Georgi Ganev*

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

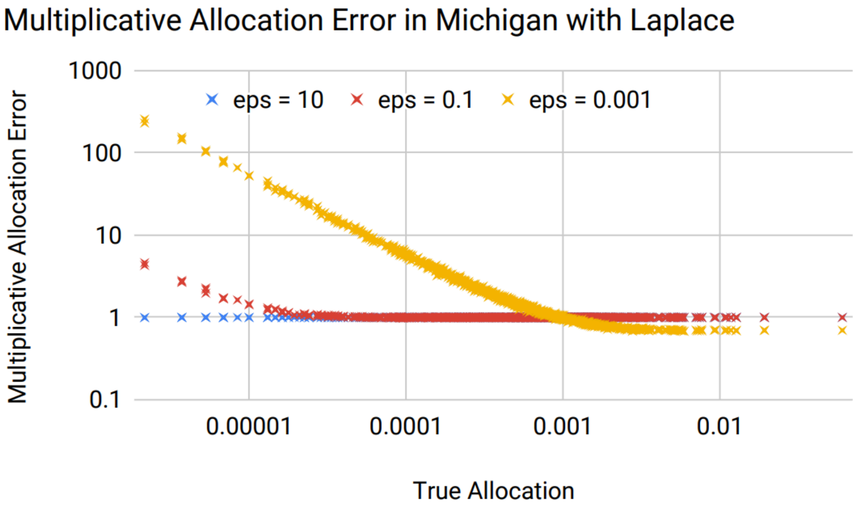
ICML 2022

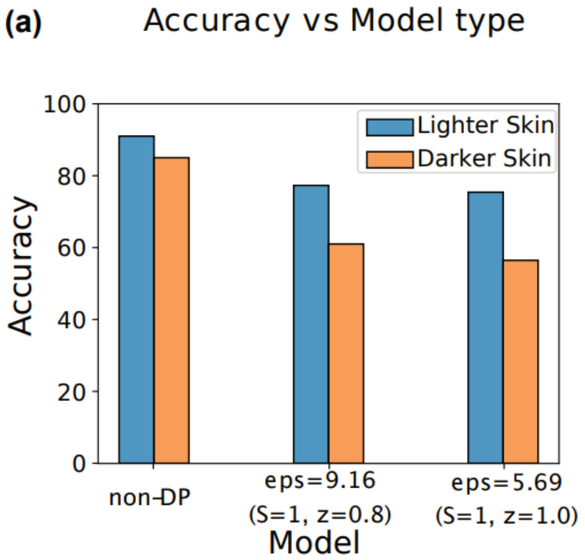
*Joint work with: Bristena Oprisanu and Emiliano De Cristofaro

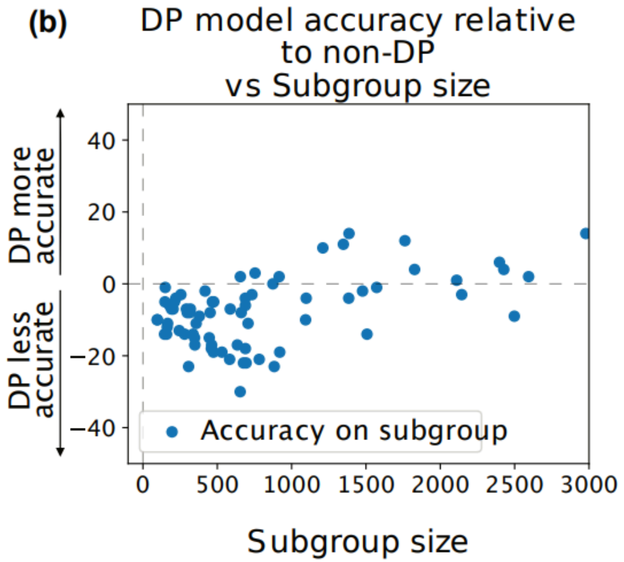
Motivation

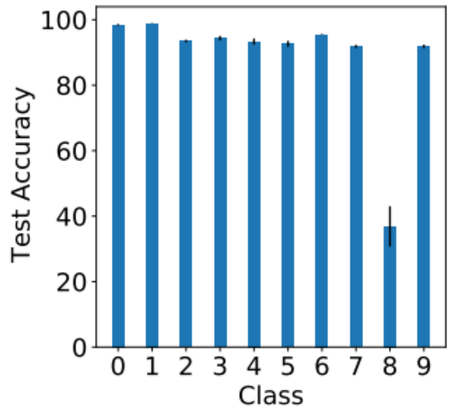
DP has a disparate effect[1] on:

- Statistics[2]
- Deep learning classifiers[3,4]

