مجلة العلوم البحثة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.ly ISSN 2521-9200
Received 03/07/2018 Revised 19/12/2018 Published online 15/01/2019

Study to use NEO4J to analysis and detection SIM-BOX fraud

Ibrahem Soliman Alsadi , *Nassir Abuhamoud
School of Communication and Computer Engineering, Sebha University, Libya
*Corresponding author: mans.Abuhamoud1@sebhau.edu.ly

Abstract The high price of incoming international calls is a common method of subsidizing telephony infrastructure in the developing world. Accordingly, international tele- phone system interconnects are regulated to ensure call quality and accurate billing. High call tariffs create a strong incentive to evade such interconnects and deliver costly international calls illicitly. Specifically, adversaries use VoIP-GSM gateways informally known as “simboxes” to receive incoming calls over wired data connections and deliver them into a cellular voice network through a local call that appears to originate from a customer’s phone. In this paper, we analyze and compare known methods of fraud detection (sim-box), explaining the advantages and defects for each method and proposed a new method. System relies on analyze CDR files using data mining technology (Neo4j), and then use known method TCG (test call generation) to increase efficiency and to be more sure to results.

Keyword: Sim-Box, Network, Fraud, Mobile, Telecommunication.

دراسة إستخدام تقنية NEO4J للتحليل و الكشف عن SIM-BOX

إبراهيم سليمان الساعدي و *ناصر أبوهمود
كلية العلوم الهندسية والتقنية –جامعة سبها، ليبيا
*للمراسلة: mans.Abuhamoud1@sebhau.edu.ly

الملخص يعد ارتفاع أسعار المكالمات الدولية الواردة طريقة شائعة لدعم البنية الهاتفية في العالم الثالث، حيت آن استحقاق الشركات لجزء من القيمة المالية للمكالمة الواردة من دول اخري يؤخذ كقيمة للتطوير والتحسين في الخدمات بدلا من ارهاق كاهل المواطنين بالمكالمات الداخلية. ولاكن كالعادة في كل الاعمال التجارية، توجد عمليات احتيال لتقوم شركات وافراد للاستحواذ على هذه المداخيل. وتعد أشهر الطرق واكثر استخداما ما يسمى بصندوق الكروت (Sim-Boxes)، حيت تستخدم بوابات VoIP-GSM لتلقي المكالمات الواردة عبر اتصالات البيانات السلكية وتسليمها إلى شبكة هاتف خلوية من خلال مكالمة محلية تبدو وكأنها ناتجة عن هاتف العميل. في هذه الورقة، نقوم بتحليل ومقارنة طرق الكشف عن الاحتيال المعروفة، موضحا مزايا وعيوب كل طريقة واقترح طريقة جديدة. يعتمد النظام على تحليل ملفات CDR باستخدام تقنية استخراج البيانات الضخمة ومن تم استدعاء نظام توليد المكالمات العشوائي للتأكد وزيادة الدقة.

الكلمات المفتاحية :-sim-box، الشبكة ،الاحتيال، الاتصالات المتنقلة.

I. Introduction

Fraud is a problem affecting operators and telecom companies around the world. This is a significant source of lost revenue from the telecommunications sector. Bypass fraud, which is used in international calls to avoid access fees and profit taking, is the most popular fraud patterns, causing heavy loss of revenue for operators and other negative impacts. As cellular network operators lose about 3% of annual revenue due to fraudulent and illegal services. Juniper Research estimated the total loss from the underground mobile network industry at $ 58 billion in 2011. [1]

The SIM-box or Subscriber Identity Module-box is the most current and pervasive fraud that most mobile networks experience, particularly in the third world, where many people travel to other countries, The simplicity of its use and the low cost of calls make it the ideal choice for most of them, despite the enormous risks to the local economy, national security and even the movement of innovation and development in the field of mobile networks.

