provider fairness, and producer fairness. Also, in situations where the roles of the user and the items are reversed, such as recommendation or people for a certain job [4], users and items need to be swapped.

such as their age and gender, which is equal to say that the resource allocation should be unbiased by protected characteristics. As such, we can observe a connection between statistical independence and treatment disparity [19, 24, 69], which embodies the idea that the system should make judgments (exposure) regardless of the individual's protected attributes.

Note that we do not claim that this set of dimensions is complete. However, the analysis in the next section shows that this dimension set is enough to capture the limitations of existing metrics in terms of scenario coverage. Therefore, we can use such a set for evaluating the generality of fairness measurement approaches. The reader is referred to Section 6 for a discussion on limitations and outlooks.

3 FAIRNESS METRICS

Wang et al. presented an interesting survey [78] where they categorized and defined many metrics for fairness evaluation in recommender systems. Rather than presenting a comprehensive catalog of metrics and their definitions, we aim at analyzing to what extent the metrics proposed in the literature are general or if they are actually limited to different particular scenarios. More specifically, we are interested in identifying those metric schemes that allow to capture diverse fairness scenarios. Unfortunately, the number of existing metrics is very large and there would not be space in the article to include their definitions. Since the focus of this article is to evaluate the scenario coverage of the metrics, we are interested in including as many as possible and at least describe their properties in terms of coverage over possible fairness scenarios.

As a result, Table 1 summarizes how each metric or measurement approach (on the rows) captures each variant under the five dimensions presented above (columns). For each column, the meaning of the symbols is as follows: Exp and Eff (D1) mean exposure and effectiveness oriented; Us and It (D2) represent user and item (provider) serving as RSs' main stakeholders; 2g (D3) represents that the metric is defined for two groups (protected and non protected), ng represents that the metric can be defined over many groups and I represent that the metric is defined only for individuals (if all granularity levels can be captured, including individuals, we use the ✓symbol); Rat, Set, and Rank (D4) mean that the exposure is set-, ranking-, or rating-based; P, S, U, and I (D5) represent that the fairness criterion is based on Parity, Size proportionality, Utility proportionality, or Independence. The tick symbol (✓) represents that the metric can be customized into all the variants of the corresponding dimension.

We describe each group of metrics in each of the following subsections, as indicated in the table; within each subsection we generally follow the same metric order as in the table, and we group existing metrics according to D1 (benefit) and D5 (fairness criterion). We independently analyze metrics that are based on exposure but consider utility proportionality as fairness criterion. We also consider separately some general frameworks and some metrics that capture the independence fairness criterion.

Since the scope of this article is the space of metrics that quantify fairness on the basis of system outputs rather than the recommendation algorithm process, we believe our work aligns more with the *outcome fairness* rather than *process fairness* [78]. The second evaluates aspects such as what data has been used, under what principles the system makes decisions, or what are the causal relationships between inputs and outputs. In contrast, outcome fairness ignores how the system works internally and focuses on the fair distribution of benefits. Finally, we feel that outcome fairness is an established topic with a large number of metrics, making this an ideal time to build a generalization; in contrast, the many perspectives on process fairness are still in the early stages of research.

Table 1. Dimensions captured by fairness measures (part I). Meaning of symbols is as follows. Exp and Eff (D1): exposure and effectiveness oriented. Us and It (D2): user and item (provider) as main stakeholders. 2g, ng, and I (D3): two groups (protected and non protected), many groups, and individuals (if all granularity levels can be captured, including individuals, we use the ✓symbol). Rat, Set, and Rank (D4): exposure is set-, ranking-, or rating-based. P, S, U, and I (D5): fairness criterion is based on Parity, Size proportionality, Utility proportionality, or Independence. ✓: the metric can be customized into all the variants of the corresponding dimension.

	D1 Benefit (Exp/Eff)	D2 Stakeholder (Us/It)	D3 Partition (2g/ng/I)	D4 Exposure (Rat/Set/Rank)	D5 Criterion (P/S/U/I)
Fairness Measures Based on Exposure (Section 3.1)					
Ranked Group Fairness Condition [86]	Exp	It	2g	Rank	P
Fairness Constraint [14]	Exp	It	2g	Rank	P
rND, rKl, rRD [83]	Exp	It	2g	Rank	S
Skew@k [29]	Exp	It	2g	Rank	S
Rank Parity [43]	Exp	It	2g	Rank	S
Disparate Exposure [10]	Exp	It	2g	Rank	S
Attention Bias Ratio [31]	Exp	It	✓	Rank	S
Product Ranking Fairness [75]	Exp	It	✓	Rank	S
NKLD [29]	Exp	It	✓	Rank	S
Inequality in Producer Exposures [60]	Exp	It	✓	Set	P
Uniform Fairness Variance [80]	Exp	It	✓	Set	P
Equity of Attention for Group Fairness [30]	Exp	It	✓	Set	P
Gini Index [27]	Exp	It	I	Set	P
Jain's fairness index [87]	Exp	It	I	Set	P
Fraction of Satisfied Producers [60]	Exp	It	✓	Set	P
Average Provider Coverage Rate [52]	Exp	It	✓	Set	P
Group Fairness Measure [55]	Exp	It	✓	Set	P
Supplier Popularity Deviation [30]	Exp	It	✓	Set	S
MAD [88]	Exp	It	2g	✓	S
Non-Parity Fairness [84]	Exp	It	2g	✓	S
Demographic Parity [69]	Exp	It	2g	✓	S
Gupta et al. [35]	Exp	It	✓	✓	S
Fairness Measures Based on Effectiveness (Section 3.2)					
Absolute Difference [88]	Eff	Us	2g	✓	S
KS statistic [88]	Eff	Us	2g	✓	S
Effectiveness Standard Deviation [60, 80]	Eff	Us	✓	✓	S
Rating Prediction Fairness [75]	Eff	✓	✓	Rat	S
Wang and Joachims [77]	Eff	Us	✓	✓	S
Yao and Huang [84]	Eff	Us	2g	Rat	S
User Bias [51]	Eff	Us	2g	✓	S
Pairwise Fairness [8]	Eff	It	ng	✓	S
Item Bias [51]	Eff	It	2g	✓	S
Disparate Relevance [10]	Eff	It	2g	Rank	S

3.1 Fairness Metrics Based on Exposure

Fairness in terms of exposure in RSs is related to previous fairness metrics for classification in Artificial Intelligence, such as *Fairness Based on Predicted Outcome* [74] (also called *Statistical Parity* [23], *Equal*

Table 1. Dimensions captured by fairness measures (part II).

	D1 Benefit (Exp/Eff)	D2 Stakeholder (Us/It)	D3 Partition (2g/ng/I)	D4 Exposure (Rat/Set/Rank)	D5 Criterion (P/S/U/I)
Fairness Measures Based on Utility-Equalized Exposure (Section 3.3)					
Supplier Popularity Deviation [2]	Exp	It	✓	Set	U
Mean Average Calibration [18]	Exp	It	✓	Set	U
JS-Divergence [56]	Exp	It	✓	Set	U
Rank Equality [43]	Exp	It	2g	Rank	U
Steck [70]	Exp	It	✓	✓	U
Equity of Amortized Attention [9]	Exp	It	✓	✓	U
Quality Weighted Fairness [80]	Exp	It	✓	✓	U
Disparate Treatment Ratio [69]	Exp	It	2g	✓	U
Flexible Fairness Measures (Section 3.4)					
Wu et al. [79]	✓	✓	✓	Set	P/S/U
Kirnap et al. [40]	Exp	It	✓	✓	P/S/U
Sacharidis et al. [64]	Exp	✓	✓	✓	P/S/U
Deldjoo et al. [19]	✓	✓	✓	✓	P/S/U
Fairness Measures Based on Independence (Section 3.5)					
Relative Opportunity [12]	Exp	✓	2g	Set	I
Bias Disparity [72]	Exp	✓	2g	Set	I

Acceptance Rate [89], and *Benchmarking* [68]). A classifier satisfies these definitions if the probability of being assigned to the positive predicted class is equal across different item groups. In the case of RSs, we can instead speak of item exposure. This set of metrics includes those that evaluate the equity of exposure across user or item groups (D1=Exp). In general, the exposure fairness in RSs is commonly defined from the item side. One reason is that the item providers are interested in gaining visibility.

We start by identifying a set of IR metrics that focus on the relative presence of protected and non-protected item groups (D2=It) in the top-k ranking positions. *Ranked Group Fairness Condition* [86] and the *Fairness Constraint* [14] specify upper and lower bounds on the number of items from each group that are allowed to appear in the top-k positions of the ranking; both are parity oriented (D5=P).

In other metrics, the fairness criterion is size-proportional (D5=S). For instance, Yang and Stoyanovich [83] proposed three metrics, namely *Normalized Discounted Difference (rND), Divergence (rKL),* and *Ratio* (rRD), that compare the distribution of the protected group above a certain ranking position with the group presence in the corpus. Geyik et al. proposed the metric Skew@k which is similar, and compares protected with non-protected groups [29]. A common feature of all these metrics is that the fairness score is averaged across ranking position thresholds. On the contrary, other rank oriented metrics such as *Rank Parity* [43] quantify exposure in terms of the cases in which items from one group are ranked above another group. The *Disparate Exposure* proposed by Boratto et al. [10] computes the difference between the minority group representation in the item catalog and the average exposure taking into account the ranking positions of items.

A common limitation of all the previous metrics is that they are defined for two groups (D3=2g). The *Attention Bias Ratio* [31] addresses this limitation by quantifying the disparity between the groups with the lowest and highest mean exposure, considering the ranking position bias. Another metric which is able to capture multiple groups is the *Product Ranking Fairness* which computes the Kullback-Leibler Divergence (KLD) between the amount of top ranked items and the item group size [75]. A similar metric is NDKL which aggregates KLD values across ranking position thresholds [29].