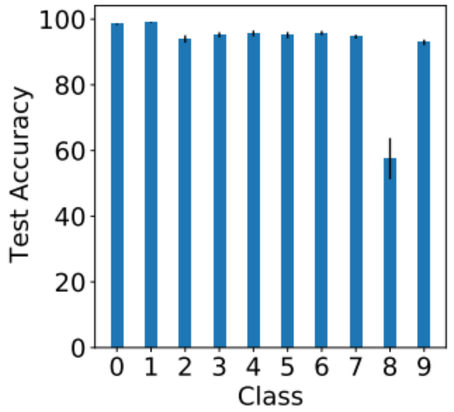








(b) DP-SGD for $\varepsilon = 5$

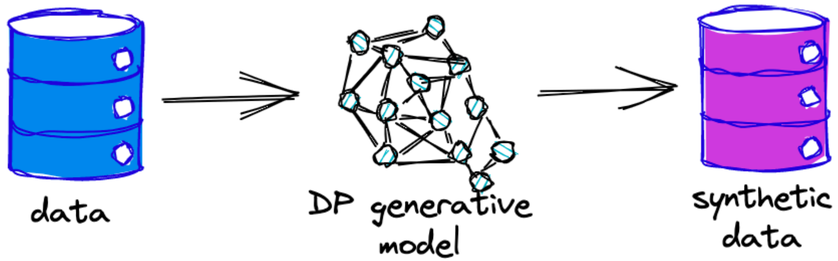


(e) PATE for $\varepsilon = 5$

[1]Fioretto et al., Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey
[2]Kuppam et al., Fair Decision Making using Privacy-Protected Data
[3]Bagdasaryan et al., Differential privacy has disparate impact on model accuracy
[4]Uniyal et al., DP-SGD vs PATE: Which Has Less Disparate Impact on Model Accuracy?

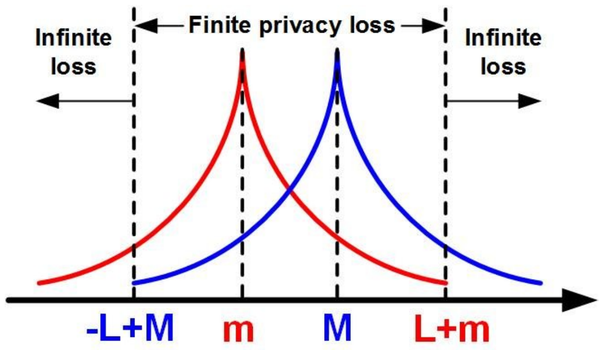
Goal

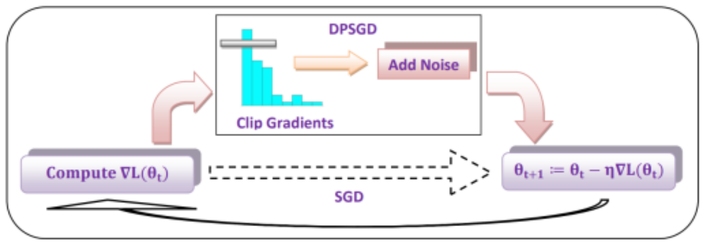
Empirically evaluate and analyze the disparate effect DP causes on generative models vis-a-vis *underrepresented* class/subgroup *size* and *accuracy.*

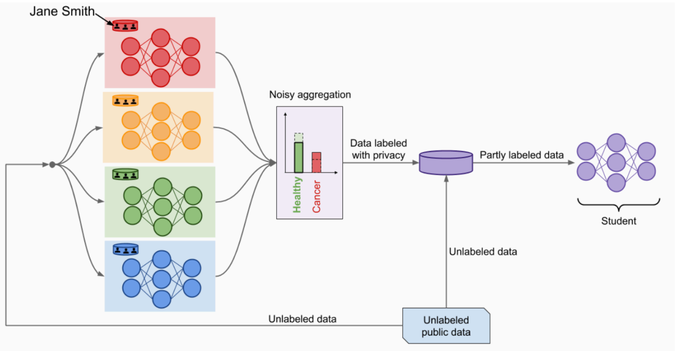


Three DP generative models:

1. PrivBayes[1] (Laplace)
2. DP-WGAN[2] (DP-SGD)
3. PATE-GAN[3] (PATE)







[1]Zhang et al., PrivBayes: Private Data Release via Bayesian Networks
[2]Alzantot et al., Differential Privacy Synthetic Data Generation using WGANs
[3]Jordon et al., PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees

Settings

Four data settings:

