Simulating SIMBox frauds for detection investigation

Anne Josiane Kouam, Aline Carneiro Viana, Alain Tchana

Simulating SIMBox frauds for detection investigation

Anne Josiane Kouam
INRIA, France
anne-josiane.kouam-djuigne@inria.fr

Aline Carneiro Viana
INRIA, France
aline.viana@inria.fr

Alain Tchana
Grenoble INP, France
alain.tchana@grenoble-inp.fr

ABSTRACT

SIMBox fraud is one of the most prevalent scams in cellular networks that cause a significant monetary loss to mobile operators. It consists of diverting international voice traffic from regulated routes and rerouting it as local calls in the destination country. *SIMBox* fraud mitigation is challenging as it requires knowledge from both regular and fraudulent parties. This paper identifies and addresses such challenges through the realistic design of a simulator, i.e., *FraudSIM*, enabling the generation of datasets enriched with regular and multiple fraudulent strategies-induced behaviors.

CCS CONCEPTS

• **Security and privacy → Mobile and wireless security**; • **Computing methodologies → Simulation environments**.

KEYWORDS

Cellular networks, CDRs, International bypass fraud

ACM Reference Format:
Anne Josiane Kouam, Aline Carneiro Viana, and Alain Tchana. 2022. Simulating SIMBox frauds for detection investigation. In *CoNEXT Student Workshop 2022 (CoNEXT-SW '22), December 9, 2022, Roma, Italy*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3565477.3569161

1 PROBLEM STATEMENT

SIMBox fraud consists of diverting the international voice traffic from the regulated routes through VoIP established links [4]. The diverted traffic is received at the level of a *SIMBox* (VoIP to GSM gateway) in the destination country and re-originated as a national mobile call to its recipient. Hence, destination mobile operators perceive national termination fees instead of international ones, which are much higher. The impact of this problem is enormous, affecting states' and operators' revenues, with a loss estimated to USD 3.11 Billion in 2021 [2]; but also, network quality, national security, and research. Yet, the *SIMBox* fraud mitigation remains little tackled by researchers: We identified only 14 fraud detection approaches in the literature since 2011, which is relatively low for an actual security problem of this importance. This is due to the following challenges:

The lack of raw cellular traffic datasets for analysis. First, due to the privacy of network users' traces (e.g., Call Data Records, i.e., CDRs or Audio records), *data access requires collaborations with operators, which is often restricted.* In particular, CDRs describe time-stamped and geo-referenced event types (i.e., data, calls, SMS) generated by mobile devices interacting with operator networks. They disclose sensitive information of users' habits, hardening their shareability [5]. Second, when available, traces are generally aggregated, for privacy compliance, limiting exploitation. Third, some frauds involve multiple national operators and having these operators' CDRs at the same period is barely attainable. As a result, such frauds are still unexplored in the literature to our knowledge.

Rapid fraud strategies evolution. *SIMBox* fraud evolves at a significant pace. Fraudsters create and refine strategies to evade existing detection solutions *by mimicking human communication behavior in terms of traffic and mobility*. Hence, the existing literature is outdated and today's fraud detection challenges will not be the same in a few years.

Limited fraudulent ground truth. Ground truth describes the known fraudulent or legitimate users in cellular datasets. Mobile operators generally own no or a low percentage of fraudulent ground-truth compared to the total amount of users in the trace. The remaining large percentage of users is considered legitimate. Thus, detection approaches built from such limited ground truth likely cause many false negatives.

A wide variety of *SIMBox* brands and fraud strategies. There are many *SIMBox* manufacturers on the international market, and each *SIMBox* brand has its specificities and fraud possibilities. Hence, many fraud strategies are likely to exist, making hard the validaton of a detection method. Indeed, researchers may be unaware of the latest developments, and *the effectiveness of a detection solution may vary from one fraud strategy to another*. Furthermore, *acquiring and deploying a single SIMBox architecture demands many resources.*

2 HOW WE TACKLE THE PROBLEM

We propose the design of a realistic open-source *SIMBox* fraud simulator, i.e., *FraudSIM*, as a first-in-the-literature contribution to ease research on the fraud detection.

