Analysis and Detection of SIMbox Fraud in Mobility Networks

Ilona Murynets*, Michael Zabarankin+, Roger Piqueras Jover* and Adam Panagia‡
*AT&T Security Research Center, New York, NY
+Stevens Institute of Technology, Hoboken, NJ
‡AT&T Financial Billing Operations, Picataway, NJ
*{ilona,roger.jover}@att.com, +mzabaran@stevens.edu, ‡adam.panagia@att.com

***Abstract*—Voice traffic termination fraud, often referred to as Subscriber Identity Module box (SIMbox) fraud, is a common illegal practice on mobile networks. As a result, cellular operators around the globe lose billions annually. Moreover, SIMboxes compromise the cellular network infrastructure by overloading local base stations serving these devices. This paper analyzes the fraudulent traffic from SIMboxes operating with a large number of SIM cards. It processes hundreds of millions of anonymized voice call detail records (CDRs) from one of the main cellular operators in the United States. In addition to overloading voice traffic, fraudulent SIMboxes are observed to have static physical locations and to generate disproportionately large volume of outgoing calls. Based on these observations, novel classifiers for fraudulent SIMbox detection in mobility networks are proposed. Their outputs are optimally fused to increase the detection rate. The operator's fraud department confirmed that the algorithm succeeds in detecting new fraudulent SIMboxes.**

I. Introduction

Cellular network operators lose about 3% of the annual revenue due to fraudulent and illegal services. 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 (SIM)-box 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], and the cellular operators of the intermediate and destination networks do not receive payments for the 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. This increases the call routing cost for the operator of the injected traffic.

Besides causing the economic loss, SIMboxes degrade the 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 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 analyze 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 [5]. By contrast, most of the SIMbox traffic, analyzed 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.

In this work, we analyze fraudulent SIMbox traffic based on anonymized communication data from one of the major tier-1 network operators in the United States. SIMboxes operate with a large number of SIM cards from foreign and national operators. If an operator detects and shuts down a fraudulent account, the fraudsters deploy a set of new SIM cards as in the Short Message Service (SMS) spam fraud [6]. Also, fraudulent SIMboxes have almost static physical locations and generate disproportionately large number of outgoing calls (100 times as many as incoming calls). Based on these observations, we introduce three classifiers for fraudulent SIMboxes and combine their outputs into a classification rule, which has a high detection rate and correctly filters out mobile network probes with traffic patterns similar to those of SIMboxes.

This paper is organized into five sections. Section II overviews voice termination fraud in mobility networks and illustrates it with some basic examples. Section III analyzes SIMbox related traffic, compares it to the legitimate traffic, and, based on the extracted features, presents a novel algorithm for SIMbox detection in mobility networks. Section IV overviews the related work, and Section V concludes the paper.

II. Voice fraud in mobile networks

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.

