Analysis and Detection of Fraud in International Calls Using Decision Tree

Ahmed Aljarray and Abdulla Abouda

Almadar Aljadid R&D Office, Libya-Misrata

Abstract. fraud is one of the most severe threats to revenue and quality of service in telecommunication networks. The advent of new technologies has provided fraudsters with new techniques to commit fraud. Subscriber identity module box (SIMbox) fraud is one of such fraud that is used in international calls and it has emerged with the use of VOIP technologies. In this paper, we propose a novel technique for detecting SIMbox fraud in international calls. The proposed technique is based in using decision tree algorithm to build a model based on six features extracted from call data record (CDR). The proposed algorithm is tested using dataset obtained from a real mobile operator (Almadar Ajadid Co.,) and it has shown 97.95% detection accuracy.

1 Introduction

Cellular network operators lose about 3% of the their annual revenue due to fraudulent and illegal services [1]. Juniper Research estimated the total losses from the underground mobile network industry to be 58 billion in 2011 [1, 2]. The impact of voice traffic termination fraud, commonly known as Subscriber Identity Module (SIMbox) fraud or bypass fraud, on mobile networks is particularly severe in some parts of the globe [2]. Recent highly publicized raids on fraudsters include those in Mauritius, Haiti, and El Salvador [3].

Fraudulent SIMboxes hijack international voice calls and transfer them over the Internet to a cellular device, which injects them back into the cellular network. As a result, the calls become local at the destination network [4]. When international call is received with the emergence of a local number on the phone screen that call should be noted as a type of fraud which causes considerable losses for the telecommunications companies. Cellular operators of the intermediate and destination networks do not receive payments for call routing and termination. Fraudulent SIMboxes also hijack domestic traffic in certain areas, e.g. in Alaska within the United States, where call termination costs are high. In some cases, the traffic is injected into a cellular network and is forwarded to the terminating country [5]. This increases the call routing cost for the operator of the injected traffic. Besides causing the economic loss, SIMboxes degrade the quality of local service where they operate. Often, cells are overloaded, and voice calls routed over a SIMbox have poor quality, which results in customer dissatisfaction. Although some vendors provide cellular anti-fraud services, the large amount of

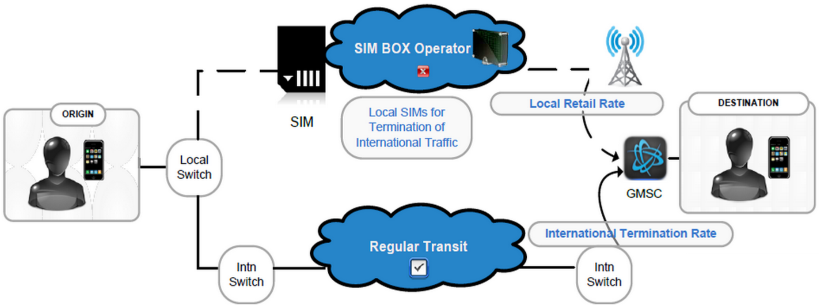


Fig. 1: Example of one-hop SIM-box bypass fraud hijacking of an international call [7]

daily cellular traffic and the number of connected mobile devices make detecting call bypassing fraud extremely challenging. Moreover, traffic patterns and characteristics of fraudulent SIMboxes are very similar to those of certain legitimate devices, such as cellular network probes. So, detecting fraudulent SIMboxes resembles searching for a few needles in a huge haystack full of small objects that look like needles. While operators of the intermediate and destination networks have high financial incentives to understand the problem, they do not have the data to analyse the international calls that are gone. Also, the absence of publicly available SIMbox related data is a major obstacle for emerging of comprehensive studies on voice bypassing fraud analysis and detection [6]. By contrast, most of the SIMbox traffic analysed in this paper is on the originating end of the communication, giving us insight on SIMbox fraud from a different perspective than most networks with a bypass problem. This work analyses fraudulent SIMbox traffic based on communication data from Almadar Aljadid company, one of the major mobile operators in Libya. It neither collects nor uses any personally identifiable information. Based on these observations, we propose using decision tree for detecting fraudulent SIMboxes. The proposed technique shows high detection rate and correctly filters out mobile network probes with traffic patterns similar to those of SIMboxes.
The rest of this paper is organized into six sections. Section II overviews fraud in international call and illustrates it with basic example. Section III presents decision tree algorithm of fraud detection in international calls. Section IV analyses SIMbox related traffic, compares it to the legitimate traffic, based on the extracted features. In Section V we describe some experiments we have performed, and Section VI concludes the paper.

