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

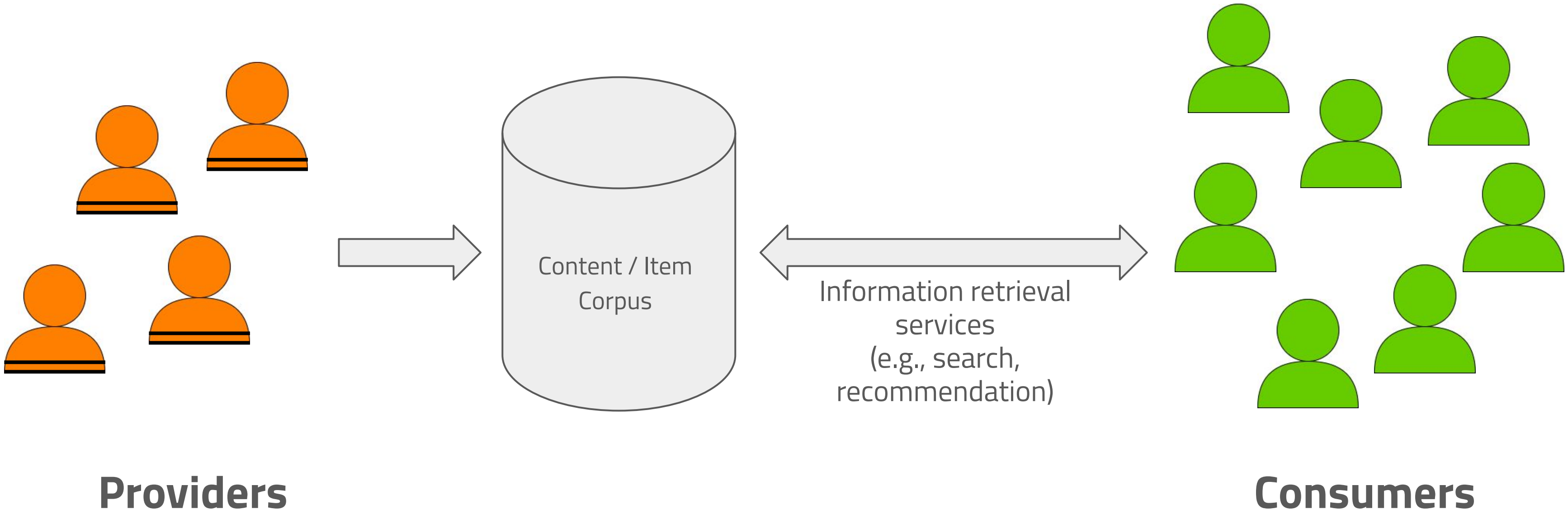
Outline

- 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

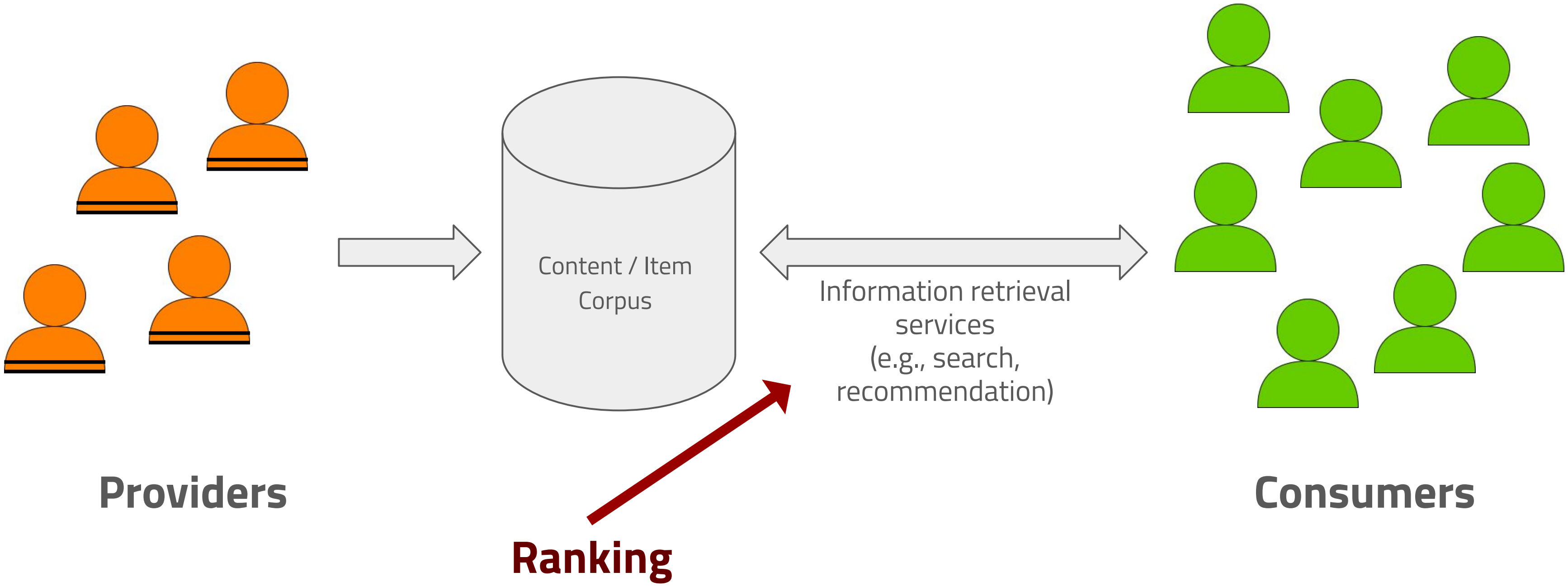
Ranking in Online Platforms

Major Online (Market) Platforms



Ranking in Online Platforms

Major Online (Market) Platforms



Examples

Platform Use Case	Example	Content/Item Ranked	Providers
E-commerce	amazon	Products	Sellers
Hiring	LinkedIn	Candidate Profiles	Job Seekers
Media	Spotify	Media Content	Artists

Examples...

Platform Use Case	Example	Content/Item Ranked	Providers	Opportunities
E-commerce	amazon	Products	Sellers	Sales
Hiring	LinkedIn	Candidate Profiles	Job Seekers	Employment
Media	Spotify	Media Content	Artists	Royalty, Ad revenue

How do platforms usually rank items/services?

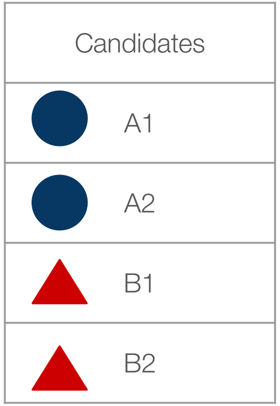
Upon query

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*"

*Stephen E Robertson. 1977. *The probability ranking principle in IR*. Journal of documentation (1977).

Ranking Example: Hiring Platforms

Candidate items = Job seekers’ profiles	Machine learned relevance scores
● A1	0.91
● A2	0.90
▲ B1	0.89
▲ B2	0.88

Ranking Example: Hiring Platforms

Ranking

Candidate items = Job seekers' profiles	Machine learned relevance scores
● A1	0.91
● A2	0.90
▲ B1	0.89
▲ B2	0.88

What's wrong with optimizing ranking utility?

Position Bias

Ranked Results	Relevance	Attention Received
A1	0.91	0.5
A2	0.90	0.3
B1	0.89	0.13
B2	0.88	0.07

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

Ranked Results	Relevance	Attention Received
A1 ●	0.91	0.5
A2 ●	0.90	0.3
B1 ▲	0.89	0.13
B2 ▲	0.88	0.07

Total Attention = **0.8**

Total Attention = **0.2**

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

$\Downarrow$

Rankings bring socio-economic opportunities for the providers.