In the context of RSs we found a set of metrics that are able to capture multiple groups (D3=✓), but limited to set-based exposure (D4=Set). Some metrics study the exposure variance across groups. Some examples are the entropy-based metric *Inequality in Producer Exposures* [60], the *Uniform Fairness Variance* [80], the *Equity of Attention for Group Fairness* [30], the Gini Index [27], or the *Jain's fairness index* [87]. The *Fraction of Satisfied Producers* [60], the *Average Provider Coverage rate* (APCR) [52], and the *Group Fairness Measure* [55] are similar but based on the number of providers (item groups) covered in single user lists. The *Supplier Popularity Deviation* [30] is also similar, but taking into account the item group size (D5=S).

On the other hand, the absolute difference between mean ratings of two groups (MAD), which is extended with the Kolmogorov-Smirnov statistic [88], captures graded exposure (including set and ranking, D4=✓), but it is defined for two groups (D3=2g). The same applies to the *Non-Parity Fairness* [84] and the Singh and Joachims's [69] *Demographic Parity*. Finally, the Gupta et al.'s [35] *Demographic Disparity* is computed as the maximum difference of exposure between group pairs, where exposure includes the ranking discount function. A common property of these previous RS fairness metrics is that they are size proportional (D5=S), i.e., group exposure is normalized according to the group size.

3.2 Fairness Metrics Based on Effectiveness

Recommendation effectiveness is a natural benefit function. This links with the classification fairness notion *Predictive Parity*: *"both protected and unprotected groups have equal probability of a subject with positive predictive value to truly belong to the positive class"* [74]. It is also equivalent to *Outcome Test* [68], *Equal opportunity* [15, 36, 44], and *False negative error rate balance* [16]: positive samples from different groups have equal probability to be classified as positive.

One major consumer-side group fairness problem is to determine whether the system provides comparable quality of service or utility to different groups of consumers. This family of metrics includes those that focus on the equity of recommendation effectiveness across user groups (D1=Eff, D2=Us). In general, since these metrics work with expected effectiveness of individual users, the fairness criterion is size proportionality (D5=S).

In the contexts of IR and RSs, a common fairness evaluation procedure in the literature consists in comparing the expected effectiveness of user groups [25, 54], for instance, using the *Absolute Difference* [88] or the Kolmogorov-Smirnov statistic [88]. The standard deviation of expected effectiveness across individuals or user groups is a common way to quantify fairness [60, 80], allowing multiple groups (D3=✓). This method is agnostic regarding the effectiveness metric, as it captures both set and ranking based exposition (D4=✓). Similarly, the *Rating Prediction Fairness* proposed by Wan et al. [75] applies the ANOVA test over the null hypothesis of independence between prediction errors and market segments. As well as accepting multiple groups (D3=✓), this method allows to define market segments over both user and items (D2=✓) but it is only rating oriented (D4=Rat).

Yao and Huang [84] defined a set of alternative metrics, namely, *Value Unfairness*, *Absolute Unfairness*, *Underestimation Unfairness*, *Overestimation Unfairness*, and *Non Parity*. They all compare, for each item, the expected score for disadvantaged and advantaged users. They are limited to two user groups (D3=2g) and top ranking heaviness is not captured since they define the recommendation problem as an item rating prediction problem (D4=Rat). Wang and Joachims [77] defined a user fairness metric that quantifies the effectiveness equity across multiple user groups through a social-welfare function. It captures multiple groups (D3=✓) and graded exposure (D4=✓). The *User Bias* proposed by Lin et al. [51] can be also applied to any effectiveness metric (D4=✓), but it is defined for only two groups (D3=2g).

From the provider perspective, one major group fairness problem is to determine whether the system provides comparable quality of service or utility to different providers, i.e., useful items from different providers have equal opportunity to be exposed. However, metrics for this aspect are not very common in RSs. An exception is the *Pairwise Fairness* which computes the probability that a useful (clicked in the user feedback) item is ranked above another useless item within a certain item group [8]. It allows to compare multiple item groups (D3=ng) but not individual items, and the target distribution is proportional to size since it is defined as a probability (D5=S). Another exception is the *Item Bias* [51], which computes the difference between effectiveness metrics over two item sets (D3=2g). Finally, the *Disparate Relevance* proposed by Boratto et al. [10] is somewhat particular; it computes the difference between the minority group representation in the item catalog and the estimated relevance of their exposed items.

3.3 Fairness Metrics Based on Utility-Equalized Exposure

In some cases a uniform exposure distribution is not fair. It is natural to think that the exposure of suppliers should be proportional to the amount of useful items they provide. This is related to classification fairness metrics such as *Equalized odds* [36], *conditional procedure accuracy equality* [7], and *disparate mistreatment* [85]: protected and unprotected groups have equal true positive rate, i.e., the probability of true instances (useful items in RSs) to be classified as true (exposed items in RSs). The benefit function is exposure (D1=Exp) but the ideal distribution is related to item utility (D5=U).

Some utility equalized exposure metrics are oriented to set exposure (D4=Set). For instance, *Supplier Popularity Deviation* [2] and *Mean Average Calibration* [18] sum the absolute differences between the ratio of recommendations and ratings that come from items of supplier. There exist some ranking oriented utility equalized metrics, like *Rank Equality* [43] that computes the number of times an item of a group is falsely given a higher rank than an item of another group. It can be applied to two item groups (D3=2g). Modani et al. used the Jensen-Shannon Divergence to compare the exposure and the utility provided by item groups [56].

Other approaches are agnostic as to the type of exposure function (set, ranking, etc.). Steck defined a metric in terms of KLD between exposure weight and utility (according to the user's previous preferences) of item groups [70]. In the context of IR, *Equity of Amortized Attention* [9] is based on the L1-norm distance between accumulated exposure and relevance of single item groups. The *Disparate Treatment Ratio* compares ratios of exposure with utility of group pairs [69]: it is applied to two item groups (D3=2g). The *Quality Weighted Fairness* computes the variance of exposure/utility ratios across item groups, capturing the rank exposure bias and multiple item groups [80].

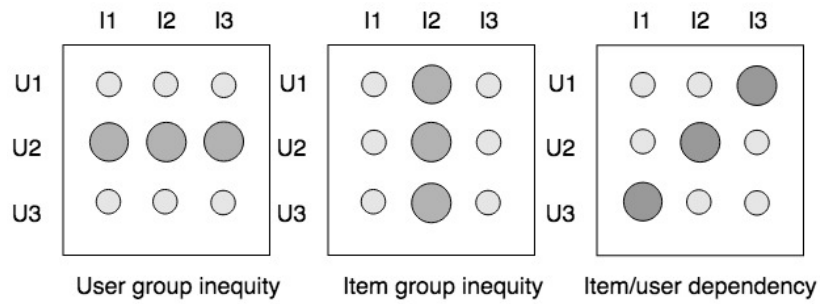

Fig. 1. The notion of independence based fairness. Each circle represents the amount of benefit (e.g., exposure) for each item (column) and user group (row).

3.4 Flexible Fairness Metrics

Some authors have proposed flexible fairness measurement models that can be instantiated to particular scenarios. The common feature of these approaches is that the user/item bidimensional space is divided according to user or item groups. Then, the benefit distribution across groups is compared against an ideal (fair) distribution. These models are flexible enough to consider any fairness criterion (D5=P/S/U) and one or many user/item groups (D3=✓).

In this line, Wu et al. [79] proposed to compute the average differences. The effectiveness benefit function is implicitly captured by considering item relevance as target exposure. The limitations of this framework is that the management of ranking exposure is not specified (D4=Set). In the context of IR, Kirnap et al. [40] proposed a general theoretical framework consisting of: (i) the exposure distribution, which decreases with rank according to decay functions in IR evaluation metrics [21]; (ii) the target distribution, which can be *parity*, *proportionality to the corpus presence*, or *proportionality to the relevance* (D5=P/S/U); (iii) the similarity between the exposure distribution of groups and the target distribution across different rank thresholds in terms of KLD or other distribution similarity metrics. In the context of RSs, the framework proposed by Sacharidis et al. [64] is very similar. It computes the KLD between the real and the desirable exposure distribution across user or item groups. However, the way of managing ranking exposure is not specified and the model is limited to the exposure benefit criterion (D1=Exp). In fact, the KLD has been used by different authors to compare the actual and the desirable benefit distribution across groups [28, 49, 71] . Deldjoo et al. [19] proposed another similar framework. The utility distribution across user/item groups is compared with the ideal distribution via the *Generalized Cross Entropy* which is more robust against outliers. The fairness metric can be instantiated into an exposure fairness metric by considering all items equally useful. In addition, the Deldjoo et al.'s model captures the effectiveness benefit function (D1=✓) and ranking exposure weighting (D3=✓).

3.5 Fairness Metrics Based on Independence

All the metrics discussed in the previous subsections refer to equity across user or item groups. However, there also exists the notion of independence between user and item groups which is related with treatment disparity. As an example, let us assume that both men and women receive the same amount of effective items (e.g., jobs), and that all item groups are equally distributed in terms of exposure quantity and quality to users. Even in this situation, it may be the case that the user genre conditions the type of recommended jobs. Figure 1 illustrates the notion of independence based fairness. Each panel represents the benefit distribution (e.g., exposure) across three user groups and three item groups. The larger the circles, the more the corresponding item group is exposed to the corresponding user group. In the left case there is no parity across user groups (the user group U2 receives more exposure). In the second distribution, there is no parity across item groups (the item group I2 is exposed to a larger extent). In the third case, there is parity across user and item groups (two small and one big circle for each column or row). However, user groups U1, U2, and U3 are biased to item groups I3, I2, and I1 respectively. In other words, since the benefit across user and item groups is not independent, there is treatment disparity.

In a more abstract way, we can say that the metrics described previously quantify the equity across user or item groups, whereas the metrics we are considering now deal with the dependence between attributes of users and items. Consequently, there is no target distribution in independence oriented fairness, but an independence requisite (D5=I).

