Recommender Systems Fairness Evaluation via Generalized Cross Entropy*

Yashar Deldjoo
Polytechnic University of Bari, Italy
yashar.deldjoo@poliba.it

Vito Walter Anelli
Polytechnic University of Bari, Italy
vitowalter.anelli@poliba.it

Hamed Zamani
University of Massachusetts
Amherst, USA
zamani@umass.edu

Alejandro Bellogín
Autonomous
University of Madrid, Spain
alejandro.bellogin@uam.es

Tommaso Di Noia
Polytechnic University of Bari, Italy
tommaso.dinoia@poliba.it

ABSTRACT

Fairness in recommender systems has been considered with respect to sensitive attributes of users (e.g., gender, race) or items (e.g., revenue in a multistakeholder setting). Regardless, the concept has been commonly interpreted as some form of *equality* – i.e., the degree to which the system is meeting the information needs of all its users *in an equal sense*. In this paper, we argue that fairness in recommender systems does not necessarily imply equality, but instead it should consider a distribution of resources based on merits and needs. We present a probabilistic framework based on generalized cross entropy to evaluate fairness of recommender systems under this perspective, where we show that the proposed framework is flexible and explanatory by allowing to incorporate domain knowledge (through an ideal fair distribution) that can help to understand which item or user aspects a recommendation algorithm is over- or under-representing. Results on two real-world datasets show the merits of the proposed evaluation framework both in terms of user and item fairness.

KEYWORDS

recommender systems; fairness; metric; Generalized cross entropy, evaluation

1 INTRODUCTION AND CONTEXT

Recommender systems (RS) are widely applied across the modern Internet, in e-commerce websites, movies and music streaming platforms, or on social media to point users to items (products or services) [16]. For evaluation of RS, accuracy metrics are typically employed, which measure how much the presented items will be of interest to the target user. One commonly raised concern is how much the recommendations produced by RS are fair. For example, do users of certain gender or race receive fair utility (i.e., benefit) from the recommendation service? To answer this question, one has to recognize the multiple stakeholders involved in such systems and that fairness issues can be studied for more than one group of participants [8]. In a job recommendation scenario, for instance, these multiple groups can be the job seekers and prospective employers where fairness toward both parties has to be recognized. Moreover, fairness in RS can be measured towards items or users; in this context, user and item fairness are commonly associated with an equal chance for appearing in the recommendation results (items) or receiving results of the same quality (users). As an example for the latter, an unfair system may discriminate against users of a particular race or gender.

One common characteristic of the previous literature focusing on RS fairness evaluation is that fairness has been commonly interpreted as some form of *equality* across multiple groups (e.g., gender, race). For example, Ekstrand et al. [16] studied whether RS produce *equal utility* for users of different demographic groups. The authors find demographic differences in measured effectiveness across two datasets from different domains. Yao and Huang [28] studied various types of unfairness that can occur in collaborative filtering models where, to produce fair recommendations, the authors proposed to penalize algorithms producing disparate distributions of prediction error. For additional resources see [8, 9, 17, 30, 31].

Nonetheless, although less common, there are a few works where fairness has been defined beyond uniformity. For instance, in [5], the authors proposed an approach focused on mining the relation between *relevance* and *attention* in Information Retrieval by exploiting the positional bias of search results. That work promotes the notion that ranked subjects should receive attention that is proportional to their *worthiness* in a given search scenario and achieve fairness of attention by making exposure proportional to relevance. Similarly, a framework formulation of fairness constraints is presented in [26] on rankings in terms of exposure allocation, both with respect to group fairness constraints and individuals. Another approach where non-uniform fairness has been used is the work proposed in [29], where the authors aim to solve the top-k ranking problem by optimizing a fair utility function under two conditions: in-group monotonicity (i.e., rank more relevant items above less relevant within the group) and group fairness (proportion of protected group items in the top-k ranking should be above a minimum threshold). In summary, even though these approaches use some notion of non-uniformity, they are applied under different perspectives and purposes.

