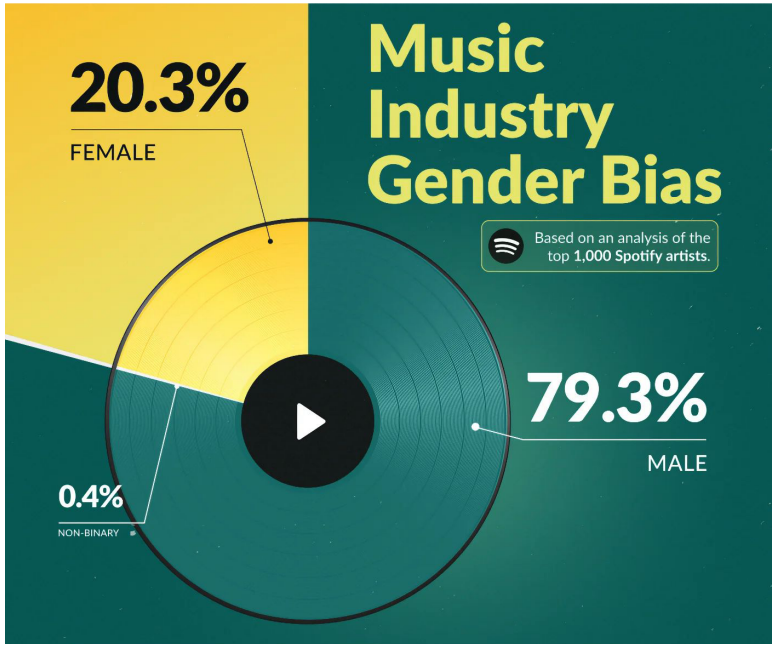
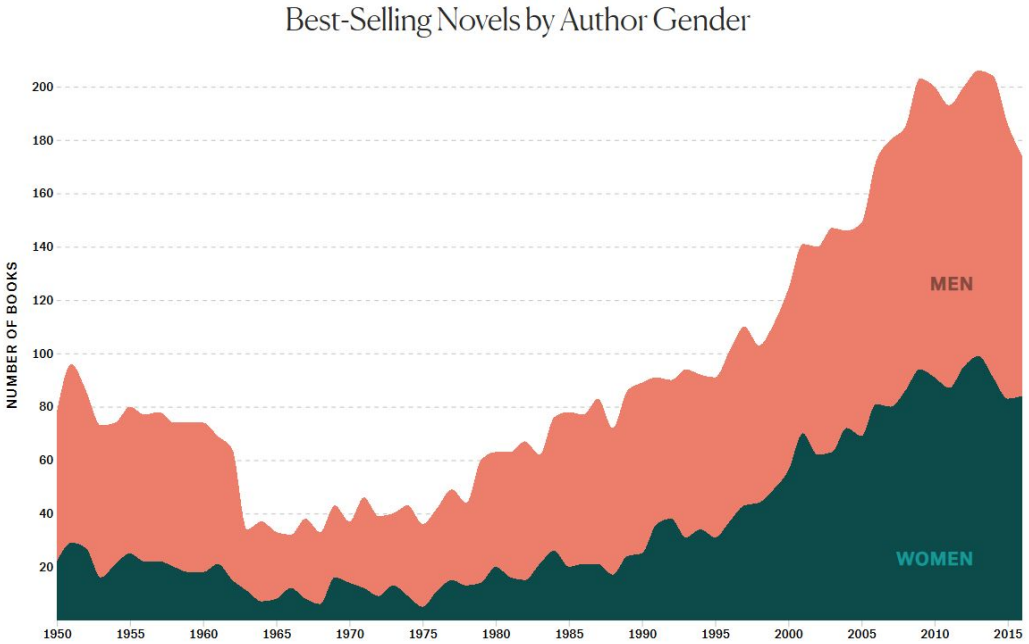
Reproducing Popularity and Gender Bias in Music Recommenders with Cross-Domain Extension to Books

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Team 2: Elif Deger
Nataliya Kharitonova

Background





Pérez Posada, S. (2025, March 13). *Exposing the music industry's gender bias*. Skoove. https://www.skoove.com/blog/music-gender-bias/
Rosie Cima, https://pudding.cool/2017/06/best-sellers/

Methodology

Replicate the methodology of Lesote et al.

