Fair ranking: a critical review, challenges, and an impact-oriented research agenda for future

AICPM 2022

Gourab Kumar Patro
IIT Kharagpur, India
L3S Research Center, Germany
Outline

- Ranking (terminology, problem setting, example)
- Fair ranking (motivation, fairness of exposure)
- Pitfalls of current fair ranking models
- Impact-oriented, long-term thinking (applied modeling and simulation)
  - Causality (necessity and sufficiency)
- Challenges in long-term studies (data and law)
Introduction to Fair Ranking
Ranking in Online Platforms

Major Online (Market) Platforms

Providers

Content / Item Corpus

Consumers

Information retrieval services (e.g., search, recommendation)

Information retrieval services (e.g., search, recommendation)
Ranking in Online Platforms

Major Online (Market) Platforms

Providers

Content / Item Corpus

Information retrieval services (e.g., search, recommendation)

Ranking

Consumers
## Examples

<table>
<thead>
<tr>
<th>Platform Use Case</th>
<th>Example</th>
<th>Content/Item Ranked</th>
<th>Providers</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-commerce</td>
<td><img src="#" alt="Amazon" /></td>
<td>Products</td>
<td>Sellers</td>
</tr>
<tr>
<td>Hiring</td>
<td><img src="#" alt="LinkedIn" /></td>
<td>Candidate Profiles</td>
<td>Job Seekers</td>
</tr>
<tr>
<td>Media</td>
<td><img src="#" alt="Spotify" /></td>
<td>Media Content</td>
<td>Artists</td>
</tr>
</tbody>
</table>
## Examples...

<table>
<thead>
<tr>
<th>Platform Use Case</th>
<th>Example</th>
<th>Content/Item Ranked</th>
<th>Providers</th>
<th>Opportunities</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-commerce</td>
<td><img src="amazon.png" alt="Amazon" /></td>
<td>Products</td>
<td>Sellers</td>
<td>Sales</td>
</tr>
<tr>
<td>Hiring</td>
<td><img src="linkedin.png" alt="LinkedIn" /></td>
<td>Candidate Profiles</td>
<td>Job Seekers</td>
<td>Employment</td>
</tr>
<tr>
<td>Media</td>
<td><img src="spotify.png" alt="Spotify" /></td>
<td>Media Content</td>
<td>Artists</td>
<td>Royalty, Ad revenue</td>
</tr>
</tbody>
</table>
How do platforms usually rank items/services?

Upon query

Given:
(a) A set of candidate items/services
(b) Relevance scores of each item/service (often output of ML model)

Task: Rank the items

Most popular guiding principle: **Probability Ranking Principle**

“Rank the items in the descending order of their probability of relevance to maximize the ranking utility”

## Ranking Example: Hiring Platforms

<table>
<thead>
<tr>
<th>Candidate items = Job seekers’ profiles</th>
<th>Machine learned relevance scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.91</td>
</tr>
<tr>
<td>A2</td>
<td>0.90</td>
</tr>
<tr>
<td>B1</td>
<td>0.89</td>
</tr>
<tr>
<td>B2</td>
<td>0.88</td>
</tr>
</tbody>
</table>
**Ranking Example: Hiring Platforms**

<table>
<thead>
<tr>
<th>Candidate items = Job seekers’ profiles</th>
<th>Machine learned relevance scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.91</td>
</tr>
<tr>
<td>A2</td>
<td>0.90</td>
</tr>
<tr>
<td>B1</td>
<td>0.89</td>
</tr>
<tr>
<td>B2</td>
<td>0.88</td>
</tr>
</tbody>
</table>
What’s wrong with optimizing ranking utility?

Position Bias

<table>
<thead>
<tr>
<th>Ranked Results</th>
<th>Relevance</th>
<th>Attention Received</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.91</td>
<td>0.5</td>
</tr>
<tr>
<td>A2</td>
<td>0.90</td>
<td>0.3</td>
</tr>
<tr>
<td>B1</td>
<td>0.89</td>
<td>0.13</td>
</tr>
<tr>
<td>B2</td>
<td>0.88</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Exposure or visibility: the amount of user attention received by a provider

- High rank ⇔ More exposure
- More #times selected in top-k ⇔ More exposure
What’s wrong with optimizing ranking utility?