In the present work, we argue that fairness does not necessarily imply equality between groups, but instead proper distribution of utility (benefits) based on merits and needs. To this end, we present a probabilistic framework for evaluating RS fairness based on attributes of any nature (e.g., sensitive or insensitive) for both items or users and show that the proposed framework is flexible enough to measure fairness in RS by considering fairness as equality or non-equality among groups, as specified by the system designer or any other parties involved in multistakeholder setting. As we shall see later, the discussed approaches are different from our proposal in that we are able to accommodate different notions of fairness, not only *ranking*, e.g., *rating*, *ranking* and *even-beyond accuracy metrics*. In fact, the main advantage of our framework is to provide the system designer with a high degree of flexibility on defining fairness from multiple viewpoints. Results on two real-world datasets show the merits of the proposed evaluation framework, both in terms of user and item fairness.

2 EVALUATING FAIRNESS IN RS

In this section, we propose a framework based on generalized cross entropy for evaluating fairness in recommender systems. Let $\mathcal{U}$ and $\mathcal{I}$ denote a set of users and items, respectively. Suppose $\mathcal{A}$ be a set of sensitive attributes in which fairness is desired. Each attribute can

Presented at the RMSE workshop held in conjunction with the 13th ACM Conference on Recommender Systems (RecSys), 2019, in Copenhagen, Denmark.

Table 1: A set of 6 users belonging to groups g_1 and g_2 and 10 items along with their true labels marked by ✓and recommended items by recommenders Rec 0, Rec 1, Rec 2. Rec 0 produces 3 and 6 relevant items for free and premium users (in total) respectively; Rec 1 generates 1 relevant item for each user; Rec 2 produces recommended items that are all relevant for all users.

		i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	**Rec 0**	**Rec 1**	**Rec 2**
a_1	**user 1**	✓		✓				✓				$\{i_1, i_6, i_8\}$	$\{i_1, i_5, i_9\}$	$\{i_1, i_3, i_7\}$
a_1	**user 2**					✓			✓			$\{i_2, i_5, i_9\}$	$\{i_2, i_5, i_7\}$	$\{i_1, i_5, i_8\}$
a_1	**user 3**		✓					✓				$\{i_1, i_6, i_7\}$	$\{i_2, i_5, i_9\}$	$\{i_2, i_7, i_9\}$
a_2	**user 4**			✓	✓					✓		$\{i_3, i_4, i_9\}$	$\{i_4, i_5, i_6\}$	$\{i_3, i_4, i_9\}$
a_2	**user 5**					✓		✓			✓	$\{i_1, i_5, i_7\}$	$\{i_1, i_2, i_{10}\}$	$\{i_5, i_7, i_{10}\}$
a_2	**user 6**	✓		✓			✓			✓		$\{i_2, i_6, i_9\}$	$\{i_1, i_5, i_8\}$	$\{i_3, i_6, i_9\}$

Table 2: Fairness of different recommenders in the toy example presented in Table 1 according to proposed GCE and individual-level accuracy metrics. Note that $p_{f_0} = [\frac{1}{2}, \frac{1}{2}]$ and $p_{f_1} = [\frac{2}{3}, \frac{1}{3}], p_{f_2} = [\frac{1}{3}, \frac{2}{3}]$ characterize the fair distribution as uniform or non-uniform distributions between two groups.

	GCE ($p_f, p, \alpha = -1$)			P@3	R@3
	p_{f_0}	p_{f_1}	p_{f_2}		
Rec 0	0.0800	0.3025	0.0025	$\frac{1}{2}$	$\frac{1}{6} \cdot \frac{19}{6} = 0.530$
Rec 1	0	0.0625	0.0625	$\frac{1}{3}$	$\frac{1}{6} \cdot \frac{9}{4} = 0.375$
Rec 2	0.0078	0.1182	0.0244	1	$\frac{1}{6} \cdot \frac{23}{4} = 0.958$

be defined for either users, e.g., gender and race, or items, e.g., item provider (or stakeholder).