As far as we know, the study of fairness as independence between attributes has been addressed by very few authors. One exception is the *Relative Opportunity* metric proposed by Burke et al. [12]. In this metric, fairness is quantified as the ratio between the relative frequency of gender-protected items in one user group and in the other user group. Thus, if the group to which the user belongs and the gender of the item are statistically independent, then the mean returns the neutral value 1. This metric of independence-based fairness is limited in that it is not generalizable to more than two groups of users and items (D3=2g) and it does not capture rank position bias (D4=Set). Another exception is the *Bias Disparity* proposed by Tsintzou et al. [72]. The authors study the ratio between the frequency of the item category in a group of users versus the overall frequency of the item category. If the category of items and the group of users is statistically independent, the metric returns 1. Furthermore, this bias is compared against the original bias in the users' preferences, thus analyzing the extent to which the system introduces biases against the original data. This metric has the same limitations as the previous one.

3.6 Discussion: To What Extent Metrics Capture Diverse Scenarios

Looking at Table 1, two observations can be made in relation to the analysis of fairness metrics. The first one is that there are predominant dimension variants in the metrics proposed and studied by the community. For example, effectiveness as a benefit function tends to be user-oriented rather than item group-oriented, and the item-oriented fairness metrics based on effectiveness are group size proportional. That is, none of them take as fairness criterion parity (D5=P) or the utility mass provided by item groups (D5=U). In other words, there exist combinations of dimension values that are not captured by specific metrics in the literature.

These alternative scenarios can be captured by flexible fairness metrics. In this respect, our second observation is that many of the generalists approaches (Steck [70], *Equity of Amortized Attention* [9], Deldjoo et al. [19], and Kirnap et al. [40]) apply KLD between benefit function distributions. In particular, the Deldjoo et al. [19] and Sacharidis et al. [64] models capture most dimension variants.

However, the outstanding issue is still the independence fairness criterion (D5=I). More specifically, *Relative Opportunity* [12] and *Bias Disparity* [72] metrics only allow comparing two groups and do not capture graded exposure. Like the information theoretic metric KLD generalizes equity fairness aspects, according to our intuition, the information theoretic metric Mutual Information (MI) is the most appropriate for independence analysis.

In sum, after analyzing the great variety of existing metrics, the hypothesis on which the framework proposed in this paper is based is that, *interpreting the space of users and items as a probabilistic sample space, the two fundamental measures in information theory (KLD and MI) can capture most possible scenarios of fairness measurement on recommendation system outputs.*

4 THEORETICAL FRAMEWORK

As discussed in the previous sections, we consider fairness as *the equity or independence of user or item groups regarding a certain benefit distribution*. According to our analysis of fairness metrics, the KLD between the benefit distribution across groups and the ideal distribution [19, 40] captures most dimensions of existing fairness metrics. In our general framework, we consider this schema for equity fairness. At theoretical level, the main particularities of the proposed framework with respect to previous approaches is that, not only exposure, but also the item exposure effectiveness is modeled as a probability distribution over single user/item pairs. The second contribution is that we also define a metric based on MI to capture independence between user/item groups and individuals regardless any target benefit distribution (independence-based fairness).

4.1 Framework Definition

Figure 2 illustrates the fairness framework and its notation. Let $\mathcal{U}$ and $\mathcal{I}$ be the sets of users and items, respectively. To denote the elements of these sets we use $u \in \mathcal{U}$ and $i \in \mathcal{I}$. Let $\mathcal{A}_{\mathcal{U}}$ and $\mathcal{A}_{\mathcal{I}}$ be the sets of user and item *attributes* (e.g., $\mathcal{A}_{\mathcal{U}} = \{\text{male}, \text{female}\}$). $\psi(u,i)$ and $\phi(u,i)$ represent the *utility* and the *exposure* of the item i for the user u respectively. Both are functions from the user/item space $\mathcal{U} \times \mathcal{I}$ to $(0,1)$. The exposure function is interpreted as item accessibility, or the probability of the user u to access i. Then, the *Exposure Effectiveness* $\texttt{Eff}(u,i)$ represents to what extent an item exposure to a user is effective and is modeled as: $\texttt{Eff}(u,i) = \phi(u,i) \cdot \psi(u,i)$. One way to interpret the above definitions is as follows: $\psi(u,i)$ is a user-defined function, while $\phi(u,i)$ is a system-driven function. While $\psi(u,i)$ answers the question of "*how much user u judges item i useful*", $\phi(u,i)$ represents "*how much the system provides opportunity to a user-item pair meet*". If one of these two functions, ϕ or ψ, decreases then $\texttt{Eff}(u,i)$ decreases as well. This effectiveness formalization is similar to the one defined in the Deldjoo et al.'s [19] model, but instead of operating on an individual user (aggregate) level, i.e., $\texttt{Eff}(u)$, it measures effectiveness for each user/item pair $\texttt{Eff}(u,i)$.

We can then normalize the functions ψ, ϕ, and $\texttt{Eff}$, obtaining three probability distributions P_θ with $\theta \in \{\psi, \phi, \texttt{Eff}\}$ over the user/item space. That is, $P_\theta(u,i) = \frac{\theta(u,i)}{\sum_{u \in \mathcal{U}, i \in \mathcal{I}} \theta(u,i)}$. Given a distribution $P_\theta(u,i)$, we can infer the probability associated to user attributes (e.g., $P_\theta(\text{male}, \mathcal{I})$), item attributes (e.g., $P_\theta(\mathcal{U}, \text{action films})$), or combinations of user and item attributes (e.g., $P_\theta(\text{male}, \text{action films})$).

Our generalized fairness measurement model is based on two parameterizable metrics. Table 2 illustrates the possibilities. First, the inequity of user or item groups is quantified via KLD between the

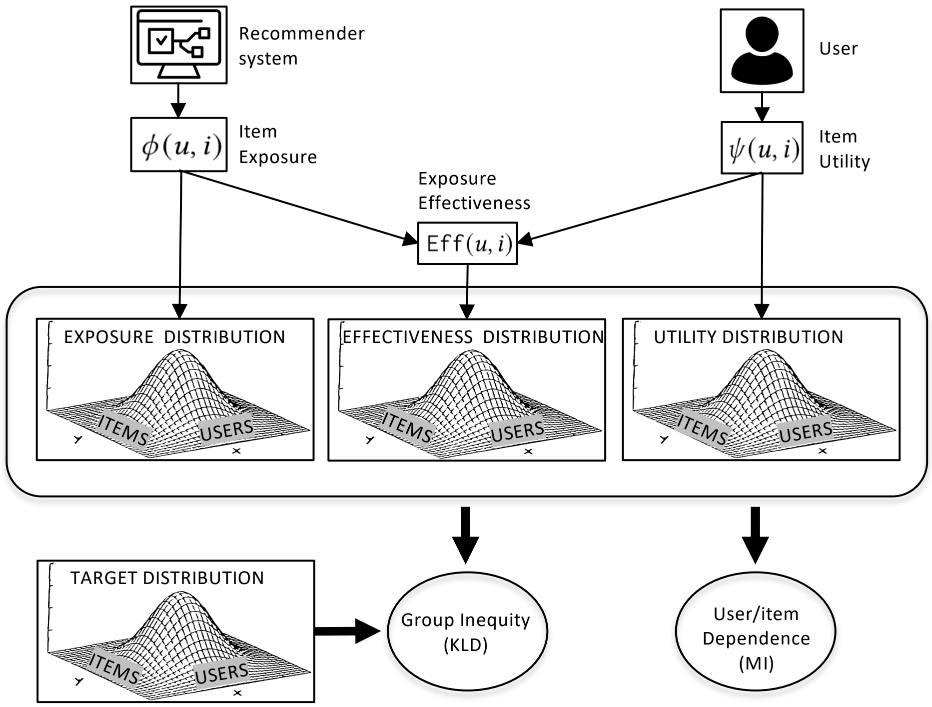

Fig. 2. The main components of the proposed fairness framework; exposure ($\phi(u,i)$), utility ($\psi(u,i)$), effectiveness ($\mathtt{Eff}(u,i)$), their distributions over users and items, the comparison with the target distribution by means of KLD, and the measure of independence by means of MI.

real (P_θ) and fair (Q) benefit distribution. Following the scheme proposed by other authors [18, 19, 40, 83]:

$$\text{Inequity}(\theta, \text{Q}, \mathcal{A}_X) = D_{KL}(P_\theta \mid\mid Q; \mathcal{A}_X) = \sum_{x \in \mathcal{A}_X} P_\theta(x) \log \frac{P_\theta(x)}{Q(x)}. \tag{1}$$

The groups partition $\mathcal{A}_X$ can be user or item oriented. The benefit function θ can be based on utility (ψ), exposure (ϕ), or effectiveness ($\mathtt{Eff}$). The target distribution Q can be parity (equal benefit, $Q(x) = {}^1/_{|\mathcal{A}_X|}$), proportional to the user group size ($Q(x) = {}^{|\{(u,i)\in x\}|}/_{|\mathcal{U}\times\mathcal{I}|}$), or proportional to utility ($Q(x) = P_\psi(x)$). In other words, being x a certain user or item attribute, in the case of parity the distribution Q is uniform, in the case of user group size the distribution Q corresponds to the group size, and in the case of proportionality to utility the distribution Q corresponds to the group utility mass.

Second, the treatment disparity is captured via group dependence, which is measured with Mutual Information (MI):

$$\text{Dependence}(\theta, \mathcal{A}_X, \mathcal{A}_Y) = I_\theta(\mathcal{A}_X; \mathcal{A}_Y) = \sum_{x \in \mathcal{A}_X, y \in \mathcal{A}_Y} P_\theta(x,y) \log \frac{P_\theta(x,y)}{P_\theta(x) \cdot P_\theta(y)}. \tag{2}$$

Table 2. Fairness metric variants according to the general fairness measurement framework. The most popular fairness notions in the literature are highlighted in bold.

