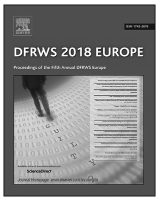
Digital Investigation 24 (2018) S131–S141

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin



DFRWS 2018 Europe — Proceedings of the Fifth Annual DFRWS Europe

Data-driven approach for automatic telephony threat analysis and campaign detection

Houssem Eddine Bordjiba*, ElMouatez Billah Karbab, Mourad Debbabi

Concordia University, Canada

Keywords:
Telephony abuse
Telephony complaints
Vishing
Spoofing
Telephony abuse campaigns

ABSTRACT

The growth of the telephone network and the availability of Voice over Internet Protocol (VoIP) have both contributed to the availability of a flexible and easy to use artifact for users, but also to a significant increase in cyber—criminal activity. These criminals use emergent technologies to conduct illegal and suspicious activities. For instance, they use VoIP's flexibility to abuse and scam victims. According to (F. I. F.. N. D. N. C. R. D. Book, Available at: https://www.ftc.gov/news-events/press-releases/2016/12/ftc-issues-fy-2016-national-do-not-call-registry-data-book, accessed on: 27 August 2017), US government revealed receiving more than 5.3 million telephony abuse complaints in 2016. Based on this report, more than 226 million phone numbers were registered on the *Do Not Call Registry* list as not to receive telemarketing calls. For instance, they use VoIP's flexibility to abuse and scam victims. A lot of interest has been expressed into the analysis and assessment of telephony cyber-threats. A better understanding of these types of abuse is required in order to detect, mitigate, and attribute these attacks. The purpose of this research work is to generate relevant and timely telephony abuse intelligence that can support the mitigation and/or the investigation of such activities. To achieve this objective, we present, in this paper, the design and implementation of a Telephony Abuse Intelligence Framework (TAINT) that automatically aggregates, analyzes and reports on telephony abuse activities. We deploy our framework on a large dataset of telephony complaints, spanning over seven years, to provide in-depth insights and intelligence about emerging telephony threats. The framework presented in this paper is of a paramount importance when it comes to the mitigation, the prevention and the attribution of telephony abuse incidents. We analyze the data and report on the complaint distribution, the used numbers and the spoofed callers' identifiers. In addition, we identify and geo-locate the sources of the phone calls, and further investigate the underlying telephony threats. Moreover, we quantify the similarity between reported phone numbers to unveil potential groups that are behind specific telephony abuse activities that are actually launched as telephony abuse campaigns.

Introduction

The Internet is commonly used by cyber-criminals to exploit the users through emails, social media networks or other vulnerabilities. However, in recent years, cyber criminals started using another channel to reach their victims, namely *the telephony network*. Being a well-established and more secure service compared to the Internet, the use of telephony for different purposes has increased. However, its service is now being abused to perpetrate various cybercrime attacks. Furthermore, Internet telephony offers a plethora of options for cyber criminals to generate noisy bulk calls, which results in disrupting telephony services as well as targeting people to monetize their activities. Therefore, efficient forms of unsolicited telemarketing and vishing (voice-phishing) campaigns, involving interactive voice response and dialing algorithms, have emerged. In addition, the trivial use of SMS/MMS messages has given the telephony abuse an epidemic trend, encouraging the propagation of scamming campaigns as well as phishing mobile technology users. Based on these facts, uncovering the key players behind telephony abuses is a real challenge, especially since abusers hide themselves behind anonymity services (Tu et al., 2016).

Problem statement. Recently, we have witnessed a significant rise in telephony abuse. In 2016, according to (H, 2017), Americans

* Corresponding author.
E-mail address: bordjiba.houssemnc@gmail.com (H.E. Bordjiba).

https://doi.org/10.1016/j.diin.2018.01.016

lost 9.5 billion dollars due to phone scams. These losses are the result of scamming campaigns that targeted approximately 32 million telephone customers. Additionally, fraudsters have been impersonating government agencies and well-known companies to craft their attacks. For example, in April 2017, fraudsters intimidated and scammed a telephone customer by impersonating her bank (Warning over phone scam that cost this woman 70, 2017). According to (R. o. I. D. D. L. o. T. S. f. t.. F. S, 2016), fraudsters posed as the Internal Revenue Service (IRS) and threatened the customers with police involvement and possible imprisonment if the person does not pay the fake tax statements. Consequently, twelve Nebrasken inhabitants lost $56,000 due to this telephone scam (Irs nebraskans lost 56000 to telephone scam, 2016).

The victims reported that the criminals created a believable scam by assigning the caller ID of IRS to their number and using the victims personal information. Therefore, it is of a paramount importance to design and implement a telephony abuse intelligence framework that will provide assistance in the detection, mitigation and attribution of scamming campaigns. In this respect, we aim to answer the following research questions:

1. How to analyze data about telephony abuse to derive situational awareness and insights about the different telephony scams?
2. How to generate timely and relevant intelligence about telephony abuses that can be used for detection, mitigation and attribution purposes?
3. How to analyze the collected data to timely detect the different scamming campaigns that are taking place on the telephony network?

To address the aforementioned research questions, we design and implement a framework that is capable of collecting, in near-real-time, telephony complaints data and analyze it to generate timely and relevant intelligence on telephony abuse activities. The main benefits of our framework are: (i) Near-real-time and worldwide situational awareness on telephony abuse activities; (ii) Generation of profiling information on top abusers by calling identifiers, service providers, and geo-locations; (iii) Identification of scamming and tele-marketing campaigns by exploring the similarities of the attributes underlying telephony abuse activities. Our analysis relies on using multiple data mining and machine learning techniques and the correlation of the data with external databases, such as the *Canadian Numbering Administrator* database (C. N. A, 2015) and the *North American Numbering Plan Administration* database (N. A. N. P. A, 2015) to enrich the open-source collected data, then profiling different phone abuse activities together with the underlying campaigns.

The main differentiating factors of our proposal with respect to the state-of-the-art contributions are: (i) We rely on multiple sources of complaints data and publicly available telephony databases; (ii) We use larger datasets compared to (Maggi, 2010). Indeed, we used a dataset that is comprised of 5 million complaints whereas (Maggi, 2010) used a dataset of 300 complaints which they collected using their developed web application; (iii) We analyze and study the various telephony abuse threats and their associated campaigns while (Miramirkhani et al., 2017) focuses their study on phone abuses in technical support scam campaigns; (iv) Our framework is an automatic and online solution, where limited human interaction is needed, and it aggregates near-real-time data and generates near-real-time intelligence whereas in (Costin et al., 2013; Gupta et al., 2015) they had an automatic collection, yet an off-line analysis of the dataset collected; (v) This is the very first research contribution, to the best of our knowledge, proposing an intelligent, robust, and large-scale framework for the timely detection of telephony abuse campaigns by exploring the similarities between the individual abuse incidents form telephony complaints data.

