Detecting and Mitigating Bias in Targeted Advertising

Cheryl Brooks

Data Science & AI Research

Joint work with Emily Dodwell, Balachander Krishnamurthy and Ritwik Mitra

Data Science & AI Research

Cautionary Examples – Unintentional Discrimination Through Machine Learning

Carnegie Mellon Study Finds Gender Discrimination In Ads Shown On Google

Study authors point out, however, cause of alleged discrimination is unclear.

Greg Sterling on July 8, 2015 at 11:33 am

Source: marketingland.com

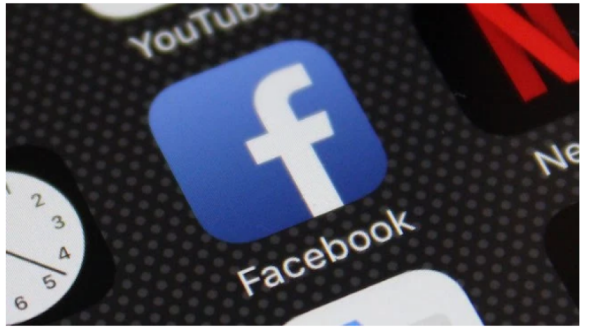
Facebook's ad-targeting problems prove how easy it is to discriminate online

Online personalization opens up significant possibilities for discrimination against vulnerable communities.

Source: nbcnews.com

Facebook is removing over 5,000 ad targeting options to prevent discriminatory ads

Sarah Perez @sarahintampa / 5 months ago Comment



Source: techcrunch.com

Our work is a proactive effort to detect and mitigate problems in this domain

Our assumptions

- We define a targeted ad as one where the advertiser selects a set of users to see the ad
- We are interested in assessing the targeted list for bias, and only consider the content of the ad to classify the ad category
- The majority of ads do not need to be checked for bias
- A small set of ad categories exist where the company does not want to appear biased towards certain demographics – we refer to these as *sensitive ad categories*
- A set of sensitive demographic attributes is known for the ad category

Bias Risk for Targeted Advertising Roles

Advertiser

- Internal models could create a biased targeted list
- Proposed Solution: Fair ML models

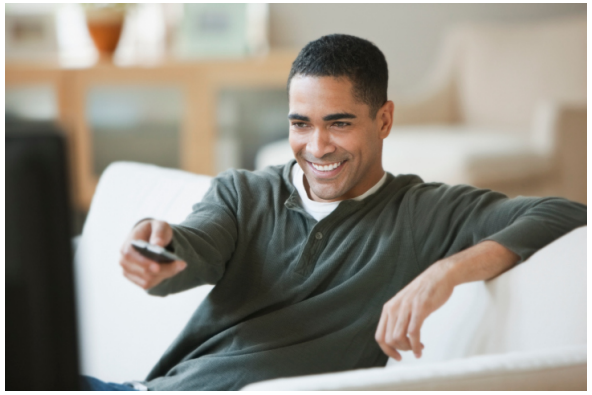
Ad Platform

1. Advertiser constructs a targeted list from features provided by the Ad Platform
   - Provided features or some combination of them could correlate with a sensitive attribute
   - Proposed Solution: Algorithmically remove sensitive information from features

2. Advertiser provides targeted list to Ad Platform
   - Provided list is biased against a demographic group
   - Proposed Solution: **Open research problem -> focus of work**

Motivating Example –Targeted Advertising on TV or Streaming Services





Consider the following scenario:

- Advertiser provides a list of households they would like to reach
- List matched to customers with TV or Streaming Service
- Final targeted list set to receive the ad

If the ad category is sensitive, how can the ad platform verify that the targeted list is not biased?

Research Goals

Develop methods using ML techniques for an Ad Platform to detect and mitigate bias in targeted advertising lists

Techniques should apply regardless of advertiser intent

- Unintentionally biased advertiser
- Malicious advertiser

Techniques should be robust to different attack scenarios

- Malicious advertiser with a simple strategy – e.g. drops targets with sensitive attribute (*focus for today's talk)
- Malicious advertiser with a more sophisticated strategy – e.g. selects non-relevant targets with sensitive attribute

Measuring Discrimination – Disparate Impact

Preliminary definitions:

- Targeted Audience (TA) – set of users that the advertiser selects to see an ad
- Relevant Audience (RA) – set of users for whom the expected benefit of viewing an ad is equivalent
- Disparate Impact – over- or under-representation of targets with a sensitive attribute measured by disparity

$$\text{Representation ratio (RR)} = \frac{|TA \cap RA_s|/|RA_s|}{|TA \cap RA_{\neg s}|/|RA_{\neg s}|}$$

$$\text{where } RA_s = \{RA \text{ with sensitive attribute}\} \text{ and } RA_{\neg s} = \{RA \text{ without sensitive attribute}\}$$

$$\text{Disparity} = \max\left(\text{RR}, \frac{1}{\text{RR}}\right)$$

Disparity > β suggests over- or under-representation, where β is a threshold set by the ad platform

- E.g. setting β = 1.25 would align with the 80% disparate impact rule (EEOC - Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.4(D), 2015)

Challenge: We only observe the targeted audience, need to infer the relevant audience

Data Set-Up

Treat targeted set as a set of known positive examples and all other examples as unlabeled

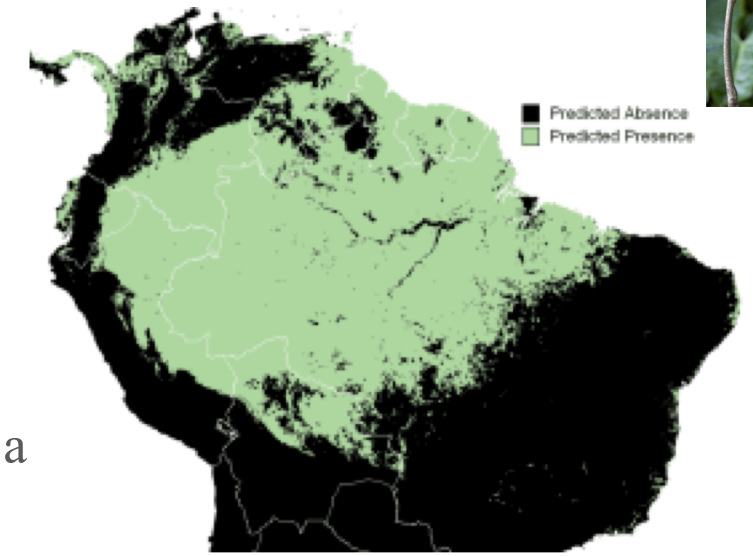
$$
\begin{array}{c|c|c}
y_1 = 1 & t_1 = 1 & x_{11}, x_{12}, \ldots, x_{1p} \\
y_2 = 0 & t_2 = ? & x_{21}, x_{22}, \ldots, x_{2p} \\
y_3 = 1 & t_3 = ? & x_{31}, x_{32}, \ldots, x_{3p} \\
y_4 = 1 & t_4 = 1 & x_{41}, x_{42}, \ldots, x_{4p} \\
y_5 = 0 & t_5 = ? & x_{51}, x_{52}, \ldots, x_{5p}
\end{array}
$$

Unobserved Relevant Audience	Observed Targeted Set	Features available to Ad Platform

Two Lines of Related Work

1 **Ecology/Statistics literature on presence-only data**

- "Presence-only data and the EM algorithm", Ward et al., Biometrics, 2009
- Application to species modeling
- Assumes we sample a set of positive examples and then sample a separate set of unlabeled examples from the population
- Methods require estimate of population prevalence



Species distribution modeling of Brown-Throated Sloth
https://clarklabs.org/species-distribution-modeling-in-terrsets-land-change-modeler/

2 **Machine Learning literature on positive unlabeled data (PU-Learning)**

- "Learning classifiers from only positive and unlabeled Data", Elkan and Noto, KDD, 2008
- Application to document classification
- Assumes one sample from entire population and observed positive labels are selected completely at random
- Key Result:

$$P(y \mid x) = \frac{P(t \mid x)}{k}, \text{ where } k \text{ is a constant } \rightarrow \text{ rank is preserved}$$

Detecting and Mitigating Discrimination

Our contributions

- We show that detecting bias in targeted advertising from the perspective of an ad platform can be framed as a PU-Learning problem
- Establish similar result to Elkan and Noto (2008), but allow for the targeted set to be biased with respect to a sensitive feature