The losers are primarily the networks – but there is also an impact on two other groups. Users of telecommunications services may find that call quality is noticeably lowered and additional functionality and services such as caller line identity can often be missing. In addition, many telecommunications regulators or governments take the issue of SIM box fraud seriously. The absence of revenues to telecommunications providers impacts the taxation that governments can raise from the industry. In a small number of territories, telecommunications regulators themselves stipulate that networks should always deploy a reputable SIM box detection service as part of their licence conditions to offer telecommunications services. Sadly this is fairly infrequent, but the revenue that can be recovered by a reliable and trusted SIM box detection service is usually far higher than the cost of detection – providing telecommunications service providers with the opportunity to recoup significant lost revenues. The key to ensuring that revenues remain buoyant is to sustain the detection service. Fraudsters are opportunists and will seek to generate illicit gains whenever they can. Network operators need to be vigilant, particularly when, according to that recent survey, more than 80% of

operators queried identified SIM box fraud occurring on their network.

II. Background

II.1. Cellular Networks:

The Global System for Mobile Communications (GSM) is a suite of standards used to implement cellular communications. It is used by the majority of carriers in Europe, Africa, and Asia. GSM is a "second generation" (2G) cellular network and has evolved into UMTS (3G) and LTE (4G) standards. We focus on GSM because it is the most available for direct experimentation in Libya. Note that the methods we present in the paper can easily be ported to other cellular standards.

II.2. VoIP:

Voice over Internet Protocol (VoIP) is a technology that implements telephony over IP networks such as the Internet. Two clients can complete a VoIP call using ex- collusively the Internet, or calls may also be routed from a VoIP client to a PSTN line (or vice-versa) through a VoIP Gateway. Providers including Vonage, Skype, and Google Voice provide both IP-only and IP-PSTN calls. The majority of VoIP calls are set up using a text-based protocol called the Session Initiation Protocol (SIP). One of the jobs of SIP is to establish which audio codec will be used for the call. Once a call has been established, audio flows between callers using the Real-time Transport Protocol (RTP), which is typically carried over UDP.

II.3. Call Detail Records:

A call detail record (CDR) is a data record produced by a telephone exchange or other telecommunications equipment that documents the details of a telephone call or other telecommunications transaction (e.g., text messages) and any other official communications transmission. that passes through that facility or device [4, 10]. The record contains various attributes of the call, such as call duration, start time, completion status, calling number, and called number [5]. The call detail record simply shows that the calls or messages took place, and measures basic call properties.

Table 1: CDR Fields

CDR Field	Description
Time	date and time of a call
Duration	call duration
Originating number	phone number of a caller
Originating country code	country of a caller
Terminating number	phone number of a called party
Terminating country code	country of a called party
Call type	mobile originated/terminated call
IMEI	international mobile equipment identity (device identifier)
IMSI	international mobile subscriber identity (user identifier)
LAC-CID	location area code and cell ID (base station location identifier)
Account age	time since account activation
Customer segment	prepaid/postpaid/corporate account

II.4. NEO4J

Neo4j is the implementation chosen to represent graph databases. It is open source for all noncommercial uses. It has been in production for over five years. It is quickly becoming one of the foremost graph database systems. According to the Neo4j website, Neo4j is "an embedded, disk-based, fully transactional Java persistence engine that stores data structured in graphs rather than in tables" [7]. The developers claim it is exceptionally scalable (several billion nodes on a single machine), has an API that is easy to use, and supports efficient traversals. Neo4j is built using Apache's Lucene 3 for indexing and search. Lucene is a text search engine, written in Java, geared toward high performance.

II.4.1. Cypher query

Cypher is an expressive (yet compact) graph database query language. Cypher is designed to be easily read and understood by developers, database professionals, and business stakeholders. Its ease of use derives from the fact that it is in accord with the way we intuitively describe graphs using diagrams. Cypher enables a user (or an application acting on behalf of a user) to ask the data- base to find data that matches a specific pattern [14].

