

Measuring Individual User Fairness with User Similarity and Effectiveness Disparity

Theresia Veronika Rampisela¹[0000–0003–1233–7690], Maria Maistro¹[0000–0002–7001–4817], Tuukka Ruotsalo^{1,2}[0000–0002–2203–4928], and Christina Lioma¹[0000–0003–2600–2701]

¹ University of Copenhagen, Copenhagen, Denmark
{thra,mm,tr,c.lioma}@di.ku.dk

² LUT University, Lappeenranta, Finland

Abstract. Individual user fairness is commonly understood as treating similar users similarly. In Recommender Systems (RSs), several evaluation measures exist for quantifying individual user fairness. These measures evaluate fairness via either: (i) the disparity in RS effectiveness scores regardless of user similarity, or (ii) the disparity in items recommended to similar users regardless of item relevance. Both disparity in recommendation effectiveness and user similarity are very important in fairness, yet no existing individual user fairness measure simultaneously accounts for both. In brief, current user fairness evaluation measures implement a largely incomplete definition of fairness. To fill this gap, we present Pairwise User unFairness (PUF), a novel evaluation measure of individual user fairness that considers both effectiveness disparity and user similarity. PUF is the only measure that can express this important distinction. We empirically validate that PUF does this consistently across 4 datasets and 7 rankers, and robustly when varying user similarity or effectiveness. In contrast, all other measures are either almost insensitive to effectiveness disparity or completely insensitive to user similarity. We contribute the first RS evaluation measure to reliably capture both user similarity and effectiveness in individual user fairness. Our code: github.com/theresiavr/PUF-individual-user-fairness-recsys.

Keywords: individual fairness · evaluation · recommender systems

1 Introduction

Recommender System (RS) evaluation has always relied heavily on effectiveness, as it directly affects user utility and satisfaction. Existing work on RS fairness evaluation often uses measures that depend on effectiveness scores, especially when it comes to individual user fairness. *Individual fairness* is traditionally defined as treating similar individuals similarly [12]. We focus on attribute-free fairness for individual users, where we assume no user attribute is available, other than the user identifiers and interactions [24, 54]. Sensitive attributes (e.g., race, age) are frequently unavailable due to privacy or data incompleteness [5].

Table 1. Overview of individual user fairness measures in RS.

| Measure | Reference | Effectiveness | User Similarity |
|--------------------|-------------------------|---------------|-----------------|
| Standard deviation | [4, 24, 31, 38, 50, 51] | ✓ | - |
| Gini index | [15, 23] | ✓ | - |
| Envy-based | [4, 11, 31] | ✓ | - |
| UF | [48] | - | ✓ |
| PUF | ours | ✓ | ✓ |

Table 2. Example of how our measure, PUF, quantifies unfairness.

| User Similarity | Effectiveness Disparity | Unfairness (PUF) |
|-----------------|-------------------------|------------------|
| High | High | High |
| High | Low | Low |
| Low | High | Low |
| Low | Low | Low |

To measure individual user fairness in RSs, the disparity in recommendation effectiveness across users is often used as a proxy of how similarly the recommendation algorithm treats the users. Recommendations across users are deemed fair if similar users have similar effectiveness scores. Yet, no existing individual user fairness measure considers both effectiveness and user similarity (Tab. 1).

User similarity and recommendation effectiveness are both important, as the former affects our expectation of how close effectiveness scores should be. For example, given two users whose past interactions are highly similar, the recommendations they receive are deemed fair if their effectiveness is similar, and vice versa. This is because fairness means similar users should get similar treatment. However, two dissimilar users cannot expect to get similar treatment, as their past interactions may differ, e.g., in terms of amount, frequency, or item type.

Current individual user fairness measures can be grouped into those that consider: (i) only effectiveness disparity; and (ii) only user similarity and recommendation similarity. For (i), different effectiveness for users is not always unfair. Some users may have only a few past interactions, and others may have very specific tastes for whom only a few items are relevant. Fairness should consider that these users are different and cannot be compared to users who, e.g., consume only popular items. For (ii), recommending similar items to similar users may not be fair if one user is satisfied with their recommendation, but another is not. In this case, the recommendation is not really fair as its effectiveness differs.

To counter the above limitations, we propose a novel evaluation measure for individual user fairness in RS: **Pairwise User unFairness** (PUF). PUF³ quantifies individual user fairness based on the disparity in recommendation relevance between user pairs, weighted by the similarity of user pairs. Our measure works as follows (Tab. 2). PUF will give a high unfairness score if highly similar users receive recommendations with a high difference in effectiveness. In contrast, PUF will give a low unfairness score for: (1) highly similar users with a low difference in effectiveness; and (2) users with low similarity. As such, PUF is aligned with the definition of individual fairness and also accounts for recommendation effectiveness, and thus, user utility. We show that compared to existing measures, PUF has higher sensitivity to changes in effectiveness scores and user similarity distribution. Overall, we contribute a new evaluation measure for individual user fairness, which considers both user similarity and disparity in recommendation effectiveness, and which does not have the same limitations as existing measures.

³ Pronounced as ‘puff’.

2 Individual User Fairness

We define individual user fairness as per [12]: let u and u' denote two users; L_u and $L_{u'}$ be the recommendation lists of these two users; and $M(\cdot)$ be a function mapping a recommendation list to a score, e.g., its effectiveness score. Any two users whose profile distance is $d(u, u')$ should receive recommendations such that recommendation effectiveness satisfies $D(M(L_u), M(L_{u'})) \leq d(u, u')$, where D is a distance measure. In other words, fairness is achieved when the difference in the users' recommendation effectiveness is at most $d(u, u')$. This definition agrees with the definitions of RS individual user fairness in [42, 53].

Next, we present all evaluation measures of attribute-free individual user fairness in RSs that have been published up to January 2025 [1, 3, 25, 32, 42, 46, 49, 53, 56]; we include their equations in an online appendix together with the code repository. All of them quantify unfairness, and the lower their score, the fairer (denoted by \downarrow). All these measures, except the distance-based measure, quantify effectiveness disparity but ignore user similarity. While the distance-based measure considers user similarity, it is detached from effectiveness. Hence, no existing individual user fairness measure in RS considers both user similarity and effectiveness disparity.

Standard Deviation (SD). In RSs, \downarrow SD and variance are often used to quantify individual user fairness. An RS is fair if it provides equal prediction accuracy to all users, and fairness is evaluated via the variance of the mean squared error (MSE) of user rating prediction [24, 38]. Other works measure fairness from user recommendation lists, e.g., via the variance of NDCG scores across users [50] or the SD of user utility [4, 31, 51].

Gini Index (Gini). \downarrow Gini is a well-known inequality measure that quantifies the extent to which a distribution deviates from a perfectly equal distribution. It has been used to measure individual user fairness from different distributions, e.g., $P@k$ (Gini-P), NDCG (Gini-NDCG), or the utility score per user [15, 23].

Envy-based measures. Envy is defined as the extra utility a user u would have received if they were given the recommendation list of user u' , $L_{u'}$ [4, 31]. Three fairness evaluation measures are based on this concept, where envy is aggregated differently: Mean Envy (\downarrow ME) [4, 31], Mean Max Envy (\downarrow MME) [11], and Proportion of ϵ -Envious User (\downarrow PEU) [11].

Distance-based measures. In [48], fairness for individual users is operationalised as any two similar users u, u' receiving similar recommendations. Recommendation disparity between pairs of users whose similarity exceed a mean-based similarity threshold is then measured with the UnFairness score (\downarrow UF). UF uses both the user similarity and the pairwise distance between the representation (e.g., embeddings) of items in the recommendation list of users u and u' . User similarity is modelled by a weighted sum of Jaccard (sim_{Jacc}) of the user's set of past interactions and JS-divergence between item feature (e.g., genre) distributions in the interactions of users u and u' . Unlike the other individual user fairness measures,⁴ UF does not consider recommendation effectiveness.

⁴ Including PUF. We further explain how PUF differs from UF in §3.

