Near Real-time SIM Box Fraud Detection using Stream Mining

Sisay Matebe
HiLCoE, Software Engineering, Ethiopia
EthiopiaSisaymatebe@gmail.com

Mesfin Kifle
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

Voice traffic termination fraud, often referred to as Subscriber Identity Module box (SIM Box) fraud, is a common illegal practice on mobile operators. It terminates international calls using a local SIM card inside the SIM Box machine and delivered to the customer as local call without operator's know how to manipulate the tariff difference. As a result, operators are losing billions of money annually. In this paper, we propose near real-time SIM Box fraud detection using stream mining analytics technique to enable the system real-time decision making by inspecting, correlating and analysing the telecom data as it is streaming into applications. Rather than singling out specific transactions, our solution analyses historical transaction data to model a system that can detect fraudulent patterns. This model is then used to analyse Call Detail Records (CDR) transactions in real-time.

We have employed Lammed architecture with seven time-sensitive data mining classification algorithms, namely Decision Tree, Extra Trees Classifier (random), Logistic Regression and Random Forest, Linear SVM (Support Vector Machine), MLP Neural Network, and Gradient Boosting Classifier. Six selected attributes from the CDR that could be used to detect fraudulent behaviour in a near-real time are extracted. The experiment set up will enable us to understand SIM Box from genuine call behaviour both in near-real-time and one-hour window (eleven attributes selected), in case near-real-time detection missed to detect. Finally, we obtained a good result from Very Fast Decision Tree (VFDT) classification for near-real time and linear SVM for one-hour detection. VFDT classification algorithm achieved an accuracy of 98.5% and false positive 0.4% with average latency of 2 minutes per detection at input event rate of 280 CDR per second with limited hardware. SVM classification achieved 99% accuracy with 0.2% false positive for one-hour window detection. Finally, we propose both near-real-time and one-hour SIM Box fraud detection solution for effective outcomes.

Keywords: Stream Event Processing; Stream Mining; Data Analytics; Call Detail Records (CDR); Call Patterns; SIM Box Fraud; Telecom Fraud

1. Introduction

Telecommunication Fraud can be defined as any activity by which service is obtained with intention of not to pay [2]. Imagine you get an unknown call from a local number, you pick up the phone and it is from a friend or a family living abroad, it does feel strange that you would receive an international call from a local number; this is basically the SIM Box fraud, or SIM Boxing. In this case, the international calls are transferred over the Internet to inject them back into the cellular network through SIM Box machine with multiple local SIM cards inside. As a result, the calls turn as local to the destination network and the fraudsters who set up these boxes pay only local rates for mobile operators after charging international rates from the source operators.

Currently available traditional fraud detection systems make detections by generating a set of features or aggregate values by querying static data over a large time window and make decisions based on such values [3, 4]. This is time consuming and by that time a fraudster can make a number of successful calls before being detected. SIM Box Fraud has a big business impact to telecom operators and demands near real time CDR analysis tool which can watch CDR streams and generate a recent or near-real time view of the incoming CDR. Such approaches support the operators to gain maximum advantage by exploiting timeliness of detected events

and save significant amount of revenue by minimizing fraudulent activities.

In this paper, a real time detection system is designed and tested in cooperation with ethio telecom. The system utilizes stream mining techniques to detect whether a SIM card is used by a normal customer or by a SIM box fraudster.

2. Literature Review

2.1 Telecom Fraud Types

The telecom industry is one of the major sectors in the world infiltrated by fraud [1, 2]. There are different types of telecommunication frauds and these can occur at various levels. Different authors categorized frauds in different ways. For instance, the work in [8] categorized subscription fraud and technical fraud in general as the most common fraud types, whereas the work in [4] classifies them into seven groups as Superimposed, Subscription, Technical, Internal Fraud, Social Engineering, Fraud based on loopholes in technology, and Fraud based on new technology. Similarly, the work in [5] lists the top three types of telecommunication frauds that cause a significant loss to the operators as International Revenue Share Fraud, Premium Rate Service Fraud and Bypass fraud. In [14], they are classified as Contractual, Hacking, technical and Procedural frauds.

a. International Revenue Share Fraud (IRSF)