Our contributions are threefold:

1. ***Design and implementation of a Telephony Abuse Intelligence Framework*** (TAINT)

We design and implement a framework that takes as input near-real-time complaints data about telephony abuse and generates timely and relevant intelligence on abusers, the nature of the abuse, the geo-locations, the call identifiers, etc. This generates important situational awareness and insights about the ongoing worldwide abuses over the telephony network.

2. ***Telephony abuse campaign detection***. We design and implement an algorithm that explores the similarities between abuse incidents in order to detect, in near-real-time, orchestrated and coordinated scamming and tele-marketing telephony campaigns.
3. ***Evaluation of the system using real-world data***. We conduct a thorough evaluation of our framework over a large dataset, which is comprised of 5 million abuse complains, spanning over 7 years. It is important to mention that the derived intelligence is instrumental in the detection, mitigation and attribution of telephony incidents. As such, it can be used by law-enforcement officers to investigate the underlying incidents and attribute them. On the other hand, it can be also used by telephony operators to mitigate telephony abuse activities.

The remainder of this paper is structured as follows: In Section Dataset, we present a description of our telephony complaints dataset. Section Framework architecture provides an overview of the architecture and design of our framework together with the algorithmics of campaign detection. Section Implementation describes and explains our back-end and front-end implementations of our framework. Section Results presents an extensive evaluation of our framework with the underlying results; and finally Section Conclusion presents some concluding remarks on this research.

Dataset

We secured complaint data in near-real-time from our partners; an average of 2000 complaints is received per day. This number increased to more than 8000 in 2016. Thus far, we received more than 5 million complaints gathered during a 7-year period. The raw received complaints contain multiple attributes such as the source phone number, the time when the complaint was made, the caller identification, and the message expressing the underlying complaint. Table 1 presents the attributes of the complaints together with their description.

Framework architecture

The goals of Telephony Abuse Intelligence Framework (TAINT) through its components are to automatically: (i) aggregate and analyze telephony abuse complaints filled up by telephony customers, (ii) identify and geo-locate scamming perpetrators and their utilized infrastructure, (iii) rank reported phone numbers in the complaints data according to their badness score, and (iv) cluster telephony abuses to unveil potential groups that are behind particular scamming campaigns. As input, it takes real-time telephony complaints that are then subjected to extensive analysis. The latter produce timely and relevant intelligence about worldwide telephony abuse activities. Such intelligence is meant to empower law enforcement investigators, and/or Telephony Service Providers (TSPs) in their efforts for the detection, mitigation and attribution of telephony abuse activities that are perpetrated by telemarketers,

Table 1
Description of the collected information.

Data Field	Description
Source Number	A string containing the calling party number
Caller Identification	A string representing the calling-line identification information if present
Time	A date/time object indicating the date-time of when the complaint was made
Complaint Text	A message expressing the underlying complaint

debit collectors, scammers, etc. In this section, we present the architecture of our framework and its main components. Fig. 1 provides a high-level overview of the framework and the interaction between its essential components. TAINT's different components are enumerated hereafter:

1. Telephony complaints text features extraction and classification
2. Correlation
3. Badness scoring of the scammers infrastructure
4. Campaign detection

Telephony complaints text features extraction and classification

The first component of TAINT aims to classify the complaints into two classes: *telephony abuse complaints*, or *non-telephony abuse complaints*. Our manual analysis of the dataset used in this work revealed that it contains a number of non-related telephone abuse complaints. Since, the collection of the complaints is through our partners' web pages, people make mistakes as to which form they should fill in their complaints. For instance, people complaining about an Internet service provider as illustrated in Fig. 2.

In addition, we assume that law enforcement agencies and telephone service providers receive numerous complaints about different problems the consumers might face. Therefore, the classification of complaints is a required step in our framework to filter the complaints so that TAINT analyzes only the telephony abuse complaints. To do so, the module goes through 3 phases: the collection phase, the learning phase and the execution phase. In the collection phase, we collect general conversation data from *20newsgroups* (T, 2017) text dataset. This dataset contains more than 18,000 conversations about 20 different general topics. We use such dataset to build baseline knowledge about general conversation in order to differentiate them from telephone complaints. In the learning phase, we split the *20newsgroups* and the complaints dataset into: (i) 70% training dataset and (ii) 30% test dataset to build a classification model that distinguishes telephony complaints from non-related complaints or text data. Then, we feed this data to a classification algorithm to build a model that will be used later on to classify the streamed complaints. In the execution phase, we pre-process the complaints from text messages to extract the features, and create a class where we label the complaints as *Valid* or *Non-Valid*. This phase is executed in 4 steps:

1. We tokenize (separate the words in a document) the complaint document.
2. We apply stemming (reduce the words to their stem) on the tokenized words, to get the root of the words to reduce our features set size and thus better accuracy of the classifier.
3. We remove the stop words and the special characters from the documents.
4. We apply Term Frequency minus Inverse Document Frequency (TF-IDF) on our preprocessed documents in order to obtain the most frequent and important words in each document in relation with the whole dataset.

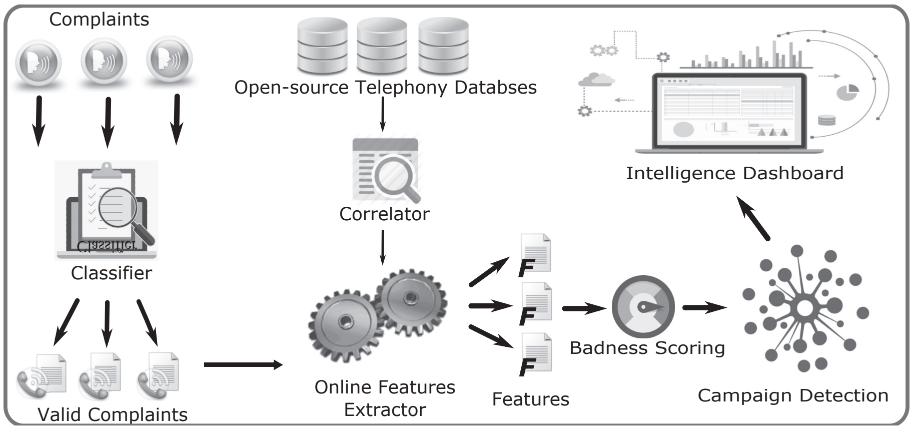
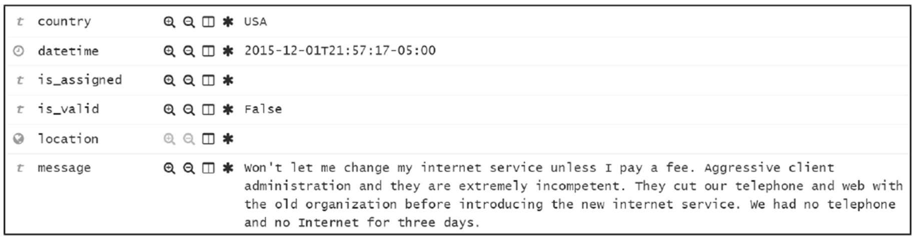


Fig. 1. TAINT framework overview.



```
t  country       USA
   datetime      2015-12-01T21:57:17-05:00
t  is_assigned
t  is_valid      False
   location
t  message       Won't let me change my internet service unless I pay a fee. Aggressive client
                 administration and they are extremely incompetent. They cut our telephone and web with
                 the old organization before introducing the new internet service. We had no telephone
                 and no Internet for three days.
```

Fig. 2. Screenshot of a non Telephony Abuse Complaint.