3 Pairwise User unFairness (PUF)

We present our evaluation measure of individual user fairness, Pairwise User unFairness (\downarrow PUF). PUF has two components: similarity between users and disparity in recommendation effectiveness, which we describe next.

(1) User similarity. Measuring similarity between users is an inherently hard problem, and there is no single ground truth of what makes two users similar [6]. When there is no user attribute, as in our case, user profiles can be modelled based on their historical interactions [18, 24, 48]. User similarity is then computed pairwise based on the user representation (e.g., click/rating matrix, user embedding), for example, with cosine similarity [39] or Jaccard [48].

(2) Effectiveness disparity. PUF quantifies individual user fairness based on disparity in recommendation effectiveness, considering user similarity. PUF measures the mean⁵ pairwise difference in the effectiveness score per user, weighted by the user pair similarity. Intuitively, PUF measures *how much the recommendation quality differs between users who interact with the same items (i.e., users with similar profiles)*. Based on (1) and (2), we define PUF as follows:

$$\text{PUF} = \frac{2}{m(m-1)} \sum_{u \in U} \sum_{u' \in U \setminus \{u\}} \text{sim}(u, u') \times |S(u) - S(u')| \quad (1)$$

where U is the set of all users ($|U| = m$, where $m \geq 2$), $\text{sim}(u, u') \in [0, 1]$ is the similarity of users u and u' , S is an effectiveness measure, e.g., $\text{P@}k$. S must range or scaled to be in $[0, 1]$, so PUF also ranges in $[0, 1]$. PUF can be used with any similarity/effectiveness measure that fulfils the range requirement.

How PUF differs from UF. At a high level, while both PUF and UF [48] consider user similarity, PUF considers the difference in the recommendation relevance, rather than only the disparity based on the representation of the recommended items, as in UF. Moreover, recommending different sets of items to two similar users may be considered unfair by UF even if both users like their recommendations, but it is fair based on PUF. In theory, users with similar tastes in the past are likely to have a similar preference in the future [41]. Yet, the recommendation problem is challenging, as even two highly similar users may not equally like their recommended items if they are given identical items. This may be due to, for example, incomplete historical data that is unrepresentative of user taste, diverging user preferences, or limited ground truth data. These limitations necessitate a look into the disparity of the recommendation relevance, which is what our measure quantifies.

Overall, our measure, PUF, aligns with the definition of individual user fairness and quantifies fairness through the disparity in recommendation effectiveness, which is more meaningful than the disparity in the representation of recommended items, as effectiveness relates more to user utility.

⁵ To get the mean, the summation is divided by the $\binom{m}{2} = \frac{m(m-1)}{2}$ user pairs.

4 Empirical Analysis

We compare PUF to existing effectiveness and individual user fairness measures.

4.1 Experimental Setup

Datasets. We use four real-world datasets from three domains: music (Lastfm [7]), video (QK-video [52]), and movie (ML-10M and ML-20M [16]). We obtain Lastfm and ML-* from [55], and QK-video from [52]. We use the ‘sharing’ interactions in QK-video. For all datasets, we remove duplicate interactions and keep the most recent. We remove users and items with < 5 interactions. For ML-* we map ratings ≥ 3 to 1, and discard the rest. The threshold 3 is chosen as the ratings range between $[0.5, 5]$. Lastfm and QK-video have unary ratings, so no mapping is required. We split the preprocessed datasets into train/val/test with a ratio of 6:2:2. The ML-* datasets are temporally split, while Lastfm and QK-video are randomly split as they have no timestamps. We split datasets globally (not user-wise) to avoid data leakage [30]. After splitting, users with < 5 interactions in the train set are removed from all splits to ensure that each user has adequate training data. The final preprocessed dataset statistics are in Tab. 3.

Recommenders. We use 7 well-known collaborative filtering recommenders: user- and item-based K -nearest neighbour (U-KNN [41] and I-KNN [10]), Bayesian Personalised Ranking (BPR [40]), Variational Autoencoder with multinomial likelihood (MVAE [26]), Neural Graph Collaborative Filtering (NGCF [44]), Neural Matrix Factorisation (NMF [17]), and Neighbourhood-enriched Contrastive Learning (NCL [27]). All models except U- and I-KNN are trained for 300 epochs with early stopping. We tune hyperparameters with grid search. The configuration with the best NDCG@10 during validation is the final model. Implementation, training, and tuning are done with RecBole [55].

Evaluation measures. We measure recommendation effectiveness (EFF) with Hit Rate (HR), MRR, Precision (P or Prec), Recall (R), MAP, and NDCG [19]. Individual user fairness (FAIR) is evaluated with our PUF measure (§3) and all existing measures (§2): standard deviation (SD) of the $P@k$ (SD-P) and NDCG@ k (SD-NDCG) scores across all users; Gini Index of $P@k$ (Gini-P) and NDCG@ k (Gini-NDCG); envy-based measures (ME, MME, and PEU); and distance-based measure, UF.

We evaluate all runs at $k = 10$. For PEU, we set $\epsilon = 0.05$ [11]. For UF, the number of all user pairs is used as the log base, so that $UF \in [0, 1]$. As we use several models with diverse item representations, for a fair comparison between all models, we represent items with the one-hot encoding of the historical interactions. We test four variants of PUF (two user similarity measures \times two effectiveness measures). We use cosine (sim_{cos}) and Jaccard (sim_{Jacc}) for similarity measures, as both are used in UF.⁶ In addition, sim_{cos} is used

⁶ For UF and PUF, we also vary the weights of user interaction and item features in computing user similarity. As we find similar conclusions, we report only interaction-based similarity; the rest is in the online appendix.

Table 3. Statistics of the preprocessed datasets.

| dataset | #users (all/test) | #items | #interactions | sparsity (%) |
|---------------|-------------------|--------|---------------|--------------|
| Lastfm [7] | 1,859/1,836 | 2,823 | 71,355 | 98.64% |
| QK-video [52] | 4,656/3,514 | 6,423 | 51,777 | 99.83% |
| ML-10M [16] | 49,378/1,523 | 9,821 | 5,362,685 | 98.89% |
| ML-20M [16] | 89,917/2,178 | 16,404 | 10,588,141 | 99.28% |

in our U-KNN model, as well as to analyse user similarity [39]. All user similarities are computed from the observed interactions in the train set (i.e., the binarised interaction matrix) because they are guaranteed to be available, even if sparse. As such, both cosine and Jaccard similarities account for the number of interactions per user. We compute PUF with P and NDCG, to represent set- and rank-based measures, as well as for consistency with other FAIR measures, which are also computed with P and NDCG. To avoid extremely small values, we min-max normalise the pairwise user similarity per dataset.

4.2 Comparison of All Evaluation Measures

Tab. 4 shows the EFF and FAIR evaluation results.⁷ To identify empirical limitations in the measures, we study their score range and computational efficiency. By doing so, we provide insights into the similarities and differences in how the various FAIR measures evaluate individual user fairness in practice.

Score range. A measure that does not have a wide score range across various fairness levels is less useful in distinguishing changes in fairness. Such measures may create an illusion of a negligible difference in fairness, due to their compressed empirical range [28, 29, 35, 37]. Across datasets, the observed range (not the theoretical range) of each FAIR measure varies, except for MME, which is extremely small ($\leq 5.2 \times 10^{-3}$). Swapping a user’s recommendation list with another user’s does not generally result in a large increase in the user’s P@k (envy), which translates to low MME. ME and PEU are unaffected by this despite being envy-based, as ME accounts for envy across all users rather than the maximum envy per user, and PEU employs an envy threshold. In short, MME is the least sensitive as it fails to discriminate between models, while other FAIR measures, including PUF, are more sensitive. The most sensitive measures, i.e., the ones with the widest observed range, are Gini and PEU.

