UNIVERSITY OF MANNHEIM

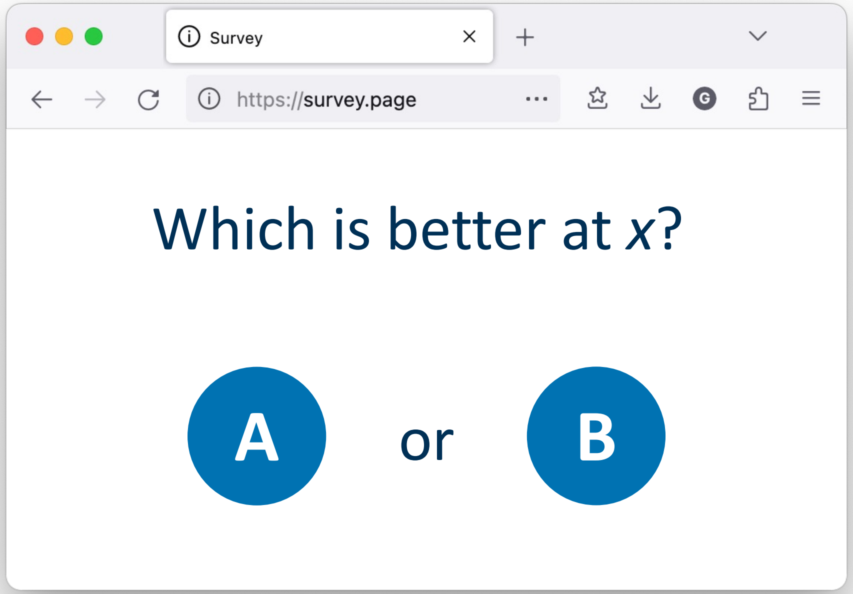
Fair Sampling for Global Ranking Recovery

Master‘s Thesis

Georg Ahnert – ahnert@uni-mannheim.de

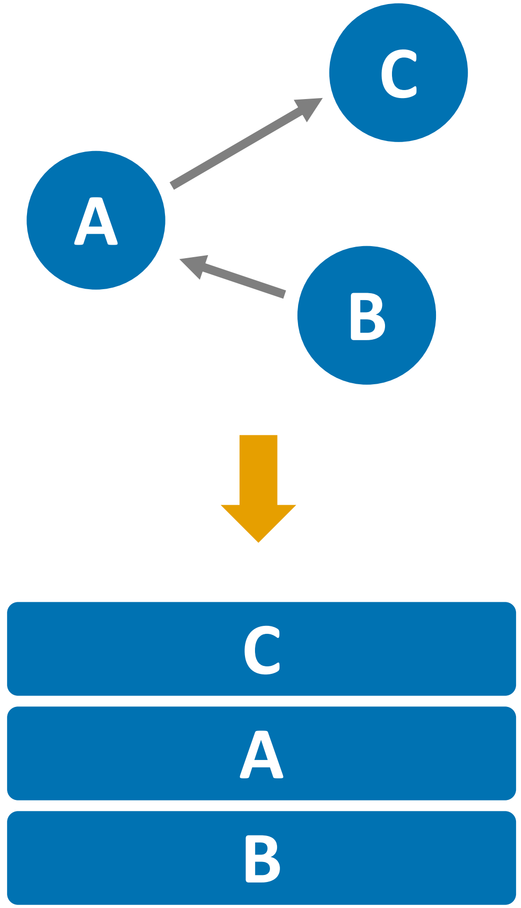
22.02.2024

Pairwise Comparisons



- More consistency (Kiritchenko & Mohammad, 2017)
- Less judgement error (Chen et al., 2013)

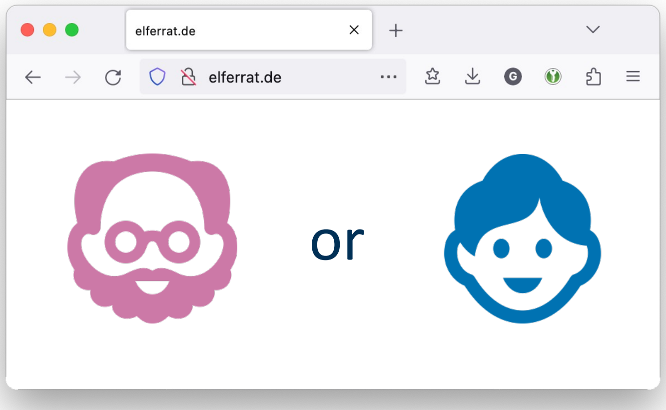
Aggregating Pairwise Comparisons



Applications

- Text readability (Crossley et al., 2023)
- Power of arguments (Loewen et al., 2012)
- Perceived ideology of US senators (Hopkins & Noel, 2022)
- Data extraction from Large Language Models (LLMs) (Wu et al., 2023)
- Human alignment of LLMs (Song et al., 2023)