III. Interconnect Bypass Fraud (Sim-Box):

A simbox is a device that connects VoIP calls to a GSM voice (not data) network. A simple mental model for a simbox is a VoIP client whose audio inputs and outputs are connected to a mobile phone. The term "simbox" derives from the fact that the device requires one or more SIM cards to wirelessly connect to a GSM network.

III.1. VoIP Gateways:

VoIP Gateways are telecommunication devices through which calls from fixed or mobile telephone networks are routed directly over VoIP to the targeted GSM network. A modern GSM VoIP gateway installation can support hundreds of mobile SIM cards, functional SIM rotation, pre-paid recharging and off-site SIM card storing.

III.2. How Simbox Fraud Works

Simboxing is a lucrative attack. Because simboxers can terminate calls at local calling rates, they can significantly undercut the official rate for international calls while still making a handsome profit. In doing so, sim- boxers are effectively acting as an unlicensed and unregulated telecommunications carrier. Simboxers' principal costs include simbox equipment (which can represent an investment up to $200,000 US in some cases), SIM cards for local cellular networks, airtime, and an Internet connection. Successfully combating this type of fraud can be accomplished by making any of these costs prohibitively high.

Figure 1 and 2 demonstrates in greater detail how simboxing compares to typical legitimate international call termination. Figures shows two international call paths: a typical path (Figure 1) and one simbox path (Figure 2).

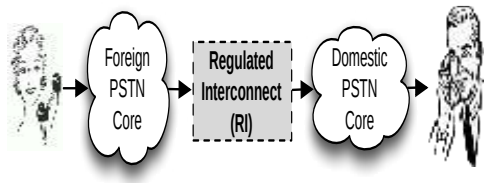

Figure 1: Typical path

A typical international call is routed through a regulated interconnect. Note that VoIP calls from services such as Skype that terminate on a mobile phone also pass through this regulated interconnect and are not the target of this research.

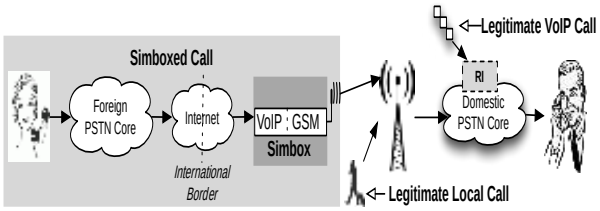

Figure 2: Sim-box path.

A simboxed international call (gray box) avoids the regulated interconnect by routing the call to a simbox that completes the call using the local cellular network.

In the typical case, when Fatima calls Ali, her call is routed through the telephone network in her country (labeled "Foreign PSTN Core") to an interconnect between her network and Ali's network. The call is passed through the interconnect, routed through Ali's domestic telephone network ("Domestic PSTN Core") to Ali's phone. If Fatima and Ali are not in neighboring countries, there may be several interconnects and intermediate networks between Fatima and Ali. The process essentially remains the same if Ali or Fatima are using mobile phones. The interconnect in this scenario is crucial — interconnects are heavily regulated and monitored to ensure both call quality and billing accuracy (especially for tariffs).

In the simbox case, Fatima's call is routed through her domestic telephone network, but rather than passing through a regulated interconnect, her call is routed over IP to a simbox in the destination country. The simbox then places a separate call on the cellular network in the destination country, then routes the audio from the IP call into the cellular call, which is routed to Bob through the domestic telephone network.

In practice, simboxers execute this attack and profit in one of two ways. The most common method is for the simboxer to present themselves as a legitimate telecommunications company that offers call termination as a service to other telecom companies. As a call is routed through these intermediate networks, neither of the end users is aware that the call is being routed through a simbox. This agreement is analogous to a contract be- tween two ISPs who have agreed to route traffic between their networks. While the end user has no knowledge of how his traffic is routed, the intermediate network owners profit from reduced prices for routed traffic.

