Mitigating Mainstream Bias in Recommendation via Cost-sensitive Learning

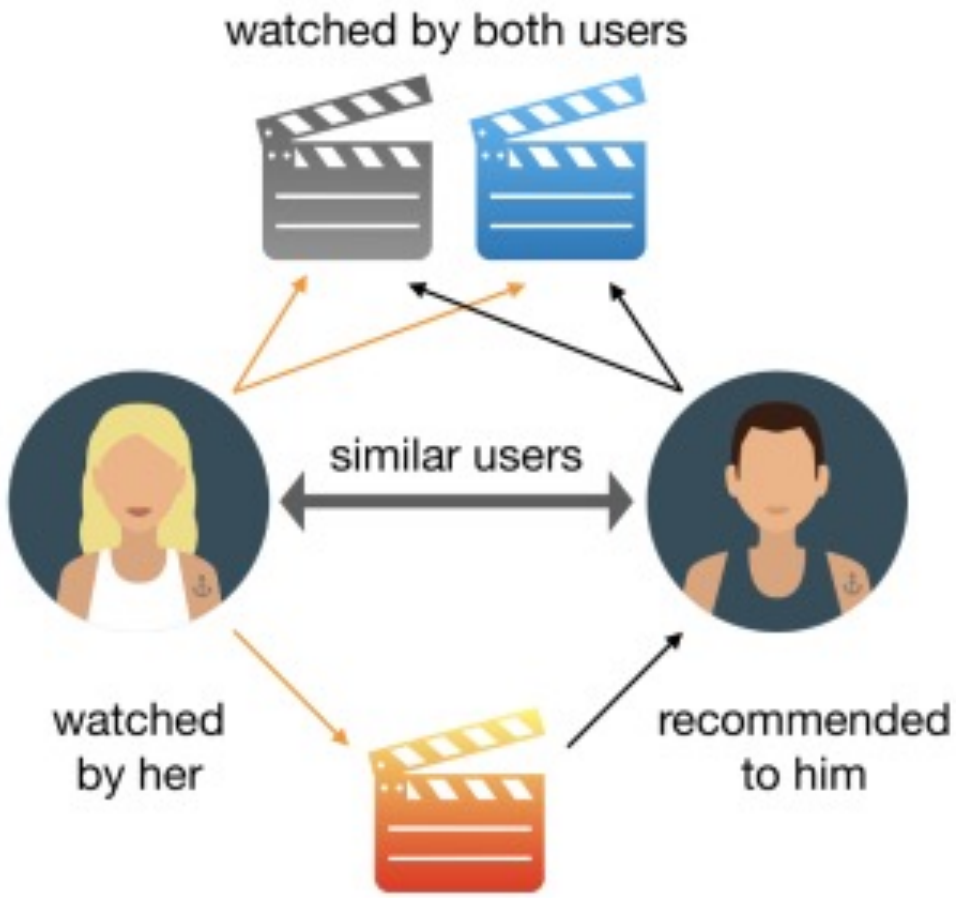
Roger Zhe Li, Julián Urbano, Alan Hanjalic