Measure computation time. A potential concern about pairwise evaluation measures such as PUF is their computational efficiency, due to quadratic time complexity. We report the average computation time of the FAIR measures in Tab. 5. All runs are done with AMD EPYC 75F3 for a fair comparison. User pairwise similarity is computed once per dataset for each PUF variant. We find

⁷ Note that using the same similarity measure sim_{cos} for both PUF and the recommender model (U-KNN) does not lead to the model being the fairest based on PUF. This may be due to the data split, e.g., two users having identical train items but different test items.

Table 4. EFF and FAIR scores at $k = 10$ for recommenders. The most effective/fair score per measure is bolded. \uparrow/\downarrow means the higher/lower the better.

| | | U-KNN | L-KNN | BPR | NMF | MVAE | NGCF | NCL | | | U-KNN | L-KNN | BPR | NMF | MVAE | NGCF | NCL |
|----------|-------------------|------------------------|---------------|---------------|--------|---------------|---------------|---------------|---------------|------------------------|---------------|---------------|--------|--------|---------------|---------------|---------------|
| Lastfm | Eff | \uparrow HR | 0.7102 | 0.7647 | 0.7729 | 0.7424 | 0.7783 | 0.7707 | 0.7930 | \uparrow HR | 0.5207 | 0.4872 | 0.5121 | 0.5141 | 0.4169 | 0.5128 | 0.5213 |
| | | \uparrow MRR | 0.4386 | 0.4835 | 0.4916 | 0.4641 | 0.4764 | 0.4954 | 0.5026 | \uparrow MRR | 0.3105 | 0.2818 | 0.2987 | 0.2915 | 0.2372 | 0.2865 | 0.3019 |
| | | \uparrow P | 0.1532 | 0.1721 | 0.1782 | 0.1649 | 0.1758 | 0.1779 | 0.1840 | \uparrow P | 0.1531 | 0.1369 | 0.1458 | 0.1433 | 0.1066 | 0.1460 | 0.1536 |
| | | \uparrow MAP | 0.1190 | 0.1369 | 0.1414 | 0.1291 | 0.1377 | 0.1418 | 0.1478 | \uparrow MAP | 0.1027 | 0.0887 | 0.0953 | 0.0911 | 0.0666 | 0.0930 | 0.1009 |
| | | \uparrow R | 0.1899 | 0.2180 | 0.2238 | 0.2069 | 0.2241 | 0.2236 | 0.2342 | \uparrow R | 0.0276 | 0.0222 | 0.0249 | 0.0246 | 0.0203 | 0.0246 | 0.0263 |
| | | \uparrow NDCG | 0.2167 | 0.2450 | 0.2520 | 0.2334 | 0.2473 | 0.2530 | 0.2611 | \uparrow NDCG | 0.1691 | 0.1501 | 0.1600 | 0.1560 | 0.1191 | 0.1579 | 0.1673 |
| | Fair | \downarrow SD-P | 0.1485 | 0.1529 | 0.1538 | 0.1505 | 0.1518 | 0.1527 | 0.1552 | \downarrow SD-P | 0.2206 | 0.2080 | 0.2125 | 0.2076 | 0.1781 | 0.2109 | 0.2195 |
| | | \downarrow SD-NDCG | 0.2040 | 0.2091 | 0.2092 | 0.2072 | 0.2084 | 0.2085 | 0.2124 | \downarrow SD-NDCG | 0.2376 | 0.2244 | 0.2297 | 0.2241 | 0.1960 | 0.2262 | 0.2367 |
| | | \downarrow Gini-P | 0.5146 | 0.4768 | 0.4668 | 0.4889 | 0.4654 | 0.4642 | 0.4554 | \downarrow Gini-P | 0.6884 | 0.7076 | 0.6906 | 0.6874 | 0.7415 | 0.6880 | 0.6853 |
| | | \downarrow Gini-NDCG | 0.5213 | 0.4789 | 0.4678 | 0.4953 | 0.4735 | 0.4646 | 0.4581 | \downarrow Gini-NDCG | 0.6889 | 0.7123 | 0.6954 | 0.6948 | 0.7460 | 0.6943 | 0.6911 |
| ML-10M | Eff | \uparrow ME | 0.0897 | 0.0827 | 0.0904 | 0.0983 | 0.0973 | 0.0897 | 0.0886 | \uparrow ME | 0.1017 | 0.1131 | 0.1290 | 0.1307 | 0.0726 | 0.1284 | 0.1192 |
| | | \uparrow MME | 0.0025 | 0.0017 | 0.0020 | 0.0025 | 0.0020 | 0.0021 | 0.0019 | \uparrow MME | 0.0039 | 0.0040 | 0.0046 | 0.0046 | 0.0025 | 0.0043 | 0.0044 |
| | | \uparrow PEU | 0.6525 | 0.6176 | 0.6318 | 0.6765 | 0.6531 | 0.6275 | 0.6193 | \uparrow PEU | 0.5936 | 0.6356 | 0.6927 | 0.7072 | 0.4668 | 0.7026 | 0.6619 |
| | | \uparrow UF | 0.6615 | 0.6616 | 0.6660 | 0.6667 | 0.6685 | 0.6675 | 0.6681 | \uparrow UF | 0.6948 | 0.6954 | 0.7148 | 0.7103 | 0.6994 | 0.7095 | 0.7053 |
| | Fair | \downarrow Prec-Cos | 0.0097 | 0.0100 | 0.0100 | 0.0096 | 0.0098 | 0.0099 | 0.0100 | \downarrow Prec-Cos | 0.0493 | 0.0445 | 0.0472 | 0.0458 | 0.0363 | 0.0469 | 0.0497 |
| | | \downarrow Prec-Jacc | 0.0073 | 0.0075 | 0.0075 | 0.0073 | 0.0074 | 0.0075 | 0.0076 | \downarrow Prec-Jacc | 0.0298 | 0.0268 | 0.0285 | 0.0276 | 0.0217 | 0.0284 | 0.0302 |
| | | \downarrow NDCG-Cos | 0.0134 | 0.0137 | 0.0136 | 0.0132 | 0.0135 | 0.0135 | 0.0136 | \downarrow NDCG-Cos | 0.0547 | 0.0495 | 0.0523 | 0.0507 | 0.0409 | 0.0515 | 0.0549 |
| | | \downarrow NDCG-Jacc | 0.0101 | 0.0103 | 0.0102 | 0.0099 | 0.0101 | 0.0101 | 0.0102 | \downarrow NDCG-Jacc | 0.0331 | 0.0298 | 0.0316 | 0.0306 | 0.0245 | 0.0311 | 0.0333 |
| | ML-10M (existing) | \uparrow HR | 0.1155 | 0.0396 | 0.0993 | 0.0857 | 0.1093 | 0.1238 | 0.1298 | \uparrow HR | 0.5092 | 0.4881 | 0.5051 | 0.4927 | 0.4890 | 0.5083 | 0.5051 |
| | | \uparrow MRR | 0.0435 | 0.0131 | 0.0390 | 0.0303 | 0.0392 | 0.0484 | 0.0481 | \uparrow MRR | 0.2970 | 0.2799 | 0.2928 | 0.2781 | 0.2593 | 0.3007 | 0.2934 |
| QK-video | Eff | \uparrow P | 0.0122 | 0.0041 | 0.0105 | 0.0088 | 0.0116 | 0.0134 | 0.0140 | \uparrow P | 0.1582 | 0.1423 | 0.1449 | 0.1433 | 0.1409 | 0.1517 | 0.1500 |
| | | \uparrow MAP | 0.0188 | 0.0046 | 0.0170 | 0.0128 | 0.0178 | 0.0218 | 0.0218 | \uparrow MAP | 0.1079 | 0.0923 | 0.0957 | 0.0934 | 0.0889 | 0.1010 | 0.0997 |
| | | \uparrow R | 0.0517 | 0.0141 | 0.0433 | 0.0369 | 0.0506 | 0.0589 | 0.0608 | \uparrow R | 0.0213 | 0.0193 | 0.0193 | 0.0186 | 0.0185 | 0.0199 | 0.0200 |
| | | \uparrow NDCG | 0.0331 | 0.0091 | 0.0290 | 0.0231 | 0.0314 | 0.0375 | 0.0381 | \uparrow NDCG | 0.1711 | 0.1538 | 0.1584 | 0.1544 | 0.1481 | 0.1649 | 0.1626 |
| | Fair | \downarrow SD-P | 0.0350 | 0.0205 | 0.0327 | 0.0293 | 0.0343 | 0.0373 | 0.0376 | \downarrow SD-P | 0.2327 | 0.2145 | 0.2157 | 0.2160 | 0.2152 | 0.2239 | 0.2228 |
| | | \downarrow SD-NDCG | 0.1089 | 0.0539 | 0.1058 | 0.0910 | 0.1079 | 0.1191 | 0.1188 | \downarrow SD-NDCG | 0.2484 | 0.2290 | 0.2333 | 0.2324 | 0.2278 | 0.2399 | 0.2385 |
| | | \downarrow Gini-P | 0.8906 | 0.9618 | 0.9060 | 0.9168 | 0.8968 | 0.8850 | 0.8790 | \downarrow Gini-P | 0.6981 | 0.7068 | 0.6991 | 0.7054 | 0.7100 | 0.6983 | 0.7001 |
| | | \downarrow Gini-NDCG | 0.9215 | 0.9733 | 0.9344 | 0.9425 | 0.9265 | 0.9156 | 0.9120 | \downarrow Gini-NDCG | 0.7021 | 0.7110 | 0.7053 | 0.7138 | 0.7203 | 0.7022 | 0.7046 |
| | | \downarrow ME | 0.0874 | 0.0957 | 0.0962 | 0.0757 | 0.0412 | 0.0899 | 0.0622 | \downarrow ME | 0.1034 | 0.1199 | 0.1361 | 0.1412 | 0.1421 | 0.1348 | 0.1337 |
| | | \downarrow MME | 0.0012 | 0.0003 | 0.0010 | 0.0014 | 0.0011 | 0.0011 | 0.0012 | \downarrow MME | 0.0038 | 0.0049 | 0.0046 | 0.0052 | 0.0049 | 0.0046 | 0.0050 |
| ML-20M | Eff | \uparrow PEU | 0.7746 | 0.9087 | 0.8426 | 0.6810 | 0.3867 | 0.8056 | 0.5626 | \uparrow PEU | 0.5877 | 0.6327 | 0.7029 | 0.6956 | 0.7071 | 0.6910 | 0.6869 |
| | | \uparrow UF | 0.5731 | 0.5751 | 0.5739 | 0.5736 | 0.5721 | 0.5734 | 0.5726 | \uparrow UF | 0.7098 | 0.7098 | 0.7339 | 0.7305 | 0.7336 | 0.7309 | 0.7246 |
| | Fair | \downarrow Prec-Cos | 0.0001 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | \downarrow Prec-Cos | 0.0553 | 0.0502 | 0.0504 | 0.0503 | 0.0497 | 0.0529 | 0.0525 |
| | | \downarrow Prec-Jacc | 0.0001 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | \downarrow Prec-Jacc | 0.0329 | 0.0298 | 0.0299 | 0.0299 | 0.0296 | 0.0314 | 0.0312 |
| | | \downarrow NDCG-Cos | 0.0003 | 0.0001 | 0.0003 | 0.0002 | 0.0003 | 0.0003 | 0.0003 | \downarrow NDCG-Cos | 0.0605 | 0.0548 | 0.0559 | 0.0552 | 0.0532 | 0.0581 | 0.0576 |
| | | \downarrow NDCG-Jacc | 0.0002 | 0.0001 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | \downarrow NDCG-Jacc | 0.0361 | 0.0326 | 0.0333 | 0.0328 | 0.0317 | 0.0346 | 0.0343 |
| | ML-20M (existing) | \uparrow HR | 0.1155 | 0.0396 | 0.0993 | 0.0857 | 0.1093 | 0.1238 | 0.1298 | \uparrow HR | 0.5092 | 0.4881 | 0.5051 | 0.4927 | 0.4890 | 0.5083 | 0.5051 |
| | | \uparrow MRR | 0.0435 | 0.0131 | 0.0390 | 0.0303 | 0.0392 | 0.0484 | 0.0481 | \uparrow MRR | 0.2970 | 0.2799 | 0.2928 | 0.2781 | 0.2593 | 0.3007 | 0.2934 |
| | | \uparrow P | 0.0122 | 0.0041 | 0.0105 | 0.0088 | 0.0116 | 0.0134 | 0.0140 | \uparrow P | 0.1582 | 0.1423 | 0.1449 | 0.1433 | 0.1409 | 0.1517 | 0.1500 |
| | | \uparrow MAP | 0.0188 | 0.0046 | 0.0170 | 0.0128 | 0.0178 | 0.0218 | 0.0218 | \uparrow MAP | 0.1079 | 0.0923 | 0.0957 | 0.0934 | 0.0889 | 0.1010 | 0.0997 |
| | | \uparrow R | 0.0517 | 0.0141 | 0.0433 | 0.0369 | 0.0506 | 0.0589 | 0.0608 | \uparrow R | 0.0213 | 0.0193 | 0.0193 | 0.0186 | 0.0185 | 0.0199 | 0.0200 |