Proposed Approach

- Consider targeted audience as labeled data and remainder of customer base as unlabeled data
- Use Positive-Unlabeled (PU) ML algorithms to construct an *inferred relevant audience (IRA)*
  - Use ML model to classify unlabeled examples, but treat targeted examples as known positives
- Compute disparity using IRA and mitigate by expanding the targeted audience when bias is detected

Empirical Study

UCI ML Repository Adult Data Set

- 45k examples collected from 1994 census database

Simulated Example – Employment Ad

- Relevant audience: All examples with a college or graduate degree
- Sensitive feature: Gender
- Targeted list construction
  - τ_s = proportion of females in the relevant audience
  - Randomly sample n_t examples from the relevant audience, where females are sampled with probability $\tau_s(1-\alpha)$
  - Vary n_t = (7500, 5000, 2500, 1000) and α = (0, 0.2, 0.4, 0.6, 0.8)

Evaluation – Detection and Mitigation

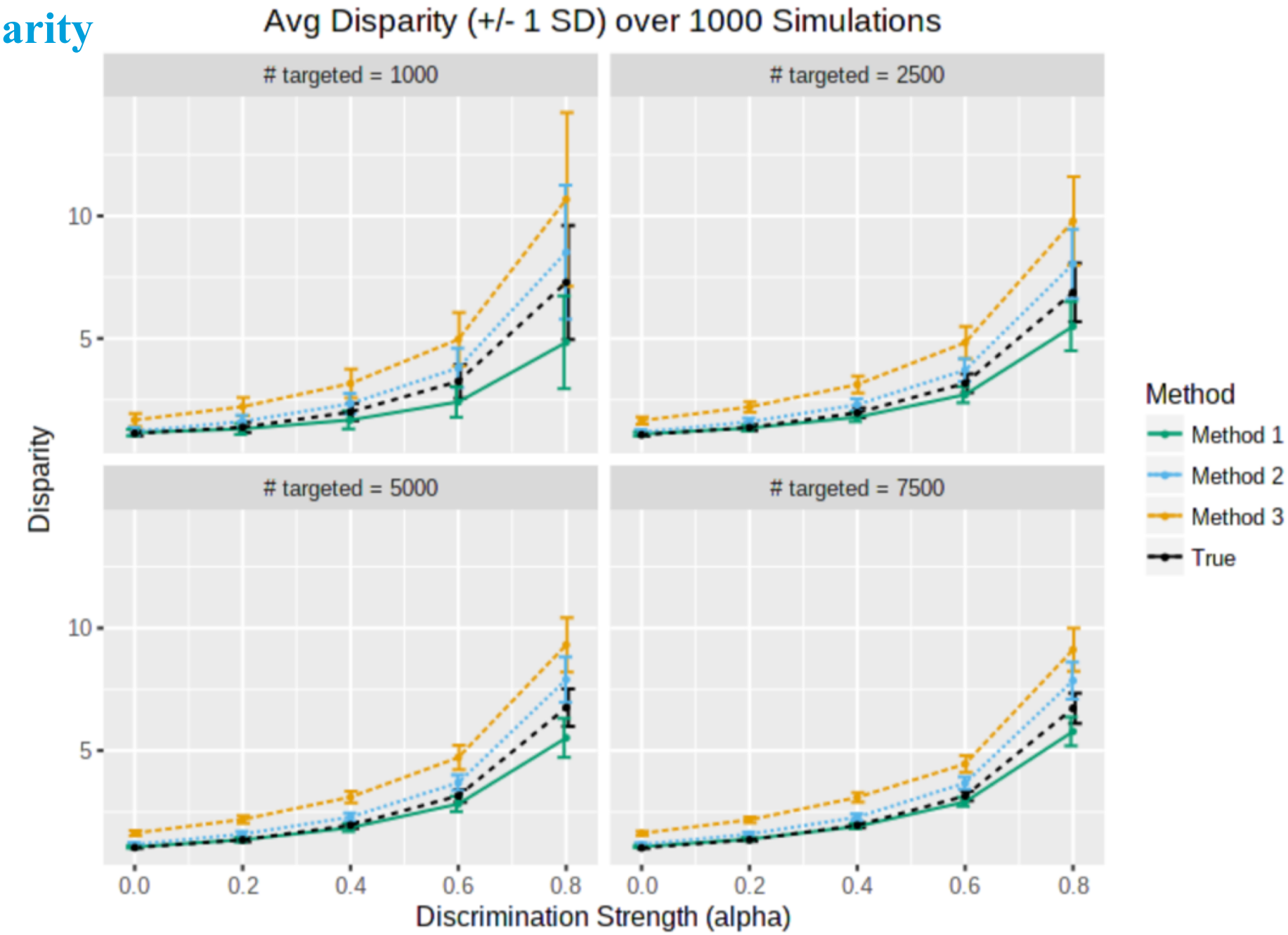
- Run 1000 simulations for each parameter specification using R ‘foreach’ package
- Train PU Logistic Regression model