July 23, 2023 @ ICTIR

Zhe_Delft

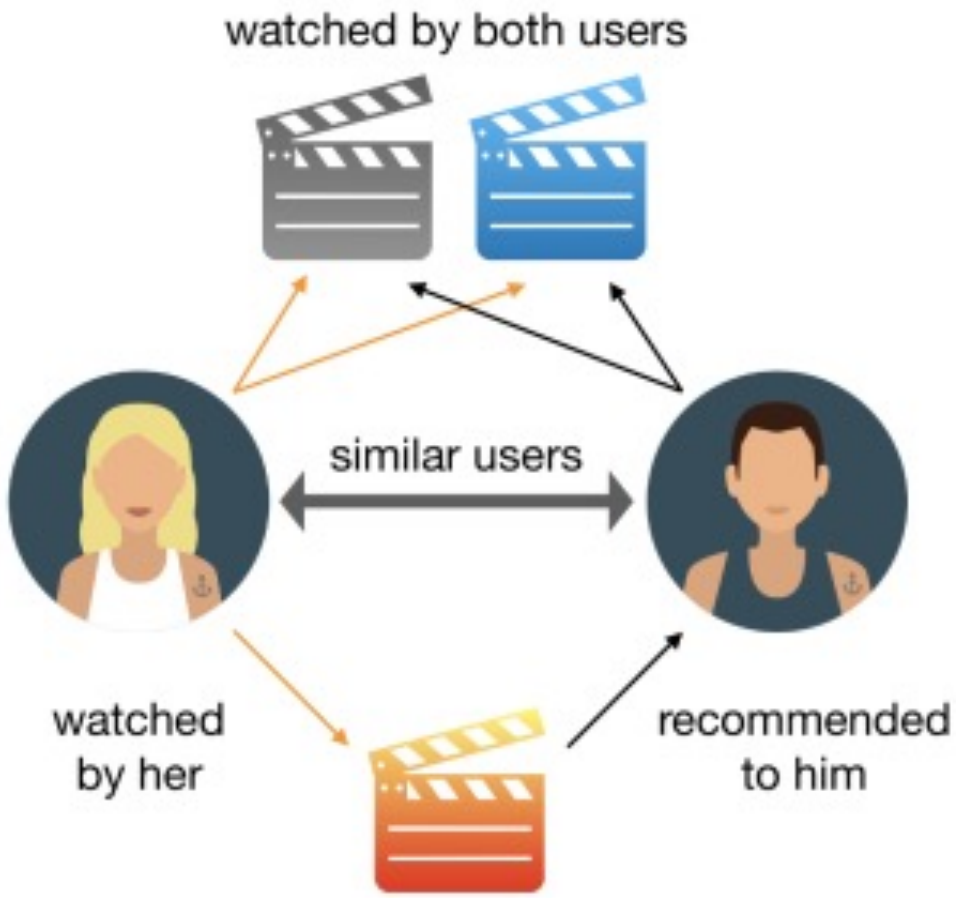
TU Delft

Collaborative Filtering & Similarity



https://vitalflux.com/recommender-systems-in-machine-learning-examples/

Collaborative Filtering & Similarity

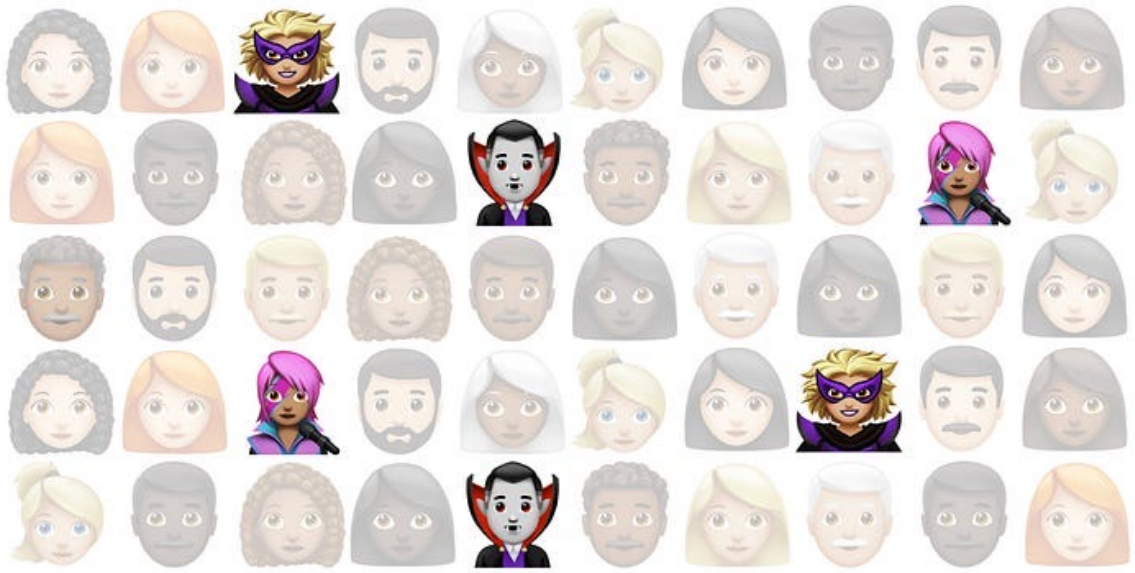


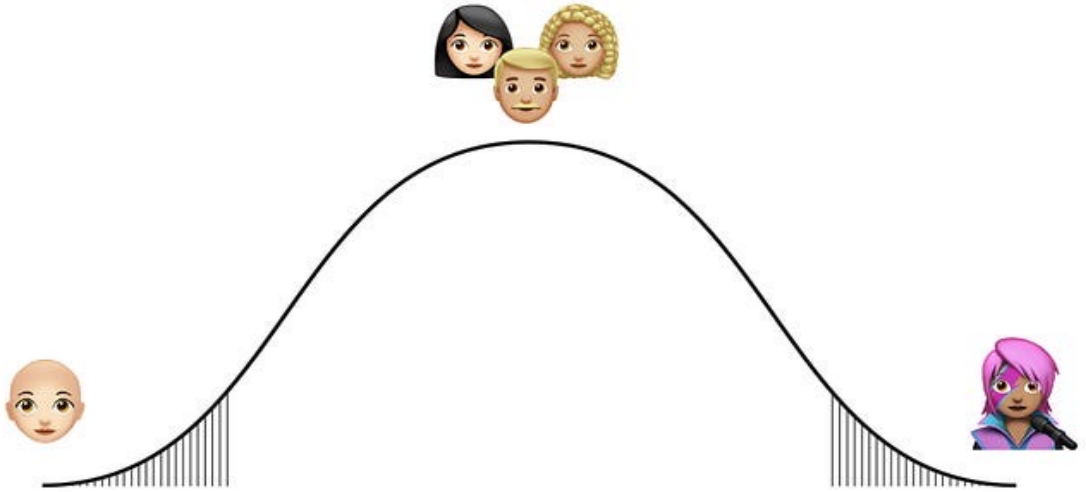
However, when…

- There are not enough data for similarity modeling
- There are users NOT similar to anyone