Afterwards, the features extracted through the text preprocessing of the complaints are then used as an input to the Support Vector Machine Classifier (SVM), and later in the other components of TAINT.

Support *Vector Machine Classifier.* The Support vector machine (SVM) is a supervised machine learning technique that is widely used for binary classification and regression. The rationale underlying the use of SVMs is that they are known to outperform other classifiers when it comes to supervised text categorization (Joachims, 1998). Furthermore, text classification relies on word vector features, which results in a multidimensional data analysis. Joachims (1998) shows that SVM do not require feature selection to build an accurate classification model comparing to the other suggested classifiers. Having a training dataset, and aiming to classify our complaints data into two main classes, we found that SVM is the most suitable machine learning technique for our problem. We also used TF-IDF to improve the performance of text classification as suggested in (Salton and Buckley, 1988).

Correlation

In order to enrich the complaint dataset, we extract the phone number reported in the complaint and we correlate them with other telephony databases. Our goal is to derive other information in our analyses, such as the geographic distribution of telephony abuse incidents. We mainly use the data sources provided by the *Canadian Numbering Administrator (CNA)* (C. N. A, 2015) and the *North American Numbering Plan Administration (NANPA)* (N. A. N. P. A, 2015) to determine the location of North American phone calls. In addition, we rely on the recommendation of the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T), namely E.164, to determine the country of origin for international calls. The correlation of the dataset results in three main classes of phone numbers according to the origin location: *North America*, *non-geographic or toll-free*, and *international*. Using the correlation tool, we extract two additional features that will be used in the subsequent components. The two additional features are:

- *Invalid phone numbers:* Invalid phone numbers are the numbers that do not conform to the North American Numbering Plan and cannot be dialed within the public switched telephone network. This may be viewed as the traditional form of spoofing, where the calling party number is relatively static and fictitious (e.g, 0123-4567-8910). Valid numbers conform to NANPA or international assignments under Recommendation ITU-T E.164. The numbers extracted may be formatted as an international numbers or as national numbers.
- *Unassigned or VoIP phone number*: These numbers use a valid phone number format, but the numbering plan area is unassigned within NANPA (e.g, 123-456-7890), which is a valid number that is not assigned to any customer.

Badness scoring of the scammers infrastructure

In order to provide law enforcement investigators and Telephony Service Provider (TSP) operators with an appreciation of the severity level of telephony abuse incidents, we elaborate a badness scoring to the phone number reported in each complaint. To do so, we rely on a training dataset that was provided to us by law enforcement officers. This dataset contain the complaint text and the source phone number along with their badness score. The badness score is a value between 1 and 100, where 1 refers to a low severity of the phone number, and 100 refers to the highest severity. In order to assign its badness score, law enforcement officers relied on the losses each phone number has caused to. As a result, this badness score will help to identify the most worthy phone numbers for investigation. We used this labeled dataset to create a regression model to assign a badness score to the new reported phone numbers. To this end, we subject our dataset to a linear regression machine learning algorithm. To train the linear regression machine learning model, we use the phone number reported in the complaint along with the word vector feature generated from the *Telephony complaints text features extraction and classification* component of TAINT as the explanatory variable X. Then, we use the badness score of phone numbers information provided by our partner as the dependent variable y. Similar to the procedure mentioned in the *Telephony complaints text features extraction and classification* component, we first train and build the regression model. Secondly, we evaluate the model on the test dataset. Lastly, we run the model on the streamed data to automatically rank the new arriving complaints.

Campaign detection

Caller Identification. Caller Identification Name (CNAM) is a feature provided by telephone carriers to identify the caller. Banks, government entities, companies and so on use the Caller Identification to authenticate themselves to their customers. This technology is very useful to both caller and receiver, yet the scammers can alter the callers identification. The action of alternating the caller identification to reach a malicious goal is called Caller ID spoofing. Criminals started using this technique to deceive and scam their victims. The increase of telephone users and the availability of Smartphones and VoIP helped in the effectiveness and the rise of this type of fraud (Mustafa et al., 2014).

Telephony Abuse Campaign: Telephony abuse campaign is a cyber attack in which one or multiple parties with a common objective (e.g., steal credit card numbers and sell them in the black market) coordinate and plan to attack or scam a group of people vulnerable to one of their attack schema. In our dataset, we have a multitude of complaints about various campaigns that victims have been subjected to. To detect telephony abusing campaigns, we apply record linkage on our dataset, then we use graph analysis as in (Christin et al., 2010). For this, we consider the CNAM feature, and the word vector feature generated from *Telephony complaints text features extraction and classification* component of TAINT to link the similar source numbers. We choose these features since abusers usually rely on the CNAM as the first attribute of their attack; for instance, fraudesters will use a *Bank Name* to scam a bank customers. Furthermore, we choose the complaint text to potentially get similar complaints, thus similar abusers.

Subsequently, we create a graph that represents the relation between these different source numbers. This graph helps in the visualization of the different campaigns, and it will be used as an input to the campaign detection algorithm. The nodes of this graph represent the phone numbers of the call sources, and the edges show that these sources use a similar CNAM, and have a similar technique in approaching their victims which indicates that it is a similar type of abuse. The result of this exercise gives us a network of phone numbers used in similar scamming campaigns, which in turn helps to identifying potential sources of fraud.

Similarity computation

Building the similarity network is an important module of TAINT framework. We generate the similarity network by computing the similarity between feature vectors. Moreover, having multiple similarities between the involved phone numbers in the similarity network shows a close relationship between these phone numbers, which indicates potential similar actors. To compute the similarity matrix for the sources in the complaints dataset, we use the caller identification as well as the extracted features from the complaint text messages. The idea is that phone numbers that use similar spoofed caller identifications and have the same pattern in approaching the victims are most likely part of the same campaign. The similarity between the different numbers is calculated using the Jaccard Similarity Index. We use the Jaccard similarity index to compute the similarity between the feature vectors, and by that group phone numbers that are involved in the same abuse campaigns. We choose Jaccard Similarity Index since it is known to be efficient when it comes to document similarities, and it has shown a promising results for our problem compared to other techniques such as Locality Sensitive Hashing (LSH). When using LSH, we noticed that many similar documents were omitted. Given a pair of phone numbers, after extracting the feature vectors, we use the Jaccard distance to compute the distance between two feature vectors *m* and *n*. The Jaccard Similarity Index is computed by first calculating the intersection of two sets *A* and *B*, which are the number of the elements that exist in both *A* and *B*. Then, computing the union of different elements in both set *A* and set *B*. Finally, the cardinality of the intersection of the two sets is divided by the cardinality of their union, as given by the following formula:

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

The final result of the similarity computation produced a heterogeneous graph, since some nodes did not have any edges whereas other nodes had multiple edges. The nodes of the network represent the phone numbers and the edges represent the similarity between these phone numbers if it exceeds a certain threshold. The threshold value is fixed after a manual testing and evaluation.