Inequity $D_{KL}(P_\theta \,||\, Q; \mathcal{A}_X)$ (Eq. 1)

Stakeholder	Benefit Function	Fairness Criterion	$\mathcal{A}_X$	θ	$Q(x)$
User Groups	Exposure Based	Parity	$\mathcal{A}_\mathcal{U}$	ϕ	$1/\lvert\mathcal{A}_\mathcal{U}\rvert$
		Size Proportionality			$\lvert\{(u,i)\in x\}\rvert/\lvert\mathcal{U}\times\mathcal{I}\rvert$
		Utility Proportionality			$P_\psi(x)$
	Effectiveness Based	Parity		Eff	$1/\lvert\mathcal{A}_\mathcal{U}\rvert$
		Size Proportionality			$\lvert\{(u,i)\in x\}\rvert/\lvert\mathcal{U}\times\mathcal{I}\rvert$
		Utility Proportionality			$P_\psi(x)$
Item Groups	Exposure Based	**Parity**	$\mathcal{A}_\mathcal{I}$	ϕ	$1/\lvert\mathcal{A}_\mathcal{U}\rvert$
		Size Proportionality			$\lvert\{(u,i)\in x\}\rvert/\lvert\mathcal{U}\times\mathcal{I}\rvert$
		Utility Proportionality			$P_\psi(x)$
	Effectiveness Based	Parity		Eff	$1/\lvert\mathcal{A}_\mathcal{U}\rvert$
		Size Proportionality			$\lvert\{(u,i)\in x\}\rvert/\lvert\mathcal{U}\times\mathcal{I}\rvert$
		Utility Proportionality			$P_\psi(x)$

Dependence $I_\theta(\mathcal{A}_X; \mathcal{A}_Y)$ (Eq. 2)

User partition	Item partition	Benefit Distribution	$\mathcal{A}_X$	$\mathcal{A}_Y$	θ
User Groups	Items	Exposure Based	$\mathcal{A}_\mathcal{U}$	$\mathcal{I}$	ϕ
		Effectiveness Based			Eff
User	Item Groups	Exposure Based	$\mathcal{U}$	$\mathcal{A}_\mathcal{I}$	ϕ
		Effectiveness Based			Eff
User Group	Item Groups	**Exposure Based**	$\mathcal{A}_\mathcal{U}$	$\mathcal{A}_\mathcal{I}$	ϕ
		Effectiveness Based			Eff

$\mathcal{A}_X$ and $\mathcal{A}_Y$ represent the user and item partitions. When considering single items ($\mathcal{A}_Y = \mathcal{I}$) and user groups ($\mathcal{A}_X = \mathcal{A}_\mathcal{U}$) we are measuring to what extent the user group does not influence the exposed items. In other words, the user group does not provide information about what items are recommended to the users in that group. In the same way, when considering single users ($\mathcal{A}_X = \mathcal{U}$) and item groups ($\mathcal{A}_Y = \mathcal{A}_\mathcal{I}$) we are measuring to what extent the item group does not influence to which users the items are exposed. When considering both user and item groups, we are checking that user and item groups do not influence each other.

4.2 Framework Generalization Power

Our framework provides 18 fairness metric instances. It includes both effectiveness and exposure oriented fairness depending on the benefit function $\theta \in \{\phi, \mathtt{Eff}\}$ (Dimension D1). In addition, our effectiveness function `Eff` generalizes ranking metrics under the scheme proposed by Carterette [13] in IR or Singh and Joachims [69] in RSs, where ϕ represents the ranking decay function such as $1/\log(\mathrm{rank}(u,i))$ in DCG or $p^{\mathrm{rank}(u,i)}$ in RBP. In the set exposure context, `Eff` generalizes classification metrics such as Accuracy ($\phi(u,i) \in \{0,1\}$ and $\psi(u,i) \in \{0,1\}$). It can also generalize Precision or Recall by normalizing the utility with respect to the amount of relevant items in the collection or group, or the amount of exposed items. It also captures both user and producer stakeholders depending on whether we split the user/item space according to user ($\mathcal{A}_X = \mathcal{A}_\mathcal{U}$) or item attributes ($\mathcal{A}_X = \mathcal{A}_\mathcal{I}$), complying with Dimension D2. The possibilities of Dimension D3 are also captured since we can consider two or more attribute values or even individuals ($\mathcal{A}_\mathcal{U} = \mathcal{U}$ or $\mathcal{A}_\mathcal{I} = \mathcal{I}$). The variants of Dimension D4 are also captured: the exposure function ϕ and the effectiveness function Eff can be adapted to ranking or set exposure schemes; the rating scenario requires to state the exposure value for each rating or the translation to a ranking scenario (sorting items by rating). Finally, the variants in Dimension D5 are captured by the flexibility of the target distribution Q (Parity, Size Proportionality, Utility Proportionality) and the application of dependence instead of inequity.

In addition, the framework captures many possibilities that are not covered in the literature. For instance, most of user group oriented inequity metrics in the literature are oriented to effectiveness and based on group-size proportionality: $D_{KL}(P_{\mathtt{Eff}} \,||\, P; \mathcal{A}_\mathcal{U})$ (D1=Eff, D2=Us, and D5=S). However, in recommendation scenarios with a variable amount of exposed items per user, metrics based on exposure (D1=Exp) could be useful. One could be also interested in distributing the effective exposure mass uniformly across user groups regardless of their size (D5=P), or proportional to their needs (D5=U).

Conversely, within the item group inequity metrics, most of exposure based metrics are based on the uniform target distribution ($D_{KL}(P_\phi \,||\, 1/|\mathcal{A}_\mathcal{I}|; \mathcal{A}_\mathcal{I})$), or the utility-equalized ($D_{KL}(P_\phi \,||\, P_\psi; \mathcal{A}_\mathcal{I})$), with the exception of Yang and Stoyanovich's approach which applies the group size proportionality ($D_{KL}(P_\phi \,||\, P; \mathcal{A}_\mathcal{I})$) [83]. We find in the literature two item group equity metrics oriented to effectiveness [8, 15]. Both use the item group size as target distribution. However, one could be interested in giving the same amount of effective exposures to all items groups regardless the amount of items they provide (parity) or in providing effective exposures to item groups with respect to their item utility (utility-equalized).

Regarding dependence based fairness (treatment fairness), the only two metrics that we found in the literature are exposure oriented and combine user and items groups [12, 72] (see Section 3.5). However, the treatment fairness in terms of effective exposures can be the focus in certain scenarios.

In sum, the proposed generalized model captures most fairness notions measured in the literature while opening the door for new fairness variants. In addition, it overcomes many limitations presented by the other metrics. For instance, the item oriented exposure fairness metrics in the literature capture ranking exposures but only for two groups, or capture many groups but only on item set exposures (Section 3.1). Existing independence-based metrics capture only two groups and item set exposure.

It should be noted that there are many aspects of fairness measurement that remain unresolved. Patro et al. [61] identify some of them as provider utility beyond position based exposure, temporal effects, cross-platform effects, or the use of positioning strategies. A positive aspect of the proposed theoretical framework is that the vast majority of these aspects can be encapsulated within the exposure or utility functions, so that the framework does not lose generality.

5 EXPERIMENTS

To validate the soundness and generality of our proposal, we perform[3] experiments on both synthetic and real data sets with state-of-the-art system outputs. Our research questions are:

- *RQ1. Do the metrics capture those aspects of fairness for which they are designed?* To answer this question, we use synthetic data. We artificially generate a distribution of users, items, and preferences, as well as seven system outputs with known biases to check that each metric captures specific aspects.
- *RQ2. Is there a trade-off between fairness and effectiveness?* Answering this question requires real data set and real systems. There are many studies in the literature that observe a certain trade-off between effectiveness and fairness. But is this true for every fairness criteria? We exploit the generality and completeness of our framework to check this.
- *RQ3. Is there a trade-off between fairness metrics?* There is work showing that there are incompatibilities between fairness metrics. That is, different fairness criteria cannot be satisfied simultaneously. Regardless of the fact that not all metrics can be maximized simultaneously, we will study on real data and systems whether improving one fairness criterion necessarily implies worsening others.
- *RQ4. Are the fairness metrics consistent across data sets?* Our framework provides 18 fairness metric instances. We hypothesize that some fairness criteria are more sensitive to particular data sets than others. We run them on two different real data sets over the same systems to check this.

5.1 Synthetic Recommendation Outputs

In the following experiment, we apply the fairness metrics derived from our framework to synthetic data and RS outputs. The aim is to answer RQ1. As a starting point we generate an oracle system output, in which the items are sorted by utility for each user, and a random system output, in which the items are sorted randomly for each user. The methodology consists of modifying artificially the oracle output or the random baseline to improve particular fairness features. Then the metric results should be consistent.

5.1.1 Data and Settings. Our synthetic data consists of 100 users and 100 items, both divided into three groups (1–10, 11–30, 31–100). The utility function is: $\psi(u,i) = \text{Max}\left({}^{1}/\sqrt{i\cdot u}, {}^{1}/\sqrt{(101-i)\cdot(101-u)}\right)$. Figure 3 illustrates the user/item utility distribution across groups. The resulting distribution is such that items 1 and 100 are more popular than the rest; groups are unbalanced (10, 20, and 70 items or users); user group A is biased toward item group I and user group C is biased toward item group III.

Table 3 displays the name, description, and hypothesized behavior of synthetic baseline systems. Each synthetic system output consists of a ranking of 100 items per user, ordered according to a certain priority function $\texttt{Pri}(u,i)$. The `Oracle` system output sorts items by utility ($\texttt{Pri}(u,i) = P_\psi(u,i)$), while the `Random` baseline sort items randomly. The rest of systems modify these baselines by multiplying them with a certain *fairness factor*.

[3]Source code for running these experiments can be found in the following GitHub repository: FairnessFramework4RecSys at abellogin.

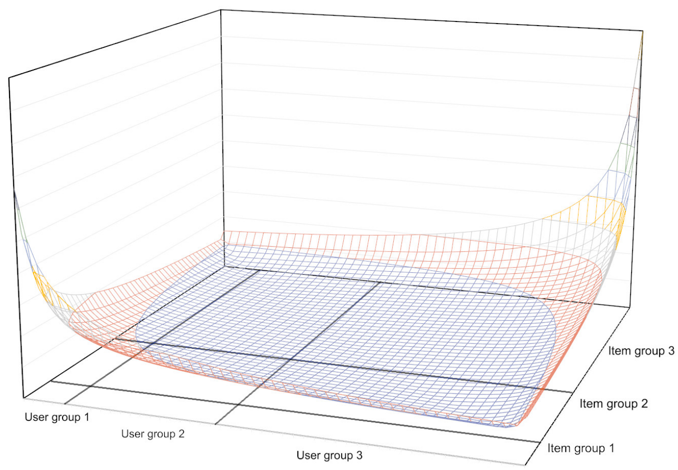

Fig. 3. Utility distribution across user and item groups in the synthetic data set.

