Responsible Machine Learning

Lecture 12: Algorithmic Debiasing

CS 4973-05

Fall 2023

Instructors: Avijit Ghosh
ghosh.a@northeastern.edu
Northeastern University, Boston, MA

Northeastern University

Agenda

1. Fair Classification
2. Fair Ranking

Northeastern University

Fair Classification

Review of Fair Classification Definitions

- **Variables**
  - Y is the true value (0 or 1 for binary classification)
  - C is the algorithm's predicted value
  - A is the protected attribute (gender, race, etc.)

Review of Fair Classification Definitions

- **Variables**
  - Y is the true value (0 or 1 for binary classification)
  - C is the algorithm's predicted value
  - A is the protected attribute (gender, race, etc.)
- **Definitions***
  - Demographic Parity: $P(C|A=0) = P(C|A=1)$
  - Equal Opportunity: $P(C|A=0,Y=1) = P(C|A=1,Y=1)$
  - Equalized Odds: $P(C|A=0,Y=y) = P(C|A=1,Y=y)$, *for* $y \in \{0,1\}$

* "A Survey on Bias and Fairness in Machine Learning" has even more definitions.

Definitions to Metrics

- **Definitions**
  - Demographic Parity: $P(C|A=0) = P(C|A=1)$
  - Equal Opportunity: $P(C|A=0,Y=1) = P(C|A=1,Y=1)$
  - Equalized Odds: $P(C|A=0,Y=y) = P(C|A=1,Y=y)$, for $y \in \{0,1\}$
- **Metrics**
  - Demographic Parity Difference: $P(C|A=1) - P(C|A=0)$
  - Equal Opportunity Difference: $P(C|A=1,Y=1) - P(C|A=0,Y=1)$
  - Equalized Odds Difference: $E[P(C|A=1,Y=y) - P(C|A=0,Y=y)]$

General Debiasing Approaches

- Pre-processing
    - Transform the training data (e.g. re-sampling, collecting more data)

General Debiasing Approaches

- Pre-processing
  - Transform the training data (e.g. re-sampling, collecting more data)
- In-processing
  - Transform the learning algorithm (e.g. different objective function, add constraints)

General Debiasing Approaches

- Pre-processing
  - Transform the training data (e.g. re-sampling, collecting more data)
- In-processing
  - Transform the learning algorithm (e.g. different objective function, add constraints)
- Post-processing
  - Transform the predictions (e.g. different thresholds)

Pre-processing Example

Sample Reweighting [1]: assign weights to individual data points, so the sample resembles what would have been generated by a "fair process"

[1] Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." *Knowledge and information systems* 33.1 (2012): 1-33.

Pre-processing Example

Sample Reweighting [1]: assign weights to individual data points, so the sample resembles what would have been generated by a "fair process"

Table 4 Sample job-application relation with weights

Sex	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	–	2
F	Non-nat.	Univ.	Education	–	0.67
F	Native	H. school	Education	–	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	–	0.67
F	Native	H. school	Board	+	1.5

[1] Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." *Knowledge and information systems* 33.1 (2012): 1-33.

Pre-processing Example

Sample Reweighting [1]: assign weights to individual data points, so the sample resembles what would have been generated by a "fair process"

Table 4 Sample job-application relation with weights

Sex	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	–	2
F	Non-nat.	Univ.	Education	–	0.67
F	Native	H. school	Education	–	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	–	0.67
F	Native	H. school	Board	+	1.5

For $a \in A, y \in Y$

$$w(a, y) = \frac{P(\text{expected})}{P(\text{observed})}$$

$$w(a, y) = \frac{P(A = a)P(Y = y)}{P(A = a \wedge Y = y)}$$

[1] Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." *Knowledge and information systems* 33.1 (2012): 1-33.

Pre-processing Example

Sample Reweighting [1]: assign weights to individual data points, so the sample resembles what would have been generated by a "fair process"

Table 4 Sample job-application relation with weights

Sex	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	−	2
F	Non-nat.	Univ.	Education	−	0.67
F	Native	H. school	Education	−	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	−	0.67
F	Native	H. school	Board	+	1.5

Expected "fair" probabilities

- $P(A = M)\ P(Y = +) = 0.3$
- $P(A = M)\ P(Y = -) = 0.2$
- $P(A = F)\ P(Y = +) = 0.3$
- $P(A = F)\ P(Y = -) = 0.2$

[1] Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." *Knowledge and information systems* 33.1 (2012): 1-33.