Adjacency Matrix Computation. The spamming campaign network is represented via an adjacency matrix, where each value is a binary value, either 0 or 1. The Matrix is used as an input to the community detection algorithm in order to unveil the different campaigns. If two phone numbers are estimated similar based on their caller identification list and word vector feature, and are above a chosen threshold, the value one is assigned to this pair of sources. On the other hand, if the similarity between the caller identification list and word vector features of two different phone numbers is below the chosen threshold, then the value of 0 is given to that pair. Using the computed adjacency matrix, we build a graph that illustrates the relational network underlying our data. The graph contains many clusters of communities. These communities most likely represent different calling campaigns. We apply the campaign detection algorithm on our complaint data represented in the similarity matrix.

Community detection

Community detection is a technique used in graph analysis to extract the most interconnected subgraphs. After applying text mining and features engineering, we built a relationship graph of the reported phone numbers. Then, to extract the subgraphs, which represents in our study the different telephony abuse campaigns, we use the Fast Unfolding Community Detection Algorithm (Blondel et al., 2008). We choose this algorithm (Blondel et al., 2008); since, it can scale to hundreds of millions of nodes and billions of links. Furthermore, through it's technique the algorithm achieves good results since it relies on measuring the modularity of communities, to unveil the different communities. The modularity is a scale value between −1 and 1 that measures the density of links inside communities as compared to links between communities. Similar to work completed in (Karbab et al., 2016), we feed our computed network to the algorithm. Then, we use a degree filtering parameter to filter all the nodes that have less degree than the chosen parameter. We chose the threshold of the Jaccard similarity index to the value of 0.3, since using these parameters we had more false positive. As we know that the investigator can easily look afterwards into a given campaign and filter those very easily. Nodes with a high connections will keep up their edges, which indicates that they are malicious telephone numbers that belongs to that particular campaign. We have fixed all the parameters in our assessments as follow. First, we use the degree 1 to filter all telephone numbers having no connection to what other telephone numbers; since they are clearly not part of any campaign. Our system filters these phone numbers as they are supposedly isolated caller. Then, we extract the different communities. This method permits an effective grouping of campaign. Lastly, we compute the average badness score generated from the *Badness Scoring of the Scammers Infrastructure* component of the phone numbers and caller identification involved in each campaign to get its badness score.

Campaign identification

Algorithm 1. Real-time campaign identification algorithm.

```
Algorithm 1: Real-time Campaign Identification Algorithm
  Data: Complaints
  Result: Campaign attribution
  List complaints;
  Last_seen = 0;
  Threshold $\tau$;
  while complaint do
      read complaint;
      if $t_i$ - Last_seen < $\tau$ then
          if ( $id_i \approx id_j$ and $m_i \approx m_j$ ) then
              $E_i$ and $E_j \in C_i$;
          else
              New campaign $C_n$
              $E_i \in C_n$
          end
      else
          New campaign $C_n$
          $E_i \in C_n$
      end
  end
```

Giving a set of campaigns: $C = \{C_1, C_2, C_3, \ldots, C_N\}$, and a call event as: $E_i = \langle s_i, d_i, t_i, m_i, id_i \rangle$, where:

- s_i is the source node in E_i,
- d_i is the destination node in E_i,
- t_i is the time when E_i happened,
- m_i is the message in E_i and,
- id_i is the Caller Identification in E_i.

We claim that E_i belongs to the campaign C_i if and only if:

- t_i-$ls < \tau$ (where ls is the last time when C_i was detected, and τ is a chosen threshold)
- Given $E_{i,j \in N}$, $m_i \approx m_j$
- Having $id_{i,j}$, where $id_i \approx id_j$

Note that Jaccard similarity index is used for the above similarity computation.

Real-time campaign identification

To detect campaigns in real-time, we follow the rules described below:

Given a call event $E_i = \langle s_i, d_i, t_i, m_i, id_i \rangle$, and a known set of campaigns $C_1 = \{E_1, E_2, E_3, \ldots, E_{N_1}\}$, $C_2 = \{E_1, E_2, E_3, \ldots, E_{N_2}\}$, until $C_n = \{E_1, E_2, E_3, \ldots, E_{N_m}\}$. We then Evaluate for each given (E,C) whether:

1 We create a new campaign, or
2 We consider that $E_i \in C_i$

Algorithm 1 details the process.

Implementation

In this section, we detail the back-end data process, and the front-end analytics.

Back-end

In this section, we detail the implementation of TAINT back-end. We present the general components of TAINT and the role of each component.

Telephony Abuse Intelligence Framework has different components that take part in analyzing the complaints from various sources (see Fig. 3). We choose each component of TAINT to suit well the objective of our solution. We enumerate the components of TAINT in the following:

- **Apache Kafka:** We first pre-process and store the complaints in near real-time using a python script. This results into a feed that is streamed to our system using *Apache* Kafka (2015), which is an Apache open-source software used as a high-throughput distributed messaging system. It provides a unified and low-latency platform for real-time handling of datasets. We use *Kafka* as a temporary feed storage, where other components of the system can asynchronously retrieve the complaints. Since our system do real-time analysis and is designed to handle big data, we choose Kafka to allow easy distribution of the processes
- **MongoDB**: These complaints are persisted into Mongodb (2015), an open-source and highly scalable document database, which is used to store the complaints in a document format. Since our framework aggregates and analyzes data from different sources, we choose Mongodb which stores the information as JSON document which can vary in structure. This allows also the easy integration of other information from other collection sources, such as, a telephony honeypot, to our Framework.
- **Neo4j**: The online features extractor component leverages the graph format provided by Neo4j and the asynchronous stream provided by Kafka to generate the features. We choose Neo4j as it allows easy manipulation and querying of graph data. We use it as well to extract the communities and hence detect the different calling campaigns.
- **Redis**: After computing the features, we store them in in-memory key-value database for high-speed operations. The key is the phone number and the value is its features. We chose Redis (R, 2015), an open source key-value cache, for this purpose.

Furthermore, we use Scikit-learn (scikit-learn Machine Learning in Python, 2016) to build the machine learning classification and regression models. Finally, the clustering system leverages the high speed of the feature-cache to get the similarity graph. Then, it applies a Fast Unfolding community detection algorithm to find the most relevant clusters based on the modularity value. TAINT then uses each cluster to classify each phone number to a scamming campaign. For new received complaints, TAINT computes the features only for the new complained phone numbers and caller IDs. Afterwards, the new computed features will overwrite the old ones in the key-value caches.

Front-end

The Front-end of TAINT Framework is a dashboard implemented using *Joomla* (J. CMS, 2015) and an interactive data analysis using *Kibana* (YW, 2015), a sophisticated web front-end and part of *Elasticsearch* (Stack, 2015) ecosystem. We choose the search and analytics engine Elasticsearch as it is a scalable solution, and provides a high quality front-end Kibana. Joomla is used in order to allow the creation of different tabs which helps in a good user interface experience.

