Data-Driven Approach for Automatic Telephony Threat Analysis and Campaign Detection

Houssem Eddine Bordjiba

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering (CIISE)

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Information System Security) at

Concordia University

Montréal, Québec, Canada

September 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Houssem Eddine Bordjiba**

Entitled: **Data-Driven Approach for Automatic Telephony Threat Analysis and Campaign Detection**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Information System Security)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Amr Youssef

_____ External Examiner
Dr. Otmane Ait-Mohamed

_____ Examiner
Dr. Chadi Assi

_____ Supervisor
Dr. Mourad Debbabi

Approved by _____
Rachida Dssouli, Chair
Department of Concordia Institute for Information Systems Engineering (CIISE)

_____ 2017

_____
Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Data-Driven Approach for Automatic Telephony Threat Analysis and Campaign Detection

Houssem Eddine Bordjiba

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. 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 thesis, the design and implementation of a Telephony Abuse Intelligence Framework (TAINT) that automatically aggregates, analyzes and reports on telephony abuse activities. Such a framework monitors and analyzes, in near-real-time, crowd-sourced telephony complaints data from various sources. 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 thesis 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.

Acknowledgments

First of all, I would like to express my heartfelt gratitude to my supervisor, Professor Mourad Debbabi, for his wise guidance and continuous support throughout my graduate studies. Thank you for giving me the opportunity to grow not only academically, but personally and professionally as well. I was fortunate enough to work under your supervision for which I am very grateful.

My gratitude extends to the examining committee members, Dr. Chadi Assi, Dr. Otmane Ait-Mohamed, and Dr. Amr Youssef for their thorough evaluation and valuable feedback. I would also like to extend my thanks to all the excellent professors who taught me throughout my studies. I cannot forget to thank them for their help in bringing the best of me and keeping me moving forward.

Furthermore, I wish to express my utmost gratitude to all my lab-mates who gave me the motivation and created such a great environment to work in. The laboratory and my learning experience would not have been the same without you. I owe a special thanks to my colleague and great friend ElMouatez Billah Karbab who provided me with a lot of help and encouragement, whenever needed throughout this journey.

Many thanks to all of my friends for their sincere concern and friendship. I want to thank you for giving me every time a lot of encouragements. Moreover, I would like to send a special and warm thanks to my whole family. In particular, I must express special gratitude and appreciation to my uncle, Faouzi Drouiche, and his family. Words can hardly express how much I appreciate everything they have done for me during my master studies.

Last but not least, I feel a very deep gratitude towards my parents and my sisters, Yasmine and Rania, for giving me their unconditional affection and support, and continuous guidance throughout my life. This accomplishment would not be possible without you. You are so wonderful beings, the best of which I could hope to ask for in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

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 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 [58].

A better understanding of these types of abuses is required for detection, mitigation and attribution purposes. One way to gain such an understanding is to rely on the complaints filed by the victims to government agencies and telephony service providers.

According to [14], 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. The complaint data provides valuable information for investigators such as the nature and the type of the abuse. In this context, we have developed a framework that automatically collects, analyzes and reports on abuse activities. Such a framework monitors and analyzes, in near-real-time, complaints data from multiple sources. By doing so, the telephony abuse intelligence framework generates complete abuse intelligence that can support mitigation and/or investigation activities.

We subject to analysis the large dataset of telephony complaints in order to provide in-depth insights and intelligence on these emerging telephony threats. In real world scenarios, such analysis is highly beneficial to both incident handlers and law enforcement officials to cope with telephony abuse incidents.

1.2 Problem Statement

Recently, we have witnessed a significant rise in telephony abuse. In 2016, according to [28], Americans lost 9.5 billion dollars due to phone scams. These losses are the result of scamming campaigns that targeted approximately thirty-two 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 [60]. According to [47], fraudsters posed as the 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 [1]. 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 collect data about telephony abuse and how to analyze it 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?

1.3 Objectives and Contributions

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 [19] and the *North American Numbering Plan Administration* database [41] 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 source raw data and publicly available telephony databases; (ii) We use larger datasets compared to [36, 24]. Indeed, we used a dataset that is comprised of 5 million complaints whereas [36] used only a dataset of 300 complaints which they collected using their own developed web application; (iii) Our framework is an automatic and online solution, where a limited human interaction is needed and it aggregates near-real-time data and generated near-real-time intelligence whereas in [24] they had an automatic collection, yet an off-line analysis of the dataset collected using a telephony honeypot; (iv) This is the very first research contribution, to the best of

our knowledge, on the 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 open-source raw 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.

1.4 Thesis Organization

The remainder of this thesis is structured as follows. Chapter 2 provides background information, describes and analyzes the different telephony threats and presents a survey on some of the major work done on telephony abuse analysis. In Chapter 3, we present a description of our telephony complaints dataset. Then, we provide an overview of the architecture and design of our framework together with the algorithmics of campaign detection. Moreover, in this chapter, we describe and explain the back-end and front-end implementations of the proposed framework. Chapter 4, provides an in-depth profiling of the telephony complaints data, which gives key insights about the telephony

abuses. In addition, it presents an extensive evaluation of our framework with the underlying results. Finally, Chapter 5 provides the concluding remarks on this research along with a discussion of future research directions.

Chapter 2

Background & Related work

In this chapter, we present the definition of terms and essential concepts used in telephony abuse analysis and investigation. In Section 2.1, we explain the technical terms that are used in this thesis. In Section 2.2, we present an analysis of telephony threats. In Section 2.3, we provide the techniques that are used by researchers and the security community to collect data and information about telephony abuses. In Section 2.4, we give a brief description of the data mining and machine learning techniques used in our work. Finally, in Section 2.5, we present an overview of the existing work related to the analysis of telephony abuses.

2.1 Definitions

In this section, we describe the main technical terms that are used in this research work, namely VoIP, Social Engineering, Robocalling, Toll Free Number, Caller ID or CNAM, and Caller ID spoofing.

