Random Isn't Always Fair:

Candidate Set Imbalance & Exposure Inequality in Recommender Systems

Amanda Bower, Kristian Lum, Tomo Lazovich,
Kyra Yee, Luca Belli
Twitter

FAccTRec 2022

Amanda Bower [speaker]
Pronouns: She/Her
@amandarg__

Kristian Lum
Pronouns: She/Her
@KLdivergence

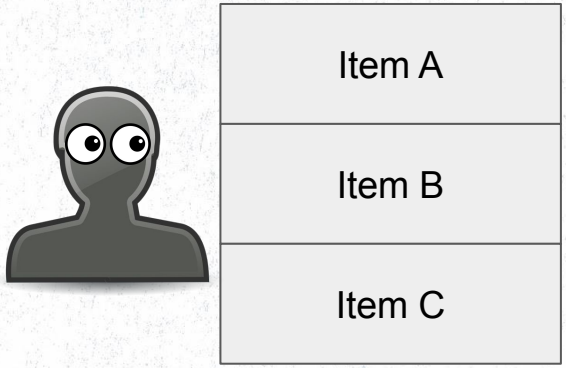
Tomo Lazovich
Pronouns: They/Them
@laughsovich

Kyra Yee
Pronouns: She/Her
@Kyra_Yee

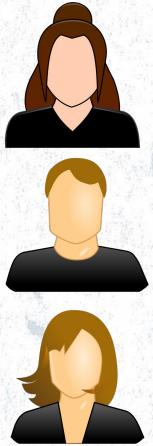
Luca Belli
Pronouns: He/Him
@__lucab

Impact of Recommender Systems

Historically items are ordered by the **Probability Ranking Principle to maximize utility to the consumer**…



…while the **producers** of the items are largely ignored….



…despite economic and social impacts to producers.

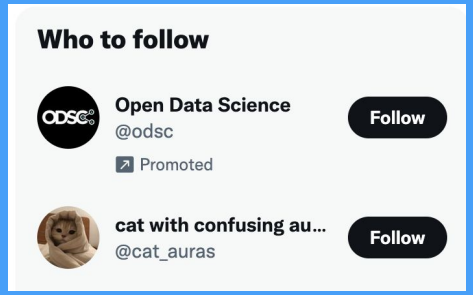
Access to Multi-billion Dollar Creator Economy

Occupational stereotyping In Image Search

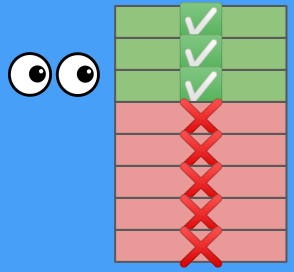
Factors Contributing to Inequalities

Naturally only a few items get exposure

Limited Rec Spots

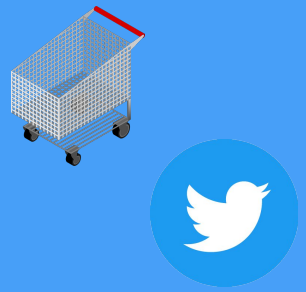


User Trust Bias



Limited User Attention

User Intention



\+

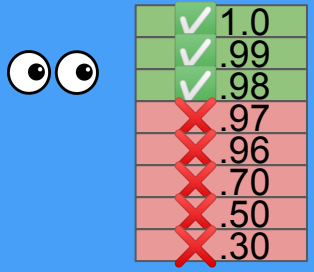
Probability Ranking Principle

(using estimated relevance scores)

1.0
.99
.98
.97
.96
.70
.50
.30

=

Inequalities



One Standard Solution: Stochasticity

Fairness of Exposure in Rankings

Ashudeep Singh
Cornell University
Ithaca, NY
ashudeep@cs.cornell.edu

Thorsten Joachims
Cornell University
Ithaca, NY
tj@cs.cornell.edu

Stochastic Rankings

D: .97
B: .99
A: 1.0
C: .98

~

vs

Deterministic Rankings
(Probability Ranking Principle)

A: 1.0
B: .99
C: .98
D: .97

So sampling rankings from a uniform distribution should be the most “fair”?