IRSF is the largest contributor to the overall fraud losses according to Communication Fraud Control Association (CFCA). It is when a fraudster makes an agreement with a local carrier in high cost destination to share profit for increasing traffic, then the fraudster hacks into any organization's public branch exchange (PBX) and gets illegal access to generate calls. After that, the fraudster generates high traffic calls to high cost destinations and gets revenue from the sharing agreements [1].

b. Premium Rate Service Fraud

Premium rate service is an agreement between some service provider and telecom companies to share revenues generated by traffic to the premium service number. It is used in TV shows, contests and entertainment services. The fraudsters try to stimulate consumers by giving them a missed call or a message, then make money from share of the call back revenues [4].

c. SIM Box Fraud

Interconnect Bypass (SIM Box) fraud is unauthorized insertion of traffic onto another carrier's network. This fraud type can be named as Interconnect fraud, Global System for Mobile Communications (GSM) Gateway fraud, or SIM Boxing. This scenario requires that the fraudsters have access to advanced technology such as VoIP, which can make international calls appear to be cheaper domestic calls, effectively bypassing the normal payment system for international calls. The fraudsters will typically sell long distance calling cards. When customers call the number on the cards, operators can switch the call to make it look like a domestic call. The most common implementation of interconnect bypass fraud is known as SIM Boxing, which is enabled by VoIP GSM gateways. SIM Box fraud transports international calls through VoIP, then routes them to the local cellular network via a collection of SIM cards inside a SIM Box device.

To explain how SIM Box fraud is committed, firstly we describe the legitimate way for international calls. Let us assume that caller A and caller B are in different countries. Caller A makes a call to caller B over the mobile operator. The mobile operator of country A takes the call and sends it through its international gate to a transient operator. The transient operator then routes the call through voice over IP (VoIP) to country B's mobile operator and pays a toll. After that, the mobile operator of country B terminates the call through its network to caller B. Figure 1 [15] shows the legitimate route.

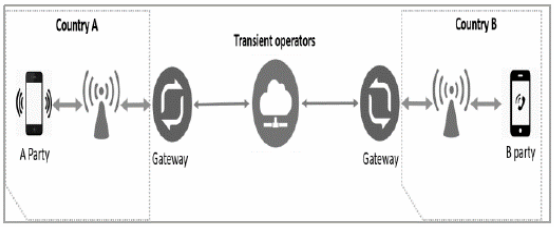

Figure 1: Legitimate Route of International Call

In bypass fraud, the transient operator routes the call through a SIM Box placed in country B using VoIP, the SIM Box then reroutes the call through country B's mobile operator and pays for just the

local call. Figure 2 [15] shows the bypass fraud route.

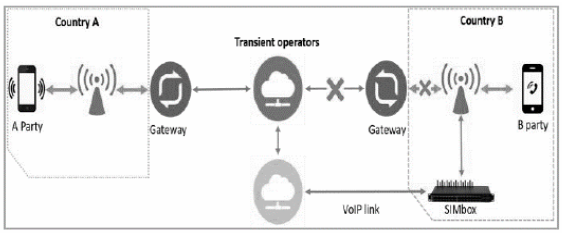

Figure 2: Bypass Fraud Route of International Call

The incentive here is that the toll charge by country B's mobile operator is much higher than the local call fee. Bypass fraud is committed when fraudsters install SIM Boxes with multiple low-cost, prepaid SIM cards. SIM Box equipment includes SIM slots and antennas. The SIM Box is connected to the Internet through an Ethernet port.

d. On-Net Vs Off-net Bypass Scenario

We can divide Bypass fraud into two major categories as On-net bypass and off net Bypass. On-net bypass means fraudsters use the SIM cards of same network of destination number. This is the common case in most of the countries as calls within same network cost much lower than charges for calls to other operators or international calls. Instead of using the costly high-quality routes that go through destination networks' International Switching Center (ISC), some wholesale carriers route calls through SIM Box operators connected through the Internet. The SIM Box dials that call via a SIM card of the same network of the destination number.