The collected information is indexed and stored in a way to achieve the maximum granularity. Moreover, through the design and implementation of the digital dashboard, we expose near-real-time analytics results to scientists, cyber risk professionals, law enforcement, etc. The digital dashboard dynamic components are automatically updated and equipped with drill-down capabilities.

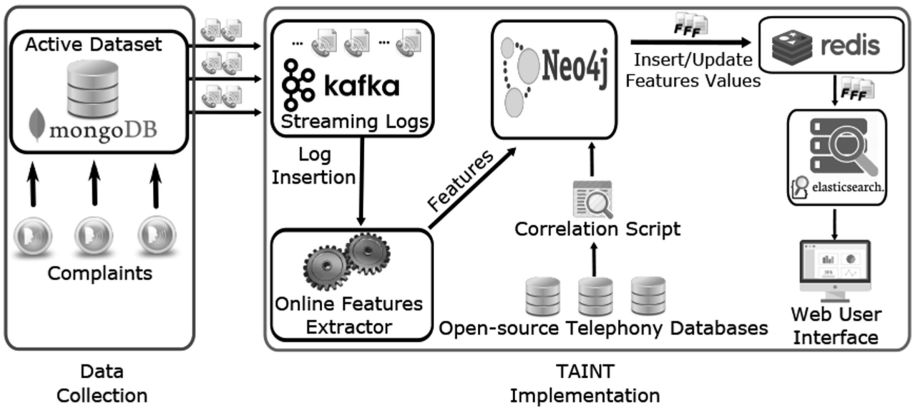


Fig. 3. Components of TAINT framework.

The distribution of calls over time (per hour, day, week, month and all time) and the various analytics are provided on the web interface in different tabs. The tabs were created using a Joomla CMS. The primary interface depicts the various analytics including the abuser's geolocation distribution map, as depicted in Fig. 4.

Results

In this section, we present the evaluation and provide some findings of our telephony abuse analysis framework. In this pursuit, we first present the evaluation of the algorithms used in TAINT. Second, we provide statistics about telephony fraud. The aim is to answer questions about the main static and dynamic characteristics of this dataset. Finally, we present and review some significant detected telephony calling campaigns.

Evaluation of the algorithms

In this section, we provide the results of the SVM classifier and the linear regression model for badness ranking. An experimental setup with two evaluation criteria namely, the False Positive Rate (FP rate) and the True Positive rate (TP rate) have been defined to evaluate the classification models in TAINT. TP rate is the percentage of positive tuples and is considered as a measure of completeness. Whereas, the FP rate is dedicated to the misclassified data (Han et al., 2011). The equations of the mentioned criteria are provided in the following:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$TP \text{ rate} = \frac{TP}{TP + FN} \tag{3}$$

and

$$FP \text{ rate} = \frac{FP}{FP + TN} \tag{4}$$

The classification of the complaints were based on text mining, and to achieve a high accuracy many experiments have been performed on the data. In our work, we have used SVM. As provided in Table 2, we can see that the execution of SVM using our selected features gave us a good accuracy of 98.85%.

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \tag{5}$$

Furthermore, the execution of the Linear regression model for badness scoring gave also a very promising results with a Mean Squared Error (MSE) of 6.4.

Statistics on telephony abuse data

In order to grasp insights on telephony abuse, we have performed an analysis on the collected data. To this end, we conducted a thorough inspection of the data to extract quantitative and qualitative insights. The collected data spans over a period of 7 years, starting from 1st of January, 2009 to 30th of December 2016.

Phone Number Counts Compared to the Number of Complaints We observed interesting statistics, which show that 836,630 phone numbers are reported only one time, whereas 260,280 phone numbers are reported many times. This results in a total of 5,003,873 complaints, as demonstrated in Table 3. This shows that a set of phone numbers are making multiple calls, which demonstrates the actual poor monitoring of telephony abuse.

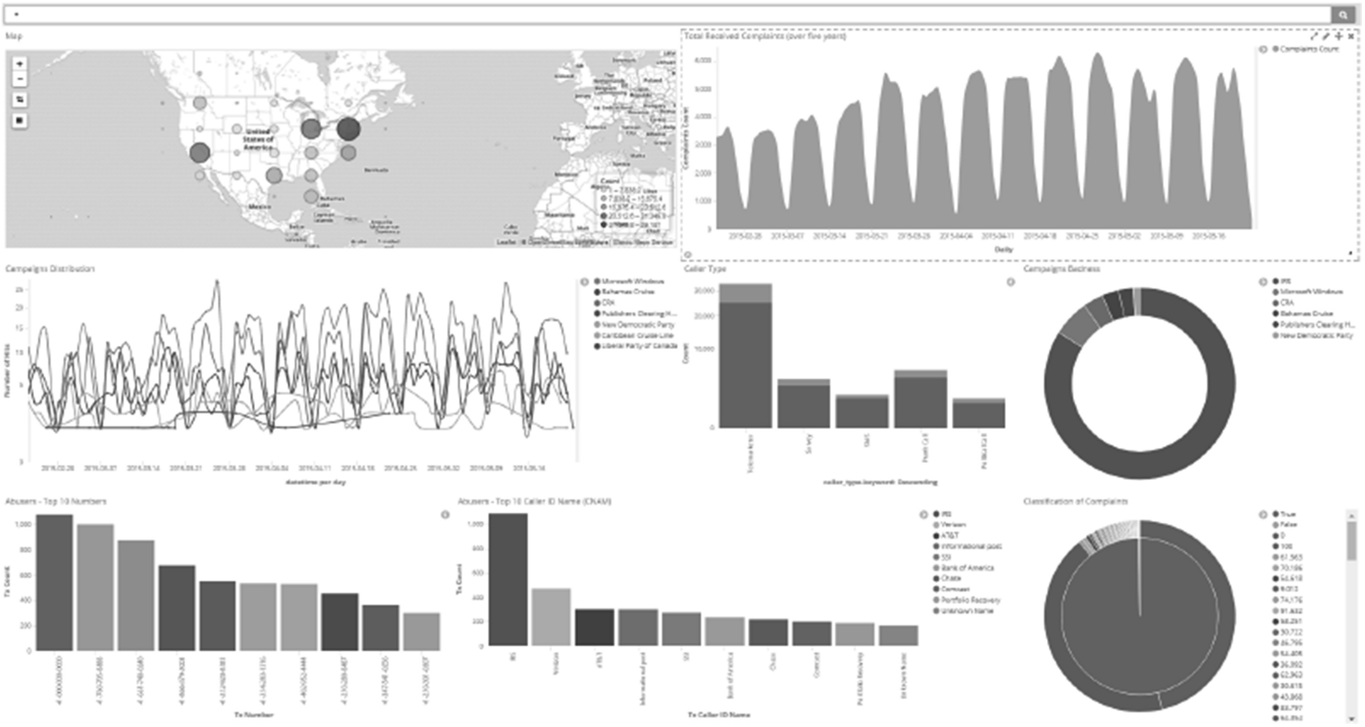


Fig. 4. Screenshot of the Near real-time Monitoring Web Interface.