Uniformly Random Rankings Can Increase Inequality

Deterministic Rankings for Consumers

	1	2	3	4	5	6	7	8	9	10
First	A	B	C	D	E	F	G	H	I	J
Second	B	C	D	E	F	G	H	I	J	A
Third	C	D	E	F	G	H	I	J	A	B
Fourth	J	J	J	J	J	J	J	A	B	C

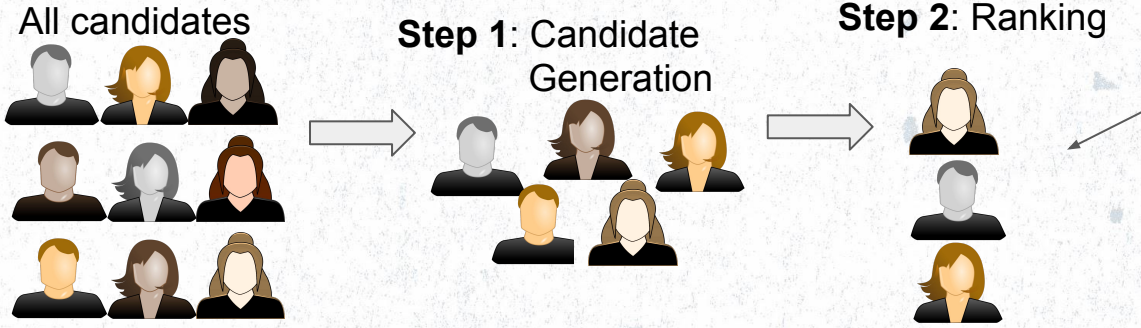
popular producer

Expected Producer Exposure

	A	B	C	D	E	F	G	H	I	J
Ranking	1	1	1	1	1	1	1	1	1	1
Random	1	1	1	3/4	3/4	3/4	3/4	3/4	3/4	2 1/2

What's going on?

In practice, ranking is typically a two-step process



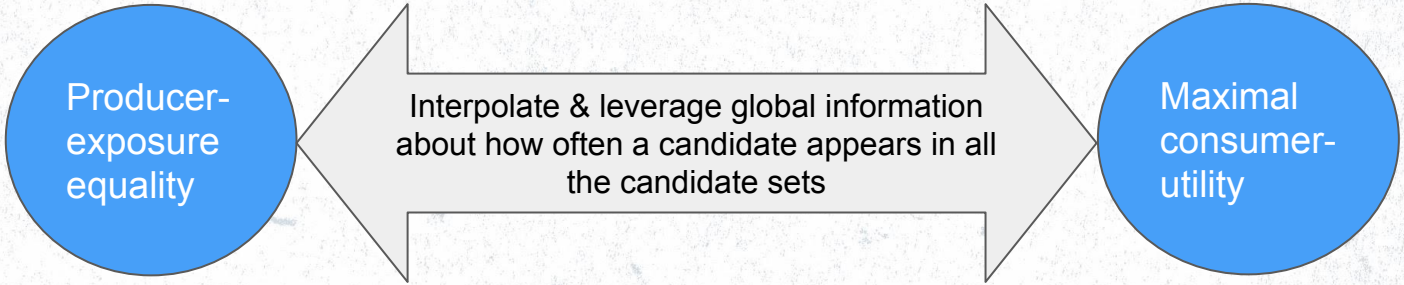
Literature focuses on step #2 with the exception of Wang & Joachims '22

& rankings are produced for many users at many points in time.

A line of literature focuses on fairness at an **individual ranking level** as opposed to the **global level**.

Our Contribution

We propose a post-processing algorithm to sample rankings from a class of ranking distributions.



Plackett-Luce Sampling

Set-up:

- n users: $U := \{u_i\}_{i=1}^{n}$
- m items: $V := \{v_i\}_{i=1}^{m}$
- For every user $u \in U$ let $\{v_{u_i}\}_{i=1}^{m_u}$ be the candidate set of m_u items that need to be ranked
- Let the corresponding set of relevance scores be given by $\{r_{u_i}\}_{i=1}^{m_u}$
- Let $\beta \in \mathbb{R}$

Algorithm:
For each user $u \in U$ sample their ranking from the (scaled) Plackett-Luce distribution where the probability of sampling the ranking $(v_{u_1}, v_{u_2}, \ldots, v_{m_u})$ is

$$\prod_{j=1}^{m_u} \frac{\exp(\beta r_{u_j})}{\sum_{\ell=j}^{n} \exp(\beta r_{u_\ell})}$$

“Evaluating Stochastic Rankings with Expected Exposure” - Diaz et al., 2020

“Joint Multisided Exposure Fairness for Recommendation” - Wu et al. 2022

Our algorithm: Plackett-Luce Sampling With Inverse Candidate Frequency Weights

Set-up:

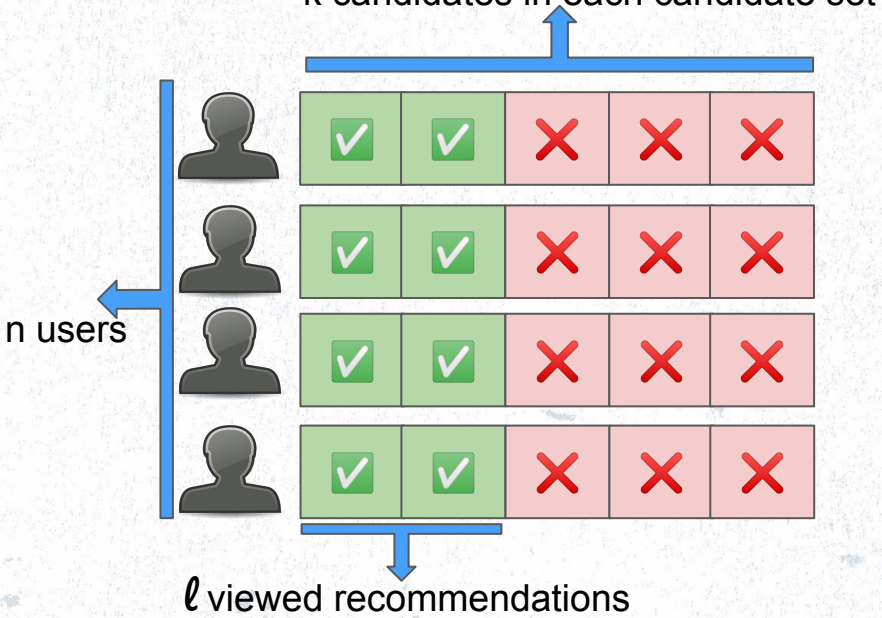
- n users: $U := \{u_i\}_{i=1}^{n}$
- m items: $V := \{v_i\}_{i=1}^{m}$
- For every user $u \in U$ let $\{v_{u_i}\}_{i=1}^{m_u}$ be the candidate set of m_u items that need to be ranked
- Let the corresponding set of relevance scores be given by $\{r_{u_i}\}_{i=1}^{m_u}$
- Let $\alpha, \beta \in \mathbb{R}$
- Let W_v be the number of candidate sets that item $v \in V$ appears in

Algorithm:
For each user $u \in U$ sample their ranking from the (scaled) Plackett-Luce distribution where the probability of sampling the ranking $(v_{u_1}, v_{u_2}, \ldots, v_{m_u})$ is

$$\prod_{j=1}^{m_u} \frac{\alpha \exp(\beta r_{u_j}) + \frac{1}{W_{v_{u_j}} + \alpha}}{\sum_{\ell=j}^{n} \alpha \exp(\beta r_{u_\ell}) + \frac{1}{W_{v_\ell} + \alpha}}$$

Experimental Set-Up

Basics



Candidate Set Construction

m candidates, each with a candidate score c_i

For all but 10 candidates, $c_i \sim B(\alpha = 1, \beta = 10)$. The other 10 candidates have a candidate score of 5.

Candidate sets of size k are sampled such that candidates with higher candidate scores are more likely to be included.

Experimental Set-Up

Algorithms

- **PL-ICFW**: Our algorithm
- **Inverse Weighted**: Our algorithm where α, □ = 0
- **Deterministic**: Candidates ordered in decreasing order of relevance score
- **Scaled-PL:** Plackett-Luce Sampling
- **Randomized**: Uniformly randomized rankings
- **PG-Rank**: In-processing algorithm Singh & Joachims ‘19

Evaluation Metrics

- **Inequality**: The percentage of all views that the top 1% of candidates (with respect to views) receive, Lazovich et al. ‘22. Referred to as **T1PS**.