Electing an *Elferrat*



Select the top 11 candidates

Goals

- Equal representation
- Equal accuracy

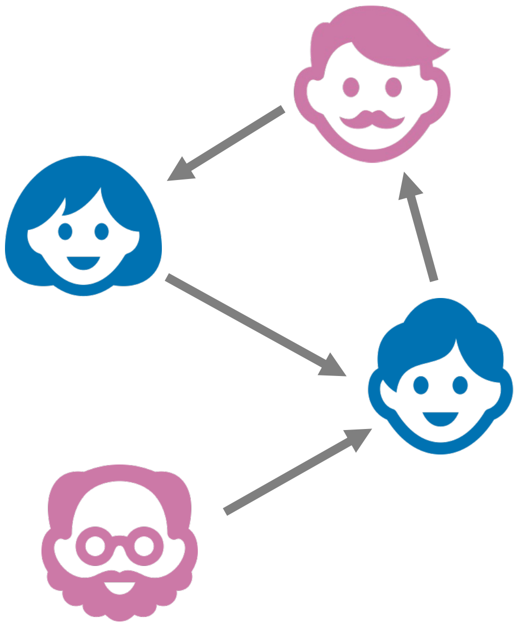
There could be fewer female candidates

- Historical bias
- Self-selection bias

Pairwise comparisons might be biased

- Systemic discrimination

Ranking Recovery

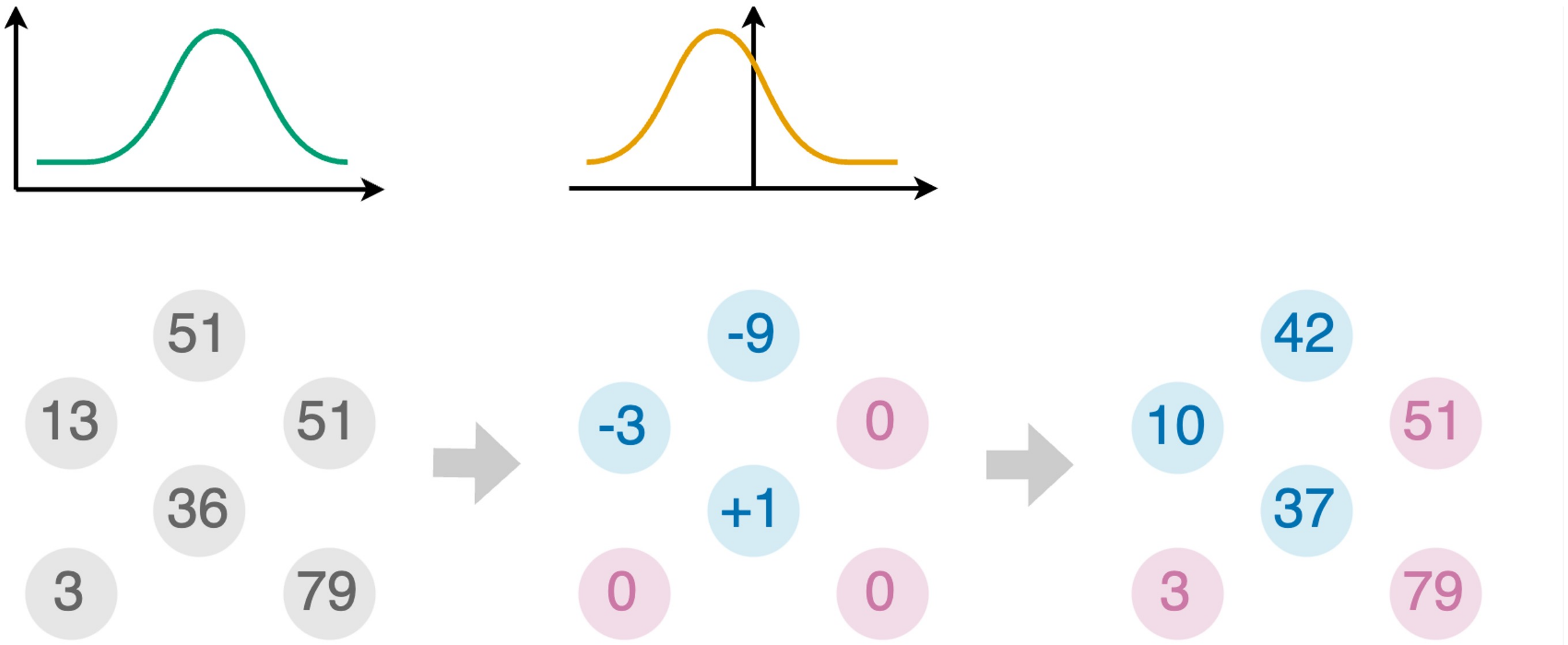


Pairwise comparisons generally are **incomplete** and **inconsistent**

- David's Score (David, 1987)
- RankCentrality (Negahban et al., 2012)
- GNNRank (He et al., 2022)