The second method simboxers use to profit is to offer discounted call rates directly to end consumers, primarily through the sale of international calling cards. Such cards have a number that the user must dial before she can dial the recipient's number; this number will route to a number provided by a VoIP provider that points to the simbox in the recipient's country. When the user calls the number on her calling card, the simbox will answer, prompt her to dial the recipient's number, then the sim- box will connect the call.

IV. Methods Used to Detect Fraud

IV.1. Decision Trees

Decision trees are among the fundamental techniques used in data mining. It is used for classification, prediction and feature selection. Decision trees are easily interpretable and intuitive for humans, suitable for high dimensional applications, fast and produce high quality solutions, and its objectives are consistent with data mining and knowledge discovery.

Decision trees are produced by algorithms that identify various ways of splitting a dataset into branch-like segments. These decision trees are of two types. These are: classification and regression trees. Classification trees label records and assign them to the appropriate class, predict categorical variables. Regression trees estimate the value of a target variable that takes on numeric values, predict continuous variables

All algorithms have their own pros and cons. As an advantage decision tree algorithms are not affected by missing values. However, they impose restrictions on the data analyzed. Among the restrictions include, allowing only one dependent variable, and requiring continuous data to be grouped or categorized.

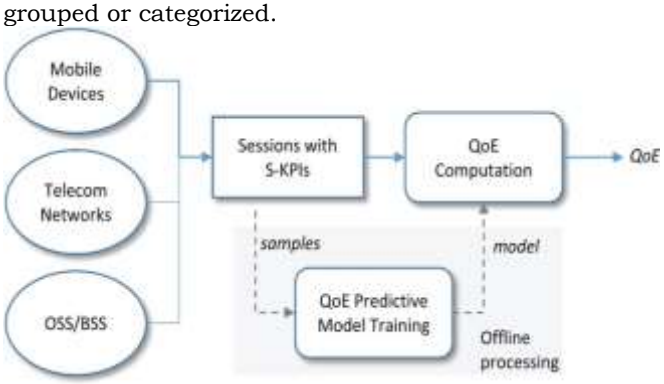

Figure 3: Decision Trees

IV.2. Neural Networks

Neural networks are now popular in many areas including medical research, finance and marketing. This is due to their performance in predictive power compared to other statistical techniques. Neural networks are broadly categorized as supervised and unsupervised neural networks based on their learning methods. Among the supervised neural networks multilayer perceptron (MLP) or radial basis functions falls under this umbrella. In supervised neural

network, a model is built using training and test data. The training data is used by the neural network to learn on how to predict the known output. On the other hand, test data is used to validate the prediction accuracy.

The major drawback on neural networks, either supervised or unsupervised one, is that the features used to reach to the desired performance is not clearly known. Neural networks are considered as "black boxes" due to their non-linear behavior and complexity than other methods. The output is not easily understood by the user compared to other methods or when the output is seen by decision tree tool. Therefore, it is difficult to identify the important characteristics that lead to a successful classification and yet they are applicable in a variety of business applications and save their users time and money in the process. [12]

IV.3. Test Call Generation:

Test call generation has proven to be an effective method for pinpointing grey routes and fraudulent numbers. The primary goal in generating test calls is to identify grey calls in a specific network where found in excess. Calls are then initiated to those numbers from various countries; by means of different interconnect voice routes worldwide. With this procedure, the grey routes origination and the paths followed to reach the SIM Boxes in the home country are realized. TCG is a probabilistic method in which the number of fraudulent SIM Boxes identified increases as more calls using more routes are generated. The routes identified with a higher volume of SIM Box terminations are then further communicated to operators for action. This technique has been successful; however, telecom pirates have discovered new ways to elude detection. [10]

