Understanding the Impact of Data Domain Extraction on Synthetic Data Privacy

Georgi Ganev[1,2], Meenatchi Sundaram Muthu Selva Annamalai[1],

Sofiane Mahiou[2], Emiliano De Cristofaro[3]

UCL[1], SAS[2], UC Riverside[3]

Problem Description

Main Question: What is the impact of data domain extraction on the privacy of Differentially Private (DP) synthetic data generation?*

Experimental Framework:

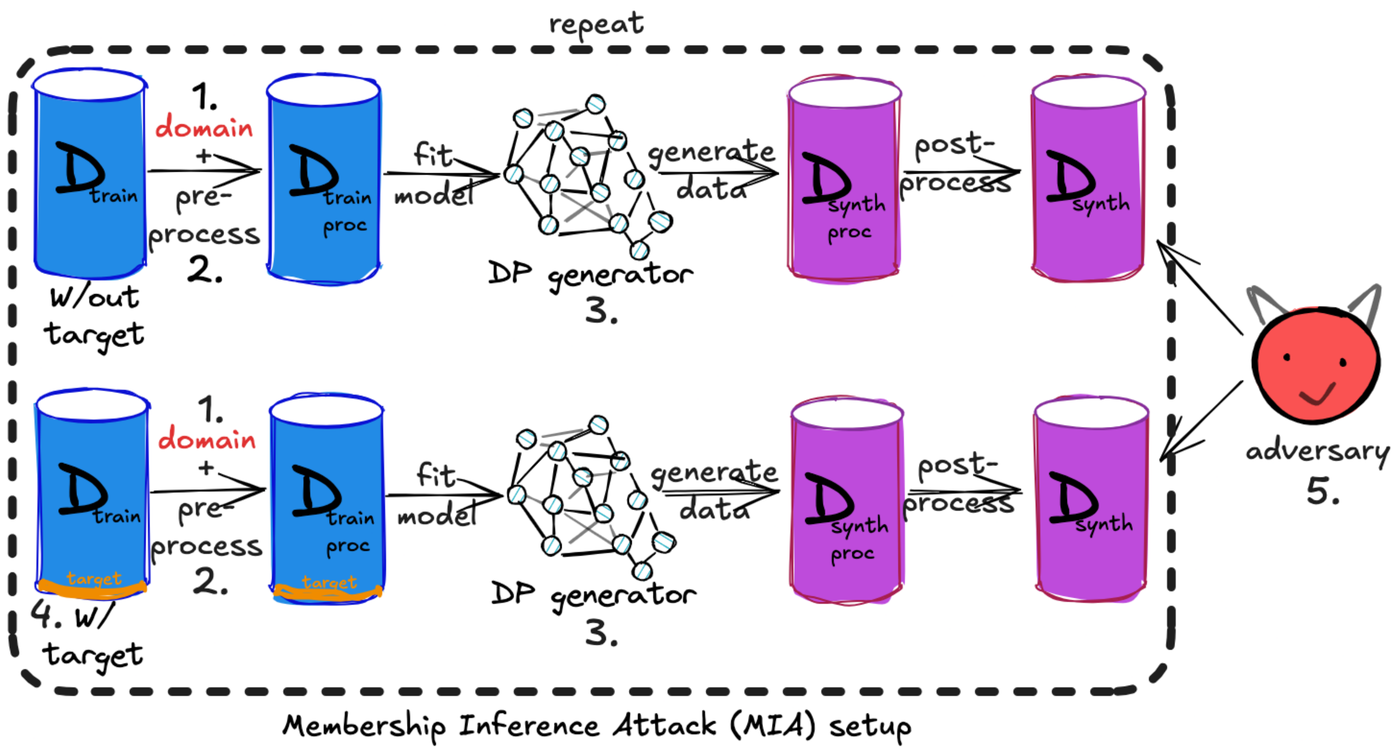
1. Three data domain extraction strategies:
   1. externally provided (ideally from public data)
   2. extracted directly from the input data (w/out DP)
   3. extracted from the input data w/ DP mechanisms