Research gap: **Fairness-aware ranking recovery from pairwise comparisons**

Research Setup – Normative Assumptions

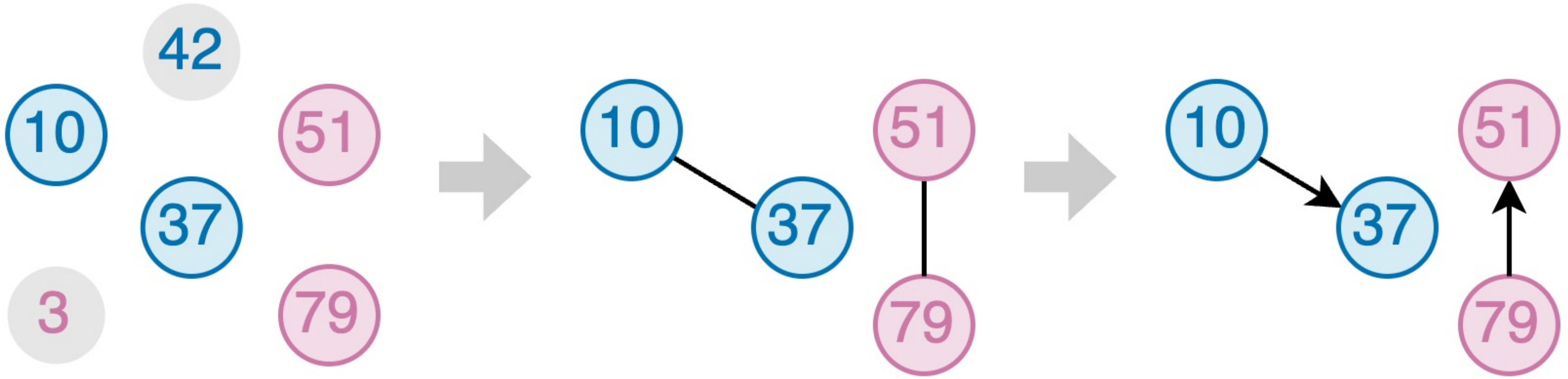


Ground-Truth Skill Score
Assuming a *we are all equal* worldview, skills are independent of group membership

Bias
We consider two groups and assume bias present against the *unprivileged* group

Average Perceived Score
…is the sum of skill score and the average bias present against this individual