https://vitalflux.com/recommender-systems-in-machine-learning-examples/

Mainstream Bias in Collaborative Filtering



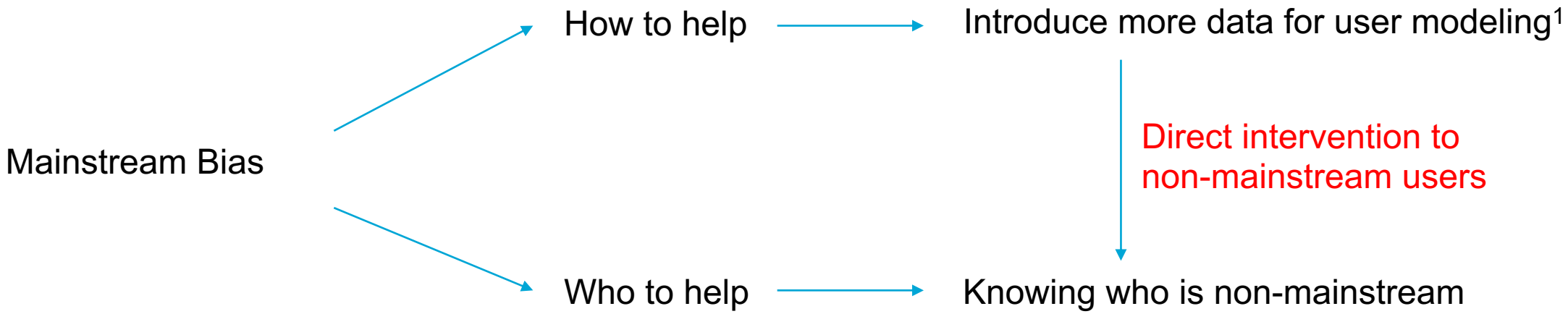


https://uxdesign.cc/the-fundamentals-of-engaging-with-extreme-users-45e0033e6b2