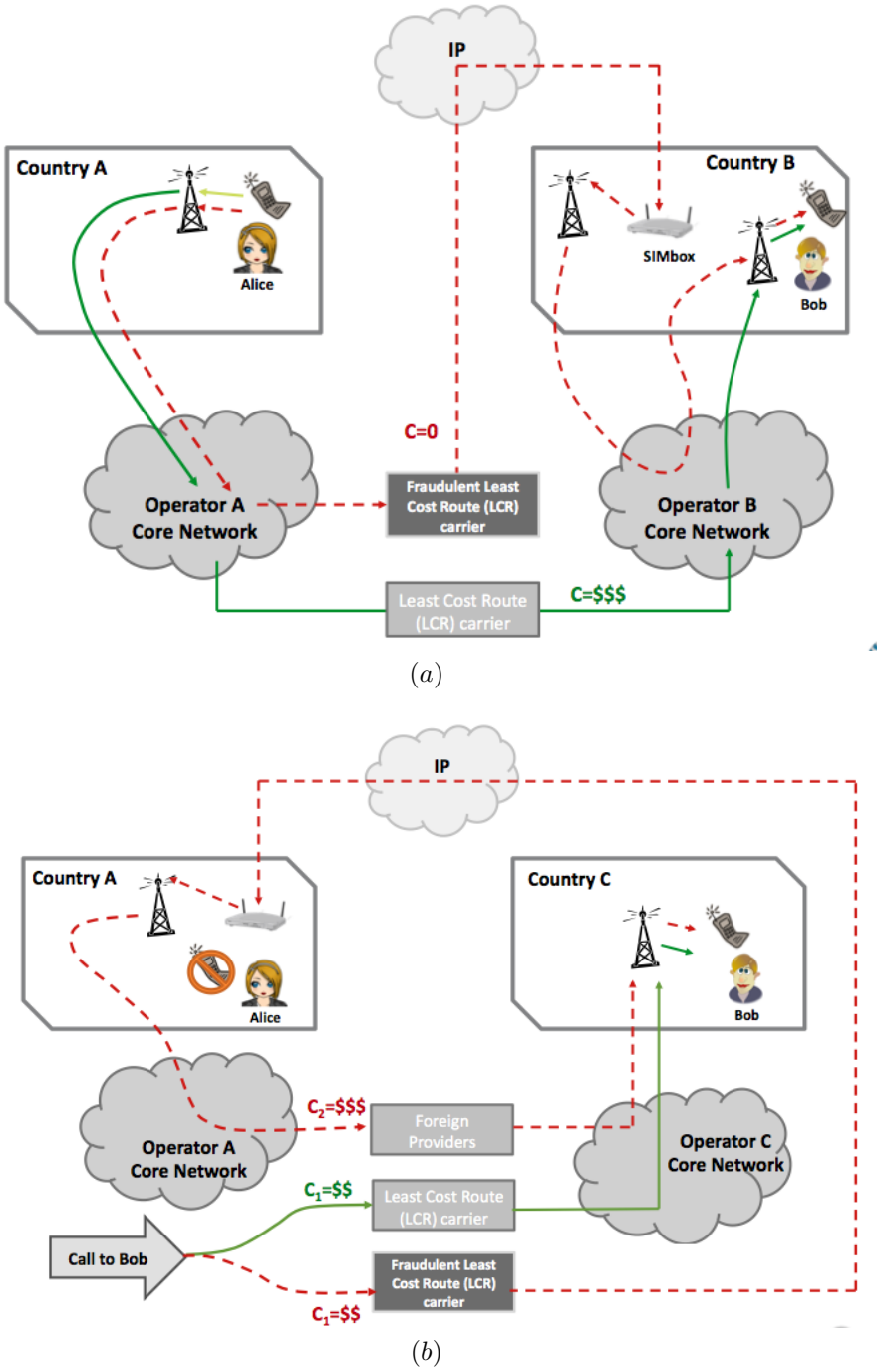

Fig. 1. Two examples of one-hope SIM-box bypass fraud: (a) hijacking of an international call (b) hijacking and re-injecting of an international call.

Fraudsters make profit by offering low-cost international and sometimes domestic voice calls to other operators. To bypass call routing fees, they buy or hijack large amounts of SIM cards, install them into an off-the-shelf hardware[1] to connect to the mobility 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 mobility network masked as mobile customer calls, and the operator pays for carrying the re-injected calls.

Figure 1 shows two examples of how SIMbox bypass fraud occurs for international phone calls. For simplicity, the examples assume that there is only one intermediate hop connection

[1]The hardware can be used for legitimate purposes, for example in machine-to-machine (M2M) applications. It has been recently reported to transmit SMS spam [7].

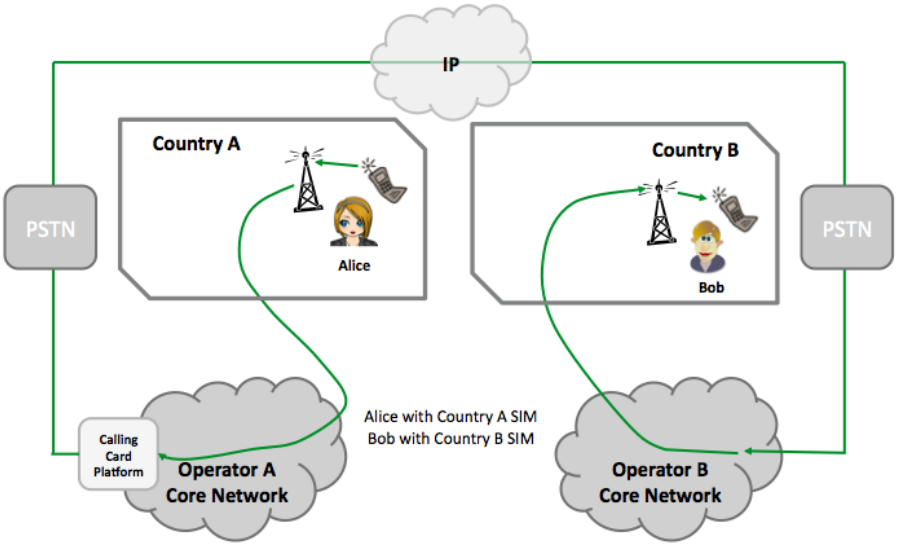

Fig. 2. Legitimate International calling card scheme

between two pairs of countries: country A to country B and country A to country C. The solid line marks a legitimate path for a phone call, whereas the dotted line indicates a fraudulent one when a SIMbox is in place. Actual SIMbox fraud is often more complex, involving multiple intermediate steps.

