Evaluating Blocking Biases in Entity Matching

- Mohammad Hossein Moslemi
- Harini Balamurugan
- Mostafa Milani

IEEE BigData 2024, Washington D.C., USA

Content

- **Background**
  - Entity matching and Blocking definition
  - Blocking methods
  - Quality of Blocking
- **Fairness and Blocking**
- **Measuring bias in Blocking**
- **Experimental Results**

Entity Matching and Blocking

- **Entity Matching (EM):** Identifies record pairs from data sources that refer to the same entity.

- *Examples*:
  - Background Checks: Airport, Loans, …
  - Healthcare
  - …

- **Blocking:**
  - Groups similar records to filter unlikely matches.
  - Reducing computational costs and time.

Blocking Methods

- **Traditional methods:**
  - Group records by attribute similarities
  - Techniques like exact matches and sorted windows.

- *Examples:* Suffix array blocking, Sorted Neighborhood, …

- **Deep learning methods:**
  - Use deep learning to identify matches.
  - Techniques like automated rule learning and threshold-based similarity.

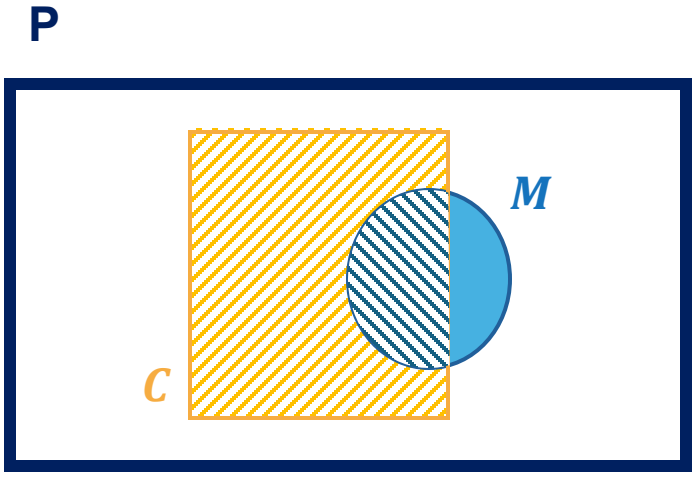
- *Examples:* AUTO-block, CTT-block, …

Quality of Blocking

- **Blocking quality:**
  - Its ability to Maximize true matches and minimizing non-matching pairs.

- For datasets D_1 and D_2:
  - P: All possible pairs: $D_1 \times D_2$
  - M: True matches
  - C: Candidate set after blocking

- **Metrics:**
  - Reduction ratio (RR): $1 - \frac{|C|}{|P|}$
  - Pair completeness (PC): $\frac{|C \cap M|}{|M|}$
  - Pair quality (PQ): $\frac{|C \cap M|}{|P|}$



EM, Blocking and Fairness

- **Examples of bias in EM:**

Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.

By Julia Carpenter

The Washington Post

Airline "no-fly" lists trample the rights of people of color. Seattle should not allow hotels to create "no stay" lists

Amy Roe, Former ACLU-WA Senior Writer

Published: Friday, July 19, 2019

ACLU Washington

- **Bias propagation:** Blocking biases affect matching; fairness in blocking is crucial.

Bias measurement in Blocking

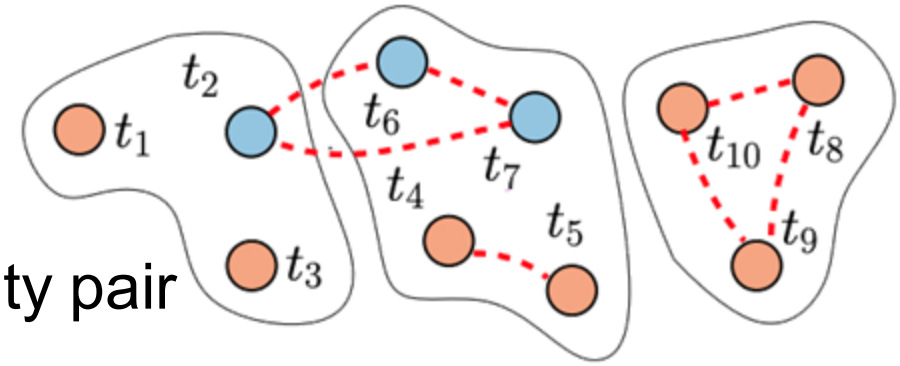
- **Minority Pair:** A pair (t_1, t_2) is minority if either 't_1' or 't_2' belongs to a minority group.

