On the challenges of studying bias in Recommender Systems: A UserKNN case study

Savvina Daniil, Manel Slokom, Mirjam Cuper, Cynthia Liem, Jacco van Ossenbruggen, Laura Hollink

Cultural AI
a lab for culturally valued AI

Introduction

Studying bias in recommender systems

Bias...

...is inherently related to **data and algorithm properties.**

Introduction

Studying bias in recommender systems

Bias...

...is inherently related to **data and algorithm properties**.

Reproducibility...

...issues are **present in RS research**, with often limited reporting of algorithm instantiation details.

Introduction

Studying bias in recommender systems

Bias...

...is inherently related to **data and algorithm properties**.

Reproducibility...

...issues are **present in RS research**, with often limited reporting of algorithm instantiation details.

Benchmarking...

...on datasets is very useful, but research is **bounded by the specific properties** of the very few available datasets.

Introduction

Studying bias in recommender systems

Bias...

...is inherently related to **data and algorithm properties**.

Reproducibility...

...issues are **present in RS research**, with often limited reporting of algorithm instantiation details.

Benchmarking...

...on datasets is very useful, but research is **bounded by the specific properties** of the very few available datasets.

Challenges?

Experiment with **algorithm configurations** & **data characteristics,** observe the effect on **popularity bias**.

Exploring the effects

Combinations of data and algorithm characteristics —> popularity bias?

Exploring the effects

Combinations of data and algorithm characteristics —> popularity bias?

Approach

- Identify data characteristics that relate to popularity bias, and **generate data** accordingly.
- Identify important **configurations** of UserKNN.
- Evaluate bias for **each of the combinations**.
- Offer **insights** on when bias can occur.

Popularity bias by UserKNN

Data characteristics that may affect popularity bias

Popularity bias by UserKNN

Data characteristics that may affect popularity bias

Data characteristics

1. **Relation between rating and popularity**; do the popular items also have high ratings?
2. **Influential users**; what do users with big profiles like?

Generate synthetic data

Define data scenarios according to the observations mentioned.

5 data scenarios

Based on Book-Crossing

Generate synthetic data

Define data scenarios according to the observations mentioned.

Scenario 1

There is no relation between popularity and rating.

5 data scenarios

Based on Book-Crossing

Generate synthetic data

Define data scenarios according to the observations mentioned.

Scenario 1

There is no relation between popularity and rating.

Scenario 2

Popular items are highly rated.

5 data scenarios

Based on Book-Crossing

Generate synthetic data

Define data scenarios according to the observations mentioned.

Scenario 1

There is no relation between popularity and rating.

Scenario 2

Popular items are highly rated.

5 data scenarios

Scenario 3

Popular items are low rated.

Based on Book-Crossing

Generate synthetic data

Define data scenarios according to the observations mentioned.

Scenario 1

There is no relation between popularity and rating.

Scenario 2

Popular items are highly rated.

5 data scenarios

Scenario 3

Popular items are low rated.

Scenario 4

Popular items are highly rated by users with big (top20%) profiles.

Based on Book-Crossing

Generate synthetic data

Define data scenarios according to the observations mentioned.

5 data scenarios

Scenario 1: There is no relation between popularity and rating.

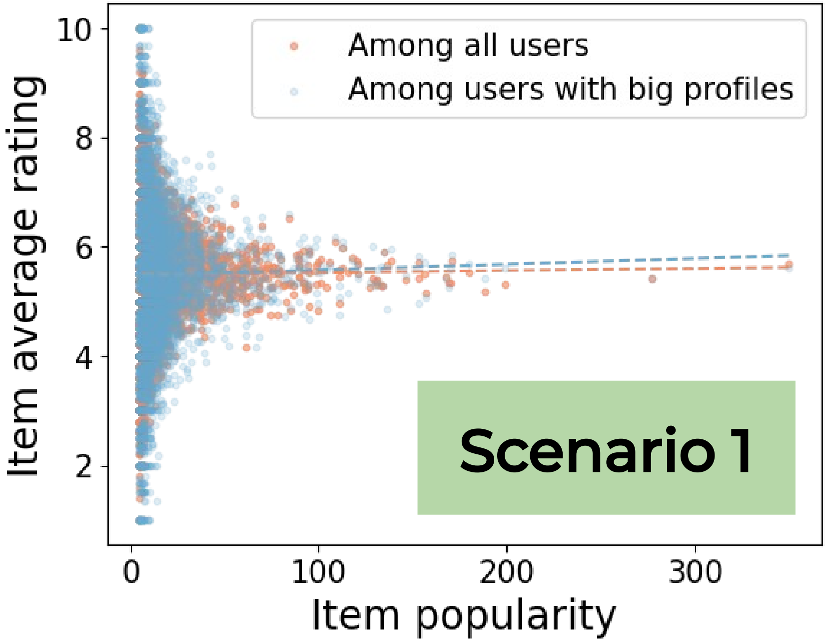
Scenario 2: Popular items are highly rated.