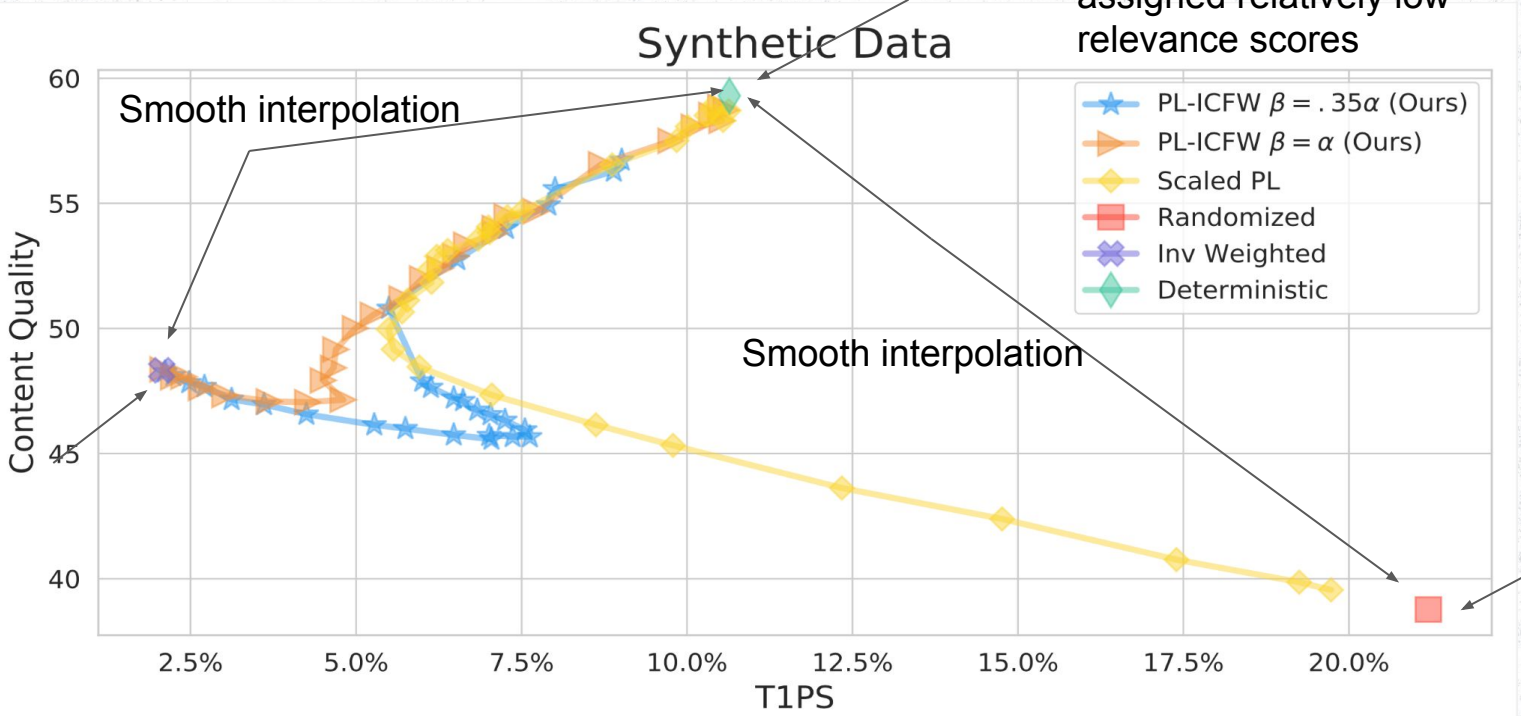
- **Model Performance**: The sum of the ground truth relevance scores of viewed candidates divided by total number of users. Referred to as **content quality**.

Synthetic Experiment Set-up

- n = 2000 users
- m = 1000 candidates
- k = 40 candidate set size
- ℓ = 10 candidates viewed per user
- Given candidate score c_i, we assign the candidate a relevance score of $r_i := \max\{0, 5 - c_i + x_i\}$ where $x_i \sim N(0, 1)$

Synthetic Experiment Results

Popular candidates are assigned relatively low relevance scores



Key Takeaways:
(1) Our approach outperforms the scaled PL baseline.The minimum T1PS that Scaled PL can achieve is 6% whereas our approach achieves 2% with nearly the same content quality.
(2) Randomized performs extremely poorly for T1PS.

German Credit Experiment Set-Up

Data:

- 1000 individuals seeking a loan from a bank
- Each labeled as good or bad risk
- Each has 29 features such as demographic information, financial history, education, employment, etc.