S1: Binary class size, precision, and recall

S2: Multi-class size, precision, and recall

S3: Single-attribute subgroup size, accuracy, and correlation

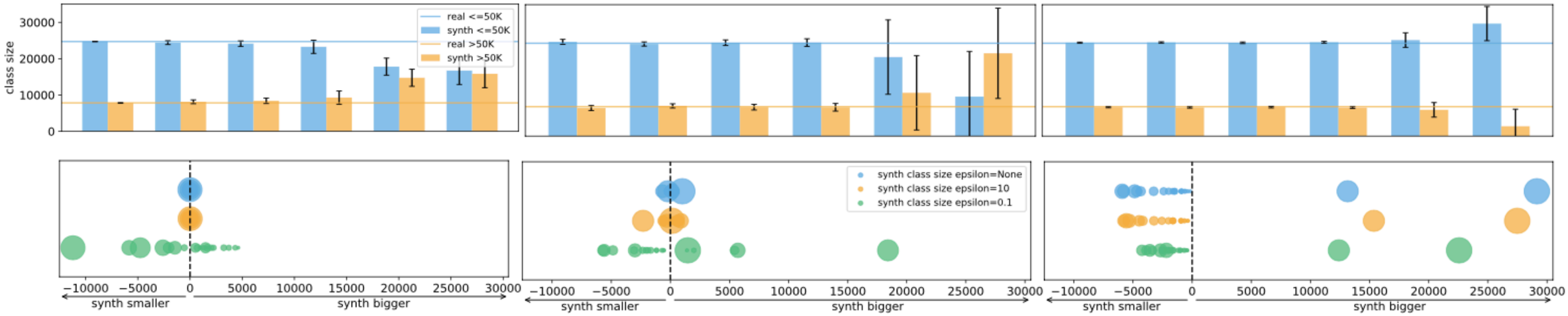
S4: Multi-attribute subgroup size, accuracy, and correlation

Various levels of subgroup imbalance and privacy budgets.

Take-Aways 1

RQ1: Do DP generative models generate data in similar classes and subgroups proportions to the real data?

— Not really. DP distorts the proportions, yielding Robin Hood vs Matthew effects depending on the DP generative model.

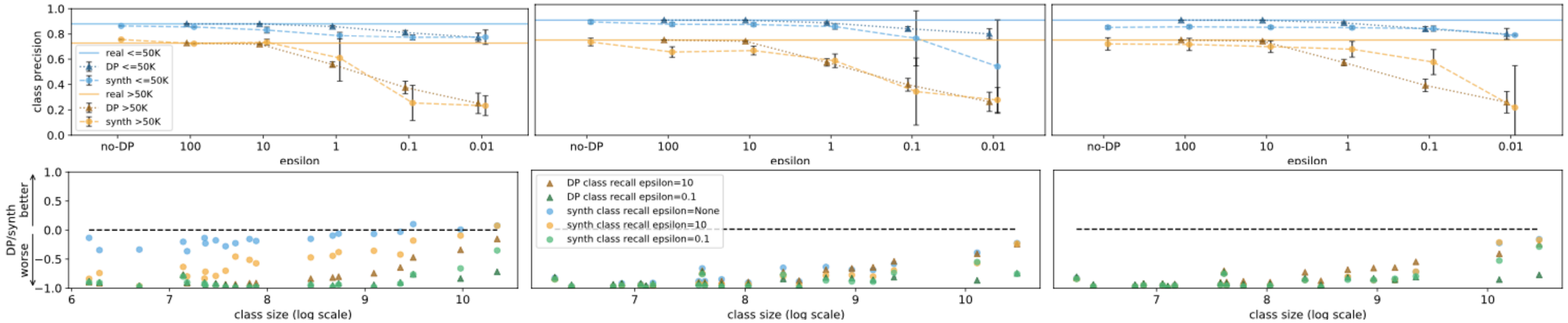


(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Take-Aways 2

RQ2: Does training a classifier on DP synthetic data lead to the same disparate impact on accuracy as training a DP classifier on the real data?

— Overall, yes. Smaller classes/subgroups suffer more similarly to DP classifiers. However, we do not see the rich get richer, the poor get poorer; everybody gets poorer. Incidentally, sometimes synthetic classifiers are better than DP classifiers.