5.1.2 Results. We consider the DCG decay exposure function in both the effectiveness and fairness measurements. That is, being $\texttt{Rank}(u, i)$ the ranking position of the item i in the user u interface according to $\texttt{Pri}(u, i)$, then $\phi(u, i) = 1/(\log(\texttt{Rank}(u,i))+1)$. Table 4 shows the fairness measurement results in all metric variants presented in Table 2. We do not consider the Exposure benefit function in user groups, given that all users receive the same amount of information in our experiment.

Numbers with colored background indicate those values that corroborate the hypotheses described in Table 3 for each of the synthetic system outputs. `Oracle` maximizes effectiveness at the cost of item group exposure inequities (KLD-$A_{\mathcal{I}}$-P-ϕ=0.198 and KLD-$A_{\mathcal{I}}$-S-ϕ=0.198) and stresses the user/item dependencies (MI fairness metrics). On the contrary, `Random` achieves the lowest effectiveness (DCG=0.221), but provides group size and utility proportional equity (KLD-$A_{\mathcal{I}}$-S-ϕ=KLD-$A_{\mathcal{I}}$-U-`Eff`=0) and user/item independence. `Popularity` provides high effectiveness and exposure user/item independence (MI-$A_{\mathcal{I}}$-$A_{\mathcal{U}}$-ϕ=MI-$A_{\mathcal{I}}$-$\boldsymbol{\mathcal{U}}$-ϕ=MI-$\mathcal{I}$-$A_{\mathcal{U}}$-ϕ=0), since all users receive the same recommendation. The cost is a higher item group exposure inequity (KLD-$A_{\mathcal{I}}$-P-ϕ, KLD-$A_{\mathcal{I}}$-P-`Eff`, KLD-$A_{\mathcal{I}}$-S-ϕ and KLD-$A_{\mathcal{I}}$-S-`Eff`). The unfairness effect of `Oracle` can be smoothed by a randomization fairness factor (`Randomized Oracle`), but at the cost of a decreased efficiency (DCG=0.275). We can also favor the uniform distribution of exposure across groups (KLD-$A_{\mathcal{I}}$-P-ϕ and KLD-$A_{\mathcal{I}}$-P-`Eff`) by dividing the utility of the items by the size of their group (`Item Group Size Normalized Oracle`). If we add the item group utility density as fairness factor (`Item Group Exposure Calibrated Oracle` and `Item Group Exposure Calibrated Random`) then we can improve the item group utility proportional equity (KLD-$A_{\mathcal{I}}$-U-ϕ=0). Finally, we can improve the exposure and effectiveness independence between user and item groups or individuals (MI based metrics) by adding as fairness factor the utility mass of user and/or item groups (`Group Debiased Oracle`, `Item Group Single User Debiased Oracle`, and `User Group Single Item Debiased Oracle`). In addition, we repeated the experiment but exposing 10 items per user (flat exposure). That is $\phi(u, i)$ is 1 if $Rank(Pri(u, i)) \leq 10$ and 0 otherwise. We obtained similar results.

Table 3. Name, description, and hypothesized behavior of synthetic baseline systems. $\texttt{Pri}(u, i)$ represents the priority of item i for user u which determines the item ranking position.

Baseline	$\texttt{Pri}(u, i)$	**Description**	**Hypothesized behavior**
`Oracle`	$P_\psi(u, i)$	Items are sorted according to user utility.	High effectiveness. Item group exposure inequities. User/item dependencies.
`Random`	$\text{Rand}()$	Items are sorted randomly.	Low effectiveness. Group size and utility proportional equity. User/item independence.
`Popularity`	$P_\psi(i)$	Items are exposed according to their popularity in ψ.	High effectiveness. Item group exposure inequity. Exposure independence.
`Randomized Oracle`	$P_\psi(i)$	Adding a random factor to `Oracle`.	Less effective than `Oracle`, but more item group exposure proportionality and independence.
`Item Group Size Norm. Oracle`	$\frac{P_\psi(i)}{\lvert g(i)\rvert}$	Items from small groups are prioritized. Item priority is its utility divided by the group size.	Less effective than `Oracle`, but more item group parity (uniform distribution).
`Item Group Exposure Cal. Oracle`	$P_\psi(u, i) \cdot \frac{P_\psi(g(i))}{\lvert g(i)\rvert}$	The item priority is its utility multiplied by the item group utility density.	Lower effectiveness than the `Oracle` but higher utility proportional item group equity.
`Item Group Exposure Cal. Random`	$P_\psi(u, i) \cdot \frac{P_\psi(g(i))}{\lvert g(i)\rvert}$	The item priority is a random value multiplied by the item group utility density.	Higher effectiveness than `Random` and higher utility proportional item group equity.
`Group Debiased Oracle`	$\frac{P_\psi(u,i)}{P_\psi(g(u),g(i))}$	It reduces the user/item group bias by dividing the `Oracle` utility by the user and item group utility mass.	Lower effectiveness than the `Oracle` but higher exposure and effectiveness independence between user and item groups.
`Item Group Single User Deb. Oracle`	$\frac{P_\psi(u,i)}{P_\psi(\mathcal{U},g(i))}$	It reduces the item group vs single user bias by dividing the `Oracle` utility by the item group utility mass.	Lower effectiveness than the `Oracle` but higher exposure and effectiveness independence between single users and item groups.
`User Group Single Item Deb. Oracle`	$\frac{P_\psi(u,i)}{P_\psi(g(u),\mathcal{I})}$	It reduces the user group bias across single items by dividing the `Oracle` utility by the user group utility mass.	Lower effectiveness than the `Oracle` but higher exposure and effectiveness independence between user groups and single items.

In conclusion, according to our experiments with synthetic data, the answer to RQ1 is positive: the metrics instantiated from the general framework capture different system output features and they are consistent with the hypothesized behavior of synthetic system outputs.

5.2 Behavior of State-of-the-art RSs

In this second experiment, we analyze the behavior of the proposed framework on a real recommendation data set and system outputs in order to answer RQ2, RQ3, and RQ4.

5.2.1 Data sets. This study evaluates the performance of Collaborative Filtering (CF) approaches within the presented fairness evaluation framework using two popular data sets including explicit or implicit preferences:

- **Netflix** (explicit). The original version of this data set is one of the largest available benchmark data sets used widely for CF algorithms today [6]. It has ratings collected over the course of

Table 4. Fairness metrics for synthetic outputs. DCG based exposure. We use the same notation as before: P, S, and U are the fairness criteria Parity, Size Proportionality, and Utility Proportionality, respectively; ϕ and `Eff` are Exposure and Effectiveness. Colored background denotes values that match hypotheses presented in Table 3, whereas italics are used to highlight best values for each metric (column), where except for `Eff`, for which higher is preferable, the lowest value indicates the best performance.

		Inequity (KLD)									Dependency (MI)					
User/Item groups:	Eff	User Groups $A_\mathcal{U}$			Item Groups $A_\mathcal{I}$						User Groups Item Groups $A_\mathcal{I}$-$A_\mathcal{U}$		Item Groups Single Users $A_\mathcal{I}$-$\mathcal{U}$		User Groups Single Items $\mathcal{I}$-$A_\mathcal{U}$	
Fairness Criterion:		P	S	U	P		S		U				Independence			
Benefit Function:	DCG	Eff	Eff	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff
Oracle	*0.292*	0.172	0.048	*0.000*	0.198	0.152	0.026	0.217	0.002	0.058	0.014	0.152	0.019	0.197	0.020	0.186
Random	0.221	0.171	0.051	*0.000*	0.283	0.166	*0.000*	*0.053*	0.036	*0.000*	*0.000*	*0.050*	0.004	0.072	0.004	0.071
Popularity	0.275	0.194	*0.036*	*0.000*	0.202	0.173	0.026	0.178	0.002	0.042	*0.000*	0.062	*0.000*	0.082	*0.000*	0.076
Randomized Oracle	0.273	0.178	0.044	*0.000*	0.215	0.151	0.016	0.169	0.006	0.036	0.008	0.119	0.012	0.163	0.012	0.152
Item Group Size Normalized Oracle	0.273	0.137	0.060	0.002	0.088	0.066	0.085	0.370	0.016	0.140	0.001	0.072	0.001	0.108	0.003	0.097
Item Group Exposure Calibrated Oracle	0.288	0.166	0.050	*0.000*	0.153	0.124	0.055	0.289	*0.000*	0.093	0.007	0.116	0.011	0.172	0.011	0.148
Item Group Exposure Calibrated Random	0.235	0.143	0.057	0.001	0.162	0.094	0.047	0.244	*0.000*	0.069	*0.000*	0.048	0.001	*0.069*	0.001	*0.061*
Group Debiased Oracle	0.279	0.151	0.055	*0.000*	0.095	0.074	0.074	0.330	0.014	0.117	0.001	0.087	0.004	0.139	0.005	0.119
Item Group Single User Debiased Oracle	0.265	*0.121*	0.066	0.005	*0.078*	*0.052*	0.090	0.399	0.021	0.160	*0.000*	0.055	*0.000*	0.078	0.001	0.073
User Group Single Item Debiased Oracle	0.281	0.188	0.039	*0.000*	0.186	0.156	0.034	0.221	*0.000*	0.060	*0.000*	0.064	0.009	0.134	*0.000*	0.076

seven years. We used a “small” variant of this data set with 9,992 users, 4,945 items, 607,803 ratings.

- **CiteULike-a** (implicit). The CiteULike data set[4] is about academic citations. CiteULike is an online platform that enables registered users to establish personal libraries by archiving relevant articles. The data set consists of the papers in the users' libraries (which are handled as "likes"), the tags provided by the users, as well as the title and abstract of the papers. CiteULike-a [76] data set contains 4,122 users, 16,908 items, and 155,588 interactions.

