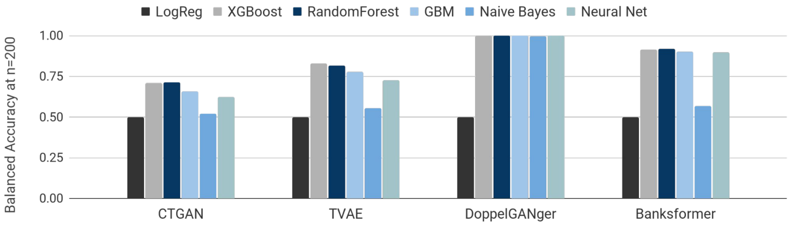
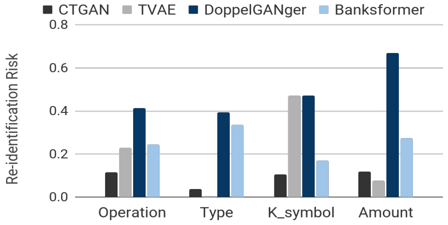
Evaluating the Privacy-Preserving Capabilities of Generated Synthetic Data

Julia L. Wang | Supervisor: Prof. Yuri Lawryshyn

Generative Models: Banksformer, CTGAN, TVAE, DoppelGANger

Dataset: Czech 1M transactions dataset with account, type, opteration, amount, k_symbol





Membership Inference Attacks: determine if an individual's data was used in the training set of a model by simulating labels an attacker may have.

Re-identification Risk: likelihood of tracing an individual back to their original data from the generated data.

Privacy Metrics: Measured k-anonymity, l-diversity, and t-closeness. Synthetic datasets maintained lower k-anonymity and l-diversity. Banksformer showed best overall performance.

Attribute Disclosure Risk: potential for sensitive info about individuals to be inferred from a dataset. Synthetic datasets exhibited lower risk overall.

Column-wise CTGAN, TVAE best distributions
Time-series Banksformer and DoppelGANger

RBC CMTE UNIVERSITY OF TORONTO

Centre for Management of Technology & Entrepreneurship

Balancing Utility and Privacy

Evaluating the Privacy-Preserving Capabilities of Generated Synthetic Data

Julia L. Wang
Supervisor: Prof. Yuri Lawryshyn
RBC Sponsor: Lucy Liu

Centre for Management of Technology & Entrepreneurship

Objectives

Evaluate the privacy-preservation of state-of-the-art generative models in synthesizing financial datasets

Investigate the trade-offs between data utility and privacy across different synthetic data generation techniques

Generative Models

CTGAN

Generating tabular data that mimic real distributions, addressing challenges of imbalanced and sparse data

DoppelGANger

Dual mechanism with MLPs and RNNs, generating metadata and time-series data while preventing mode collapse

TVAE

Triplet-based Variational Autoencoder, enhances data representation in latent space to capture complex relationships

Banksformer

A transformer-based approach for generating sequence data, focusing on temporal dynamics and patterns

Dataset: Czech transactions

Over 1M Transactions from 4500 accounts

Timestamps from Jan 1, 1993 to December 31, 1998

Features: account, type, operation, amount, k_symbol

Data Utility

Column-wise statistical analysis: CTGAN and TVAE had the most similar distributions to the real data for each individual column.

Time series statistical analysis: DoppelGANger and Banksformer performed better at replicating temporal relationships. Banksformer was the best model overall.

Privacy Metrics

01

K-Anonymity

each individual is indistinguishable from at least **k–1** others for every identifiable attribute set

02

L-Diversity

for every group of individuals, there are at least **l** diverse values for each sensitive attribute

03

T-Closeness

distribution of a sensitive attribute is no further than **t** from the dist of the attribute in the entire dataset

Results

- Real data maintained high k-anonymity and l-diversity, synthetic data had low
- Banksformer showed closer distribution proximity compared to others

Re-identification Risk

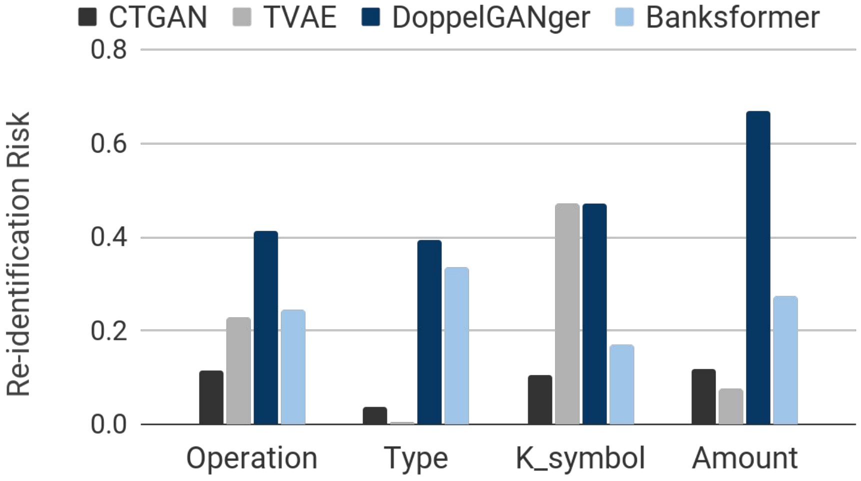
The likelihood that an individual's data in a synthetic dataset can be traced back to that individual in the original dataset

Frequency Dists

Compare the frequency distributions of categorical variables between the real and synthetic

Kolmogorov-Smirnov (KS) Test

Similarity between the distribution of numerical variables, quantifying the max discrepancy between their cumulative distribution functions