Research Setup – Sampling & Comparison



1. Sampling of Individuals
Performed randomly, by group membership, or by previous success

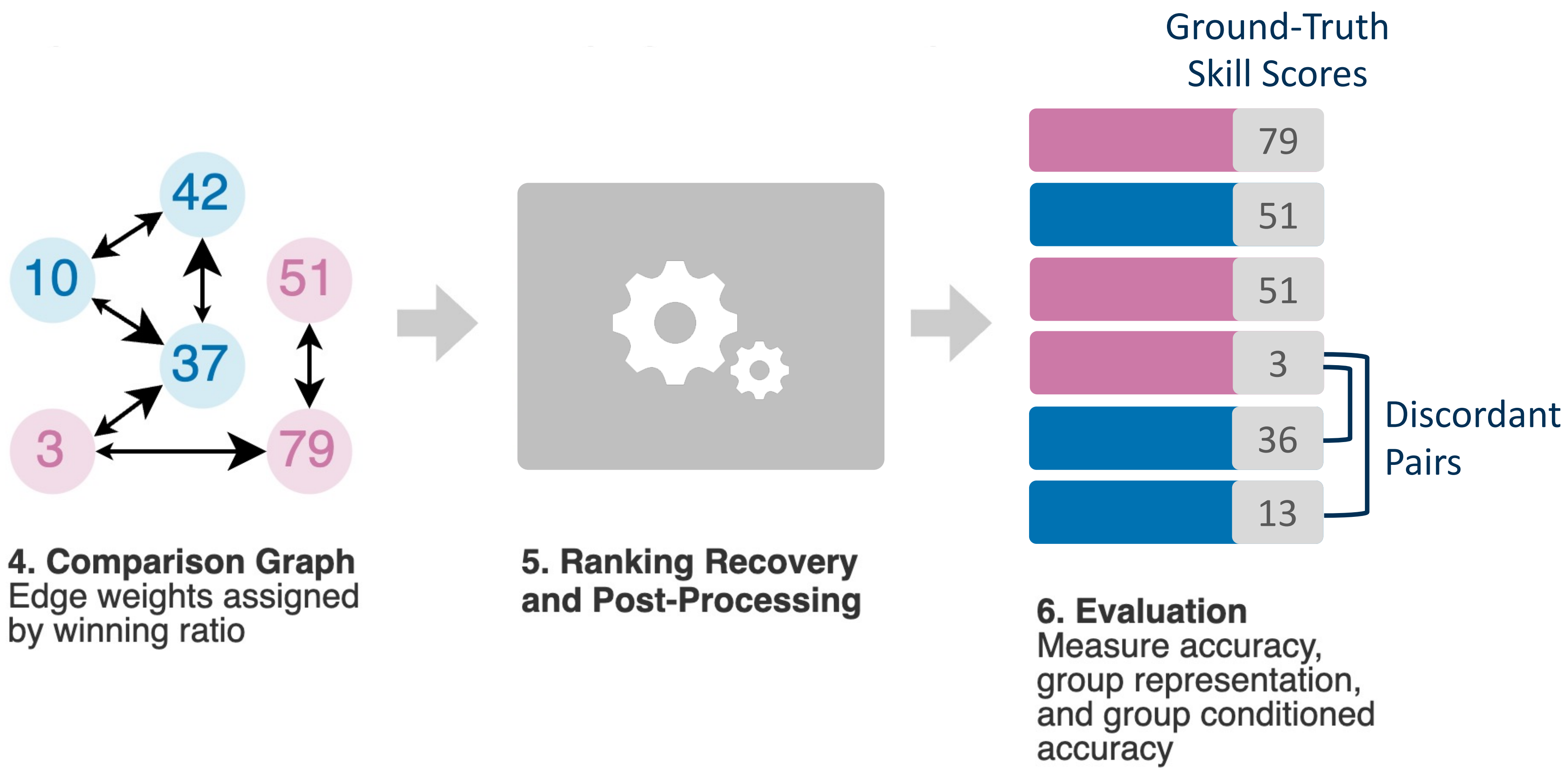
2. Pairing of Individuals

3. Comparison
Performed by human annotators or with the BTL-model

Repeat for *n* iterations – allow multiple probabilistic comparisons of the same pairs

Research Setup – Ranking Recovery

UNIVERSITY OF MANNHEIM



Measuring Accuracy & Fairness

Desiderata

- Measured against ground-truth
- Higher penalties for gross differences
- Consider sub-groups

Group-Conditioned Weighted Kemeny Distance

$$D_G := \sqrt{\frac{\sum_{G\text{'s discordant pairs}} (\text{score difference})^2}{\sum_{\text{all pairs that involve } G} (\text{score difference})^2}}$$

Group Representation measured as **Exposure** (Singh & Joachims, 2018)

Datasets

Desiderata

- Ground-truth values & pairwise comparisons
- Incomplete & probabilistic comparisons
- 2 groups, existence of bias

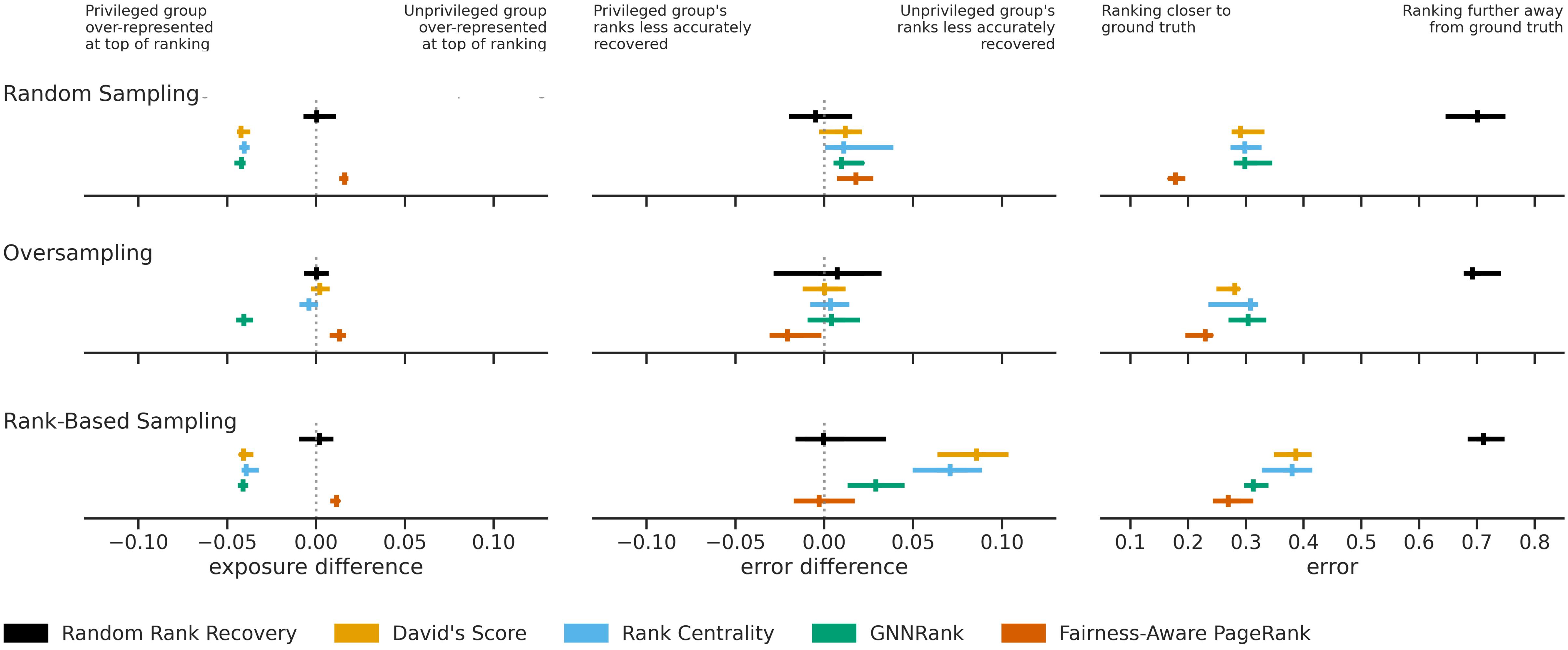
Synthetic Data

- 200+200 individuals with normally distributed skills & bias
- Compared using the Bradley-Terry-Luce model (Bradley & Terry, 1952)