- Evaluate 7 algorithms (e.g., RAND, POP, ALS, BPR, SLIM, VAE, Item-KNN) on the Last.FM (LFM-2b) music dataset.
- Analyze popularity and gender bias by comparing user listening histories with model recommendations.

Evaluate results

Apply same methodology to book domain

Bias mitigation

- Select 3 representative algorithms based on performance: Best, middle, and worst-performing
- Apply targeted bias mitigation strategies to these models
- Evaluate impact of mitigation

Research Question(s)

Main Research Question:

"To what extent are the findings on gender and popularity bias in music recommender systems by Lesota et al. (2021) reproducible, and how generalizable are these findings to domain of books?"

Sub-questions

1. How consistent are the original findings when reproduced with the same dataset, algorithms and metrics?
2. How do popularity and gender biases appear in book recommendations using the same methodology?
3. Are patterns of algorithm-induced bias domain-specific, or do they generalize across book domain?

Lesota, O., Melchiorre, A., Rekabsaz, N., Brandl, S., Kowald, D., Lex, E., & Schedl, M. (2021, September). Analyzing item popularity bias of music recommender systems: are different genders equally affected?. In *Proceedings of the 15th ACM conference on recommender systems* (pp. 601-606).

Datasets

Name:	Last.FM dataset	Book Crossing Dataset
What is it:	contains users' music listening history	contains users' book ratings and details
Variables:	user ID, artist, track, album, date, and time	user ratings of books along with book details (ISBN, title, author) and user information
Gender:	Includes additional datasets with user and artist gender info	extracted from the author's name (via gender-guesser (python))
Dataset:	166,153 entries	278,858 entries

Results - Last FM 2b

Table 1: All Results for LFM-2b

Alg.	Users	%ΔMean	%ΔMedian	%ΔVar.	%ΔSkew	%ΔKurtosis	KL	Kendall's τ	NDCG@10
RAND	All	-94.70	-94.34	-99.64	0.00	-92.42	3.56	0.18	0.0001
	ΔFemale	+0.88	+1.27	+0.04	-4.85	+7.60	+0.31	+0.02	—
	ΔMale	-0.37	-0.59	-0.02	+0.00	-2.13	-0.08	-0.01	—
POP	All	956.08	2321.62	310.19	-23.68	-97.02	5.68	0.62	0.0203
	ΔFemale	+138.43	+579.05	+57.31	-5.89	+4.56	+0.53	+0.00	—
	ΔMale	-74.30	-254.76	-63.26	+3.03	+0.94	-0.23	+0.04	—
ALS	All	+3.35	+79.87	-48.00	-28.96	-100.88	5.02	0.63	0.0204
	ΔFemale	-17.81	-6.35	-34.77	-11.27	-3.87	+0.52	-0.04	—
	ΔMale	+3.54	+0.88	+10.63	+5.20	+0.54	-0.11	+0.00	—
BPR	All	249.78	677.16	152.22	-45.42	-104.49	5.78	0.61	0.0117
	ΔFemale	+59.65	+172.71	+71.89	-3.07	+0.33	+0.59	-0.01	—
	ΔMale	-23.94	-71.35	-26.15	+1.28	-0.21	-0.19	+0.04	—
ItemKNN	All	223.97	389.08	159.65	-26.03	-99.16	5.19	0.58	0.1573
	ΔFemale	-19.98	-7.60	-51.50	-10.39	+1.14	+0.76	-0.05	—
	ΔMale	+9.48	+3.66	+14.67	+3.03	-0.58	-0.03	+0.02	—
SLIM	All	468.28	1157.50	378.16	-27.31	-97.03	5.57	0.61	0.0750
	ΔFemale	+53.39	+218.87	+54.50	-5.65	+0.46	+0.54	-0.01	—
	ΔMale	-22.72	-110.25	-38.73	+1.65	+0.00	-0.21	+0.04	—
VAE	All	-94.90	-94.44	-99.65	0.00	-92.44	3.72	0.18	0.3944
	ΔFemale	+0.69	+1.26	+0.03	-2.40	+5.18	+0.04	+0.01	—
	ΔMale	-0.34	-0.57	-0.00	+0.00	-2.17	-0.11	-0.01	—

Comparison to Paper's Results:

- Low popularity bias. Similar to paper (Lesota et al., 2021)
- Extremely biased. Similar to paper
- Reversal of gender effect; much less biased than paper
- Females more affected. More biased than the paper
- Reversal of gender effect; stronger popularity bias than paper
- Reversal of gender effect; stronger popularity bias than paper
- Paper has moderate bias. Ours has no strong popularity trend or gender bias

Comparison Summary: Our results vs. The Papers (Lesota et al.)