2. Four (DP) discretization strategies:
   1. uniform; 2. quantile; 3. k-means; 4. PrivTree

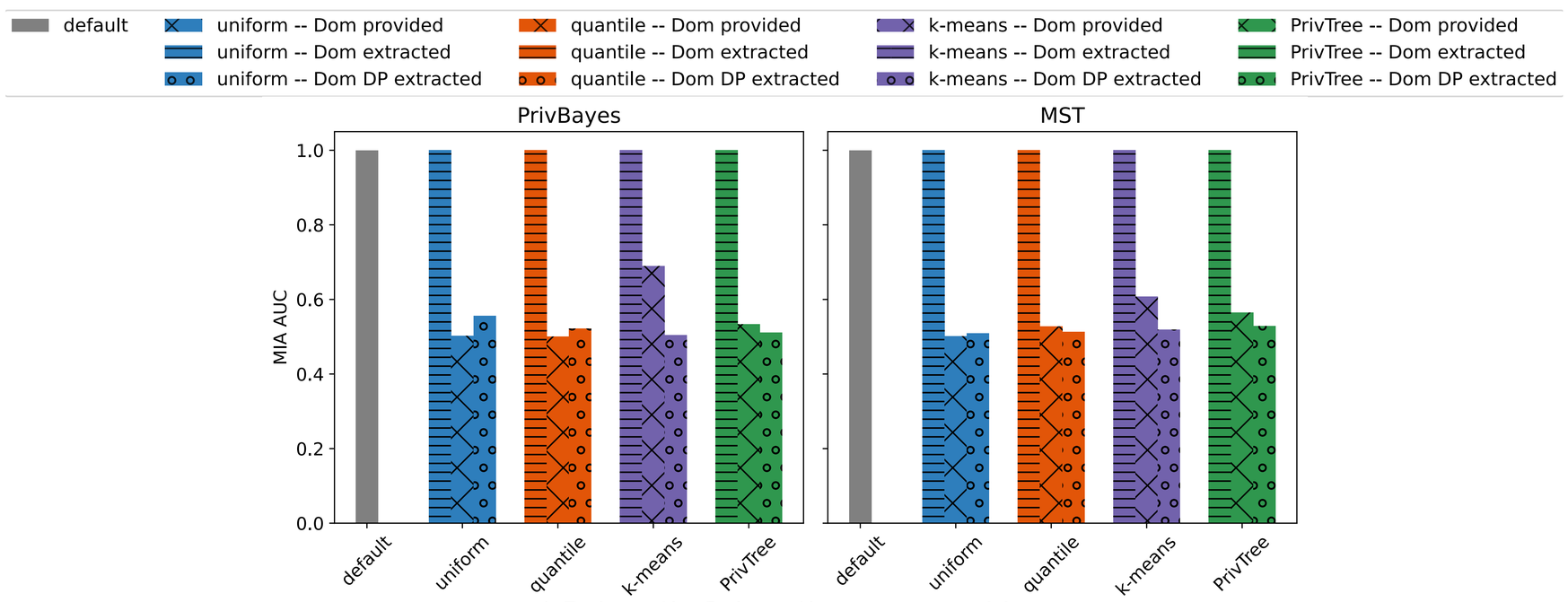
3. Two DP generative models:
   1. PrivBayes; 2. MST

4. Two types of target records:
   1. outside the domain of the remaining training data
   2. inside the domain of the remaining training data

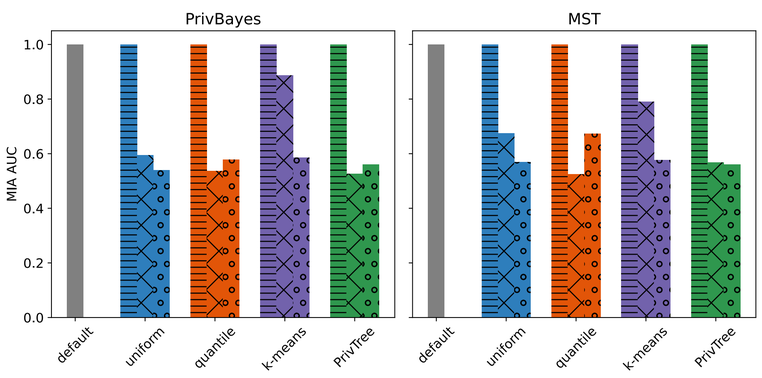
5. One Membership Inference Attack (MIA) – GroundHog