2.1.1 Social Engineering

Cyber criminals use multiple techniques to make a fraud or to perpetrate their attacks. Usually, this attacks have a schema composed of different steps needed to succeed in their attempt. The attacker relies on two factors, the human factor and the technology factor. For the human factor, criminals use Social Engineering which involves tricking people into performing specific actions that benefit the attackers such as disclosing personal and private information or providing access to an unauthorized

machine and so forth. Cyber criminals rely heavily on this technique in all types of cyber frauds and in cyber attacks.

2.1.2 Robocalling

Robocalling involves prerecorded phone messages that are used to make automatic calls. This method is usually used in telemarketing or political campaigns. Perpetrators leverage this option in telephony network to generate telephony spams, to simplify the spread of their attacks, and to be able to commit efficient phishing attacks.

2.1.3 Toll Free Number

Toll-Free Numbers are a particular phone numbers usually owned by businesses, government agencies, or individuals that offer services to distant customers, as they have a particular billing system. Contrary to the regular phone numbers, calls to toll-free numbers are charged on the callee rather than the caller which can satisfy the need of enterprises to serve their customers without them having to pay long distance calls. According to the Federal Communication Commission [4], in the USA, specific Numbering Plan Area (NPA) Code are allocated to these Toll free numbers, which are $800, 888, 877, 866, 855$ and 844.

2.1.4 Caller ID Name or CNAM

Caller Identification Name (Caller ID or CNAM) is a feature provided by telephone carriers to identify the caller. Banks, government entities, companies, etc. use the Caller Identification to authenticate themselves to their customers. This technology is very useful to both caller and callee; however, the scammers can alter the Caller Identification to substitute their identity.

2.1.5 Caller ID Spoofing

The action of altering the Caller ID to reach a malicious goal is called Caller ID spoofing. Criminals use this technique to deceive and scam their victims. The increase of telephony users and the availability of the Smartphone, and the many features of Voice over IP contributed to the effectiveness and the rise of this type of fraud [40].

2.2 Telephony Threat Analysis

Throughout time, cyber criminals used different techniques to abuse their victims, including but not limited to phishing (attempt to obtain sensitive information often for malicious reasons, by disguising as a trustworthy entity), pharming (fraudulent practice of directing Internet users to a malicious website that mimics the appearance of a legitimate one, in order to obtain sensitive information), the use of spyware, the use of malware, etc. To execute their illegal plans, cyber criminals use telephony as a new alternative medium to communicate with their victims. In the sequel, we present the techniques that cyber criminals use to scam and abuse their victims throughout telephony lines. Besides, scams of this kind are even more exacerbated with the large deployment of Voice over IP (VoIP) services. The cheap and free telephone calls encouraged these criminals to use these attacks to scam people by taking advantage of their trust of their seemingly-secure telephone services. Due to the increasing prevalence of such attacks, [45] explains to telephone users how they can differentiate between scammers and legitimate callers.

2.2.1 Voice Phishing or Vishing

Voice phishing, also known as *vishing*, is a technique used by criminals to scam and abuse telephone users. This technique is derived from phishing on the Internet. Phishing is a type of Internet scam where the attacker tries to impersonate a legal entity to steal Internet users' information and then use the information illegally. In essence, vishing is a form of phishing conducted through telephone technologies. It consists of calling random people and claiming to be an enterprise or a legal company in order to steal people's personal information. It is also the technique of leveraging Internet telephony technologies to craft a social engineering attack, by calling or sending a voice message to the victims and asking them for personal information.

2.2.2 SMS Phishing

This is another form of phishing where the criminals pretend to be a bank or any legal entity. They send to victims through SMS links to execute scripts or download malware on their mobile phones. Afterwards, the attacker uses malware or scripts to collect sensitive information like banking

credentials, credit cards, emails, contacts information and social insurance numbers. *Smishing* is the act of phishing through short message services. Authorities have received a large volume of complaints about such abuse [32].

2.2.3 Spam over Internet Telephony

The emergence and the huge recent advances in telephony technologies have also led to the appearance of the spam over Internet telephony (SPIT). Criminals make a massive number of calls to victims to mislead them into purchasing an abused service or brand items. Spammers use different features of VoIP, such as interactive voice response, simultaneous calling, and many others to craft these attacks.

2.2.4 Telephony Denial of Service

Network denial of service attack is the act of flooding a victim's network with illegitimate malicious traffic to prevent real users from accessing it. Similarly, since organizations use telephone services to promote and sell their products, criminals started to use Telephony Denial of Service (TDoS) attacks against their victims. In such attacks, they flood their victim's telephony network with a substantial number of calls so as to make it unavailable (saturated) for some period of time, which results in undesirable consequences, financial losses or even the endangerment of human lives [9].

2.2.5 Telephony Scam Examples

Debt Collector

Debt collector scams are used by some criminal organizations to call their victims and ask them to pay money they do not actually owe [5]. These are actual scams as they are directed to debt-free persons, as stated in [44]. The authors of these calls have abusive, aggressive, and harassing behaviors. In Canada, according to the same source [7], various complaints have been filled by many people to Canadian Radio-television and Telecommunications Commission (CRTC) and Royal Canadian Mounted Police (RCMP) against the perpetrators of such harassing phone calls. In [6], it is reported that one of the big collection agencies in Canada was fined $500,000 after CRTC received many complaints from users that have been harassed by automated phone calls without having debts.

Telemarketer

Telemarketing calls can be divided into two types according to [56]: Regular telemarketers, which are business companies that try to sell or promote their products by calling different potential customers, and exempt telemarketers, which are political parties, charities, survey, research companies, etc. For instance, in Canada, CRTC [3] defines how a person can make a complaint about a telemarketer if they violates certain rules, mentioned in [48]. According to [24], telemarketers call customers, try to sell their products and dial off when they feel that the deal will not work. However, from the complaints that have been filed [43], we can see that abusers try to harass customers and steal their personal or financial information such as their credit card numbers during these phone communications.

2.3 Telephony Abuse Datasets

The increasing damage caused by cyber criminals through telephony network in the last decade has pushed researchers and security specialists to study and analyze telephony abusers activities. In this pursuit, researchers used different techniques to collect relevant telephony abuse datasets, which then became crucial to combat telephony abusers. Among the datasets used to infer intelligence on telephony abuser's exercises, there are manually collected datasets from telephony complaint websites, and automatically collected datasets through built-in collection frameworks, honeypots, and honeycards. In this section, we present different telephony abuse datasets, and the various techniques used to generate and collect these datasets for research work and other aims.