- Compare performance of 3 methods for choosing the classification threshold
  1. F1-measure
  2. Minimum predicted probability for targeted examples
  3. 5^{th} percentile of predicted probability for targeted examples

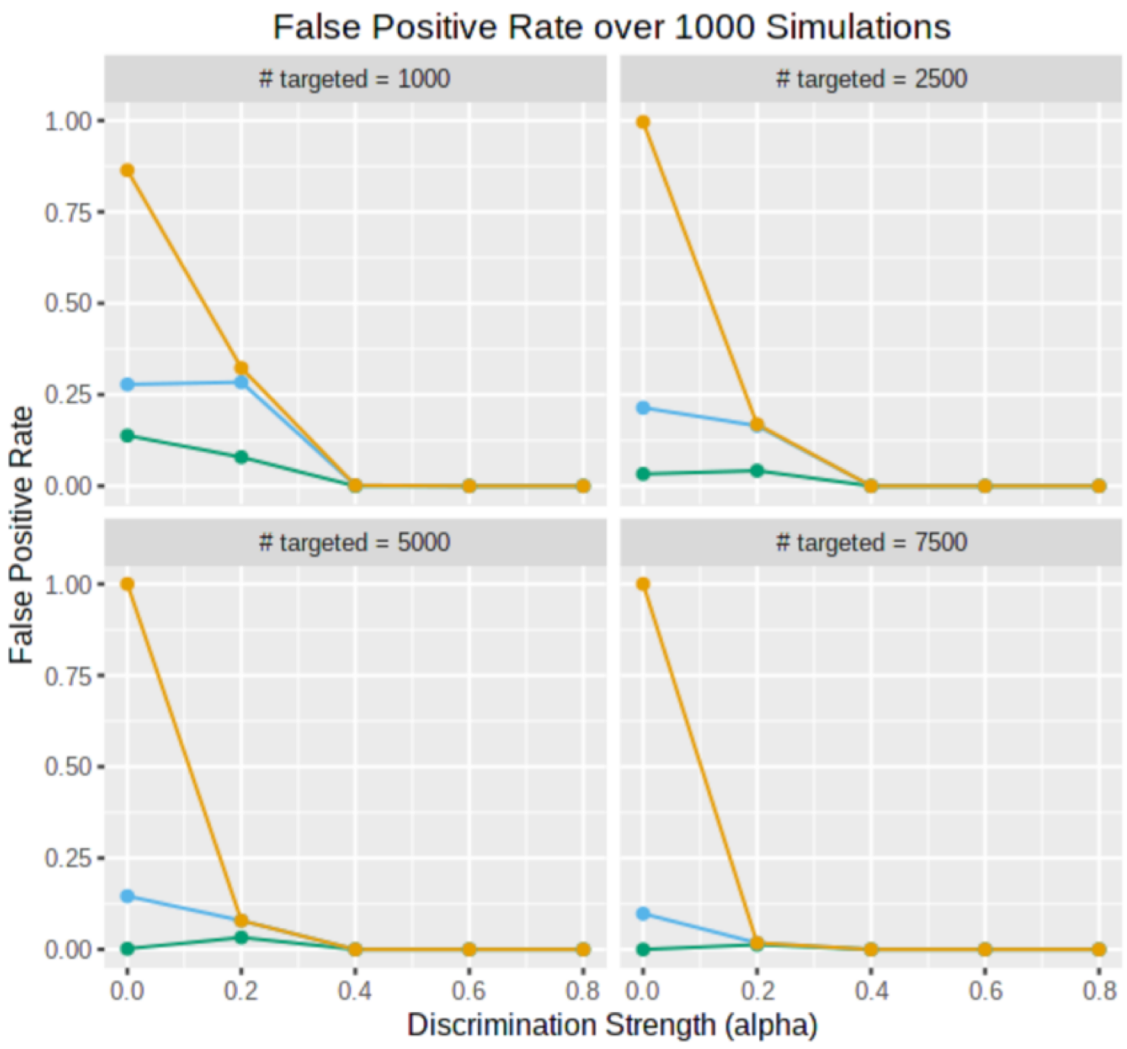
- Test set performance metrics
  - Inferred disparity
  - False Positive Rate/False Negative Rate (using $\beta = 1.25$)
  - Accuracy on non-targeted examples
  - Accuracy of expanded targeted audience