Empirical Data

- IMDB-WIKI-SbS dataset: 9,150 images in 250,249 pairs (Pavlichenko & Ustalov, 2021)
- Pre-processed using image captions crawled from IMDB.com & FairFace (Karkkainen & Joo, 2021)

Results



Privileged group over-represented at top of ranking
Unprivileged group over-represented at top of ranking
Privileged group's ranks less accurately recovered
Unprivileged group's ranks less accurately recovered
Ranking closer to ground truth
Ranking further away from ground truth
Random Sampling
Oversampling
Rank-Based Sampling
−0.10 −0.05 0.00 0.05 0.10
exposure difference
−0.10 −0.05 0.00 0.05 0.10
error difference
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
error
Random Rank Recovery
David's Score
Rank Centrality
GNNRank
Fairness-Aware PageRank

Main Take-Aways

UNIVERSITY OF MANNHEIM

- Under random sampling, **GNNRank offers little benefit** over David's Score
- **Oversampling is unreliable** for bias mitigation
- Fairness-Aware ranking recovery both **improves accuracy & decreases bias**
- **FairPageRank** or **GNNRank + FA*IR** are viable options (with drawbacks)

» Potential for dedicated fairness-aware ranking recovery algorithms

Fair Sampling for Global Ranking Recovery

Master‘s Thesis — Georg Ahnert

UNIVERSITY OF MANNHEIM

Contributions

- Introduced fairness-aware ranking recovery from pairwise comparisons
- Proposed research setup & group-conditioned accuracy measure
- Investigated representative ranking recovery & post-processing methods

Python package under MIT license: github.com/wanLo/fairpair

Contact me: ahnert@uni-mannheim.de
georgahnert.de

References

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, 39(3/4), 324-345.

Chen, X., Bennett, P. N., Collins-Thompson, K., & Horvitz, E. (2013, February). Pairwise ranking aggregation in a crowdsourced setting. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 193-202).

Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. Behavior Research Methods, 55(2), 491-507.

David, H. A. (1987). Ranking from unbalanced paired-comparison data. Biometrika, 74(2), 432-436.

He, Y., Gan, Q., Wipf, D., Reinert, G. D., Yan, J., & Cucuringu, M. (2022, June). GNNRank: Learning global rankings from pairwise comparisons via directed graph neural networks. In International Conference on Machine Learning (pp. 8581-8612). PMLR.

Hopkins, D. J., & Noel, H. (2022). Trump and the shifting meaning of “conservative”: Using activists’ pairwise comparisons to measure politicians’ perceived ideologies. American Political Science Review, 116(3), 1133-1140.

Karkkainen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1548-1558).

Kiritchenko, S., & Mohammad, S. M. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. arXiv preprint arXiv:1712.01765.

Loewen, P. J., Rubenson, D., & Spirling, A. (2012). Testing the power of arguments in referendums: A Bradley–Terry approach. Electoral Studies, 31(1), 212-221.

Negahban, S., Oh, S., & Shah, D. (2012). Iterative ranking from pair-wise comparisons. Advances in neural information processing systems, 25.

Pavlichenko, N., & Ustalov, D. (2021). IMDB-WIKI-SbS: An evaluation dataset for crowdsourced pairwise comparisons. arXiv preprint arXiv:2110.14990.

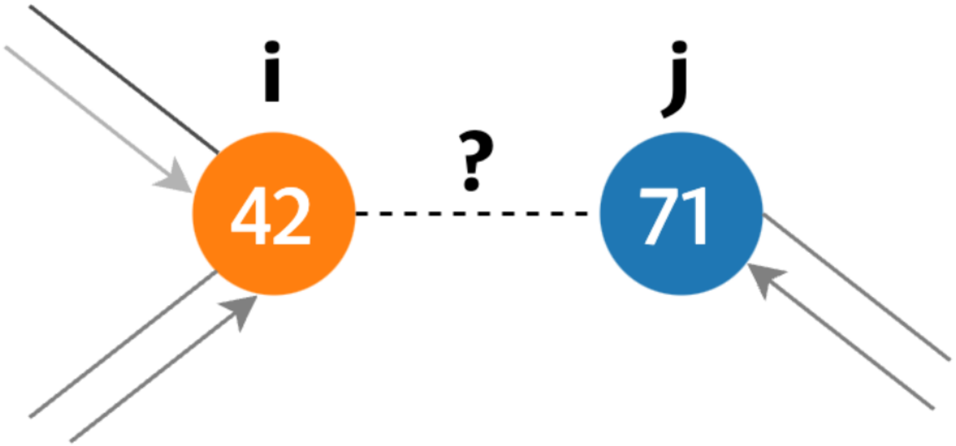
Singh, A., & Joachims, T. (2018, July). Fairness of exposure in rankings. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2219-2228).