2.3.1 Complaints Data

Complaint data is a valuable source of information for enterprises to improve their customers service [22]. In our research, the complaint data refers to a set of complaints made by real world victims regarding any given issue. The complaint contains different information about the victim, the abusers, and the abuse. Researchers [24, 39] use this source of data to analyze and understand different abuses for detection, prevention and mitigation purposes. For this thesis, the complaint data was obtained from two different sources. The first one comes from the public complaint websites. The second one comes from specialized collection frameworks built by researchers.

Complaints Websites

Complaints websites are public websites that are created for the purpose of receiving complaints from end-users. Telephony abuse complaints websites were used by researchers [24, 39] for the analysis of telephony threats. These open source websites [43, 55, 35] contain valuable information about the telephony abusers. Victims usually provide the phone number, the caller identification, and a message describing the behavior of the caller.

Collection Frameworks

Collection frameworks are systems or web applications built by researchers and security professionals to collect information about a given problem. The information collected through these systems allows for the understanding of the nature of the problem and the behavior of the attackers, and helps to further analyze this issue in order to improve future detection and prevention [36].

2.3.2 Honeypots

Telephony Honeypot

The telephony honeypot is a set of unassigned/unused phone numbers that collects data from various telephony service providers. The telephony honeypot is a solution to collect abuse data over voice channels. The information collected through this honeypot can be aggregated to identify trends and provide insights into unsolicited calling patterns in order to better understand the origins of spoofed calls and other telephony abuses. In addition, it helps in the early detection of telephony abuses and helps to fingerprint malicious calls [24].

Mobile Telephony Honeypot

The Mobile telephony honeypot (MobiPot) is a honeypot made specifically to monitor and collect data about abuses targeting mobile users. The drastic increase of mobile subscribers, which already exceeded the world population, made mobile phones a great medium for cybercriminals to reach their victims. In contrast to common telephony honeypots [24], MobiPot allows the security community

to discover attacks targeting mobile phone numbers, in addition to its ability to have an automatic engagement with the suspicious callers [12].

2.4 Data Mining

Data mining is the process of extracting valuable knowledge and information from a significant amount of data. The term mining refers to the grave harvesting of data to extract valuable results. In the literature, data mining is referred to as Knowledge discovery from data, which better reflects the process and the objective of the technique [26]. The knowledge discovery process goes through different stages: data cleaning, data integration, data selection, data transformation, knowledge discovery, pattern evaluation and knowledge presentation. In this thesis, we have designed and implemented our framework using multiple data mining techniques, in order to extract valuable intelligence from people's complaints. The intelligence generated by our system can be used by security professionals and law enforcement to detect, mitigate and prevent telephony abuses.

2.4.1 Classification

Classification is a data mining technique aimed to divide data into different classes. The classification algorithms rely on a labeled training dataset to build the classification model, which will predict the classes of new data based on the training dataset. In this thesis, we use text classification, which is the classification of text documents [26].

2.4.2 Regression

Regression is a supervised machine learning technique that divides the dataset into different classes. The regression model is built based on a training dataset relying on two variables: the explanatory variable x, which is the feature vector, and the dependent variable y, which represents the class of items in the dataset [26].

2.4.3 Frequent Pattern Mining

Frequent Pattern Mining is a technique that studies and detects the pattern of items in a dataset. This technique provides two applications, (i) Frequent itemset, and (ii) Rule association mining. Frequent itemset provides the appearance frequency of items in a dataset in a given time span. Rule association mining presents items co-occurring together in the dataset. In order to control the results of these techniques, different parameters such as minimum support, which defines how much the appearance threshold of an item to be taken into consideration, and association rule confidence, which sets a threshold of items co-occurrence [26], are used.

2.4.4 Text mining

Text mining is a data analysis field that focuses on extracting knowledge, valuable information, patterns, etc. from the analysis of text documents. It also involves the use of different data mining and machine learning techniques such as classification, clustering, and different other algorithms such as topic modeling, sentimental analysis, document summarization, etc. to achieve the analysis purpose [26]. Text mining process goes through multiple steps before the final results. First, the text is parsed and structured. Then, text preprocessing is completed using different techniques such as tokenization, stemming, and removal of stop words. Afterward, features engineering is applied to the preprocessed documents, which in turn involves different techniques related to text such as, Term Frequency or Term Frequency minus Inverse Document Frequency (TF-IDF) to create the word feature vector. Finally, machine learning, data mining or other data analysis techniques are applied to the data for classification, clustering or other purposes.

2.5 Assessment of Telephony Abuse

The drastic increase of telephony abuse has led to different studies to understand its nature and its impacts in order to develop appropriate mitigation techniques. However, due to the complexity of the telephony infrastructure, which is composed of telephony landline, mobile telephony, and telephony over Internet Protocol, the security community is far from controlling and defending against this threat. Abusers have been using different techniques and tricks to scam telephony

customers. Examples of these types of scams are the IRS scam, credit card scams, and many others, which are launched sometimes into campaigns. Therefore, the analysis of different telephony scams and the explanation of how they work is a primordial step in detecting and mitigating them.

In this section, we present the major works done on the assessment of telephony fraud and compare it with work done on email fraud. Then, we present the techniques and proposed work to defend against telephony threats. Furthermore, we explore other cyber security areas, where telephony was used as a mean for scams or fraud. Lastly, we describe the work completed on the analysis of the telephony abuse and the results characteristics of the variable threats.