Pre-processing Example

Sample Reweighting [1]: assign weights to individual data points, so the sample resembles what would have been generated by a "fair process"

Table 4 Sample job-application relation with weights

Sex	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	−	2
F	Non-nat.	Univ.	Education	−	0.67
F	Native	H. school	Education	−	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	−	0.67
F	Native	H. school	Board	+	1.5

Observed probabilities:

- $P(A = M \wedge Y = +) = 0.4$
- $P(A = M \wedge Y = -) = 0.1$
- $P(A = F \wedge Y = +) = 0.2$
- $P(A = F \wedge Y = -) = 0.3$

[1] Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." *Knowledge and information systems* 33.1 (2012): 1-33.

Pre-processing Example

Sample Reweighting [1]: assign weights to individual data points, so the sample resembles what would have been generated by a "fair process"

Table 4 Sample job-application relation with weights

Sex	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	–	2
F	Non-nat.	Univ.	Education	–	0.67
F	Native	H. school	Education	–	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	–	0.67
F	Native	H. school	Board	+	1.5

Weights

- w(A = M, Y = +) = 0.3/0.4 = 0.75
- w(A = M, Y = -) = 0.2/0.1 = 2
- w(A = F, Y = +) = 0.3/0.2 = 1.5
- w(A = F, Y = -) = 0.2/0.3 = 0.67

[1] Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." *Knowledge and information systems* 33.1 (2012): 1-33.

In-processing Example

Adversarial Debiasing [2]: maximize the model's ability to predict the output, while minimizing the adversary's ability to predict the protected attribute

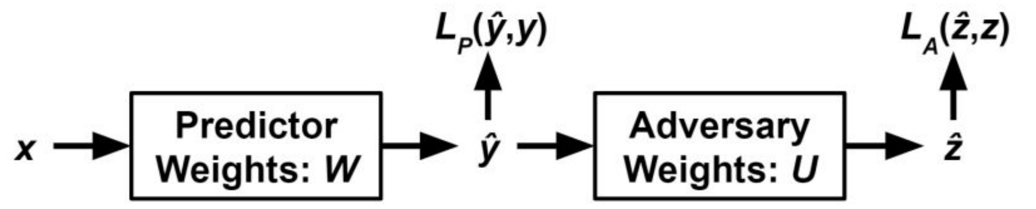


Figure 1: The architecture of the adversarial network.

[2] Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018.

In-processing Example

Adversarial Debiasing [2]: maximize the model's ability to predict the output, while minimizing the adversary's ability to predict the protected attribute

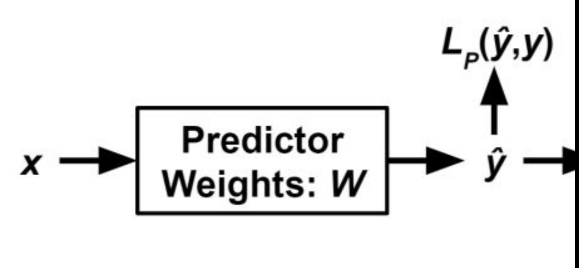


Figure 1: The architecture of t

Predictor Model

- learns function $y = f(x)$
- minimizes loss $L_p(\hat{y}, y)$

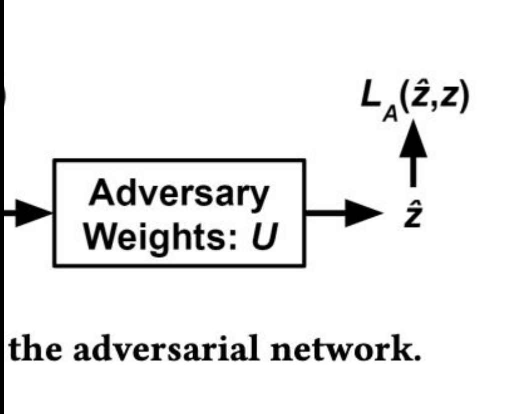
[2] Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018.