Scenario 3: Popular items are low rated.

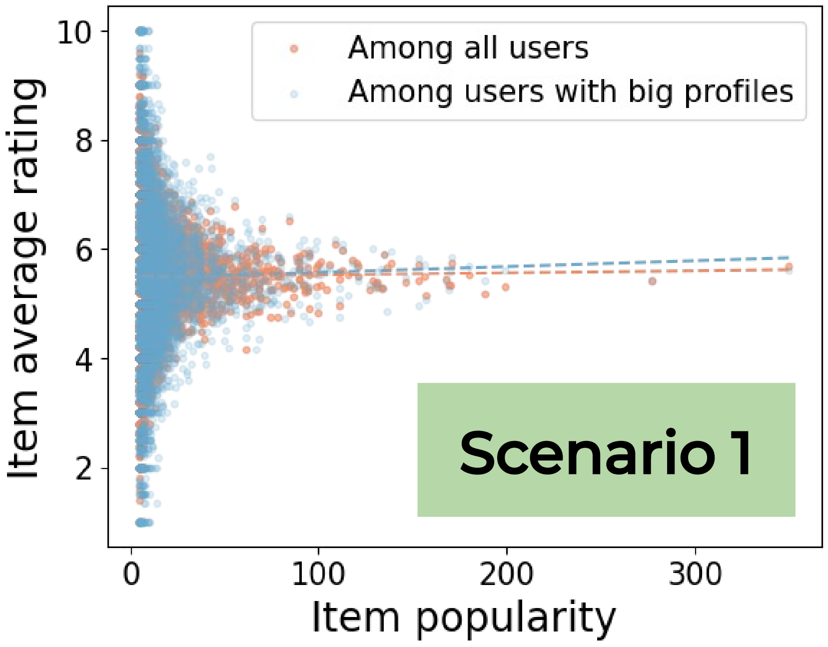
Scenario 4: Popular items are highly rated by users with big (top20%) profiles.

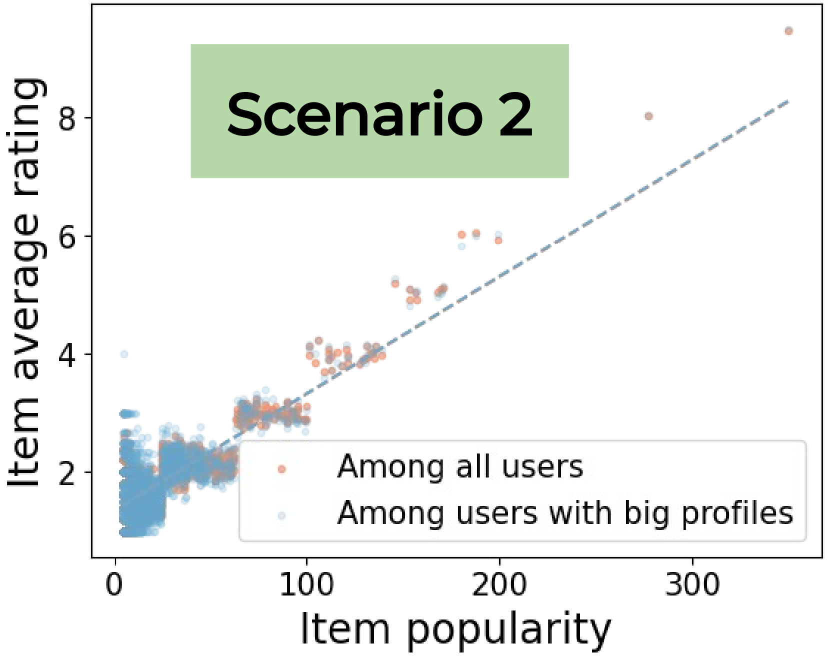
Scenario 5: Popular items are low rated by users with big profiles.

