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

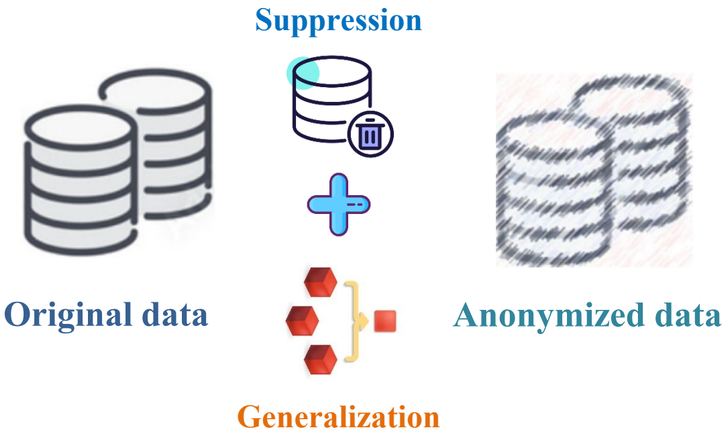
Héber H. Arcolezi[†], Mina Alishahi[‡], Adda-Akram Bendoukha[⁋], Nesrine Kaaniche[⁋]

[†]Inria [‡]Open Universiteit [⁋]Télécom SudParis

heber.hwang-arcolezi@inria.fr

Paper GitHub

"How does anonymization impact fairness metrics in ML?"



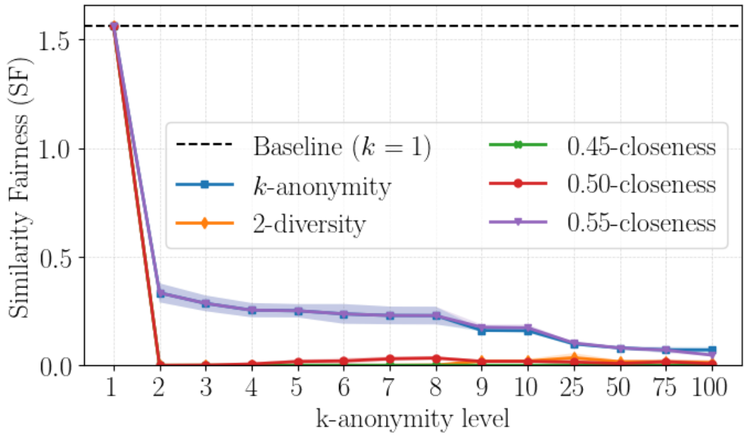
YES! BUT...

"Safe in the crowd"

Alters data distributions

Individual fairness generally **improves** (smoother, more homogeneous data)

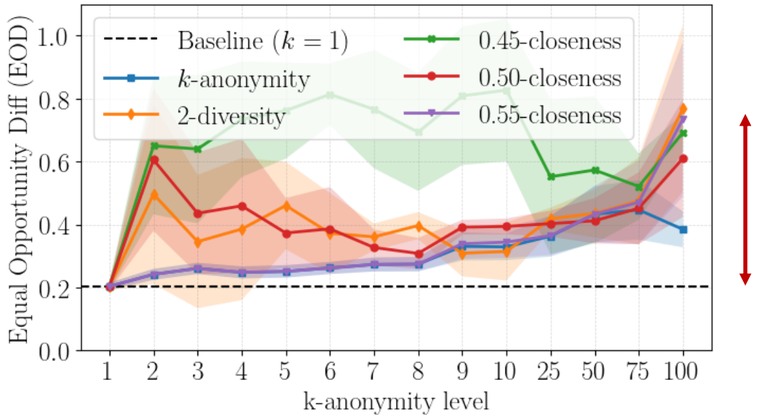
"Similar individuals → similar predictions"
(measures local consistency of model outputs)



More anonymity

Group fairness generally **worsens** (up to 4× degradation)

"Equal outcomes across demographic groups"
(measures parity between protected groups)



More anonymity