The goal is to find an *unfairness measure* I that produces a non-negative real number for a recommender system. A recommender system $\mathcal{M}$ is considered less unfair (i.e., more fair) than $\mathcal{M}'$ with respect to the attribute $a \in \mathcal{A}$ if and only if $|I(\mathcal{M},a)| < |I(\mathcal{M}',a)|$. Previous works have used *inequality* measures to evaluate algorithmic unfairness, however, we argue that fairness does not always imply equality. For instance, let us assume that there are two types of users in the system – regular (free registration) and premium (paid) – and the goal is to compute fairness with respect to the users' subscription type. In this example, it might be more fair to produce better recommendations for paid users, therefore, equality is not always equivalent to fairness. We define fairness of a recommender system as the *generalized cross entropy* (GCE) for some parameter $\alpha \neq 0,1$:

$$I(\mathcal{M},x) = \frac{1}{\alpha(1-\alpha)} \left[\int p_f^{\alpha}(x) p^{(1-\alpha)}(x)\, \mathrm{d}x - 1 \right] \tag{1}$$

where p and p_f respectively denote the probability distribution of the system performance and the fair probability distribution, both with respect to the attribute $x = a$ [6]. The unfairness measure I is minimized with respect to attribute $x = a$ when $p = p_f$, meaning that the performance of the system is equal to the performance of a fair system. In the next sections, we discuss how to obtain or estimate these two probability distributions. If the attribute a is discrete or categorical (as typical attributes, such as gender or race), then the unfairness measure is defined as:

$$I(\mathcal{M},a) = \frac{1}{\alpha(1-\alpha)} \left[\sum_{a_j} p_f^{\alpha}(a_j) p^{(1-\alpha)}(a_j) - 1 \right] \tag{2}$$

2.1 Fair Distribution p_f

The definition of a fair distribution p_f is problem-specific and should be determined based on the problem or target scenario in hand. For example, a job or music recommendation website may want to ensure that its premium users, who pay for their subscription, would receive more relevant recommendations. In this case, p_f should be non-uniform across the user classes (premium versus free users). In other scenarios, a uniform definition of p_f might be desired. Generally, when fairness is equivalent to equality, then p_f should be uniform and in that case, the generalized cross entropy would be the same as generalized entropy (see [27] for more information).

2.2 Estimating Performance Distribution p

The performance distribution p should be estimated based on the output of the recommender system on a test set. In the following, we explain how we can compute this distribution for item attributes. We define the recommendation gain (rg_i) for each item i as follows:

$$rg_i = \sum_{u \in \mathcal{U}} \phi(i, Rec_u^K)\, g(u,i,r) \tag{3}$$

where Rec_u^K is the set of top-K items recommended by the system to the user $u \in U$. $\phi(i, Rec_u^K) = 1$ if item i is present in Rec_u^K; otherwise $\phi(i, Rec_u^k) = 0$. The function $g(u,i,r)$ is the gain of recommending item i to user u with the rank r. Such gain function can be defined in different ways. In its simplest form, when $g(u,i,r) = 1$, the recommendation gain in Eq. 3 would boil down to recommendation count (i.e., $rg_i = rc_i$). A binary gain in which $g(u,i,r) = 1$ when item i recommended to user u is relevant and $g(u,i,r) = 0$ otherwise, is another simple form of the gain function based on **relevance**. The gain function g can be also defined based on **ranking** information, i.e., recommending relevant items to users in higher ranks is given a higher gain. In such case, we recommend the discounted cumulative gain (DCG) function that is widely used in the definition of NDCG [20], given by $\frac{2^{\mathrm{rel}(u,i)-1}}{\log_2(r+1)}$ where rel(u,i) denotes the relevance label for the user-item pair u and i. We can further normalize the above formula based on the ideal DCG for user u to compute the gain function g.

The performance probability distribution p is then proportional to the recommendation gain for the items associated to an attribute value a_j. Formally, the performance probability $p(a_j)$ used in Eq. (2) is computed as: $p(a_j) = \sum_{i \in a_j} rg_i / Z$ where Z is a normalization factor set equal to $Z = \sum_i rg_i$ to make sure that $\sum p(a_j) = 1$. Under an analogous formulation, we could define a variation of fairness for users based on Eq. (3):

$$rg_u = \sum_{i \in I} \phi(i, Rec_u^K)\, g(u,i,r) \tag{4}$$

where in this case, the gain function cannot be reduced to 1, otherwise, all users would receive the same recommendation gain rg_u.