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., and Wang, H. Preference ranking optimization for human alignment. arXiv preprint arXiv:2306.17492, 2023.

Wu, P. Y., Nagler, J., Tucker, J. A., & Messing, S. (2023). Large language models can be used to estimate the latent positions of politicians. arXiv preprint arXiv, 2303.

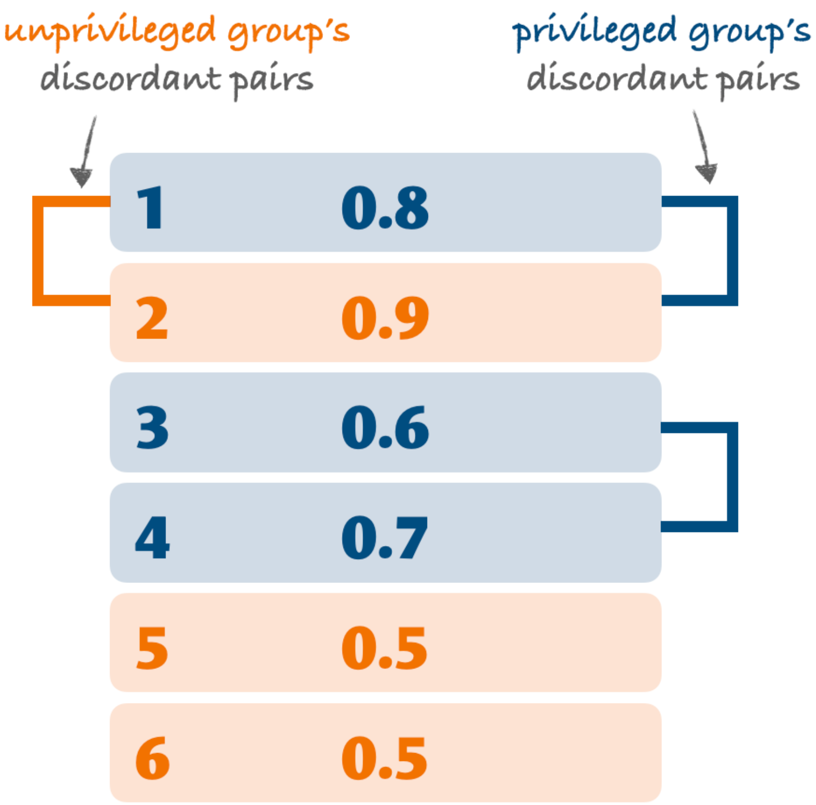
Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017, November). FA*IR: A fair top-k ranking algorithm. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 1569-1578).

Extra:
The Bradley-Terry-Luce Model



$$P\left(i \text{ beats } j\right) := \frac{e^{s_i}}{e^{s_i} + e^{s_j}}$$

Extra:
Group-Conditioned Weighted Kemeny Distance



$$D_G := \sqrt{\frac{\sum_{G\text{'s discordant pairs}} (\text{score difference})^2}{\sum_{\text{all pairs that involve } G} (\text{score difference})^2}}$$

$$D_{\text{unpriv}} = \sqrt{\frac{0.01}{0.74}} \approx 0.12 \qquad D_{\text{priv}} = \sqrt{\frac{0.02}{0.46}} \approx 0.21$$

$$D_{\text{diff}} \approx 0.12 - 0.21 = -0.09$$

Extra:
The Exposure Measure (Group-Representation)

