

The Demographics of Cool

Popularity and Recommender Performance for Different Groups of Users

GOAL

- Explore the role of demographics in recommender evaluation and decision-making

BACKGROUND

- Largest demographic groups dominates overall statistics
- Resulting decisions optimize performance for dominant group
- Per-group evaluation can yield better insights into recommender behavior [2]

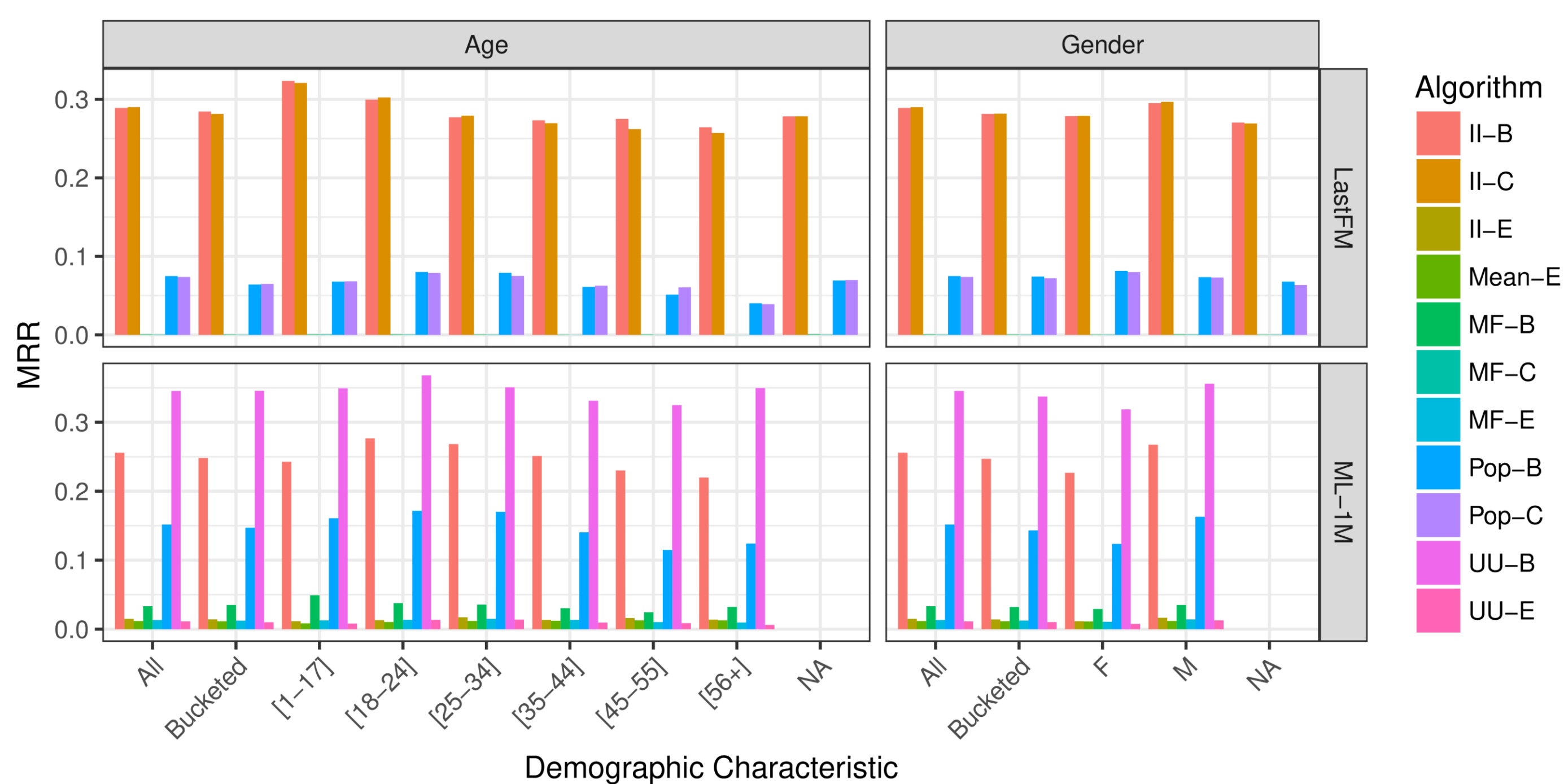
RESEARCH QUESTIONS

- What changes about our assessment of relative or absolute recommender effectiveness when we consider performance for different subgroups of users?
- Does popularity bias exacerbate demographic bias effect?
- How do popularity bias mitigations affect the demographic bias?