Conclusion:

- The original findings are **partially consistent**.
- While some trends (example: RAND, POP) could be reproduced, others (especially gender effects in ALS, SLIM) differ, indicating bias sensitivity to implementation details and data handling.

Results - Last FM 2b (Mitigation)

Table 2: Bias Mitigation for LFM-2b Dataset

Algorithm	Bias Reduction	NDCG@10 Before	NDCG@10 After	Overall Verdict
RAND	Minimal	0.0001	0.0000	Already fair; mitigation hurts
ItemKNN	High	0.1573	0.0035	Strong bias fix, but poor utility
VAE	Moderate	0.3944	0.1704	Best balance of fairness + quality

1. **VAE with Popularity-Weighted Loss**

Popular items were down-weighted, and less popular items up-weighted based on their frequency

2. **Item-KNN with Popularity-Penalized Similarity**

Scaled item similarity scores by the inverse log-frequency of item popularity => reduces the influence of highly popular items

3. **RAND with Inverse-Popularity Sampling**

Applied inverse-popularity sampling - items are selected with probabilities inversely proportional to their frequency

Question to class :)

Will Book Dataset follow the same popularity bias trend as LFM-2b, or surprise us?

Results - Book dataset

Table 3: All Results for Book Crossing Dataset

Alg.	Users	%ΔMean	%ΔMedian	%ΔVar.	%ΔSkew	%ΔKurtosis	KL	Kendall's τ	NDCG@10
RAND	All	-57.50	-66.67	0.00	0.00	-189.60	1.61	0.48	0.0001
	ΔFemale	+4.51	+0.00	0.00	0.00	-17.06	+0.00	+0.00	—
	ΔMale	-5.14	+0.00	0.00	0.00	+8.75	+0.35	+0.00	—
POP	All	1463.66	1303.85	0.00	0.00	-109.99	0.00	0.72	0.0072
	ΔFemale	-93.01	-52.68	0.00	0.00	-9.23	0.00	0.00	—
	ΔMale	+73.17	+65.38	0.00	0.00	+3.30	0.00	0.00	—
ALS	All	170.20	139.29	0.00	0.00	-95.45	0.00	1.00	0.0018
	ΔFemale	-21.49	-3.30	0.00	0.00	-1.40	0.00	0.00	—
	ΔMale	+17.08	+8.04	0.00	0.00	+2.00	0.00	0.00	—
BPR	All	128.07	65.82	0.00	0.00	-89.18	0.00	0.73	0.0047
	ΔFemale	-37.34	-32.57	0.00	0.00	+1.15	0.00	0.00	—
	ΔMale	+28.07	+19.18	0.00	0.00	-2.05	0.00	0.00	—
ItemKNN	All	-67.46	-66.67	0.00	0.00	-116.54	2.30	0.34	0.0112
	ΔFemale	+5.87	+0.00	0.00	0.00	-0.24	0.00	-0.01	—
	ΔMale	-0.80	+0.00	0.00	0.00	-0.66	0.00	+0.00	—
SLIM	All	380.31	472.73	0.00	0.00	-88.09	0.00	1.00	0.0104
	ΔFemale	-42.55	+6.06	0.00	0.00	-4.03	0.00	0.00	—
	ΔMale	+39.06	+24.73	0.00	0.00	+3.71	0.00	0.00	—
VAE	All	-63.33	-66.67	0.00	0.00	-147.67	1.39	0.48	0.0103
	ΔFemale	+2.54	+0.00	0.00	0.00	-12.89	0.00	0.00	—
	ΔMale	-3.33	+0.00	0.00	0.00	+5.58	0.00	0.00	—

Comparison to LFM-2b Results:

- No visible gender bias (similarly to LFM 2)
- Both highly biased, favours males (in contrast to LFM 2)
- Higher overall bias, favours males (similarly to LFM 2)
- Both highly biased, favours males (in contrast to LFM 2)
- Weaker gender bias, near- random effect
- Similar popularity bias, favours males (similar to LFM 2)
- Similar to LFM 2 (No strong popularity trend or gender bias)

Comparison Summary: LFM-2b & Book-Crossing

Conclusion:

- The original findings are only **partially generalizable**
- While some trends (example: POP bias amplification, VAE's poor alignment with user history) are consistent, key differences (especially in variance, skew, and gender effects) suggest that popularity bias is **domain-dependent**.
- The music domain shows stronger, more variable bias than books, **limiting cross-domain generalization.**

Question: Will Book Dataset follow the same popularity bias trend as LFM-2b, or surprise us?