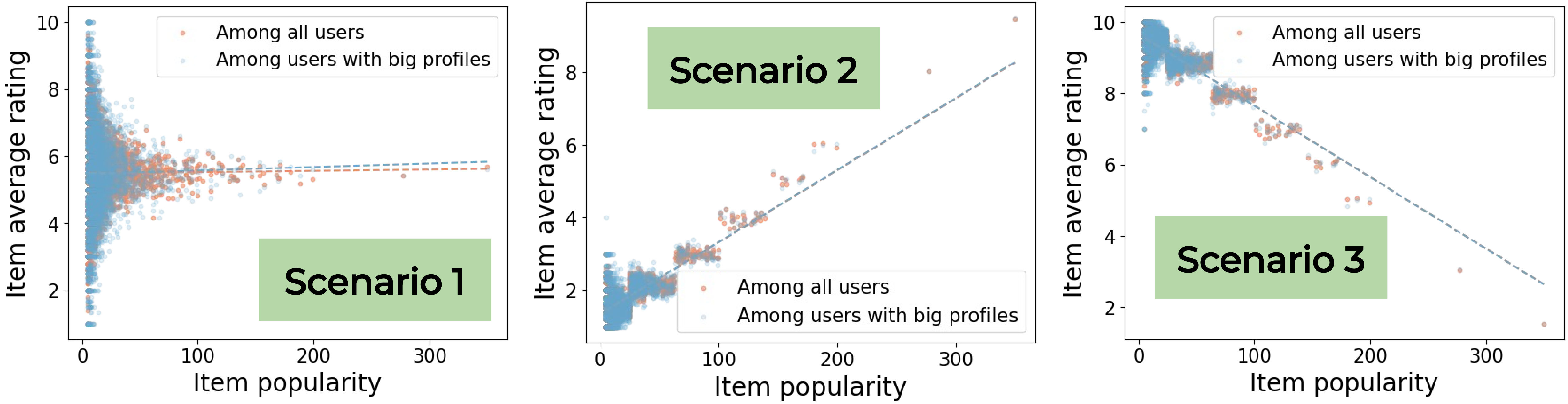
Based on Book-Crossing



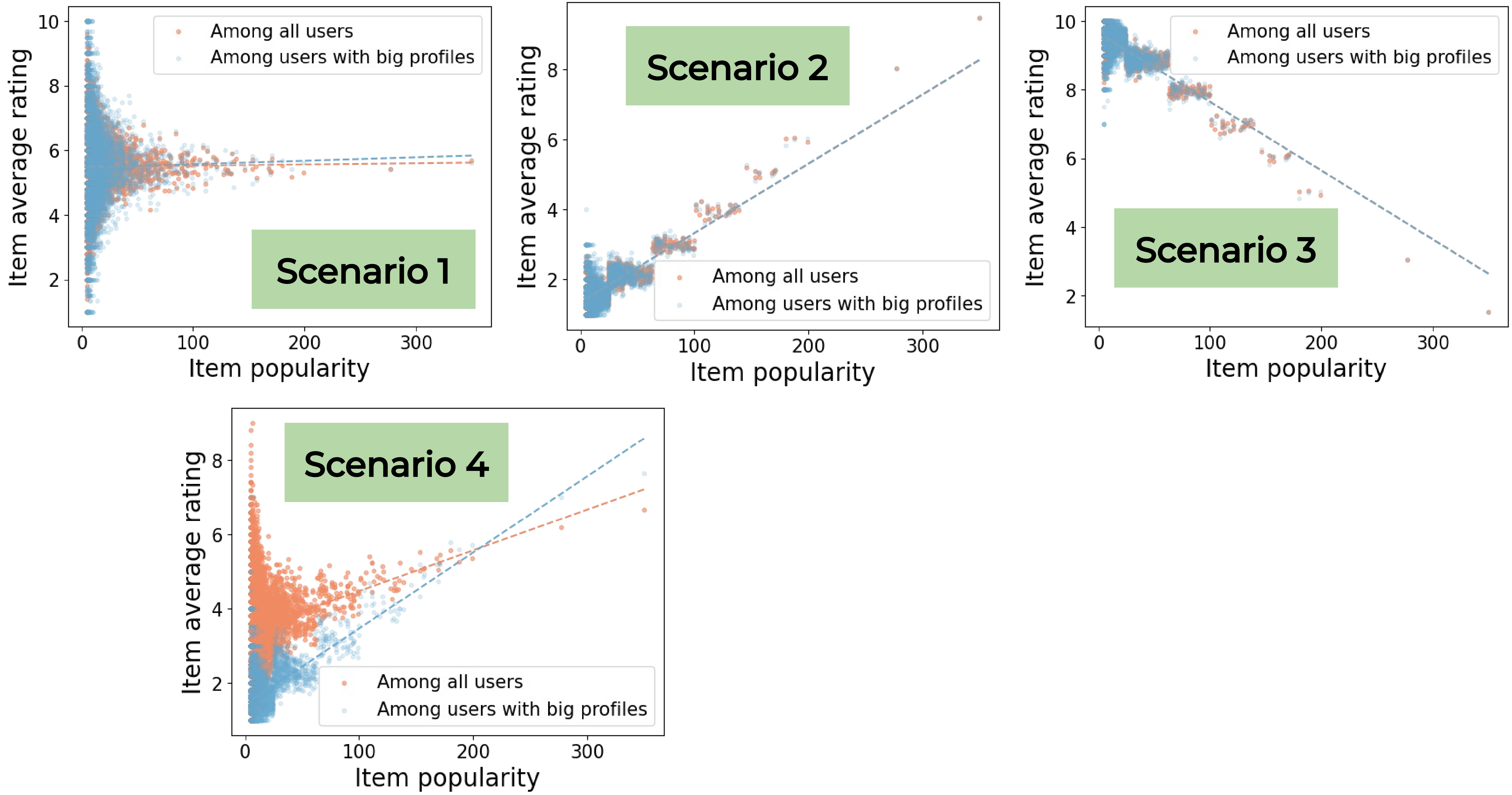
Among all users
Among users with big profiles
Scenario 1
Item average rating
Item popularity
10
8
6
4
2
0
100
200
300