- **Group-wise metrics:**
  - P_g: All pairs in group $g \in \{a, b\}$
  - Similarly, C_g and M_g.

$$RR_g = 1 - \frac{|C_g|}{|P_g|}$$
$$PC_g = \frac{|C_g \cap M|}{|M_g|}$$
$$\Rightarrow$$
$$\Delta RR = RR_b - RR_a$$
$$\Delta PC = PC_b - PC_a$$

Bias measurement in Blocking

- **Example:** A blocking with three blocks.
  - Minority entitles in blue, majority in orange.
  - True pairs with red dash lines.



- *Before Blocking:*
  - Total initial pair: $\frac{10\times9}{2} = 45$
  - 21 Majority pair, 24 Minority pair

- *After Blocking:*
  - Total Pairs: 12
  - 5 Majority pair, 7 Minority pair

- $RR_a = 1 - \frac{7}{24} \approx 0.71,\ RR_b \approx 0.76 \rightarrow\ \boldsymbol{\Delta RR \approx 0.05}$
- $PC_a = \frac{1}{3} \approx 0.33,\ PC_b = 1 \rightarrow\ \boldsymbol{\Delta PC \approx 0.67}$

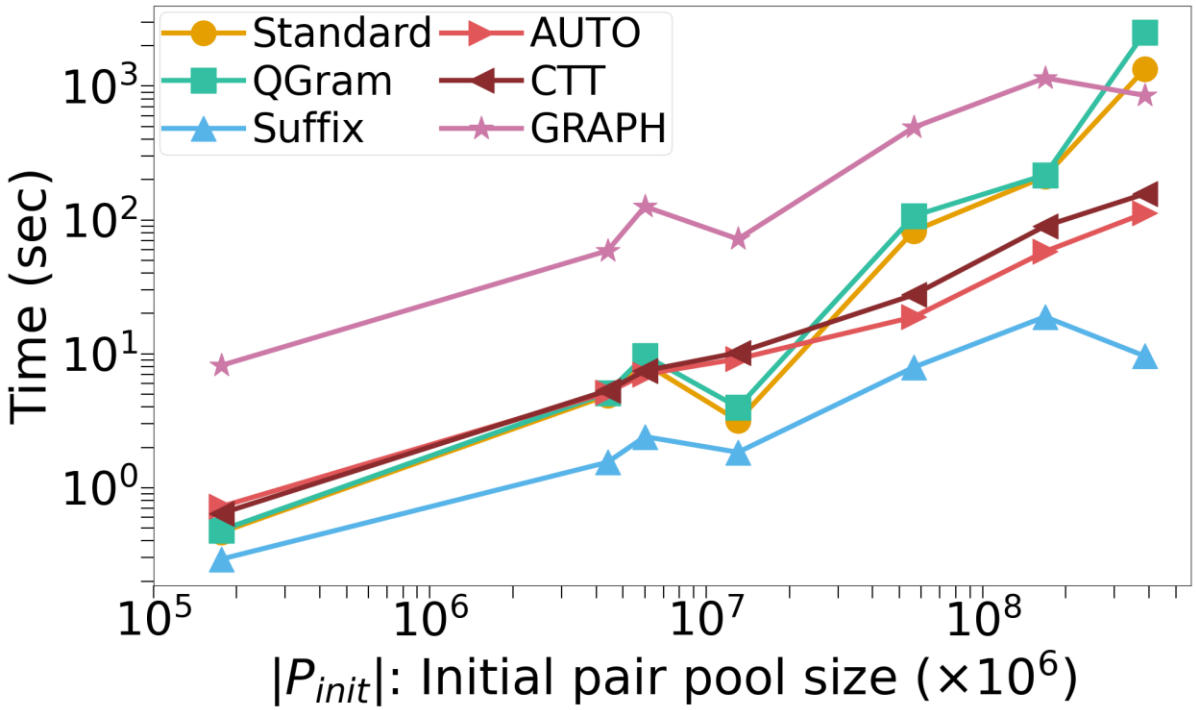
Experiments

- **Datasets:** 7 EM benchmark datasets:
  - *Amazon-Google (AMZ-GOO)*
  - *Walmart-Amazon (WAL-AMZ)*
  - *DBLP-GoogleScholar (DBLP-GOO)*
  - *DBLP-ACM (DBLP-ACM)*
  - *Beer (BEER)*
  - *Fodors-Zagat (FOD-ZAG)* → *|P|= 180k pairs*
  - *iTunes-Amazon (ITU-AMZ)* → *|P| = 382M pairs*