**Position Bias**

<table>
<thead>
<tr>
<th>Ranked Results</th>
<th>Relevance</th>
<th>Attention Received</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.91</td>
<td>0.5</td>
</tr>
<tr>
<td>A2</td>
<td>0.90</td>
<td>0.3</td>
</tr>
<tr>
<td>B1</td>
<td>0.89</td>
<td>0.13</td>
</tr>
<tr>
<td>B2</td>
<td>0.88</td>
<td>0.07</td>
</tr>
</tbody>
</table>

**Exposure or visibility:** the amount of user attention received by a provider

- High rank ⇔ More exposure
- More #times selected in top-k ⇔ More exposure
What’s wrong with optimizing ranking utility?

- The ML models which estimate relevance scores can reflect the bias inherent in data.
- Optimal ranking based on biased estimates of relevance scores can lead to unfairness for the providers.

Hiring, Gig-economy, Admissions, E-commerce, Marketplaces

Rankings bring socio-economic opportunities for the providers.

Unfairness in ranking \(\Rightarrow\) significant societal harm
Fairness in Rankings

- **Common motivation**: Fair access to opportunity for the providers.
- Both *individual fairness* and *group fairness* are used.
- Top-k ranks = Finite scarce resource
  - **Fair ranking** ⇔ *Proportional presence of group members in top-k*
- Total available exposure = Finite scarce resource
  - **Fair ranking** ⇔ *Equitable allocation of exposure*
- See *Fairness in Ranking: A Survey* by Zehlike et al. for an exhaustive analysis of fair ranking methods together with the normative assumptions behind these methods
Fairness in Rankings

- **Common motivation:** Fair access to opportunity for the providers.
- Both *individual fairness* and *group fairness* are used.
- Top-k ranks = Finite scarce resource
  - *Fair ranking* ⇔ *Proportional presence of group members in top-k*
- Total available exposure = Finite scarce resource
  - *Fair ranking* ⇔ *Equitable allocation of exposure*
- See *Fairness in Ranking: A Survey* by Zehlike et al. for an exhaustive analysis of fair ranking methods together with the normative assumptions behind these methods

Let’s see some fair ranking approaches
Fair top-k ranking

For every prefix of top-k (i.e., top-i, where 1≤i≤k),
\[
\text{proportion(\text{group 1})} \approx \text{proportion(\text{group 2})}
\]

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>A1</td>
</tr>
<tr>
<td>0.90</td>
<td>A2</td>
</tr>
<tr>
<td>0.89</td>
<td>B1</td>
</tr>
<tr>
<td>0.88</td>
<td>B2</td>
</tr>
</tbody>
</table>

*Zehlike et al. CIKM 2017, Celis et al. ICALP 2018, Geyik et al. KDD 2019*
For every prefix of top-k (i.e., top-i, where 1 ≤ i ≤ k),
proportion(group 1) ≈ proportion(group 2)

*Zehlike et al. CIKM 2017, Celis et al. ICALP 2018, Geyik et al. KDD 2019*
Fair top-k ranking

For every prefix of top-k (i.e., top-i, where 1≤i≤k),
proportion(group 1) ≈ proportion(group 2)

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Rank</th>
<th>Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>B1</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>A2</td>
<td>3</td>
<td>0.13</td>
</tr>
<tr>
<td>B2</td>
<td>4</td>
<td>0.07</td>
</tr>
</tbody>
</table>

*Zehlike et al. CIKM 2017, Celis et al. ICALP 2018, Geyik et al. KDD 2019*
**Fairness of Exposure**

**Stochastic ranking to balance relevance and attention**

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>A1</td>
</tr>
<tr>
<td>0.90</td>
<td>A2</td>
</tr>
<tr>
<td>0.89</td>
<td>B1</td>
</tr>
<tr>
<td>0.88</td>
<td>B2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.13</td>
</tr>
<tr>
<td>4</td>
<td>0.07</td>
</tr>
</tbody>
</table>

*Singh et al. KDD 2018*
### Stochastic ranking to balance relevance and attention

<table>
<thead>
<tr>
<th>Rank</th>
<th>Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.13</td>
</tr>
<tr>
<td>4</td>
<td>0.07</td>
</tr>
</tbody>
</table>

\[ p_1 + p_2 + p_3 + p_4 = 1 \]