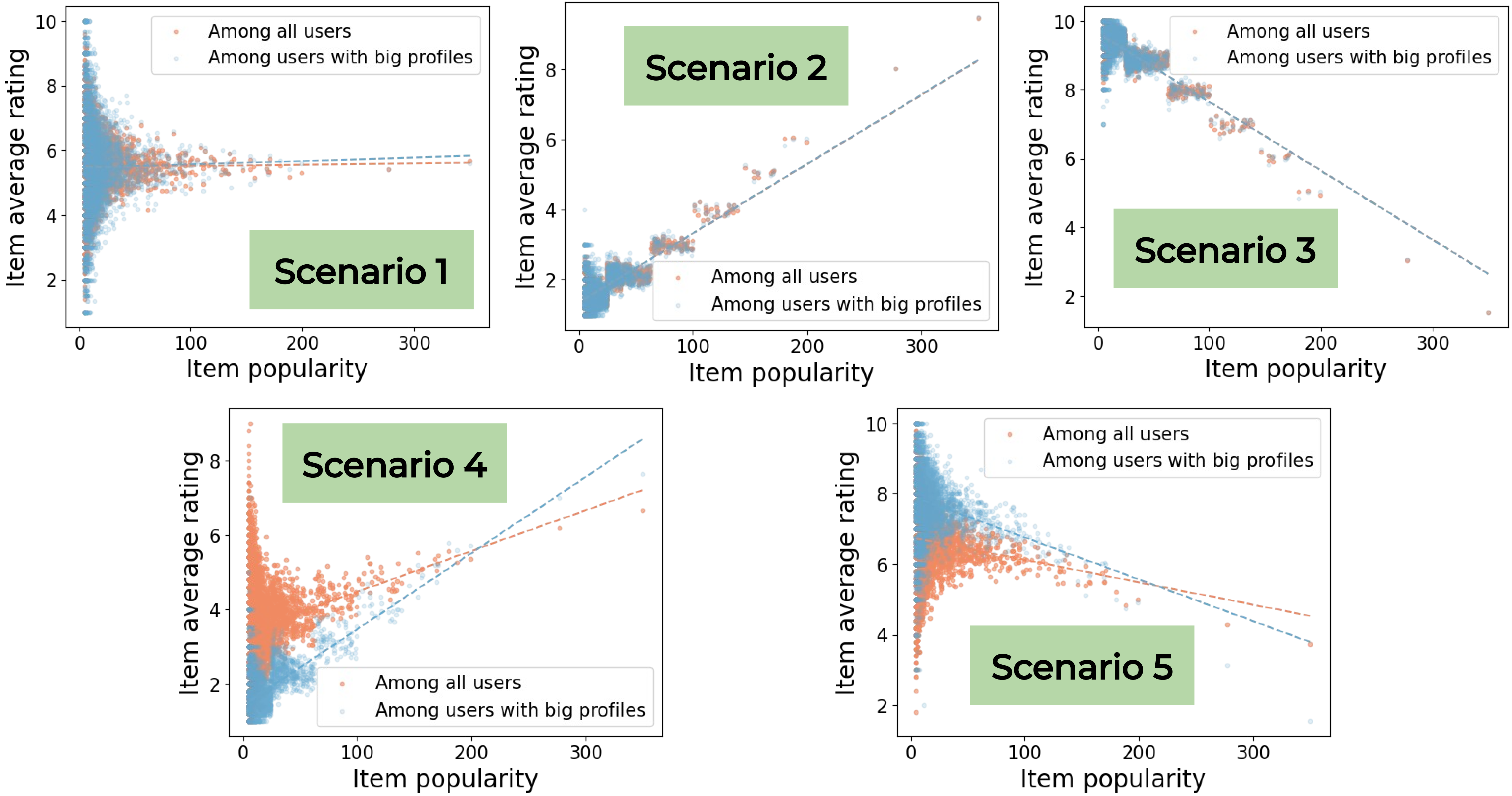

Among all users
Among users with big profiles
Scenario 1
Item average rating
Item popularity
10
8
6
4
2
0
100
200
300


Scenario 2
Among all users
Among users with big profiles
Item average rating
Item popularity
8
6
4
2
0
100
200
300



Item average rating
Item popularity
Among all users
Among users with big profiles
Scenario 1
Scenario 2
Scenario 3


Scenario 1
Among all users
Among users with big profiles
Item average rating
Item popularity
Scenario 2
Among all users
Among users with big profiles
Item average rating
Item popularity
Scenario 3
Among all users
Among users with big profiles
Item average rating
Item popularity
Scenario 4
Among all users
Among users with big profiles
Item average rating
Item popularity


Scenario 1
Scenario 2
Scenario 3
Scenario 4
Scenario 5
Among all users
Among users with big profiles
Item average rating
Item popularity

Popularity bias by UserKNN

UserKNN configurations that differ between frameworks

Popularity bias by UserKNN

UserKNN configurations that differ between frameworks

UserKNN configurations

Configuration choice	Lenskit for Python	Cornac
Minimum neighbors	Configurable (defaults to 1)	1
Minimum similarity	Configurable (defaults to 0)	-1
Items for similarity	All items	Common items

How many neighbours are needed?

How similar can neighbours be?

What if we have rated only one item in common?

Experimental setup (A)

- **Configurations tested**:
  - Min. similarity 0, over all items, 1 min. neighbour. — LKPY
  - Min. similarity 0, over all items, 2 min. neighbours. — LKPY
  - Min. similarity -1, over all items, 1 min. neighbour. — LKPY
  - Min. similarity -1, over all items, 2 min. neighbours. — LKPY
  - Min. similarity -1, over common items, 1 min. neighbour. — Cornac

Experimental setup (B)

- **For every configuration tested:**
  - Optimize the rest of the hyperparameters.
  - Train and test rating prediction (5-fold cross validation).
  - Rank based on rating and recommend the top 10.
  - Measure:
    - RMSE, NDCG@10
    - Popularity Correlation (PopCorr)
    - Average Recommendation Popularity (ARP)
    - Popularity Lift (PL)