1
2
3
4
5
6

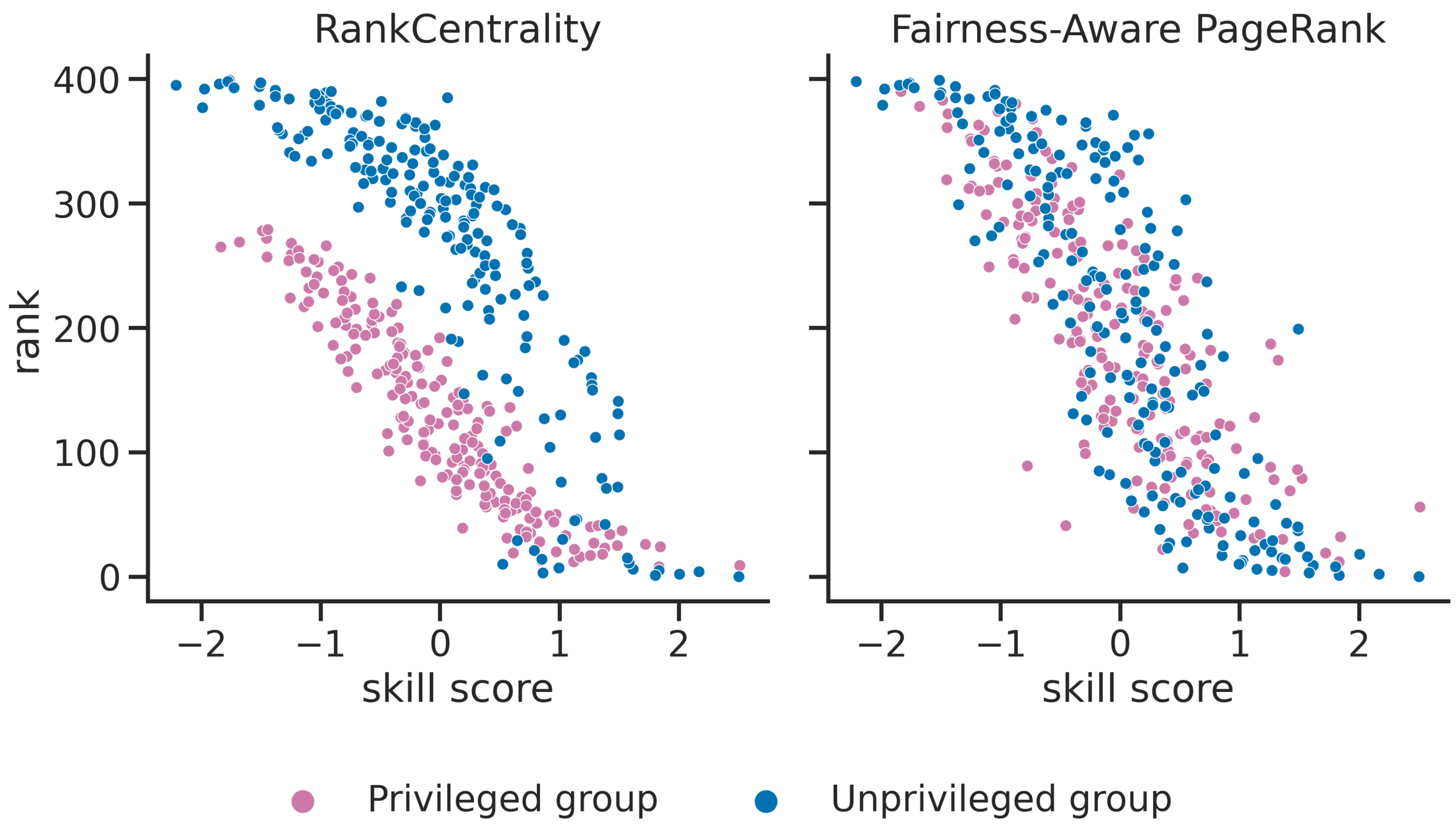
$$\mathrm{Exp}_G := \frac{1}{|G|} \sum_{\text{individuals in } G} \frac{1}{\log_2(\mathrm{rank}+1)}$$

$\mathrm{Exp}_{\mathrm{unpriv}} \approx 0.46$

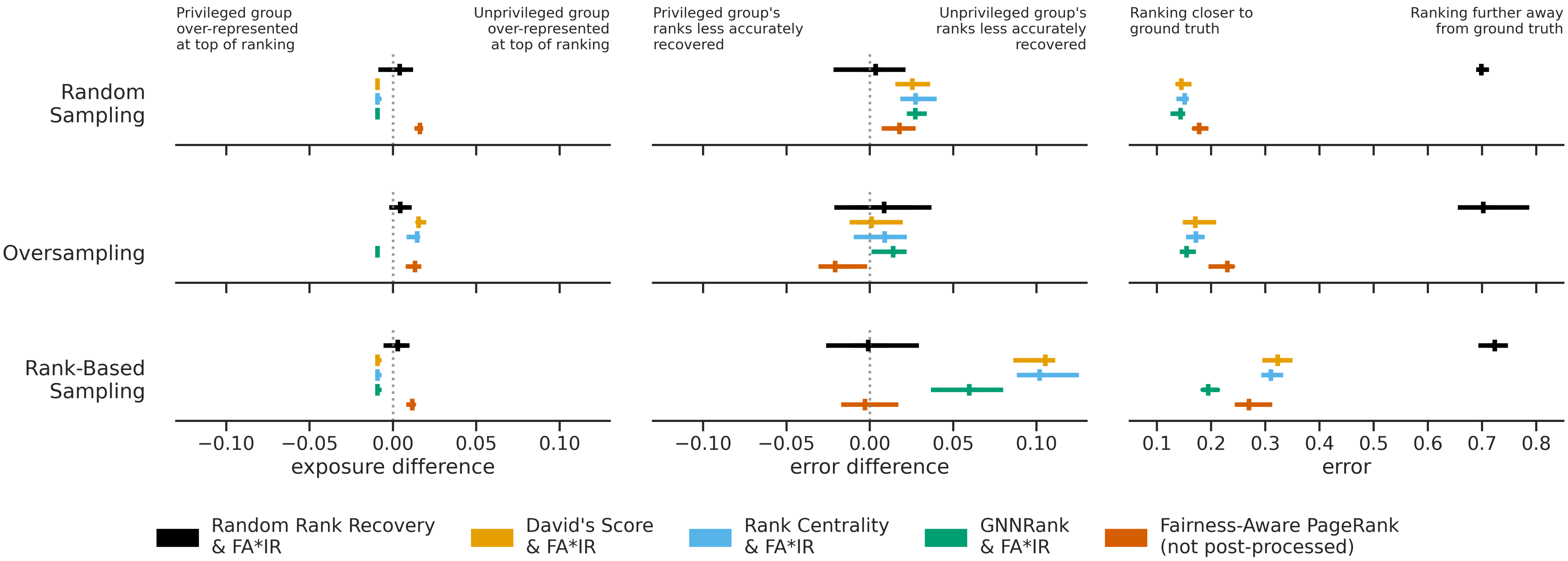
$\mathrm{Exp}_{\mathrm{priv}} \approx 0.64$

