Tutorial on Fairness of Machine Learning in Recommender Systems

Yunqi Li
Rutgers University
yunqi.li@rutgers.edu

Yingqiang Ge
Rutgers University
yingqiang.ge@rutgers.edu

Yongfeng Zhang
Rutgers University
yongfeng.zhang@rutgers.edu
Learning Objective

• Raise awareness on the importance of considering fairness in recommendation
• Get the background knowledge of fairness works in general machine learning
• Learn about the taxonomies of fairness concepts in recommendation
• Know about some datasets, evaluation protocols to assess fairness in recommendation
• Understand the challenges and opportunities of fairness research in recommendation
Outline

• Introduction:
  – Social Impact of Recommender System and Fairness
  – Motivation of Fairness in Recommender Systems
  – Relationship with AI Ethics
  – Beyond Ethics: a Utilitarian Perspective

• Fairness in Machine Learning:
  – Fairness in Classification
  – Fairness in Ranking

• Fairness in Recommendation:
  – Introduction
  – Taxonomy
  – Dataset and Evaluation
  – Challenge and Opportunity
Outline

• Introduction:
  – Social Impact of Recommender System and Fairness
  – Motivation of Fairness in Recommender Systems
  – Relationship with AI Ethics
  – Beyond Ethics: a Utilitarian Perspective

• Fairness in Machine Learning:
  – Fairness in Classification
  – Fairness in Ranking

• Fairness in Recommendation:
  – Introduction
  – Taxonomy
  – Dataset and Evaluation
  – Challenge and Opportunity
Recommender Systems are Everywhere

- For example

  - **E-commerce Systems**
    - Product Recommendation
  - **Social Networks**
    - Newsfeed Recommendation
    - Friend Recommendation
  - **Healthcare Systems**
    - Doctor Recommendation
    - Patient-doctor matching
  - **Online Entertainment Systems**
    - Movie/Video Recommendation
    - Music Recommendation
  - **Trip Planning Systems**
    - Hotel Recommendation
    - Air Ticket Recommendation
  - **Financial Applications**
    - Investment Recommendation
  - **Cyber-Physical Systems**
    - Driver Recommendation
    - Route Recommendation
  - **Talent Recruiting Systems**
    - Job Recommendation
    - Candidate Recommendation
Social Impacts of Recommender Systems

• Recommender Systems are far more than just information seeking tools
  – They control how resources are allocated among different parties
    • Resources can be exposure opportunities, products, jobs, information, etc.
    • Usually RS works in two-sided markets/environments [1]

The *Prosumer* Paradigm:

*Consumers – items – Producers*

Buyers – Goods – Sellers
Freelancer – Jobs – Employers
Borrowers – Money – Lenders
Passengers – Services – Drivers

Why Fairness in RecSys? Resources Could be Limited

Recommendation slot positions are limited, which producers’ items should be recommended and get the exposure opportunity to users?

User attention is a limited resource, whose twite should get exposure on the timeline?

Passengers are limited, which driver should get the task and make money?

Interview opportunities are limited, which candidate(s) should get an interview opportunity?
Why Fairness in RecSys? Data Could be Biased

- Most RecSys models are ML models trained on some training data
  - Training data may encode social bias
  - Recommendation models may learn "shotcuts" for decision making
  - Model may echo or even reinforce the bias in training data

Just data debias is not enough because AI doesn't know which are sensitive features (e.g., gender) and the approach of fairness is effect-based [2]. Explicit intervention on model is needed.
Potential Consequences of Unfairness in RecSys

Information Asymmetry
Knowing a piece of valuable information (e.g., a job opportunity) could change one's life.

Matthew Effect
Advantaged users, items, or groups get further propagated by recommendations, sometimes not because of their good quality but because the recommendation model is dominated by their data.

Echo Chambers
Unfair, undiversified exposure of news, messages, tweets, etc. may create echo chamber. Makes it difficult to explore new ideas and opinions different from one's own. Makes people feel like the whole world thinks the same way as they think. May even reinforce someone's extremist ideas.
Fairness in RecSys: an AI Ethics Perspective

- Recommender systems as responsible AI
  - Provide fair decisions for users, item providers, and platform

7 Principles of EU GDPR Regulation

- Fairness often appears together with other responsible AI perspectives
  - e.g., transparency/explainability (honesty) of algorithmic decisions is the foundation of fairness
Fairness in RecSys: Beyond Ethics, a Utilitarian Perspective

- RecSys platforms should consider fairness for the sake of themselves
  - Not only for legal regulations, but for the sustainable/long-term development of the platform

An e-commerce example
Big retailers vs. Small retailers

If products from small retailers (e.g., family workshops) do not have fair exposure opportunity by e-commerce recommender system, they may eventually leave since they cannot survive in the platform, making the platform unsustainable.

A social network example
Star accounts vs. Grassroot accounts

Videos from famous accounts (e.g., a film star) usually get more attention, but if videos created by grassroot accounts do not have any exposure opportunity to users, they may leave the platform, making the platform's contents less diversified and even boring.
What exactly is Fairness in RecSys?

Many different perspectives:

• Group Fairness vs. Individual Fairness
• User Fairness vs. Item Fairness
• Associative Fairness vs. Causal Fairness
• Single-sided Fairness vs. Multi-sided Fairness
• Static Fairness vs. Dynamic Fairness
• Short-term Fairness vs. Long-term Fairness
• Populational Fairness vs. Personalized Fairness
Outline

• Introduction:
  – Social Impact of Recommender System and Fairness
  – Motivation of Fairness in Recommender Systems
  – Relationship with AI Ethics
  – Beyond Ethics: a Utilitarian Perspective

• Fairness in Machine Learning:
  – Fairness in Classification
  – Fairness in Ranking

• Fairness in Recommendation:
  – Introduction
  – Taxonomy
  – Dataset and Evaluation
  – Challenge and Opportunity
Fairness in Machine Learning — Motivations

- Fairness matters because it has impact on everyone’s benefit.
Fairness in Machine Learning — Causes

Data Bias
- Statistical Bias: non-random sample; record error
- Historical Bias: biased decision
- ...