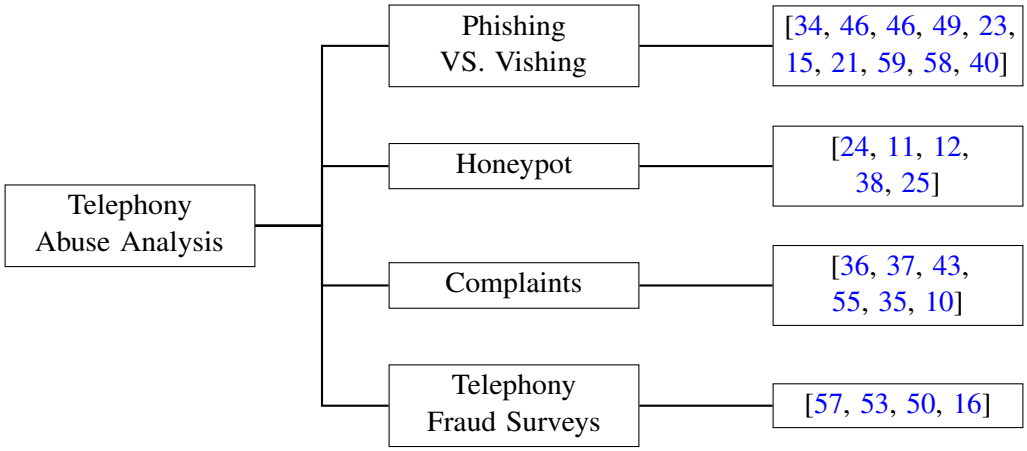

Figure 2.1: Classification of Telephony Abuse Analysis Studies

2.5.1 Comparaison Study between Email and Phone Fraud

Many research papers have discussed the security issues related to email spam and phishing attacks, such as in [34] and [46]. In [34], Khade and Shinde used data mining to detect phishing websites. On the other hand, Pandey and Ravi [46] used text mining to detect email spam. The authors of these works explained the difficulty of preventing phishing attacks since criminals use social engineering techniques, which are by nature difficult to detect and combat. In addition to the fact that Internet users lack awareness on the techniques these criminals use. These kinds of attacks went from using the Internet to explore other channels, such as telephony lines. Figure 2.2 illustrates the percentage of abuser's method of contacting their victims between 2012 and 2016 according to the Federal Trade Commission [49] reports; this bar graph shows how abusers are increasingly adopting the phone

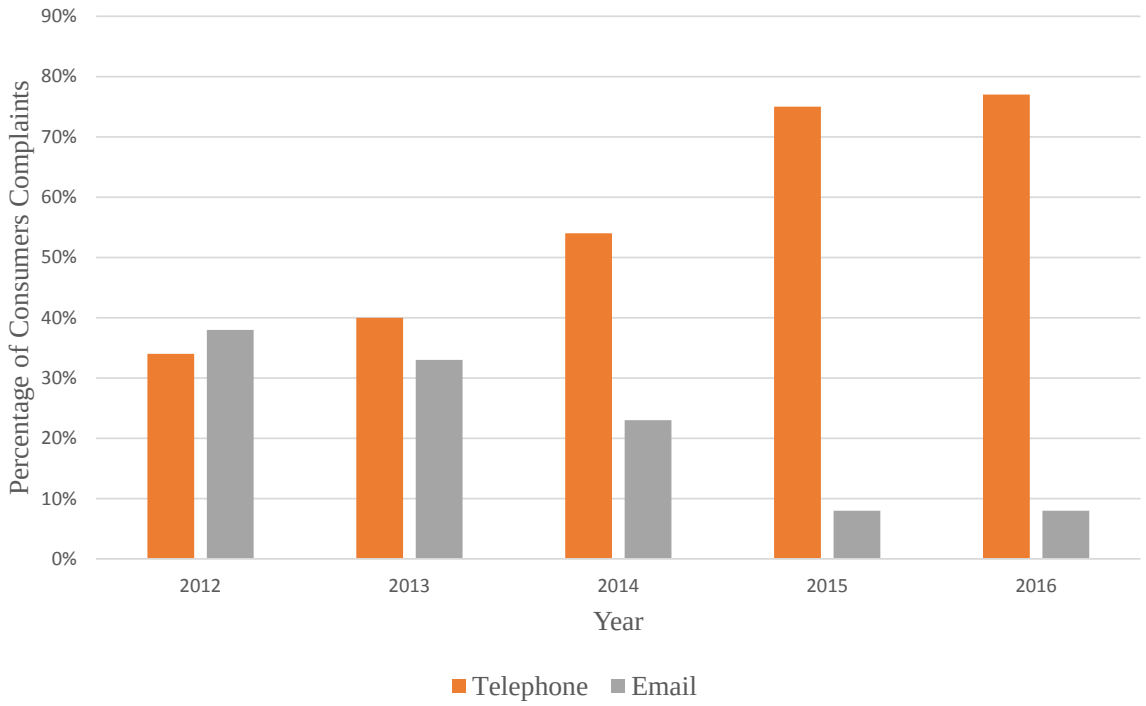

Figure 2.2: Statistics on Channel used to fraud victims– Phone vs Email [49]

service as mean of approaching their targets instead of the traditional email. Moreover, very few studies have discussed these new telephony threats. Therfore, in this thesis, a key objectives is to understand these threats as well as to provide further intelligence on these kinds of abuse. In order to attain these objectives, a framework was built and real-life data was used. Griffing and Rackley [23] explored the new alternative medium for Internet phishing, namely voice phishing (vishing). The authors showed how, with little knowledge and cost, a vishing attack could be conducted. The work also explains that the financial risk-to-reward ratio of vishing attack was high. The authors explained the techniques that could be added to the attack scheme to make it highly successful, and discussed how attackers spoof large banks or internet providers, as examples. Furthermore, in [15], Chanvarasuth conducted an experiment where he compared the effects of phishing and vishing on Thai students. The results showed that students are vulnerable to both phishing and vishing attacks. Also, many other researchers have studied the security issues related to telephony scams, such as in [21] and [59]. In [58], Tu et al. proposed an authentication scheme to control Caller ID and stop unwanted spoofed calls. The objective of the proposed scheme is to offer the possibility of securing the calling line identification Q.731.3, and help to guide in the future development of a standardized

scheme in authenticating SS7 identities. In 2014, Mustafa et al. [40] created an android mobile application *CallerDec* that detects calls that uses a spoofed caller ID in combination with Public Switched Telephony Network (PSTN landline) and cellular network information.