Mainstream Bias Mitigation

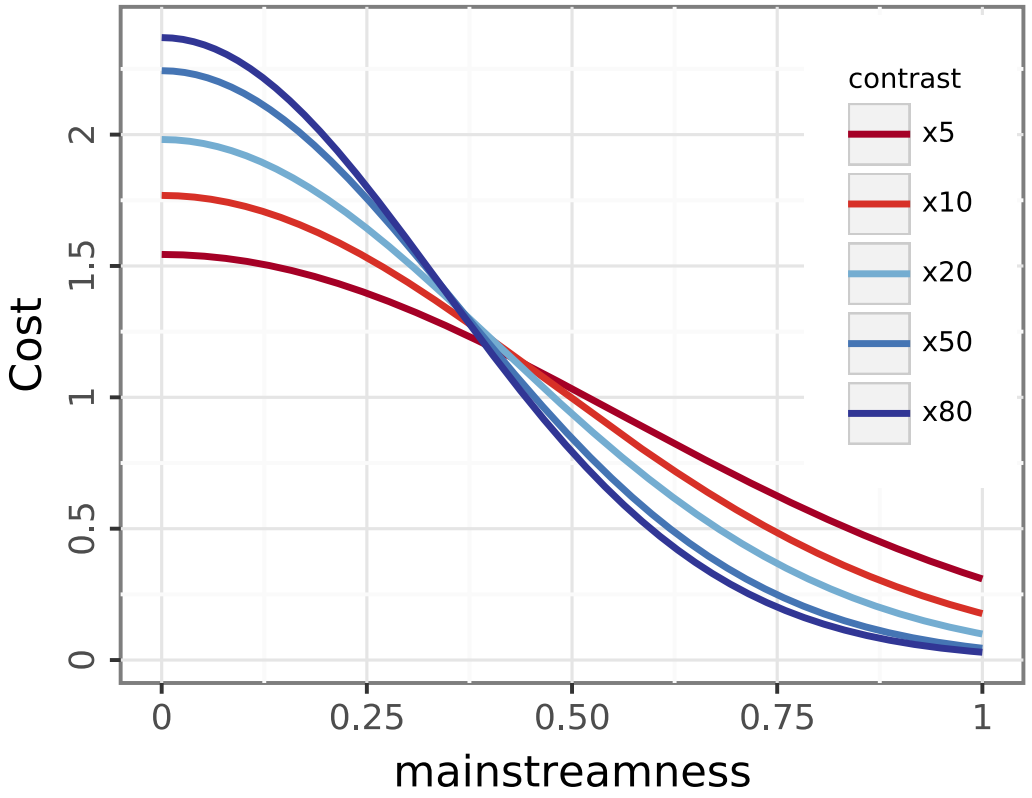
TU Delft

Pathways on Mainstream Bias Mitigation



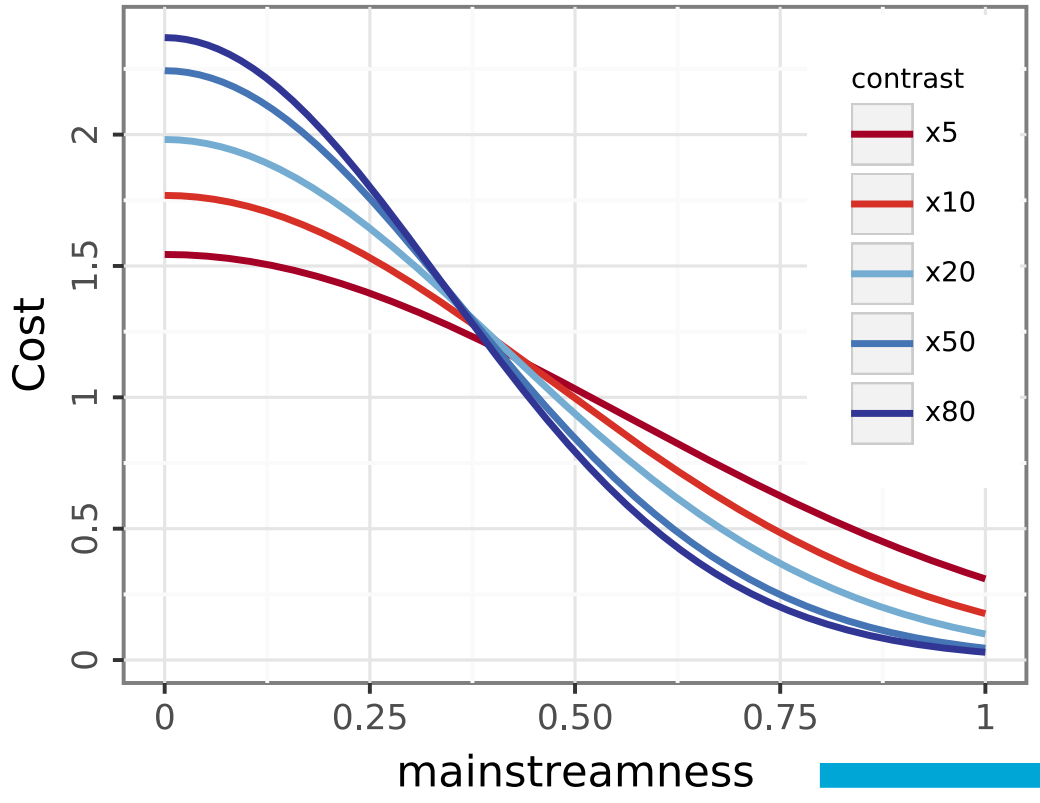
1 Leave No User Behind: Towards Improving the Utility of Recommender Systems for Non-mainstream Users