$\Downarrow$

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),
proportion(group 1) ≈ proportion(group 2)

Relevance	Candidates
0.91	A1
0.90	A2
0.89	B1
0.88	B2

*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)

Relevance	Candidates
0.91	A1
0.90	A2
0.89	B1
0.88	B2

make fair

Candidate	Rank	Attention
A1	1	0.5
B1	2	0.3
A2	3	0.13
B2	4	0.07

*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)

Relevance	Candidates
0.91	A1 ●
0.90	A2 ●
0.89	B1 ▲
0.88	B2 ▲

make fair

Candidate	Rank	Attention
A1 ●	1	0.5
B1 ▲	2	0.3
A2 ●	3	0.13
B2 ▲	4	0.07

Group	**Attention**
●	**0.63**
▲	**0.37**

*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

Relevance	Candidates
0.91	A1
0.90	A2
0.89	B1
0.88	B2

Rank	Attention
1	0.5
2	0.3
3	0.13
4	0.07

*Singh et al. KDD 2018

Fairness of Exposure

Stochastic ranking to balance relevance and attention

Relevance	Candidates
0.91	A1
0.90	A2
0.89	B1
0.88	B2

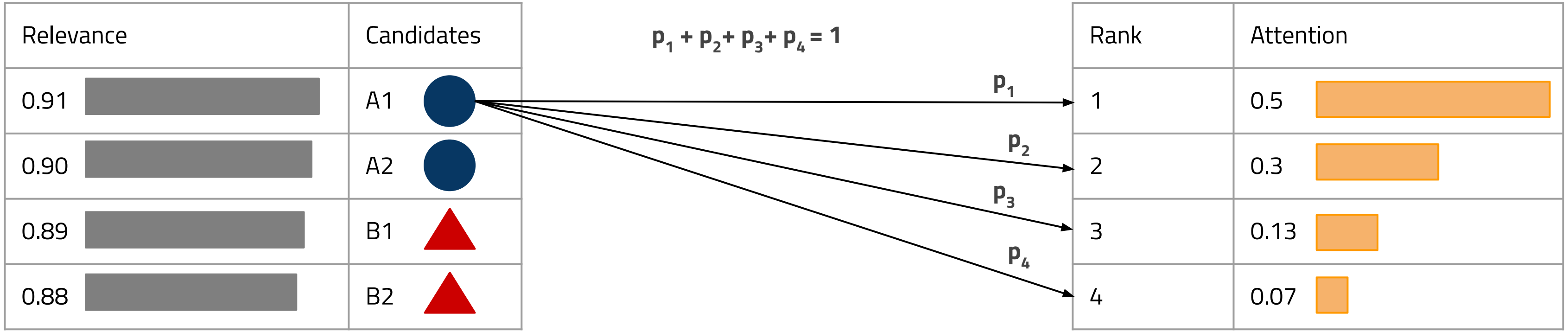
$p_1 + p_2 + p_3 + p_4 = 1$

Rank	Attention
1	0.5
2	0.3
3	0.13
4	0.07

*Singh et al. KDD 2018

Fairness of Exposure

Stochastic ranking to balance relevance and attention



$$\frac{\text{Expected attention}\ (\bullet)}{\text{Expected attention}\ (\blacktriangle)} = \frac{\text{Mean relevance}\ (\bullet)}{\text{Mean relevance}\ (\blacktriangle)}$$

*Singh et al. KDD 2018

Equity of Amortized Attention

Balancing relevance and attention amortized over time

Relevance	Candidates
0.91	A1
0.90	A2
0.89	B1
0.88	B2

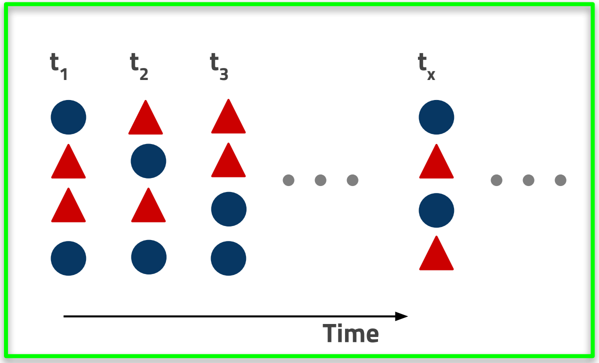
Rank	Attention
1	0.5
2	0.3
3	0.13
4	0.07

*Biega et al. SIGIR 2018

Equity of Amortized Attention

Balancing relevance and attention amortized over time

Relevance	Candidates	Rank	Attention
0.91	A1	1	0.5
0.90	A2	2	0.3
0.89	B1	3	0.13
0.88	B2	4	0.07