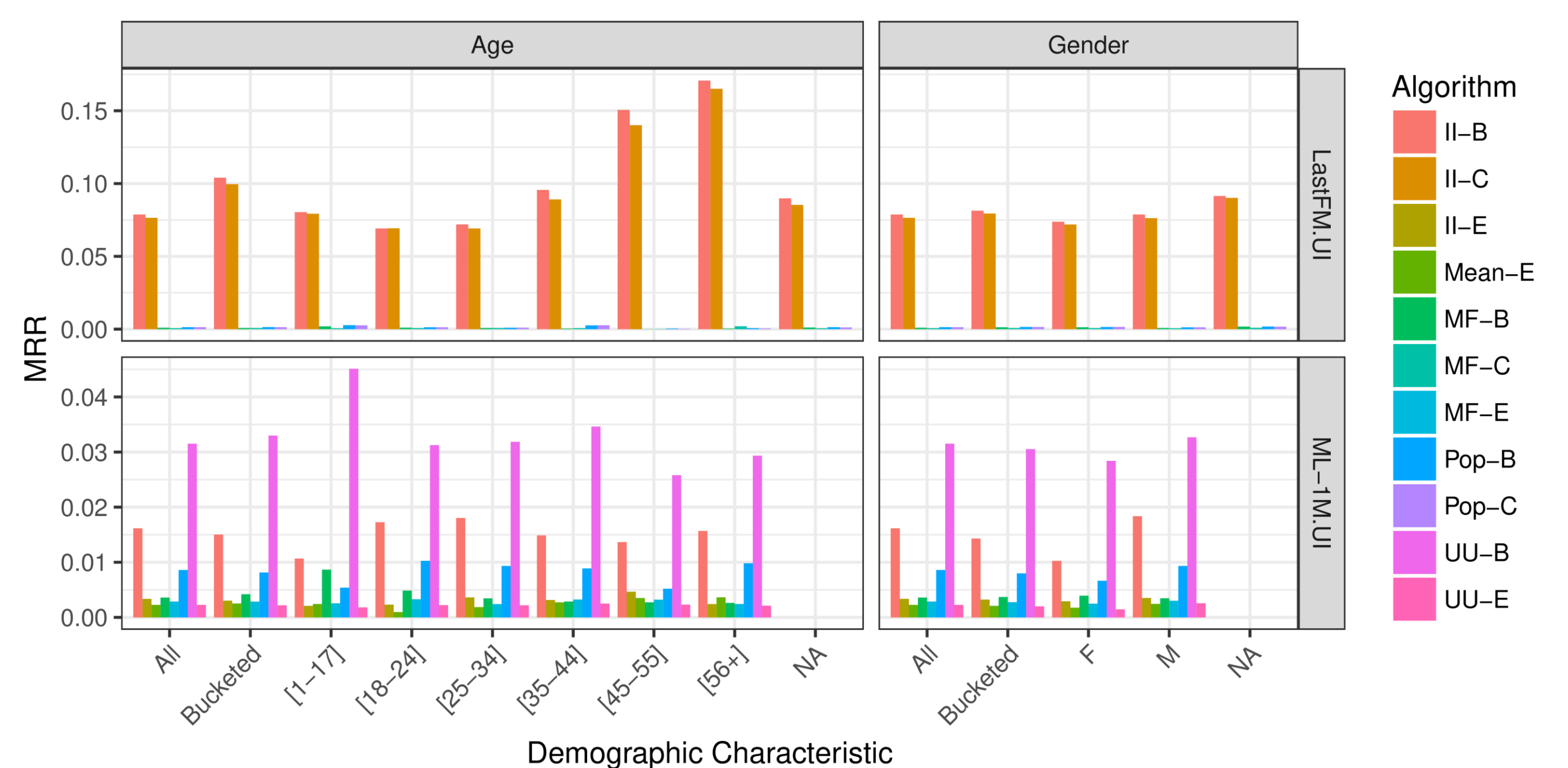
DATA & METHODS

- Datasets:** MovieLens-1M (last with demographics) & LastFM 360K
- Metric:** Mean Reciprocal Rank (MRR)
- Assessment:** 5-fold cross-validation on (1) LensKit's default strategy and (2) Bellogin's UAR strategy [1]
- Algorithms:** **Popular**, recommending the most frequently rated or played items; **Item-Item**, an item-based collaborative filter; **User-User**, a user-based collaborative filter; and **FunkSVD**, based on gradient descent matrix factorization technique
- Variations:** '-E' for explicit-feedback recommenders (MovieLens); '-B' for binary implicit-feedback recommenders (item was rated or played); and '-C' for implicit-feedback recommenders that consider the number of times an artist was played (LastFM)

FINDINGS



Results of basic runs of experiments



Results of basic runs of UAR experiments

- We find that recommenders are not equally good for all users in predictable and socially-relevant ways:
 - A number of strategies achieve moderately higher accuracy metric values for dominant demographic groups, causing an algorithm's performance to increase without delivering benefit to smaller user subgroups
 - Demographic bias has a complex interaction with mitigation strategies for other offline evaluation ailments
 - A uniform item strategy results in disproportionately higher accuracy values for users in some smaller subgroups
- Assigning equal weight to user groups can change configuration decisions
- There is a need for careful and multi-faceted consideration of recommender system behavior across a range of both users and items.

REFERENCES

- [1] A. Bellogin. Performance prediction and evaluation in Recommender Systems: an Information Retrieval perspective. PhD thesis, UAM, 2012.
- [2] R. Mehrotra, A. Anderson, F. Diaz, A. Sharma, H. Wallach, and E. Yilmaz. Auditing search engines for differential satisfaction across demographics. In Proc. WWW '17 Companion, 2017.