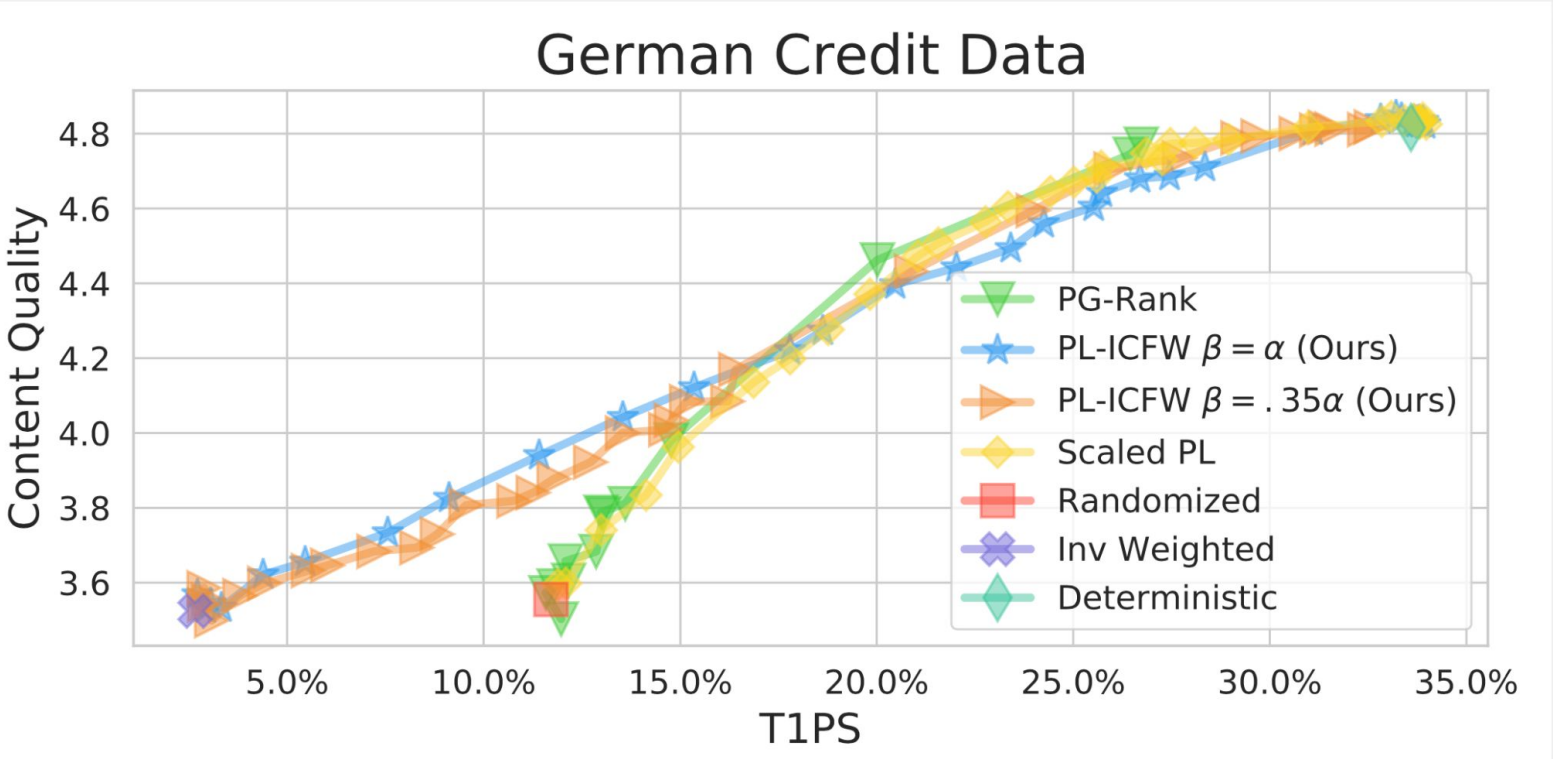
Relevance scores and baselines:

- Linear model trained to get relevance scores for all methods except PG-Rank
- PG-Rank trained for individual fairness with various hyperparameters
- In both cases, we used queries of size 10 such that in expectation 4 people have a positive label in each query

Simulation:

- m = 200 candidates
- n = 2000 users
- k = 15 candidate set size
- ℓ = 5 seen recommendations per user
- Popular items are the 10 items with the lowest 50-59 predicted scores

German Credit Experiment Results



Key Takeaways:
(1) Our approach outperforms the other baselines when T1PS < 20%.
(2) Randomized performs poorly for T1PS.

Limitations

1. Sensitivity of W_v, number of times candidate v appears in a candidate set, over different time periods.

2. Optimality of our algorithm unknown.

3. Sensitivity of our algorithm's hyperparameters to relevance scores.

4. If candidate sets have too much inequality, mitigating candidate set inequality may require intervention at the candidate generation step.

Key Takeaways:

- Common-sense solutions to “fair ranking” can behave unexpectedly when candidate sets are imbalanced.
- We proposed a simple, computationally inexpensive post-processing algorithm that interpolates between consumer-utility and producer-side exposure.