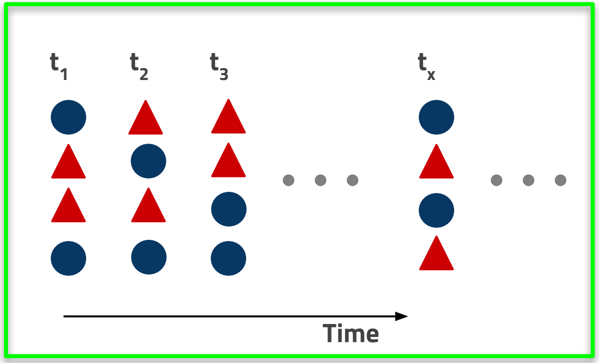
*Biega et al. SIGIR 2018

Equity of Amortized Attention

Balancing relevance and attention amortized over time

Relevance	Candidates
0.91	A1 ●
0.90	A2 ●
0.89	B1 ▲
0.88	B2 ▲

Rank	Attention
1	0.5
2	0.3
3	0.13
4	0.07



$$\frac{\text{Sum of attention over time (●)}}{\text{Sum of attention over time (▲)}} = \frac{\text{Sum of relevance over time (●)}}{\text{Sum of relevance over time (▲)}}$$

*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

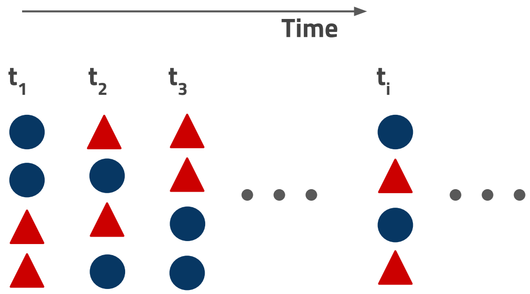
*A. J. Biega et al., 2018; A. Singh and T. Joachims, 2018; M. Zehlike and C. Castillo, 2020

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



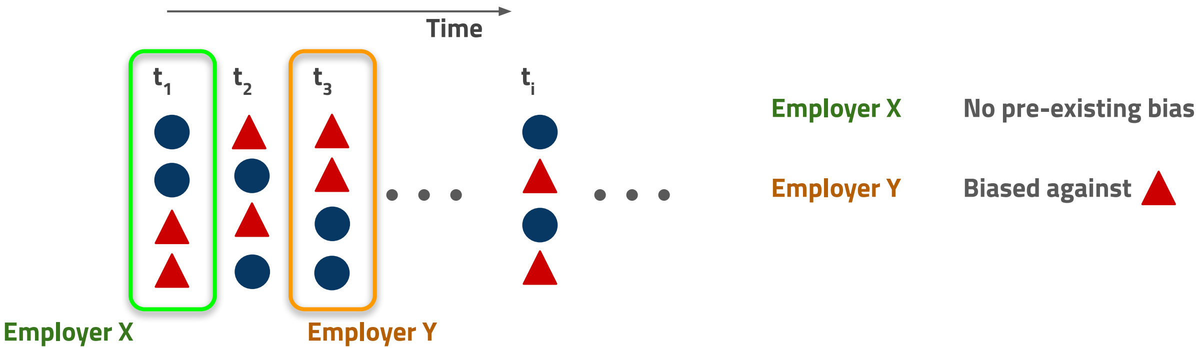
*A. J. Biega et al., 2018; A. Singh and T. Joachims, 2018; M. Zehlike and C. Castillo, 2020

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



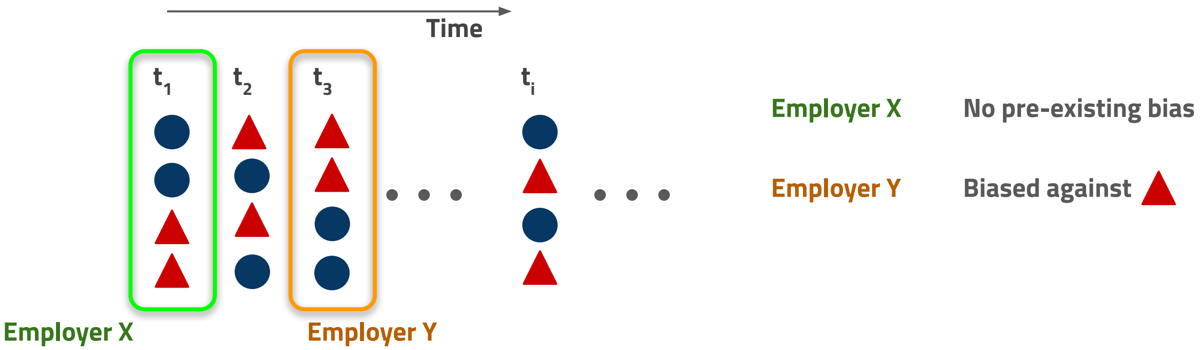
*A. J. Biega et al., 2018; A. Singh and T. Joachims, 2018; M. Zehlike and C. Castillo, 2020