In-processing Example

Adversarial Debiasing [2]: maximize the model's ability to predict the output, while minimizing the adversary's ability to predict the protected attribute

Adversary Model

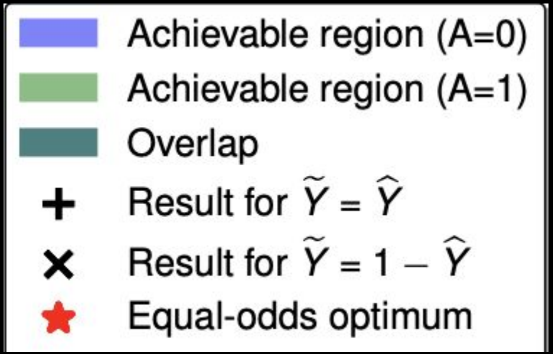
- learns function $z = g(y)$
- minimizes loss $L_A(\hat{z}, z)$

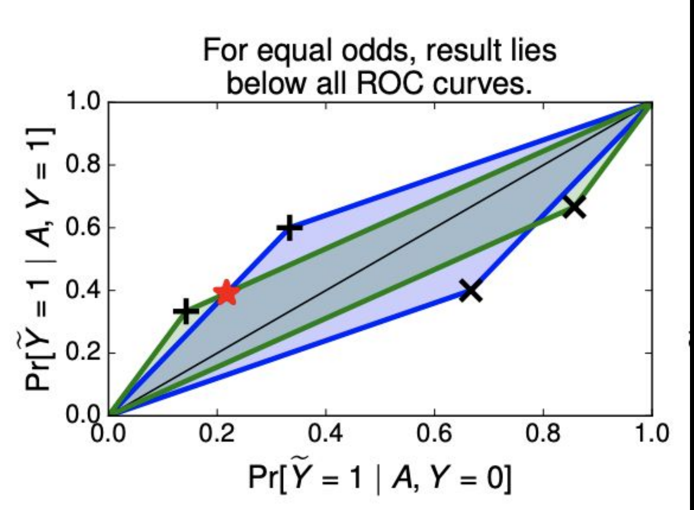


[2] Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018.

Post-processing Example

Equalized Odds Post-processing [3]: optimize a constrained linear program that is a function of Y, C (they call it Ŷ) , and A

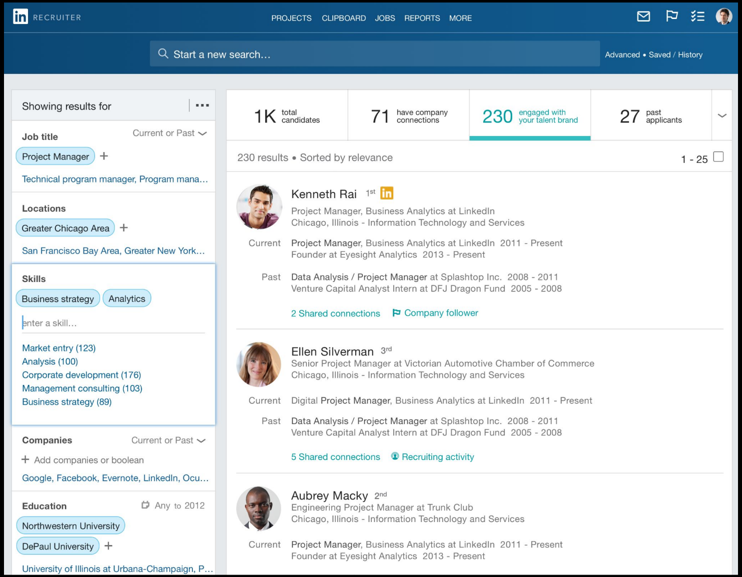


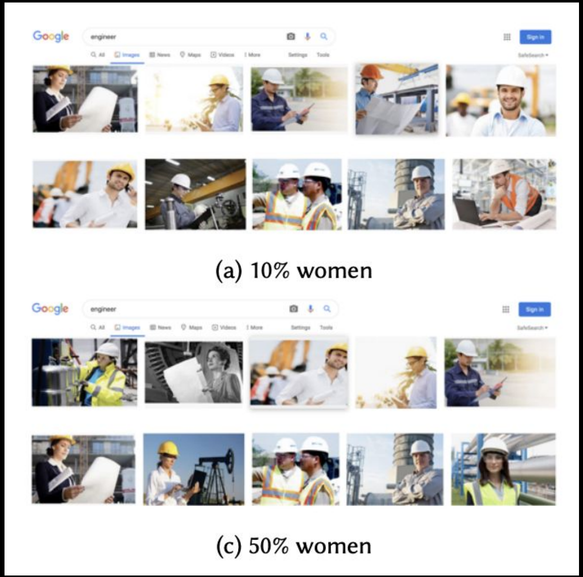


[3] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." *Advances in neural information processing systems* 29 (2016).

Fair Ranking

Fair Ranking Motivation





(a) 10% women

(c) 50% women

Fair Ranking Differences

What are differences between classification and ranking that might be important for fairness?