Table 2
Results of the classification model on the complaints dataset.

Run	TP rate	FP rate	Accuracy
Support Vector Machine	98.9%	2%	98.85%

Table 3
Complaints details.

Total Complaints		5,003,873
Phone Numbers	One time	836,630
	Multiple times	260,280

Complaints distribution

The chart in Fig. 5 depicts the distribution of the complaints over time. This figure provides insights on the trending of telephony abuse during the last years. We notice that in 2010, we recorded less than 2500 complaints per week, and the volume kept increasing over the years to reach more than 30,000 complaints per week in 2016. This shows that criminals are adopting more and more the telephony network infrastructure to reach and scams telephony customers.

Geographic analysis

In order to enrich the complaint dataset, we correlate it with other telephony data sources. Our goal is to leverage other information in our analyses, such as the geographic distribution. We mainly use the data sources provided by *Canadian Numbering Administrator* (C. N. A, 2015) and *North American Numbering Plan Administration* (N. A. N. P. A, 2015) to determine North American phone calls. In addition, we rely on the recommendation of the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T), namely E.164, to determine the country of origin. The dataset contains three main classes of phone numbers based on the origin location: *North America*, *non-geographic or toll-free*, and *international* (see Table 4). It is evident that a large portion of calls come from North America (see Fig. 6), as the complainers are mostly from the United States and Canada. The second largest portion of calls comes from toll-free or non-geographic phone numbers, for instance phone numbers starting with *800*-xxx-xxxx.

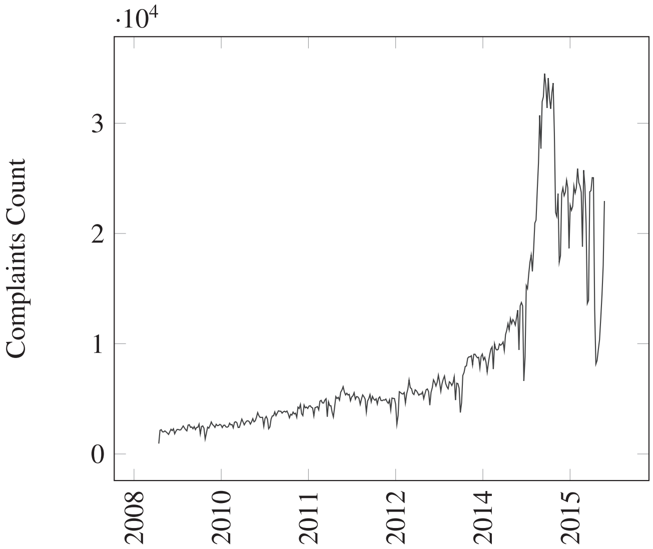


Fig. 5. Distribution of Complaints over Time (year on the abscissa).

Table 4
Calls distribution based on the geographic location.

United States	Toll-Free	Canada	International
61%	28.4%	3.3%	7.3%

Use cases: campaign detection

In our analysis, we applied our campaign detection approach to the complaint dataset. Subsequently, we detect many calling campaigns. Hereafter, we present a subset of the calling campaigns, which we will discuss and analyze in the sequel. The identification of the calling campaigns is a tedious task given the complexity of identifying similar telephony abuses. Fig. 7 shows a network graph representing calling campaigns. Within this graph, we have observed that some calling numbers are shared among campaigns and others are specific to some campaigns. In Table 5, we present some of the top discovered calling campaigns, as well as the number of complaints related to each one of them. In addition, we provide the first and last time of when the campaign appeared.

In the following section, we present more details on some of the top detected campaigns.

Tax fraud campaign

One of the big calling campaigns that we discovered through our approach is the tax scam calling campaign. Actually, it is one of the biggest campaigns that we detected in the analyzed complaint data. This campaign targets American citizens since the callers are pretending to be either from the Internal Revenue Service (IRS), the U.S. Treasury, or the Department of Legal Affairs. A similar type of campaign is found targeting Canadian citizens since the callers are claiming that they are from Revenue Canada or the Canadian Revenue Agency (CRA).

IRS Scam. Using our framework, we uncovered a telephony abuse campaign where the victims reported that the person calling them was claiming that he/she was from the Internal Revenue Service, the U.S. Treasury, or the Department of Legal Affairs. According to the complaint messages, the victims reported that the callers harassed them to immediately pay a tax that they owe to the government via wire transfer or check, otherwise the police would come to their home or some of their governmental papers would be revoked. An example of these messages were: *[…]I am officer Lauren Matthew from Internal Revenue Service, and the hotline to my division is 415-992-8009, I repeat, it's 415-992-8009. Don't disregard this message and do return the call before we take any action against you. Good bye and take care![…]*. Furthermore, we find that fraudsters are calling from different numbers several times claiming to be from the IRS. Fraudsters spoofed the caller identification of the organizations mentioned previously. They took advantage of the possibility of spoofing their caller identification to be one of the entities mentioned previously. We found that the most reported phone number, +1-213-**1-**63, was from Washington and has been calling victims under the name of the IRS 679 times between the 29th of June 2010 and end of 2016. This enormous number of calls from different numbers claiming to be the same person is an indicator of a dangerous calling campaign that must be dealt with. The IRS scam campaign was reported 26,024 times. According to our dataset, this campaign has been active using different source numbers since the first day of our data, January, 1st 2010, to April, 24th 2016. In 2010, TAINT recorded 14,876 complaints about the IRS scam campaign, with an average of 35 complaints per day. In 2016,

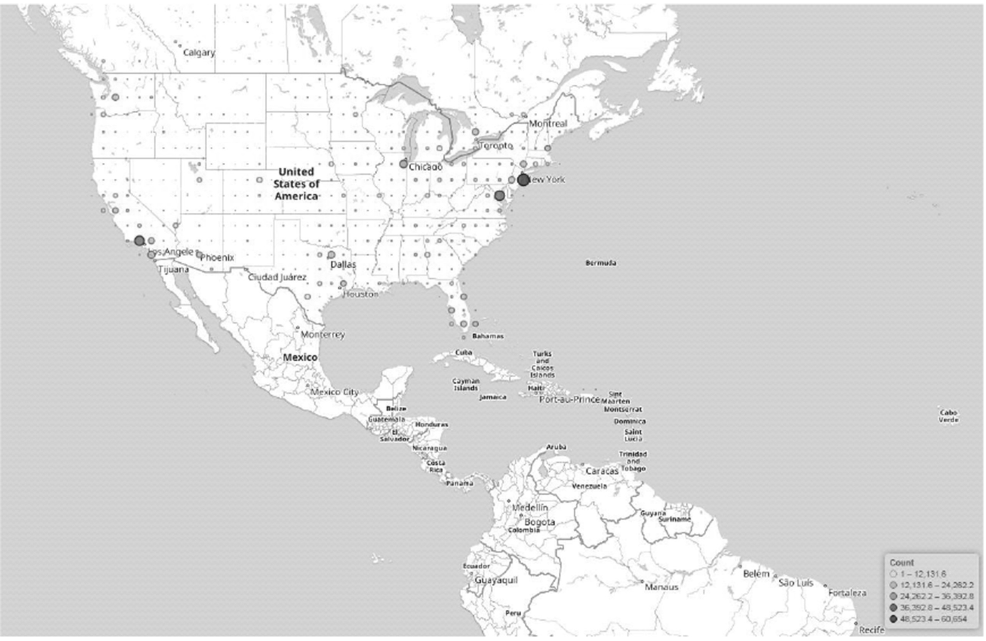


Fig. 6. Geographic distribution of reported source phone numbers for June 2017.