Data Scenario	Min Sim	Over Common	Min Nbrs	Pop Corr↑	ARP↑	PL↑	RMSE↓	NDCG @10↑
Scenario 1	-1	False	1	0.018	0.002*	-32.285*	3.502	0.001
			2	0.418	0.004*	21.252*	3.352	0.003
		True	1	0.004	0.002*	-35.746*	3.337	0.001
	0	False	1	0.101	0.003*	-12.827*	3.624	0.002
			2	**0.615**	**0.005**	**65.440**	3.464	0.005
Scenario 2	-1	False	1	0.596	0.021*	426.621*	1.188	0.019
			2	**0.614**	0.022*	447.618*	1.190	0.021
		True	1	0.604	0.015*	305.197*	1.150	0.013
	0	False	1	0.552	**0.027**	**632.300**	1.040	0.023
			2	0.562	0.027	591.966	1.026	0.025
Scenario 3	-1	False	1	0.559	0.008*	187.197*	1.182	0.002
			2	**0.728**	**0.008**	**192.127**	1.182	0.002
		True	1	0.522	0.006*	151.686*	1.151	0.001
	0	False	1	0.025	0.002*	-35.765*	1.044	0.001
			2	0.161	0.003*	-13.100*	1.034	0.004
Scenario 4	-1	False	1	0.253	0.003*	23.063*	2.502	0.001
			2	**0.772**	0.006*	97.490*	2.404	0.004
		True	1	0.184	0.003*	8.669*	2.458	0.001
	0	False	1	0.588	0.008*	164.549*	2.500	0.004
			2	0.701	**0.014**	**297.047**	2.386	0.010
Scenario 5	-1	False	1	0.087	0.003*	-7.924*	2.880	0.001
			2	**0.623**	0.005*	**57.969**	2.776	0.003
		True	1	0.057	0.002*	-16.243*	2.783	0.001
	0	False	1	0.136	0.003*	-16.122*	2.914	0.003
			2	0.612	**0.005**	42.849	2.794	0.006

Observations

- Scenarios impact popularity bias

Data Scenario	Min Sim	Over Common	Min Nbrs	Pop Corr↑	ARP↑	PL↑	RMSE↓	NDCG @10↑
Scenario 1	-1	False	1	0.018	0.002*	-32.285*	3.502	0.001
			2	0.418	0.004*	21.252*	3.352	0.003
		True	1	0.004	0.002*	-35.746*	3.337	0.001
	0	False	1	0.101	0.003*	-12.827*	3.624	0.002
			2	**0.615**	**0.005**	**65.440**	3.464	0.005
Scenario 2	-1	False	1	0.596	0.021*	426.621*	1.188	0.019
			2	**0.614**	0.022*	447.618*	1.190	0.021
		True	1	0.604	0.015*	305.197*	1.150	0.013
	0	False	1	0.552	**0.027**	**632.300**	1.040	0.023
			2	0.562	0.027	591.966	1.026	0.025
Scenario 3	-1	False	1	0.559	0.008*	187.197*	1.182	0.002
			2	**0.728**	**0.008**	**192.127**	1.182	0.002
		True	1	0.522	0.006*	151.686*	1.151	0.001
	0	False	1	0.025	0.002*	-35.765*	1.044	0.001
			2	0.161	0.003*	-13.100*	1.034	0.004
Scenario 4	-1	False	1	0.253	0.003*	23.063*	2.502	0.001
			2	**0.772**	0.006*	97.490*	2.404	0.004
		True	1	0.184	0.003*	8.669*	2.458	0.001
	0	False	1	0.588	0.008*	164.549*	2.500	0.004
			2	0.701	**0.014**	**297.047**	2.386	0.010
Scenario 5	-1	False	1	0.087	0.003*	-7.924*	2.880	0.001
			2	**0.623**	0.005*	**57.969**	2.776	0.003
		True	1	0.057	0.002*	-16.243*	2.783	0.001
	0	False	1	0.136	0.003*	-16.122*	2.914	0.003
			2	0.612	**0.005**	42.849	2.794	0.006

Observations

- Scenarios impact popularity bias
- Increasing minimum neighbours generally increases popularity bias