- **Blocking methods:**
  - Traditional:
    - Standard, Qgram, EXT-Qgram, Suffix, EXT-Suffix
  - Deep learning:
    - AUTO, CTT, Semantic Graph

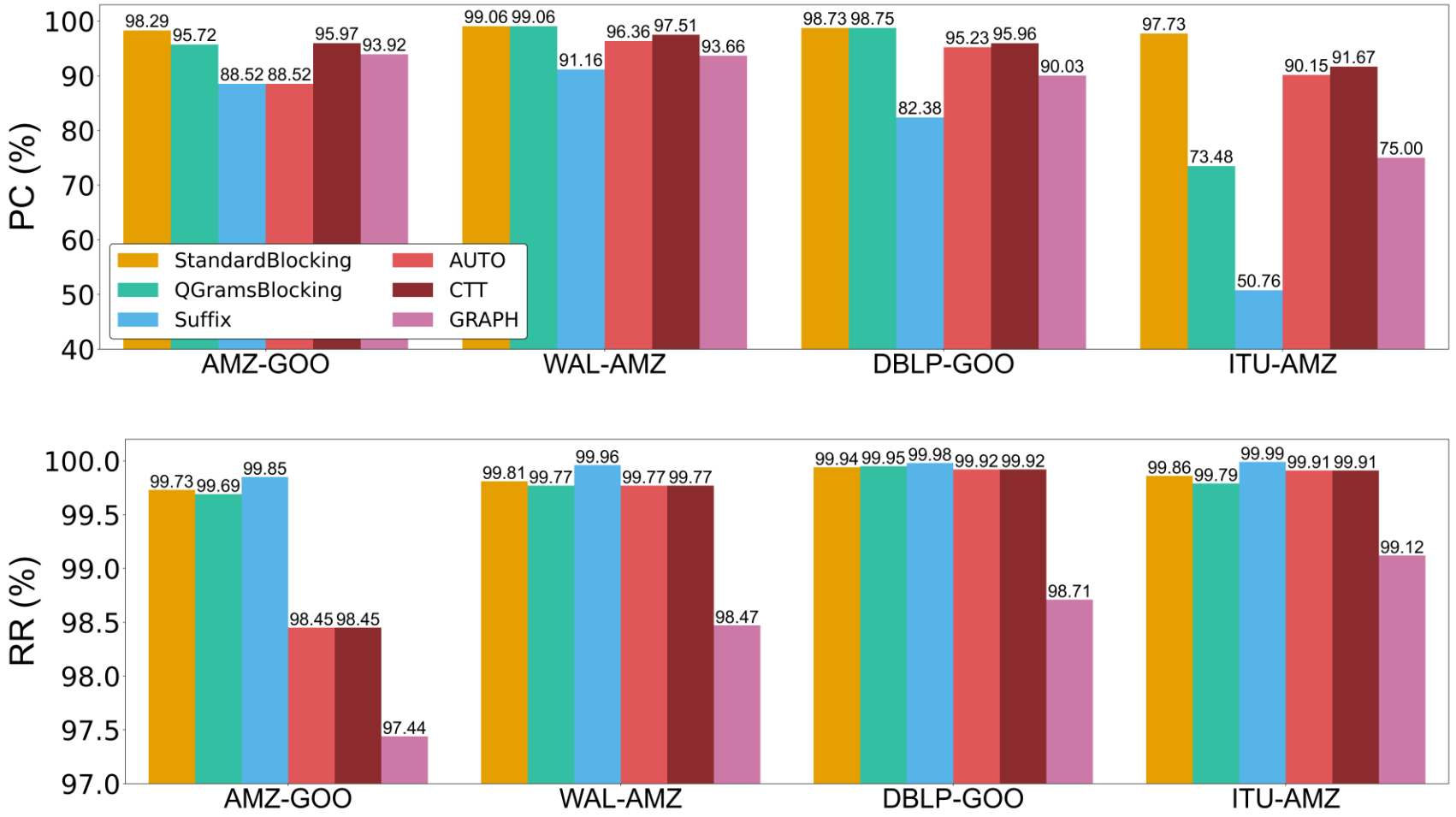
Experiments

- **Runtime evaluation:**



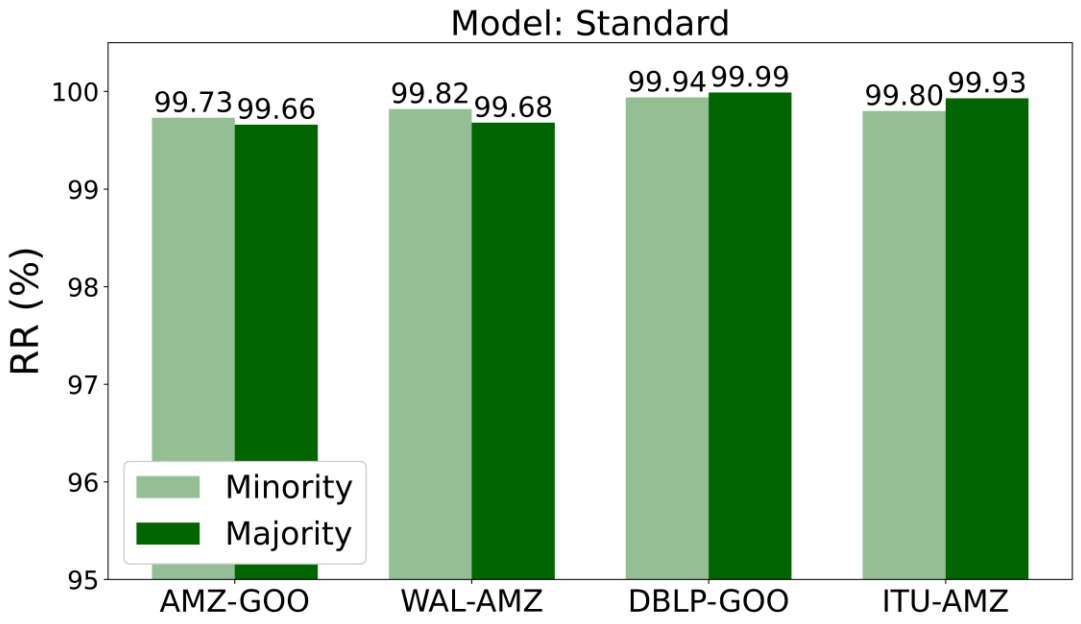