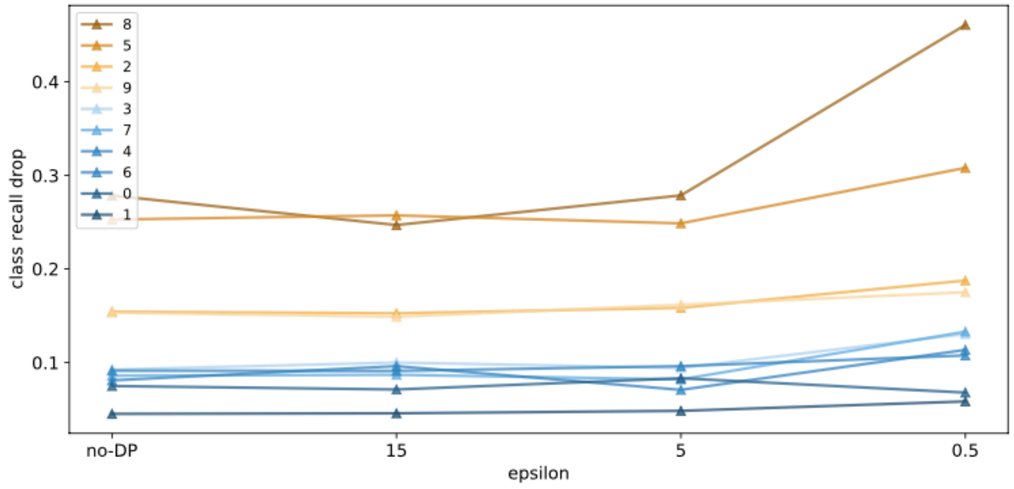


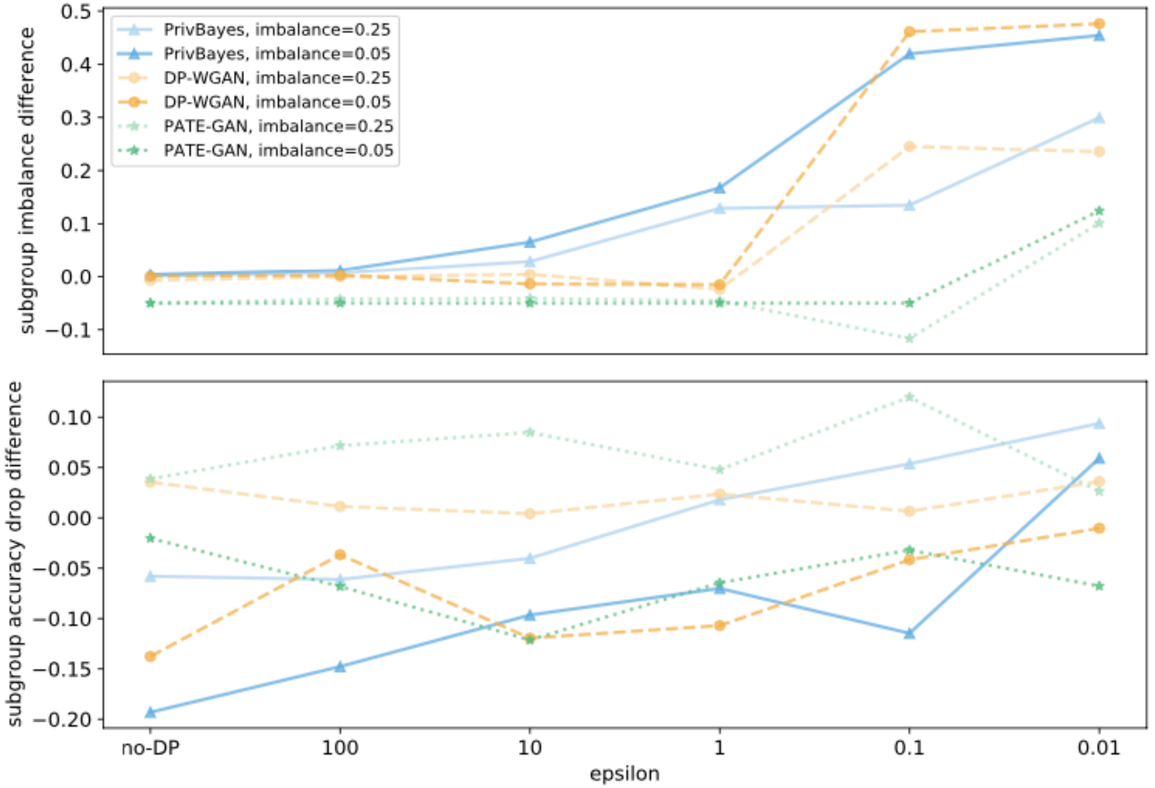
(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Take-Aways 3

RQ3: Do different DP mechanisms for DP synthetic data behave similarly under different privacy and data imbalance levels?

— No, different DP generative models behave differently. For example, PATE-GAN performs better than DP-WGAN, with some very specific exceptions, while PrivBayes is the only one that manages to maintain the data utility for the multi-class tabular data Purchases.





Thank you!