Algorithmic Bias
- Ranking Bias: exposure allocation
- Evaluation Bias: inappropriate benchmarks
- ...

User Interaction
- Behavioral Bias
- Presentation Bias
- ...

Data
- Historical Bias
- Social Bias
- ...

Algorithm
- Popularity Bias
- Ranking Bias
- ...

## Fairness in Machine Learning — Definitions

<table>
<thead>
<tr>
<th>Type of Fairness</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Fairness</td>
<td>Counterfactual fairness</td>
</tr>
<tr>
<td>Group Fairness</td>
<td>Statistical parity $P(\hat{Y}</td>
</tr>
<tr>
<td>Subgroup Fairness</td>
<td>Fairness holds over a large collection of subgroups defined by a class of functions</td>
</tr>
</tbody>
</table>
### Fairness in Machine Learning — Methods

<table>
<thead>
<tr>
<th>Pre-processing</th>
<th>In-processing</th>
<th>Post-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Try to transform the data so that the underlying discrimination is removed.</td>
<td>Try to modify the learning algorithms to remove discrimination during the model training process.</td>
<td>Perform after training by accessing a holdout set which was not involved during the training of the model.</td>
</tr>
</tbody>
</table>

Fairness in Machine Learning — Evaluation

The evaluation usually depends on the requirement of fairness.

- **Disparate Impact**: $P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$
  - Evaluation: $DI = |P(\hat{y} = 1|z = 0) - P(\hat{y} = 1|z = 1)|$

- **False Positive Rate**: $P(\hat{y} \neq y|y = -1, z = 0) = P(\hat{y} \neq y|y = -1, z = 1)$
  - Evaluation: $DM_{FPR} = P(\hat{y} \neq y|z = 0, y = -1) - P(\hat{y} \neq y|z = 1, y = -1)$

- **False Negative Rate**: $P(\hat{y} \neq y|y = 1, z = 0) = P(\hat{y} \neq y|y = 1, z = 1)$
  - Evaluation: $DM_{FNR} = P(\hat{y} \neq y|z = 0, y = 1) - P(\hat{y} \neq y|z = 1, y = 1)$
Fairness in Machine Learning — Basic tasks

Fairness in Classification

Fairness in Ranking
**Objective:** Avoid unethical interference of protected attributes into the decision-making process.

**Binary Classification:** Fairness metrics can be expressed by *rate constraints* to regularize the classifier’s positive or negative rates over different protected groups.

- **Statistical parity:**
  \[ P(\hat{Y} = 1 | Z = 0) = P(\hat{Y} = 1 | Z = 1) \]

- **Equality of Opportunity:**
  \[ P(\hat{Y} = 1 | Z = 0, Y = 1) = P(\hat{Y} = 1 | Z = 1, Y = 1) \]

...
Fairness in Classification — Method

Pre-processing: [3][4][5][6]...

Pros:
The transformed dataset can be used to train any downstream algorithm.

Cons:
Unpredictable loss in accuracy;
May not remove unfairness on the test data.

In-processing: [7][8][9][10]...

Pros:
Good performance;
May higher flexibility for the trade-off.

Cons:
A non-convex optimization problem and not guarantee optimality.

Post-processing: [11][12][13]...

Pros:
No need to modify classifier;
Relatively good performance especially fairness measures.

Cons:
Cannot be used in cases where sensitive feature information is unavailable.
## Fairness in Classification — Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Aera</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCI adult dataset</td>
<td>48,842 income records</td>
<td>Social</td>
<td>[14]</td>
</tr>
<tr>
<td>German credit dataset</td>
<td>1,000 credit records</td>
<td>Financial</td>
<td>[15]</td>
</tr>
<tr>
<td>Pilot parliaments benchmark</td>
<td>1,270 images</td>
<td>Facial images</td>
<td>[16]</td>
</tr>
<tr>
<td>WinoBias</td>
<td>3,160 sentences</td>
<td>Coreference resolution</td>
<td>[17]</td>
</tr>
<tr>
<td>Communities and crime</td>
<td>1,994 crime records</td>
<td>Social</td>
<td>[18]</td>
</tr>
<tr>
<td>COMPAS dataset</td>
<td>18,610 crime records</td>
<td>Social</td>
<td>[19]</td>
</tr>
<tr>
<td>Recidivism in juvenile justice</td>
<td>4,753 crime records</td>
<td>Social</td>
<td>[20]</td>
</tr>
<tr>
<td>Diversity in faces dataset</td>
<td>1 million images</td>
<td>Facial images</td>
<td>[21]</td>
</tr>
</tbody>
</table>

Fairness in Classification

- **Fairness Concerns**: Introduce a flexible constraint-based framework to enable the design of fair margin-based classifiers.