3 TOY EXAMPLE

For the illustration of the proposed concept, in Table 1 we provide a toy example on how our approach for fairness evaluation framework could be applied in a real recommendation setting. A set of six users belonging to two groups (each group is associated with an attribute value a_1 (red) or a_2 (green)) who are interacting with a set of items are shown in Table 1. Let us assume the red group represents users with a *regular* (free registration) subscription type on an e-commerce website while the green group represents users with a *premium* (paid)

subscription type. A set of recommendations produced by different systems (**Rec0**, **Rec1**, and **Rec2**) are shown in the last columns. The goal is to compute fairness using the proposed fairness evaluation metric based on GCE given by Eq. (2). The results of evaluation using three different evaluation metrics are shown in Table 2. The metrics used for the evaluation of fairness and accuracy of the system include: (i) GCE (absolute value), (ii) Precision@3 and (iii) Recall@3. Note that $GCE = 0$ means the system is completely fair, and the closer the value is to zero, the more fair the respective system is.

By looking at the recommendation results of **Rec0**, one can note that *if fairness is defined in a uniform way between two groups*, defined through fair distribution $p_f = [\frac{1}{2}, \frac{1}{2}]$, then **Rec0** is not a completely fair system, since $GCE = 0.08 \neq 0$. In contrast, *if fairness is defined as providing recommendation of higher utility (usefulness) to green users who are users with paid premium membership type*, (e.g., by setting $p_f = [\frac{1}{3}, \frac{2}{3}]$) then since $GCE \approx 0$, we can say that recommendations produced by **Rec0** are fair. Both of the above conclusions are drawn with respect to attribute "subscription type" (with categories free/paid premium membership). This is an interesting insight which shows the evaluation framework is flexible enough to capture fairness based on the interest of system designer by defining what she considers as fair recommendation through the definition of p_f. While in many application scenarios we may define fairness as equality among different classes (e.g., gender, race), in some scenarios (such as those where the target attribute is not sensitive, e.g., regular v.s. premium users) fairness may not be equivalent to equality.

Furthermore, by comparing the performance results of **Rec1** and **Rec2**, we observe that, even though precision and recall improve for **Rec2** and becomes the most accurate recommendation list, it fails to keep a decent amount of fairness with respect to any parameter settings of GCE, as in both cases it is outperformed by the other methods. Moreover, GCE never reaches the optimal value, which in this case is attributed to the unequal distribution of resources among classes, since there are more relevant items on green than red users. This evidences that optimizing an algorithm to produce relevant recommendations does not necessarily result in more fair recommendation rather, conversely, a trade-off between the two evaluation properties can be noticed.

4 EXPERIMENTS AND RESULTS

In the section, we discuss our experimental setup and the results.

4.1 Data Descriptions

We conduct experiments on two real-world datasets, Xing job recommendation dataset [2] and Amazon Review dataset [1]. The datasets represent different item recommendation scenarios for job and e-commerce domains. We used Xing dataset to study the **item-related** notion of fairness, while Amazon is used to study the **user-related** notion of fairness.

Xing Job Recommendation Dataset (Xing-REC 17): The dataset was first released by XING as part of the ACM RecSys Challenge 2017 for a job recommendation task [2]. The dataset contains 320M of interactions happened in over 3 months. The reason for choosing this dataset is that it provides several user-related attributes, such as *membership types* (regular vs. premium), *education degree*, and *working country*, that can be useful for the study of fairness. For example, membership type allows us to study the non-equal (non-uniform) notion of fairness, as a recruiter may want to ensure premium users obtain better quality in their recommendations.