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Candidates</th>
<th>(p_1)</th>
<th>(p_2)</th>
<th>(p_3)</th>
<th>(p_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>A1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.89</td>
<td>B1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.88</td>
<td>B2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Singh et al. KDD 2018*
**Fairness of Exposure**

**Stochastic ranking to balance relevance and attention**

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Candidates</th>
<th>Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>A1</td>
<td>0.5</td>
</tr>
<tr>
<td>0.90</td>
<td>A2</td>
<td>0.3</td>
</tr>
<tr>
<td>0.89</td>
<td>B1</td>
<td>0.13</td>
</tr>
<tr>
<td>0.88</td>
<td>B2</td>
<td>0.07</td>
</tr>
</tbody>
</table>

\[ p_1 + p_2 + p_3 + p_4 = 1 \]

*Singh et al. KDD 2018*
Equity of Amortized Attention

Balancing relevance and attention amortized over time

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>A1</td>
</tr>
<tr>
<td>0.90</td>
<td>A2</td>
</tr>
<tr>
<td>0.89</td>
<td>B1</td>
</tr>
<tr>
<td>0.88</td>
<td>B2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.13</td>
</tr>
<tr>
<td>4</td>
<td>0.07</td>
</tr>
</tbody>
</table>

*Biega et al. SIGIR 2018
## Equity of Amortized Attention

### Balancing relevance and attention amortized over time

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Candidates</th>
<th>Rank</th>
<th>Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>A1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>0.90</td>
<td>A2</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>0.89</td>
<td>B1</td>
<td>3</td>
<td>0.13</td>
</tr>
<tr>
<td>0.88</td>
<td>B2</td>
<td>4</td>
<td>0.07</td>
</tr>
</tbody>
</table>

*Balancing relevance and attention amortized over time*  

![Diagram of amortized attention over time](image)

*Biega et al. SIGIR 2018*
Equity of Amortized Attention

Balancing relevance and attention amortized over time

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Candidates</th>
<th>Rank</th>
<th>Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>A1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>0.90</td>
<td>A2</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>0.89</td>
<td>B1</td>
<td>3</td>
<td>0.13</td>
</tr>
<tr>
<td>0.88</td>
<td>B2</td>
<td>4</td>
<td>0.07</td>
</tr>
</tbody>
</table>

What is wrong with optimizing ranking utility? (2)

*Biega et al. SIGIR 2018*
Pitfalls of Fair Ranking Models

Delayed Impacts
- Factors Beyond Exposure
- Temporal Significance
- Spillover Effects

Ecosystem Dynamics
- Strategic Behaviour

Uncertain Outcomes
- Uncertainties
Pitfall 1: Delayed Impact

Factors Beyond Exposure

(Ranking position-based) Exposure is often used as a proxy for provider utility *

- Prior Beliefs, User Perceptions, User Activity, User Preferences

*A. J. Biega et al., 2018; A. Singh and T. Joachims, 2018; M. Zehlike and C. Castillo, 2020*
Pitfall 1: Delayed Impact

Factors Beyond Exposure

(Ranking position-based) Exposure is often used as a proxy for provider utility *

- Prior Beliefs, User Perceptions, User Activity, User Preferences

* A. J. Biega et al., 2018; A. Singh and T. Joachims, 2018; M. Zehlike and C. Castillo, 2020
Pitfall 1: Delayed Impact

Factors Beyond Exposure

(Ranking position-based) Exposure is often used as a proxy for provider utility *

- Prior Beliefs, User Perceptions, User Activity, User Preferences

* A. J. Biega et al., 2018; A. Singh and T. Joachims, 2018; M. Zehlike and C. Castillo, 2020
Pitfall 1: Delayed Impact

Factors Beyond Exposure

(Ranking position-based) Exposure is often used as a proxy for provider utility *

- Prior Beliefs, User Perceptions, User Activity, User Preferences

\[ t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow t_4 \]

Employer X

Employer Y

No pre-existing bias

Biased against ▲

Resultant opportunity distribution = Unfair

*A. J. Biega et al., 2018; A. Singh and T. Joachims, 2018; M. Zehlike and C. Castillo, 2020*
Pitfall 1: Delayed Impact

Temporal Significance

In fast moving domains, items or services may only be relevant for a short period of time*

- Temporal Degradation of Value: (e.g. News), Temporal Urgency (e.g. Takeaway)

