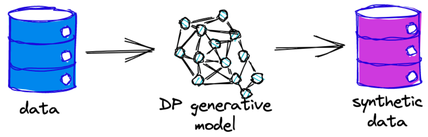
dpart: Differentially Private Autoregressive Tabular, a General Framework for Synthetic Data Generation

Sofiane Mahiou[1], Kai Xu[1,2], Georgi Ganev[1,3]

Hazy[1], University of Edinburgh[2], University College London[3]

Problem description

Goal: Propose and open source a general, flexible, and scalable framework *dpart* (*D*ifferetially *P*rivate *A*uto*R*egressive *T*abular) for synthetic data generation with Differentially Private (DP) guarantees.



Code: https://github.com/hazy/dpart,
a Python library with a MIT license.

Intended use and users

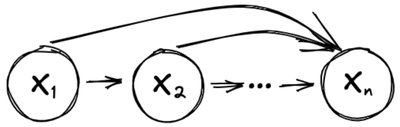
Use: Quickly and efficiently experiment to achieve competitive baseline results, not meant to get state-of-the-art results.

Users:

- beginners: making their first steps in privacy-preserving synthetic data
- domain knowledge: encoding their domain expertise into the data modelling
- privacy knowledge: using custom methods or privacy mechanism

Approach

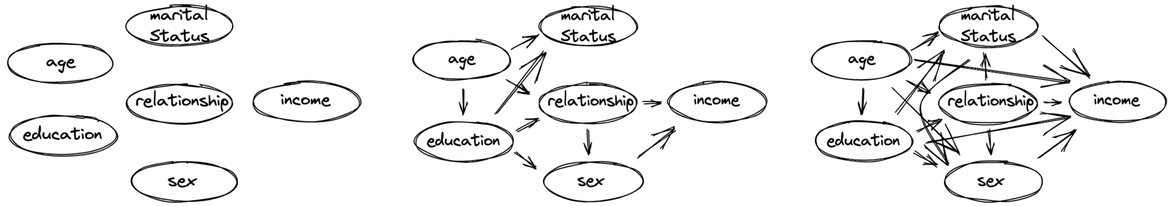
The overall training flow relies on autoregressive generative modelling and could be broken down to two steps:



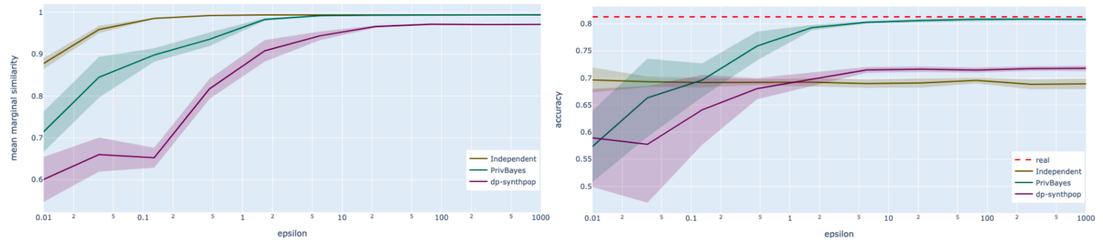
1) dependency: identifying/specifying a visit order or a prediction matrix that describes how the joint distribution is broken down to a series of lower-dimensional conditionals
2) sampler methods: given the series of conditionals, they are sequentially estimated by fitting predictive models

To generate synthetic data, the fitted sampler methods are used to generate one column at a time.

Specific *dpart* instances and comparison



(a) *Independent* (b) *PrivBayes* with *n_parents=2* (c) *dp-synthpop*



(a) Similarity (b) Accuracy

Other contributions

- as a specific use case of *dpart*, we modify and improve the speed of PrivBayes by 20x.
- as another use case, we propose a DP version of synthpop, which we name dp-synthpop.

So what?

With *dpart*, we aim to lower the entry barriers for new researchers and practitioners to facilitate experimentation and further development of synthetic data with DP guarantees.

We are looking forward to keep improving on this framework with the help of a growing community.

Full paper:

sofiane@hazy.com

8th ICML Workshop on Theory and Practice of Differential Privacy (TPDP), part of ICML 2022