2.2 SIM Box Fraud Detection Techniques

Detection of SIM cards used for SIM Box fraud is a challenging task for mobile operators. The major methods used in battling SIM Box fraud currently in use include Test Call Generation (TCG), CDR analysis and controlling distribution of SIM Cards.

a. TCG (Test Call Generation)

TCG is used as an active method to detect bypass fraud where operators test different international routes to their network and analyze whether or not calls are coming through an international number or local number routes [16]. If it came from a local number, definitely it is associated with some SIM card used in a SIM Box that granted the fraud department to process it. TCG is one of the effective methods to detect SIMs used in SIM Boxes. However, this method costs the operator to generate test calls to the entire international routes. Also, since this method is based on generating calls to random numbers, getting SIM cards used by SIM Boxes is probabilistic.

b. Fraud Management System (Role -based analysis)

FMS is based on predefined rules and threshold that are manually defined by domain experts aimed to segregate SIM Box fraudsters. These rules and threshold are summarized features constructed from CDR data of subscribers. These are used to identify abnormal behaviour of SIM cards by comparing with analyzed subscriber role. Through this process it is able to distinguish fraudulent from legitimate usage. If abnormal subscriber behaviour is observed, then the system generates an alarm and experts perform analysis on it.

c. Non-technical Methods

SIM cards are indispensable to SIM Box fraudsters to survive since fraudsters need to maintain sufficient supply of SIM cards to be in business. However, SIM card distribution control will make this process difficult. Requiring government IDs and limiting the number of SIM cards per ID will prevent fraudsters from obtaining a large number of SIM cards to install in their SIM Boxes.

2.3 Challenges for Existing Defection Solutions

Fraudsters and anti-fraud are in eternal battle. Every time detection technology improves, fraudsters develop new methods to avoid detection and increase profit. This section describes different methods used by fraudsters to avoid SIM blocking.

a. Anti-Spam (Test Call Detection)

One of the effective methods to detect SIMs used in SIM Boxes is generating test calls (TCG) using different routes to known local network numbers. The incoming call will indicate weather it is coming from a local number or from an international number. If it was coming from a local number then it must be associated with some SIM card used in a SIM Box and easily processed by the fraud department. However, the fraudsters analyze the voice call traffic coming to their SIM Boxes and based on usage and other patterns they could determine whether the calls

were real subscriber calls, or they originated from a TCG system. They could then either block the test calls and prevent them from reaching the SIM Box or reroute the calls to a legitimate route to avoid detection.

b. Human Behaviour Simulation (HBS)

Smart SIM Boxes are designed to imitate the behaviour of normal customers by using HBS [14]. This technique makes detection of fraudsters very difficult if no advanced detection algorithms are used. HBS encompasses the following.

SIM Migration (Movability): Fraudsters deploy many gateways in different locations, for example, one in a city center and another in a shopping mall or some other crowded place and once in a while they swap the SIM cards between the gateways, so it would look like the user is moving. The swapping operation could be done manually or automatically using software.

SIM Rotation: SIM Boxes can be detected easily if fraudsters operate their SIMs around the hour excessively, so they limit their usage by rotation of the SIMs as workers shifts. This will make SIMs operate in limited hours a day, which simulates the behaviour of ordinary customers.

Usage of Other Network Services: Most of the SIM Boxes use just voice services and that makes them vulnerable to detection. In order to mitigate this issue smart SIM Boxes make calls and send SMS to each other. Also, sometimes they use some Internet services provided by the network operator. HBS makes dealing with SIM Box fraud harder and time consuming. Advanced measures must be taken to tackle this problem. In this work, real-time SIM Box fraud detection system is used to detect the bypass fraud by analysing CDR data.

3. Related Work

Various solutions were proposed by several authors to detect and prevent telecom fraud. A learning system has been one of the methods proposed for fraud detection tools [13]. These tools encode knowledge about fraud attacks as if-then rules which can represent and reason about some knowledge rich domain with a view to solving problems and giving advice.

The European Institute for Research and Strategic Studies in Telecommunications (EURESCOM) project, which was carried out in 2002, is the most significant contribution to the area of telecom fraud detection [13]. The project looked at how intelligent techniques from the domains of learning, personalization and data mining could improve fraud detection in a telecom service.

A technique that was proposed is the emerging pattern detection (EPD) [7]. EPD attempts to detect new forms of fraud by comparing two data sets, one with clean customer usage records and another one containing fraudulent records to identify patterns recurring in the second data set which were not present in the first data set.

A similar and powerful data analysis method is also proposed to discover SIM Box fraud patterns using integration of TCG and FMS [6, 7]. The suggestion is to incorporate TCG detection technology into the fraud management system. The integration attempts to detect new forms of SIM Box fraud using FMS CDR Analysis engine data sets by taking fraudulent behaviours of the already detected fraudulent records from TCG. In [7], a supervised learning algorithm is used to detect SIM Box fraud. The dataset is gathered from the operator and it includes CDRs from both legitimate subscribers and a fraudulent SIM Box. The proposed classifier has scored 98.7% accuracy in identifying the SIM cards that are used in the SIM Box device.