In Figure 1(a), Alice, who lives in country A, calls Bob, who lives in country B. In the legitimate case, once she dials Bob's number, the call is routed through the cellular infrastructure of operator A to a *least cost route* (LCR) carrier. Based on an agreement between operator B and the LCR carrier, the call is routed to operator B's cellular core network. The LCR carrier pays operator B a fee in order to have the call terminated. Then the call is routed through the operator B's cellular infrastructure and is delivered to Bob.

The fraud occurs when a fraudulent LCR carrier hijacks Alice's call and forwards it to country B over the Internet, e.g. via VoIP. Then in country B, a SIMbox (an associate of the fraudulent LCR carrier) transforms the incoming VoIP flow into a local mobile call to Bob, and the operator B looses the termination fee for the hijacked call.

Figure 1(b) shows a more elaborate SIMbox fraud scheme, which consists of three stages. In the first stage, the fraudsters hijack legitimate SIM cards from customers of operator A and put them into a SIMbox. For example, they call Alice and trick her into providing her account information. Then they impersonate Alice and link her wireless account to their own SIM card. As a result, Alice's phone will be unable to connect, whereas the SIMbox traffic will be charged to her account.

In the second stage, suppose that Bob lives in country C and has a SIM card from operator C. International roaming agreements and standard industry practices specific for country C prescribe to pass call traffic from either other network operators or the Public Switched Telephone Network (PSTN) (low priority traffic) through an LCR carrier and to pass the traffic of operator A's retail customers (high priority traffic) via foreign operators (high cost routes). Suppose a low priority call, originating in country A, is intended to reach Bob. In the legitimate case, this call is routed over an LCR and is terminated in operator C's network. In this stage, the fraud occurs when an illegitimate LCR routes the traffic over IP to the SIMbox in country A.

Finally, in the third stage, the SIMbox injects the traffic destined to Bob into the cellular network of operator A. At this point, the communication becomes a wireless call from the SIM card of retail customer Alice, so that the communication is routed to country C over a high cost foreign provider. In this case, operator C always receives the call termination fee either from the legitimate LCR or from the foreign providers, whereas when Alice reports her stolen identity, operator A becomes liable for the cost of the call.

The following example illustrates the difference between legitimate international calling cards (PennyTalk [8], ZapTel [9], Vonage [10], etc.) and SIMbox fraud that reroutes voice calls via VoIP. Suppose that Alice, who is in country A, purchased an international calling card to call Bob in country B (see Figure 2). She dials either a local or toll free number that connects her to the calling card platform through operator A's cellular network. The platform requests and verifies Alice's card access code, and as soon as she enters Bob's phone number, it forwards her call over the PSTN to country B, and Bob receives a local low cost call connecting him to Alice. A portion of the call path might be routed over IP. When Alice uses the international calling card, she is aware that her call will be routed via VoIP, and so, she agrees to get a low call quality. In this case, despite no termination fees, the operators of Alice's and Bob's providers make profit from the legitimate

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

TABLE I
VOICE CALL DETAIL RECORD FIELDS

agreement between them and the calling card.

Yet another SIMbox fraud is bypassing interconnection fees in a high cost domestic traffic, which uses the schemes described in the above examples.

III. SIMBOX FRAUD ANALYSIS AND DETECTION

A. Data feeds and anonymization

We analyze samples of fully anonymized *call detail records* (CDRs) from a tier-1 cellular operator in the United States between February 2012 and June 2013. CDRs are logs of all phone calls, text messages, and data exchanges in the network. If two communicating parties (caller and receiver) belong to the same cellular provider, two records are stored. The *mobile originated* (MO) and *mobile terminated* (MT) records store data of the caller and receiver, respectively.

At a CDR pre-processing step, all individual identifiers are removed: caller and receiver phone numbers are replaced by integer hashes (anonymized), the *international mobile subscriber identity* (IMSI) is parsed and hashed, and only the first 8 digits of the *international mobile equipment identity* (IMEI) are preserved and anonymized.[2] Table I summarizes the CDR fields used in the analysis.

All results are normalized and aggregated to obfuscate total counts of calls and the number of devices in the network.

[2]This first segment of the IMEI, known as the Type Allocation Code (TAC), determines the manufacturer and model of a wireless device. For a phone, the TAC identifies the phone manufacturer and model (e.g. Samsung Galaxy S4), whereas for an M2M connected device, it identifies the embedded cellular modem (e.g. Sierra Wireless Q2687).