Fair Ranking Differences

- Selecting a ranked list instead of making individual classifications
- Evaluating items relatively instead of independently

[4] Zehlike, Meike, Ke Yang, and Julia Stoyanovich. "Fairness in ranking: A survey." *arXiv preprint arXiv:2103.14000* (2021).

Ranking Bias Metrics

Representation Based

$$Skew_{group,k} = \frac{Fraction\ of\ group\ members\ in\ top\ K}{Fraction\ of\ group\ members\ overall}$$

$$NDKL = Normalised\ Discounted\ KL\ divergence$$
$$between\ the\ group\ distributions\ in\ top\ K\ and\ overall\ population$$

The ideal value for Skew is 1, and NDKL is 0

Setup: Evaluation Metrics

Exposure Based

$$\text{Attention}_p@k(\tau) = 100 \times (1 - p)^{k-1} \times (p)$$

$$ABR = \frac{\textit{Attention of group with min. avg attention}}{\textit{Attention of group with max. avg attention}}$$

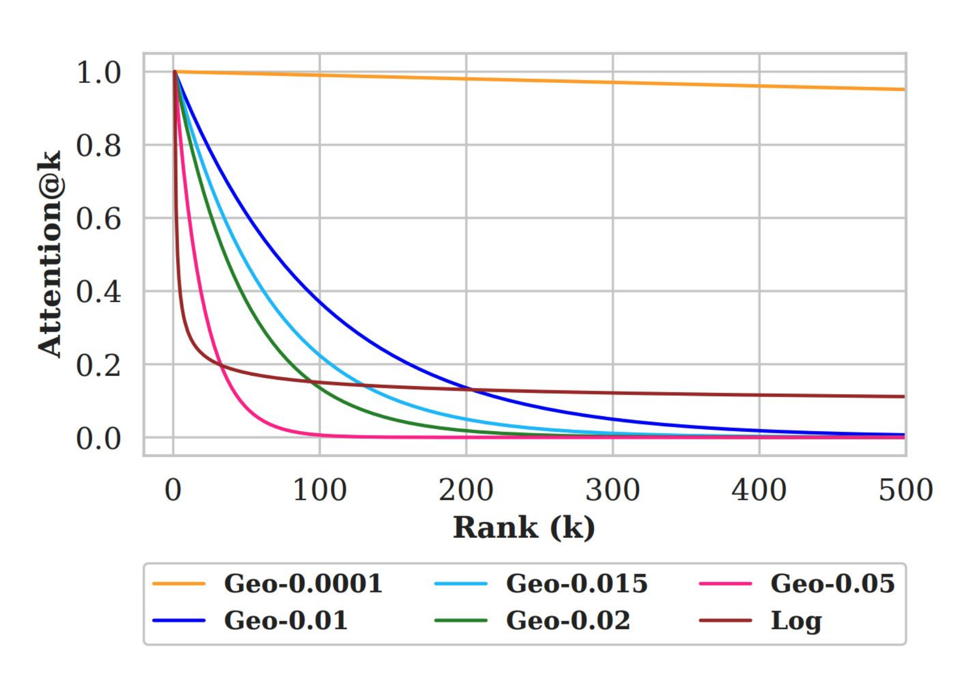
Setup: Evaluation Metrics

Exposure Based

$$\text{Attention}_p@k(\tau) = 100 \times (1-p)^{k-1} \times (p)$$

$$ABR = \frac{Attention\ of\ group\ with\ min.\ avg\ attention}{Attention\ of\ group\ with\ max.\ avg\ attention}$$

The ideal value for ABR is 1



Setup: Evaluation Metrics

Ranking Quality

$$DCG_n = \sum_{i=1}^{n} \frac{rel_i}{\log_2^{i+1}},$$

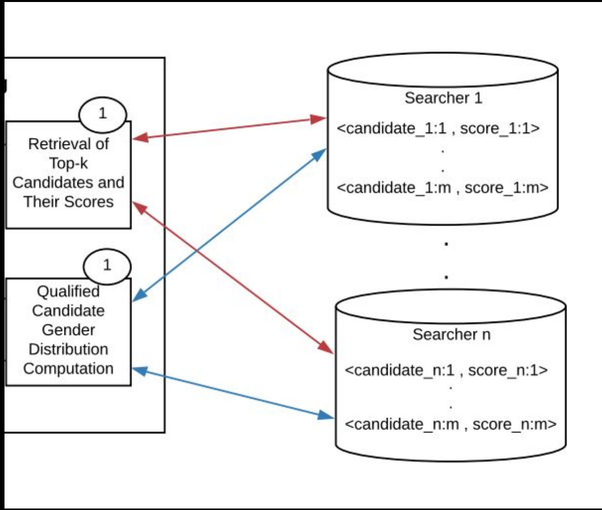
$$NDCG_n = \frac{DCG_n}{IDCG_n},$$

The ideal value for NDCG is this case is 1

NDCG - Normalized Discounted Cumulative Gain, very popular in IR Literature to measure ranking quality.

