Differentially Private Synthetic Data Generation for Mobile Money Fraud Detection

PhD Candidate

Denish Azamuke

Supervisors

Prof. Engineer Bainomugisha, PhD Dr. Marriette Katarahweire, PhD

Department of Computer Science, Makerere University

November 14, 2025

Abstract

We live in an era where routine transactions ranging from paying domestic bills to buying groceries are carried out using mobile financial services. However, the rapid growth and uptake of these services has led to amplified security and privacy risks, including SIM swap attacks, identity fraud, data theft, refund fraud, and unauthorized fees. Advances in machine learning (ML) show potential for detecting financial fraud in mobile money transactions, yet this requires access to large volumes of transaction data. Research on mobile money fraud has been hindered by data sensitivity and privacy concerns that restrict access to such datasets. In addition, real mobile money datasets are class-imbalanced, with far fewer frauds than legitimate transactions, biasing ML models against the minority class. This thesis presents a differentially private synthetic data generation approach for mobile money transaction datasets to support financial modeling and fraud detection.

Developing a synthetic data generation model for tabular data that preserves the intricate, high-order correlations that drive fraud while guaranteeing differential privacy remains notoriously difficult. This challenge stems from calibration fragility in high-dimensional spaces and a parameter search space that expands exponentially, requiring thousands of stochastic runs for model convergence. Existing synthetic data generation methods do not accurately model sparse, event-driven features, while simpler resampling techniques risk leaking private information and struggle to capture evolving fraud tactics in real mobile money ecosystems.

This thesis develops synthetic data generation techniques to investigate these limitations. This study introduces a multi-agent-based simulation model *MoMTSim*, which simulates interactions among clients, merchants, and banks. MoMTSim is calibrated using transaction aggregates derived from a real mobile money transaction dataset. Its fidelity is assessed using the sum of squared errors, Kolmogorov–Smirnov tests, and visual diagnostics such as Bland–Altman plots and kernel density estimates. The results show a close resemblance to real data, with a total error of 2.0010 at the 100 000-client benchmark. We present *MoMTSimDP*, a differentially private extension of MoMTSim that applies the Gaussian mechanism and satisfies a $(1.0, 10^{-6})$ privacy guarantee. MoMTSimDP maintains high fidelity, achieving a comparable total error of 2.0070 at 100 000 clients.

The machine learning utility of the synthetic datasets is assessed using logistic regression, decision tree, random forest, and XGBoost classifiers. Performance evaluation shows that random forest and XGBoost remain resilient under differential privacy, achieving an AUC of at least 0.79. This study embodies the simulation model within *MoMTLab*, an analytics platform built using graph-based data structures that enables visual analysis of mobile money transaction patterns without requiring extensive data science expertise.