Experiments

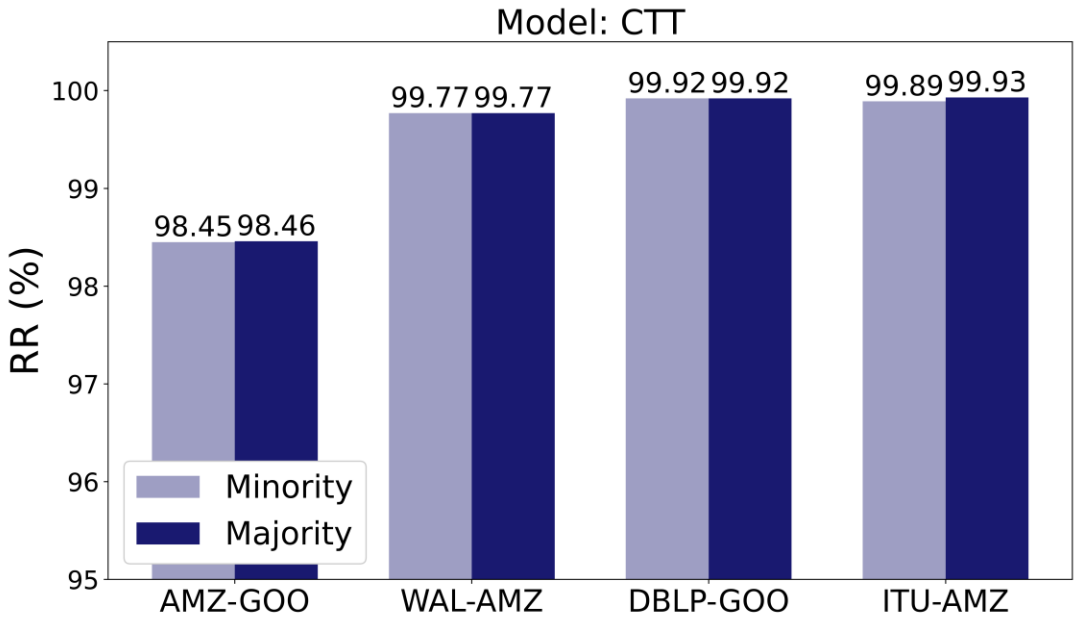
- **Blocking quality:**



Experiments

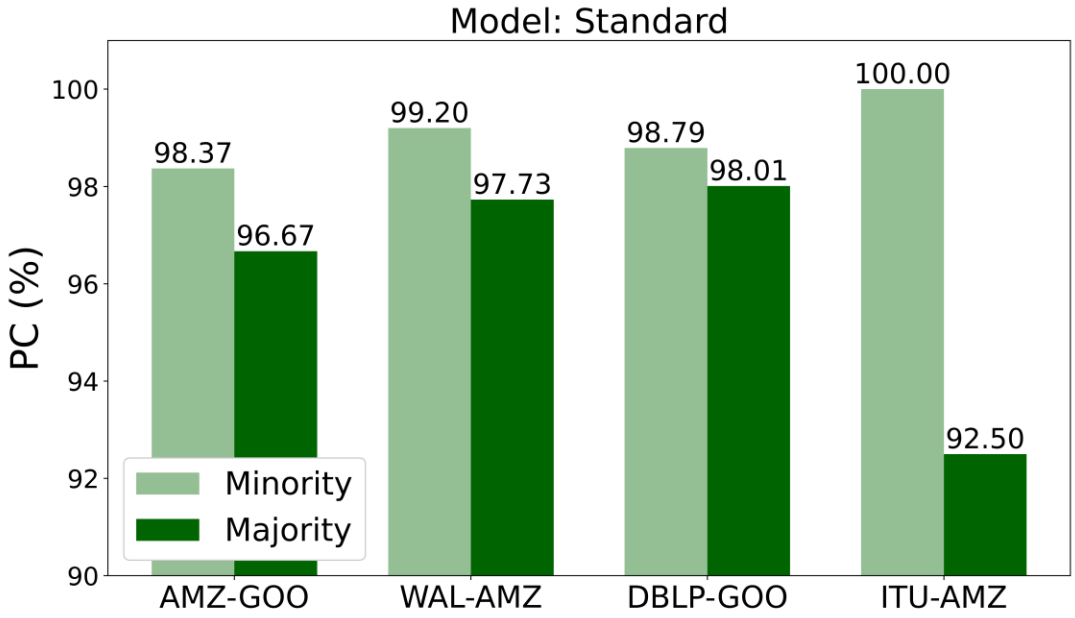
- **Bias analysis (RR):**

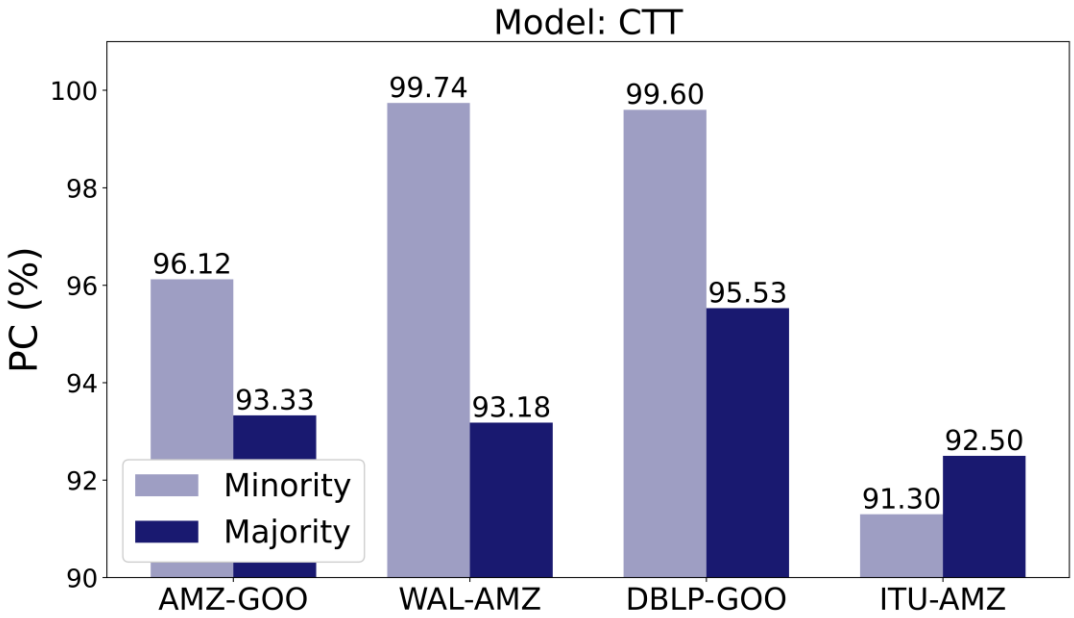




Experiments

- **Bias analysis (PC):**





Experiments

- **Bias propagation to EM:**

Model	AMZ-GOO
StdBlck	1.70 (98.37, 96.67)
QGram	-1.01 (95.66, 96.67)
XQGram	6.16 (94.49, 88.33)
Suffix	16.01 (89.34, 73.33)
XSuffix	18.15 (84.82, 66.67)
AUTO	8.98 (88.98, 80.00)
CTT	2.79 (96.12, 93.33)
GRAPH	**0.62** (93.95 , 93.33)