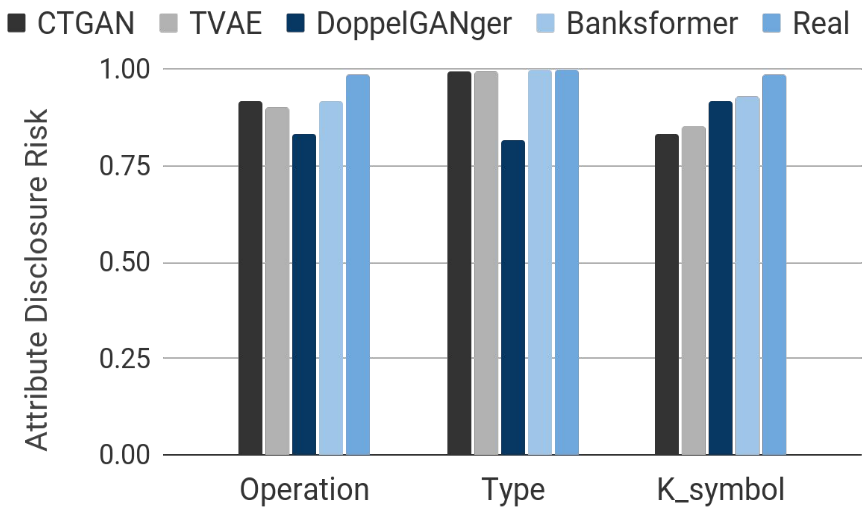
Attribute Disclosure Risk

Potential for sensitive information about individuals to be inferred from a dataset

- Random Forest classifiers to predict sensitive attributes based on other data attributes

Results

Models trained on synthetic data typically yielded lower prediction accuracies, indicating that the synthetic data doesn’t retain the same attribute relationships



Membership Inference Attacks

Combine Data

Real and synthetic combined and labelled: creates a dataset that reflects potential knowledge an attacker might possess

Attack Models

Binary classifiers trained to differentiate between real and fake

- logistic regression, Naive Bayes, Random Forest, XGBOOST, GBM, and a feedforward neural network

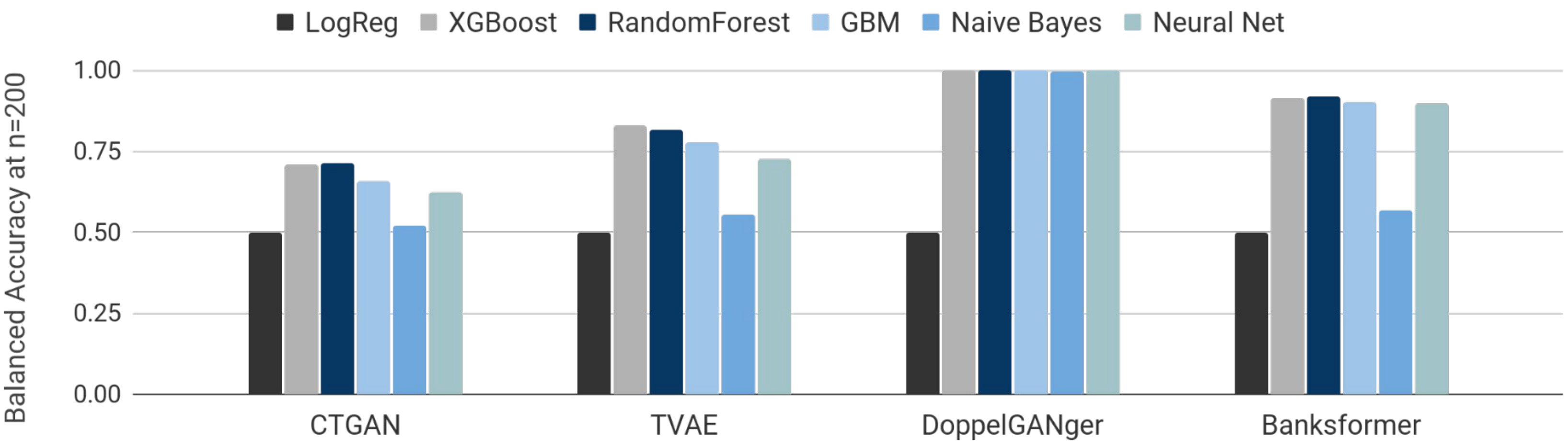
Performance Metrics

Balanced accuracy, precision, recall, and F1 scores using 5-fold cross-validation

Different Splits

Varying n, the number of accounts to which an attacker has access to labels = 1, 10, 25, 50, 100, 250, 500, 750, 1000

Membership Inference Attack Results



- As n increases, accuracies converged (around n=200)
- Logistic Regression overfit: predits 1 class → balanced accuracy of around 50%
- Banksformer and DoppelGANger: high accuracies → significant privacy concerns
- CTGAN and TVAE: better resistance to attacks, with accuracies closer to the ideal 50%

Conclusions

Generative Model Effectiveness

CTGAN and TVAE showed better results in protecting privacy
Banksformer and DoppelGANger exhibited vulnerabilities that could lead to privacy breaches

Data Utility and Privacy Balance

Further development and refinement of generative models are required to enhance their ability to produce useful yet non-revealing datasets

Further Research

Integration of differential privacy or encryption techniques directly into the data generation process and the exploration of the trade-off between utility and privacy

Thanks!

Special thanks to Prof. Lawryshyn, Lucy Liu, Peter Miasnikof, and RBC!