B. Data sample selection and labeling

The data set contains CDRs of 500 IMEIs of fraudulent SIMboxes and of about 93000 legitimate accounts. The fraudulent SIMbox accounts were investigated by the operator's fraud department and cancelled due to their malicious activity. The legitimate accounts consist of fully anonymized post-paid family plans, unlikely to be involved in fraudulent activities [7], 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. [11]. 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 [12]. Also, probing devices reuse the same IMSI for multiple physical devices as part of the probing infrastructure. This results in CDRs from each IMSI originating in multiple locations.

The data set is split into two parts: 66% and 34% of the labeled accounts are used for the training and testing, respectively.

C. Call traffic features

CDR fields in Table I (collected during one week in 2013) are transformed into 48 features characterizing voice call communication patterns of legitimate and fraudulent IMEIs. Features such as average MO and MT call durations, account age, customer segment are obtained from the corresponding CDR fields. The total number of outgoing and incoming calls along with their corresponding destinations and origins (international and domestic) are counted based on MO and MT time stamps and based on originating and terminating country codes. Since SIMboxes typically use multiple SIMs, one of the features is the number of IMSIs operated per IMEI, counted for both one week in 2013 and the period of February 2012 – June 2013.

The geo-location feature is the number of base stations (MO and MT) that IMEI connected to during that week and is obtained from the LAC-CID field. Some features are derived from the others, e.g. the ratio of the number of destination to the total number of calls, the ratio of international calls to the total number of calls, etc.

D. SIMbox data analysis

This sub-section analyzes the traffic characteristics of fraudulent SIMboxes based on the features described in Section III-C. Figure 3(a) plots the number of MO calls versus the number of locations from which these calls originate. Dots and triangles correspond to legitimate and fraudulent SIMbox accounts, respectively. To avoid disclosing sensitive information, the four plots are normalized with the same arbitrary positive integer. Figure 3(a) shows that SIMboxes are physically static