Answer: Yes, it does surprise us - at least partially 😊

Results - Book Dataset (Mitigation)

Table 4: Bias Mitigation for Book Crossing Dataset

Algorithm	Bias Reduction	NDCG@10 Before	NDCG@10 After	Overall Verdict
RAND	Moderate	0.0001	0.0000	Already fair; mitigation reduces utility
ItemKNN	High	0.0112	0.0017	Bias fixed but recommendation quality collapses
VAE	Moderate	0.0103	0.0057	Best trade-off of fairness and utility

1. **VAE with Popularity-Weighted Loss**

Popular items were down-weighted, and less popular items up-weighted based on their frequency

2. **Item-KNN with Popularity-Penalized Similarity**

Scaled item similarity scores by the inverse log-frequency of item popularity => reduces the influence of highly popular items

3. **RAND with Inverse-Popularity Sampling**

Applied inverse-popularity sampling - items are selected with probabilities inversely proportional to their frequency

Answers to Research Question(s)

Research Focus	Answer / Conclusion
Sub-question 1: How consistent are the original findings when reproduced?	**Partially consistent** - trends like RAND and POP reproduce well, but gender effects in ALS and SLIM vary, showing sensitivity to implementation and data handling.
Sub-question 2: How do popularity and gender biases appear in books?	**Biases are weaker and more stable** - popularity bias exists but with less variance, gender bias is minor or negligible across most algorithms.
Sub-question 3: Are algorithmic bias patterns domain-specific?	**Yes** - biases are **domain-dependent**. Music shows stronger, more variable biases; findings do **not fully generalize** to books.
Main Research Question Reproducibility and generalizability of gender and popularity bias findings from music to books (Lesota et al., 2021)	Findings are **partially reproducible** and **only partially generalizable**, key patterns differ between music and book domains.

Limitations

Lack of Code Availability:

The authors of original paper did not provide access to the source code, and the methodological descriptions were abstract and non-specific, limiting exact reproducibility.

Resource Constraints:

Due to limited computational resources on local machines, it was not feasible to replicate all experimental steps described in the original study, especially those requiring large-scale training or tuning

Thank You!
Any Questions?

References

Borges, R., & Stefanidis, K. (2021). On mitigating popularity bias in recommendations via variational autoencoders. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing (SAC '21)* (pp. 1383–1389). Association for Computing Machinery. https://doi.org/10.1145/3412841.3442123

Cima, R. (2017, June). Bias, she wrote: The gender balance of The New York Times best seller list. *The Pudding*.Retrieved June 26, 2025, from https://pudding.cool/2017/06/best-sellers/

Ekstrand, M. D., Tian, M., Kazi, M. R. I., Mehrpouyan, H., & Kluver, D. (2018). Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)* (pp. 242–250). Association for Computing Machinery. https://doi.org/10.1145/3240323.3240373

Lesota, O., Melchiorre, A., Rekabsaz, N., Brandl, S., Kowald, D., Lex, E., & Schedl, M. (2021, September). Analyzing item popularity bias of music recommender systems: Are different genders equally affected? In *Proceedings of the 15th ACM conference on recommender systems* (pp. 601–606).

Liu, B., Chen, E., & Wang, B. (2023). Reducing popularity bias in recommender systems through AUC-optimal negative sampling. *arXiv preprint arXiv:2306.01348*. https://arxiv.org/abs/2306.01348

Melchiorre, A. B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., & Schedl, M. (2025). Investigating gender fairness of recommendation algorithms in the music domain [GitHub repository]. https://github.com/CPJKU/recommendation_systems_fairness

Pérez Posada, S. (2025, March 13). Exposing the music industry's gender bias. *Skoove*.https://www.skoove.com/blog/music-gender-bias/

Xv, G., Lin, C., Li, H., Su, J., Ye, W., & Chen, Y. (2022). Neutralizing popularity bias in recommendation models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)* (pp. 2623–2628). Association for Computing Machinery. https://doi.org/10.1145/3477495.3531907

Appendix 1

Table 5: Results from Lesota et al.