5.2.2 Systems. We investigated a variety of latent factors CF models, which have been employed in previous and ongoing works of RS research to achieve excellent performance in rating and ranking tasks [17, 20, 41, 58].

- `MF` [41]: A classical Matrix Factorization (MF) approach; in this case, the user and item factor are learned through Stochastic Gradient Descent, despite the availability of other techniques [38]. The predicted rating in MF is computed as $\hat{r}_{ui} = \mathbf{q}_i^T \mathbf{p}_u$, where $\mathbf{p}_u \in \mathbb{R}^H$ and $\mathbf{q}_i \in \mathbb{R}^H$ are the learned H-sized latent vectors for the user u and item i, respectively.
- `PMF` [66]: A Maximum A Posteriori approach is used to factorize the matrix in light of a probabilistic linear model containing Gaussian noise.
- `BPR-MF` [41, 63]: BPR is the state-of-the-art method for personalized ranking, especially on data sets containing implicit feedback. MF is used as the predictor in BPR-MF. It is important to note that this algorithm tends to recommend popular items more often than other methods [5].
- `WMF` [39, 59]: Classic weighted MF model for implicit feedback data. It assumes the independence of the latent features of two items and gives lower weights to negative samples. The equivalent ALS-based approach [39] can reduce inference complexity.

[4] http://www.citeulike.org/

Table 5. Fairness metrics for CiteULike data set. DCG based exposure. Same notation as in Table 4.

		Inequity (KLD)										Dependency (MI)					
User/Item groups:	Eff	User Groups $A_{\mathcal{U}}$			Item Groups $A_{\mathcal{I}}$						User Groups Item Groups $A_{\mathcal{I}}$-$A_{\mathcal{U}}$		Item Groups Single Users $A_{\mathcal{I}}$-$\mathcal{U}$		User Groups Single Items $\mathcal{I}$-$A_{\mathcal{U}}$	
Fairness Criterion:		P	S	U	P		S		U		Independence					
Benefit Function:	DCG	Eff	Eff	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff
Oracle	*3.362*	0.120	0.127	0.039	0.275	*0.275*	0.141	*0.141*	*0.000*	*0.000*	0.002	0.002	0.221	0.221	0.406	0.406
Random	0.002	0.210	0.219	0.005	0.023	0.357	*0.000*	0.204	0.171	0.005	*0.000*	0.131	0.091	0.642	0.305	0.789
Popularity	0.038	0.227	0.237	*0.003*	1.000	1.000	0.760	0.760	0.316	0.316	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.032*
MF	0.002	0.118	0.125	0.040	*0.001*	0.409	0.036	0.244	0.376	0.014	*0.000*	*0.000*	0.091	0.590	0.042	0.881
PMF	0.030	*0.064*	*0.069*	0.089	0.975	0.977	0.736	0.739	0.296	0.298	*0.000*	0.003	0.019	0.022	0.031	0.423
BPR-MF	0.038	0.227	0.237	*0.003*	1.000	1.000	0.760	0.760	0.316	0.316	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.032*
WMF	0.121	0.186	0.195	0.011	0.501	0.707	0.319	0.494	0.039	0.127	0.009	0.014	0.152	0.237	0.246	0.646
NeuMF	0.166	0.108	0.115	0.047	0.786	0.884	0.564	0.653	0.171	0.232	0.002	0.002	0.125	0.108	0.115	0.370
VAECF	0.167	0.102	0.109	0.051	0.864	0.916	0.634	0.682	0.219	0.254	*0.000*	0.006	0.087	0.078	0.067	0.340

- NeuMF [37]: Using multi-layer perceptron and MF, this approach learns users and item features, and then uses non-linear activation functions to train a mapping between these features.
- VAECF [50]: The method relies on variational autoencoders, which present a multinomial likelihood generative model and employ Bayesian inference for parameter estimation.

We consider the same baseline approaches as in the previous experiment (Oracle, Random, and Popularity). Note that Random has some dependency between user groups and items in effectiveness due to the original bias of the data. It also has dependencies between groups of items and individual users due to the random effect. That is, not all individual users have a uniform distribution of item groups and vice versa. The effectiveness and fairness metrics are exactly the same as in the previous experiment (Section 5.1).

5.2.3 Evaluation setup. For each considered recommendation model, we ran them at their default hyper-parameter values according to their implementation in the Cornac recommender framework [65]. The results of the recommendation were generated based on a hold-out setting (80%-20% training-test split).

5.2.4 Results. The results are shown in Tables 5 (for CiteULike) and 6 (for Netflix), commented in the following with particular emphasis on the values highlighted in color. The answers to our three research questions RQ2–RQ4 can be synthesized as follows.

- *RQ2. Is there a trade-off between fairness and effectiveness?* The answer for this question is that it depends on the fairness metric. For instance, the highest DCG (3.362 and 17.212 respectively in each data set) is achieved by Oracle with a perfect fairness (zero KLD) in terms of item group utility-proportional exposure equity (KLD-$A_{\mathcal{I}}$-U-ϕ). In both data sets, the neural based systems (VAECF and NeuMF) achieve higher DCG values (0.166, 0.167, 1.454, and 1.605) than MF-based systems and also are more fair in terms of KLD-$A_{\mathcal{I}}$-U-ϕ (0.047, 0.051, 0.275, and 0.189). On the contrary, it seems that there exists a trade-off between effectiveness and size-proportional item group exposure inequity (KLD-$A_{\mathcal{I}}$-S-ϕ). As an intriguing observation, we could connect

Table 6. Fairness metrics for Netflix data set. DCG based exposure. Same notation as in Table 4.

		Inequity (KLD)									Dependency (MI)					
User/Item groups:	Eff	User Groups $A_{\mathcal{U}}$			Item Groups $A_{\mathcal{I}}$						User Groups Item Groups $A_{\mathcal{I}}$-$A_{\mathcal{U}}$		Item Groups Single Users $A_{\mathcal{I}}$-$\mathcal{U}$		User Groups Single Items $\mathcal{I}$-$A_{\mathcal{U}}$	
Fairness Criterion:		P	S	U	P		S		U		Independence					
Benefit Function:	DCG	Eff	Eff	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff	ϕ	Eff
Oracle	*17.212*	0.082	0.306	0.062	0.328	0.345	1.816	1.853	*0.000*	0.001	0.002	0.002	0.174	0.181	0.101	0.101
Random	0.038	0.307	0.666	*0.001*	0.450	*0.313*	*0.000*	*1.782*	1.593	*0.000*	*0.000*	0.019	0.094	0.642	0.051	0.584
Popularity	1.296	0.026	0.185	0.145	1.000	1.000	2.977	2.977	0.300	0.300	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.017*
MF	0.366	0.344	0.718	0.006	0.108	0.757	0.162	2.623	0.867	0.142	0.020	*0.000*	0.161	0.221	0.067	0.066
PMF	0.361	0.172	0.464	0.011	*0.012*	0.564	0.414	2.291	0.508	0.054	0.011	0.001	0.222	0.402	0.046	0.135
BPR-MF	1.320	0.043	0.226	0.110	1.000	1.000	2.977	2.977	0.300	0.300	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	0.020
WMF	1.043	0.034	0.206	0.126	0.497	0.818	2.167	2.718	0.032	0.176	*0.000*	0.002	0.214	0.161	0.150	0.274
NeuMF	1.454	*0.000*	*0.085*	0.275	0.868	0.933	2.794	2.888	0.207	0.250	*0.000*	*0.000*	0.094	0.062	0.092	0.156
VAECF	1.605	0.012	0.143	0.189	0.914	0.971	2.861	2.941	0.237	0.278	*0.000*	*0.000*	0.071	0.027	0.115	0.171

this practical and general result to the well-known **accuracy-diversity** or **accuracy-novelty** trade-off phenomenon in the community [73], and now we could observe a similar trade-off of effectiveness-item fairness. The Random system minimizes this inequity in both data sets (zero KLD) at the cost of DCG (0.002 and 0.038). On the other hand, Oracle maximizes effectiveness by increasing the inequity (0.141 and 1.816). The neural based systems (NeuMF and VAECF) are more effective than the others, but highly unfair in terms of item group size-proportional exposure equity (KLD-$A_{\mathcal{I}}$-S-ϕ) in both data sets, (0.564, 0.634, 2.794, and 2.861).

- *RQ3. Is there a trade-off between fairness metrics?* In view of the results, we cannot state that there is a trade-off between fairness metrics. However, we see that different metrics express different characteristics of the systems. For example, regarding the dependence-based fairness metrics (MI), all systems keep the independence between user and item groups (MI-$A_{\mathcal{I}}$-$A_{\mathcal{U}}$-ϕ and MI-$A_{\mathcal{I}}$-$A_{\mathcal{U}}$-Eff are zero or almost zero for all systems). However, WMF seems to state a certain dependence between single users and item groups and between single items and user groups (MI-$A_{\mathcal{I}}$-$\mathcal{U}$-ϕ, MI-$A_{\mathcal{I}}$-$\mathcal{U}$-Eff, MI-$\mathcal{I}$-$A_{\mathcal{U}}$-ϕ, and MI-$\mathcal{I}$-$A_{\mathcal{U}}$-Eff); this suggests a higher personalizing degree in the recommendation. On the other hand, although BPR-MF presents item group inequities (KLD-$A_{\mathcal{I}}$-S-ϕ and KLD-$A_{\mathcal{I}}$-S-Eff are 0.760 and 2.977), it keeps the independence between user/item individuals and groups in both data sets (MI-$A_{\mathcal{I}}$-$\mathcal{U}$-ϕ, MI-$A_{\mathcal{I}}$-$\mathcal{U}$-Eff, MI-$\mathcal{I}$-$A_{\mathcal{U}}$-ϕ, and MI-$\mathcal{I}$-$A_{\mathcal{U}}$-Eff are all close to zero).
- *RQ4. Are the fairness metrics consistent across data sets?* Not every metric is consistent across data sets. For instance, NeuMF is more fair than VAECF in terms of size-proportional user group effectiveness (KLD-$A_{\mathcal{U}}$-S-Eff) in the CiteULike data set, but not in the Netflix data set. In addition, for this user oriented metric, the Popularity baseline is unfair in CiteULike but not in Netflix. This suggests that user group effectiveness fairness is sensitive to the evaluation benchmark. We hypothesize that the nature of systems is more determinant in item group fairness than in user group fairness which is highly related with the distribution of user preferences.

