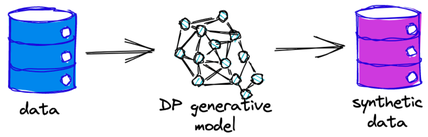
Robin Hood and *Matthew* Effects: Differential Privacy Has Disparate Impact on Synthetic Data

Georgi Ganev[1,2], Bristena Oprisanu[1], Emiliano De Cristofaro[1]

University College London[1], Hazy[2]

Problem description

Goal: Empirically evaluate and analyze the disparate effect training generative models with Differential Privacy (DP) guarantees has on the resulting synthetic data. More specifically, on underrepresented classes/subgroups (e.g., age, sex, and race) 1) size and 2) classification tasks.



Experimental settings

DP generative models:

1) PrivBayes (Laplace mechanism)
2) DP-WGAN (DP-SGD)
3) PATE-GAN (PATE)

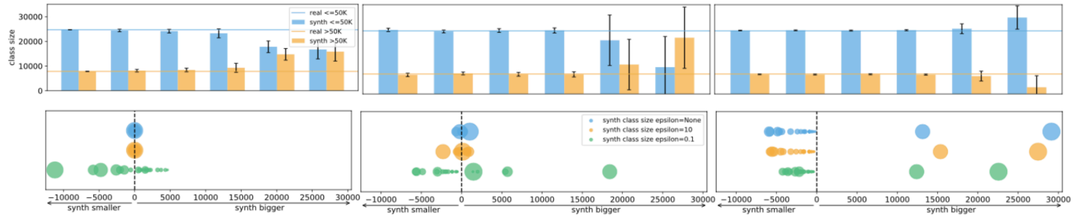
Data settings:

S1: Binary class size, precision, and recall
S2: Multi-class size, precision, and recall
S3: Single-attribute subgroup size, accuracy, and correlation
S4: Multi-attribute subgroup size, accuracy, and correlation

Various levels of subgroup imbalance and privacy budgets.

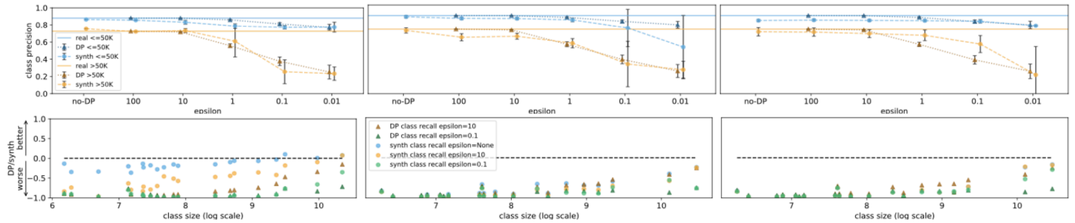
Main findings

Size: DP distorts size, yielding Robin Hood vs Matthew effects depending on the specific model and mechanisms. PrivBayes evens the imbalance, PATE-GAN increases it, while DP-WGAN has mixed results.



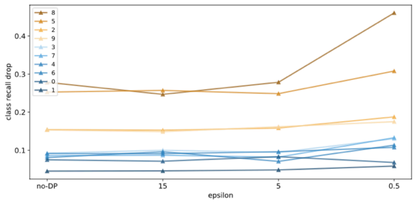
(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Classification 1: However, irrelevant of size in the synthetic data (or mechanism/model), classes/subgroups that were underrepresented in the real data suffer disproportionate drop in utility.

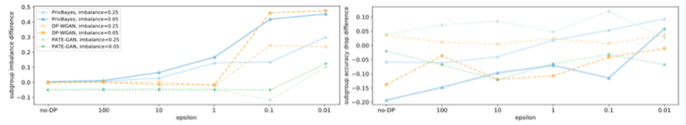


(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

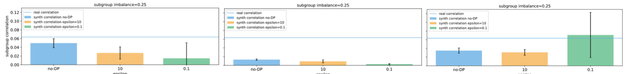
Classification 2: Unexpectedly, majority classes with similar characteristics to minority classes could also observe a more severe utility drop.



Imbalance: The magnitude of these effects increases when stronger privacy guarantees are imposed. Higher data imbalance levels further intensify them.



Correlation: While for PrivBayes and DP-WGAN imposing stronger privacy guarantees result in lower correlation between the subgroups and target columns, PATE-GAN could create undesirable artifacts in the synthetic data.



(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

So what?

Analyzing/training models on DP synthetic data could result in:

- treating different subpopulations unevenly
- unreliable/unfair conclusions with real societal costs

Full paper (+ further analysis and experiments):

georgi.ganev.16@ucl.ac.uk

International Conference on Machine Learning (ICML) 2022