Data Scenario	Min Sim	Over Common	Min Nbrs	Pop Corr↑	ARP↑	PL↑	RMSE↓	NDCG @10↑
Scenario 1	-1	False	1	0.018	0.002*	-32.285*	3.502	0.001
			2	0.418	0.004*	21.252*	3.352	0.003
		True	1	0.004	0.002*	-35.746*	3.337	0.001
	0	False	1	0.101	0.003*	-12.827*	3.624	0.002
			2	**0.615**	**0.005**	**65.440**	3.464	0.005
Scenario 2	-1	False	1	0.596	0.021*	426.621*	1.188	0.019
			2	**0.614**	0.022*	447.618*	1.190	0.021
		True	1	0.604	0.015*	305.197*	1.150	0.013
	0	False	1	0.552	**0.027**	**632.300**	1.040	0.023
			2	0.562	0.027	591.966	1.026	0.025
Scenario 3	-1	False	1	0.559	0.008*	187.197*	1.182	0.002
			2	**0.728**	**0.008**	**192.127**	1.182	0.002
		True	1	0.522	0.006*	151.686*	1.151	0.001
	0	False	1	0.025	0.002*	-35.765*	1.044	0.001
			2	0.161	0.003*	-13.100*	1.034	0.004
Scenario 4	-1	False	1	0.253	0.003*	23.063*	2.502	0.001
			2	**0.772**	0.006*	97.490*	2.404	0.004
		True	1	0.184	0.003*	8.669*	2.458	0.001
	0	False	1	0.588	0.008*	164.549*	2.500	0.004
			2	0.701	**0.014**	**297.047**	2.386	0.010
Scenario 5	-1	False	1	0.087	0.003*	-7.924*	2.880	0.001
			2	**0.623**	0.005*	**57.969**	2.776	0.003
		True	1	0.057	0.002*	-16.243*	2.783	0.001
	0	False	1	0.136	0.003*	-16.122*	2.914	0.003
			2	0.612	**0.005**	42.849	2.794	0.006

Observations

- Scenarios impact popularity bias
- Increasing minimum neighbours generally increases popularity bias
- Minimum similarity: see scenario 3

Data Scenario	Min Sim	Over Common	Min Nbrs	Pop Corr↑	ARP↑	PL↑	RMSE↓	NDCG @10↑
Scenario 1	-1	False	1	0.018	0.002*	-32.285*	3.502	0.001
			2	0.418	0.004*	21.252*	3.352	0.003
		True	1	0.004	0.002*	-35.746*	3.337	0.001
	0	False	1	0.101	0.003*	-12.827*	3.624	0.002
			2	**0.615**	**0.005**	**65.440**	3.464	0.005
Scenario 2	-1	False	1	0.596	0.021*	426.621*	1.188	0.019
			2	**0.614**	0.022*	447.618*	1.190	0.021
		True	1	0.604	0.015*	305.197*	1.150	0.013
	0	False	1	0.552	**0.027**	**632.300**	1.040	0.023
			2	0.562	0.027	591.966	1.026	0.025
Scenario 3	-1	False	1	0.559	0.008*	187.197*	1.182	0.002
			2	**0.728**	**0.008**	**192.127**	1.182	0.002
		True	1	0.522	0.006*	151.686*	1.151	0.001
	0	False	1	0.025	0.002*	-35.765*	1.044	0.001
			2	0.161	0.003*	-13.100*	1.034	0.004
Scenario 4	-1	False	1	0.253	0.003*	23.063*	2.502	0.001
			2	**0.772**	0.006*	97.490*	2.404	0.004
		True	1	0.184	0.003*	8.669*	2.458	0.001
	0	False	1	0.588	0.008*	164.549*	2.500	0.004
			2	0.701	**0.014**	**297.047**	2.386	0.010
Scenario 5	-1	False	1	0.087	0.003*	-7.924*	2.880	0.001
			2	**0.623**	0.005*	**57.969**	2.776	0.003
		True	1	0.057	0.002*	-16.243*	2.783	0.001
	0	False	1	0.136	0.003*	-16.122*	2.914	0.003
			2	0.612	**0.005**	42.849	2.794	0.006

Observations

- Scenarios impact popularity bias
- Increasing minimum neighbours generally increases popularity bias
- Minimum similarity: see scenario 3
- Items consider for similarity: see scenario 4