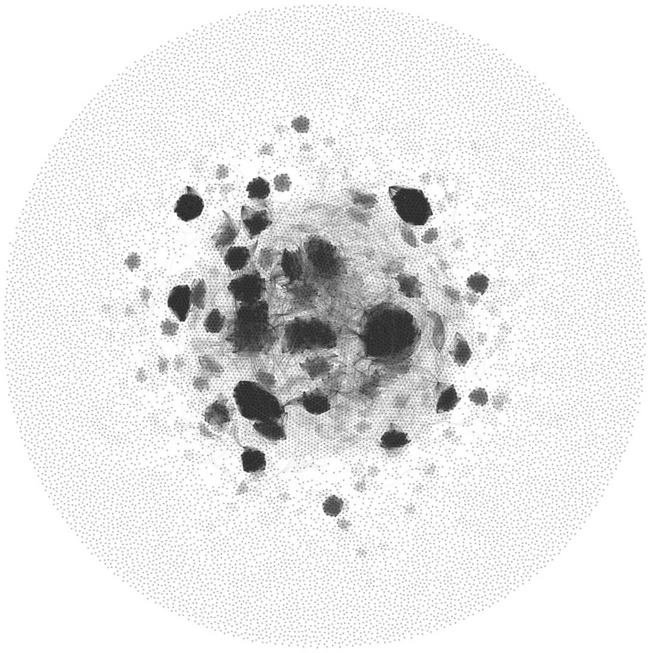


Fig. 7. Graph of the detected campaigns in 2016.

the number of complaints increased to reach 22,025 complaints with an average of 356 complaints per day.

Canadian Revenue Agency Scam. In this detected campaign, abusers have been calling Canadians while impersonating as the officers at Security Investigations Department of CRA and attempting to steal the victim's money or personal information. The victims also complained about receiving a voicemail for the investigation from a set of numbers, and they were asked to call back otherwise they would be arrested. We believe that raising awareness and informing people about this scam is something that should be taken into consideration. According to our data, this scam has been targeting people since 2010. Scammers used 626 distinct phone numbers to launch their attacks. Ottawa topped the list of the sources from where these numbers are generated with 145 different numbers which are involved in 1430 complaints. Considering our sample size, the high number of complaints, being 19,046 complaints, compared to the population of Canada shows that this is a very serious threat. We should also consider that there are many victims who, despite being scammed and having paid out money, do not make complaints, as well as individuals whose personal private information gets stolen and they remain unaware that they have been scammed. The theft of personal information can be dangerous and may be used by fraudsters for other criminal activities.

Telemarketing scam

For many decades, telephony abusers have been spoofing telemarketing companies. Fraudsters call customers and claim that they are from a given company and try to sell products cheaper if the customer pays immediately, or attempt to steal customers' bank information or even access customers’ computers (such as in the Microsoft Windows technical support scam). Hereafter, we provide more details and insights about one of the most severe telemarketing campaigns.

Criminals have gone from using simple telephony scams to more sophisticated scams. Fraudsters usually try to social engineer their victims in order to coerce them into divulging personal information or to pay an amount of money; however, fraudsters have now started using different tricks to achieve their goals. For a better and more efficient scam, scammers claim to be from a well-known company whose product is used by a maximum number of people to obtain better results. For instance, according to our results, a significant calling campaign has been spoofing Microsoft Windows in the past few years. Scammers claim that the victim has a problem with their product, has downloaded a dangerous virus, or has been hacked and attempt to gain access to the victim's computer. The scammers then use this privilege to steal the victim's information or to use their computer as a botnet node or for other malicious use. We identified 364 distinct phone numbers that abused 13,154 telephone customers.

Table 5
Subset of the detected calling campaigns.

Nature of the Campaign	Detected Campaign	# complaints	First Seen	Last Seen
Fraud Campaigns	Treasury Department/IRS/Department of Legal Affairs	26,024	2010-01-05	2016-04-24
	Canadian Revenue Agency	1961	2011-05-04	2016-01-30
	Air Canada	327	2013-10-10	2016-01-14
	Federal Express (FEDEX)	3496	2015-12-19	2015-12-24
	Ottawa/BC/Toronto/Quebec Hydro	958	2015-01-14	2015-12-14
Telemarketing Campaigns	Microsoft Windows	13,154	2010-02-15	2016-04-24
	Clearing House Publishers	1943	2014-12-15	2016-04-15
	Credit Card Services	5638	2010-01-05	2016-04-20
	Reward Redemption	2368	2014-06-12	2016-04-23
Political Campaigns	Canadian Parties Political Campaigns	942	2015-02-22	2015-10-19
Scam Campaigns	Caribbean Cruise Lines	340	2010-02-02	2016-03-27
	Global Lifestyles	390	2014-02-08	2015-11-30
	America West	272	2012-02-08	2016-04-06
	Powerball Lottery	24	2015-11-28	2015-12-09
	Carnival Cruise line	104	2010-04-27	2016-04-19
	Can you hear me?	5638	2010-01-05	2016-04-20
	Nigerian Scam	472	2014-12-16	2015-11-21
	Vanuatu Scam	27	2015-03-07	2016-01-12
	Prize Notification Center	315	2014-11-13	2016-04-05
	Bahamas Cruise	4222	2010-01-02	2016-04-21

Credit Card Service Scams. We identified ten credit card services fraud campaigns resulting in 5638 complaints. Customers received calls from fraudsters claiming that they are from cardholders services in order to steal victims' private bank information or to make victims pay a fee for a fake service. Other scammers have been telling their victims that their credit card account has been deactivated and they must provide their private information to reactivate it. Criminals usually use robo-calling to contact customers and then get the call and try to trick people using multiple techniques. We observed that they call various individuals numerous times and that there is a campaign involving 523 different numbers.

Political campaign

Political parties are using phones to reach electorate and promote their campaigns. However, we observed in our data that people are complaining about the randomness and high number of these calls. These calls have actually turned into abuse, as people are receiving an excessively high amount of robocalls.

Canadian Political Campaigns. Recent political calling campaigns were detected in Canada. Many individuals have been complaining about receiving numerous repeated calls from Canadian political parties. People have been complaining that firstly, these political parties are calling randomly and abusing people via multiple calls, and secondly, these parties are not even bothering to put a human on the call and are instead using a robocaller. We identified 942 complaints related to 4 different political promotion campaigns in Canada. These campaigns were active during the period of election. The very first complaint about these campaigns was seen on February 22, 2015, and the last complaint was seen on October 19, 2015.