- **Fairness Definitions**:
  - No disparate treatment: \( P(\hat{y}|x, z) = P(\hat{y}|x) \)
  - No disparate impact: \( P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1) \)
  - No disparate mistreatment:
    - False positive rate: \( P(\hat{y} \neq y|y = -1, z = 0) = P(\hat{y} \neq y|y = -1, z = 1) \)
    - False negative rate: \( P(\hat{y} \neq y|y = 1, z = 0) = P(\hat{y} \neq y|y = 1, z = 1) \)
    - …
Fairness in Classification

- **Method:**

\[
\begin{align*}
\text{minimize} & \quad L(\theta) \\
\text{subject to} & \quad P(\cdot|z=0) = P(\cdot|z=1) \\
\end{align*}
\]  

Classifier loss function  
Fairness constraints

- No disparate impact:  
\[P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)\]

\[
\text{Cov}_{DI}(z, d_\theta(x)) = \mathbb{E}[(z - \bar{z})d_\theta(x)] - \mathbb{E}[(z - \bar{z})]\bar{d}_\theta(x) \approx \frac{1}{N} \sum_{(x,z) \in \mathcal{D}} (z - \bar{z}) d_\theta(x)
\]

- Objective function for no disparate impact:

\[
\begin{align*}
\text{minimize} & \quad L(\theta) \\
\text{subject to} & \quad \frac{1}{N} \sum_{(x,z) \in \mathcal{D}} (z - \bar{z}) d_\theta(x) \leq c \\
& \quad \frac{1}{N} \sum_{(x,z) \in \mathcal{D}} (z - \bar{z}) d_\theta(x) \geq -c
\end{align*}
\]

Fairness in Classification

• Simulate **disparate impact** in classification outcomes.

• Generate **two synthetic datasets** with different levels of correlation between a sensitive attribute and class labels.

• Train **logistic regression** classifiers on both the datasets.

---

Fairness in Classification

- Lingxiao Huang and Nisheeth Vishnoi. “Stable and Fair Classification.” ICML 2019
Fairness in Ranking — Introduction

**List-wise** definitions for fairness: depend on the entire list of results for a given query

Unsupervised criteria: the average **exposure** near the top of the ranked list to be **equal for different groups** [71][72][75]

Supervised criteria: the average **exposure** for a group to be proportional to the average **relevance** of that group’s results to the query [65][67]
Fairness in Ranking — Method

## Fairness in Ranking — Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Sensitive Features</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>AirBnB</td>
<td>10,201 houses</td>
<td>gender of host</td>
<td>[29]</td>
</tr>
<tr>
<td>COMPAS</td>
<td>7,214 people</td>
<td>gender, race</td>
<td>[30]</td>
</tr>
<tr>
<td>CS departments</td>
<td>51 departments</td>
<td>department size, geographic region</td>
<td>[31]</td>
</tr>
<tr>
<td>Engineering students</td>
<td>5 queries, 650 students per query</td>
<td>gender, high school type</td>
<td>[32]</td>
</tr>
<tr>
<td>SAT</td>
<td>1.6M students</td>
<td>gender</td>
<td>[33]</td>
</tr>
<tr>
<td>German credit</td>
<td>1,000 people</td>
<td>gender, age</td>
<td>[34]</td>
</tr>
<tr>
<td>Forbes richest U.S.</td>
<td>400 people</td>
<td>gender</td>
<td>[35]</td>
</tr>
<tr>
<td>XING</td>
<td>40 candidates</td>
<td>gender</td>
<td>[36]</td>
</tr>
</tbody>
</table>
Fairness in Ranking

- **Fairness Concerns:** A conceptual and computational framework that allows the formulation of fairness constraints on rankings in terms of **exposure allocation**.

- Job seeker example: a small difference in **relevance** can lead to a large difference in **exposure** (an opportunity) for the group of females.

Fairness in Ranking

- **Method:** \( r = \text{argmax}_r U(r|q) \) s.t. \( r \) is fair

- **Exposure** for a document \( d_i \) under a probabilistic ranking \( P \) as:

  \[
  \text{Exposure}(d_i|P) = \sum_{j=1}^{N} P_{i,j} v_j \\
  \text{Exposure}(G_k|P) = \frac{1}{|G_k|} \sum_{d_i \in G_k} \text{Exposure}(d_i|P)
  \]

- **Demographic Parity Constraints:**

  \[
  \text{Exposure}(G_0|P) = \text{Exposure}(G_1|P) \iff f^T P v = 0
  \]

  (with \( f_i = \frac{1}{|G_0|} \mathbb{1}_{d_i \in G_0} - \frac{1}{|G_1|} \mathbb{1}_{d_i \in G_1} \))
Fairness in Ranking

- Figure (a) is optimal unfair ranking that maximizes DCG.
- Figure (b) is optimal fair ranking under demographic parity.
- Compared to the DCG of the unfair ranking, the optimal fair ranking has slightly lower utility with a DCG.
Fairness in Ranking

Outline

• Introduction:
  – Social Impact of Recommender System and Fairness
  – Motivation of Fairness in Recommender Systems
  – Relationship with AI Ethics
  – Beyond Ethics: a Utilitarian Perspective

• Fairness in Machine Learning:
  – Fairness in Classification
  – Fairness in Ranking

• Fairness in Recommendation:
  – Introduction
  – Taxonomy
  – Dataset and Evaluation
  – Challenge and Opportunity
Fairness in Recommendation — Motivation

- Recommender systems are gaining critical impacts on human decision making.
Fairness in Recommendation — Motivation

• It is crucial to address the potential unfairness problems in recommendations.
Fairness in Recommendation — Challenges

- More Perspectives
- Multiple Models And Goals
- Extreme Data Sparsity
- Dynamics
Taxonomies

- Group vs. Individual
- User vs. Item
- Association vs. Causality
- Single-sided vs. Multi-sided
- Static vs. Dynamic
Group Fairness vs. Individual Fairness

Group fairness requires that the protected groups should be treated similarly to the advantaged group.