2 Fraud in International Calls (Bypass Fraud)

SIMbox voice fraud occurs when the cost of terminating domestic or international calls exceeds the cost of a local mobile-to-mobile call in a particular region or country. Fraudsters make profit by offering low-cost international and sometimes domestic voice calls to other operators. To bypass call routing fees, they buy large amounts of SIM cards, install them into an off-the-shelf hardware to connect to a cellular network, which essentially becomes a SIMbox. Then the fraudsters transfer a call via the Internet to a SIMbox in the area of call recipient to deliver the call as local. As a result, the operators serving the called party do not receive the corresponding call termination fees. In other cases SIMboxes re-inject telecom voice traffic into the cellular network masked as mobile customer calls, and the operator pays for carrying the re-injected calls. Figure 1 shows example of how SIMbox bypass fraud occurs in international phone calls. For simplicity, the example assumes that there is only one intermediate hop connection between two countries. The lower path marks a legitimate path for a phone call, whereas the upper path indicates a fraudulent one when a SIMbox is in place. Actual SIMbox fraud is often more complicated, involving multiple intermediate steps. In the legitimate case, once the origin customer dials destination customer number, the call is routed through the cellular infrastructure of operator 1 to international switch (Regular Transit). Based on an agreement between operator 2 and the international switch, the call is routed to cellular core network of operator 2. The international switch pays operator 2 a fee in order to have the call terminated. Then the call is routed through cellular infrastructure of operator 2 and is delivered to destination customer. The fraud occurs when a fraudulent international switch hijacks call of origin customer and forwards it to operator 2 over the Internet (e.g. via VoIP). Then in the country of destination customer a SIMbox transforms the incoming VoIP flow into a local mobile call to the destination customer, and operator 2 loses the termination fee for the hijacked calls.

3 Decision Tree Algorithm

Machine learning is a technique which computer learns from a set of data given to it, and then it becomes able to predict the result of new data similar to the training data. The machine learning algorithm is meant to identify patterns based on different characteristics or features and then make predictions on new, unclassified data based on the patterns learned earlier. The input data is usually numerous instances of relations between the different variables or features relevant to the data.

There are various different approaches to machine learning namely decision trees, random forests, neural networks, clustering, bayesian networks, reinforcement learning, support vector machines, genetic algorithms, and many more. Decision tree learning is a method commonly used in data mining. Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. The goal is to create a model that predicts the value of a target feature based on several input features. Figure 2 shows general criteria in decision tree. A decision tree represents a multi-stage decision process, where a binary decision is made at each stage. The tree is made up of nodes and branches, with

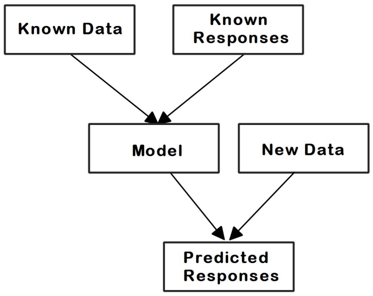


Fig. 2: General criteria in decision tree

nodes being designated as an internal or a terminal (leaf) node. Internal nodes are the ones that split into two children. Each internal node corresponds to one of the input features, there are edges to children for each of the possible values of that input feature, while terminal nodes do not have any children. A terminal node has a class label associated with it, such that observations that fall into the particular terminal node are assigned to that class. To use a decision tree, a feature vector is presented to the tree. If the value for a feature is less than a defined number, then the decision is to move to the left child. If the answer to that question is no, then we move to the right child. We continue in that manner until we reach one of the terminal nodes, and the class label that corresponds to the terminal node is the one that is assigned to the pattern. Decision tree induction algorithms are function recursively. First, a feature must be selected as the root node. In order to create the most efficient tree (i.e., smallest tree), the root node must effectively split the data. Each split attempts to pare down a set of instances (the actual data) until they all have the same classification. The best split is the one that provides what is termed the most information gain [8–10].

The tree grows by recursively splitting each node using the feature which gives the best information gain until the leaf is consistent.

