Protecting Individual Privacy Using Differential Privacy
Thomas Butler, Justina Zou, Yunyao Zhu, Yikai Wu
David Pujol, Ashwin Machanavajjhala

Differential privacy can improve fairness of binary classification models when trained on balanced data.

Motivation
Census Bureau will use differential privacy for the 2020 Census. Potential compromise on fairness of prediction making.

Goal
Find out how differential privacy would impact fairness of binary prediction models for predicting housing mortgage approval.

Definitions
Differential Privacy: Output should hardly change when a single individual joins or leaves the dataset.
Privacy quantified by epsilon (ε): Larger ε → less noise → more privacy
Smaller ε → more noise → less privacy

Fairness
Similar individuals/groups should be treated similarly.
Fairness can be measured by Approval Disparate Impact (DI): 
\[\text{DI} = \frac{\text{Pr(loan approval | protected group)}}{\text{Pr(loan approval | unprotected group)}}\]
0.8 < DI < 1.2 (mandated by US labor law)

Results
Finding #1: As a ML model becomes more private, AUC can decrease by up to 15%.
Finding #2: When training data is balanced by both action and race, both regression and neural net model achieve optimal fairness.
Finding #3: When training with data balanced by action, as privacy guarantee increases, neural net model becomes more fair.
Finding #4: As privacy increases, the features that the private neural network deem important vary more than the non-private model, potentially showing why accuracy decreases.

Future work
How does changing the dataset (and thereby the definition of fairness) affect the fairness of decisions?
How do private inputs affect the fairness of decisions?
How do different private ML models affect the fairness of decisions?

Methods
Data Preprocessing
● Balancing
● Dimension reduction
● Transformation
● Binarization

Machine Learning Models
Neural Net - TFP:Privacy
Regression - PrivateLR
Tested 6 epsilons [2, 0.0625] with 50 runs each

References