Fair Ranking: LinkedIn Example

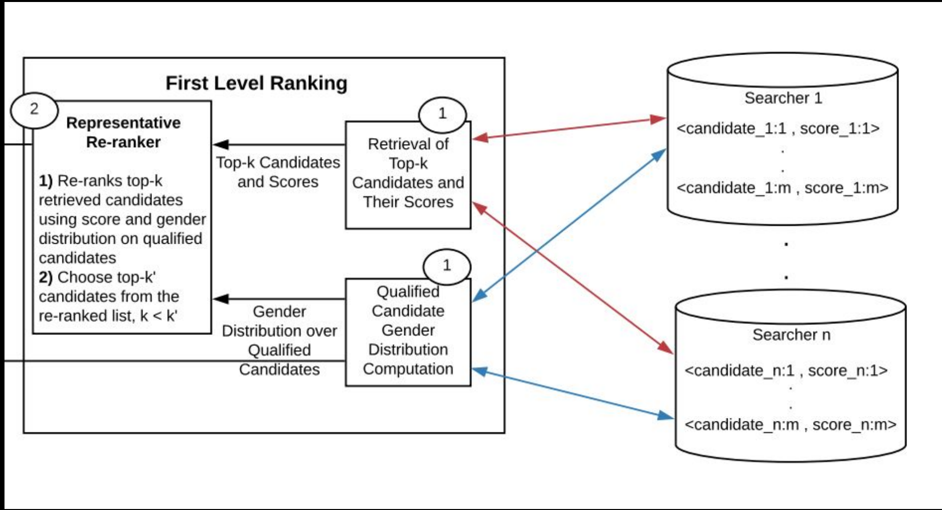
Step 1: retrieve top-k candidates, compute their gender distribution



[4] Geyik, Sahin Cem, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search." *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2019.

Fair Ranking: LinkedIn Example

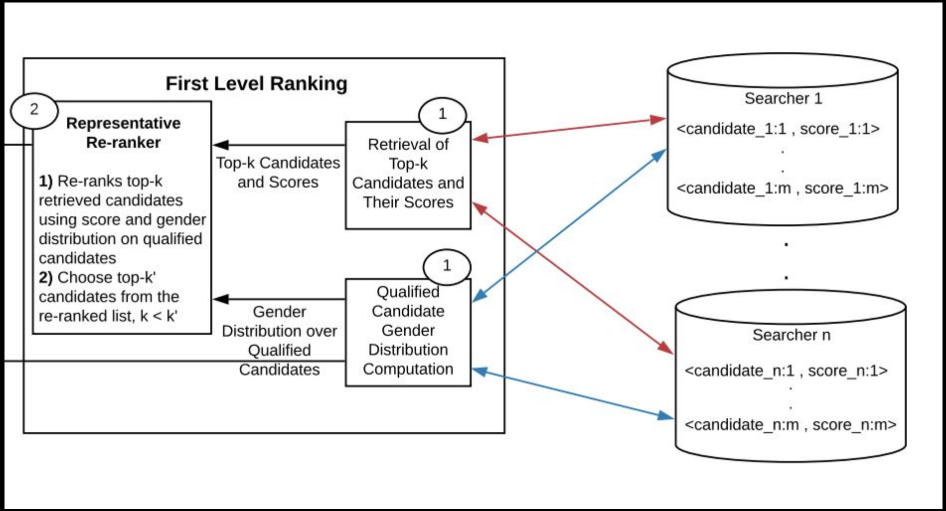
Step 2: re-rank top-k candidates so exposure of groups matches gender distribution



[4] Geyik, Sahin Cem, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search." *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2019.

Fair Ranking: LinkedIn Example

Step 2: re-rank top-k candidates so exposure of groups matches gender distribution



Caveat: LinkedIn's algorithm only intervenes with respect to gender!

[4] Geyik, Sahin Cem, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search." *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2019.

Fair Classification and Ranking Challenges

- **What if we don't have access to demographic labels?**
- We want to achieve fairness with respect to multiple, intersectional protected attributes.
- We often want to prioritize underrepresented groups, instead of simply equalizing a metric across groups

Thank You!