Intuition on Direct Intervention: Cost-sensitive Importance



$$\mathcal{L} = \sum_{u \in \mathcal{U}} \omega(u) \mathcal{L}_R(u)$$

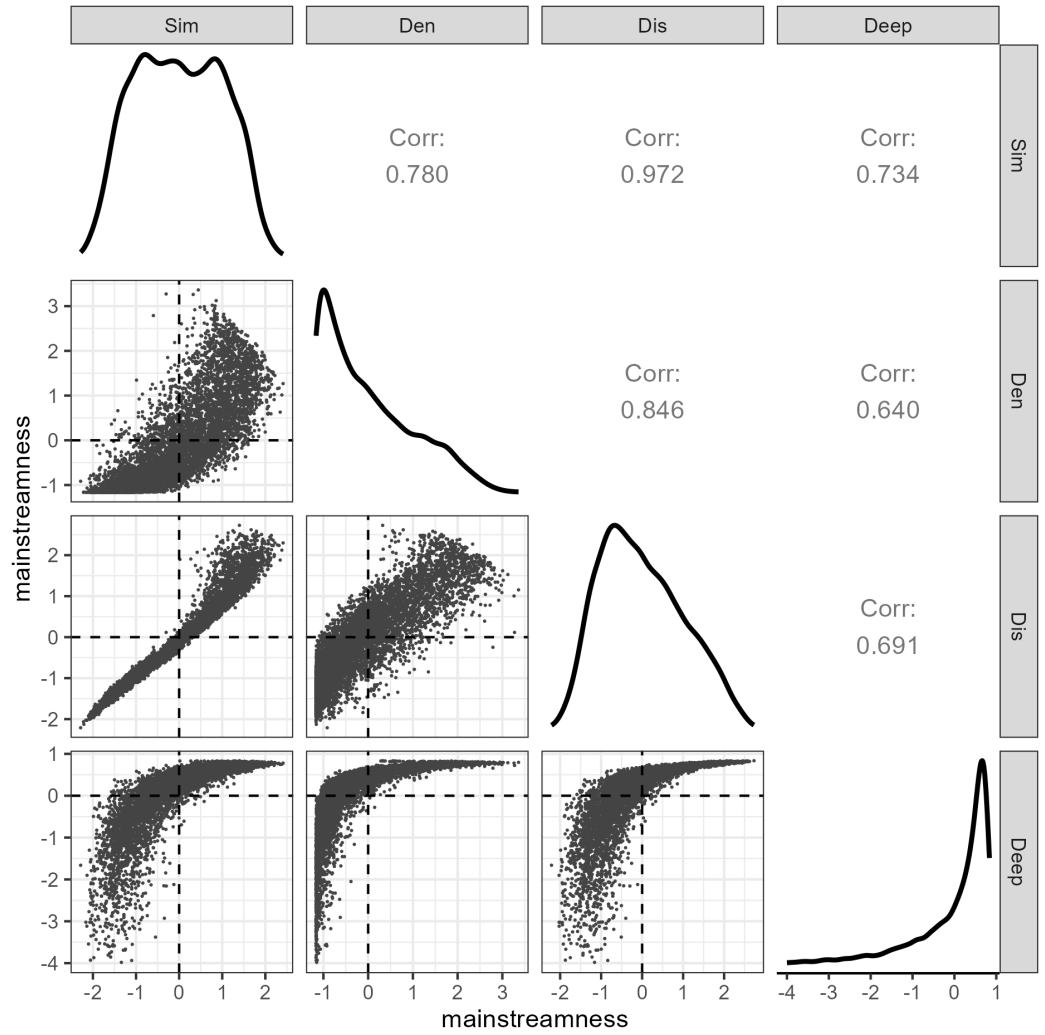
Intuition on Direct Intervention: Cost-sensitive Importance



$$\mathcal{L} = \sum_{u \in \mathcal{U}} \omega(u) \mathcal{L}_R(u)$$

HOW TO DEFINE?

Option 1: Explicit Measure/Quantify Mainstreamness



- Multiple different definitions;
- Inconsistency of user mainstreamness identification;
- Should the measurement of mainstreamness be model-agnostic?

Zhu, Ziwei, and James Caverlee. “Fighting mainstream bias in recommender systems via local fine tuning.” In WSDM 2022.

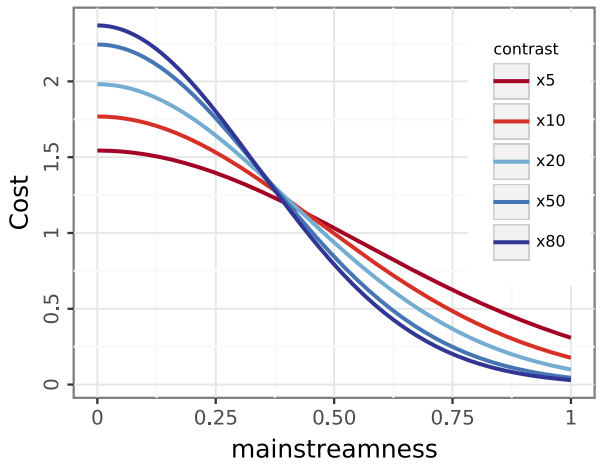
Option 2: Get a Proxy of Mainstreamness

- Expected outcome of mainstream bias: non-mainstream users suffer lower recommendation accuracy;
- Bias is intrinsic, but pronounced by the model.