Table 5. Mean computation time (s) of FAIR measures per model.

| | Existing measures | | | | | | | | PUF | | | |
|----------|-------------------|---------|--------|-----------|---------|---------|---------|---------|----------|-----------|----------|-----------|
| | SD-P | SD-NDCG | Gini-P | Gini-NDCG | ME | MME | PEU | UF | Prec-Cos | Prec-Jacc | NDCG-Cos | NDCG-Jacc |
| Lastfm | <0.01 | <0.01 | <0.01 | <0.01 | 233.88 | 233.91 | 233.91 | 1297.78 | 5.30 | 9.79 | 4.32 | 9.56 |
| QK-video | 0.01 | 0.02 | 0.01 | 0.03 | 1867.81 | 1867.88 | 1867.88 | 873.52 | 20.68 | 35.63 | 16.01 | 35.10 |
| ML-10M | <0.01 | 0.01 | <0.01 | 0.02 | 871.17 | 871.20 | 871.20 | 1020.16 | 4.00 | 7.22 | 3.38 | 7.40 |
| ML-20M | <0.01 | <0.01 | <0.01 | <0.01 | 2020.60 | 2020.70 | 2020.70 | 1358.52 | 8.25 | 15.65 | 6.72 | 15.19 |

that half of the existing measures, i.e., ME, MME, PEU, and UF, are computationally expensive (> 4 mins.), while PUF is significantly faster (< 40 s), despite also being a pairwise measure. ME/MME/PEU/UF need additional extensive computation per pair, which makes them expensive: UF does nested pairwise comparisons, while ME/MME/PEU recompute the effectiveness score for each user pair. In short, PUF is computationally more efficient and thus more practical to compute than most existing FAIR measures.

4.3 Measure Agreement

An important aspect when comparing evaluation measures is how much they agree when their scores are used to rank models from best to worst. If one measure can be used to estimate the rank ordering given by another, there is no point in using both measures if we are only interested in ranking models. To study this, we compute Kendall's τ correlation for all EFF and FAIR measures.

Kendall’s τ can handle ties (unlike Spearman’s ρ) and is more robust to nonlinear relationships (unlike Pearson’s coefficient). If two measures have $\tau \geq 0.9$, we consider their rankings equivalent [43]. Fig. 1 shows the agreement between (i) EFF and FAIR measures, and (ii) among FAIR measures.

Agreement between EFF and FAIR measures. Across datasets, the agreement between FAIR and EFF measures varies from strong disagreement (e.g., SD with $\tau \in [-1, -0.52]$) to moderate-to-strong agreement (e.g., Gini with $\tau \in [0.43, 1]$). Our PUF consistently disagrees with EFF measures ($\tau \leq -0.71$), even if the disagreement is weaker for Lastfm, $\tau \in [-0.71, -0.33]$. As no FAIR measure consistently has $|\tau| \geq 0.9$ with any EFF measure for all datasets, their model orderings cannot be precisely inferred from that of EFF measures.