*P. G. Campos et al., 2014; Q. Yuan et al., 2013
Pitfall 1: Delayed Impact

Temporal Significance

In fast moving domains, items or services may only be relevant for a short period of time:

- Temporal Degradation of Value: (e.g. News), Temporal Urgency (e.g. Takeaway)

*P. G. Campos et al., 2014; Q. Yuan et al., 2013
Pitfall 1: Delayed Impact

Temporal Significance

In fast moving domains, items or services may only be relevant for a short period of time*

- Temporal Degradation of Value: (e.g. News), Temporal Urgency (e.g. Takeaway)

*P. G. Campos et al., 2014; Q. Yuan et al., 2013
Pitfall 1: Delayed Impact

Temporal Significance

In fast moving domains, items or services may only be relevant for a short period of time*

- Temporal Degradation of Value: (e.g. News), Temporal Urgency (e.g. Takeaway)

*P. G. Campos et al., 2014; Q. Yuan et al., 2013
Pitfall 1: Delayed Impact

Temporal Significance

In fast moving domains, items or services may only be relevant for a short period of time*

- Temporal Degradation of Value: (e.g. News), Temporal Urgency (e.g. Takeaway)

*P. G. Campos et al., 2014; Q. Yuan et al., 2013
Pitfall 1: Delayed Impact

Spillover Effects

Substantial externalities (or, Spillover Effects), in addition to the first-order effects (i.e. immediate effect of an item’s ranking position)

- Compounding Popularity (First-Exposed-Advantage)\(^1\), Spillover across Products & Ranking Types \(^2\), Cross-Platform/Competition Effects \(^3\)

---

Pitfall 2: Ecosystem Dynamics

Strategic Behaviours

Current fair ranking mechanisms often fail to consider that providers themselves could be strategic players who actively maximize their utilities*

- Adversarial Attacks
  - Shilling Attacks, Data Poisoning Attacks, Doppelganger Bot Attacks
  - Content Duplication Attacks, Purposeful Information Provision/Withholding

- Strategic Offerings
  - Quality Improvements
  - Shift of Offering Strategy

- Platforms as Providers

*G. Bahar et al., 2016; M. Tennenholtz and O. Kurland, 2019
Pitfall 3: Uncertain Outcomes

Uncertainty

Fairness-aware ranking mechanisms often make assumption on uncertainties and noises, which are rarely available in real-world settings

- Sensitive Data (e.g. Demographics) ¹
- Position Bias ²
- Differential Uncertainties ³
- Relevance

¹ M. Andrus et al., 2021; M. Bogen et al., 2020; ² A. Agarwal et al., 2019; P. Chandar and B. Carterette, 2018; ³ V. Emelianov et al., 2020; N. Garg et al., 2021
Towards Impact-oriented Fairness
Towards Impact-oriented Fairness

Key aspects to move forward Algorithmic Impact Assessment (AIA)

1. Move beyond discrete moments of decision making [Vecchione et al., 2021]
   → Simulation and Applied Modeling to Study Long-term Effects and Context-specific Dynamics.

2. Participation of every suitable stakeholder (systems developers, affected communities, external experts, and public agencies) [Metcalf et al., 2021]
Simulation-based environments can help in:

1. Studying relationships between systems’ usage - users’ behavior
   → Homogenization effects [Chaney et al., 2018]

2. Detecting new forms of relationships
   → Performance paradox [Zhang et al, 2020]

3. Replicating results obtained in empirical studies
   → Popularity bias [Fortunato et al., 2006]
Simulation-based environments can help in:

1. Studying relationships between systems’ usage - users’ behavior
   → Homogenization effects [Chaney et al., 2018]

2. Detecting new forms of relationships
   → Performance paradox [Zhang et al., 2020]

3. Replicating results obtained in empirical studies
   → Popularity bias [Fortunato et al., 2006]
Simulation Tasks

Simulation-based environments can help in:

1. Studying relationships between systems’ usage - users’ behavior
   → Homogenization effects [Chaney et al., 2018]

2. Detecting new forms of relationships
   → Performance paradox [Zhang et al., 2020]

3. Replicating results obtained in empirical studies
   → *Popularity bias* [Fortunato et al., 2006]
Simulation Frameworks

