Scenario-based Synthetic Dataset Generation for Mobile Money Transactions

Denish Azamuke, Marriette Katarahweire, and Engineer Bainomugisha
denishazamuke@gmail.com,kmarriette@gmail.com,baino@mak.ac.ug
Department of Computer Science, Makerere University
Kampala, Uganda

ABSTRACT

There is limited availability of mobile money transaction datasets from Sub-Saharan Africa for research because transaction data records are sensitive in nature and therefore raise privacy concerns. This has in turn hindered the potential to study fraudulent patterns in mobile money transactions so as to propose realistic mitigation measures based on Machine Learning Approaches to the prevailing financial fraud challenges in the region. This research presents mobile money scenarios that should be considered in order to implement a simulator that can harness synthetic datasets for mobile money transactions from Sub-Saharan Africa so as to carry out fraud detection research. These scenarios include the definition of a mobile money ecosystem with processes used by actors such as mobile money agents, clients, merchants and banks to interact with each other in mobile money operations. There is also a need for a real mobile money dataset to extract statistical information and diverse fraudulent behaviours of actors and fraud examples in mobile money markets. This research uses the design considerations to examine process-driven techniques such as numerical simulation, agent-based modeling, and data-driven techniques such as neural networks that can be leveraged to generate synthetic datasets for mobile money transactions. Common data generation toolkits like PaySim, AMLSim, RetSim and ABIDES that are based on these techniques have been examined. The design considerations are used to design a realistic model known as *MoMTSim* based on real mobile money processes and agent-based modeling techniques that can be implemented to generate synthetic transaction datasets for mobile money with fraud instances. This will facilitate fraud detection research. The synthetic datasets eliminate data privacy risks, are easy and faster to obtain, and are cheap to experiment with. With the proposed model, different research groups can move to the implementation stage to realise a model for synthetic data generation for mobile money transactions from the Sub-Saharan region.

FAMECSE '22, June 7–8, 2022, Cairo-Kampala, Egypt

ACM ISBN 978-1-4503-9663-9/22/06…$15.00
https://doi.org/10.1145/3531056.3542774

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Model verification and validation**; • **Applied computing** → **Electronic funds transfer**.

KEYWORDS

Mobile money, datasets, agent-based modeling, fraud detection, synthetic data

ACM Reference Format:
Denish Azamuke, Marriette Katarahweire, and Engineer Bainomugisha. 2022. Scenario-based Synthetic Dataset Generation for Mobile Money Transactions. In *Federated Africa and Middle East Conference on Software Engineering (FAMECSE '22), June 7–8, 2022, Cairo-Kampala, Egypt*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3531056.3542774

1 INTRODUCTION

Mobile money systems enable access to financial services through the use of feature or smart phones without having an account at a bank [12]. Through the mobile phone, users are able to send or receive money and pay for goods and services such as domestic bills and in-store purchases. Most recently, during the COVID-19 pandemic, mobile money systems have been leveraged to disburse funds to the vulnerable population by governments and non-governmental organisations [11]. The Global System for Mobile Communications Association (GSMA) defines mobile money as all financial services that can be accessed using a phone. This definition thus includes mobile banking which is concerned with individuals performing transactions between bank accounts and mobile money accounts [39].

Mobile money platforms are spurring financial inclusion in Sub-Saharan Africa (SSA) with 548 million registered mobile money accounts in SSA [19], US $490 billion transaction value (a growth of 23%) and in East African countries such as Uganda, 43% of the population have mobile money accounts compared to 11% with bank accounts (BoU). In Kenya, 72% of the population have mobile money accounts compared to 29% with bank accounts.

Unfortunately, mobile money systems are vulnerable to financial fraud and laundering targeting end-users, mobile money agents, and mobile network operator systems, that if not appropriately dealt with are likely to discourage usage among the population, potentially reversing years of progress on achieving financial inclusion. Reported millions of US dollars are lost in the fraud [7].

Mobile money financial fraud types range from simple to sophisticated including split deposits and withdrawal of funds carried out by mobile money agents, parallel money transfers on the network, money laundering on mobile money financial service platform by

agents and the registration of non-existent customers by mobile money system administrators during customer acquisition stage [3].

Mobile financial services generate a large volume of data that is complex and varied. The data generated is mostly stored within the companies for business needs. Access to datasets of mobile money transactions is often restricted or not possible at all due to privacy concerns, regulations and business reasons. This could explain the limited research activities, for example on machine learning approaches and solutions for financial inclusion, fraud detection, and money laundering.

Given the difficulties in accessing real datasets on mobile financial transactions, the value of synthetic datasets can not be overstated. Synthetic data is artificially manufactured information that is based on real-world processes and scenarios but not on real-world actors/events [4]. Sectors such as financial services and healthcare could greatly benefit from synthetic data generation to aid new research. The key motivation for synthesizing financial data includes but is not limited to restrictions on the usage of internal data imposed by companies which might slow down technical teams yet those teams can rely on synthetic data to continue with research and development. Furthermore, the lack of historical data to study certain events, tackling class imbalance for example where traditional machine learning and anomaly detection approaches fail to detect fraud, a case where one needs to train advanced machine learning models and the need for data sharing among research communities in order to develop solutions for technical problems faced by the financial institutions, push for the need for synthetic datasets [26].

Synthetic data generated should have properties of the real data while taking into account the privacy of the parties involved. Synthetic data can also be referred to as artificially manufactured data that learn the properties and attributes of the real data. Synthetic data should not be mapped back to the real data [4].