Agreement among FAIR measures. Regarding the agreement between PUF and existing FAIR measures, we find that only SD aligns with PUF ($\tau \geq 0.62$), while the rest tend to disagree or have a weak correlation with PUF, $\tau \in [-1, 0.24]$. Both UF and PUF consider user similarity, yet UF weakly correlates to PUF ($\tau \in [-0.14, 0.14]$) for all datasets except QK-video ($\tau = -0.52$). This weak relationship may be due to UF not considering item relevance, while PUF does. Even if we only compare the best model based on the measures instead of the model rankings, UF and PUF never agree. Thus, using UF and PUF to gauge fairness may differ in the conclusion on which model is the fairest. In contrast, despite not being similarity-based, SD correlates the strongest with PUF. Yet, most of its rankings are not equivalent to PUF ($\tau < 0.9$), which means that fairness evaluation with SD can lead to a conclusion that misaligns with PUF.

Overall, FAIR measures often disagree on model orderings with PUF, regardless of whether the measure accounts for user similarity. While SD has the most similar conclusions to PUF, it still does not give equivalent rankings to PUF.

4.4 Varying Relevance Score Distribution

It is important that a fairness measure captures differences in recommendation quality across users, as fairness is related to the disparity in recommendation (relevance). So, any change in EFF scores should also be reflected in the FAIR scores. We thus study how varying item relevance affects EFF and FAIR scores. All FAIR measure equations, except UF, explicitly account for effectiveness, but it is unknown how sensitive FAIR measures are to changes in EFF scores.

Procedure. The change of relevance scores is artificially done as follows. For all users, we start by recommending $k = 10$ relevant items (based on the test set). For users with more than k relevant items, we select the relevant items randomly. For users with fewer than k relevant items, we fill the remaining slots with random irrelevant items to ensure that each user receives exactly k items. In each iteration, we replace the recommendation of 10% of the users with all irrelevant items and compute the measures. We expect maximum fairness at the start (as all users have the maximum EFF scores⁸) and at the end (as all users have 0 EFF scores). We expect maximum unfairness when half the users get all

⁸ P@ k -based scores may not be optimal at the start as some users have $< k$ test items.

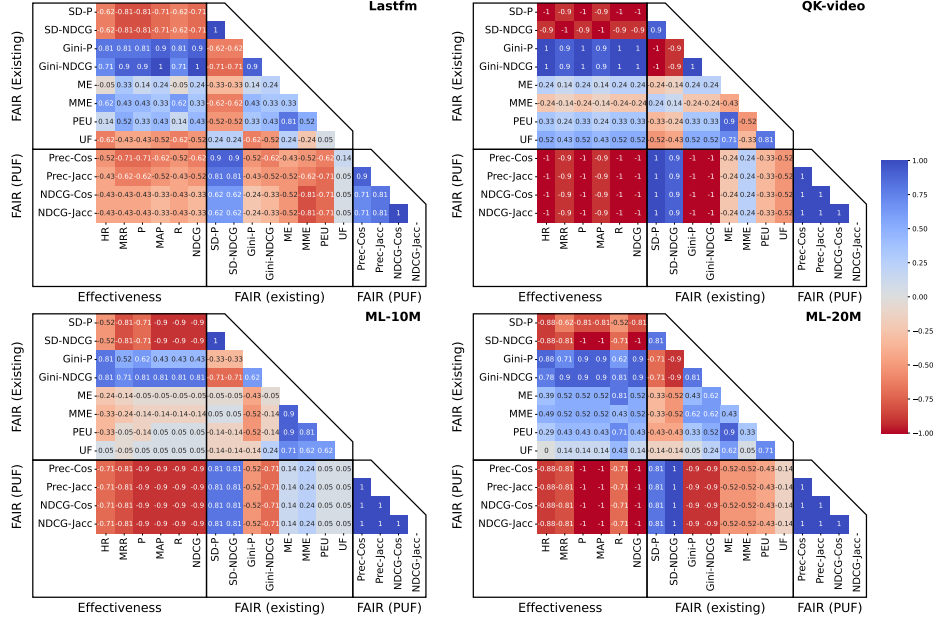


Fig. 1. Kendall's τ correlation between all measures (EFF, FAIR, and PUFs).

relevant items, and the rest get irrelevant items, as this leads to one of the most uneven EFF score distributions. We only compute PUF with sim_{Jacc} , as the model orderings given by PUF-*Cos are equivalent to PUF-*Jacc (§4.3). User similarities are computed with the (non-artificial) interactions in the train sets.

Results. Fig. 2 shows the results for QK-video and ML-20M, which represent the overall trends in all our datasets (the rest are in the online appendix). As expected, PUF has decreasing fairness followed by increasing fairness, shown by the inverted parabolas. This trend is more pronounced for the ML-* datasets, as the mean user pairwise similarity is higher than for the other datasets. Among the FAIR measures, only SD follows this expectation; the others show undesirable tendencies: as EFF drops, Gini and PEU notably become less fair, and ME also, but to a minor extent. The overall fairness drop in these measures is undesirable, as the scores closely resemble decreasing effectiveness instead of the disparity in the EFF scores distribution. Even worse, MME and UF are almost invariant to the change in recommendation effectiveness: MME tends to score extremely low to begin with, while UF does not depend on item relevance. Fig. 2 (first and last columns) also shows the possibility of having high effectiveness and low unfairness with PUF. Hence, optimising for PUF may not directly hurt utility.

To sum up, PUF and SD quantify fairness based on the disparity in recommendation effectiveness. All other FAIR measures ignore disparity; they just reflect effectiveness drops or are insensitive to changes in effectiveness. Next, we ask if changes in user similarity distribution are reflected in the FAIR measures.

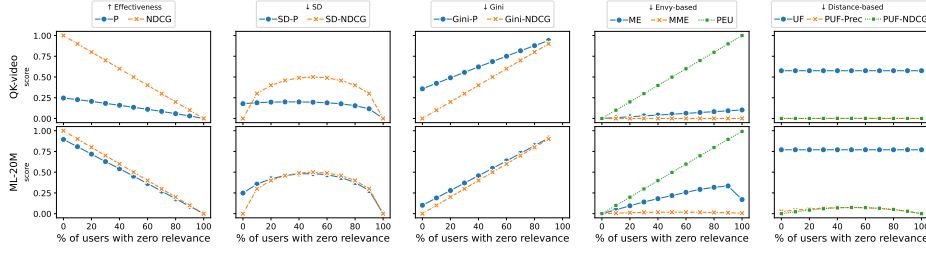


Fig. 2. Effectiveness (EFF) and fairness (FAIR) scores of QK-video and ML-20M, when artificially varying % of users with all irrelevant items (zero relevance), and the rest of the users receiving all relevant items. All PUF variants overlap. Gini is missing points at 100% users with zero relevance as it is undefined when each user has zero EFF scores.

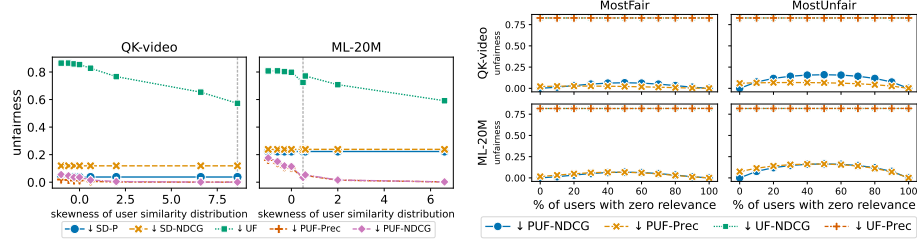


Fig. 3. Artificially varying the skewness of the user similarity distribution for QK-video and ML-20M. Vertical grey lines denote the skewness corresponding to sim_{Jacc} observed in the dataset. The distribution skewness differs across datasets.