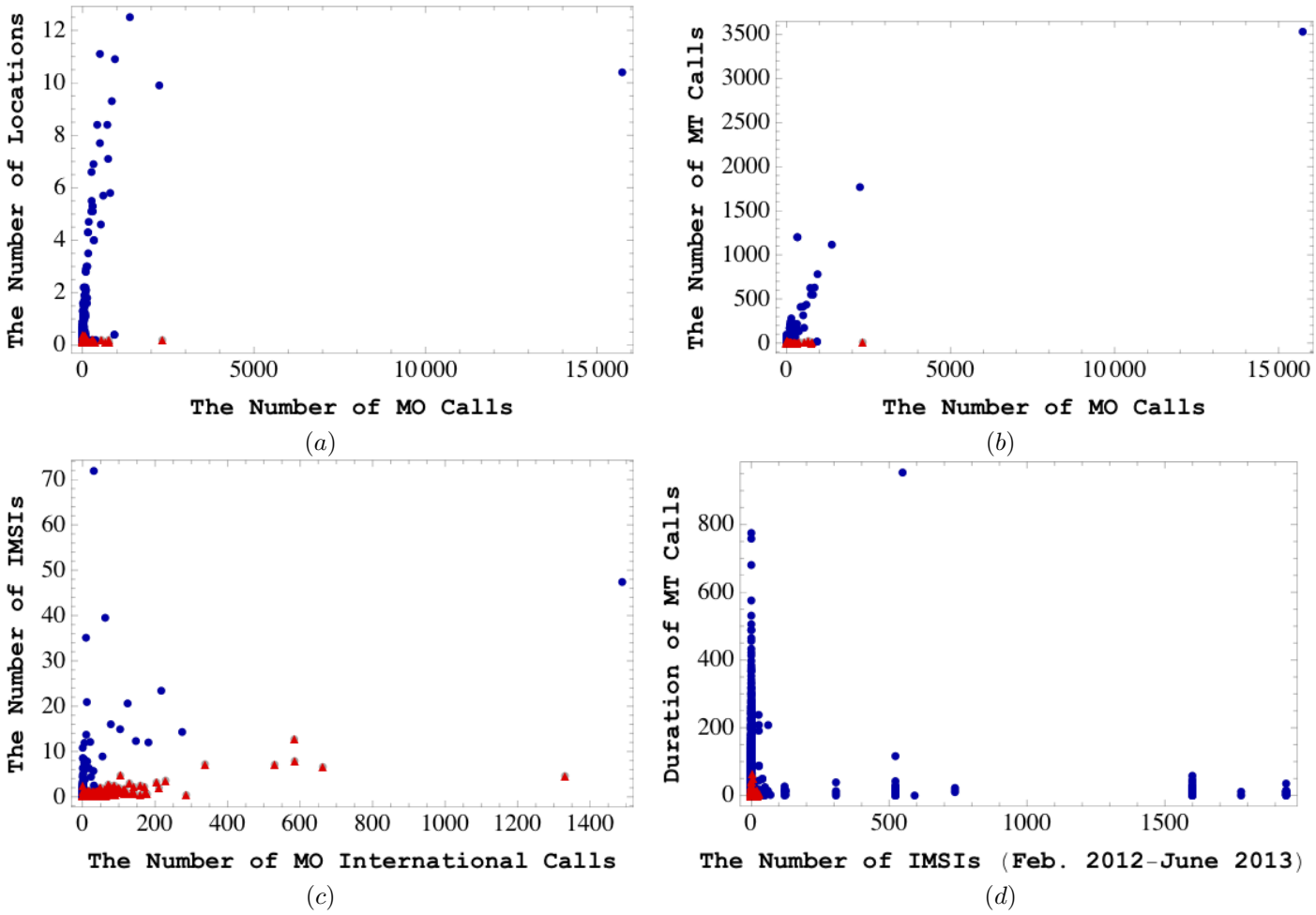

Fig. 3. Voice call traffic characteristics of SIMboxes (dots) and legitimate (triangles) accounts

and connect to a very small number of base stations.[3] By contrast, legitimate customers are highly mobile and connect to multiple base stations. In particular, network probing devices have a very large number of locations. Figure 3(b) plots the number of MO calls versus the number of MT calls. SIMboxes initiate thousand times as many calls as they receive and of significantly longer duration. By contrast, legitimate accounts have about the same number of initiated and received calls, which are of about the same duration.

Figure 3(c) plots the total number of MO international calls versus the number of IMSIs. Majority of legitimate customers are clustered around low values of both IMSI and international calls. Several network probing devices operate even more SIMs than fraudulent SIMboxes, however, they make less international calls.

Figure 3(d) shows the number of IMSIs for the period February 2012 – June 2013 versus the average duration of MT calls. MT call durations of fraudulent SIMboxes are much shorter than those of legitimate customers. Also, in contrast to some network probing devices, SIMboxes operate less IMSI during longer periods due to frequent account cancellation by the operator.

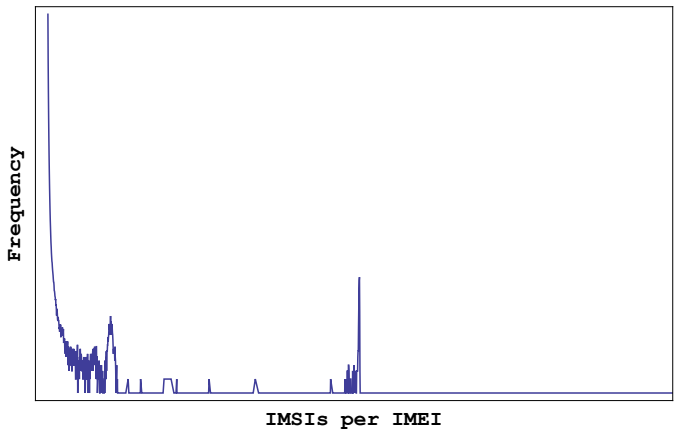

Fig. 4. The distribution of the number of IMSIs per IMEI

[3]This number is not necessarily one because many cellular devices connect to the network by means of a third generation (3G) technology based on wideband code multiple division access (WCDMA). In this technology, a device can be physically connected to up to 6 sectors at the same time, combining the signal at the receiver [13]. Depending on the channel conditions and fading, the serving base station might fluctuate throughout the set of 6 base station IDs. As a result, CDR records from the same static device will come from up to 6 different sectors.

The distribution of the number of IMSIs per IMEI from February 2012 till June 2013 is obtained by parsing CDRs of all the operator's active subscribers (see Figure 4, where the actual values on each axis are omitted to hide sensitive information). As expected, the majority of devices, connected to the mobile network, operate with one or just a few SIM cards (IMSIs). An IMEI operating two SIM cards can represent, for example, a used device which was bought on eBay. Also, Figure 4 shows two big spikes of IMEIs operating with many SIM cards. Further analysis identifies those IMEIs as legitimate cellular network probes.

E. Classification algorithm: training, testing, and implementation

To detect fraudulent SIMboxes, we use three classifiers: (i) alternating decision tree [14], [15], (ii) functional tree [16], and (iii) random forest [17], each of which is, in fact, a combination of "simple" classifiers. A classification rule is then presented as a linear combination of the three classifiers with weight coefficients found from minimization of the total prediction error on the training dataset.