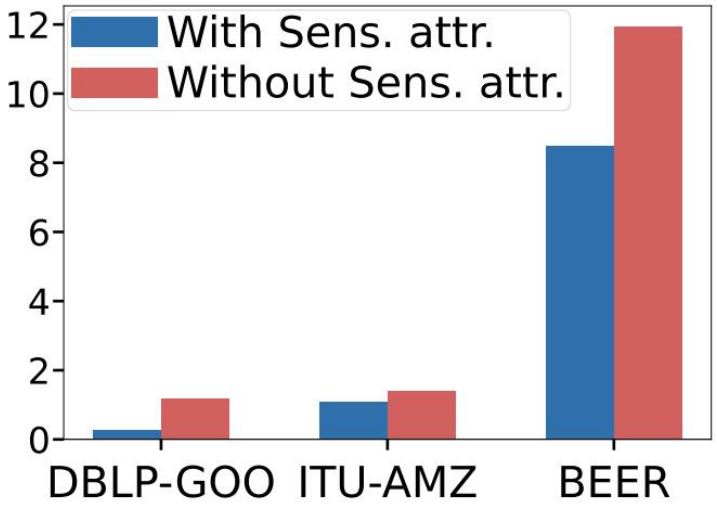
PC Bias on amazon-google

Metric	DP (%)	EO (%)	EOD (%)
QGram	4.42×10^{-3}	1.01	1.01
XSuffix	8.11×10^{-3}	18.16	18.16

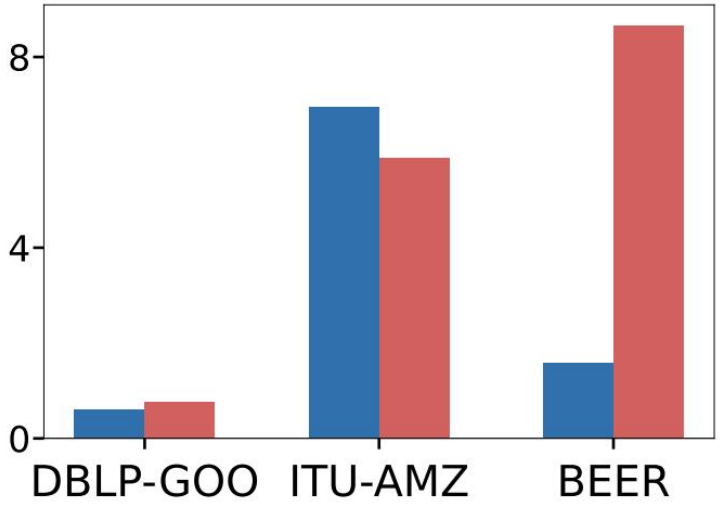
Propagated bias with a perfect matcher

Experiments

- **Removing sensitive attribute:**



a) Suffix



b) AUTO

Conclusion and Future work

- **Blocking Bias Impact**
  - Blocking in EM simplifies complexity but can introduce significant biases.
- **Method Variability**
  - No single blocking method consistently reduces disparities across datasets.
- **Future Directions**
  - Develop debiasing methods for blocking and extend them across the EM pipeline.

References

- S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, “Deep learning for entity matching: A design space exploration,” in SIGMOD, 2018.
- C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in ITCS, 2012.
- G. Papadakis, D. Skoutas, E. Thanos, and T. Palpanas, “Blocking and filtering techniques for entity resolution: A survey,” CSUR, 2020
- A. Zeakis, G. Papadakis, D. Skoutas, and M. Koubarakis, “Pre-trained embeddings for entity resolution: an experimental analysis,” PVLDB, 2023.
- N. Shahbazi, J. Wang, Z. Miao, and N. Bhutani, “Fairness-aware data preparation for entity matching,” in ICDE. IEEE, 2024.