Amazon: We used the toy and games subset which contains 53K preference scores by 1K users for 24K items, with a sparsity of 99.8%. We wanted the training set to be as close as possible to an on-line real scenario in which the recommender system is deployed, with this goal in mind we used a time-aware splitting. The most rigorous one would be the fixed-timestamp splitting method [10, 18]. In these experiments, however, we adopted the methodology proposed in [4] where a single timestamp is chosen, which represents the moment when test users are on the platform waiting for recommendations. The training set corresponds to the past interactions, and the performance is evaluated with data which correspond to future interactions. The splitting timestamp is selected to maximize the number of users involved in the evaluation according to two constraints: the training should retain at least 15 ratings, and the test set should contain at least 5 ratings.

Table 3: Results of applying the proposed fairness evaluation metrics on Xing-REC 17 winner submission to identify *item-centered* fairness for the attribute membership type (regular v.s. premium). Note that in this case, it is desired to increase the utility of recommendation for premium (paid) users. $p_{f_0} = [1/2,1/2]$ (uniform) v.s. $p_{f_2} = [1/3,2/3]$ (non-uniform).

	Membership type		GCE ($p_f, p, \alpha = 2$)	
	regular	premium	p_{f_0}	p_{f_2}
RSC Winner	4,108,771	547,029	0.2926	0.6786
Random	4,209,878	445,759	0.3269	0.7335

Table 4: Results of applying the proposed fairness evaluation metrics on Xing-REC 17 winner submission to identify *item-centered* fairness. GCE1 and GCE2 have associated fair probability distributions equal to $p_{f_0} = [0.25,0.25,0.25,0.25]$, $p_{f_1} = [0.7,0.1,0.1,0.1]$, $p_{f_2} = [0.1,0.7,0.1,0.1]$, $p_{f_3} = [0.1,0.1,0.7,0.1]$, $p_{f_4} = [0.1,0.1,0.1,0.7]$ where p_{f_0} defines fair distribution as uniform distribution while the rest define it as favoring each of groups

	Country				GCE ($p_f, p, \alpha = 2$)				
	German	Austrian	Swiss	Other	p_{f_0}	p_{f_1}	p_{f_2}	p_{f_3}	p_{f_4}
winner	3.9M	156.1K	329.4K	186.4K	-0.979	0.061	3.194	3.177	3.192
random	3.8M	253.6K	319.6K	239.8K	-0.883	0.038	2.945	2.937	2.946
	Education				GCE ($p_f, p, \alpha = 2$)				
	NA	BSc	MSc	PhD	p_{f_0}	p_{f_1}	p_{f_2}	p_{f_3}	p_{f_4}
winner	2.8M	607.2K	1.0M	158.3K	0.398	0.0974	1.673	1.547	1.741
random	3.0M	428.4K	1.0M	203.4K	0.450	0.0887	1.838	1.670	1.866

4.2 Experimental Setup

Two recommendation scenarios are considered to evaluate the effectiveness of the proposed fairness evaluation framework with respect to **item-centric** or **user-centric** notion of fairness [8].

Item fairness evaluation: It applies the proposed fairness evaluation metrics based on GCE on the winner of the ACM RecSys Challenge 2017. The challenge was formulated as "*given a job posting, recommend a list of candidates that are suitable candidates for the job*". As such, the user candidates are considered as target items for recommendation. In order to compute GCE, we used Eq. (3) by considering a simplified case $g(u,i,r) = 1$, in which the recommendation gain rg_i boils down to recommendation count rc_i for item i, i.e., the number of times each user appears in the recommendation lists of all jobs.

We compare two recommendation approaches: the winner submission and a random submission, and evaluate the systems' fairness from the perspective of users membership types, education, and location. As for membership type, premium users (or paid members) are expected to receive better quality of recommendation.

User fairness evaluation: Here, we experiment with the more traditional item recommendation task where we study the user fairness dimension. We consider a scenario where a business owner may want to ensure superior recommendation quality for its more engaged users over less engaged (or new) users (or vice versa). In order

to have a more intuitive sense about how fair different recommendation models are recommending to users of different classes, we study the fairness of different CF recommendation models with respect to users' interactions, defined in 4 categories: (i) very inactive (VIA), (ii) slightly inactive (SIA), (iii) slightly active (SA), and (iv) very active (VA). For each user, we compute the score $n_R(u)$ that corresponds to the total number of ratings provided by user u. We group the users in four groups according to the quartile that this score belongs to.

