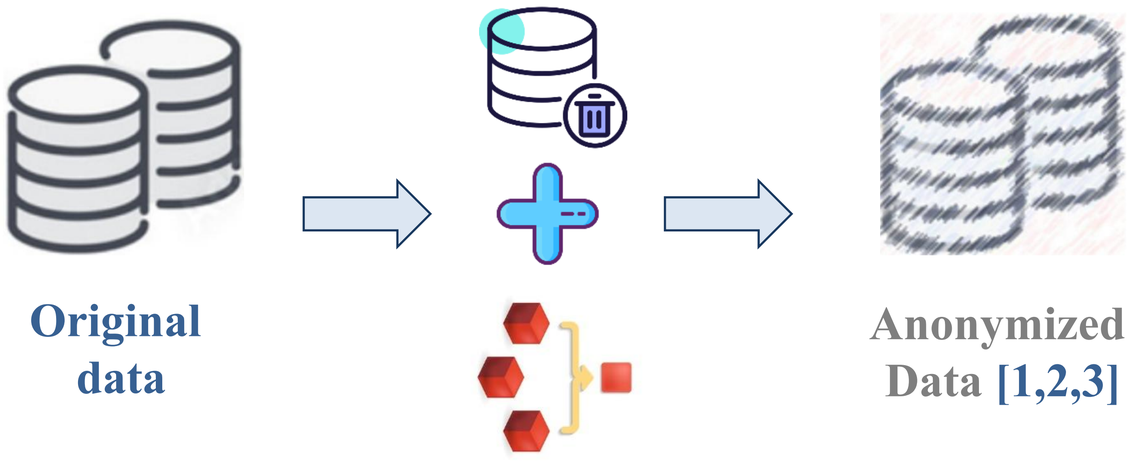
Fair Play for Individuals, Foul Play for Groups? Auditing Anonymization's Impact on ML Fairness

Héber H. Arcolezi† Mina Alishahi‡ Adda-Akram Bendoukha⁋ Nesrine Kaaniche⁋
†Inria ‡Open Universiteit ⁋Télécom SudParis
heber.hwang-arcolezi@inria.fr

Paper Code

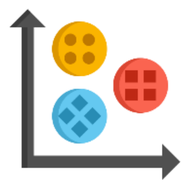
1. FROM ANONYMIZATION TO DATA DISTORTION

Record suppression



YES!

"Safe in the crowd"



BUT...

Alters data distributions

Distorted data may bias ML fairness…?