A boosting method combines "weak" classifiers into a "strong" classifier by repeatedly re-weighting training data. It can focus not only on the majority of the training data, which is usually classified correctly, but also on the outlier records, which are often misclassified. An *alternating decision tree* is derived from a variant of the boosting method and combines decision stumps, i.e. single question decision trees. It has two types of nodes: test nodes and predictor nodes and alternates between them starting from the root node, which is a predictor node. In contrast to a *standard decision tree*, the alternating one allows a data record to follow multiple paths, whose weights (could be positive and negative) are then summed up with the sign of that sum determining the predicted value. The *random forest* uses a random subset of features to generate multiple decision trees from a sampled training set with replacement. The prediction is then determined by the majority rule (common vote) of the generated decision trees. At each node, a *standard decision tree* performs a simple value test with a single feature, whereas to improve classification accuracy, a *functional tree* linearly combines several features at both decision nodes and leaves, which is advantageous for large data sets. We use the Waikato environment for knowledge analysis (WEKA) data mining tool [18], which implements these decision trees.

Let the alternating tree, functional tree, and random forest be classifiers 1, 2, and 3, respectively, and let $x_{ki} \in \{0, 1\}$ be the prediction of classifier k for testing record $i \in \{1, \ldots, n\}$, where $x_{ki} = 1$ if classifier k predicts record i as a SIMbox, and $x_{ki} = 0$ if it predicts the record as a legitimate customer. Let $y_i \in \{0, 1\}$ be an actual label of data record i ($y_i = 0$ and $y_i = 1$ if record i is a legitimate account and a fraudulent SIMbox, respectively). To improve classification accuracy, the data labels can be regressed with respect to the predictions of the three classifiers:

$$\widehat{y}_i = w_1\, x_{1i} + w_2\, x_{2i} + w_3\, x_{3i}, \tag{1}$$

where $\widehat{y}_i$ is the prediction of data label y_i and $w_1 + w_2 + w_3 = 1$. The absence of the intercept in (1) and the constraint $w_1 + w_2 + w_3 = 1$ guarantee that the cases when all three classifier predict either 0 or 1 will be classified as a legitimate customer and a SIMbox, respectively. The regression coefficients w_1, w_2, and w_3 are found by the least squares method

$$\min_{w_1, w_2, w_3} \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2 \quad \text{s.t.} \quad w_1 + w_2 + w_3 = 1, \tag{2}$$

which yields a system of four linear equations with respect to w_1, w_2, w_3, and the Lagrange multiplier corresponding to the constraint $w_1 + w_2 + w_3 = 1$. Now, the classification rule is given by

$$y_i^*(x) = \begin{cases} 1, & w_1\, x_{1i} + w_2\, x_{2i} + w_3\, x_{3i} > \alpha, \\ 0, & \text{otherwise}, \end{cases}$$

where the threshold α is found from minimizing the classification error

$$\min_{\alpha} \sum_{i=1}^{n} (y_i^* - y_i)^2. \tag{3}$$

Observe that the regression coefficients w_1, w_2, and w_3 and the threshold α can be found in "one shot" from the optimization problem

$$\min_{w_1, w_2, w_3, \alpha} \sum_{i=1}^{n} (I_{\{\widehat{y}_i > \alpha\}} - y_i)^2 \quad \text{s.t.} \quad w_1 + w_2 + w_3 = 1, \tag{4}$$

where $I_{\{\cdot\}}$ is the indicator function equal to 1 if the condition in curly brackets is true and zero otherwise. However, (4) is a nonlinear problem of four variables, whereas (2) has a closed-form solution, and the problem (3), though being similar to (4), has only a single variable.

Algorithm 1 summarizes the suggested approach.

Algorithm 1 Classification rule

1: τ = Time interval for collecting CDRs
2: S = Set of indices of IMEIs of all known fraudulent SIMboxes (investigated by the fraud department)
3: L = Set of indices of IMEIs of known legitimate devices
4: **for all** $j \in S \cup L$ **do**
5: Generate features for the period τ
6: Generate alternating decision tree, functional tree, and random forest and compute their optimal weights w_1, w_2, and w_3 along with corresponding threshold α
7: **end for**

Alternatively, given the predictions (x_{1i}, x_{2i}, x_{3i}), $i = 1, \ldots, n$, the data labels can be classified by the *logistic regression* and a *support vector machine (SVM)* [19]. However, the numerical experiments with the training data show that both these methods perform worse than the regression problem (2) with (3). This could be partially explained by the fact

TABLE II
CLASSIFICATION RESULTS

Method	False Positive	False Negative	Accuracy (%)
Alternating decision tree	0.0005	0.1	99.91
Functional tree	0.0007	0.07	99.90
Random forest	0.0001	0.16	99.93
Classification rule	0.0001	0.09	99.95

that the overwhelming majority of triplets (x_{1i}, x_{2i}, x_{3i}), being either $(0, 0, 0)$ or $(1, 1, 1)$, correspond to correct training data labels (0 or 1, respectively), so that they do not affect the objective function in (2), whereas they skew the logistic regression. For the training data, the *hard margin* SVM is not separable, whereas the *soft margin* SVM requires to specify the extent of misclassifications (through an additional parameter) and involves $n + 4$ variables, which makes it less attractive from the computational perspective.