Empirical Cumulative Distribution Function

$$\omega(u) = \text{cost}(\text{ecdf}_{\mathcal{U}}(m_u))$$

Recommendation accuracy



Experiments & Results

TU Delft

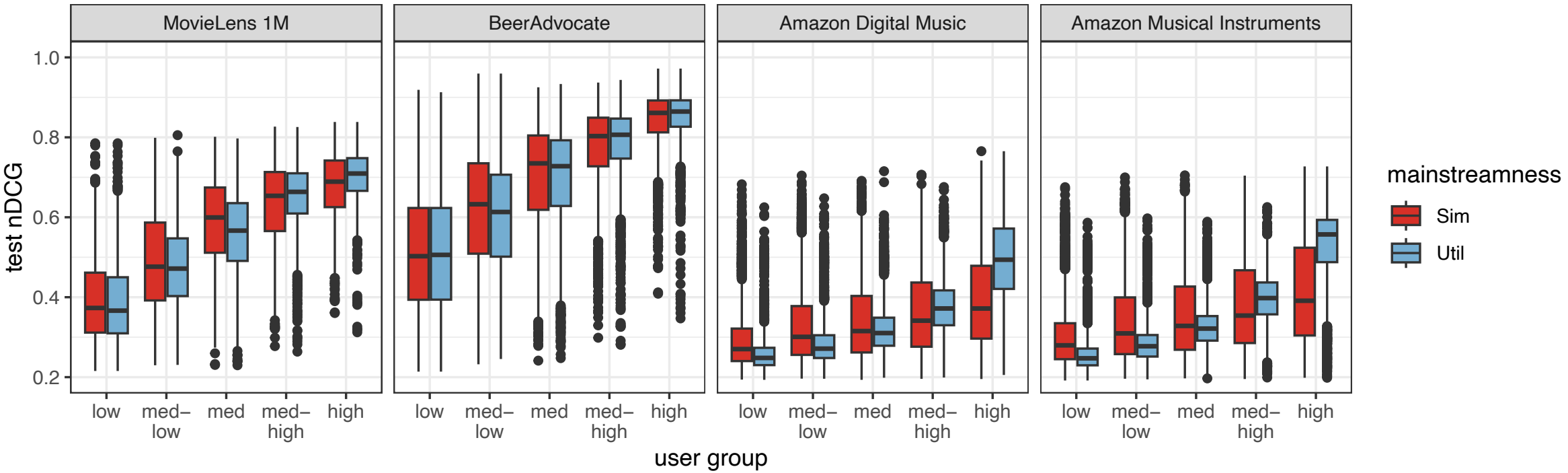
Datasets & Baselines

Dataset	#users	#items	#ratings	Density
MovieLens 1M [6]	6,040	3,609	562,957	2.583%
BeerAdvocate [13]	8,821	43,663	780,752	0.203%
Amazon Digital Music [15]	14,057	379,171	619,673	0.011%
Amazon Musical Instruments [15]	15,270	585,766	862,798	0.010%

- Binarization: all ratings are regarded relevant;
- FM as the model to get m_u;
- nDCG as the metric to measure m_u;
- Always with at least 5 items per user for train/val/test. Datasets are evenly split;
- Baselines: FM and one of the four model-agnostic strategies (Sim);
- Measure the recommendation performance in 5 buckets, as per the Sim or nDCG (Util) scores on vanilla FM models. Lowest nDCG scores are interpreted as least mainstream, and higher nDCG scores mean more mainstream.