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



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)

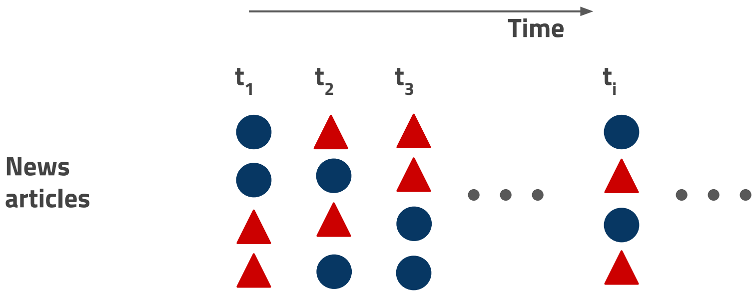
*P. G. Campos et al., 2014; Q. Yuan et al., 2013

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)



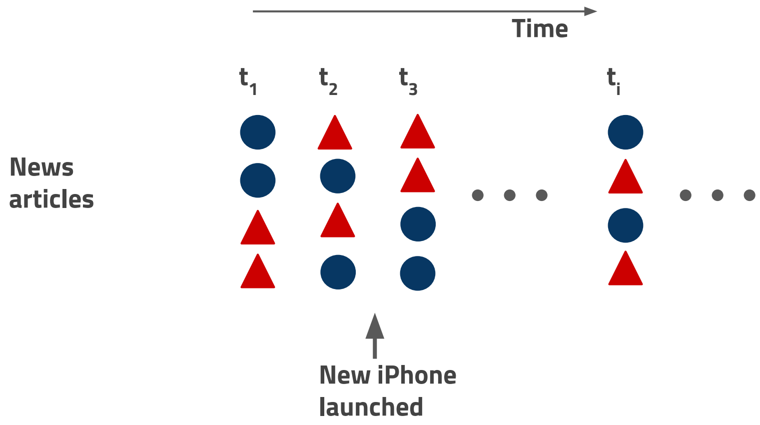
*P. G. Campos et al., 2014; Q. Yuan et al., 2013

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)



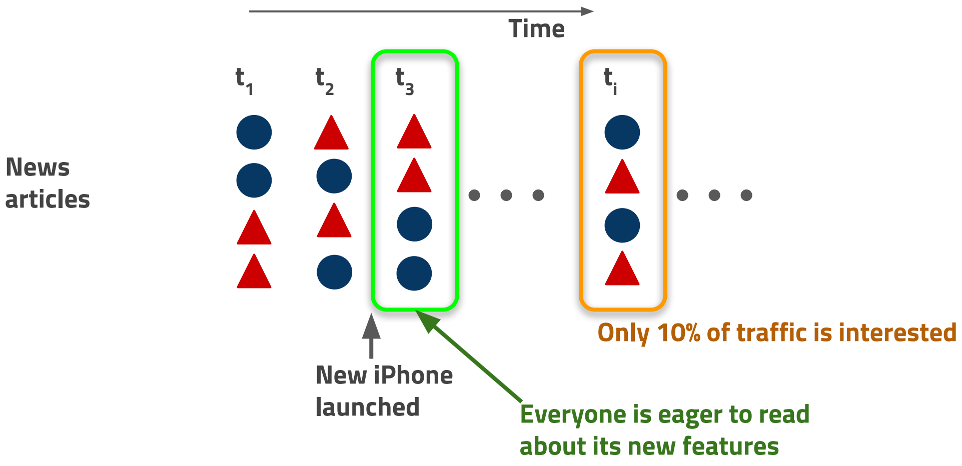
*P. G. Campos et al., 2014; Q. Yuan et al., 2013