MARS-Gym: https://github.com/deeplearningbrasil/mars-gym

ML-fairness-gym: https://github.com/google/ml-fairness-gym

Accordion: https://github.com/jamesmcinerney/accordion

RecLab: https://github.com/berkeley-reclab/RecLab

RecSim NG: https://github.com/google-research/recsim_ng

SIREN: https://github.com/dbountouridis/siren

T-RECS: https://github.com/elucherini/t-recs

RecoGym: https://github.com/mindis/rnd-reco-gym

Virtual-Taobao: https://github.com/eyounx/VirtualTaobao
Simulation Frameworks

MARS-Gym: https://github.com/deeplearningbrasil/mars-gym
ML-fairness-gym: https://github.com/google/ml-fairness-gym
Accordion: https://github.com/jamesmcinerney/accordion
RecLab: https://github.com/berkeley-reclab/RecLab
RecSim NG: https://github.com/google-research/recsim_ng
SIREN: https://github.com/dbountouridis/siren
T-RECS: https://github.com/elucherini/t-recs
RecoGym: https://github.com/mindis/rnd-reco-gym
Virtual-Taobao: https://github.com/eyounx/VirtualTaobao

Assumptions? Worldviews?
Friedler et al. (2021)
Causality and Impact Assessment in Retrieval Systems

Retrieval from a causal inference perspective

- **Ranking as an intervention**
  - Goal: Estimate the effect of new interventions (new ranking mechanisms)

- **Counterfactuals** for analyzing the impact of associated services (complementary recsys)
  - Goal: Estimate the additional traffic created by the associated service

*Schnabel et al. ICML 2016, Sharma et al. EC 2015*
Will causality be enough for impact assessment?

- Short answer: No
- Long answer: It depends on what one considers as an impact. If one strives to estimate only the instantaneous click behaviour, then may be yes. But for an impact-oriented, long-term study, the answer is no.

- What more is required?
  - Behavioural modeling for strategic behaviour
  - Temporal modeling for temporal variations

Schnabel et al. ICML 2016
Temporal, behavioural and causal models can be integrated to ensure

- Ecosystem parametrization
- Stakeholder behaviour
- System pay-offs

are representative of the real-world.
Towards Impact-oriented Fairness

1. Data bottlenecks
2. Legal bottlenecks
Challenge 1: Data bottlenecks

Datasets Suitable for Impact-oriented Fairness Analysis

- Complemented with additional contextual information, to understand broader environment and underlying dynamics. Such as user interface design characteristics, whether a user was directed from an affiliate link etc

- Move from static to temporal datasets. Such as monitoring effects of previous rankings over time, temporal variations in rankings, and modelling ranking trajectory of new entrants onto the platform.

- Incorporating modelling of uncertainty. For example uncertainty in gender/race, and in the genuineness of ratings and reviews. Also keeping in mind the legal implications.
Challenge 2: Legal bottlenecks

Information asymmetry between online platforms using rankings, and individuals/organisations wanting to understand or audit the ranking system.

A legal framework can:

i) Enable access to ranking information. However,
   - Differences across jurisdictions.
   - Privacy/transparency tradeoff.

ii) Provide a method by which to challenge ranking system implementation. However:
   - Clear guidance needed on how to apply laws to long-term fairness scenarios.
   - Consideration needs to be given to differences in laws that are applicable to ranking of individuals vs ranking of products.
Pitfalls of Existing Fair Ranking Models

Delayed Impacts
- Factors Beyond Exposure
- Temporal Significance
- Spillover Effects

Ecosystem Dynamics
- Strategic Behaviour

Uncertain Outcomes
- Uncertainties in position bias, demographic info, relevance

Summary

Legal bottlenecks
- Simulations
- Applied modeling
- Impact assessment

Needs
- Impact-oriented and long-term perspective

Data bottlenecks
Fair ranking: a critical review, challenges, and future directions (FAccT 2022)
Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, Nikhil Garg

**Delayed Impacts**
- Factors Beyond Exposure
- Temporal Significance
- Spillover Effects

**Ecosystem Dynamics**
- Strategic Behaviour

**Uncertain Outcomes**
- Uncertainties in position, bias, demographic info, relevance

**Legal bottlenecks**
- Simulations
- Applied modeling
- Impact assessment

**Needs**
- Impact-oriented and long-term perspective

**Data bottlenecks**
Thank you!