Data Scenario	Min Sim	Over Common	Min Nbrs	Pop Corr↑	ARP↑	PL↑	RMSE↓	NDCG @10↑
Scenario 1	-1	False	1	0.018	0.002*	-32.285*	3.502	0.001
			2	0.418	0.004*	21.252*	3.352	0.003
		True	1	0.004	0.002*	-35.746*	3.337	0.001
	0	False	1	0.101	0.003*	-12.827*	3.624	0.002
			2	**0.615**	**0.005**	**65.440**	3.464	0.005
Scenario 2	-1	False	1	0.596	0.021*	426.621*	1.188	0.019
			2	**0.614**	0.022*	447.618*	1.190	0.021
		True	1	0.604	0.015*	305.197*	1.150	0.013
	0	False	1	0.552	**0.027**	**632.300**	1.040	0.023
			2	0.562	0.027	591.966	1.026	0.025
Scenario 3	-1	False	1	0.559	0.008*	187.197*	1.182	0.002
			2	**0.728**	**0.008**	**192.127**	1.182	0.002
		True	1	0.522	0.006*	151.686*	1.151	0.001
	0	False	1	0.025	0.002*	-35.765*	1.044	0.001
			2	0.161	0.003*	-13.100*	1.034	0.004
Scenario 4	-1	False	1	0.253	0.003*	23.063*	2.502	0.001
			2	**0.772**	0.006*	97.490*	2.404	0.004
		True	1	0.184	0.003*	8.669*	2.458	0.001
	0	False	1	0.588	0.008*	164.549*	2.500	0.004
			2	0.701	**0.014**	**297.047**	2.386	0.010
Scenario 5	-1	False	1	0.087	0.003*	-7.924*	2.880	0.001
			2	**0.623**	0.005*	**57.969**	2.776	0.003
		True	1	0.057	0.002*	-16.243*	2.783	0.001
	0	False	1	0.136	0.003*	-16.122*	2.914	0.003
			2	0.612	**0.005**	42.849	2.794	0.006

Takeaway: Reporting on bias is complicated.

Takeaway: Reporting on bias is complicated.

Bias evaluation

Estimation and interpretation of (popularity) bias is **highly dependent on the context** — impact of non configurable parameters.

Takeaway: Reporting on bias is complicated.

Bias evaluation

Estimation and interpretation of (popularity) bias is **highly dependent on the context** — impact of non configurable parameters.

Reporting

It is necessary to **document the research process** carefully, be explicit about **relevant factors** and consequent **experimentation,** and acknowledge the unique aspects and limitations of the reported results.

Current and future work

How can we extend this approach?

Current and future work

How can we extend this approach?

Extension

- Incorporate more state-of-the-art algorithms.
  - Formulate data scenarios based on UserKNN properties, as before — do the conclusions hold?
- Incorporate benchmarking (“real”) datasets.
- Examine other commonly used frameworks.

Current and future work

How can we extend this approach?

Extension

- Incorporate more state-of-the-art algorithms.
  - Formulate data scenarios based on UserKNN properties, as before — do the conclusions hold?
- Incorporate benchmarking ("real") datasets.
- Examine other commonly used frameworks.

Savvina Daniil, Manel Slokom, Mirjam Cuper, Cynthia Liem, Jacco van Ossenbruggen, Laura Hollink

CWI VU TUDelft KB

Cultural AI
a lab for
culturally
valued AI

Appendix

Additional slides

Generate data

Define data scenarios according to the observations mentioned.

Generate data

Define data scenarios according to the observations mentioned.

Data scenarios

1. **There is no relation between popularity and rating**: For each interaction, draw a rating value between 1 and 10 uniformly at random.
2. **Popular items are generally rated higher by the users**: For each interaction, draw a rating value between 1 and 10 from a normal distribution, where the mean is the popularity of the item normalized between 1 and 10.
3. **Popular items are generally rated lower by the users**: For each interaction, draw a rating value between 1 and 10 from a normal distribution, where the mean is the opposite of the popularity of the item normalized between 1 and 10.
4. **Only users with big profiles rate popular items higher**: For each interaction, draw a rating value between 1 and 10 uniformly at random. For the users with the 20% largest profiles, replace by drawing from a Poisson distribution where the mean is the popularity of the item normalized between 1 and 10.
5. **Only users with big profiles rate popular items lower**: For each interaction, draw a rating value between 1 and 10 uniformly at random. For the users with the 20% largest profiles, replace by drawing from a Poisson distribution where the mean is the opposite of the popularity of the item normalized between 1 and 10.