There are several approaches for generating synthetic datasets. Synthetic financial time series have been used in finance to generate synthetic datasets. This involves the use of simple statistical models such as the autoregressive or GARCH (Generalized Autoregressive Conditional Heteroscedasticity) models [4].

Multi-Agent-Based Simulation (MABS) is a technique that can be used to model and implement the schema of a diverse real mobile money service based on the MASON (Multi-Agent Simulation of Neighbourhood) toolkit which is core in Java [30]. This technique allows the expression of many behaviours of different actors as they are in a real-world mobile money service. MASON has the advantage of being fast, portable, and can handle large custom simulations as compared to other simulation tools such as Teambots, Ascape and RePast [13].

A number of opportunities exist for the use of synthetic data in finance. These include two broad classes of financial data such as retail banking (retail financial data) which is concerned with transaction data, loan applications, and customer service. This kind of data often has privacy concerns and synthetic data generation can be leveraged to carry out research in this domain. The second one is market microstructure data (market data) which is in the form of time series describing say a stock price over a given period [4]. Generators such as PaySim [25] which primarily generates mobile money financial payments and BankSim [28] for bank payment simulation have been developed to facilitate research in synthetic data generation in finance. Details on the different simulators previously designed have been presented in Section 4 of the paper. Most of these simulators are limited in the possible scenarios and actors compared to the current use case for mobile money. There is a need to access diverse, local context fraud data that could be used to calibrate them so as to generate synthetic financial datasets that represent real fraud scenarios in financial systems.

This research aims at describing design considerations for synthetic data generation for mobile money transactions. These include among others the need for real mobile money transaction datasets, common examples of fraud scenarios in the mobile money markets, appropriate design options, techniques for synthetic data generation and tools that have been previously developed to facilitate synthetic data generation in finance. It also examines the limitations of the different techniques and toolkits so as to propose a realistic model that can be implemented and used to generate diverse mobile money transaction datasets.

The rest of the paper is organised as follows: Section 2 describes the mobile money ecosystem and common fraud scenarios in mobile money services in SSA. In Section 3, a design of the synthetic data simulator for mobile money transactions is presented. Section 4 highlights work done previously, describes existing simulation tools and techniques for synthetic data generation, it also highlights the gaps in the tools. The paper is concluded with an insight into future work in Section 5.

2 MOBILE MONEY ECOSYSTEM

The mobile money ecosystem consists of several actors as illustrated in Figure 1. Some of the common actors include Clients, Agents, Banks, Merchants and Service Providers.

The clients hold registered accounts (mobile wallet) with service providers. They can credit and carry out transactions from their mobile wallet (M-wallet) through the transfer of funds to other clients or simply a deposit/withdraw transaction via a mobile money agent. These clients have got sub-divisions based on the activities they perform. Some of the clients behave as fraudsters in the mobile money system whereby they tend to exploit the weaknesses of the service for their own gains. Others simply behave normally, carrying out the standard transaction activities defined by the service providers.

The agents are an interface between the clients and the service providers. They get paid by the service providers, in form of a commission for the services they offer in mobile money. The agents register new clients on behalf of the service providers, receive cash to credit a client's m-wallet (cashin) and pay cash from a client's mobile wallet (cashout) [5]. They are able to share float among themselves in the agent network and load electronic money into the system with the help of the service providers. Some agents also carry fraudulent transactions in the mobile money systems.

The banks mainly facilitate the debit/credit transactions that involve the movement of electronic money between a mobile money account and a bank account. Within the mobile money system, very many intermediate money transfers can happen from one mobile money account to other mobile money accounts before the money is eventually cashed out of the system via an agent [40].