Example:

Applying decision tree rules on node A for a tree model is shown in Figure 3 where M is the number of SIMbox training samples, N is the number of legitimate training samples. Four next steps are used to calculate $I.G$ for one feature with one condition:

1- Calculate entropy at node A:

$$H(S) = -\left(\frac{M}{M+N}\right)\log_2\left(\frac{M}{M+N}\right) - \left(\frac{N}{N+M}\right)\log_2\left(\frac{N}{N+M}\right) \quad (1)$$

2- The data set is split into two branches by different feature, the entropy for each branch is calculated:

$$H_a = H(m,n)$$

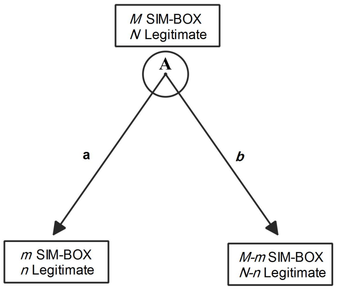


Fig. 3: Model example

$$H_a \quad = \quad -\left(\frac{m}{m+n}\right)\log_2\left(\frac{m}{m+n}\right) \quad - \quad \left(\frac{n}{n+m}\right)\log_2\left(\frac{n}{n+m}\right) \tag{2}$$

$$H_b = H(M-m, N-n)$$

$$\begin{aligned} H_b = -\left(\frac{M-m}{(M-m)+(N-n)}\right)\log_2\left(\frac{M-m}{(M-m)+(N-n)}\right) \\ -\left(\frac{N-n}{(N-n)+(M-m)}\right)\log_2\left(\frac{N-n}{((N-n)+(M-m)}\right) \end{aligned} \tag{3}$$

3- The entropy for each branch is added proportionally to get total entropy for the split:

$$H(S|A) = P_a H_a + P_b H_b$$

$$H(S|A) = \left(\frac{m+n}{M+N}\right)H_a + \left(\frac{(M-m)+(N-n)}{(M+N)}\right)H_b \tag{4}$$

where P_a is the number of samples at node (a) per the number of samples at node (A), P_b is the number of samples at node (b) per the number of samples at node (A).
4- The resulting entropy is subtracted from the entropy before the split and the result is the information gain or decrease in entropy:

$$I.G(S,A) = H(S) - H(S|A) \tag{5}$$

Table 1 summarises the decision tree algorithm specialized to learning boolean-valued functions. Decision tree is a greedy algorithm that grows the tree top-down. At each node selecting the features that best classifies the local training samples. This process continues until the tree perfectly classifies the training samples, or all features have been used [11].

Table 1: Summary of decision tree algorithm

```
Decision tree (data samples, target-feature, features-list)

Data samples are the training samples (building samples). Target-feature is the feature whose value is to be predicted by the tree. Feature-list is a list of other features that may be tested by the Decision tree.

Create a Root node for the tree

  – If all samples are SIMbox, Return the single-node tree Root, with label =SIMbox
  – If all samples are Legitimate, Return the single-node tree Root, with label = Legitimate
  – If features-list are empty, Return the single-node tree Root, with label = most common value of Target-feature in samples
  – Otherwise Begin
      • A is the feature from features-list with condition that gives best classifies samples with best(the feature that gives the biggest I.G)
      • The decision feature for Root is A
      • For each possible value, $v_i$, of A,
          ∗ Add a new tree branch below Root, corresponding to the test A = $v_i$
          ∗ Let $samples_{v_i}$ be the subset of samples that have value $v_i$ for A
          ∗ If samples of $v_i$ is empty
              · Then below this new branch add a leaf node with label = most common value of Target-feature in samples.
              · Else below this new branch add the subtree Decision tree($samples_{v_i}$, Target-feature, features-list without A).

End.

Return Root
```

4 SIMbox Fraud Analysis

4.1 Data feeds

We analyse samples of fully anonymous call data records (CDRs) from a tier-1 cellular operator in Libya (Almadar Aljadid Co.,). Data collected between October 2014 and November 2014. 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.