2.5.2 Studies Relying on Honeypots

One of the most used techniques to collect data on cyber security events is Honeypot systems. Gupta et al. [24] enumerated the drawbacks of the crowd data sources as follows. First, even though the victims reported numerous complaints on these websites, nothing can confirm their *completeness* since it is hard, if not impossible, to assume that all the victims reported their complaints on these websites, owing to the facts that some victims do not have Internet access and other users do not know the existence of such websites. Second, making a complaint in these websites is not something difficult to do, which affects the *accuracy* of the data. People can make complaints that are not related to telephony onslaught; besides, it is difficult to demonstrate who are the people making these complaints since it could be the attackers themselves, trying to confuse the data. Lastly, victims usually report the abuses after being harassed by numerous calls, which causes a delay between the time when the call was made and when it was reported; hence, one cannot confirm the *timeliness* of the user's supplied data. Subsequently, the authors implemented a telephony honeypot to collect data about telephony abuses. The authors discussed the benefits and the limitation of such honeypot. According to the authors, the honeypot collected much more data about abusers than what has been found in the reported complaints to FTC. The honeypot collected and reported the abuses of telephone lines. This telephony honeypot collected call data from unassigned phone numbers. By doing so, the telephony honeypot generated complete, timely and accurate telephony abuse data in comparison with crowd sourced datasets. Nevertheless, in this study, the researchers used manual analysis by performing some off-line statistics on the collected data.

Bladuzzi et al. [11, 12] studied telephony fraud by focusing on mobile telephony threats in Asia, specifically China. They presented a mobile telephony honeypot namely *Mobipot*. Similar to the work done by Gupta et al. [24], this honeypot was introduced to collect suspicious calls and SMS messages targeting mobile users. Mobipot collected more than two thousand calls and SMS messages

after more than seven months of deployment. The authors analyzed the collected results and showed how more than fifty per cent of the calls were suspicious. Afterward, they identified patterns of the behavior of criminals attacking mobile users.

Marzuoli et al. [38] presented threat intelligence that was extracted from the analysis of telephony scam data. Their dataset was collected from a telephony honeypot that received more than one million calls. The results of this analysis showed that half of the robocalls were only generated from 8 distinct abusers. Gupta et al. [25] built a system that collects and correlates information on telephony customers from multiple applications, such as, social media and email, in order to craft phishing attacks. The system then determines which over-the-top content (OTT) messaging application they can reach the victims through, to execute their vishing scheme. This showed the effectiveness of such a system and the simplicity of launching such attacks by relying on phone numbers.

2.5.3 Studies Relying on Users Complaints

Crowdsourced datasets are playing a big part in combating telephony threats. Researchers used these large databases to investigate different telephony related attacks. For instance, *800notes.com* [43], *Callercomplaints.com* [55], *FTC complaints* [10], *whocallsme.com* [35], and some others are used to investigate suspicious sources. The different datasets contain complaints about telephony abuses where the victims indicate the telephone numbers of the sources and report information about these numbers. In 2010, Maggi [36] designed and implemented a framework to collect complaints about voice phishing attacks. Contrary to honeypot collected datasets, this study vetted the customer's complaints by accepting it after investigating the complainer, which gave it more accuracy. Besides, the researchers asked the victims to provide the information they wanted to generate; such as, the caller ID, the spoken language, the country of the source, and to transcribe the conversation. This study permitted the collection and provided statistics about telephony phishers. Through this implementation, they succeeded in collecting only a small dataset of 300 complaints compared to the crowdsourced datasets. Therefore, more advanced analysis on a larger dataset would be of keen interest and would yield beneficial intelligence. Finally, Maggi et al. [37] implemented a system that contains two small modules to collect phishing information. The first module captures telephone

conversations about voice phishing, and the second module captures phishing emails. Then the authors compared both techniques to understand their potential applications. Their work showed the efficiency of voice phishing over traditional email phishing techniques. Nevertheless, the main drawback in all of these studies remains the small size of the dataset.

2.5.4 Surveys on Telephony Fraud

Sahin et al. [50] carried out a comprehensive study and provided a taxonomy of fraud in the telephony network. The taxonomy divides the features of telephony fraud into, the causes, the weaknesses, the techniques, the fraud scheme and the fraud benefits. In their work, Tu et al. [57] tackled the issue of Spam in telephony network (SPIT). The authors surveyed the existing techniques to combat telephony Spam and compared it to email Spam. Their work showed that there is no effective solution to prevent SPIT yet. However, some methods or combination of some techniques showed a good level of effectiveness. Other studies have applied systems that present countermeasures to telephony threats, such as in [53], where the author proposed a framework to detect spoofed calls. Recently, J. Chaudhry and S. A. Chaudhry, [16] enumerated the different techniques to prevent Caller ID spoofing, and provided a comparison between these methods. Then, they presented some areas to control Caller ID spoofing, and proposed a solution that rely on strengthening the identification security in telecommunication networks.

Chapter 3

Chasing Telephony Abuse: Analysis of Users' Complaints to Profile Threats and Identify Campaigns

3.1 Introduction

The emergence of technology and the Internet have changed everything in our daily life, from business services and learning to social interactions. Nonetheless, although the digital world made our life much easier since we can engage in all our activities using different technologies, we are exposed to a new lifestyle, with little knowledge concerning its risks and dangers. Precisely, the appearance of commercial online transactions led to new types of crimes called "Cybercrimes." Among these, we can name phishing, pharming, spamming activities, etc. For instance, phishing is a criminal act where the phisher pretends being a legitimate person from an enterprise, bank, or government agencies in order to scam people into giving their private personal information or sensitive data. A credit card number, social insurance number, and other personal information are used by these criminals illegally afterward. Fraudsters applied some of the cyber-attacks by sending their victims spoofed emails that link them to fraudulent websites. The later can look like the legitimate website of a bank or a company, but by doing so, they trick people into divulging their

financial or personal information.