A similar study was done by Feven Fesseha [10], who used predictive model using data mining technology to detect SIM Box fraud. Twelve Selected CDR attributes were used to train the model. The study was conducted using WEKA software with five data mining algorithms for classification and the BFTree is chosen as the best to SIM Box fraud detection with an accuracy of 98%.

Frehiwot Mola [11] did a similar research on SIM Box fraud detection. The algorithms used to detect and predict SIM box fraudulent numbers are PART, J48, multilayer perceptron and hybrid. Using voting combination mechanism and by taking the strength of J48 decision tree and PART from rule based classification algorithms a hybrid algorithm is built. Nine features extracted from CDR data are utilized to

build a model that can be used to distinguish between legitimate and SIM box fraud calls. The feature includes calling number, called number, call fee, call duration, date and time, location number, total number of calls per day, SMS and GPRS call detail records. PART classification from rule based induction resulted in 99.49% accuracy and performs better than multilayer perceptron, J48 and Hybrid algorithms (J48 and PART).

In [12], the researcher analysed CDR data and identified twelve features to distinguish SIM-Box fraudulent from legitimate subscribers. The author proposed three dataset profiles (4 hour, daily, and monthly aggregated), to detect within possible short period of time. The researcher selected three classifiers: RF, ANN and SVM. The RF model scores slightly higher than the other two on the 4 hour dataset profiling. The accuracy of the model built with the 4 hour datasets are 95.98% and on the daily and monthly dataset is 98% and 99.05% respectively.

Tigistu Debebe [9] used ethio telecom as a case study. The objective is not detection but management principles on global fraud detection approaches. Finally the author proposed that ethio telecom needs advanced fraud management system with different technological options to detect fraud proactively.

4. The Proposed System and Findings

We have followed Lambda architectures to meet the research objectives. The Lambda Architecture is the new paradigm for big data that helps in data processing with a balance on throughput, latency and fault tolerance. The Lambda Architecture defines a set of analytic layers for building a complete big data system: Real-time Layer, Batch Layer and Serving Layer. Each layer satisfies a set of properties and builds upon the functionality provided by the layers under it. Figure 3 shows the high-level system architecture of the proposed real-time SIM Box fraud detection system.

The proposed detection system design was done in four stages:

a. Data Collection and Cleaning

Call Detailed Records were obtained from ethio telecom core systems. Basically there are two main data sources, namely, Local and International CDRs that are used in our design. Local CDR means transaction logs for calls originated by operator's own subscribers. These records are generated at Mobile Switching Centers (MSCs). International CDR records correspond to calls that originate to or terminated from foreign operators and detailed version of these CDRs are generated there. As seen from Figure 4, a new file is generated in less than a minute in MSC. Those CDRs files are originally generated as Abstract Syntax Notation One (ASN.1) formatted files and then converted to Comma-separated values (CSV) format for our processing. This data can then be blended with historical or other enterprise data sources including Equipment Identification Register (EIR) and history data from Fraud management database (FMS) to enrich the feed before further processing by downstream systems. The CDR's sensitive data like user numbers (MSISDN) or identities were modified. CDR data was processed in order to remove missing values, duplicate information and useless fields.

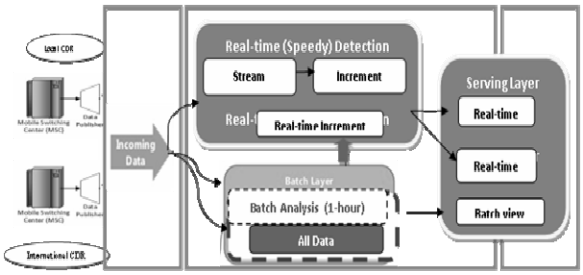

Figure 3: High-level System Architecture

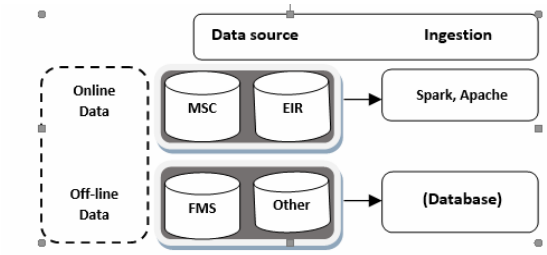

Figure 4: Data Sources and Ingestion

The real time Ingestion Service is the entry point to the streaming architecture as it decouples and manages the flow of information from MSC and EIR to the processing and storage components by providing reliable, high throughput, and low latency capabilities.

b. Feature Extraction and Engineering

To prepare the data for the machine learning model, informative features for each CDR were extracted. Also, features had been engineered in order to get the best features for training the model.