4.2 Data sample

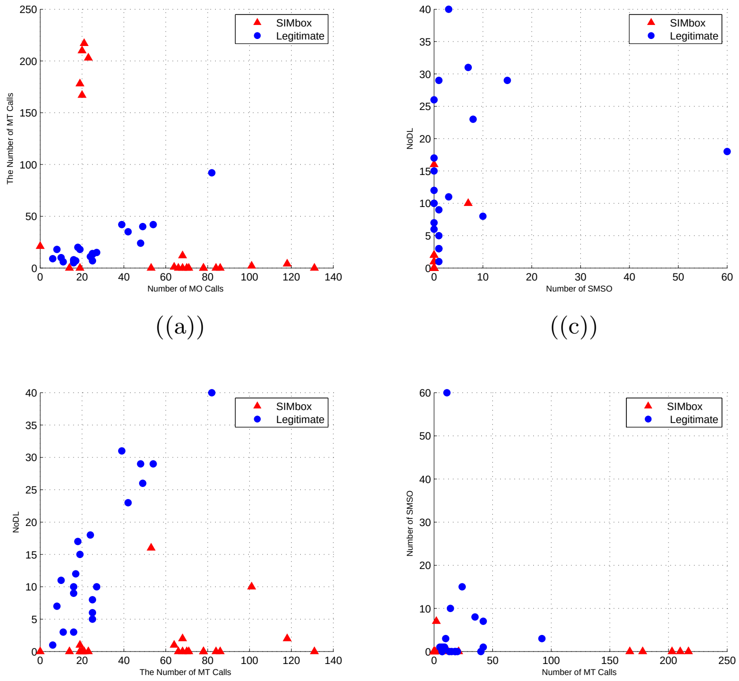
The data set contains CDRs of 34 known fraudulent SIMboxes account and of about 273 legitimate accounts. The legitimate accounts consist of fully anonymized post-paid family plans, unlikely to be involved in fraudulent activities, corporate accounts, and mobile network probing devices. It is a common 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.

4.3 Call traffic feature

CDR fields (collected during five days in 2014) are transformed into 6 features characterizing voice call communication patterns of legitimate and fraudulent users. The six features are: The total number of outgoing and incoming calls are counted based on MO and MT, the total number of SMS originating and SMS terminating, the total number of hand over and the total number of different locations (NoDF) number of calls have different between first and last location at the same call and summing with number of calls that have different between last location of call and first location of the next call. Customer details are obtained from the corresponding CDR fields.

4.4 SIMbox data Analysis

This sub-section analysis the traffic characteristics of fraudulent SIMboxes based on the features described in the previous subsection (Section 4.3).
Figure 4 a plots the number of MO calls versus the number of MT calls. It can be noticed that most of SIMboxes are clustered around two areas without any legitimate account and legitimate accounts are clustered around another area. It can be observed that most of SIMboxes have originating calls more than terminating ones while legitimate accounts have comparable number of originating and terminating calls. That is because SIMboxes are used mainly to regenerate the calls received from the VOIP branch and make them GSM calls again. This feature is very useful to distinguish between SIMboxes and legitimate accounts.
Figure 4 b present the number of MT calls versus the number of different locations (NoDL). It can be clearly seen that legitimate accounts have large number of terminating calls and higher mobility than SIMbox accounts. This is due to the fact that legitimate users are usually not tight to a specific location while SIMboxes are installed to one location and could be moved from time to time. This feature is very attractive to utilize in order to detect SIMbox accounts.
Figure 4 c plots number of SMS originating (SMSO) versus NoDL. The number of locations feature has split most of samples and here it has been used with SMS originating. We can notice that most of SIMboxes have a small number of SMSO bounded in a small level but legitimate accounts have number of SMSO larger than number of SMSO of SIMboxes.
Figure 4 d plots the number of MT calls versus the number of SMSO. It can be seen that legitimate and SIMboxes accounts have similar behaviour and it is



((a))

((c))

((b))

((d))

Fig. 4: Analysis the traffic characteristics of fraudulent SIMboxes based on the features

hard to distinguish between them based on this feature.
Based on the analysis above we can conclude that from the six explored feature the number of locations feature can give the highest distinguish rate between SIMboxes and legitimate accounts. In other words the number of locations feature results in the highest information gain and therefore, it should be used in the first stage.

5 Experimental results

The practical performance of the decision tree algorithm described in the previous section was tested using another data sample (that used for testing) that consist of 12 samples of SIMboxes and 251 samples of legitimate accounts. According to the information gain measure, the Number of different locations provides the best prediction of the target feature (kind of account) over the training samples. Therefore, the number of different locations is selected as the decision feature for the root node, and branches are created below the root for each of its