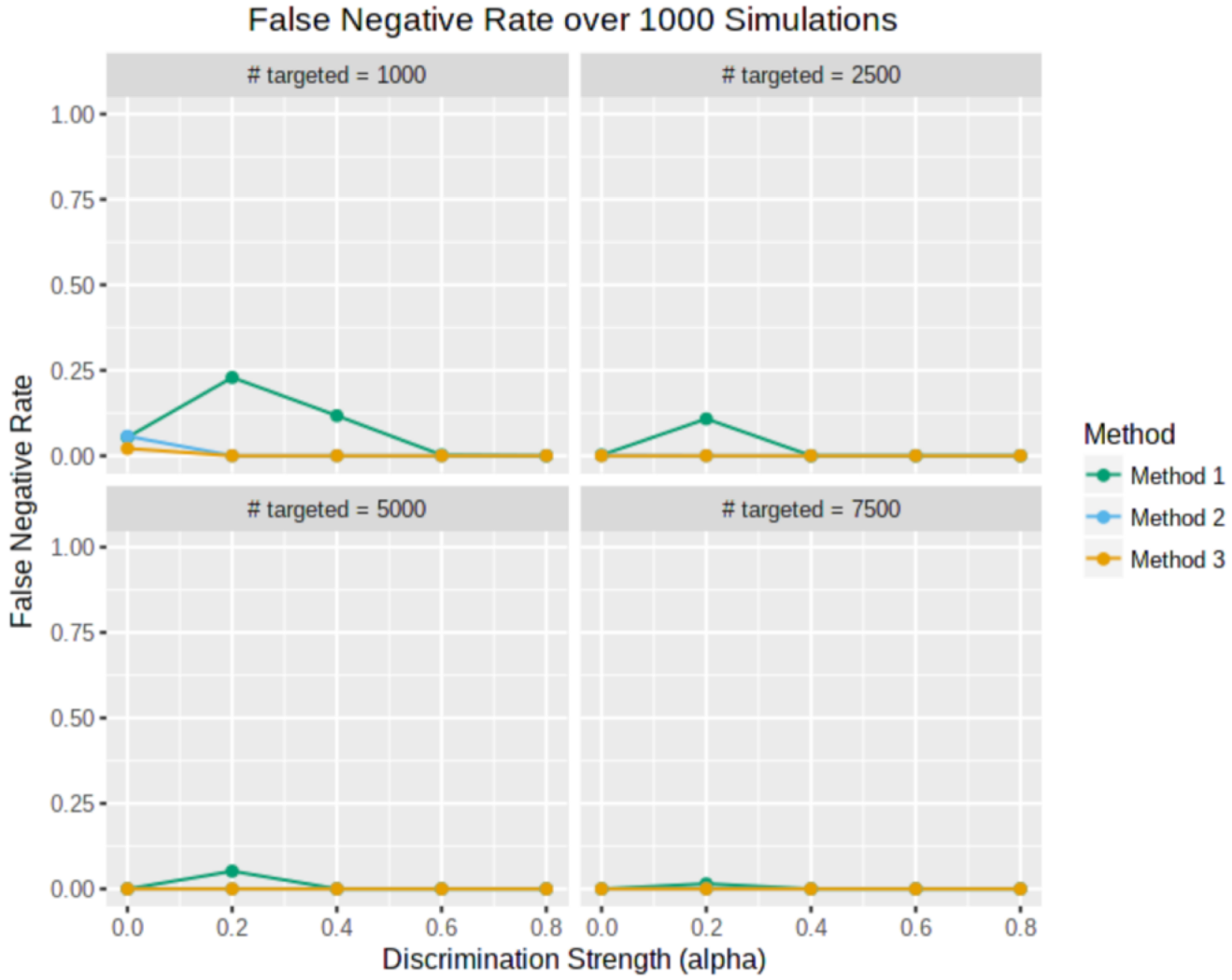
Results - Inferred Disparity

Inferred disparity converges to the true disparity as the number of targeted examples increases