Group = Male
Advantaged

Group = Female
Protected

Require the same acceptance rate for both male and female job applicants
Group Fairness vs. **Individual Fairness**

- Individual fairness requires that the similar individual should be treated similarly.

Group Fairness in Recommendation

- **Fairness concerns:** The unfair recommendation quality between user groups with different activity levels, e.g., number of interactions.
- Unfairness of current recommender systems:
  - Active users only account for a small proportion of users.
  - The average recommendation quality on the small group (active) is significantly better than that on the remaining majority of users (inactive) for all baselines.

Group Fairness in Recommendation

**Fairness-aware Algorithm:** A re-ranking method with user-oriented group fairness constrained on the recommendation lists generated from any base recommender algorithm.

\[
\max_{W_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{N} W_{ij} S_{i,j} \quad \text{Preference of user } i \text{ in terms of item } j
\]
\[
\text{s.t. } \quad \text{UGF}(Z_1, Z_2, W) < \varepsilon \quad \text{Fairness constraint}
\]
\[
\sum_{j=1}^{N} W_{ij} = K, W_{ij} \in \{0, 1\} \quad \text{Top-K list}
\]

**Experiment Results:** Improve fairness; Improve recommendation quality of overall and disadvantaged users. However, the performance of advantaged users is reduced to satisfy our fairness requirement.

<table>
<thead>
<tr>
<th>BiasedMF</th>
<th>F1 Orig. Fair</th>
<th>F1 Fair</th>
<th>Adv Orig.</th>
<th>Adv Fair</th>
<th>Disadv Orig.</th>
<th>Disadv Fair</th>
<th>UGF Orig.</th>
<th>UGF Fair</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiasedMF</td>
<td>14.27</td>
<td>30.68</td>
<td>12.77</td>
<td>17.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiasedMF</td>
<td><strong>15.06</strong></td>
<td>19.18</td>
<td><strong>14.68</strong></td>
<td><strong>4.50</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiasedMF</td>
<td>43.25</td>
<td>67.79</td>
<td>41.00</td>
<td>26.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiasedMF</td>
<td><strong>43.97</strong></td>
<td>52.51</td>
<td><strong>43.19</strong></td>
<td><strong>9.32</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Group Fairness in Recommendation

Other works:

• Fu et al. [37] require to impair the group unfairness problem in the context of explainable recommendation over knowledge graphs with a fairness constrained approach.

• Both [38] and [39] categorize different types of multi-stakeholder platforms and the different group fairness properties they desired.
Individual Fairness in Recommendation

- **Fairness concerns**: the position bias which leads to disproportionately less attention being paid to low-ranked subjects.

- No single ranking can achieve individual attention fairness.

- **Equity of Amortized Attention**: A sequence of rankings \{1, 2, \ldots, m\} offer equity of amortized attention if each subject \(u_i\) receives cumulative attention proportional to her cumulative relevance:

\[
\frac{\sum_{l=1}^{m} a_{i1}^l}{\sum_{l=1}^{m} r_{i1}^l} = \frac{\sum_{l=1}^{m} a_{i2}^l}{\sum_{l=1}^{m} r_{i2}^l}, \forall u_{i1}, u_{i2}
\]

Individual Fairness in Recommendation

- **Method (Offline optimization):**
  
  \[
  \text{minimize } \sum_i |A_i - R_i| \quad \text{Fairness}
  \]
  
  subject to \[\text{NDCG-quality}@k(\rho^j, \rho^{j*}) \geq \theta, \ j = 1, \ldots, m.\quad \text{Ranking quality}\]

- **Experiment Results:**
  - **Improving equity of attention is crucial:** the discrepancy between the attention received and the deserved attention can be substantial.
  - Improving equity of attention can often be done **without sacrificing much quality** in the rankings.

Individual Fairness in Recommendation

Other Works:

• Patro et al. [40] view individual fairness from both producers and customers sides, and response to the question of the long-term sustainability of two-sided platforms.

• Li et al. [80] consider personalized fairness for users in recommendations, i.e., users’ personalized demands for fairness. For example, some users may care more about gender, while others care more about age.
User Fairness vs. Item Fairness

Fairness on user side: Fairness requirements in recommender systems may come from users.
User Fairness vs. Item Fairness

• Fairness on Item side: Fairness requirements in recommender systems may come from items (Products/Producers).

• For example, we search for “phone case” but the system ranks accessories for iPhone on top but quite few for other brands, which is an item-side unfairness.
User Fairness in Recommendation

- **Group Recommendation**: recommend items to groups of users whose preferences can be different from each other.

- **Fairness Concerns**: maximize the satisfaction of each group member while minimizing the unfairness (the imbalance of user utilities inside the group) between them.

- **Fairness Definitions**:
  - Least Misery: $F_{LM}(g, I) = \min \{U(u, I), \forall u \in g\}$
  - Variance: $F_{Var}(g, I) = 1 - Var\{U(u, I), \forall u \in g\}$
  - Jain’s Fairness: $F_J(g, I) = \frac{(\sum_{u \in g} U(u, I))^2}{|U| \cdot \sum_{u \in g} U(u, I)^2}$
  - Min-Max Ratio: $F_M(g, I) = \frac{\min \{U(u, I), \forall u \in g\}}{\max \{U(u, I), \forall u \in g\}}$
User Fairness in Recommendation

- **Method:**
  - The **Social Welfare** ($SW(g, I)$): overall utility of all users inside the group $g$ given a group recommendation $I$.
  - The **Fairness** ($F(g, I)$): a function of $U(u, I), \forall u \in g, \forall I$.
  - Multi-Objective Optimization: $\lambda \cdot SW(g, I) + (1 - \lambda) \cdot F(g, I)$

- **Experiment Results:** The results indicate that considering fairness can improve the quality of group recommendation.