Alg.	Users	%ΔMean	%ΔMedian	%ΔVar.	%ΔSkew	%ΔKurtosis	KL	Kendall's τ	NDCG@10
RAND	*All*	−91.8	−87.2	−99.5	11.5	15.3	3.904	0.165	0.000
	Δ*Female*	−1.8	−3.5	−0.2	+0.0	−3.5	+0.976	−0.189	−0.000
	Δ*Male*	+0.5	+1.1	+0.1	−0.0	+1.3	−0.281	+0.053	+0.000
POP	*All*	432.5	975.2	487.0	−58.0	−87.0	6.023	0.057	0.045
	Δ*Female*	+11.0	+282.1	−172.2	−2.1	−1.9	+1.626	−0.033	+0.003
	Δ*Male*	−2.8	−115.8	+55.9	+0.5	+0.5	−0.380	+0.016	−0.001
ALS	*All*	121.8	316.6	72.6	−25.2	−43.9	4.368	0.046	0.184
	Δ*Female*	+9.9	+27.4	−7.1	−3.2	−5.4	+0.467	+0.110	−0.017
	Δ*Male*	−2.7	−6.6	+1.6	+0.8	+1.5	−0.121	−0.023	+0.005
BPR	*All*	−49.0	−3.7	−87.4	−14.8	−29.4	1.202	0.268	0.129
	Δ*Female*	+5.2	+7.7	+2.1	−1.4	−3.9	+0.476	−0.043	−0.011
	Δ*Male*	−1.1	−1.9	−0.6	+0.4	+1.1	−0.110	+0.010	+0.003
ItemKNN	*All*	9.6	4.6	5.7	−14.3	−29.0	0.175	0.423	0.301
	Δ*Female*	+2.0	+5.8	−2.6	−2.1	−3.2	+0.128	−0.037	−0.042
	Δ*Male*	−0.5	−1.3	+0.9	+0.8	+0.9	−0.020	+0.008	+0.012
SLIM	*All*	49.8	99.8	56.0	−12.5	−26.0	0.424	0.189	0.365
	Δ*Female*	−6.4	−13.1	−17.4	−1.7	−4.6	+0.217	+0.052	−0.048
	Δ*Male*	+1.9	+3.9	+5.6	+0.6	+1.1	−0.029	−0.012	+0.014
VAE	*All*	303.9	736.3	351.0	−45.2	−70.1	4.823	−0.028	0.191
	Δ*Female*	+10.1	+56.4	−69.3	−6.2	−6.6	+0.633	+0.146	−0.020
	Δ*Male*	−2.3	−20.4	+17.3	+1.8	+2.1	−0.161	−0.042	+0.006

Appendix 2
Glossary

- **Random Item (RAND):** A baseline algorithm that recommends for each user random items. It avoids recommending already consumed items.
- **Most Popular Items (POP):** A baseline that implements a heuristic-based algorithm that recommends the same set of overall most popular items to each user.
- **Item k-Nearest Neighbors (ItemKNN):** A neighborhood-based algorithm that recommends items based on item-to-item similarity. Specifically, an item is recommended to a user if the item is similar to the items previously selected by the user. ItemKNN uses statistical measures to compute the item-to-item similarities.
- **Sparse Linear Method (SLIM) :** Also a neighborhood-based algorithm, but instead of using predefined similarity metrics, the item-to-item similarity is learned directly from the data with a regression model.
- **Alternating Least Squares (ALS) :** A matrix factorization approach that learns user and item embeddings such that the dot product of these two approximates the original user-item interaction matrix.
- **Matrix factorization with Bayesian Personalized Ranking (BPR) :** Learns user and item embeddings, however, with an optimization function that aims to rank the items consumed by the users according to their preferences (hence, personalized ranking) instead of predicting the rating for a specific pair of user and item.
- **Variational Autoencoder (VAE) :** An autoencoder-based algorithm that, given the user's interaction vector, estimates a probability distribution over all the items using a variational autoencoder architecture.