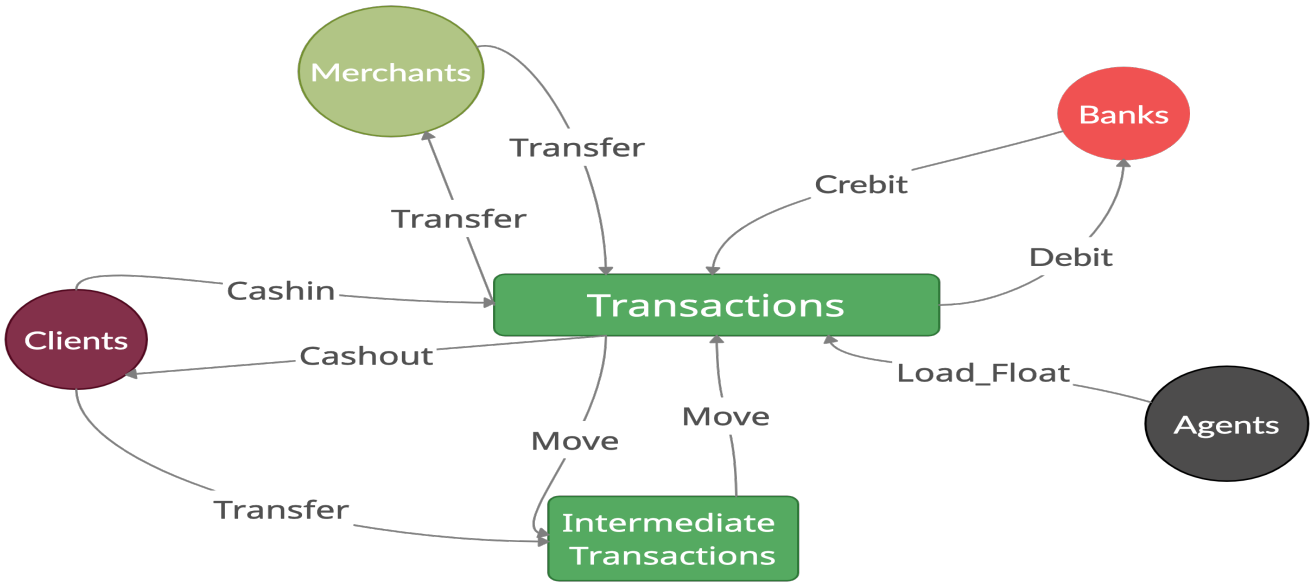

Figure 1: Mobile money actors

2.1 Categories of Mobile Money Agents

Different categories of mobile money agents exist in the mobile money market. This categorization is based on the amount of float for mobile money transactions that an agent can afford. They include high float agents who handle large transactions per day and keep records of all transactions in a digital system for purposes of a financial audit. High float agents are however few in markets and are located in big business centers. Medium float agents handle relatively higher amounts of transactions per day than low float agents and record most of the transactions they handle in a book. They are many across the region, they are found in most business centers and local markets. Low float agents handle a few transactions per day, rarely keep transaction records, and are many compared to high float agents. They are located in villages and small markets [40]. These different categories of mobile money agents in the market have been summarised in Table 1.

Table 1: Categories of Mobile Money Agents

Agent	Volume of Transactions	Status
High Float	Very High	Few agents
Medium Float	High	Very many agents
Low Float	Low	Many agents

Mobile money transactions include cashin which involves a client loading electronic money to their account via an agent, cashout which is concerned with removing money from a mobile money account via an agent, and merchant transfers which enable small and medium enterprises (SMEs) to receive payments. Mobile banking is a technology that involves transfers from bank accounts to M-wallet and vice versa. It facilitates debit/credit transactions between a mobile money account and a bank account. Another transfer operation involves a client moving funds from his account to another account within the mobile money network [42].

2.2 Fraud Scenarios in Mobile Money Systems

Tremendous work has been done in assessing and evaluating fraud scenarios and key security issues associated with mobile money systems in Uganda [3]. Security issues can be categorized as direct and indirect attacks. Direct attacks include authentication attack which is concerned with attackers taking advantage of weak procedures that have been put in place for mobile money users to reset their personal identification number (PIN). The PIN is an entry point to every successful mobile money transaction.

Identity theft is common among the workers of a mobile money agent shop. Most times in SSA, the agents put up other businesses together with the mobile money business and in an event of dishonest workers at the premises, they can take advantage and execute unauthorised transactions on behalf of their colleagues [33].

A smishing attack involves fraudsters sending emotionally hallucinating SMSs to victims requesting their mobile money PIN [34].

In a Brute-force attack scenario, fraudsters guess the numeric access codes of the mobile money systems using the machine-readable zone information thereby allowing them to carry out unauthorised mobile money transfers [20].

Denial-of-Service (DoS) attack involves attackers disrupting the mobile money network from a weak link in the network/system and carrying out SIM swaps to defraud the victims [9].

There is also a Man-in-the-Middle Attack involving the interception of a mobile money message by an attacker, altering it for their own gains before it gets delivered to the respective parties involved in the transaction [7].

A Salami Attack involves an employee of a telecommunication company or bank installing third-party software on the system to deduct small sums of money from mobile money transactions/accounts [6].

Indirect attacks on the other hand include an inside threat that is facilitated by employees of mobile money service providers and banks. Sometimes dishonest employees of banks and/or service providers collude with mobile money agents. These employees generate float (e-money) that is not backed by physical cash and this money is transferred to colluding agents, and clients and is eventually cashed out of the mobile money network. This affects the service providers since the money generated does not have evidence of cash in the bank account of the service provider [9].

Agent-driven fraud includes a mobile money agent taking advantage of the illiteracy or fear of their clients to carry out mobile money transactions themselves thereby allowing mobile money agents to carry out transaction approvals on behalf of their clients, potentially granting a chance for unscrupulous agents to carry out unauthorised transactions and money transfers on the accounts of the clients.

Malware arises from third-party libraries used by software developers in applications that have mobile money services to facilitate payments. These third-party libraries institute security vulnerabilities in the applications making mobile money transactions insecure. Also, there are vulnerabilities inherent in mobile phones such as unencrypted PINs, and USSD technology vulnerabilities such as improper data validation causing leakage of sensitive mobile money payments information [3].

With a fraud detection strategy in place, these indirect attacks could be noticed as soon as they happen. However, an efficient approach to mitigating these security concerns could start with securing the weakest link and identifying the most common fraud scenarios and behaviours that can be modeled using Machine Learning approaches.

The weakest link in mobile money service is the vulnerability of the entry point (PIN) and the behaviours of actors in the system. Security of mobile money transactions has mainly been enabled through a PIN which is normally a 4 or 5-digit number. The security of the mobile money operations is not always guaranteed since the PIN is vulnerable to fraudsters whose behaviours are more adaptive than the current financial fraud controls.

These fraudulent behaviours in mobile money systems do happen at different stages and these stages include the customer acquisition stage where a service provider is registering new clients through their agents. Many service providers often want to register as many clients as possible in order to increase their service coverage and revenue. They usually pay their agents commission to carry out the registration exercise.

The transaction activation stage follows customer acquisition and it is an instance where a client is granted rights to start making transactions through mobile money. In this stage, the service providers do encourage customers to make transactions using their mobile money services [34].

The customer loyalty stage is a scenario where clients have preferences for service providers of their choice. In this stage, customers can pay for many services such as water bills, satellite television subscriptions, and electricity bills using their preferred mobile money service.

