dpmm: Differentially Private Marginal Models, a Library for Synthetic Tabular Data Generation

Sofiane Mahiou[1], Amir Dizche[1], Reza Nazari[1], Xinmin Wu[1],

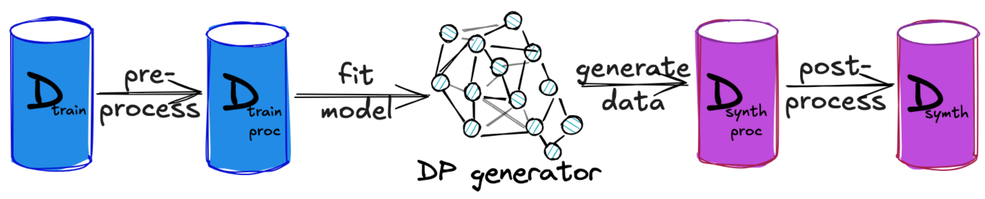
Ralph Abbey[1], Jorge Silva[1], Georgi Ganev[1,2]

SAS[1], UCL[2]

Problem Description

Goal: Propose a lightweight library for DP synthetic data generation, containing three state-of-the-art marginal models (PrivBayes, MST, and AIM) with: 1) superior utility, 2) richer functionality, 3) end-to-end DP guarantees.

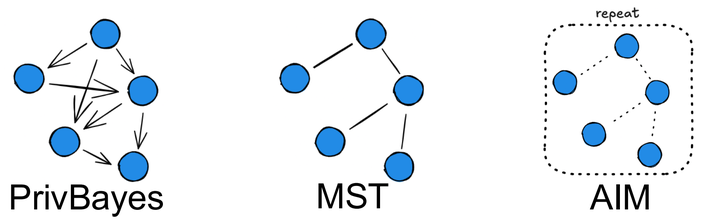
Code: https://github.com/sassoftware/dpmm
(a Python library; pip installable; Apache-2.0 license)



Overview of *dpmm*

Marginal Models: *dpmm* contains three DP generative marginal models relying on the select-measure-generate paradigm and Private-PGM:

- PrivBayes: builds an optimal Bayesian network
- MST: builds a maximum spanning tree
- AIM: adaptively and iteratively optimizes marginals



DP Guarantees: *dpmm* adopts best practices from various scientific papers and DP libraries to provide end-to-end DP guarantees:

- Data Domain: either provided as input or extracted with DP
- Data Preprocessing: uses PrivTree, DP tree-based discretization method
- Floating-Point Precision: uses OpenDP for Gaussian mechanism sampling

Library	Marginal Model			DP Guarantee		
	PrivBayes [50]	MST [32]	AIM [33]	Data Domain	Data Preprocessing	Floating-Point Precision
***dpmm* (ours)**	✓	✓	✓	✓	✓	✓
private-pgm [29]	✘	✓	✓	∼[1]	✘	✘
OpenDP [41]	✘	✓	✓	∼[2]	∼[3]	✓
synthcity [43]	✓	✘	✓	✘	✘	✘

Functionality: *dpmm* offers rich functionality across all models:

- Mixed Data Support: supports both numerical and categorical data
- Conditional Generation: satisfies any conditions at generation time
- Public Pretraining: models can be pretrained on public data
- Structural Zeros: can be configured or suppressed
- Max Model Size: the trained model size can be controlled
- Serialization: trained models can be saved and reloaded

Library	Mixed Data Support	Conditional Generation	Public Data Pretraining	Structural Zeros	Max Model Size	Serialization
***dpmm* (ours)**	✓	✓	✓	✓	✓	✓
private-pgm [29]	✘	✘	✘	MST	AIM	✘
OpenDP [41]	✓	✘	✘	MST	AIM	✘
synthcity [43]	✓	✘	✘	✘	AIM	✓