Overview. *FraudSIM* allows (i) *the scalable simulation of numerous real fraud strategies inspired by actual market SIMBox features and acquired SIMBox equipment*, (ii) the realistic implementation of a cellular network infrastructure configurable in multiple scenarios involving multi-operators, legitimate users, and fraud architectures of varying sizes and locations, and (iii) yields CDRs traces of legitimate and fraudulent user interactions. Related to (i), *FraudSIM* design is drawn from a thorough review of all *SIMBox* equipment from the six major *SIMBox* manufacturers in the international market. It thus provides the flexibility to reproduce and investigate innumerable *SIMBox* frauds without purchasing hardware or setting up running architectures. Besides, *FraudSIM* modularity facilitates its

CoNEXT-SW '22, December 9, 2022, Roma, Italy

ACM ISBN 978-1-4503-9937-1/22/12…$15.00
https://doi.org/10.1145/3565477.3569161

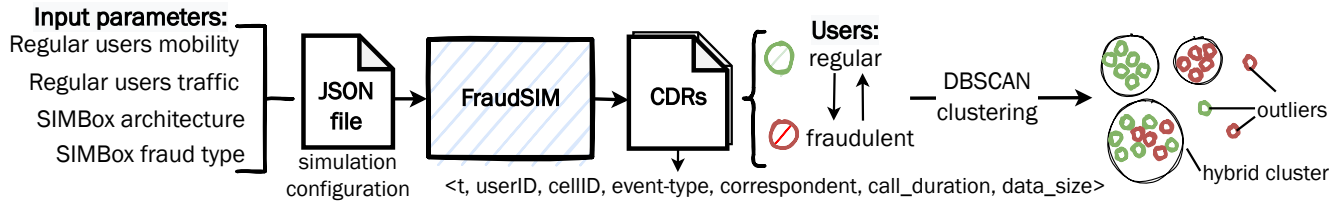

Figure 1: *FraudSIM* design and validation pipeline.

extension with new *SIMBox* functionalities from *SIMBox* evolution or forecasting purposes. *FraudSIM*, therefore, provides key elements to unblock *SIMBox* mitigation challenges above-mentioned.

Design. *FraudSIM* is constituted by four modules each performing a key role in the simulation: the *Simulator*, the *NetworkManager*, the *TrafficManager*, the *MobilityManager*.

As depicted in Fig. 1, a *FraudSIM* simulation is done by reading a user configuration file providing simulation parameters (i.e., duration, number of operators, etc.) along with choices of the regular users' traffic and mobility behaviors, and the fraudulent strategy to be executed. From these inputs, a simulation is done as follows: (i) The *Simulator* configures the simulation duration and the stop time accordingly. (ii) The *NetworkManager* creates the network infrastructure (i.e., network cells) and mobile operators entities. (iii) It then defines traffic and mobility models of legitimate users, which are responsible for the timely generation of regular users' network events (calls, SMS, or data), incoming international calls, and handovers (i.e., movements through network cells). (iv) Then, the *NetworkManager* creates all legitimate User Equipment (UEs) and SIM cards implementing these models through the *TrafficManager* and the *MobilityManager*, respectively. (v) Next, regarding the fraud simulation, the *NetworkManager* creates the *SIMBox* architecture (i.e., equipments and fraudulent SIM cards), and sets its running policies given by the chosen fraud strategy. This latter determines the communication behavior of simulated fraudulent users. (vi) Finally, the *Simulator* launches and manages the chronological execution of events scheduled through the steps above.

FraudSIM is implemented in C++ and comprises approximately 19,000 code lines.

3 HOW TO PERFORM VALIDATION

The validation of *FraudSIM* consists of confirming its ability to simulate fraud strategies of different levels in terms of mimicking human behavior. Indeed, the more a fraud strategy induces fraudulent SIM card behaviors close to real human ones, the more the strategy is efficient, i.e., difficult to detect. To this end, we propose the implementation of five representative fraud strategies, described in Table 1, allowing to evaluate the significance of each traffic behavioral feature for the fraud's efficiency.