Fraud scenarios and behaviours happen across all stages and include among others, split withdrawal and split deposit of funds initiated by agents. Split withdrawals involve an agent defrauding the service provider by dividing the withdrawal of funds from the mobile money system into small chunks so as to earn more commission from the transactions [3]. Similarly, some agents also split deposits into small values and carry out deposits for the small values separately, aiming at increasing their earnings.

Cases of a client depositing money directly into another client's account so as to skip the charges of transferring funds from one account to another have denied revenue to the service providers of mobile money [34].

Some dishonest agents transfer funds using the agent-to-agent network that is normally free thereby denying the chance for service providers to earn revenue from transfers that involve clients. Other agents use their relationships with service providers to launder money on the mobile money platforms [3].

Registration of non-existent customers and the creation of fake users in the mobile money systems is a behaviour reported to be practised by some dishonest mobile money system administrators [34].

The collision between mobile money agents and fraudsters to register an individual as a business so as to enable the individual to receive large sums of money from the population has been a common practice in East Africa.

The magnitude of the different fraud scenarios at the customer acquisition stage, transaction activation stage and at the stage where these customers are loyal to the service providers has been recorded as shown in Table 2.

Even though some manual fraud reporting mechanisms have been put by the local security unit in case of fraud, service providers have implemented sensitisation programs. These sensitisation programs are carried out through radio station announcements, sending messages using short messaging service (SMS) to clients at regular intervals and issuing fraud warnings over television and other print media [31].

Using these unique fraud scenarios and other design considerations, a design of a synthetic data simulator that is multi-agent-based and is tailored to mobile money systems based on real scenarios and processes can be formulated to aid fraud detection research in mobile money systems and transactions from SSA.

3 MODEL DESIGN FOR MoMTSim

This section details a design of a model based on the different fraud scenarios that exist in mobile money systems in Sub-Saharan Africa.

Fraud awareness programs have been set up, especially for mobile money financial fraud. However, these approaches are manual and leverage enforcement from security agencies that are already occupied with other security concerns in many other sectors. In finance, the sensitivity of financial records posed challenges in the study and modeling of diverse fraud scenarios in financial systems given that fraudsters are more adaptive to control measures than

Table 2: Local Context Fraud Scenarios

Fraud	Actor	Customer Acquisition Stage	Transaction Activation Stage	Customer Loyalty Stage
Split withdrawals	Agents	Low	Medium	High
Split deposits	Agents	Low	High	Medium
Direct deposits	Clients and Agents	Low	High	High
Parallel money transfer on the network	Agents	High	High	High
Money laundering on mobile financial service platform	Agents	Low	Medium	Low
Registration of customers with fake details	Agents	High	Medium	Low
Registration of individuals as business	Agents	Low	Medium	High
Registration of non-existent customers	Mobile money system administrators	High	High	Low
Creation of fake / non-existent users	Mobile money system administrators	Low	High	High

financial institutions adapt to mitigate emerging fraud scenarios. The financial data privacy restrictions then led this research to study techniques and tools that have been developed to synthetically generate financial datasets so as to aid fraud detection. There are several limitations in the state of the art as discussed in Section 4. This research presents a design that incorporates diverse, local fraud dynamics and examples in real-world operations of mobile money services in SSA.

This research proposes a model, *MoMTSim* that leverages the MABS technique to express processes in mobile money operations. The *MoMTSim* model design is based on the design considerations ranging from the use of information gathered from real mobile money transaction datasets to different fraud information captured from local mobile money markets so as to aid the generation of realistic synthetic datasets for mobile money transactions. The goal is to express a realistic design of a multi-agent platform that is tailored to mobile money transactions based on the design considerations, mobile money processes and local context scenarios that can be implemented so as to generate realistic datasets that are as close as possible to the original datasets.

The different fraud scenarios in mobile money transactions detailed in Subsection 2.2 are incorporated and used in the model so as to aid the generation of synthetic datasets with diverse fraud examples depicting what happens in a real-world mobile money system. The synthetic output results can then be used to study fraud detection approaches using different machine learning techniques such as Logistic regression, Decision tree and Neural networks such that real fraud threats can be automatically detected in mobile money systems.

The design of the model is shown in Figure 2 with the flow of synthetic data generation processes for mobile money transactions.

Mobile money has got several operational processes, so the model is based on an analysis of a large batch of transaction records of mobile money. From this analysis, the main actor in the model is a client who carries out deposits, and withdrawal of funds from the system and the client transfers funds from one account to another. These client operations are aided by a mobile money agent who primarily facilitates the conversion of hard cash into electronic money in the mobile money network. The analysis thus yields a profile file for the client that is used as initial input together with an aggregated file containing statistical properties of mobile money transactions to simulate the synthetic datasets. All these input files form transaction data metrics that are fed as input parameters to the simulator. The simulator contains other actors that are defined to act similar to known behaviours of actors in real mobile money operations.

Additionally, real-world behaviours of customers in mobile money operations such as having a limit on the daily amount of transactions are derived from the statistical analysis of the sample transaction datasets and are added to the model. With this statistical information, the future participation of a client in mobile money transactions is determined.

The simulator generates synthetic output results using the transaction data metrics which include a synthetic client profile and synthetic aggregated transaction file that is calibrated based on real mobile money processes defined in the implementation of the simulator.