<table>
<thead>
<tr>
<th>$\lambda$, RG</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>F@K</td>
<td>0.0260</td>
<td>0.0817</td>
<td>0.0877</td>
<td>0.0953</td>
<td>0.1019</td>
<td>0.1041</td>
<td>0.1046</td>
<td>0.1053</td>
<td>0.1058</td>
<td>0.1062</td>
<td>0.1062</td>
</tr>
<tr>
<td>NDCG@K</td>
<td>0.0697</td>
<td>0.2200</td>
<td>0.2287</td>
<td>0.2334</td>
<td>0.2394</td>
<td>0.2423</td>
<td>0.2440</td>
<td>0.2421</td>
<td>0.2459</td>
<td>0.2478</td>
<td>0.2476</td>
</tr>
</tbody>
</table>
User Fairness in Recommendation

Other Works:

- Leonhardt et al. [41] quantify the user unfairness caused by the post-processing algorithms which have the original goal of improving diversity in recommendations.

- Abdollahpouri et al. [42] see the problem from the users’ perspective with finding how popularity bias causes the recommendations to deviate from what the user expects to get from the recommender system.
Item Fairness in Recommendation

- **Fairness Concerns:** focus on the risk to groups of items from being under-recommended
- **Pairwise Accuracy:**
  \[ P(g(f_\theta(q, v_j)) > g(f_\theta(q, v_{j'}))) | y_{q,j} > y_{q,j'}, j, j' \in \mathcal{R}_q) \]
- **Pairwise Fairness:**
  \[ P(c_q(j, j') | y_{q,j} > y_{q,j'}, s_j = 0, z_{q,j} = \tilde{z}) = P(c_q(j, j') | y_{q,j} > y_{q,j'}, s_j = 1, z_{q,j} = \tilde{z}), \forall \tilde{z} \]
- **Inter-Group Pairwise Fairness:**
  \[ P(c_q(j, j') | y_{q,j} > y_{q,j'}, s_j = 0, s_{j'} = 1, z_{q,j} = \tilde{z}) = P(c_q(j, j') | y_{q,j} > y_{q,j'}, s_j = 1, s_{j'} = 0, z_{q,j} = \tilde{z}), \forall \tilde{z} \]
- **Intra-Group Pairwise Fairness:**
  \[ P(c_q(j, j') | y_{q,j} > y_{q,j'}, s_j = s_{j'} = 0, z_{q,j} = \tilde{z}) = P(c_q(j, j') | y_{q,j} > y_{q,j'}, s_j = s_{j'} = 1, z_{q,j} = \tilde{z}), \forall \tilde{z} \]
Item Fairness in Recommendation

• **Method:**

\[
\min_{\theta} \left( \sum_{(q, j, y, z) \in D} L(f_\theta(q, v_j), (y, z)) \right) + |\text{Corr}\varphi(A, B)|
\]

- Recommender Loss
- Fairness Penalty

\[
A = (g(f_\theta(q, v_j)) - g(f_\theta(q, v_{j'})))(y - y')
\]
\[
B = (s_j - s_{j'})(y - y')
\]

• **Experiment Results:**
  - The subgroup items are significantly under-ranked relative to the non-subgroup items.
  - The regularization effectively closes the gap in the inter-group pairwise fairness metric.

Beutel, Alex, et al. “Fairness in Recommendation Ranking through Pairwise Comparisons” SIGKDD’19
Item Fairness in Recommendation

Other Works:

- Many works about the **popularity bias** problem in recommendations.

- Often solved by increasing the number of recommended unpopular items (long-tail items) or otherwise the overall catalog coverage in these researches [43-45].

Associative Fairness vs. Causal Fairness

Find the **discrepancy of statistical metrics** between individuals or sub-populations.

In **binary classification**, fairness metrics can be represented by regularizing the classifier's positive or negative rates over different protected groups.
Associative Fairness vs. Causal Fairness

- Fairness cannot be well assessed only based on association notions [46-49].

- Difference:
  - Reason about the causal relations between the protected features and the model outcomes.
  - Leverage prior knowledge about the world structure in the form of causal models, help to understand the propagation of variable changes in the system.
Causal Fairness

- **Methods:**
  - Intervention
  - Counterfactual

- **Causal graph:** A directed acyclic graph which is used to capture the causal relations between variables, where nodes represent variables and directed edges represent a causal influence between the corresponding variables.

- A “what if” statement in which the “if” portion is *unreal* or *unrealized*, is known as a counterfactual.
Causal Fairness

• Disparate Impact:
  – Total Effect: \( TE_{a_1,a_0} (y) = P(y_{a_1}) - P(y_{a_0}) \)
  – Effect of Treatment on the Treated: \( ETT_{a_1,a_0} (y) = P(y_{a_1} | a_0) - P(y | a_0) \)
  – ...

• Disparate Treatment:
  – Direct Effect: the causal effect along the causal path from the sensitive feature to the final decision
  – Indirect Effect: the causal effect along the causal path through proxy features
  – Path-Specific Effect: the causal effect over specific paths.

Counterfactual fairness

• Counterfactual fairness is an individual-level causal-based fairness notion. It requires that for any possible individual, the predicted result of the learning system should be the same in the counterfactual world as in the real world.
Associative Fairness in Recommendation

• **Fairness Concerns:** study fairness in collaborative-filtering recommender systems; propose four new metrics that address different forms of unfairness.