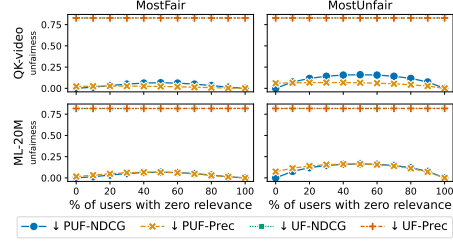


Fig. 4. Artificially varying the % of users with zero relevance for QK-video and ML-20M. Lower EFF score difference is assigned to user pairs with higher similarity (MostFair), and to lower similarity (MostUnfair). Both UF overlap.

4.5 Varying User Similarity Distribution

Individual user fairness is defined based on user similarity; it is important to know how user similarity affects fairness. While two/more recommender models should be evaluated under the same similarity distribution, a desirable individual user fairness measure should be able to distinguish a single model’s performance across different similarity distributions, which may arise from various ways of modelling user similarity. To this end, we investigate how the FAIR measures respond to (artificial) variations of user similarity distributions.

Procedure. User similarity distributions tend to be right-skewed (many dissimilar users) for random users [39]. Further, users are often dissimilar, as some users are new to the systems or do not engage much, leading to discrepancies in the number of interactions among users and potentially affecting user similarity. Considering the above, we create synthetic user similarity scores by sampling from the Weibull distribution [47], to model skewed distributions. It has been used to model user rating distributions and sampling user neighbour candidates in RSs [2, 20, 21]. Its probability density function is $p(x) = \lambda x^{\lambda-1} \exp(-x^\lambda)$.

To obtain various right- and left-skewed distributions to represent possible user similarity distributions, we set $\lambda \in \{0.5, 1, 2, 5, 10, 50\}$. We also sample from the normal distribution $\mathcal{N}(0, 1)$, which has zero skew (i.e., equal portions of user pairs with similarity below/above the mean). We then min-max normalise the sampled similarity values to rescale them in the $[0, 1]$ range and randomly assign them to user pairs. We analyse non-random assignment in §4.6.

While the user similarity is artificial, we use the actual recommendation lists and scores from the NCL model runs as they perform relatively well. To save computation time (Tab. 5), we only compute SD to represent all similarity-independent measures; theoretically, these measures will remain constant given no change in their input. We also compute PUF and UF, the two similarity-based measures. We compare the measure scores with the scores corresponding to the user similarity distribution observed in the datasets based on sim_{Jacc} (§4.2).

Results. Fig. 3 shows the results for QK-video and ML-20M (the rest have similar trends and are in the online appendix). We see that the similarity-based measures PUF and UF become fairer as skewness increases. Increasing skewness means a higher proportion of user pairs with low similarity, hence \downarrow PUF and \downarrow UF tend to be lower. Conversely, but as expected, the similarity-independent SD remains constant despite the change in skewness. For highly skewed similarities (skewness > 6), which is a realistic similarity distribution as seen in QK-video, SD-NDCG is somewhat unfair (≈ 0.2) for all datasets. Yet, PUF is almost perfectly fair (≈ 0). Therefore, simply using SD may lead to the underestimation of fairness. While we only compute SD here, we expect the other FAIR measures to show the same invariance as they are similarity-independent.

Between the two similarity-based measures, PUF is more sensitive to negatively skewed similarity distributions than UF. As skewness decreases, the mean user similarity increases at a slower rate. UF only considers user pairs above the mean-based similarity threshold; thus, the number of user pairs contributing to its log sum decreases slower than in the right-skewed distributions. Another concern of UF is its relatively high unfairness compared to PUF and SD, even when most users are dissimilar. This may be because UF computes the pairwise distance of the representation of the recommended items of a user pair. Minimising this distance is hard as each recommended item of a user must have a similar representation to each of the other user’s items, regardless of item relevance.

To sum up, PUF can distinguish fairness levels across various similarity distributions, while non-similarity-based measures cannot. This shows the strength of PUF over SD (and indirectly over other non-similarity-based measures, i.e., all FAIR measures except UF). We find that disregarding user similarity can also lead to misinterpretation of the fairness level. Next, we compare the two similarity-based measures and show the strengths of PUF over UF.

4.6 PUF and UF under Extreme Cases

We compare the similarity-based measures (PUF and UF) under extreme scenarios: can their scores reflect the difference in maximum and minimum fairness,

across differences in the recommendation quality? Given an artificial set of pairwise user similarities and an artificial set of pairwise EFF score differences, to simulate the fairest case (MostFair), we sort these values and assign a higher similarity value to user pairs with lower EFF score differences. Separately, we assign a higher similarity score to pairs with a higher EFF score difference to mimic the unfair case (MostUnfair). For MostFair, a desirable measure would score close to 0 (the fairest). For MostUnfair, it should exhibit an inverted U-shape when the EFF score distribution is varied (i.e., similar to Fig. 2), as the maximum unfairness happens when the EFF distribution is the most uneven.

Procedure. We use an artificial, right-skewed user similarity distribution sampled from the Weibull distribution with $\lambda = 2$ (§4.5). We use the P@ k and NDCG@ k scores per user from the artificial runs in §4.4. To compute PUF- and UF-NDCG, we assign the user similarities to user pairs following the sorted pairwise difference of NDCG, and likewise for PUF- and UF-Prec.

Results. Fig. 4 shows the results for QK-video and ML-20M (the rest have similar trends and are in the online appendix). The MostFair assignment yields PUF scores that are overall close to the fairest (0) across varying recommendation effectiveness, but UF remains constantly unfair (≈ 0.8) nonetheless. This emphasises a mismatch between fairness computed based on the disparity of item representation and based on EFF score differences. PUF scores are more unfair for the MostUnfair than the MostFair scenario, while UF is almost invariant to the change in similarity assignment between MostFair and MostUnfair. This confirms the insensitivity of UF towards varying recommendation effectiveness seen in Fig. 2. In the MostUnfair case, \downarrow PUF ranges in $[0, 0.2)$ (close to the fairest) for all datasets. This is because the overall user similarity is low, which can happen in real-world scenarios. Hence, the maximum unfairness is expected to be relatively low, as PUF relies on user similarity to quantify fairness.

To sum up, **PUF correctly scores close to the fairest for MostFair, while UF does not.** PUF is notably more unfair for the MostUnfair than the MostFair case, while UF is almost constant for both cases and across varying effectiveness. In both cases, UF overestimates effectiveness-based unfairness by constantly scoring much higher than PUF. Overall, PUF can reliably measure extreme (un)fairness.

5 Related Work

In §2, we overviewed existing individual user fairness measures. Here, we discuss work that studies RS fairness measures empirically [33–35, 37] and proposes pairwise individual fairness measures for ranking [14, 45]. Our work is close to [33], which studies item group fairness measures in RS, and [34, 35, 37], which study evaluation measures of individual item fairness. Similar to [33, 35, 37], we find that fairness measures may disagree in their model ordering, and that some measures are more sensitive than others, given decreasing effectiveness and disparity in the recommendations. Among the individual fairness measures in [34, 35, 37], the similarity criterion of individual fairness is often ignored. Recent

work [14, 45] also propose pairwise individual fairness measures, but these are for items, whereas our pairwise measure, PUF, is for individual users. Further, PUF considers user similarity, while the measure in [14] does not. The measure in [45] is similar to UF, as both employ thresholding of user similarities. Instead of applying a threshold, which can be arbitrary, PUF is weighted by user similarities, which introduces degrees of user similarity in the fairness computation.

There exists also work on user group fairness (e.g., [13, 57]) or counterfactual fairness (e.g., [9]). Most such work requires sensitive attributes (e.g., gender), but public recommendation datasets with sensitive attributes tend to lack user representations (e.g., only binary genders) [8, 16, 52], and grouping users may require discretising the attribute (e.g., age) [6]. We focus on attribute-free individual fairness, rather than group or counterfactual fairness, to better assess distribution across all individuals [22], which is often hidden in group fairness evaluation [14, 36]. Nevertheless, evaluating only individual fairness may not provide a comprehensive view of fairness. For example, PUF as an individual fairness measure could deem fair a system that simultaneously gives bad recommendations to all users who like niche items and gives good recommendations to all users who prefer mainstream items. Meanwhile, a group fairness measure could detect this disparity. As group and individual fairness measures serve different purposes, they should not replace each other [36].