Known fraud patterns are then injected during periodic calibration of the simulator so as to output diverse synthetic datasets that contain local context mobile money fraud examples. The simulation process continues using different seeds of input data like the way real mobile money systems operate. Therefore, several synthetic mobile money datasets are produced as final outputs. The results can then be verified and validated using scientific methods such as error-rate measurement, and the comparison of data distribution between the synthetic datasets and the original dataset using statistical methods. However, the validation of these results is beyond the scope of this paper.

4 RELATED WORK

This section focuses on discussing the available techniques used for synthetic data generation. It also discusses tools that have been developed to facilitate synthetic data generation and the limitations of the tools.

Different types of synthetic datasets exist. For instance, text synthetic data which is simply an artificially generated text, common in the domain of Natural Language Processing (NLP) where machine learning models are being used to generate text from natural language systems.

Media synthetic data include synthetic videos, images and sound that are generated by rendering media with attributes that are close to those of real data. These datasets are useful in vision recognition research [32].

Tabular synthetic data is based on real scenarios that produce synthetic datasets representing properties of real-life data that are kept in table format [35]. The data include financial transaction logs, healthcare datasets, and retail store data. Diverse financial data can be generated in tabular formats.

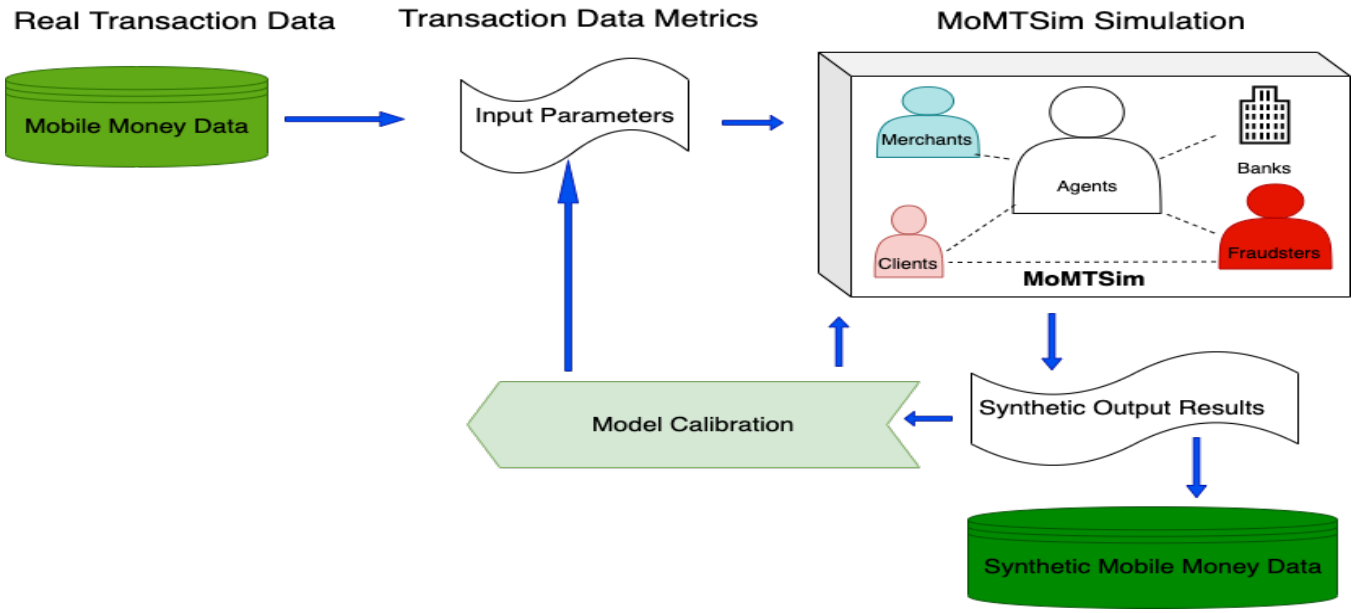

Figure 2: Design of MoMTSim simulator

4.1 Techniques for Synthetic Data generation

The techniques for synthetic data generation fall under two categories namely process-driven approaches and data-driven methods [17]. Process-driven approaches generate synthetic datasets based on real-world processes that require domain-specific knowledge to model. They do not rely directly on real datasets to synthesize new datasets thus avoiding re-identification risks.

Process-driven data generation examples include but are not limited to numerical simulations, discrete-event simulations, agent-based modeling and Monte Carlo simulations [24].

Data-driven approaches such as Neural Networks use real data to aid synthetic data generation. These methods are not based on domain expertise and are thus easier to scale [24].

Assefa et. al [4] highlight opportunities, challenges and pitfalls in generating synthetic datasets in finance. They discussed a number of approaches for generating synthetic datasets with privacy guarantees including the use of tabular data which encompasses XML-based synthetic data definition language (SDDL) from which synthetic data may be generated using classical machine learning classifiers such as Support Vector Machines, and Random Forest.

Agent-based modeling (ABM) is a technique that has been used in the context of synthesizing payment data [16]. This approach has been commonly applied in modeling bank payments since it has the ability to express the behaviours of real actors in a banking/mobile payment phenomenon. Different actors in payment systems are modeled to interact with each other and new scenarios are simulated between the different actors based on physical processes thus giving rise to synthetic datasets. However, an automated quantitative method for calibrating the models can cause data leakages [37].

Assefa et. al [4] also detail the synthetic financial time series technique which is concerned with using simple statistical models such as the autoregressive or GARCH (Generalized Autoregressive Conditional Heteroscedasticity) models for financial time series. Much as GARCH models are easy to fit by the use of maximum-likelihood, they rely on assumptions and are unable to reproduce many of the statistical features of financial time series.