Experimental Evaluation

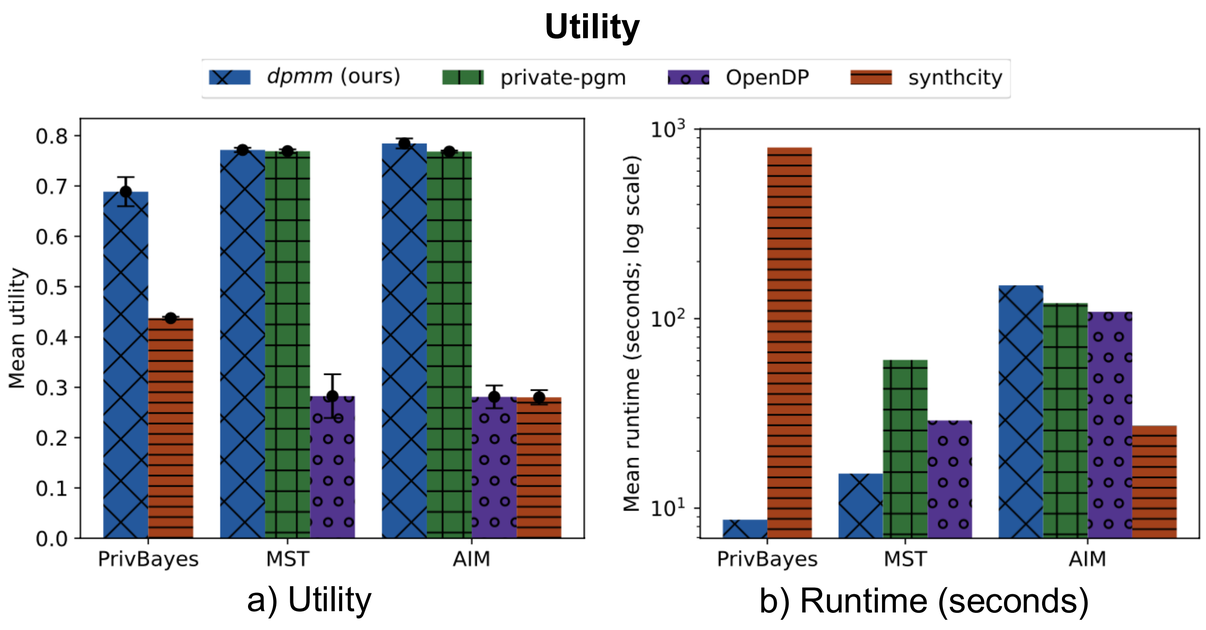


Figure 1: Comparison between *dpmm* and other libraries (ε = 1, δ = 10−5).

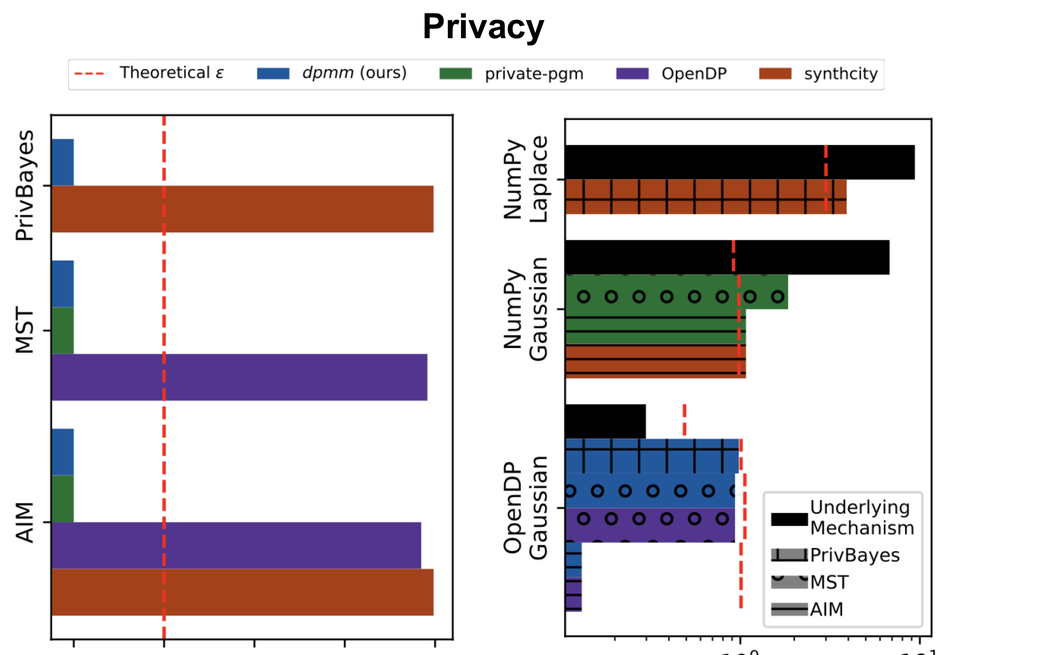


Figure 3: DP auditing of *dpmm* and other libraries.

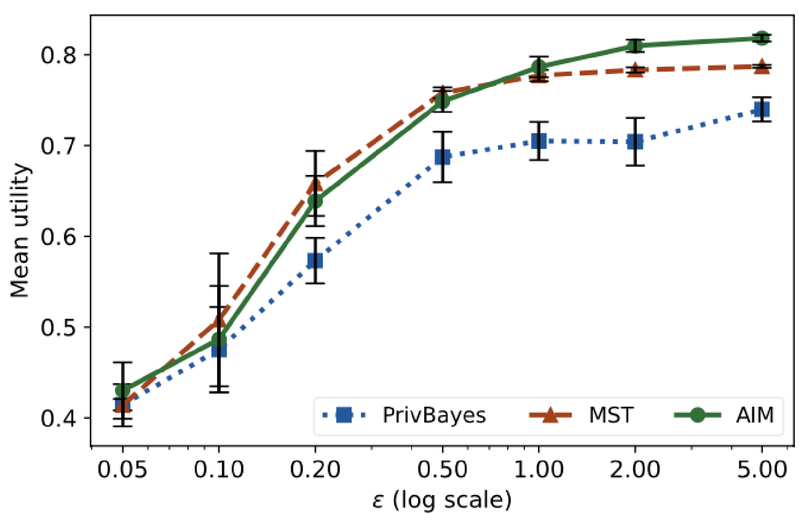


Figure 2: Utility-privacy tradeoffs of *dpmm*.

Main Take-Aways

1. We implement and open source *dpmm*, a lightweight library for end-to-end DP synthetic data generation, containing three popular marginal models (PrivBayes, MST, and AIM) with rich functionality.

2. *dpmm* achieves higher utility than previous implementations -- on average 1.5% higher than private-pgm and 147% than OpenDP/synthcity

3. *dpmm* contains state-of-the-art DP auditing procedures and effectively addresses known DP-related vulnerabilities.

sofiane.mahiou@sas.com

11th Theory and Practice of Differential Privacy Workshop (TPDP 2025)