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)



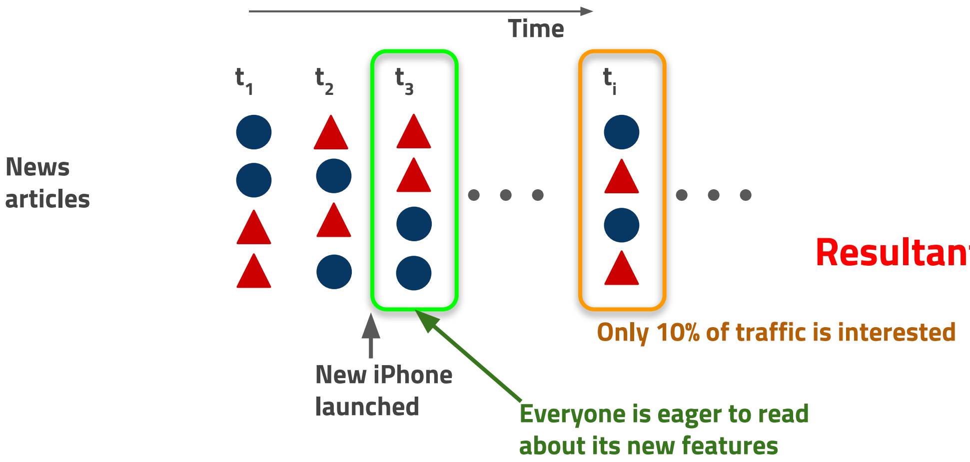
*P. G. Campos et al., 2014; Q. Yuan et al., 2013

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)



Resultant opportunity distribution = Unfair

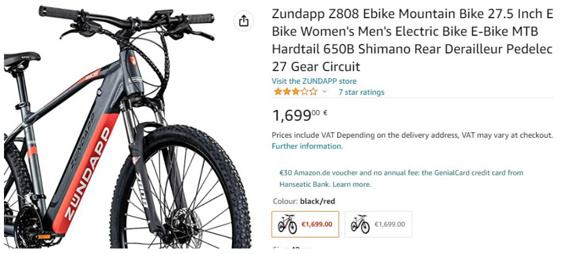
*P. G. Campos et al., 2014; Q. Yuan et al., 2013

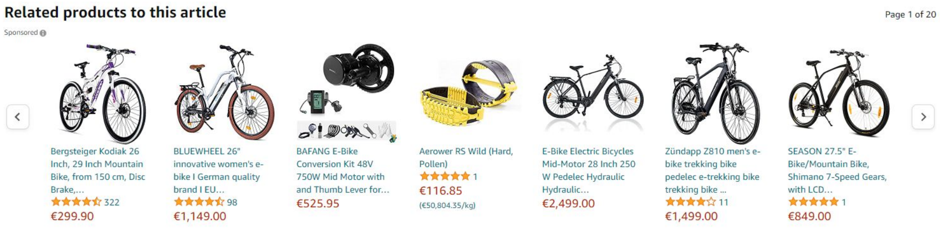
Pitfall 1: Delayed Impact

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]





1: F. Figueiredo et al., 2014; H. Abdollahpouri et al., 2017; 2: C. Liang et al., 2019; M. J. Pazzani and D. Billsus, 2007;
3. A. Farahat and T. Bhatia, 2016; H. Krijestorac et al., 2020

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 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 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 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/berk

RecSim NG: https://github.com/g

SIREN: https://github.com/dbountou

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**

Expose a user to items in ranking (online platform study)

Expose a patient to a treatment (medical study)

  - 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

Applied Modeling

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.

Summary

Delayed Impacts

- Factors Beyond Exposure
- Temporal Significance
- Spillover Effects

Ecosystem Dynamics

- Strategic Behaviour

Uncertain Outcomes

- Uncertainties in position bias, demographic info, relevance

Legal bottlenecks

Needs
Impact-oriented and long-term perspective

- Simulations
- Applied modeling
- Impact assessment

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

Needs Impact-oriented and long-term perspective

- Simulations
- Applied modeling
- Impact assessment

Data bottlenecks

Thank you!

MD4SG

Mechanism Design for Social Good