We then define a metric, namely *blending_metric*, evaluating how well fraudulent users can blend in the crowd of legitimate ones. This metric is obtained by performing a clustering to simulated fraudulent and legitimate users with respect to the set of features used in the literature for *SIMBox* fraud detection (e.g., number of calls per day [6], number of unique cell Ids [1], etc.). As hybrid clusters indicate a behavior similarity between fraudulent and legitimate users, we define *blending_metric* as *the number of fraudulent users in a hybrid cluster over the total number of fraudulent users*. For instance, in Fig. 1, the value of *blending_metric* is $\frac{4}{11} = 0.36$. Table 1 reports preliminary results of *blending_metric* obtained for a simulation with 6000 legitimate users from a real-world anonymized CDRs and 50 fraudulent ones, in which we consider 3% of incoming international calls. Specifically, we apply a DBSCAN clustering, known to be extremely efficient in detecting outliers and not requiring to specify a number of clusters beforehand, as in K-means. We can see that the fraudulent traffic behavior is more impactful than other behavioral features when the number of fraudulent users is low. Also, *all_advanced* strategy yields all fraudulent users indistinguishable from legitimate ones.

Table 1: Sample fraud strategies description

Fraud strategy	Description	*Blending metric*
All_naive	No human behavior mimicking: - only outgoing calls all day long - stationary fraudulent users - no max. number of contacts	0
Advanced_traffic	Naive mobility and social behaviors Human traffic behavior mimicking: - all event types: calls, data, SMS - no traffic at night periods	0.96
Advanced _mobility	Naive traffic and social behaviors Human mobility behavior mimicking: - mobility given by the realistic Working Day Mobility model [3] - neighborhood movements in the evening	0
Advanced_social	Naive traffic and mobility behaviors Human social behavior mimicking: - history based call routing - max. number of contacts	0
All_advanced	Includes all advanced config. of each behavioral feature	1

4 WHAT WE PLAN NEXT

FraudSIM will be released open-source to ease and promote research on SIMBox fraud mitigation. We believe such a tool is indispensable for research in this field where data is intrinsically private. In future works, we plan to deepen *FraudSIM* validation by analyzing *blending_metric* while varying the number of fraudulent users

and simulation scenarios. Moreover, we will implement literature *SIMBox* fraud detection approaches and assess their performance and limitations against the above fraud strategies through a comprehensive evaluation given by multiple parameters. We will then define a more resilient detection approach.

REFERENCES

[1] M. R. AlBougha. 2016. *Comparing Data Mining Classification Algorithms in Detection of Simbox Fraud.* Master's thesis. St. Cloud State University.

[2] Communications Fraud Control Association. 2021. *Fraud Loss Survey.* Technical Report. https://cfca.org/wp-content/uploads/2021/12/CFCA-Fraud-Loss-Survey-2021-2.pdf

[3] F. Ekman, A. Keränen, J. Karvo, and J. Ott. 2008. Working Day Movement Model. In *Proceedings of ACM SIGMOBILE Mobility Models Workshop.*

[4] A. J. Kouam, A. C. Viana, and A. Tchana. 2021. SIMBox Bypass Frauds in Cellular Networks: Strategies, Evolution, Detection, and Future Directions. *IEEE Communications Surveys & Tutorials* 23, 4 (2021), 2295–2323.

[5] Yves-Alexandre Montjoye, Cesar Hidalgo, Michel Verleysen, and Vincent Blondel. 2013. Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific reports* 3 (03 2013), 1376. https://doi.org/10.1038/srep01376

[6] R. Sallehuddin, S. Ibrahim, A. Zain, and A. Elmi. 2015. Detecting SIM Box Fraud by Using Support Vector Machine and Artificial Neural Network. In *Jurnal Teknologi.*