• **Fairness Definitions:**
  - Value Fairness: \( U_{val} = \frac{1}{n} \sum_{j=1}^{n} \left| (E_{g \cdot y} - E_{g \cdot r}) - (E_{\neg \cdot y} - E_{\neg \cdot r}) \right| \)
  - Absolute Fairness: \( U_{abs} = \frac{1}{n} \sum_{j=1}^{n} \left| E_{g \cdot y} - E_{g \cdot r} - E_{\neg \cdot y} + E_{\neg \cdot r} \right| \)
  - Underestimation unfairness: \( U_{under} = \frac{1}{n} \sum_{j=1}^{n} \left( \max\{0, E_{g \cdot r} - E_{g \cdot y}\} - \max\{0, E_{\neg \cdot r} - E_{\neg \cdot y}\} \right) \)
  - Overestimation unfairness: \( U_{over} = \frac{1}{n} \sum_{j=1}^{n} \left( \max\{0, E_{g \cdot r} - E_{g \cdot y}\} - \max\{0, E_{\neg \cdot r} - E_{\neg \cdot y}\} \right) \)
Associative Fairness in Recommendation

• **Method:** $\min_{P,Q,u,v} J(P,Q,u,v) + U$

  Loss for recommender model  Fairness constraint

• **Experiment Results:** the experiments on synthetic and real data show that minimization of these forms of unfairness is possible with no significant increase in reconstruction error.

<table>
<thead>
<tr>
<th>Unfairness</th>
<th>Error</th>
<th>Value</th>
<th>Absolute</th>
<th>Underestimation</th>
<th>Overestimation</th>
<th>Non-Parity</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.887 ± 1.9e-03</td>
<td>0.234 ± 6.3e-03</td>
<td>0.126 ± 1.7e-03</td>
<td>0.107 ± 1.6e-03</td>
<td>0.153 ± 3.9e-03</td>
<td>0.036 ± 1.3e-03</td>
</tr>
<tr>
<td>Value</td>
<td>0.886 ± 2.2e-03</td>
<td><strong>0.223 ± 6.9e-03</strong></td>
<td>0.128 ± 2.2e-03</td>
<td><strong>0.102 ± 1.9e-03</strong></td>
<td><strong>0.148 ± 4.9e-03</strong></td>
<td>0.041 ± 1.6e-03</td>
</tr>
<tr>
<td>Absolute</td>
<td>0.887 ± 2.0e-03</td>
<td>0.235 ± 6.2e-03</td>
<td><strong>0.124 ± 1.7e-03</strong></td>
<td>0.110 ± 1.8e-03</td>
<td>0.151 ± 4.2e-03</td>
<td>0.023 ± 2.7e-03</td>
</tr>
<tr>
<td>Under</td>
<td>0.888 ± 2.2e-03</td>
<td>0.233 ± 6.8e-03</td>
<td>0.128 ± 1.8e-03</td>
<td><strong>0.102 ± 1.7e-03</strong></td>
<td>0.156 ± 4.2e-03</td>
<td>0.058 ± 9.3e-04</td>
</tr>
<tr>
<td>Over</td>
<td><strong>0.885 ± 1.9e-03</strong></td>
<td>0.234 ± 5.8e-03</td>
<td><strong>0.125 ± 1.6e-03</strong></td>
<td>0.112 ± 1.9e-03</td>
<td><strong>0.148 ± 4.1e-03</strong></td>
<td>0.015 ± 2.0e-03</td>
</tr>
<tr>
<td>Non-Parity</td>
<td>0.887 ± 1.9e-03</td>
<td>0.236 ± 6.0e-03</td>
<td>0.126 ± 1.6e-03</td>
<td>0.110 ± 1.7e-03</td>
<td>0.152 ± 3.9e-03</td>
<td><strong>0.010 ± 1.5e-03</strong></td>
</tr>
</tbody>
</table>

Yao, Sirui, and Bert Huang. “Beyond Parity: Fairness Objectives for Collaborative Filtering” NIPS’17
Causal Fairness in Recommendation

- **Fairness Concerns**: Counterfactual fairness for users in recommendations.

- **Definition**: A recommender model is *counterfactually fair* if for any possible user \( u \) with features \( X = x \) and \( Z = z \), for all \( L \), and for any value \( z' \) attainable by \( Z \):

\[
P(L_Z | X = x, Z = z) = P(L_{z'} | X = x, Z = z)
\]

Sensitive features

Insensitive features

Top-N recommendation list

for user \( u \) with sensitive features \( z \)
Causal Fairness in Recommendation

- **Method:** Generate feature independent user embeddings through *adversary learning*.
  - **Filter Module:** filter the information about sensitive features from user embeddings
  - **Discriminator module:** predict the sensitive features from the learned user embeddings.

- **Experiment Results:**
  - Improve fairness
  - A little sacrifice on recommendation performance

<table>
<thead>
<tr>
<th>MoiveLens</th>
<th>AUC-G</th>
<th>AUC-A</th>
<th>AUC-O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orig.</td>
<td>0.7697</td>
<td>0.8428</td>
<td>0.6024</td>
</tr>
<tr>
<td>PMF</td>
<td>0.5389</td>
<td>0.5560</td>
<td>0.5289</td>
</tr>
<tr>
<td>SM</td>
<td>0.5532</td>
<td>0.5951</td>
<td>0.5396</td>
</tr>
<tr>
<td>CM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Single-sided vs. Multi-sided Fairness

• Most research on the fairness of recommender systems is conducted either from the perspective of customers or from the perspective of product (or service) providers, which is also known as single-sided fairness.