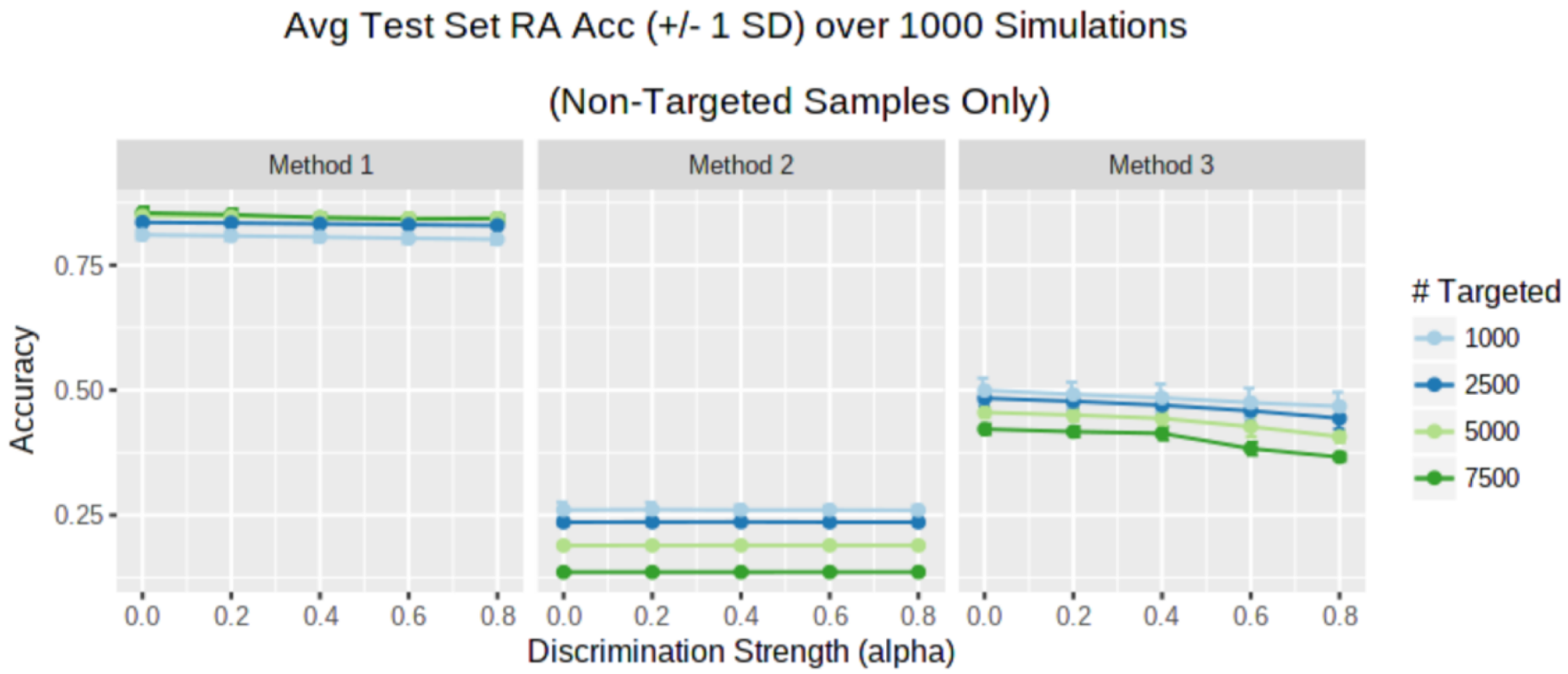
Results – FPR/FNR





Method 1 has lower FPR than competing methods, but higher FNR when there's a small number of targeted examples.