In summary, these results confirm that the different instantiations of fairness metrics in real data sets give us different information about system output bias and, in some cases, this information is sensitive to the particularities of the data set.

6 CONCLUSIONS AND FUTURE WORK

Contributions. In this paper we have defined a formal, broad, and unified framework for measuring fairness in RSs, and validated it experimentally. The proposed framework captures the five dimensions that characterize existing fairness metrics in the literature. The practical implications of this model are essentially: (i) a tool to identify the most appropriate metric in a given scenario, (ii) the unification of fairness evaluation criteria for the comparison of results in different research works, and (iii) the identification of formal aspects of fairness that have not yet been explored, such as the statistical independence of the benefit between user and item attributes. We hope that these contributions will allow a better understanding of fairness measurement and, in perspective, to overcome the limitations imposed by the current fragmented landscape of fairness definitions and metrics.

In general, we expect both researchers and practitioners to benefit from these contributions, especially those concerned about measuring and assessing fairness from novel dimensions. This is because our framework, as defined and demonstrated throughout the paper, is two-sided (it allows capturing the notion of fairness on users and items at the same time without the need of having an *ideal* preconceived notion of fairness), flexible (because it is possible to boil down to many existing notions of fairness), and reliable (as it is focused on *independence* rather than *equity*).

Limitations and Outlook. While we make no claim that the proposed five dimensions for fairness measurement are exhaustive (as we anticipated at the end of Section 2), we believe they can serve as a useful start point for practitioners, students, and scholars. Nonetheless, we briefly outline several other dimensions that could be taken into account. For example, different scales can be defined for the utility of items (binary, rating, preferences, continuous, etc.), and the benefit distribution can be defined in terms of groups or the user past behavior itself (the notion of *calibration* [70]). However, from our point of view both aspects can be encapsulated within the notion of item utility, to which the fairness model should be agnostic. Another dimension that could be taken into account is the possibility of considering degrees of membership of items or users in groups, with a non-binary group membership function. We have not included this dimension as it is very rare in the literature, although we do take it into account in the definition of our theoretical framework.

It should be noted that some notions of fairness are not captured by our five dimensions. For example, *Fairness Through Unawareness* [15, 44, 74] represents to what extent certain attributes are not explicitly used in the training process. However, in this paper we focus on the evaluation of recommender output, regardless of how the system has been trained. In addition, our five dimensions focus on *group fairness* rather than *individual fairness* [62]. However, due to definition of group fairness that incorporate as input protected features, more attention has been paid to group fairness; also, individual fairness requires a certain (arbitrary) similarity function between users or items [15, 26].

Although we believe that our experimental results are representative, in the future we aim to perform a more complete experimental activity, with more RSs and on more data sets. In addition, we remark that our approach seems general to be applied to any kind of information system including IR systems; we plan to do so in future work.

Still about future work, this framework offers us a uniform tool to comprehensively study the theoretical and empirical trade-off between different fairness criteria. Although there is work in the literature in this respect, the lack of a general framework for measuring fairness has not yet allowed a comprehensive analysis of the problem. Being even more ambitious, we intend to exploit this theoretical tool to identify a single measure that, even at the cost of effectiveness, ensures maximum fairness levels in all metrics. One candidate measure could be the multi-variate entropy, but this conjecture requires further study.

REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Augusto Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Model. User Adapt. Interact.* 30, 1 (2020), 127–158. DOI: http://dx.doi.org/10.1007/s11257-019-09256-1

[2] Himan Abdollahpouri and Masoud Mansoury. 2020. Multi-sided Exposure Bias in Recommendation. In *International Workshop on Industrial Recommendation Systems (IRS2020) in Conjunction with ACM KDD 2020.* arXiv:2006.15772 https://arxiv.org/abs/2006.15772

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward C. Malthouse. 2021. User-centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, Judith Masthoff, Eelco Herder, Nava Tintarev, and Marko Tkalcic (Eds.). ACM, 119–129. DOI: http://dx.doi.org/10.1145/3450613.3456821

[4] Fabian Abel, Yashar Deldjoo, Mehdi Elahi, and Daniel Kohlsdorf. 2017. RecSys Challenge 2017: Offline and Online Evaluation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, August 27-31, 2017*, Paolo Cremonesi, Francesco Ricci, Shlomo Berkovsky, and Alexander Tuzhilin (Eds.). ACM, 372–373. DOI: http://dx.doi.org/10.1145/3109859.3109954

[5] Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, Dietmar Jannach, and Claudio Pomo. 2022. Top-N Recommendation Algorithms: A Quest for the State-of-the-Art. In *UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4 - 7, 2022*, Alejandro Bellogín, Ludovico Boratto, Olga C. Santos, Liliana Ardissono, and Bart Knijnenburg (Eds.). ACM, 121–131. DOI: http://dx.doi.org/10.1145/3503252.3531292

[6] James Bennett and Stan Lanning. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. Citeseer, 35.

[7] Richard A. Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50 (2017), 3 – 44.

[8] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.). ACM, 2212–2220. DOI: http://dx.doi.org/10.1145/3292500.3330745

[9] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 405–414. DOI: http://dx.doi.org/10.1145/3209978.3210063

[10] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Model. User Adapt. Interact.* 31, 3 (2021), 421–455. DOI: http://dx.doi.org/10.1007/s11257-021-09294-8

[11] Robin Burke. 2017. Multisided Fairness for Recommendation. In *2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017).*

[12] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, 202–214. http://proceedings.mlr.press/v81/burke18a.html

[13] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.). ACM, 903–912. DOI: http://dx.doi.org/10.1145/2009916.2010037

[14] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic (LIPIcs)*, Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella (Eds.), Vol. 107. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 28:1–28:15. DOI: http://dx.doi.org/10.4230/LIPIcs.ICALP.2018.28

[15] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and Debias in Recommender System: A Survey and Future Directions. *CoRR* abs/2010.03240 (2020). arXiv:2010.03240 https://arxiv.org/abs/2010.03240

[16] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. DOI: http://dx.doi.org/10.1089/big.2016.0047

[17] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010*, Xavier Amatriain, Marc Torrens, Paul Resnick, and Markus Zanker (Eds.). ACM, 39–46. DOI: http://dx.doi.org/10.1145/1864708.1864721

[18] Diego Corrêa da Silva, Marcelo Garcia Manzato, and Frederico Araújo Durão. 2021. Exploiting personalized calibration and metrics for fairness recommendation. *Expert Syst. Appl.* 181 (2021), 115112. DOI: http://dx.doi.org/10.1016/j.eswa.2021.115112

[19] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, and Tommaso Di Noia. 2021. A flexible framework for evaluating user and item fairness in recommender systems. *User Model. User Adapt. Interact.* 31, 3 (2021), 457–511. DOI: http://dx.doi.org/10.1007/s11257-020-09285-1

[20] Yashar Deldjoo, Alejandro Bellogín, and Tommaso Di Noia. 2021. Explaining recommender systems fairness and accuracy through the lens of data characteristics. *Inf. Process. Manag.* 58, 5 (2021), 102662. DOI: http://dx.doi.org/10.1016/j.ipm.2021.102662

[21] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 275–284. DOI: http://dx.doi.org/10.1145/3340531.3411962

[22] Qiang Dong, Shuang-Shuang Xie, and Wen-Jun Li. 2021. User-Item Matching for Recommendation Fairness. *IEEE Access* 9 (2021), 130389–130398. DOI: http://dx.doi.org/10.1109/ACCESS.2021.3113975

[23] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, Shafi Goldwasser (Ed.). ACM, 214–226. DOI: http://dx.doi.org/10.1145/2090236.2090255

[24] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Found. Trends Inf. Retr.* 16, 1-2 (2022), 1–177. DOI: http://dx.doi.org/10.1561/1500000079

[25] Michael D. Ekstrand and Vaibhav Mahant. 2017. Sturgeon and the Cool Kids: Problems with Random Decoys for Top-N Recommender Evaluation. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017*, Vasile Rus and Zdravko Markov (Eds.). AAAI Press, 639–644. https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15534

[26] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, 172–186. http://proceedings.mlr.press/v81/ekstrand18b.html

[27] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards Long-term Fairness in Recommendation. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 445–453. DOI: http://dx.doi.org/10.1145/3437963.3441824

[28] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane J. Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward Pareto Efficient Fairness-Utility Trade-off in Recommendation through Reinforcement Learning. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (2022).

[29] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.). ACM, 2221–2231. DOI: http://dx.doi.org/10.1145/3292500.3330691

[30] Alireza Gharahighehi, Celine Vens, and Konstantinos Pliakos. 2021. Fair multi-stakeholder news recommender system with hypergraph ranking. *Inf. Process. Manag.* 58, 5 (2021), 102663. DOI: http://dx.doi.org/10.1016/j.ipm.2021.102663

[31] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When Fair Ranking Meets Uncertain Inference. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1033–1043. DOI: http://dx.doi.org/10.1145/3404835.3462850

[32] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. 2022. Provider fairness across continents in collaborative recommender systems. *Inf. Process. Manag.* 59, 1 (2022), 102719. DOI: http://dx.doi.org/10.1016/j.ipm.2021.102719

[33] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Mirko Marras. 2021. The Winner Takes it All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1808–1812. DOI: http://dx.doi.org/10.1145/3404835.3463235

[34] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 265–308. DOI: http://dx.doi.org/10.1007/978-1-4899-7637-6_8

[35] Ananya Gupta, Eric Johnson, Justin Payan, Aditya Kumar Roy, Ari Kobren, Swetasudha Panda, Jean-Baptiste Tristan, and Michael L. Wick. 2021. Online Post-Processing in Rankings for Fair Utility Maximization. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 454–462. DOI: http://dx.doi.org/10.1145/3437963.3441724

[36] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3315–3323. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

[37] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 173–182. DOI: http://dx.doi.org/10.1145/3038912.3052569

[38] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*. IEEE Computer Society, 263–272.