Every time a call is placed on a telecommunications network, descriptive information about the call is saved as CDR. CDR information can be used to analyse the characteristics of each call. The dataset used for this work are listed in Table 1.

Table 1: Data Set

	1-houd window	*Near Real-time (<2min)*
Training data	11573	11573 (375 separated by 2min)
Known fraud	333	141 in each 2min

First dataset comprises of 11573 CDR entries of which 332 were known fraud and the rest are for testing purpose. Further we have Split subsets of data to train the model, test it and further to validate against new dataset. All features of the CDR are not used directly for data mining, since the goal of data mining is to extract knowledge at the customer level. Thus, the CDR associated with customers must be summarized into a single record that describes the customer's calling behaviour. All relevant data for each subscriber was assembled in the form of columns and each raw dataset represents a unique subscriber. The choice of features is critical in order to obtain a useful description of the subscriber. So, a total of 12 features have been identified to be useful in detecting SIM Box fraud in 1-hour window and 6 future for near real time detection. Table 2 shows the list of these features and their corresponding data types. The selected features are based on the literature studied on the typical characteristics of SIM Box fraud subscribers as well as contribution from the experience of the staff who work on telecom fraud.

Table 2: Features Identified for 1-Hour Window Detection

Attribute	*Description*
Sub_No	Calling Number (Primary Key)
Idd_Call_Times	Toltal International Reciving Call
Idd_Calling_Times	Total International Outgoing Call Count Originated By Given Subscriber
Sms_Send_Cnt	Count Of Sms Sent By Number
Sms_Receive_Cnt	Count Of Sms Recived By Number
Idd_Sms_Cnt	Total International Sms Sent By Number
Calling_Duration	Total Outgoing Call Duration By Number
Cell_Count	Total Number Of Distinct Cell Coun By Number
Risk_Area_Coun	Number Of Orginating Number Risk Area Count
Bn_Missed_Idd_Calling	Count Of International Miised Call Made By Bn.
Calling_Imei _Cnt	Number Of Different Imei Numbers Used By Calling Party.
Is_Fraud	Known Fraud Numbers For Traning

We use custom ratio to split it into the two subsets, use the first 90% of data for training and the subsequent 10% of data for testing.

For the near real time detection, we rejected fields contain any instance of ***extreme usage*** related attributes. Extreme usage scenarios are relatively invalid compared to real-time SIM Box fraud instances. Therefore, we have taken CDRs for past instances of non-extreme usage related fraud cases and used those data to build our near real time logics.

Locating near real time patterns SIX CDR attribute are identified see table 3. Some attributes are selected from the two minutes collected CDR of unsuccessful call from the international CDR.

Table 3: Features Selected for Near Real Time Detection

Field Name	*Description*
Prod_ID	This is the Subscriber Identity Module (SIM) number which will be used as the identity field
MSISDN_Prefifix	number the First 5 prifix numbers
AN_Cell_ID_RISK_AREA	Cell site risk area flag
BN_IDD_SMS_in_2_min	The total international SMS initiated by BN with in 2_min
BN_Unsuccessful_IDD_Call	Total unsuccessful international Calls made in 2 minute
AN_Cloned-IMEI_Flag	AN_IMEI registered fraud flag
IS_Fraud	Known fraud numbers for training purpose

c. Model Training

The features extracted in the previous stage were used to train machine learning models. In this stage,

we used two different models. Besides real time analysis, we trained one-hour detection scenario to detect SIMs that failed to be detected in the real time scenario. Decision Tree, Extra Trees Classifier (random), Logistic Regression and Random Forest, Linear SVM, MLP Neural Network and Gradient Boosting Classifier had been used as supervised learning algorithms to train the model.

d. Model Evaluation and Testing

The final stage was testing and comparing the performance of the designed algorithms in terms of accuracy. To achieve this objective we used stream mining algorithms for near real time and IBM spss modular for one-hour decisions making. Figure 4 shows integrated stream mining to be applied for the proposed SIM Box fraud detection solution.