Results – Accuracy on Non-Targeted Samples

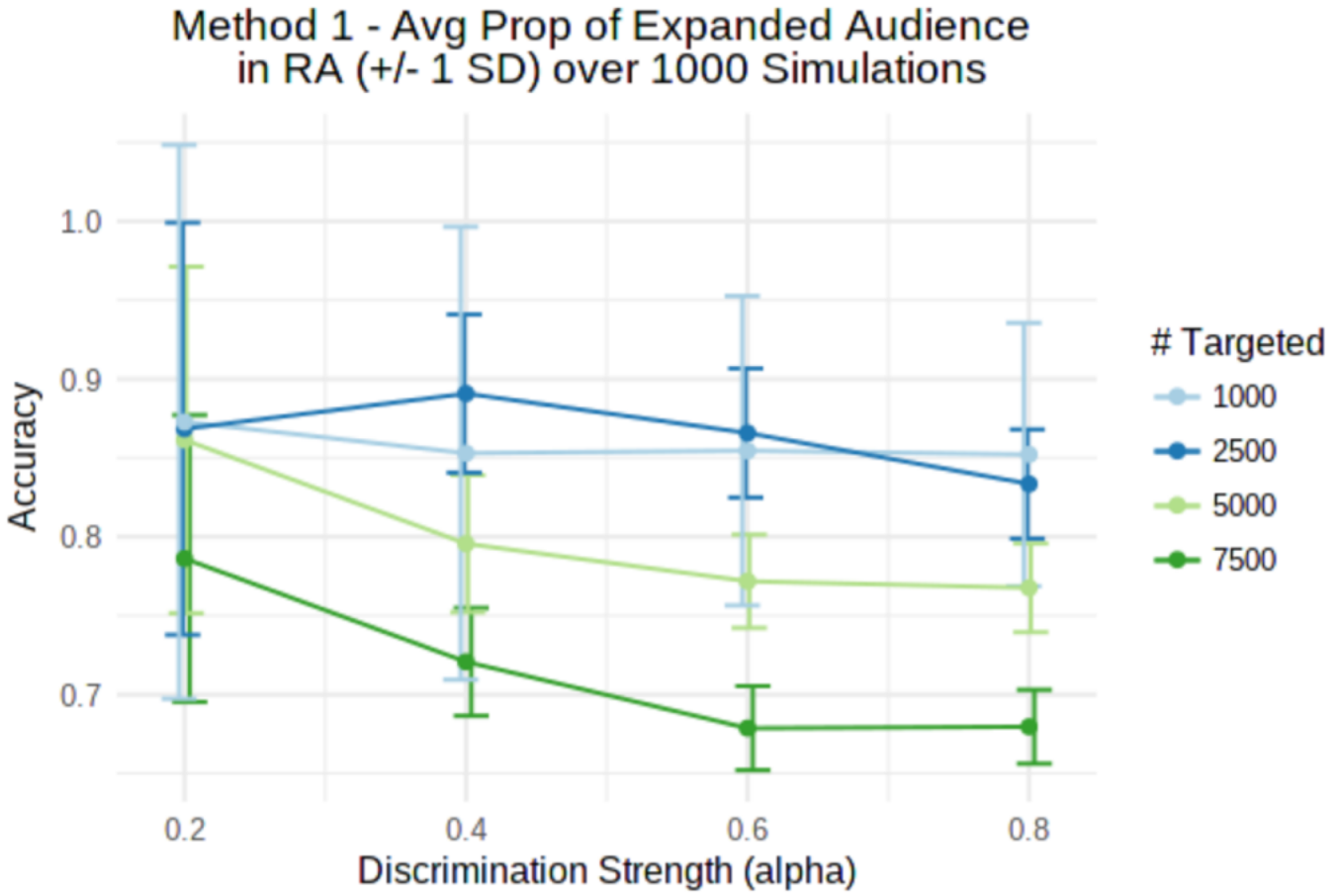


Method 1 clear winner based on accuracy

Results – Proportion of Expanded Audience in Relevant Audience

Mitigation strategy

- In cases where bias is detected, compute the number of sensitive examples needed to correct the disparity measure
- Suggest examples with highest likelihood of belonging to the relevant audience as options for expanding the targeted audience



Note: The # of examples that need to be added increases as the size of the targeted set or the discrimination strength increases

Summary

Preliminary results suggest that bias can be detected with high probability when the targeted list is sufficiently large and that expanding the targeted set is a reasonable mitigation strategy, but further investigation is needed to assess the validity of this approach.

Current work

- Studying model performance under different attack scenarios
- Investigating impact of correlation of sensitive and non-sensitive features
- Researching model tuning procedure designed specifically for PU-Learning

Acknowledgements:

Krishna Gummadi, Vedant Nanda and Till Speicher

Max Planck Institute