We have experimented with several recommendation models such as UserKNN [7], ItemKNN [25] (considering binarized and cosine similarity metric, Jaccard coefficient [15], and Pearson correlation [19]), SVD++ [21, 22], BPRMF [22, 24], BPRSlim [23], and two non-personalized models, most-popular and random recommender.

For comparison with the proposed GCE metric, we include two complementary baseline metrics based on the absolute deviation between the mean ratings of different groups as defined in [31] $MAD(R^{(i)}, R^{(j)}) = \left| \frac{\sum R^{(i)}}{|R^{(i)}|} - \frac{\sum R^{(j)}}{|R^{(j)}|} \right|$ where $R^{(i)}$ denotes the predicted ratings for all user-item combinations in group i and $\left|R^{(i)}\right|$ is its size. Larger values for MAD mean larger differences between the groups, interpreted as unfairness. Given that our proposed GCE in user-fairness evaluation is based on NDCG, we adapt this definition to also compare between average NDCG for each group. We refer to these two baselines as **MAD-rating** and **MAD-ranking**. Finally, the reported MAD corresponds to the average MAD between all the pairwise combinations within the groups involved, i.e., $MAD = \text{avg}_{i,j}(MAD(R^{(i)},R^{(j)}))$.

4.3 Results and Discussion

We start our analysis with the results for the item fairness evaluation as described in Section 4.2, presented in Tables 3 and 4. The counts in these tables represent the total number of users with a given category that each submission recommends. We observe in Table 4 that recommendations produced by the RecSys Challenge winner performs better with p_{f_0} than with p_{f_2}, since the GCE value is closer to 0. This evidences that the proposed winner system produces balanced recommendations across the two membership classes. This is in contrast to our expectation that premium users should be provided better recommendations. Therefore, even though the winning submission could produce higher recommendation quality *from a global perspective*, it does not comply with our expectation of a fair recommendation for this attribute, which is to recommend better recommendations to premium users.

Furthermore, in Table 4 we present the recommendation fairness evaluation results using GCE across two other attributes: Country and Education; each of these attributes takes 4 categories. We define five variations of the fair distribution p_f: while p_{f_0} considers all attribute categories equally important, the others give one attribute category a higher importance compared to the rest. After applying the GCE on the winner submission, we observe that with respect to the Country attribute, the lowest value of GCE (best case) is produced for the German companies ($GCE = 0.061$) while for the Education attribute the category Unknown ($GCE = 0.97$) produces the best outcome, in both cases, these categories are the most frequently recommended by the analyzed submission. These results show that for a given target application, if the system is looking for candidates with certain nationality (in this case, German) or education-level (here any), the system recommendations coming from the winner submission are closer to a fair system. In fact, due to the inherent biases in the dataset, the random submission is obtaining better results according to our definition of fairness for several of the fair distributions analyzed. However, it is worth mentioning that if the system designer wants to promote those users with BSc or PhD, the GCE would show that the winner submission provides better recommendations to those users than the random submission. To the best of our knowledge, this is the first fairness-oriented evaluation metric that allows to capture these nuances, which as a consequence, helps on understanding how the recommendation algorithms work on each user group.

Now Table 5 shows the results for the proposed user fairness evaluation as described in Section 4.2. We observe in this table that each recommender obtains a GCE value on a different range, an obvious consequence of the different performance obtained in each case for the different groups (as we observe in the NDCG@10 columns for each user type). For instance, BPRMF is the one found by GCE to perform in a fair way when assuming uniformity with respect to the user groups (p_{f_0}), however, if the system designer aims to promote those recommenders that provide better suggestions to the most active users then Random, followed by ItemKNN and SVD++ are the most fair algorithms.