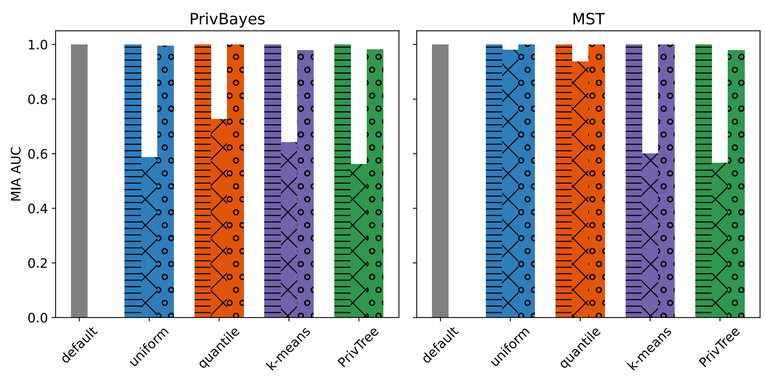
Experimental Evaluation



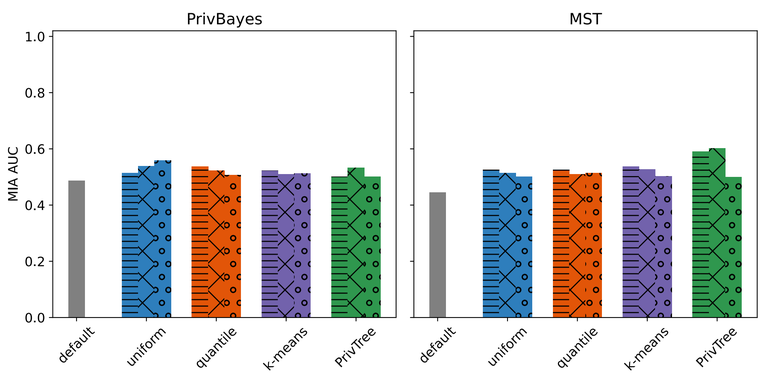
a) D (ϵ = 1), G (ϵ = 1), target *outside*



b) D (ϵ = 100), G (ϵ = 100), target *outside*



c) D (ϵ = 1,000), G (ϵ = 1,000), target *outside*



d) D (ϵ = 1), G (ϵ = 1), target *inside*

Figure 1: Privacy leakage with *provided* and *extracted* domain (w/ and w/o DP) for the four DP discretizers (D) and two DP generative models (G) on a target record outside/*inside* the domain of the remaining data, Wine dataset

Main Take-Aways

1. Strategy 2 (extracting the domain directly from the input data) breaks the end-to-end DP guarantees of generators and exposes outliers to MIAs

2. Both Strategy 1 and 3 (provided data domain and extracting it with DP, up to ϵ = 100) successfully protect outliers from specific MIAs

3. Adopting Strategy 3 could address many previously identified DP vulnerabilities in open-source implementations and libraries

4. The GroundHog MIA may be more effective at detecting issues with data domain extraction than with vulnerabilities of the generators

So What?

1. Developers and practitioners should be more conscious of the implementation details surrounding DP synthetic tabular data generation, or these pipelines could be left exposed to serious privacy vulnerabilities

2. We highlights the need for further analysis of MIAs against DP generative models

Full paper:

* In "The Importance of Being Discrete: Measuring the Impact of Discretization in End-to-End Differentially Private Synthetic Data." *arXiv:2504.06923 (2025)*, we examine the broader question of discretization in end-to-end DP generative models, primarily focusing on utility. This paper closely relates to RQ4.

georgi.ganev.16@ucl.ac.uk

1st Synthetic Data × Data Access Problem workshop (SynthData), part of ICLR 2025