TU Delft

Effectiveness of Accuracy as Mainstreamness Proxy

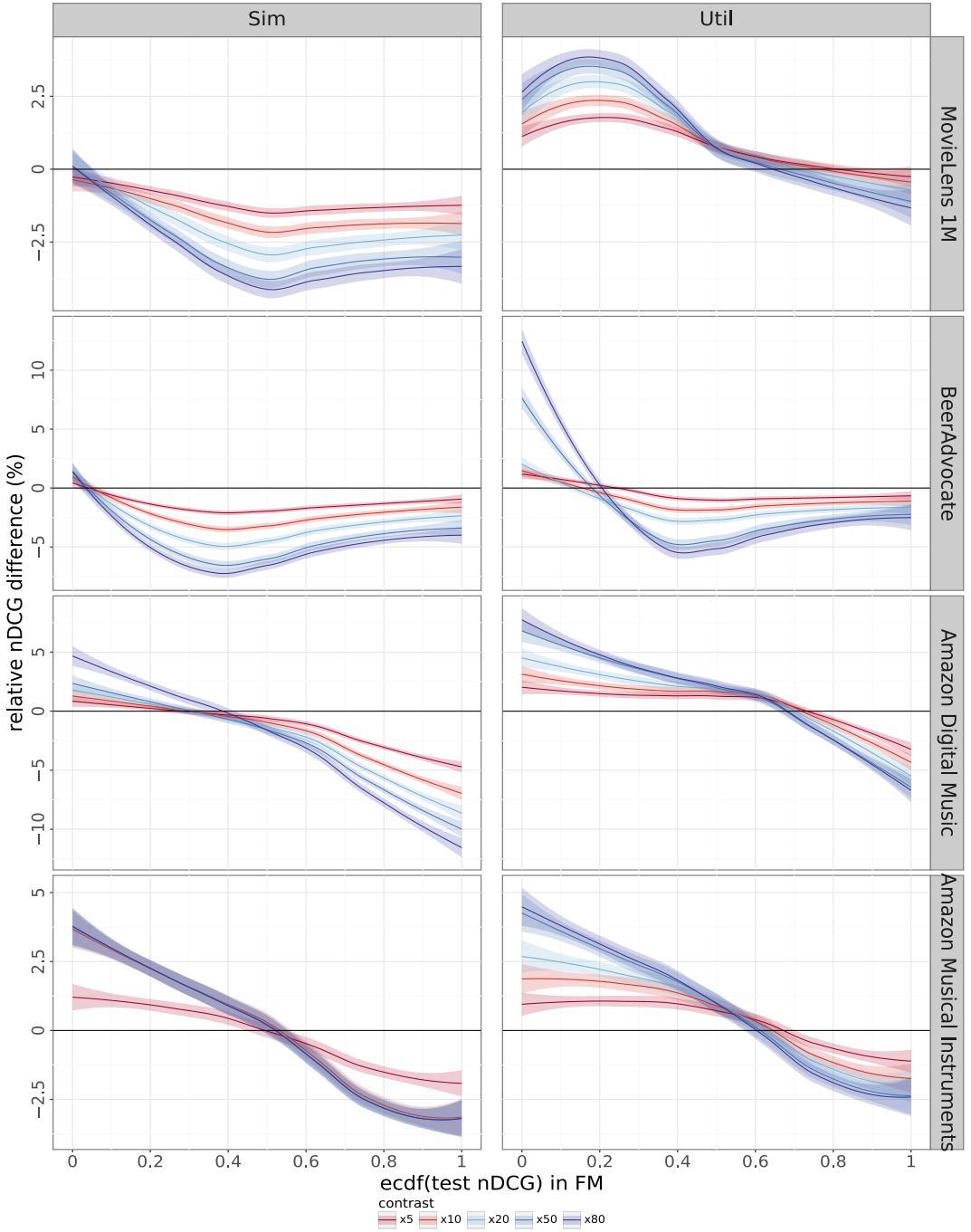


Effect of Bias Mitigation: Split by User Accuracy (Util)

		MovieLens 1M						BeerAdvocate						Amazon Digital Music						Amazon Musical Instruments					
		Overall	low	med-low	med	med-high	high	Overall	low	med-low	med	med-high	high	Overall	low	med-low	med	med-high	high	Overall	low	med-low	med	med-high	high
FM		.5531	.3284	.4621	.5753	.6613	.7388	.6887	.4144	.6051	.7301	.8132	.8809	.3456	.2324	.2695	.3145	.3828	.5289	.3606	.2348	.2772	.3276	.4085	.5552
Sim	x5	.5465	-0.36	-0.89	-1.62	-1.3	-1.33	.6792	-0.52	-1.72	-1.83	-1.44	-1.13	.3395	+0.65	+0.05	-0.61	-1.8	-4.45	.3581	+1.61	+0.77	-0.09	-1.06	-2.5
	x10	.5437	-0.47	-1.32	-2.23	-1.87	-1.94	.6734	-0.81	-2.87	-2.94	-2.27	-1.8	.3368	+0.99	+0.1	-1	-2.67	-6.27	.3577	+3.29	+1.44	+0.08	-1.68	-3.54
	x20	.541	-0.53	-1.76	-2.85	-2.41	-2.51	.6666	-1.35	-4.23	-4.13	-3.22	-2.56	.3347	+1.31	+0.09	-1.3	-3.46	-7.69	.3576	+3.33	+1.44	+0.04	-1.72	-3.58
	x50	.5376	-0.67	-2.28	-3.64	-3.06	-3.26	.6588	-2.01	-5.65	-5.56	-4.29	-3.54	.3328	+1.63	+0.07	-1.6	-4.13	-8.92	.3576	+3.37	+1.45	+0.03	-1.75	-3.6
	x80	.5359	-0.67	-2.55	-4.06	-3.39	-3.59	.6548	-2.55	-6.39	-6.17	-4.84	-4.08	.3316	+3.66	+0.9	-1.78	-5.07	-10.62	.3576	+3.39	+1.45	+0.02	-1.77	-3.61
Util	x5	.5567	+1.67	+1.81	+0.63	+0.16	-0.13	.6846	+0.63	-0.41	-0.94	-0.83	-0.77	.3454	+1.59	+1.43	+1.2	+0.4	-2.58	.3607	+1.1	+0.96	+0.62	-0.04	-1.19
	x10	.5574	+2.38	+2.34	+0.7	+0.11	-0.27	.6807	+0.44	-1.19	-1.66	-1.39	-1.27	.3453	+2.54	+2.09	+1.53	+0.18	-3.5	.3607	+2	+1.52	+0.78	-0.32	-1.8
	x20	.5579	+3.05	+2.87	+0.73	0	-0.48	.6762	+0.54	-2.11	-2.67	-2.09	-1.74	.3454	+3.59	+2.78	+1.64	+0.11	-4.25	.3607	+2.63	+1.93	+0.8	-0.53	-2.08
	x50	.5579	+3.62	+3.31	+0.68	-0.21	-0.84	.6722	+2	-3.09	-3.86	-2.89	-2.32	.3458	+4.84	+3.67	+1.92	-0.11	-4.9	.3608	+3.94	+2.48	+0.85	-0.83	-2.66
	x80	.5577	+3.89	+3.47	+0.63	-0.32	-1.05	.6715	+2.98	-3.2	-4.25	-3.14	-2.54	.346	+5.45	+4	+2.02	-0.18	-5.15	.3608	+4.48	+2.64	+0.88	-0.97	-2.86