Table 2: Information Gain for the features at each node

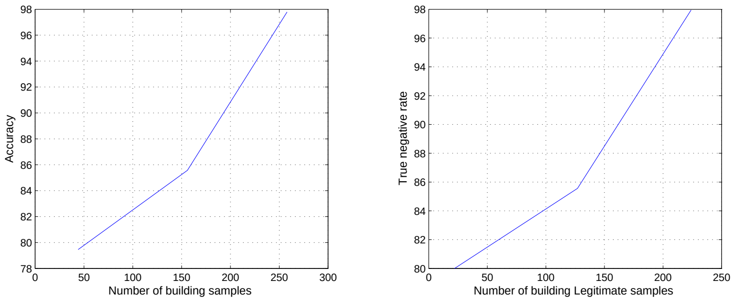
MO	MT	SMSO	SMST	NoDL	node
0.052596	0.207557	0.097026	0.088234	0.276763	Root(R)
0.514704	0.245623	0.118183	0.133216	0.181276	R-Left(L)
0.066197	0.136376	0.040580	0.174136	0.072861	R-L-L
0.311689	0.311689	0	0	0.141619	R-L-L-L
0.027740	0.257678	0	0	0.242697	R-L-L-L-L
0.144484	0.078982	0	0	0.144484	R-L-L-L-L-L
0.122556	0.122556	0	0	0.811278	R-L-L-L-L-L-r
0.093531	0.111687	0.138122	0.185579	0.012461	R-L-r
0.970950	0.970950	0.321928	0.170950	0	R-L-r-r
0.013723	0.053982	0.002601	0.006265	0.008751	R-r
0.918295	0.251629	0.251629	0.918295	0.918295	R-r-L

possible values. Table 2 summarises the information gain for the six features at each node. where R is root node, L is a node on the left and r is a node on the right. There are two types of testing to determine the accuracy of the algorithm, true negative rate test and true positive rate test. The true negative rate test is the proportion of legitimate accounts classified as legitimate (It's inverse of The false positive rate), whereas true positive rate is the proportion of SIMboxes classified as SIMbox accounts (It's inverse of the false negative rate).

Figure 5 a shows the prediction accuracy of the proposed algorithm as a function of number of building samples. It can be clearly seen that as the number of samples increases the accuracy of the algorithm improves. When the full number of samples were used the classification accuracy has reached 97.95%. Figure 5 b shows true negative rate versus the number of legitimate building samples when using decision tree algorithm to predict status of the SIM-Card. It can be seen that the prediction accuracy improves with changing the number of samples for legitimate users. The improvement is due to the fact increasing the number of legitimate building samples improves the understanding of the of behaviour of legitimate users.

6 Conclusions

In this paper six features extracted from CDR data are utilized to build decision tree that can be used to distinguish between legitimate and SIMbox accounts. The features include the total number of outgoing and incoming calls, the total number of SMS originating and SMS terminating, the total number of hand over and the total number of different locations. The proposed decision tree algorithm has shown accuracy up to 97.95% when it was tested using testing samples data from Almadar Aljadid company.



((a)) ((b))

Fig. 5: Total Accuracy of Algorithm and True Negative rate of Algorithm

References

1. H. Windsor, "Mobile Revenue Assurance Fraud Management," Juniper Research, http://goo.gl/GX7G4.
2. M. Yelland, "Fraud in mobile networks," *Computer Fraud* & *Security*, vol. 2013, no. 3, pp. 5-9, 2013.
3. "Raids on SIM Box/GSM Gateway Fraudsters Save Mobile Operators Millions," Reuters, http://goo.gl/pHCpK.
4. "Fraud in the Mobile World," Revector, http://goo.gl/Uobx6.
5. I. Murynets, M. Zabarankin, R.P. Jover and A. Panagia, "Analysis and detection of SIMbox fraud in mobility networks," *INFOCOM, 2014 Proceedings IEEE*, pp. 1519-1526, May 2014.
6. A. 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.
7. N2B Risk Management, http://www.zira.com.ba/products/risk-managemet/n2b-fraud-management-system/sim-box.
8. G. Kesavaraj, S. Sukumaran, "A study on classification techniques in data mining," *International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1-7, July 2013.
9. Wendy L. Martinez , Angel R. Martinez, "Computational Statistics Handbook with MATLAB,", 2002.
10. T. M. Mitchellz, "Machine Learning,",Published by McGraw-Hill, March 1997.
11. Y. Freund, "The alternating decision tree learning algorithm," *in Machine Learning: Proceedings of the Sixteenth International Conference*, March 1999.
12. I. Murynets and R. Piqueras Jover, "Crime scene investigation: SMS spam data analysis," *in Proceedings of the 2012 ACM conference on Internet measurement.* ACM, pp. 441-452, 2012.
13. "RCATS - Remote Cellular Active Test System," JDSU, http://goo.gl/VEbMA.
14. A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509-512, 1999.