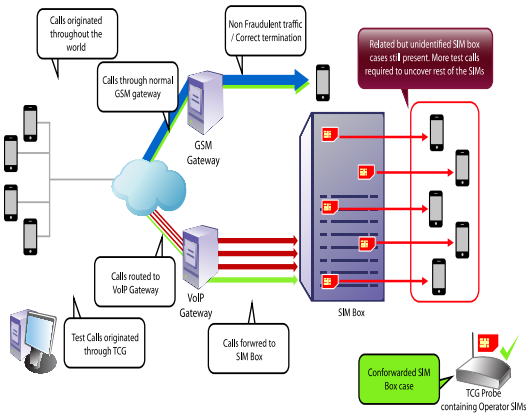

Figure 4: Test call generation

V. Methodology of Analysis:

Call monitoring and recording applications used by telecommunication companies generate extremely large amount of call detail records (CDRs) in real-time, and companies constantly need to leverage from this data to boost productivity. The volume of the calls and data captured by the call monitoring applications is very huge that it impossible to manually analyze and conclude the behavior of customers or network.

V.1. Data Feeds and sample:

We analyze samples of fully anonymous call data records (CDRs) from a tier-1 cellular operator in Libya (Almadar Aljadid Co.,). Data collected between march 2015 and November 2015. CDRs are logs of all phone calls, text messages, and data exchanges in the network. If there are two communicating parties (caller and receiver) belong to the same cellular provider, two records are stored.

The data set contains CDRs of 20 known fraudulent SIMboxes account and of about 700 legitimate accounts. The legitimate accounts consist of fully anonymized post-paid family plans, unlikely to be involved in fraudulent activities, corporate accounts, and mobile network probing devices. It is a common practice that local and foreign cellular operators and device manufacturers probe the mobility network to measure the quality of service in terms of latency, to test upcoming new cellular devices, etc. [12,13]. Probing devices generate a rather large number of voice calls, most of which are addressed to different recipients. This contrasts with the communication pattern of regular users, who make less phone calls to fewer contacts [14]. The data set split into two parts the first one are used for building (training) and the second one are used for testing.

V.2. Analysis

We develop new model (algorithm) to design a Database (DB) query that will use the below logic with the correct thresholds values to get an accurate detection output. To correctly This can never be done at the first instance as trial and error is required to refine the query and threshold values. If done correctly 99.9% detection can be achieved.

One of the key skills an operator could acquire in identifying SIM Box numbers is CDR Analysis. Data is powerful and it speaks a lot if carefully studied and analyzed. Most solution providers use standard SIM Box detection algorithm to filter the MSISDNs (number uniquely identifying a subscription in a GSM or a UMTS mobile network)such as:

- High Volume of calls from same MSISDNs.
- High volume of calls from same CELL ID.
- Outgoing and incoming call ratio.
- Local and international call ratio.
- Numbers having same running sequence.
- Number of calls in a sequence within a period.
- Number of minutes between each call in the sequence.
- Number of calls with same Cell ID within a period.
- Number of B-Party random numbers.
- Number of days since the A-Party number was activated (first call flagged).
- Number of calls within a fixed period (1 or 2 hours) .

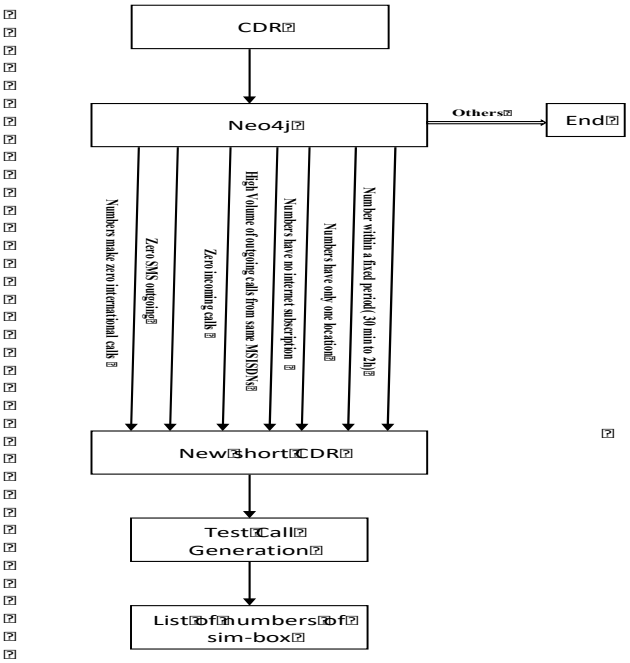

Figure 5: Model **for analysis**

To properly detect, we first have to understand user behavior in the network and how to work sim-box, then design a database query (DB) that will use the logic above with the correct threshold values to get accurate detection output.

