RPI

Measuring Privacy Risks and Tradeoffs in Financial Synthetic Data Generation

Michael Zuo, Inwon Kang, Stacy Patterson, Oshani Seneviratne
Rensselaer Polytechnic Institute, Department of Computer Science

Motivations

- Sensitive data: privacy concerns, regulatory constraints (e.g. GDPR/ CCPA) make direct sharing and reuse of individual-level data risky.
- Even synthetic data can reveal information about real data used to create it
- Standard privacy mechanisms (differential privacy) add noise to obscure individual records
- Noise introduced to data generally degrades utility
- **How well do state-of-the-art generators simultaneously provide quality/utility and privacy?**

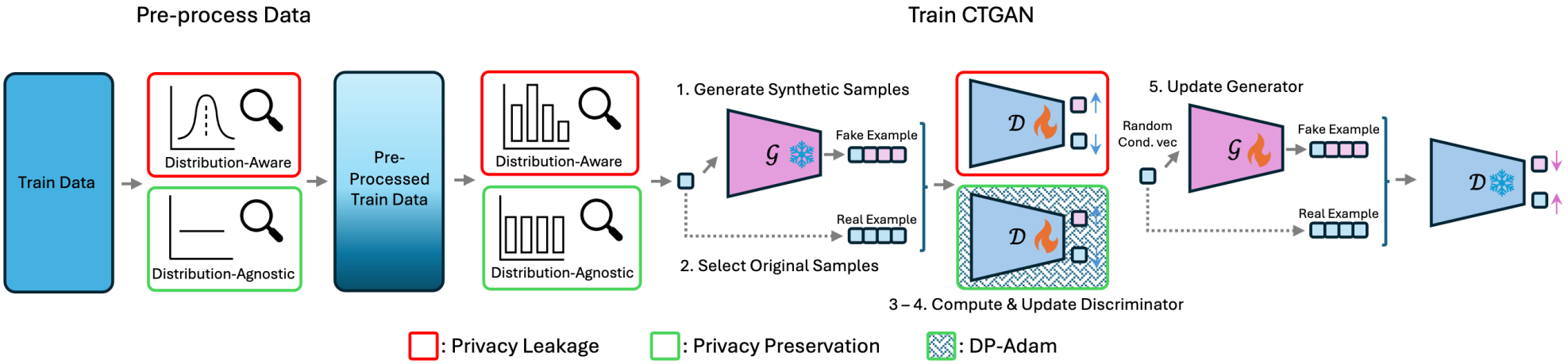
Synthetic Data Generators

- Generator models selected from four classes:
  - **Gaussian Copula** [Patki 2016]: classical statistical model
  - **TabDiff** [Shi 2025]: diffusion model for mixed tables
  - **CTGAN** [Xu 2019]: conditional GAN on tabular data
  - **TVAE** [Xu 2019]: tabular variational autoencoder
- None of these models give formal privacy guarantees

[24] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The Synthetic Data Vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 399–410.
[28] Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. 2025. TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation. In *The Thirteenth International Conference on Learning Representations*.
[36] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling Tabular Data Using Conditional GAN. In *Advances in Neural Information Processing Systems*, Vol. 32.

DP Synthetic Data Generation

- We introduce differentially private implementations for CTGAN and TVAE
  - Existing implementations add DP to training step without providing end-to-end protection



Patch preprocessing leak: replace mode-specific normalization with uniform binning

Patch conditioning leak: replace selection of real samples computing condition vectors from uniform samples

Train with DP-Adam

Evaluation Metrics

- Datasets: 6 financial classification datasets from the TabArena suite

- **Quality**:
  - SDMetrics column shapes and pair trends scores
  - Minority class %
- **Utility**:
  - XGBoost downstream task model trained on synthetic data
  - Evaluated on balanced accuracy over real test split
- **Privacy**:
  - SDMetrics distance-to-closest-record (DCR) baseline and overfitting protection
  - Shadow model membership inference attack [Stadler 2020]

[30] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic Data – Anonymisation Groundhog Day. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 1451–1468.

Utility—Downstream Task Accuracy

- Even non-private generators lose some utility
- DP-CTGAN performs well but degrades with added privacy

- For some datasets, classifier learns better on synthetic than real data

Dataset	Original	Non-DP				DP-CTGAN				DP-TVAE			
		Gauss.	TabDiff	CTGAN	TVAE	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = \infty$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = \infty$
AD	0.818	0.516	0.684	0.718	0.769	0.681	0.709	0.734	0.784	0.508	0.5	0.5	0.753
BC	0.721	0.539	0.707	0.670	0.650	0.523	0.556	0.562	0.648	0.499	—	—	0.619
BM	0.596	0.506	0.576	0.586	0.603	0.528	0.574	0.579	0.580	0.499	0.5	0.5	0.500
CC	0.658	0.546	0.642	0.609	0.635	0.554	0.614	0.613	0.673	0.499	0.5	0.5	0.635
CR	0.686	0.533	0.614	0.507	0.552	0.501	0.494	0.505	0.555	0.518	0.515	0.524	0.553
GM	0.594	0.502	0.583	0.628	0.666	0.541	0.633	0.591	0.583	0.509	0.5	0.593	0.543

Privacy—Membership Inference Attack

- Identification rate of dataset generated with a canary sample in training set
- 0.5 success rate is the equivalent of random guessing
- Membership inference attacks find limited signal on all synthetic generators

Dataset	Non-DP				DP-CTGAN				DP-TVAE			
	Gauss.	TabDiff	CTGAN	TVAE	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = \infty$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = \infty$
AD	0.491	0.507	0.505	0.496	0.513	0.497	0.512	0.475	0.509	0.484	0.502	0.508
BC	0.536	0.49	0.426	0.501	0.504	0.499	0.499	0.514	0.474	0.495	0.529	0.483
BM	0.523	0.482	0.505	0.515	0.491	0.487	0.49	0.488	0.51	0.519	0.494	0.507
CC	0.498	0.549	0.412	0.489	0.494	0.499	0.471	0.494	0.5	0.524	0.518	0.491
CR	0.530	0.502	0.497	0.488	0.517	0.495	0.516	0.485	0.518	0.476	0.513	0.519
GM	0.48	0.528	0.453	0.522	0.532	0.516	0.477	0.474	0.465	0.479	0.494	0.478

Key Takeaways

- Private data processing implementations require end-to-end accounting
  - We identify privacy leak sites in preprocessing, sampling, and training
  - We redesign CTGAN/TVAE around safe transformations and DP-Adam

- CTGAN and TVAE produce high-utility synthetic data
  - At moderate privacy levels (ϵ = 5, 10), DP-CTGAN maintains most data utility

- Membership inference attacks do not give tight measure of privacy
  - Need for new methods for empirical privacy validation

Evaluation code: https://github.com/brains-group/ppsdg-supplementary
Email: zuom@rpi.edu