$\mathrm{Exp}_{\mathrm{diff}} \approx 0.18$

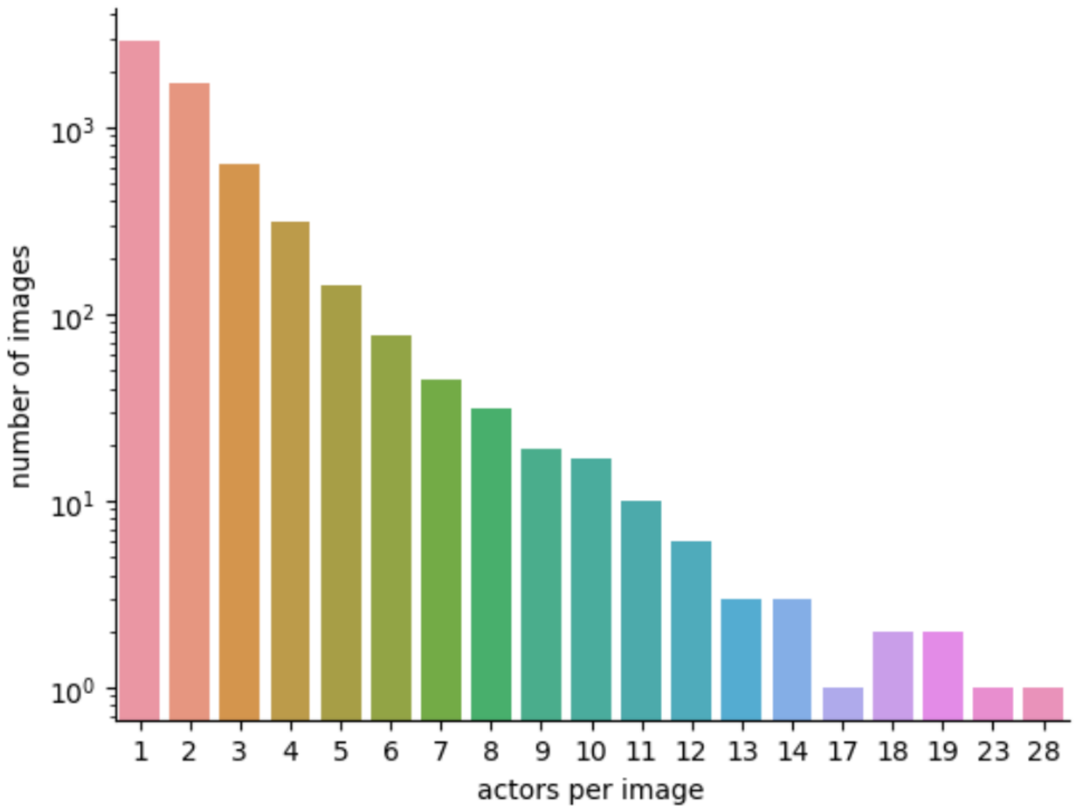
Extra:
The Oversampling Anomaly



Extra: Post-Processing Results



Extra:
Issues with the IMDB-WIKI-SbS dataset



Extra:
Pre-Processing the IMDB-WIKI-SbS dataset

filename:
nm0001627_rm4164078592_1927-2-20_1997.jpg

Old annotation:
Sidney Poitier, male, 70

IMDB url:
https://www.imdb.com/name/nm0001627/mediaviewer/rm4164078592

Actors in image caption on IMDB.com:
(Richard Gere, nm0000152),
(Sidney Poitier, nm0001627)

Additional data from Wikidata:
(nm0000152, male, born 1949),
(nm0001627, male, born 1927)

Age at time of image:
48, 70

Detected by FairFace (Karkkainen & Joo, 2021):
male, white, 50–59

Exact match on gender & closest match on age & age in FairFace range +-5 years

New annotation:
Richard Gere, male, 48

Extra:
Empirical Results