The three classifier were trained on the training dataset, for which their optimal weight coefficients are $w_1 = 0.31$, $w_2 = 0.26$, and $w_3 = 0.43$ with the corresponding threshold to be $\alpha = 0.39$. Table II shows predictions of the three classifiers and of their optimal linear combination on the testing dataset. The false positive is the proportion of legitimate customers classified as SIMboxes, whereas false negative is the proportion of SIMboxes classified as legitimate accounts. The "accuracy" column in Table II shows the proportions of the total correct classifications. Among the three classifiers, the random forest has the lowest false positive and the highest false negative, whereas the functional tree has the lowest false negative and the highest false positive. The optimal linear combination of the three classifiers improves both the false positive and false negative and has the highest accuracy of 99.95%.

For a real mobility network, generating of the set of features for each account requires processing hundreds of millions of CDRs, in which case, accounts with less than 10 IMSIs per IMEI that are unlikely to be very active SIMboxes are filtered out.[4] Additionally, all known legitimate accounts such as network probing and corporate accounts are also filtered out. Thus, after the pre-processing, only 0.02% of all active accounts remains for feature extraction and classification. Algorithm 2 incorporates Algorithm 1 and presents a scheme for detecting new fraudulent SIMboxes in the real mobility network. The operator's fraud department confirmed that Algorithm 2 successfully identifies new SIMboxes.

Algorithm 2 SIMbox detection

loop
  Run Algorithm 1 for the period τ
  N = Set of indices of all IMEIs in the network
  $\Omega = \emptyset$
  Θ = Set of indices of IMEIs corresponding to known network probing devices and corporate accounts
  h = Filtering threshold
  for all $i \in N \setminus \Theta$ **do**
    m_i = number of IMSIs operated by IMEI$_i$
    if $m_i \geqslant h$ **then**
      $\Omega = \Omega \cup \{i\}$
    end if
  end for
  for all $i \in \Omega$ **do**
    Generate features, then apply the alternating tree, functional tree, and random forest obtained at step 2 to the features and compute the prediction y_i^*
  end for
  return Detected SIMboxes to the fraud department
  update S and L based on the feedback from the fraud department
end loop

IV. RELATED WORK

The rise of the mobile communication technology for the past decade is mirrored by the range and sophistication of illegal activities on mobile networks. SMS spam is by far the most prevailing illegal activity that has attracted considerable attention from both industry and academy [20]. In this fraud, mobile users are tricked to visit phishing urls and to provide sensitive information. As SIMbox fraudsters, spammers also operate large numbers of SIM cards from each IMEI [6]. The volume of SMS spam is expected to grow at the annual rate of 500% [21]. For a detailed analysis of SMS spam on mobility networks, see [7].

Smart phone GGtracker malware [22] is yet another example of recent fraudulent activities. Being embedded into a legitimately-looking app, it silently subscribes users to premium number services such as a horoscope for $10 per month and hides all communications with those premium numbers. The users learn that they are victims of the fraud only in the end of a billing cycle when they see excessive charges on their bills. For an overview of other malware on Android platforms, see [23].

Subscription fraud occurs when fraudsters steal customer's identification and use it to subscribe to a mobility network [24], [25]. With a low cost technique, a fraudster can sniff traffic from a GSM (Global System for Mobile Communications) mobility network and break its encryption [26]. As a

[4]Section III shows that fraudulent SIMboxes use a large number of SIM cards. Thus, first we calculate the total number of SIMs used by each active IMEI in the network within a week. The threshold of 10 IMSIs is arbitrary.

result, he can obtain the IMSI and the secret key of any victim in his vicinity and then can use wireless service at the victim's expense. As with the GGTracker malware, victims learn about the fraud only when their bills arrive.

Several security firms offer their services for detection and prevention of SIMbox fraud [27], [28], however, details of their detection techniques are not disclosed. To the best of our knowledge, there is only one publicly available work [5] on this subject. It uses artificial neural networks (multi layer perception method) to detect fraudulent SIMboxes based on 9 voice call communication features for 6415 subscribers from one Cell ID (234,324 calls in total). The method detects SIMboxes with 98.71% accuracy.