Comparing MAD against GCE, we observe that MAD-ranking produces lower results when NDCGs in each class are close to each other (e.g., in the case of Random recommender), which corresponds to the already discussed notion of fairness as equality/uniformity; similarly, MAD-rating obtains better results for the random algorithm because, as expected, such method has no inherent bias with respect to the defined user groups, but also for SVD++, probably because this recommender tends to predict ratings in a small range. In both cases, it becomes evident that MAD, in contrast to our proposed GCE metric, cannot incorporate other definitions of fairness in its computation, hence, its flexibility is very limited.

In summary, we have shown that our proposed fairness evaluation metric is able to unveil whether a recommendation algorithm satisfies our definition of fairness, where we argue that it should emphasize a proper distribution of utility based on merits and needs. We demonstrate this in both notions of fairness: based on users and based on items. Therefore, we conclude that this metric could help better explaining the results of the algorithms towards specific groups of users and items, and as a consequence, it could increase the transparency of the recommender systems evaluation.

5 CONCLUSION

Fairness-aware recommendation research requires appropriate evaluation metrics to quantify fairness. Furthermore, fairness in RS can be associated with either items or users, even though this complementary view has been underrepresented in the literature. In this work, we have presented a probabilistic framework to measure fairness of RS under the perspective of users and items. Experimental results on two real-world datasets show the merits of the proposed evaluation framework. In particular, one of the key aspects of our proposed evaluation metric is its transparency and flexibility, since it allows to incorporate domain knowledge (by means of an ideal fair distribution) that helps on understanding which item or user aspects the recommendation algorithms are over- or under-representing.

In the future, we plan to exploit the proposed fairness and relevance aware evaluation system to build recommender systems that directly optimize for this objective criterion. Also, it is of our interest to consider studying various fairness of recommendation under various content-based filtering or CF models using item content as side information [11, 12] on different domains (e.g., tourism [3], entertainment [14], social recommendation among others). Finally, we are considering to investigate the robustness of CF models against shilling attacks [13] crafted to undermine not only the accuracy of recommendations but also fairness of these models.

Table 5: Results of applying the proposed fairness evaluation metrics on Amazon dataset to identify user-centered fairness. The fair probability distributions are defined as p_{f_i} so that $p_{f_i}(j) = 0.1$ when $j \neq i$ and 0.7 otherwise, except for p_{f_0} that denotes the uniform distribution; i.e., just as in Table 4. Types of users (VIA/SIA/SA/VA) as defined in Section 4.2. Best values per column in bold.

	NDCG@10				GCE (p_f, p, $\alpha = -1$)					MAD	
	VIA	SIA	SA	VA	p_{f_0}	p_{f_1}	p_{f_2}	p_{f_3}	p_{f_4}	rating	ranking
Random	0.0000	0.0000	0.0000	0.0005	1.5000	4.5000	4.5000	4.5000	**0.2143**	**0.0000**	**0.0003**
MostPopular	0.0000	0.0006	0.0013	0.0014	0.2435	1.3586	1.2289	0.6714	0.5825	0.1864	0.0008
ItemKNN	0.0023	0.0021	0.0016	0.0036	0.0487	0.6218	0.6722	0.7537	0.2636	0.0254	0.0011
UserKNN	**0.0031**	**0.0040**	**0.0037**	**0.0053**	0.0214	0.6483	0.5379	0.5783	0.3319	0.0375	0.0012
BPRMF	0.0022	0.0025	0.0028	0.0016	**0.0191**	0.5496	**0.4767**	0.3881	0.6642	0.2078	0.0006
BPRSlim	0.0027	0.0023	0.0035	0.0017	0.0353	**0.5377**	0.6150	**0.3267**	0.7267	9.0009	0.0010
SVD++	0.0025	0.0025	0.0025	0.0042	0.0324	0.6336	0.6382	0.6361	0.2750	0.0027	0.0009

6 ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by project TIN2016-80630-P (MINECO). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] 2014 (accessed June 5, 2019). *Amazon product data.* http://jmcauley.ucsd.edu/data/amazon/.