Trip scam

Numerous calling campaigns have been reported trying to scam victims by offering them a cheap trip or by claiming that they have won a vacation. After analyzing the complaint data, we found some of these campaigns that will be explained in what follows.

Caribbean and Bahamas Cruise Line Scam. One of the detected scamming campaigns are the trip scam campaigns. A set of numbers and spoofed caller identifications have been calling people and offering them a Bahamas cruise trip; other numbers were offering a Caribbean cruise line trips. The victims of this scam have been reporting that callers are offering them a cheap trip if they pay immediately; others have been saying that the callers inform them that they won a trip, but in order to get the tickets, they have to pay an amount of money and provide their personal information. Fraudsters have been using this technique also to get people's money and confidential information. We observed that this campaign involves 121 numbers. Some of the numbers were originating from the Bahamas, which indicates that either the fraudsters have been spoofing numbers from the Bahamas so that the calls look legitimate, or that the scam really is from the Bahamas.

Conclusion

Internet telephony technologies have enabled new types of abuses among which telephony abuse is a prominent one. Criminals nowadays exploit extensively this channel in order to scam their victims. Complaints about telephony scams have been dramatically increasing over the last years. Scammers are using different characteristics of telephony networks such as caller identification and phone number spoofing, taking advantage of the low cost and the possibility of the spamming campaign to easily reach and deceive many telephone service customers. Different studies investigated email spamming and developed techniques to combat it. However, with the recent advent of telephony spamming, not enough research has been conducted on this important problem.

In this paper, we presented TAINT, an automatic framework for the near-real-time aggregation and analysis of telephony complaints to understand spammers activities and to detect telephony abuse campaigns. The system has been evaluated on a real large-scale dataset of more than five million telephony complaints from different sources. TAINT automatically aggregated and analyzed the data in order to extract patterns from telephony abuse, the geolocation distribution as well as the underlying campaigns by exploring the similarities among the abuse incidents. TAINT detected 1519 different calling campaigns that have been reported and that are causing their victims a lot of losses. We found that most of the calls were generated from the United States and Canada, which is reasonable since the complaints data were mostly collected from North American customers. Furthermore, we discovered that most of the calling campaigns were appearing continuously, such as IRS, CRA, and the technical support campaigns; however, some campaigns depended on the events, e.g., the political propaganda campaigns, which happened on a specific time. Some other scamming campaigns happened in a short period of time till people became aware of that, namely, *Nigerian scam*, *Vanuatu scam*, *Can You Hear Me?* Criminals used multiple phone numbers to attack their targets, and most of the detected scams were launched as campaigns. We observed that Caller ID spoofing was the main technique attackers relied on to craft their attacks. Finally, the insights gained from this research enhances our understanding of telephony abuse and our framework generates valuable intelligence that can be used to reduce this type of abuse.

References

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory. Exp. 2008 (10), P10008.

C. N. A. Consortium, Available at: http://www.cnac.ca/, (Accessed 11 January 2015).

Christin, N., Yanagihara, S.S., Kamataki, K., 2010. Dissecting one click frauds. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, ACM, pp. 15–26.

Costin, A., Isacenkova, J., Balduzzi, M., Francillon, A., Balzarotti, D., 2013. The role of phone numbers in understanding cyber-crime schemes. In: Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on, IEEE, pp. 213–220.

Gupta, P., Srinivasan, B., Balasubramaniyan, V., Ahamad, M., 2015. Phoneypot: data-driven understanding of telephony threats. In: 22nd Annual Network and Distributed System Security Symposium, NDSS 2015, San Diego, California, USA, February 8-11, 2014. http://www.internetsociety.org/doc/phoneypot-data-driven-understanding-telephony-threats.

H. How much phone scams cost Americans last year, Available at: http://www.marketwatch.com/story/heres-how-much-phone-scams-cost-americans-last-year-2017-04-19, (Accessed 23 August 2017).

Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques: Concepts and Techniques. Elsevier.

Irs nebraskans lost 56000 to telephone scam, Available at: http://nebraskaradionetwork.com/2016/02/17/irs-nebraskans-lost-56000-to-telephone-scam/, (Accessed 13 September 2016).

J. CMS, Available at: https://www.joomla.ca/, (Accessed 14 June 2015).

Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: European conference on Machine learning: ECML-98, pp. 137–142.

A. Kafka. Available at: http://kafka.apache.org/. (Accessed 19 February.2015).

Karbab, E.B., Debbabi, M., Derhab, A., Mouheb, D., 2016. Cypider: building community-based cyber-defense infrastructure for android malware detection. In: Proceedings of the 32nd Annual Conference on Computer Security Applications, ACM, pp. 348–362.

Maggi, F., 2010. Are the con artists back? a preliminary analysis of modern phone frauds. In: CIT, IEEE Computer Society, pp. 824–831. http://dblp.uni-trier.de/db/conf/IEEEcit/IEEEcit2010.html#Maggi10a.

Miramirkhani, N., Starov, O., Nikiforakis, N., 2017. Dial one for scam: a large-scale analysis of technical support scams. In: Proceedings of the 24th Network and Distributed System Security Symposium (NDSS).

Mongodb, Available at: https://www.mongodb.org/, accessed on: 21 January.2015.

Mustafa, H., Xu, W., Sadeghi, A.R., Schulz, S., 2014. You can call but you can't hide: detecting caller id spoofing attacks. In: Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on, IEEE, pp. 168–179.

N. A. N. P. A. NANPA, Available at: http://www.nanpa.com/, (Accessed 17 January 2015).

R. data structure store, http://redis.io, (Accessed 16 February.2015).

R. o. I. D. D. L. o. T. S. f. t.. F. S. Phone Scams Continue to be a Serious Threat, Available at: https://www.irs.gov/uac/newsroom/, (Accessed 13 February 2016).

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process Manag. 24 (5), 513–523.

scikit-learn Machine Learning in Python, Available at: http://scikit-learn.org/stable/, (Accessed 21 August.2016).

E. Stack, Available at: https://www.elastic.co/, (Accessed 12 January.2015).

T.. Newsgroups text dataset, Available at: http://scikit-learn.org/stable/datasets/twenty_newsgroups.html, (Accessed 13 August.2017).

Tu, H., Doupé, A., Zhao, Z., Ahn, G.-J., 2016. Toward authenticated caller id transmission: the need for a standardized authentication scheme in q. 731.3 calling line identification presentation. In: Kaleidoscope, I.T.U. (Ed.), ICTs for a Sustainable World (ITU WT), 2016, IEEE, pp. 1–8.

Warning over phone scam that cost this woman 70, Available at: http://www.telegraph.co.uk/money/consumer-affairs/warning-phone-scam-cost-woman-70000/, accessed on: 21 July 2017.

Y. W. into the Elastic Stack, Available at: https://www.elastic.co/products/kibana, (Accessed 16 January 2015).