Monte Carlo simulation is a technique that can be used to generate synthetic datasets from real datasets with known distributional parameters and the distribution of the datasets can be fit using machine learning models such as decision trees that can model non-classical distributions [22].

Stream data with privacy guarantees, first addressed by Dwork [14] is not a popular technique used for synthetic data generation and it is not certainly the most natural model of privacy for time series data. In this technique, stream data is described as a bit string with a continuous counter of the number of ones observed being indicated.

Unstructured data presented by Abay et al. [1] includes images and audio which are best approached using neural networks that are evolving rapidly. The approaches are still new in the field of synthetic data generation and they suffer from the usual problems that neural networks experience.

Graph theory has been used to synthesize transaction data by building a rule-based semantic graph to aid the data generation process. Information such as the dependency between zip codes and phone numbers is obtained from real data. This information constitutes the rules that the graph structure has to fulfil so as to realise synthetic datasets [23].

Modern data-driven methods such as the use of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have become popular in the domain of synthetic data generation [17].

GAN uses an interactive mode of operation involving a discriminator and a generator that is fed real sample data in order to generate synthetic outputs. The discriminator validates the results

by comparing the synthetic outputs with real datasets using a set of defined conditions [21].

VAE is an unsupervised machine learning approach which involves an encoder compressing original datasets into compact structures and transferring this information to a decoder. The decoder produces synthetic datasets that are close to the original datasets and the system is trained by optimising the correlations between the input datasets and the synthetic output datasets [21].

MedGAN is concerned with the synthesis of discrete variables with a variational autoencoder and a GAN [10]. On the other hand, CorrGAN involves the use of a convolutional neural network for generating synthetic datasets [36].

4.2 Tool-kits for Synthetic Data Generation

Multi-Agent-Based Simulation (MABS) of Financial Transactions for Anti Money Laundering (AML) [27] simulated the behaviour of several clients interacting in a mobile money system with the aim of producing a log of transactions. The simulation was built based on the MASON toolkit that expressed most of the behaviours of actors in the real mobile money service and it included some fraud scenarios. However, the validation of the model developed was limited by the lack of real mobile money datasets. This made the researchers rely on the opinions of experts and engagements with users. The model was not statistically validated and it could not be relied on for simulating diverse scenarios of real mobile money service.

Mason toolkit [30] is a fast, discrete event, multi-agent simulation purely implemented in Java programming language designed to handle simulations having a variety of multi-agent tasks with considerations of complex environments. This makes the formulation of real-world scenarios in multi-agent systems very possible. The toolkit eases the creation of multi-agent simulations and it is very portable. This makes it a preferred choice for modeling diverse processes of mobile money service. The toolkit requires a lot of customisation in order to harness its potential for creating complex real-world processes.

PaySim [25] is a data generator for synthesizing mobile money payment datasets that are as close as possible to sample mobile money datasets. PaySim uses the MABS technique and it was revised several times so as to generate realistic datasets. The simulator fits in the domain of generating mobile money datasets however it has several limitations raging from its inability to model diverse fraud scenarios that are ever emerging in the financial markets. It is limited to straightforward fraudulent scenarios which is not really the case with real-world mobile money operations.

RetSim [29], a retail store simulator is another simulation that applies the concept of MABS to model the behaviours of a shoe store staff and its customers. The model was calibrated using real shoe store datasets and social network analysis of the staff-customer interactions and relationships. Limited fraud scenarios in the retail store were modeled thus limiting the ability of the simulator to cope with generating synthetic data with diverse fraudulent behaviours of the actors in a retail store.

AMLSim [41] is a prototype data simulator that uses MABS and dynamics noticed in real financial data to generate synthetic datasets tailored to only detect money laundering. The patterns of fraud described in the model are specific to money laundering that represents real-world money laundering. Financial institutions experience many different fraud scenarios and thus the utility of the simulator leaves behind many scenarios that are present in financial transactions.

BankSim [28] present a bank payment simulator based on the MABS concept and analysis of aggregated transaction data to produce synthetic datasets. The association of customers and merchants were used to calibrate the model so as to generate datasets that relate to the real dataset. However, the model did not include some realistic transaction processes such as deposits, withdraws and transfers that are core in banking. Also, it didn't take into consideration, different fraud scenarios that are present in the domain due to the lack of diverse banking datasets.

ABIDES [8] is an Agent-Based Interactive Discrete Event Simulation that aids the simulation of many trading agents interacting with an exchange agent to make possible transactions. The goal of this research was to support AI agent research in market applications. However, at the current state, the simulation does not support complex trading agents and it is not based on the MABS technique that can model complex agents.

Gaber [15] presents a related technique for producing synthetic logs for fraud detection. However, the major difference is that they had real data for the calibration of their results and also for performing comparisons based on the quality of the results obtained from the simulator. The research mainly focused on generating testing data that researchers can be able to use to evaluate different approaches in fraud detection.

IncidentResponseSim [18] is a tool used for simulating data that enables the assessment of the risk of online banking services. The simulator primarily estimates the economic impact of current and future threats modeled with the aid of an incident response tree in combination with a qualitative model.

Rieke [38] presents Predictive Security Analyzer (PSA) tool-kit with the goal of identifying cases of fraud in a stream of events from a mobile money transfer service. PSA used a dataset of 4.5 million logs from a mobile money service for a period of nine months. The motivation behind this work was the limitation and knowledge of existing fraud in the current logs. PSA was intended to detect money laundering cases caused by the interactions of many users in the simulator by noting suspicious cases of money laundering with the goal of automatically blocking fraudulent transactions.