Cyber criminals exploited the Internet to communicate and social engineer their victims by sending them emails or interacting with them through social networks. However, a few years ago, cyber-criminals started using another channel to reach their victims, namely telephony.
This is mainly due to the fact that Internet telephony technologies and services have lowered the costs of voice services and provided a platform for innovative services and applications. Internet telephony has also enabled abuses such as new forms of automated bulk call generation that can disrupt telephony services. In addition, it has enabled more effective and efficient forms of unsolicited telemarketing that integrate interactive voice response systems and dialing algorithms. Moreover, cyber criminals use phishing techniques by calling or sending SMS to telephony users as a mean to steal their credentials. These challenges are further complicated by the inability of telephone users – virtually everyone – to verify the calling party information, which provides a degree of anonymity for abusers. Furthermore, the new availability of Voice over Internet Protocol (VoIP) helped in the increase of swindler's activities due to the protocol's many features; such as the anonymity of the offender and the difficulty of tracing the number back. Moreover, the telephony abuses are increasing over time due also to the relatively low rates of the telephony communications. Many studies were done previously to assess different types of cyber-attacks, but only a few studies were done on telephony abuses. As a result, a better understanding of these types of abuses is required for detection, mitigation and attribution purposes.

To this end, the primary objective of this thesis is to develop new techniques, together with algorithms and tools, to contribute to the analysis of this new emergent cyber threat. This Thesis lies in the domain of telephony abuse analysis and campaign detection using newly developed and designed systems, along with a set of existing data mining techniques. The telephony abuse analysis framework can be an effective solution to analyzes, investigates, and potentially mitigates the abuses over voice channels. However, the complaints data filled to government agencies and to telephone service providers can be overwhelming and noisy especially if it is aggregated from different sources. The information collected with this framework will be aggregated and analyzed to identify trends

and provide insights into unsolicited calling patterns to understand the origins of spoofed calls and other telephony abuses. Furthermore, in this thesis, we explain how our automatic framework take as an input the complaint's data coming from different sources, and output clusters of phone numbers having some different extracted features, to help in the early detection of telephony abuses, and identify malicious calling campaigns. We suppose the importance of such framework since to the best of our knowledge no research work did such an automatic and real-time analysis on telephony threats. We believe that such work can assist investigator and legal entities to deal with these new type of onslaughts. Besides, it will help the security community to combat more effectively such emergent threats.

3.2 Dataset

We secured complaint data in near-real-time from our partners; an average of 2,000 complaints is received per day. This number increased to more than 8,000 in 2016. Thus far, we gathered more than 5 million complaints during a 7-year period. Our dataset contains more than five million complaints. The script logs that received complaints contain multiple attributes such as the source phone number, the victim phone number, the time when the complaint was made, the call type, the caller identification, and the message expressing the underlying complaint. Table 3.1 presents the attributes of the complaints together with their description.

3.3 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

Data Field	Description
Source Number	A string containing the Calling Party Number
Victim Number	A string containing the Callee Party Number
Call Type	A string containing the Call Type assigned by the victiim
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

Table 3.1: Description of the Collected Information

efforts for the detection, mitigation and attribution of telephony abuse activities that are perpetrated by telemarketers, debit collectors, scammers, etc. In this section, we present the architecture of our framework and its main components. Figure 3.1 provides a high-level overview of the framework and the interaction between its essential components. TAINT different components are enumerated hereafter:

(1) Telephony complaints text features extraction and classification

(2) Correlation

(3) Online features extractor

(4) Badness scoring of the scammers infrastructure

(5) Campaign detection

3.3.1 Telephony Complaints Text Features Extraction and Classification

The first component of TAINT aims to classify the complaints into two classes: *telephony abuse related complaints*, or *other subject matter complaints*. Our manual analysis of the dataset used

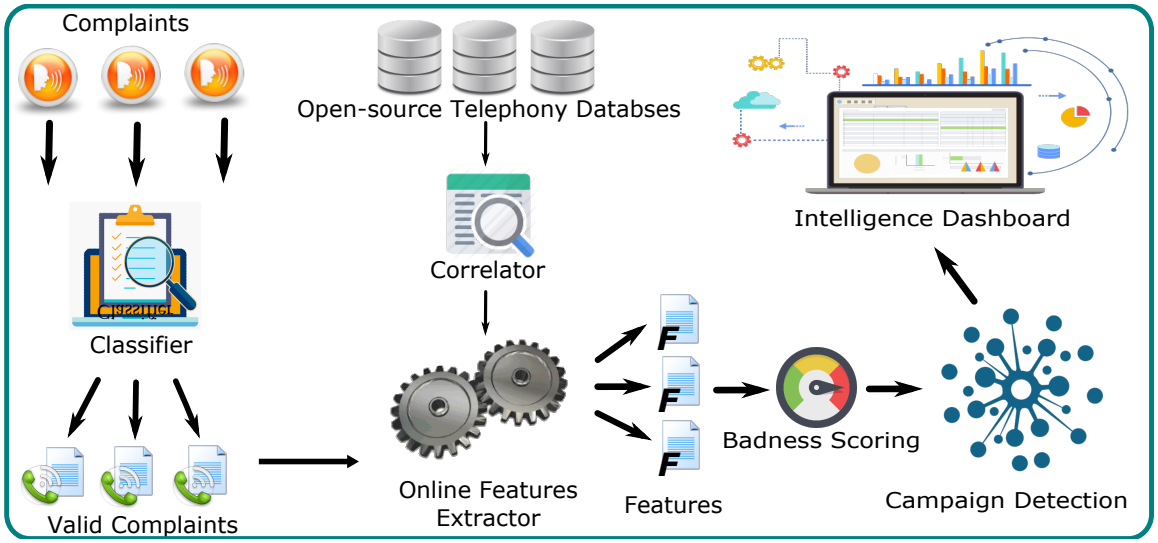

Figure 3.1: TAINT Framework Overview

in this work revealed that it contains a number non-related telephone abuse complaints. Since, the collection of the complaints is through a web page, 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 figure 3.2.

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.

Figure 3.2: Screenshot of a None Telephony Abuse Complaint

In addition, we assume that law enforcement agencies and telephone service providers receive numerous complaints about different problems their clients 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*[42] text dataset. This dataset contains more than 18000 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 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.

Afterwards, the features extracted through the text pre-processing 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 [30]. Furthermore, text classification relies on word vector features, which results in a multidimensional data analysis. Joachims [30] 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 [51].