[2] Fabian Abel, Yashar Deldjoo, Mehdi Elahi, and Daniel Kohlsdorf. 2017. RecSys Challenge 2017: Offline and Online Evaluation. In *Proc. RecSys (RecSys '17).* ACM, New York, NY, USA, 372–373.

[3] Jens Adamczak, Gerard-Paul Leyson, Peter Knees, Yashar Deldjoo, Farshad Bakhshandegan Moghaddam, Julia Neidhardt, Wolfgang Wörndl, and Philipp Monreal. 2019. Session-Based Hotel Recommendations: Challenges and Future Directions. *arXiv preprint arXiv:1908.00071* (2019).

[4] Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Azzurra Ragone, and Joseph Trotta. 2019. Local Popularity and Time in top-N Recommendation. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I.* 861–868.

[5] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *Proc. SIGIR, Ann Arbor, MI, USA, July 08-12, 2018.* ACM, 405–414.

[6] Zdravko I Botev and Dirk P Kroese. 2011. The generalized cross entropy method, with applications to probability density estimation. *Methodology and Computing in Applied Probability* 13, 1 (2011), 1–27.

[7] John S. Breese, David Heckerman, and Carl Myers Kadie. 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, University of Wisconsin Business School, Madison, Wisconsin, USA, July 24-26, 1998.* Morgan Kaufmann, 43–52.

[8] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[9] 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),* Vol. 81. PMLR, 202–214.

[10] Pedro G. Campos, Fernando Díez, and Iván Cantador. 2014. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Model. User-Adapt. Interact.* 24, 1-2 (2014), 67–119.

[11] Yashar Deldjoo, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. 2018. Audio-visual encoding of multimedia content for enhancing movie recommendations. In *Proc. RecSys, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018.* ACM, 455–459.

[12] Yashar Deldjoo, Maurizio Ferrari Dacrema, Mihai Gabriel Constantin, Hamid Eghbal-zadeh, Stefano Cereda, Markus Schedl, Bogdan Ionescu, and Paolo Cremonesi. 2019. Movie genome: alleviating new item cold start in movie recommendation. *User Model. User-Adapt. Interact.* 29, 2 (2019), 291–343.

[13] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2019. Assessing the Impact of a User-Item Collaborative Attack on Class of Users. In *Workshop on the Impact of Recommender Systems (ImpactRecSys'19) - 13th ACM Conference of Recommender Systems.*

[14] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2018. Content-Based Multimedia Recommendation Systems: Definition and Application Domains. In *Proceedings of the 9th Italian Information Retrieval Workshop, Rome, Italy, May, 28-30, 2018. (CEUR Workshop Proceedings),* Vol. 2140.

[15] Wei Dong, Charikar Moses, and Kai Li. 2011. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web.* ACM, 577–586.

[16] 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.* 172–186.

[17] Golnoosh Farnadi, Pigi Kouki, Spencer K. Thompson, Sriram Srinivasan, and Lise Getoor. 2018. A Fairness-aware Hybrid Recommender System. *CoRR* abs/1809.09030 (2018). arXiv:1809.09030

[18] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook.* Springer, 265–308.

[19] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2002. An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms. *Inf. Retr.* 5, 4 (2002), 287–310.

[20] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[21] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008.* ACM, 426–434.

[22] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.

[23] Xia Ning and George Karypis. 2011. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011.* IEEE Computer Society, 497–506.

[24] 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.* AUAI Press, 452–461.

[25] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2000. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM conference on Electronic commerce.* ACM, 158–167.

[26] 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.* ACM, 2219–2228.

[27] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* ACM, 2239–2248.

[28] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems.* 2921–2930.

[29] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo A. 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.* ACM, 1569–1578.

[30] Yong Zheng, Tanaya Dave, Neha Mishra, and Harshit Kumar. 2018. Fairness In Reciprocal Recommendations: A Speed-Dating Study. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP 2018, Singapore, July 08-11, 2018.* ACM, 29–34.

[31] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-Aware Tensor-Based Recommendation. In *Proc. CIKM, CIKM 2018, Torino, Italy, October 22-26, 2018.* ACM, 1153–1162.