The work of Alexandre and Balsa [2] used simulated data to evaluate the method they developed to detect fraud based on intelligent agents. The method carries out tasks that a security officer does manually with limited data.

A summary of the tools that have been developed to model and generate synthetic datasets for some of the use cases in the real world is shown in Table 3.

In conclusion, the limitations of the tools reviewed range from the unavailability of real datasets that could be used to statistically calibrate the models. Also, the lack of diverse fraud examples at the disposal of the researchers posed a risk of calibration of the tools that aimed at detecting different financial fraud. Many of the tools could not generate realistic datasets that could be used to study fraud patterns. This research looked at the limitations of the previously designed tools and explored the design considerations so

Table 3: Simulation Tools For Fraud Detection

Tool	Purpose	Limitations
ABIDES [8]	Simulates diverse trading agents in market applications	It does not model complex trading agents
AMLSim [41]	Simulates financial transactions that can be used to detect patterns of money laundering	It is designed for purposes of studying money laundering only with limited data
BankSim [28]	Models bank payments using MABS	Lacks real world banking processes such as deposits, withdraws and transfers
Mason [30]	A MABS toolkit that models fast, portable, and large custom purpose simulations	Easy to use but requires a lot of customisation in order to incorporate processes and behaviours of agents of a real world phenomenon
PaySim [25]	Simulates mobile money transactions with straightforward fraud scenarios	Models a single type of fraud
RetSim [29]	Simulates retail store processes using MABS	It does not model mobile payments. Social network analysis and sample retail shoe store data were used for its calibration

as to propose a realistic design of a model that has been presented in Section 3.

5 CONCLUSION AND FUTURE WORK

This paper has been a step toward establishing requirements for realising *MoMTSim*; a multi-agent-based simulation model for generating synthetic datasets for mobile money transactions from Sub-Saharan Africa. It presents mobile money processes and examples of fraud scenarios in mobile money systems from Sub-Saharan Africa that can be modeled to generate synthetic datasets for automated fraud detection using machine learning algorithms. After establishing privacy concerns and sensitivity of mobile money financial transactions, a review of previous works that focused on generating synthetic datasets in finance was done. Based on the design considerations that address the limitations of the existing toolkits and also incorporate the diverse, local context fraud scenarios in mobile money transactions, a realistic simulator design was proposed for mobile money transactions to aid synthetic data generation.

It is evident in the local mobile money markets that fraudulent behaviours keep on evolving as time goes on. Future work for this research will focus on implementing the design with common fraud scenarios obtained from the mobile money markets. Verification and validation of the simulation results using appropriate scientific methods shall also be performed in future. The resulting datasets shall be made publicly available to the research community to inform and stimulate new research in the domain of financial fraud detection.

ACKNOWLEDGMENTS

This research has been possible with funding from JPMorgan Chase & Co AI Research under Faculty Research Award 2021.

REFERENCES

[1] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. 2018. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 510–526.

[2] Cláudio Alexandre and Joao Balsa. 2015. A Multiagent Based Approach to Money Laundering Detection and Prevention. In *In Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART-2015)*. 230–235.

[3] Guma Ali, Mussa Ally Dida, and Anael Elikana Sam. 2020. Evaluation of Key Security Issues Associated with Mobile Money Systems in Uganda. *Information* 11, 6 (2020), 309.

[4] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. 2020. Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance, ICAIF '20*. Association for Computing Machinery, Article 44, 8 pages.

[5] Karthik Balasubramanian, David F Drake, et al. 2015. *Service quality, inventory and competition: An empirical analysis of mobile money agents in Africa*. Technical Report. Harvard Business School Cambridge, MA.

[6] S Balasubramanian. 2018. Study of Cybercrime in Banking and Financial Sectors. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 3, 1 (2018), 1205–1212.

[7] Mercy Wangari Buku and Rafe Mazer. 2017. *Fraud in Mobile Financial Services: Protecting Consumers, Providers, and the System*. Technical Report. The World Bank. Accessed: 2022-03-27.

[8] David Byrd, Maria Hybinette, and Tucker Hybinette Balch. 2019. ABIDES: Towards High-Fidelity Market Simulation for AI Research. *CoRR* abs/1904.12066 (2019).

[9] Sam Castle, Fahad Pervaiz, Galen Weld, Franziska Roesner, and Richard Anderson. 2016. Let's talk money: Evaluating the security challenges of mobile money in the developing world. In *Proceedings of the 7th Annual Symposium on Computing for Development*. 1–10.

[10] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*. PMLR, 286–305.

[11] Sonja Davidovic, Soheib Nunhuck, Delphine Prady, and Herve Tourpe. 2020. Beyond the COVID-19 crisis: a framework for sustainable government-to-person mobile money transfers. (2020).

[12] Asli Demirgüç-Kunt, Leora Klapper, Dorothe Singer, Saniya Ansar, and Jake Hess. 2020. The Global Findex Database 2017. (2020).

[13] Alexis Drogoul, Diane Vanbergue, and Thomas Meurisse. 2002. Multi-agent based simulation: Where are the agents?. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*. Springer, 1–15.

[14] Cynthia Dwork. 2010. Differential privacy in new settings. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 174–183.

[15] Chrystel Gaber, Baptiste Hemery, Mohammed Achemlal, Marc Pasquet, and Pascal Urien. 2013. Synthetic logs generator for fraud detection in mobile transfer services. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 174–179.