[39] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*. IEEE Computer Society, 263–272. DOI: http://dx.doi.org/10.1109/ICDM.2008.22

[40] Ömer Kirnap, Fernando Diaz, Asia Biega, Michael D. Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of Fair Ranking Metrics with Incomplete Judgments. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 1065–1075. DOI: http://dx.doi.org/10.1145/3442381.3450080

[41] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. DOI: http://dx.doi.org/10.1109/MC.2009.263

[42] Marina Krakovsky. 2022. Formalizing fairness. *Commun. ACM* 65, 8 (2022), 11–13. DOI: http://dx.doi.org/10.1145/3542815

[43] Caitlin Kuhlman, MaryAnn Van Valkenburg, and Elke A. Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking using Pairwise Error Metrics. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 2936–2942. DOI: http://dx.doi.org/10.1145/3308558.3313443

[44] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4066–4076. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[45] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 624–632. DOI: http://dx.doi.org/10.1145/3442381.3449866

[46] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2022. Fairness in Recommendation: A Survey. *CoRR* abs/2205.13619 (2022). DOI: http://dx.doi.org/10.48550/arXiv.2205.13619 arXiv:2205.13619

[47] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards Personalized Fairness based on Causal Notion. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1054–1063. DOI: http://dx.doi.org/10.1145/3404835.3462966

[48] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. 2021. Tutorial on Fairness of Machine Learning in Recommender Systems. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2654–2657. DOI: http://dx.doi.org/10.1145/3404835.3462814

[49] Yangkun Li, Mohamed-Laid Hedia, Weizhi Ma, Hongyu Lu, Min Zhang, Yiqun Liu, and Shaoping Ma. 2022. Contextualized Fairness for Recommender Systems in Premium Scenarios. *Big Data Research* 27 (2022), 100300. DOI: http://dx.doi.org/https://doi.org/10.1016/j.bdr.2021.100300

[50] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 689–698. DOI: http://dx.doi.org/10.1145/3178876.3186150

[51] Chen Lin, Xinyi Liu, Guipeng Xv, and Hui Li. 2021. Mitigating Sentiment Bias for Recommender Systems. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 31–40. DOI: http://dx.doi.org/10.1145/3404835.3462943

[52] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 467–471. DOI: http://dx.doi.org/10.1145/3298689.3347016

[53] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (2021), 115:1–115:35. DOI: http://dx.doi.org/10.1145/3457607

[54] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna M. Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 626–633. DOI: http://dx.doi.org/10.1145/3041021.3054197

[55] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.).

ACM, 2243–2251. DOI: http://dx.doi.org/10.1145/3269206.3272027

[56] Natwar Modani, Deepali Jain, Ujjawal Soni, Gaurav Kumar Gupta, and Palak Agarwal. 2017. Fairness Aware Recommendations on Behance. In *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II (Lecture Notes in Computer Science)*, Jinho Kim, Kyuseok Shim, Longbing Cao, Jae-Gil Lee, Xuemin Lin, and Yang-Sae Moon (Eds.), Vol. 10235. 144–155. DOI: http://dx.doi.org/10.1007/978-3-319-57529-2_12

[57] Hervé Moulin. 2003. *Fair division and collective welfare.* MIT Press.

[58] Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Yashar Deldjoo. 2022. CPFair: Personalized Consumer and Producer Fairness Re-ranking for Recommender Systems. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022.* ACM, 770–779. DOI: http://dx.doi.org/10.1145/3477495.3531959

[59] Rong Pan, Yunhong Zhou, Bin Cao, Nathan Nan Liu, Rajan M. Lukose, Martin Scholz, and Qiang Yang. 2008. One-Class Collaborative Filtering. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy.* IEEE Computer Society, 502–511.

[60] Gourab K. Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. 2020. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 1194–1204. DOI: http://dx.doi.org/10.1145/3366423.3380196

[61] Gourab K. Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair ranking: a critical review, challenges, and future directions. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022.* ACM, 1929–1942. DOI: http://dx.doi.org/10.1145/3531146.3533238

[62] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2022. Fairness in rankings and recommendations: an overview. *VLDB J.* 31, 3 (2022), 431–458. DOI: http://dx.doi.org/10.1007/s00778-021-00697-y

[63] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, Jeff A. Bilmes and Andrew Y. Ng (Eds.). AUAI Press, 452–461.

[64] Dimitris Sacharidis, Kyriakos Mouratidis, and Dimitrios Kleftogiannis. 2019. A Common Approach for Consumer and Provider Fairness in Recommendations. In *Proceedings of ACM RecSys 2019 Late-Breaking Results co-located with the 13th ACM Conference on Recommender Systems, RecSys 2019 Late-Breaking Results, Copenhagen, Denmark, September 16-20, 2019 (CEUR Workshop Proceedings)*, Marko Tkalcic and Sole Pera (Eds.), Vol. 2431. CEUR-WS.org, 1–5. http://ceur-ws.org/Vol-2431/paper1.pdf

[65] Aghiles Salah, Quoc-Tuan Truong, and Hady W. Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *Journal of Machine Learning Research* 21, 95 (2020), 1–5. http://jmlr.org/papers/v21/19-805.html

[66] Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic Matrix Factorization. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis (Eds.). Curran Associates, Inc., 1257–1264.

[67] Tefko Saracevic. 1975. RELEVANCE: A review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.* 26, 6 (1975), 321–343. DOI: http://dx.doi.org/10.1002/asi.4630260604

[68] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. 2016. The Problem of Infra-Marginality in Outcome Tests for Discrimination. *Econometrics: Econometric & Statistical Methods - General eJournal* (2016).

[69] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Yike Guo and Faisal Farooq (Eds.). ACM, 2219–2228. DOI: http://dx.doi.org/10.1145/3219819.3220088

[70] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 154–162. DOI: http://dx.doi.org/10.1145/3240323.3240372

[71] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. *WWW '18: Proceedings of the 2018 World Wide Web Conference* (04 2018), 923–932. DOI: http://dx.doi.org/10.1145/3178876.3186140

[72] Virginia Tsintzou, Evaggelia Pitoura, and Panayiotis Tsaparas. 2019. Bias Disparity in Recommendation Systems. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, September 20, 2019 (CEUR Workshop Proceedings)*, Robin Burke, Himan Abdollahpouri, Edward C. Malthouse, K. P. Thai, and Yongfeng Zhang (Eds.), Vol. 2440. CEUR-WS.org. http://ceur-ws.org/Vol-2440/short4.pdf

[73] Saul Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius (Eds.). ACM, 109–116. https://dl.acm.org/citation.cfm?id=2043955

[74] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, Yuriy Brun, Brittany Johnson, and Alexandra Meliou (Eds.). ACM, 1–7. DOI: http://dx.doi.org/10.1145/3194770.3194776

[75] Mengting Wan, Jianmo Ni, Rishabh Misra, and Julian J. McAuley. 2020. Addressing Marketing Bias in Product Recommendations. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 618–626. DOI: http://dx.doi.org/10.1145/3336191.3371855

[76] Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, Chid Apté, Joydeep Ghosh, and Padhraic Smyth (Eds.). ACM, 448–456. DOI: http://dx.doi.org/10.1145/2020408.2020480

[77] Lequn Wang and Thorsten Joachims. 2021. User Fairness, Item Fairness, and Diversity for Rankings in Two-Sided Markets. In *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, Faegheh Hasibi, Yi Fang, and Akiko Aizawa (Eds.). ACM, 23–41. DOI: http://dx.doi.org/10.1145/3471158.3472260

[78] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2022. A Survey on the Fairness of Recommender Systems. *ACM Transactions on Information Systems* abs/2206.03761 (2022).

[79] Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. 2022. Joint Multisided Exposure Fairness for Recommendation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 703–714. DOI: http://dx.doi.org/10.1145/3477495.3532007

[80] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. 2021. TFROM: A Two-sided Fairness-Aware Recommendation Model for Both Customers and Providers. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1013–1022. DOI: http://dx.doi.org/10.1145/3404835.3462882

[81] Bruna D. Wundervald. 2021. Cluster-based quotas for fairness improvements in music recommendation systems. *Int. J. Multim. Inf. Retr.* 10, 1 (2021), 25–32. DOI: http://dx.doi.org/10.1007/s13735-020-00203-0

[82] Emre Yalcin and Alper Bilge. 2021. Investigating and counteracting popularity bias in group recommendations. *Inf. Process. Manag.* 58, 5 (2021), 102608. DOI: http://dx.doi.org/10.1016/j.ipm.2021.102608

[83] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*. ACM, 22:1–22:6. DOI: http://dx.doi.org/10.1145/3085504.3085526

[84] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 2921–2930. https://proceedings.neurips.cc/paper/2017/hash/e6384711491713d29bc63fc5eeb5ba4f-Abstract.html

[85] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 1171–1180. DOI: http://dx.doi.org/10.1145/3038912.

3052660

[86] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li (Eds.). ACM, 1569–1578. DOI: http://dx.doi.org/10.1145/3132847.3132938

[87] Qiliang Zhu, Qibo Sun, Zengxiang Li, and Shangguang Wang. 2020. FARM: A Fairness-Aware Recommendation Method for High Visibility and Low Visibility Mobile APPs. *IEEE Access* 8 (2020), 122747–122756. DOI: http://dx.doi.org/10.1109/ACCESS.2020.3007617

[88] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-Aware Tensor-Based Recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 1153–1162. DOI: http://dx.doi.org/10.1145/3269206.3271795

[89] Indre Zliobaite. 2015. On the relation between accuracy and fairness in binary classification. In *2nd workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[90] Indre Zliobaite. 2017. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.* 31, 4 (2017), 1060–1089. DOI: http://dx.doi.org/10.1007/s10618-017-0506-1