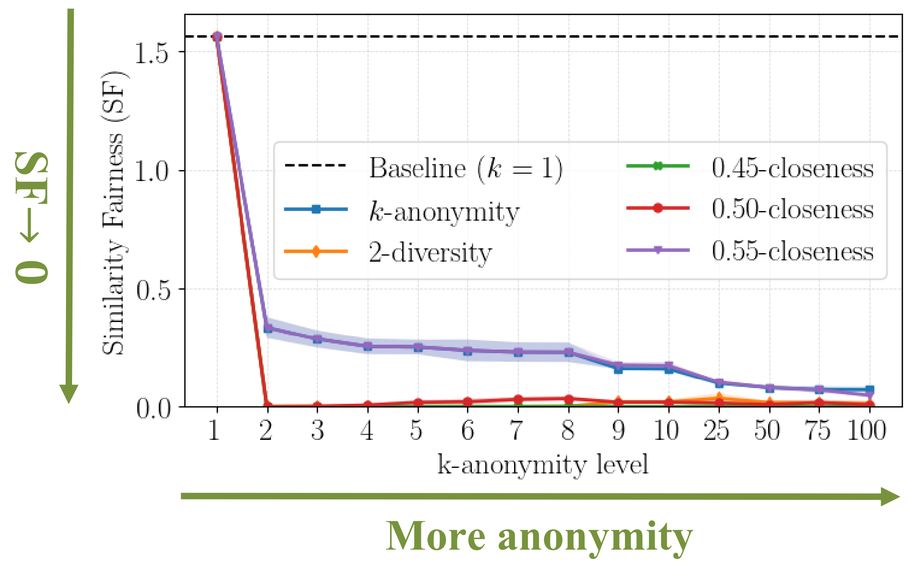
2. FAIRNESS EFFECTS VARY BY LEVEL OF ANALYSIS

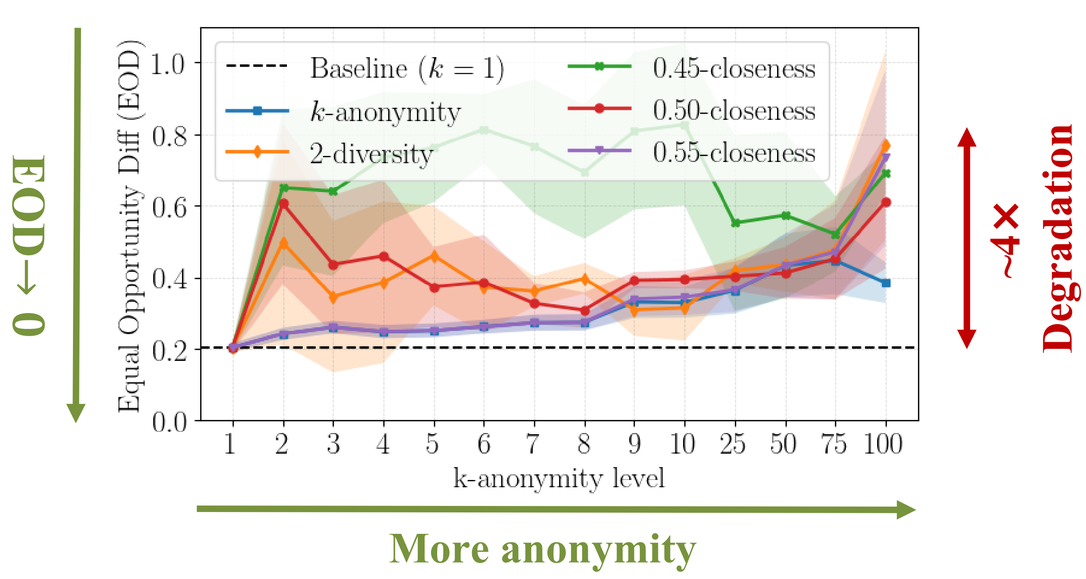
Individual fairness [4]
"Similar individuals → similar predictions"
(measures local consistency of model outputs)

VS

Group fairness [5]
"Equal outcomes across demographic groups"
(measures parity between protected groups)

Anonymization tends to ***smooth predictions*** (↑ **individual fairness**)
but to ***amplify group disparities*** (↓ **group fairness**)





3. ROBUSTNESS AND GENERALIZATION

✅ **Consistent results across**

- **Datasets** → Adult, Compas, ACSIncome
- **Protected attributes** → Gender, Race
- **Models** → XGBoost, RF, LGBM, MLP
- **Metrics** → Utility & individual and group fairness

🔍 **Further research questions**

- **Record suppression** → minorities **most affected?**
- **Target distribution variation** → **fairness varies?**
- **Data size variation** → **reduces stability?**

4. KEY TAKEAWAYS & PERSPECTIVES

🟩 **Individual fairness ↑**

→ homogeneous data… "**fairness washing**" [6] ⚠

🟥 **Group fairness ↓**

→ up to 4× disparity

⚖ **Utility ↓ slightly**

🔮 **Future work**

- **Fairness-aware anonymization**
- Joint optimization of **privacy-fairness-utility**
- Extend to **multiclass classification**, **regression**

REFERENCES

[1] Sweeney, Latanya. "k-anonymity: A model for protecting privacy". International journal of uncertainty, fuzziness and knowledge-based systems, 2022.

[2] Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity". ACM TKDD, 2007.

[3] Li, Ninghui, et al. "t-closeness: Privacy beyond k-anonymity and l-diversity". ICDE, 2007.

[4] Dwork, Cynthia, et al. "Fairness through awareness". Proceedings of the 3rd innovations in theoretical computer science conference. 2012.

[5] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". NeurIPS, 2016

[6] Aïvodji, Ulrich, et al. "Fairwashing: the risk of rationalization". ICML, 2019.

ACKNOWLEDGEMENTS

This work has been partially supported by the French National Research Agency (ANR), under contracts "ANR-24-CE23-6239" (AI-PULSE), "ANR-22-CE39-0002" (EQUIHID), and "ANR 22-PECY-0002" (IPoP).