• Fairness, that considers both customer-side fairness and provider-side fairness, is known as multi-sided fairness.
Multi-sided Fairness

• Why multi-sided fairness?
  – When fairness is guaranteed for one side, the fairness and rights of the other side might sacrifice [83,84]
Multi-sided Fairness

• How to approach multi-sided fairness?
  – Usually, the two-sided objective is a linear interpolation of consumer and producer fairness metrics [81,82,83].

\[
\begin{align*}
\text{minimize} & \quad \lambda \cdot \text{inequality}_D(M) + (1 - \lambda) \cdot \text{inequality}_C(M) \\
\text{subject to} & \quad \text{constraints ensuring a correct matching.}
\end{align*}
\]
Static vs. Dynamic Fairness

- **Static fairness** means the protected attribute or group labels (i.e., gender or race) are fixed throughout the entire ranking or recommendation process.

- **Dynamic fairness** considers the dynamic factors in the environment, such as the changes in item utility or attributes in recommendation environment, and learns a strategy to accommodate such dynamics.
Static vs. Dynamic Fairness

- Why consider dynamic fairness?
  - In real world, biases are usually dynamic rather than static. For example,
    - New items will come into the item pool;
Static vs. Dynamic Fairness

• Why consider dynamic fairness?
  – In real world, biases are usually dynamic rather than static. For example,
    • Users experience many new items and may change their preferences;
Static vs. Dynamic Fairness

• Why consider dynamic fairness?
  – In real world, biases are usually dynamic rather than static. For example,
    • The recommendation system will update its recommendation strategy periodically.
Static vs. Dynamic Fairness

- How to achieve dynamic fairness?

Ge et al. [91] proposed to model the dynamic long-term fairness in recommendation with respect to **dynamically changing group labels** through a fairness-constrained reinforcement learning framework.

D becomes much popular than C, while its label is still long-tailed, creating a **NEW Matthew Effect** in long-tailed item group.
Static vs. Dynamic Fairness

The problem is formulated as **Constrained MDP** (Markov Decision Process).

**State**: state $s_t$ of a user
- $H_t$ - user’s most recent positive interaction history
- Demographic information (if exists).

**Action**: a recommendation list $a_t = \{a_t^1, \ldots, a_t^K\}$ with current state $s_t$.

**Reward**: the immediate feedback $R(s_t,a_t)$ given the action $a_t$ and the user state $s_t$
- Typical user feedback includes click, skip, or purchase, etc.

**Cost**: a cost value $C(s_t,a_t)$ given by the problem-specific cost function
- i.e., #items that come from the sensitive group

**Discount rate**: $\gamma_r$ and $\gamma_c$:
- $\gamma_r \in [0,1]$ is for long-term rewards
- $\gamma_c \in [0,1]$ is for long-term costs.
Static vs. Dynamic Fairness

How to achieve dynamic fairness?

Define Exact-K fairness,

\[
\frac{\text{Exposure}_t (G_0)}{\text{Exposure}_t (G_1)} \leq \alpha \\
\text{Exposure}_t (G_0) \leq \alpha \text{Exposure}_t (G_1)
\]
Static vs. Dynamic Fairness

• How to achieve dynamic fairness?

Finally, they used Constraint Policy Optimization to solve the above problem.

Using two critics – a fairness critic and a utility critic to learn fairness and utility in a dynamic reinforcement learning framework.

Ge et al. “Towards Long-term Fairness in Recommendation”. WSDM’21
Applications of Fairness-aware RecSys

- Ride-hailing (Uber, Lyft): Sühr et al. [81]
- Ecommerce (Amazon, Etsy): Patro et al. [82]
- Content streaming (Spotify, YouTube): Htun et al. [85]
- Social Media (Twitter, LinkedIn): Vasudevan et al. [86], Geyik et al. [87]
- Cyber-Physical Systems (e-Vehicle charging): Wang et al. [90]
Evaluation of Fairness

- **Measuring User-side Fairness:** Based on the general definition of user fairness, Fu et al. [89] defined GRU as a measurement.

- **Group Recommendation Unfairness (GRU)**

\[
GRU(G_1, G_2, Q) = \left| \frac{1}{|G_1|} \sum_{i \in G_1} F(Q_i) - \frac{1}{|G_2|} \sum_{i \in G_2} F(Q_i) \right|
\]

$F$ refers to a metric that scores the recommendation quality such that $F(Q_i)$ denotes the recommendation quality for user $u_i$, invoking a metric such as NDCG@K or F1 score.
Evaluation of Fairness

• Measuring Item-side Fairness: Several measures for evaluating the fairness of a ranked list have been explored in the information retrieval literature [88].

• Normalized discounted difference (rND)

\[
rND(\tau) = \frac{1}{Z} \sum_{i=10,20,\ldots}^{N} \frac{1}{\log_2 i} \left| \frac{|S_{1\ldots i}^+|}{i} - \frac{|S^+|}{N} \right|
\]

Normalizer Z is computed as the highest possible value of rND for the given number of items N and protected group size |S+|.

rND computes the difference in the proportion of members of the protected group (S+) at top-i and in the over-all population.
Evaluation of Fairness

• Measuring Item-side Fairness: Several measures for evaluating the fairness of a ranked list have been explored in the information retrieval literature [88].
• Normalized discounted difference (rND)

The figure plots the behavior of rND on synthetic datasets of 1000 items, with 200, 500 and 800 items in S+, as a function of fairness probability.