In contrast to [5], our classification rule is trained and tested on a larger data sample of accounts distributed nationwide, and our features are computed per IMEI (device identifier) rather than per subscriber identifier. Since SIMboxes operate with multiple SIMs, 500 IMEIs of fraudulent SIMboxes correspond to thousands of fraudulent SIMbox subscriber identifiers. Only few of our features coincide with those in [5]. For example, the number of locations, which we consider, is quite important but is not relevant for [5], since all subscribers in [5] were in one Cell ID. Other important new features include the number of SIM cards, the total number of international calls and its ratio to the total number of calls. Also, we show that network probing devices have communication patterns similar to those of SIMboxes.

V. Conclusions

We have analyzed voice call communication features of fraudulent SIMboxes in the mobility network of a major tier-1 network operator in the United States and have identified call traffic patterns distinguishing fraudulent SIMboxes from legitimate devices. Those patterns include high number of IMSIs per IMEI, large number of international phone calls, imbalance between MO and MT traffic (international and domestic) and static physical location. Based on the features, we have proposed three classifiers of fraudulent SIMboxes in mobility networks: alternating decision tree, functional tree, and random forest. The random forest and functional decision tree provide the lowest false positive and the lowest false negative, respectively. The false positive of the alternating decision tree is lower than that of the functional tree, and its false negative is lower than that of the random forest. The predictions of the three classifiers have been linearly combined into a classification rule, where classifiers' weight coefficients have been found from minimization of the total classification error on the training dataset. The random forest has the largest weight coefficient followed by that of the alternating decision tree. The accuracy of the classification rule is 99.95%. For large data sets, the scalability of the algorithm can be improved by filtering out accounts with less than 10 IMSIs (99.98% of all active subscribers). The operator's fraud department has confirmed that the proposed algorithm detects new fraudulent SIMboxes with a low false positive.

Acknowledgements

We are grateful to Angus MacLellan and Richard Becker for their help, comments, and valuable suggestions.

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] 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.
[6] I. Murynets and R. Piqueras Jover, "Analysis of SMS Spam in Mobility Networks," in *International Journal of Advanced Computer Science (IJACSci)*, vol. 3, num.7, July 2013.
[7] I. Murynets and R. Piqueras Jover, "Crime scene investigation: SMS spam data analysis," in *Proceedings of the 2012 ACM conference on Internet measurement*. ACM, 2012, pp. 441–452.
[8] "PennyTalk," http://www.pennytalk.com/.
[9] "ZapTel Calling Cards," http://www.zaptel.com/.
[10] "Vonage Calling Cards," http://lp.vonage.com/callingcard.
[11] "RCATS - Remote Cellular Active Test System," JDSU, http://goo.gl/VEbMA.
[12] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.
[13] Universal Mobile Telecommunications System (UMTS), "Physical layer procedures (FDD). 3GPP TS 25.214," vol. v3.17.0, 1999.
[14] Y. Freund and L. Mason, "The alternating Decision Tree Learning Algorithm," 1999.
[15] G. Holmes, B. Pfahringer, R. Kirkby, E. Frank, and M. Hall, "Multiclass Alternating Decision Trees," 2002.
[16] J. Gama, "Functional trees," *Machine Learning*, no. 55, 2004.
[17] L. Breiman, "Random Forests," vol. 45.
[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA Data Mining Software: An Update," in *SIGKDD Explorations*, vol. 11.
[19] M. Zabarankin and S. Uryasev, *Statistical Decision Problems: Selected Concepts and Portfolio Safeguard Case Studies*. Springer, 2013, to appear.
[20] N. Perlroth, "Spam Invades a Last Refuge, the Cellphone," The New York Times, April 2012, http://preview.tinyurl.com/7nwvm3g.
[21] A. Bobotek, "Threat of Mobile Malware and Abuse," Messaging Anti-Abuse Working Group (MAAWG), October 2010, http://goo.gl/Ay57e.
[22] T. Strazzere, "GGTracker Technical Tear Down," Lookout Mobile Security, http://goo.gl/IuVfm.
[23] Y. Zhou and X. Jiang, "Dissecting android malware: Characterization and evolution," in *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 2012, pp. 95–109.
[24] Y. Moreau, H. Verrelst, and J. Vandewalle, "Detection of mobile phone fraud using supervised neural networks: A first prototype," in *Artificial Neural NetworksICANN'97*. Springer, 1997, pp. 1065–1070.
[25] P. Barson, S. Field, N. Davey, G. McAskie, and R. Frank, "The detection of fraud in mobile phone networks," *Neural Network World*, vol. 6, no. 4, pp. 477–484, 1996.
[26] K. Nohl and S. Munaut, "Wideband GSM sniffing," in *In 27th Chaos Communication Congress*, 2010, http://goo.gl/wT5tz.
[27] "SIM Box Detection Service," Telekom Austria, http://goo.gl/Ac12d.
[28] "SIMbox detector," Xintec, http://goo.gl/AUZbe.