3.3.2 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)* [19] and the *North American Numbering Plan Administration (NANPA)* [41] 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.

3.3.3 Online Feature Extractor

The complaints are streamed to the feature extractor, which is a near-real-time component in TAINT framework with the main role of online features computation. There are many features that need to be computed from the complaints stream to provide various analytics. Some of the features do not inherently support online computation such as *frequent pattern mining* features [8, 27]. Accordingly,

we implement other frequent item set algorithms [17], which rely on graph representation in memory. The main difference between our implementation and the one proposed in [17] is the use of a persistent graph database instead of a graph-based representation. Further details on the implementation of the frequent item set will be presented in the implementation Section.

The output of the features extractor is two streams of features: one stream for phone numbers and the other stream is for caller identification. These streams of features are used to feed a clustering system. The aim of using a clustering system is to automatically group the complaints data into clusters that give a clear grouping in each cluster rather than grouping all the items, which can be overwhelming. On the other hand, the generated clusters represent calling type categories such as invalid phone numbers, and spoofed phone numbers.

Extracting Features for Clustering

In order to be able to group relevant suspicious phone numbers and the caller identification, we need to identify clustering features. The latter are used to cluster the phone numbers according to defined criteria. The features need to be extracted on the fly from the complaints stream. The complaints input stream is generated after receiving a complaint from a victim. We perform an offline analysis on a seven years dataset from the complaints feed. We study the static and the dynamic behaviors of the calling phone numbers to determine the features that are more likely related to the level of suspiciousness. Evidently, the earlier the system detects a relevant threat the sooner the threat can be mitigated.

The result of analyzing a relatively large dataset of the telephony complaints is the extraction of **17** different features that we use to provide various analytics and estimate the damage caused by the reported phone numbers and caller identification.

There are **8** statistical features shared between the phone numbers and caller identifications. We define **17** features for phone numbers and **12** features for caller identifications, as we will present in the next section.

Feature Set	#	*Feature Name*
Temporal Features	1	Daily occurrence
	2	Week day occurrence
	3	Hourly occurrence
	4	Day hours occurrence
Targeted Victims Features	5	Number of distinct values
	6	Number of distinct regions
	7	Average calls of the targets
	8	STD calls of the targets

STD = Standard Deviation.

Table 3.2: Common Features

Features Selection

To determine the relevant features for our analysis of the telephony complaints, we analyzed the complaints dataset and the involved phone numbers and Caller Identifications. Then, we extracted **17** features that can be used to generate intelligence from the phone numbers and caller Identifications that are relevant for investigations. There are **13** features (Table 3.3) to group the phone numbers, and **12** features (Table 3.4) for caller IDs. Among these features, **8** features (Table 3.2) are shared among phone numbers and caller IDs. Clearly, the grouping can drastically decease the overwhelming and nosiness of the collected complaints. In the following, we present the statistical features used in TAINT. We first present the common features between the phone numbers and the caller IDs, as shown in Table 3.2. Afterwards, we present the phone number features and caller IDs features, as depicted in Table 3.3 and Table 3.4 respectively.

Common Ranking Features

In what follows, we present the common statistical features between the caller IDs and phone numbers which can be categorized into: Temporal features, and targeted victims.

Feature Set	#	*Feature Name*
Caller ID Features	9	Number of distinct values
	10	Average calls
	11	Calls STD
Phone String Features	12	Source and target list similarity
	13	Source and target set similarity

STD = Standard Deviation.

Table 3.3: Phone Numbers Features

Feature Set	#	*Feature Name*
Source Phone String Features	14	Number of distinct values
	15	Number of distinct regions
	16	Distinct values average calls
	17	Distinct values STD

STD = Standard Deviation.

Table 3.4: Caller Identifications Features

Temporal Features

In this category of features, we analyze the time, in which the complaints have been made. The complaint time does not provide enough information to detect the pattern of a calling campaign. For instance, a TDOS campaign will call a phone number multiple time in a short period; whereas a telemarketing campaign may call a phone number the same number of times but within a larger period. The distribution of complaints about time, in which the system received more complaints reporting a particular phone number or caller identification. In particular, we analyze the distribution of the complaints that are reporting a specific phone number or caller identification in four different time spans as described in the following:

- *The Daily Occurrence:* The percentage of the days in which the system received complaints reporting a phone number or caller identification. From our manual analyses, we found that there are phone numbers reported same number of time. However, the distribution is very

different; some number were reported only over one day and others have sparse distribution. Using this feature, we can distinguish between the temporal behavior of each phone number.

- *The Week Days Occurrence:* The percentage of the days of the week in which the telephony system received complaints reporting a specific phone number or caller identification. In this feature, we focus on the weekly distribution of the complaints reporting a phone number. Similar to what [24] reported, the telephony abuse intelligence framework receives complaints on all weekdays. However, only a small amount of complaitns is received during the weekend. Using this feature, we differentiate between phone numbers that are abusing victims during the weekend and those that are not, which could distinguish the purpose of the call.

- *The Hourly Occurrence :* The percentage of the hours from all the complaints life span hours in which the telephony complaints websites received a complaint from the phone number or the caller identification. This feature shares the same goal with the *daily occurrence* feature but with more granularity. This helps in generating quick alarms about specific detected abuse or campaign.

- *The Day Hours Occurrence:* The percentage of the hours of the day when the system received complaints reporting a phone number or caller identification. The hourly occurrence during the day spots the light on the distribution of the calls over the day. This feature could be very insightful concerning the time zone of the phone numbers. We notice during our manual analysis that the peak calling time for phone numbers could be different.

There are various data mining algorithms for extracting frequent patterns, such as Apriori [8], FP-growth [27], and ECLAT [61]. However, these algorithms compute the frequent item set on the whole dataset, which is very expensive. We leverage the frequent item set algorithms to extract temporal features; however, we implement a lighter version of a frequent item set in order to build an online frequent item set extraction. This will be further discussed in the implementation section.