Yang et al. “Measuring Fairness in Ranked Outputs”. 2016
Evaluation of Fairness

- Measuring Item-side Fairness: Several measures for evaluating the fairness of a ranked list have been explored in the information retrieval literature [88].
- **Normalized discounted KL-divergence (rKL)**

\[ P = \left( \frac{|S^+_{1\ldots i}|}{i}, \frac{|S^-_{1\ldots i}|}{i} \right), Q = \left( \frac{|S^+|}{N}, \frac{|S^-|}{N} \right) \]

\[ rKL(\tau) = \frac{1}{Z} \sum_{i=10,20,\ldots}^{N} \frac{D_{KL}(P||Q)}{\log_2 i} \]

- P is the presented exposure distribution; Q is the desired distribution
- It uses KL-divergence to compute the expectation of the difference between protected group membership at top-i vs. in the over-all population

Yang et al. “Measuring Fairness in Ranked Outputs”. 2016
Evaluation of Fairness

• Measuring Item-side Fairness: Several measures for evaluating the fairness of a ranked list have been explored in the information retrieval literature [88].

• Normalized discounted KL-divergence (rKL)

The figure plots the behavior of rKL on synthetic datasets of 1000 items, with 200, 500 and 800 items in S+, as a function of fairness probability.

Yang et al. “Measuring Fairness in Ranked Outputs”. 2016
Evaluation of Fairness

• Measuring pairwise fairness by comparing utility and prediction errors [92,93,94]

**DEFINITION**  
**Rank Calibration Error**

\[ R_{cal_{A_i}}(\rho, \hat{\rho}) = \frac{\phi^D_i(X)}{\phi(X) - \phi(A_j)} \]

Where \( \phi^D_i(X) \) denotes the number of discordant pairs containing at least one object from the target group \( A_i \).

Kuhlman et al. “Fare: Diagnostics for fair ranking using pairwise error metrics”. WWW ’19
Evaluation of Fairness

- Measuring Item-side Fairness: Researchers also use Gini Index to do the evaluation at an individual level [84,91].

\[
G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2 \sum_{i=1}^{n} \sum_{j=1}^{n} x_j} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n \sum_{j=1}^{n} x_j} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}}
\]
# Fairness in Recommendation — Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Interactions</th>
<th>Sensitive Features</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ModCloth</td>
<td>99,893</td>
<td>Gender</td>
<td>[50]</td>
</tr>
<tr>
<td>RentTheRunway</td>
<td>192,544</td>
<td>Age</td>
<td>[51]</td>
</tr>
<tr>
<td>MovieLens</td>
<td>1,000,000</td>
<td>Age; Gender; Occupation</td>
<td>[52]</td>
</tr>
<tr>
<td>Insurance</td>
<td>5,382</td>
<td>Gender; Marital status; Occupation</td>
<td>[53]</td>
</tr>
<tr>
<td>Post</td>
<td>71,800</td>
<td>Gender</td>
<td>[54]</td>
</tr>
<tr>
<td>Coat</td>
<td>11,600</td>
<td>Age; Gender</td>
<td>[55]</td>
</tr>
<tr>
<td>Sushi</td>
<td>50,000</td>
<td>Age; Gender</td>
<td>[56]</td>
</tr>
</tbody>
</table>
Challenges and Opportunities

- No Consensus on Definition
- Transparent Fairness
- Fairness-Utility Relationship
- Better Evaluation
- ...

84
Challenges and Opportunities

- No Consensus on Definition
- Transparent Fairness
- Fairness-Utility Relationship
- Better Evaluation
- ...

Rutgers
Challenges and Opportunities

No Consensus on Definition

Transparent Fairness

Fairness-Utility Relationship

Better Evaluation

…
Challenges and Opportunities

- No Consensus on Definition
- Transparent Fairness
- Fairness-Utility Relationship
- Better Evaluation

...
Summary

Introduction and Background:
- Social impact of recommender system and fairness.
- Motivation of fairness.
- Relationship with AI Ethics & Beyond Ethics.

Fairness in Machine Learning:
- Fairness in Classification
- Fairness in Ranking

Fairness in Recommendation:
- Introduction
- Taxonomy
- Dataset and Evaluation
- Challenge and Opportunity
Questions?

Yunqi Li, Yingqiang Ge, Yongfeng Zhang
Rutgers University
yunqi.li@rutgers.edu
yingqiang.ge@rutgers.edu
yongfeng.zhang@rutgers.edu

Tutorial website: https://fairness-tutorial.github.io/
Reference

Reference

Reference

Reference

• [29] AirBnB. https://insideairbnb.com
• [36] XING. https://github.com/MilkaLichtblau/xing_dataset
Reference

Reference

- [51] Rishabh Misra, Mengting Wan, Julian McAuley, “Decomposing fit semantics for product size recommendation in metric spaces” RecSys, 2018
- [52] https://grouplens.org/datasets/movielens/
- [53] https://www.kaggle.com/mrmorj/insurance-recommendation
- [54] https://www.kaggle.com/vatsalparsaniya/post-pecommandation
Reference

• [63] Lingxiao Huang and Nisheeth Vishnoi. "Stable and Fair Classification." ICML 2019
Reference

- [76] Li, Y et al. “User-oriented Fairness in Recommendation” WWW’21.
- [77] Xiao, Lin, et al. "Fairness-aware group recommendation with pareto-efficiency." Recsys’17
- [78] Beutel, Alex, et al. “Fairness in Recommendation Ranking through Pairwise Comparisons” SIGKDD’19
- [79] Yao, Sirui, and Bert Huang. “Beyond Parity: Fairness Objectives for Collaborative Filtering” NIPS’17
- [80] Li, Y et al. “Towards Personalized Fairness based on Causal Notion” SIGIR’21
Reference


- [86] Sriram Vasudevan and Krishnaram Kenthapadi. 2020. “The LinkedIn Fairness Toolkit” (LiFT)


Reference