[16] Marco Galbiati and Kimmo Soramäki. 2011. An agent-based model of payment systems. *Journal of Economic Dynamics and Control* 35, 6 (2011), 859–875.

[17] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. 2020. Generation and evaluation of synthetic patient data. *BMC medical research methodology* 20, 1 (2020), 1–40.

[18] Dan Gorton. 2015. IncidentResponseSim: An agent-based simulation tool for risk management of online Fraud. In *Nordic Conference on Secure IT Systems*. Springer, 172–187.

[19] GSMA. 2021. *State of the Industry Report on Mobile Money 2021*. Technical Report. Accessed: 2022-03-27.

[20] Md Arif Hassan, Zarina Shukur, Mohammad Kamrul Hasan, and Ahmed Salih Al-Khaleefa. 2020. A review on electronic payments security. *Symmetry* 12, 8 (2020), 1344.

[21] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.

[22] Jan F Kiviet et al. 2012. Monte Carlo simulation for econometricians. *Foundations and Trends® in Econometrics* 5, 1–2 (2012), 1–181.

[23] Pengyue J Lin, Behrokh Samadi, Alan Cipolone, Daniel R Jeske, Sean Cox, Carlos Rendón, Douglas Holt, and Rui Xiao. 2006. Development of a synthetic data set generator for building and testing information discovery systems. In *Third International Conference on Information Technology: New Generations (ITNG'06)*. IEEE, 707–712.

[24] Mikael Ljung. 2021. *Synthetic Data Generation for the Financial Industry Using Generative Adversarial Networks*. Ph.D. Dissertation.

[25] Edgar Lopez-Rojas, Ahmad Elmir, and Stefan Axelsson. 2016. PaySim: A financial mobile money simulator for fraud detection. In *28th European Modeling and Simulation Symposium, EMSS, Larnaca*. Dime University of Genoa, 249–255.

[26] Edgar Alonso Lopez-Rojas and Stefan Axelsson. 2012. Money laundering detection using synthetic data. In *Annual workshop of the Swedish Artificial Intelligence Society (SAIS)*. Linköping University Electronic Press, Linköpings universitet.

[27] Edgar Alonso Lopez-Rojas and Stefan Axelsson. 2012. Multi agent based simulation (mabs) of financial transactions for anti money laundering (aml). In *Nordic Conference on Secure IT Systems*. Blekinge Institute of Technology.

[28] E. A. Lopez-Rojas and Stefan Axelsson. 2014. BankSim: a bank payments simulator for fraud detection research. In *Proceedings of the European Modeling and Simulation Symposium, EMSS*. 144–152.

[29] E. A. Lopez-Rojas, Dan Gorton, and Stefan Axelsson. 2013. RETSIM: A shoe store agent-based simulation for fraud detection. In *25th European Modeling and Simulation Symposium, EMSS 2013*. 25–34.

[30] Sean Luke, Claudio Cioffi-Revilla, Liviu Panait, Keith Sullivan, and Gabriel Balan. 2005. Mason: A multiagent simulation environment. *Simulation* 81, 7 (2005), 517–527.

[31] Aaron Martin. 2019. Mobile money platform surveillance. *Surveillance & Society* 17, 1/2 (2019), 213–222.

[32] Andrea Britto Mattos, Dario Augusto Borges Oliveira, and Edmilson da Silva Morais. 2018. Improving CNN-based viseme recognition using synthetic data. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[33] Adam B Mtaho. 2015. Improving mobile money security with two-factor authentication. *International Journal of Computer Applications* 109, 7 (2015).

[34] Mudiri Joseck Luminzu. 2013. Fraud in Mobile Financial Services. https://www.microsave.net/files/pdf/RP151_Fraud_in_Mobile_Financial_Services_JMudiri.pdf. Accessed: 2022-04-21.

[35] Alexander G Ororbia II, Fridolin Linder, and Joshua Snoke. 2016. Privacy protection for natural language: Neural generative models for synthetic text data. (2016).

[36] Shreyas Patel, Ashutosh Kakadiya, Maitrey Mehta, Raj Derasari, Rahul Patel, and Ratnik Gandhi. 2018. Correlated discrete data generation using adversarial training. *CoRR* (2018).

[37] Donovan Platt. 2020. A comparison of economic agent-based model calibration methods. *Journal of Economic Dynamics and Control* 113 (2020), 103859.

[38] Roland Rieke, Maria Zhdanova, Jürgen Repp, Romain Giot, and Chrystel Gaber. 2013. Fraud detection in mobile payments utilizing process behavior analysis. In *2013 International Conference on Availability, Reliability and Security*. IEEE, 662–669.

[39] Marina Solin and Andrew Zerzan. 2010. Mobile money: Methodology for assessing money laundering and terrorist financing risks. https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2012/03/amlfinal35.pdf. Accessed: 2022-04-21.

[40] Peter Tobbin. 2011. Understanding mobile money ecosystem: ROLES, structure and strategies. In *2011 10th International Conference on Mobile Business*. IEEE, 185–194.

[41] Mark Weber, Jie Chen, Toyotaro Suzumura, Aldo Pareja, Tengfei Ma, Hiroki Kanezashi, Tim Kaler, Charles E. Leiserson, and Tao B. Schardl. 2018. Scalable Graph Learning for Anti-Money Laundering: A First Look. *CoRR* abs/1812.00076 (2018).

[42] Idongesit Williams. 2013. Regulatory frameworks and implementation patterns for mobile money in Africa: The case of Kenya, Ghana and Nigeria. In *Ghana ICT Conference*.