6 Conclusion

Current evaluation measures of individual user fairness in RSs consider either only disparity in recommendation effectiveness or user similarity, but never both jointly. None of them aligns with both the individual fairness definition and user utility as a key objective of RSs. To address this issue, we introduced PUF, a novel evaluation measure that quantifies user fairness through pairwise difference in effectiveness scores and considers similarity between users. While PUF is simple, it is a novel, intuitive measure that is sensitive to various effectiveness disparities and user similarities, which sets it apart from existing metrics, addressing a crucial gap in RS fairness evaluation. We recommend using PUF to evaluate individual user fairness due to its alignment with a common definition of individual fairness, its computational efficiency, as well as its sensitivity to varying levels of user similarities and recommendation effectiveness in both typical and extreme cases. Future work includes integrating graded relevance or other ways of modelling user similarity (e.g., from user metadata), as well as testing the statistical power of the fairness measures.

Acknowledgments. The work is supported by the Algorithms, Data, and Democracy project (ADD-project), funded by the Villum Foundation and Velux Foundation. We thank the anonymous reviewers who have provided helpful feedback to improve earlier versions of the manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

References

1. Aalam, S.W., Ahanger, A.B., Bhat, M.R., Assad, A.: Evaluation of Fairness in Recommender Systems: A Review. In: Balas Valentina E., Sinha, G.R., Agarwal Basant, Sharma Tarun Kumar, Dadheech Pankaj, Mahrishi Mehul (eds.) *Emerging Technologies in Computer Engineering: Cognitive Computing and Intelligent IoT*. pp. 456–465. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-07012-9_39
2. Adamopoulos, P., Tuzhilin, A.: On over-specialization and concentration bias of recommendations: probabilistic neighborhood selection in collaborative filtering systems. In: *Proceedings of the 8th ACM Conference on Recommender Systems*. pp. 153–160. RecSys '14, Association for Computing Machinery, New York, NY, USA (2014). <https://doi.org/10.1145/2645710.2645752>
3. Amigó, E., Deldjoo, Y., Mizzaro, S., Bellogín, A.: A unifying and general account of fairness measurement in recommender systems. *Information Processing & Management* **60**(1), 103115 (1 2023). <https://doi.org/10.1016/J.IPM.2022.103115>
4. Biswas, A., Patro, G.K., Ganguly, N., Gummadi, K.P., Chakraborty, A.: Toward Fair Recommendation in Two-sided Platforms. *ACM Trans. Web* **16**(2) (12 2021). <https://doi.org/10.1145/3503624>
5. Bruun, S.B., Lioma, C., Maistro, M.: Recommending Target Actions Outside Sessions in the Data-poor Insurance Domain. *ACM Transactions on Recommender Systems* **3**(1), 1–24 (8 2024). <https://doi.org/10.1145/3606950>
6. Buyl, M., De Bie, T.: Inherent Limitations of AI Fairness. *Commun. ACM* **67**(2), 48–55 (1 2024). <https://doi.org/10.1145/3624700>
7. Cantador, I., Brusilovsky, P., Kuflik, T.: 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011). In: *Proceedings of the 5th ACM conference on Recommender systems*. RecSys 2011, ACM, New York, NY, USA (2011). <https://doi.org/10.1145/2043932.2044016>
8. Celma Herrada, O.: Music recommendation and discovery in the long tail. Ph.D. thesis, Universitat Pompeu Fabra (2009), <http://www.tdx.cat/TDX-0612109-190038>
9. Chen, W., Wu, Y., Zhang, Z., Zhuang, F., He, Z., Xie, R., Xia, F.: FairGap: Fairness-Aware Recommendation via Generating Counterfactual Graph. *ACM Trans. Inf. Syst.* **42**(4) (2 2024). <https://doi.org/10.1145/3638352>
10. Deshpande, M., Karypis, G.: Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems* **22**(1), 143–177 (1 2004). <https://doi.org/10.1145/963770.963776>
11. Do, V., Corbett-Davies, S., Atif, J., Usunier, N.: Online Certification of Preference-Based Fairness for Personalized Recommender Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* **36**(6), 6532–6540 (6 2022). <https://doi.org/10.1609/aaai.v36i6.20606>
12. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference* pp. 214–226 (2012). <https://doi.org/10.1145/2090236.2090255>

13. Ekstrand, M.D., Tian, M., Azpiaz, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D., Pera, M.S.: All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In: Friedler, S.A., Wilson, C. (eds.) *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. *Proceedings of Machine Learning Research*, vol. 81, pp. 172–186. PMLR (10 2018), <https://proceedings.mlr.press/v81/ekstrand18b.html>
14. Fabris, A., Silvello, G., Susto, G.A., Biega, A.J.: Pairwise Fairness in Ranking as a Dissatisfaction Measure. In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. pp. 931–939. WSDM '23, Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3539597.3570459>
15. Fu, Z., Xian, Y., Gao, R., Zhao, J., Huang, Q., Ge, Y., Xu, S., Geng, S., Shah, C., Zhang, Y., De Melo, G.: Fairness-Aware Explainable Recommendation over Knowledge Graphs. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* pp. 69–78 (7 2020). <https://doi.org/10.1145/3397271.3401051>
16. Harper, F.M., Konstan, J.A.: The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (2015). <https://doi.org/10.1145/2827872>
17. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S.: Neural collaborative filtering. In: *26th International World Wide Web Conference, WWW 2017*. pp. 173–182. *International World Wide Web Conferences Steering Committee* (2017). <https://doi.org/10.1145/3038912.3052569>
18. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 230–237. *SIGIR '99*, Association for Computing Machinery, New York, NY, USA (1999). <https://doi.org/10.1145/312624.312682>
19. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20(4), 422–446 (10 2002). <https://doi.org/10.1145/582415.582418>
20. Kermany, N.R., Zhao, W., Batsuuri, T., Yang, J., Wu, J.: Incorporating user rating credibility in recommender systems. *Future Generation Computer Systems* 147, 30–43 (2023). <https://doi.org/doi.org/10.1016/j.future.2023.04.029>
21. Kermany, N.R., Zhao, W., Yang, J., Wu, J.: ReInCre: Enhancing Collaborative Filtering Recommendations by Incorporating User Rating Credibility. In: *Web Information Systems Engineering*. pp. 64–72. Springer Singapore, Singapore (2020). https://doi.org/10.1007/978-981-15-3281-8_7
22. Lazovich, T., Belli, L., Gonzales, A., Bower, A., Tantipongpipat, U., Lum, K., Huszár, F., Chowdhury, R.: Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *Patterns* 3(8) (8 2022). <https://doi.org/10.1016/j.patter.2022.100568>