TU Delft

Effect of Bias Mitigation: Split by User Accuracy (Util)

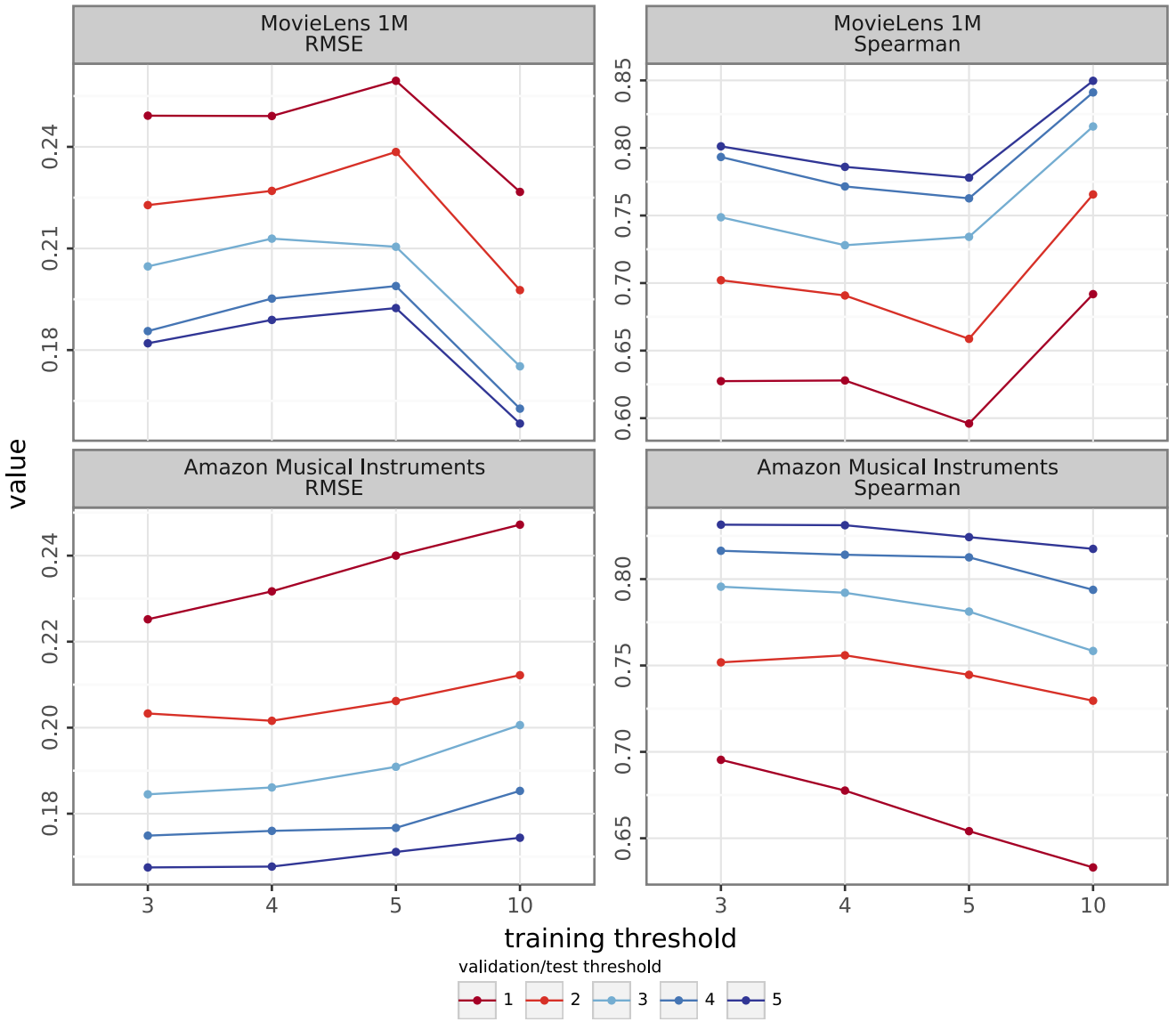


Discussion: Are the Results on the Individual User Level Reliable?

TU Delft

Data for producing reliable results on the indivisual user level

- Enough items for each user to be well trained;
- Enough items for validation/evaluation;
- Achieve higher user coverage.



Conclusions

- Direct intervention by focusing more on non-mainstream users can help mitigate mainstream bias;
- User recommendation accuracy on a traditional recommendation model is an implicit but effective proxy of mainstreamness;
- The reliable results on the research of individual user performance are subject to sufficient data for training/validation/testing.

Code & Data: https://github.com/roger-zhe-li/ictir23-cost-sensitive

TU Delft