The training dataset used for real time SIM Box fraud detection consists of 11752 distinct CDR records and 331 known fraud cases. After several cycles of fine tuning, the model fits to detect. The test dataset contained all 5670 subscribers out of which 155 are known fraudulent calls. The proposed system detected 154 fraud instances correctly with one false positive and 12 false negatives. The same data sets with different features were used and out of it 139 frauds were detected using near real time detection scenario whereas the remaining 15 additional detections were made using one-hour window detection. As a result, near real time detection scenario acquired highest percentage of 81% of detection rate whereas one-hour window scores 99%. Detected in near real-time before huge impact is felt to operator with 98.5% over all detection accuracy. Following the research result, we consulted telecom fraud subject experts. They confirmed that the research result is promising to support telecom operators by detecting SIM Box fraudsters in a very short time.

- Linear SVM classification model is chosen as the best by IBM modular classifier. It scored 99.61% accuracy. This algorithm examined all the 11 fields in the training data set.
- VFDT classification model is chosen as the best real time classification by AutoAI IBM classifier. It scored 98.5% accuracy. This algorithm examined all the 6 fields in the training data set.

5. Conclusion and Future Work

The focus of this research is to detect SIM Box fraud in telecom networks in near real-time using streaming analysis techniques on call patterns indicated from the CDR. As a result, we followed the Lambda architecture and developed a system architecture that comprises batch (1-hour) and real-time layers, as it is well suited for applications which perform both near real-time and batch analytics. IBM cloud solution was used for both batch and real time SIM Box detection experiment due to its ability to perform high-speed processing. As a result, we have significantly increased the detection speed of SIM Box fraud. As per our observations, the proposed system performed the needed functionality.

As a future work, more refinement of data features will be done to improve performance and accuracy of the real time detection technique, for instance, including very sensitive data like signaling data will highly advance the detection capability.

References

[1] Global Fraud Loss Survey, C. F. C. Association et al., “Global fraud loss survey,” Press Release, http://www.cfca.org/fraudlosssurvey.

[2] A. Abdallah, M. A. Maarof, and A. Zainal, “Fraud detection system: A survey,” Journal of Network and Computer Applications, Vol. 68, pp. 90–113, 2016.

[3] M. Yelland, “Fraud in mobile networks,” Comput. Fraud Secur., Vol. 2013, No. 3, pp. 5–9, 2013.

[4] James Le, 2017, “Blue Ocean Thinker” https://jameskle.com/.

[5] P. Gosset and M. Hyland, “Classification, detection and prosecution of fraud in mobile networks,” Proceedings of ACTS mobile summit, Sorrento, Italy, 1999.

[6] Airn Vipin, ”Analysis and detection of SIM Box”, 2018. www.ijariit.com Combating Against Telecom Fraud: http://www.mmtelcom.com/webdex/fraud_prev.html.

[7] Latro services, 2016, "Beat Fraud before Fraud Happens", http://www.latroservices.com/

[8] Farvaresh, H. and Sepehri, M. M., "A data mining framework for detecting subscription fraud in telecommunication", Engineering Applications of Artificial Intelligence, 2011.

[9] Tigistu Debebe, “Detecting fraud in a mobile telephony network, ethio telecom as a case study", Unpublished Masters Thesis, HiLCoE, 2017.

[10] Feven Fesseha, “Predictive SIM Box Fraud Detection Model for ethio telecom”. Unpublished Masters Thesis, HiLCoE, 2017.

[11] Frehiwot Mola, “Analysis and Detection Mechanisms of SIM box Fraud in The Case of Ethio Telecom”, Unpublished Masters Thesis, Addis Ababa University, 2017.

[12] Kahsu Hagos, “SIM Box Fraud Detection Using Data Mining Techniques: The Case of ethio telecom”, 2018.

[13] Bradley Reaves, Ethan Shernan, and Adam Bates, "Blocking Cellular Interconnect Bypass Fraud at the Network Edge", 2015.

[14] Murynets, M. Zabarankin, R. P. Jover, and A. Panagia, “Analysis and detection of SIM box fraud in mobility networks,” Proc. IEEE INFOCOM, pp. 1519–1526, 2014.