23. Leonhardt, J., Anand, A., Khosla, M.: User fairness in recommender systems. In: Companion Proceedings of the The Web Conference 2018. p. 101–102. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). <https://doi.org/10.1145/3184558.3186949>
24. Li, J., Ren, Y., Sanderson, M., Deng, K.: Explaining Recommendation Fairness from a User/Item Perspective. *ACM Trans. Inf. Syst.* (10 2024). <https://doi.org/10.1145/3698877>
25. Li, Y., Chen, H., Xu, S., Ge, Y., Tan, J., Liu, S., Zhang, Y.: Fairness in Recommendation: Foundations, Methods and Applications. *ACM Transactions on Intelligent Systems and Technology* (1 2023). <https://doi.org/10.1145/3610302>
26. Liang, D., Krishnan, R.G., Hoffman, M.D., Jebara, T.: Variational autoencoders for collaborative filtering. In: Proceedings of the 2018 World Wide Web Conference. p. 689–698. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). <https://doi.org/10.1145/3178876.3186150>
27. Lin, Z., Tian, C., Hou, Y., Zhao, W.X.: Improving Graph Collaborative Filtering with Neighborhood-enriched Contrastive Learning. In: WWW 2022 - Proceedings of the ACM Web Conference 2022. pp. 2320–2329. Association for Computing Machinery, Inc, Virtual Event, Lyon, France (4 2022). <https://doi.org/10.1145/3485447.3512104>
28. Lioma, C., Simonsen, J.G., Larsen, B.: Evaluation measures for relevance and credibility in ranked lists. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. p. 91–98. ICTIR '17, Association for Computing Machinery, New York, NY, USA (2017). <https://doi.org/10.1145/3121050.3121072>
29. Maistro, M., Chaves Lima, L., Grue Simonsen, J., Lioma, C.: Principled multi-aspect evaluation measures of rankings. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. p. 1232–1242. CIKM '21, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3459637.3482287>
30. Meng, Z., McCreadie, R., MacDonald, C., Ounis, I.: Exploring Data Splitting Strategies for the Evaluation of Recommendation Models. In: RecSys 2020 - 14th ACM Conference on Recommender Systems. pp. 681–686. Association for Computing Machinery, Inc, Virtual Event, Brazil (9 2020). <https://doi.org/10.1145/3383313.3418479>
31. Patro, G.K., Biswas, A., Ganguly, N., Gummadi, K.P., Chakraborty, A.: FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In: The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020. pp. 1194–1204. Association for Computing Machinery, Inc (4 2020). <https://doi.org/10.1145/3366423.3380196>
32. Pitoura, E., Stefanidis, K., Koutrika, G.: Fairness in rankings and recommendations: an overview. *VLDB Journal* **31**(3), 431–458 (5 2022). <https://doi.org/10.1007/s00778-021-00697-y>

33. Raj, A., Ekstrand, M.D.: Measuring Fairness in Ranked Results. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 726–736. ACM, New York, NY, USA (7 2022). <https://doi.org/10.1145/3477495.3532018>
34. Rampisela, T.V., Maistro, M., Ruotsalo, T., Lioma, C.: Evaluation Measures of Individual Item Fairness for Recommender Systems: A Critical Study. *ACM Trans. Recomm. Syst.* **3**(2) (11 2024). <https://doi.org/10.1145/3631943>, <https://doi.org/10.1145/3631943>
35. Rampisela, T.V., Maistro, M., Ruotsalo, T., Scholer, F., Lioma, C.: Relevance-aware individual item fairness measures for recommender systems: Limitations and usage guidelines. *ACM Trans. Recomm. Syst.* (Sep 2025). <https://doi.org/10.1145/3765624>
36. Rampisela, T.V., Maistro, M., Ruotsalo, T., Scholer, F., Lioma, C.: Stairway to fairness: Connecting group and individual fairness. In: Proceedings of the Nineteenth ACM Conference on Recommender Systems. p. 677–683. RecSys '25, Association for Computing Machinery, New York, NY, USA (2025). <https://doi.org/10.1145/3705328.3748031>
37. Rampisela, T.V., Ruotsalo, T., Maistro, M., Lioma, C.: Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 271–281. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3626772.3657832>
38. Rastegarpanah, B., Gummadi, K.P., Crovella, M.: Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. p. 231–239. WSDM '19, Association for Computing Machinery, New York, NY, USA (2019). <https://doi.org/10.1145/3289600.3291002>
39. Reisz, N., Servedio, V.D.P., Thurner, S.: Quantifying the impact of homophily and influencer networks on song popularity prediction. *Scientific Reports* **14**(1), 8929 (2024). <https://doi.org/10.1038/s41598-024-58969-w>
40. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian Personalized Ranking from Implicit Feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. pp. 452–461. UAI '09, AUAI Press, Arlington, Virginia, USA (2009). <https://doi.org/10.5555/1795114.1795167>
41. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. pp. 175–186. CSCW '94, Association for Computing Machinery, New York, NY, USA (1994). <https://doi.org/10.1145/192844.192905>
42. Smith, J.J., Beattie, L., Cramer, H.: Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective. In: Proceedings of the ACM Web Conference 2023. pp. 3648–3659.

- WWW '23, Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3543507.3583204>
43. Voorhees, E.M.: Evaluation by Highly Relevant Documents. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 74–82. SIGIR '01, Association for Computing Machinery, New York, NY, USA (2001). <https://doi.org/10.1145/383952.383963>
 44. Wang, X., He, X., Wang, M., Feng, F., Chua, T.S.: Neural graph collaborative filtering. In: SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 165–174. Association for Computing Machinery, Inc (7 2019). <https://doi.org/10.1145/3331184.3331267>
 45. Wang, X., Wang, W.H.: Providing Item-side Individual Fairness for Deep Recommender Systems. In: ACM International Conference Proceeding Series. vol. 22, pp. 117–127. Association for Computing Machinery (6 2022). <https://doi.org/10.1145/3531146.3533079>
 46. Wang, Y., Ma, W., Zhang, M., Liu, Y., Ma, S.: A Survey on the Fairness of Recommender Systems. *ACM Trans. Inf. Syst.* **41**(3), 1–43 (2 2023). <https://doi.org/10.1145/3547333>
 47. Weibull, W.: A Statistical Distribution Function of Wide Applicability. *Journal of Applied Mechanics* (1951), <https://hal.science/hal-03112318>
 48. Wu, K., Erickson, J., Wang, W.H., Ning, Y.: Equipping Recommender Systems with Individual Fairness via Second-Order Proximity Embedding. In: Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 171–175. ASONAM '22, IEEE Press (2023). <https://doi.org/10.1109/ASONAM55673.2022.10068703>
 49. Wu, Y., Cao, J., Xu, G.: Fairness in Recommender Systems: Evaluation Approaches and Assurance Strategies. *ACM Trans. Knowl. Discov. Data* **18**(1) (8 2023). <https://doi.org/10.1145/3604558>
 50. Wu, Y., Cao, J., Xu, G., Tan, Y.: TFROM: A Two-sided Fairness-Aware Recommendation Model for Both Customers and Providers. SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval pp. 1013–1022 (7 2021). <https://doi.org/10.1145/3404835.3462882>
 51. Xu, L., Lin, Z., Wang, J., Chen, S., Zhao, W.X., Wen, J.R.: Promoting two-sided fairness with adaptive weights for providers and customers in recommendation. In: Proceedings of the 18th ACM Conference on Recommender Systems. p. 918–923. RecSys '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3640457.3688169>
 52. Yuan, G., Yuan, F., Li, Y., Kong, B., Li, S., Chen, L., Yang, M., YU, C., Hu, B., Li, Z., Xu, Y., Qie, X.: Tenrec: A Large-scale Multipurpose Benchmark Dataset for Recommender Systems. *Advances in Neural Information Processing Systems* **35**, 11480–11493 (12 2022), <https://www.tencent.com/en-us/>

53. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems. *ACM Comput. Surv.* **55**(6) (12 2022). <https://doi.org/10.1145/3533380>
54. Zeng, H., He, Z., Yue, Z., McAuley, J., Wang, D.: Fair Sequential Recommendation without User Demographics. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 395–404. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3626772.3657703>
55. Zhao, W.X., Mu, S., Hou, Y., Lin, Z., Chen, Y., Pan, X., Li, K., Lu, Y., Wang, H., Tian, C., Min, Y., Feng, Z., Fan, X., Chen, X., Wang, P., Ji, W., Li, Y., Wang, X., Wen, J.R.: RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In: *International Conference on Information and Knowledge Management, Proceedings*. pp. 4653–4664. ACM, New York, NY, USA (2021). <https://doi.org/10.1145/3459637.3482016>
56. Zhao, Y., Wang, Y., Liu, Y., Cheng, X., Aggarwal, C.C., Derr, T.: Fairness and Diversity in Recommender Systems: A Survey. *ACM Trans. Intell. Syst. Technol.* (5 2024). <https://doi.org/10.1145/3664928>
57. Zhu, Z., Hu, X., Caverlee, J.: Fairness-Aware Tensor-Based Recommendation. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. pp. 1153–1162. CIKM '18, Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3269206.3271795>