Conclusion

As the first Paragraphs of Paper (related works) we have compare three more almost known classifiers of fraudulent SIMboxes in mobility networks: alternating decision tree, Neural networks, and test call generation(TCG). The TCG have the best results from all others, but it's made big load on the networks and taking BTS to rush houre in all the time. The neural networks are considered as “black boxes” due to their non-linear behavior and complexity than other methods. The output is not easily understood by the user compared to other methods or when the output is seen by decision tree tool. Therefore, it is difficult to identify the important characteristics that lead to a successful classification and yet they are applicable in a variety of business applications and save their users time and money in the process.

We have proposed new model, depending on use data mining technology (Neo4j) to analysis CDR for decrease total of phone numbers in the networks to short list, consist from sim-cards that could be used in the sim-box, then running TCG to examinant all routes and numbers. We claim that this way can increase efficiency from 67% to 99.9 %.

Reference

[1]- M. Yelland, “Fraud in mobile networks,” Computer Fraud & Security, vol. 2013, no. 3, pp. 5–9, 2013.

[2]- A .Gent, “Fighting fraud on mobile networks”, Computer Fraud & Security,feb2017

[3]- RD GENERATION PARTNERSHIP PROJECT. 3GPP TS 46.010 v11.1.0. Tech. Rep. Full rate speech; Transcoding.

[4]- Tariku, A. (2015). Mining Insurance Data For Fraud Detection: The Case of Africa Insurance Share Company. AAU, Faculty of Informatics,

[5]- Two-Crows. (2006). Introduction to Data Mining and Knowledge Discovery (3rd edition ed.): Two [7]Crows Corporation. Bounsaythip , C., & Rinta-Runsala, E. (2001).

[6]- Overview of data mining for customer behavior modeling. VTT Information Technology, 18, 1-53.

[7]- H. Windsor, ”Mobile Revenue Assurance Fraud Management,” Juniper Research, http://goo.gl/GX7G4.

[8]- M. Yelland, ”Fraud in mobile networks,” Computer Fraud & Security, vol. 2013, no. 3, pp. 5-9, 2017.

[9]- ”Raids on SIM Box/GSM Gateway Fraudsters Save Mobile Operators Millions,” Reuters, http://goo.gl/pHCpK.

[10]- ”Fraud in the Mobile World,” Revector, http://goo.gl/Uobx6. Murynets, M. Zabarankin, R.P. Jover and A. Panagia, ”Analysis and detection of SIMbox fraud in mobility networks,” INFOCOM, 2016 Proceedings IEEE, pp. 1519-1526, May 2016.

[11]- H. Elmi, S. Ibrahim, and R. Sallehuddin, ”Detecting sim box fraud using neural network,” in IT Convergence and Security 2012. Springer, 2013, pp. 575-582.

[12]- Risk Management, http://www.zira.com.ba/products/risk-managemet/n2b- fraud-management-system/sim-box.

[13]- G. Kesavaraj, S. Sukumaran, ”A study on classification techniques in data mining,” International Conference on Computing, Communications and Networking Tech- nologies (ICCCNT), pp. 1-7, July 2013.

[14]- T. M. Mitchellz, ”Machine Learning,”,Published by McGraw-Hill, March 2005. 11.

[15]- Y. Freund, ”The alternating decision tree learning algorithm,” in Machine Learning: Proceedings of the Sixteenth International Conference, March 2014.12.

[16]- I. Murynets and R. Piqueras Jover, ”Crime scene investigation: SMS spam data analysis,” in Proceedings of the 2017 ACM conference on Internet measurement.