Targeted Phone Number Features

The complaints records generated from the complaints dataset consist of the source and the target of the abuse. The source is the phone number that was reported by the complainer. In this category of features, we focus on the targeted victims. Clearly, these features could give an insight about how suspicious the phone number or the caller identification. It is evident that calling many people is more suspicious than calling one targeted phone number. Moreover, this could give a more clear idea about the nature of the abuse. In the following, we define the features related to the targeted victims:

- *Number of Distinct Values:* We believe that the number of customers phone numbers that are targeted is an important feature for two main reasons. First, it could help detect the abused very early when the source phone number or caller ID is calling multiple phone numbers. Second, it could spot the light on the nature of the abuse. For instance, we found in our manual analysis a call pattern, which is manifested by calling one targeted phone number intensively for a period of time; This could be a TDOS (Telephony Denial Of Service) attack. Another example is addressed in [24], where the authors give an example how we could use the number of targeted phone numbers to distinguish between the telemarketers and debt collectors.

- *Number of Distinct Regions :* To enhance the previous feature, we define the distinct regions of the targeted phone number that have been called. Relying on this feature, we could gain information about the geographic region being targeted that comes from specific phone numbers or a caller identification. We found some reported phone numbers calling multiple targets in the same region, while others call multiple targets phone number in different geographic regions.

- *Average Calls of The Targets:* To compute this feature, we divide the total number of calls on the *The number of distinct targets phones* which give the average calls to each victim phone numbers. We consider this feature and the next feature *the standard deviation* as complementary features. The aim is to gains insight on the number of different victims targeted by the given number.

- *The Standard Deviation of the Targets Calls:* By computing the standard deviation, we could

define if the phone number or the caller identification is calling the targets phone number uniformly or randomly. The value of this feature indicates how much the call varies from the average calls – the previous feature – that have been made to the target phone numbers.

Phone Numbers Features:

In addition to the common features, there are some specific features to the source phone numbers. We define two main categories: the caller identification features and the phone string features, as we are going to present in the following:

Caller Identification Features

The caller identification, which the source phone number is using, is more likely a key of defining the characteristics of a specific phone number. The following features focus on caller ID as a part of the phone number ranking features:

- *Number of distinct values :* How many Caller IDs the phone number used, is an important metric to understand the abuse of the phone number. We found, from our manual analysis, that there are phone numbers with different caller identification. This shows that these phone numbers were use in different calling campaigns, thus these numbers are more suspicious and should be investigated. On the other hand, other phone numbers use the same caller identifications, such as a bank name, and are involved temporary in one particular calling campaign.

- *Number of distinct regions :* The geographic spreading of phone numbers is an important feature to evaluate the abuse of a given caller identification. The wider is the geographic spread indicates a bigger campaign.

- *Average calls :* In this feature, we compute the average calls that has been made by each caller identifications. To compute this feature, we divide the number of total calls of a phone number on the number of its caller identifications. This feature helps to easily identify whether a phone number is involved in one campaign or many.

- *Calls STD :* Using the standard deviation of calls on the caller identification, we could estimate whether the abuser is generating the caller identification randomly or not.

Phone String Features

This category of features is for only source phone numbers. We use the string of both reported source and victim phone numbers to compute the similarity. We define two type of similarity for phone numbers as presented in the following:

- *List similarity:* We compute this similarity by matching the source and victim phone numbers using the *Jaccard similarity index* as will presented in the implementation section, which produces a value between zero and one. The bigger is the value, the more the target and the source are similar. The value one means that the source and the target are identical.

- *Set similarity:* To compute this feature, we extract the unique digits from source and compare them with the unique digits of the target phone number. The similarity is the number of intersection digits divided by the number of union digits.

3.3.4 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.

3.3.5 Campaign Detection

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. 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.

Record Linkage

As described above, we proceed with the detection of potential telephony scamming campaigns. We began our analysis by obtaining all reported phone numbers along with their caller identifications and word vectors. We then create a list of all caller IDs used by each number and the aggregated word vector. This provides information about the types of abuse in which these numbers are employed. In the analysis of our dataset of complaints, we build a connection network of security events that are attributed to the same type of telephony abuse. Security events are visualized, in our representation, as a graph that refers to a multitude of telephony campaigns, such as vishing, telemarketing, political campaign, etc. Our ultimate goal is to obtain better insight about the instigators of telephony abuse activities. The graph reported in Figure 4.7 provides a graphical representation of the detected calling campaigns. The graph shows phone numbers that have been used to make independent calls that are not connected to other numbers. However, the graph depicts other numbers that are related to other source numbers as they are using a set of similar caller identification information, and have been reported abusing victims in the the same way. As such, these numbers are considered to be part of the same calling campaign.

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 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:

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

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

To extract the different calling campaigns, we use the Fast Unfolding Community Detection Algorithm [13]. We choose this algorithm [13]; 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 [33], 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 Jacquard 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

Giving a set of campaigns: $C = \{C_1, C_2, C_3, ..., C_N\}$, 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 $C_1 = \{E_1, E_2, E_3, ..., E_{N_1}\}$, $C_2 = \{E_1, E_2, E_3, ..., E_{N_2}\}$, until $C_n = \{E_1, E_2, E_3, ..., E_{N_m}\}$. We then Evaluate for each given (E,C) whether:

(1) We create a new campaign, or

(2) We consider that $E_i \in C_i$

The following algorithm 1 details the process:

Algorithm 1: Real-time Campaign Identification Algorithm

```
Data: Complaints
Result: Campaign attribution
List complaints;
Last_seen = 0;
Threshold τ;
while complaint do
    read complaint;
    if t_i - Last_seen < τ then
        if ( id_i ≈ id_j and m_i ≈ m_j) then
            E_i and E_j ∈ C_i;
        else
            New campaign C_n
            E_i ∈ C_n
        end
    else
        New campaign C_n
        E_i ∈ C_n
    end
end
```

3.4 Implementation

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

3.4.1 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. 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-proccess 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* [31], 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* [2], 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. Moreover, the feature extraction relies, for some features, on external data sources such as *Canadian*

Numbering Administrator [19] and *North American Numbering Plan Administration* [41] for North American phone numbers, or *International Assignments under Recommendation* ITU-T E.164 for international phone numbers. 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* [20], an open source key-value cache, for this purpose.

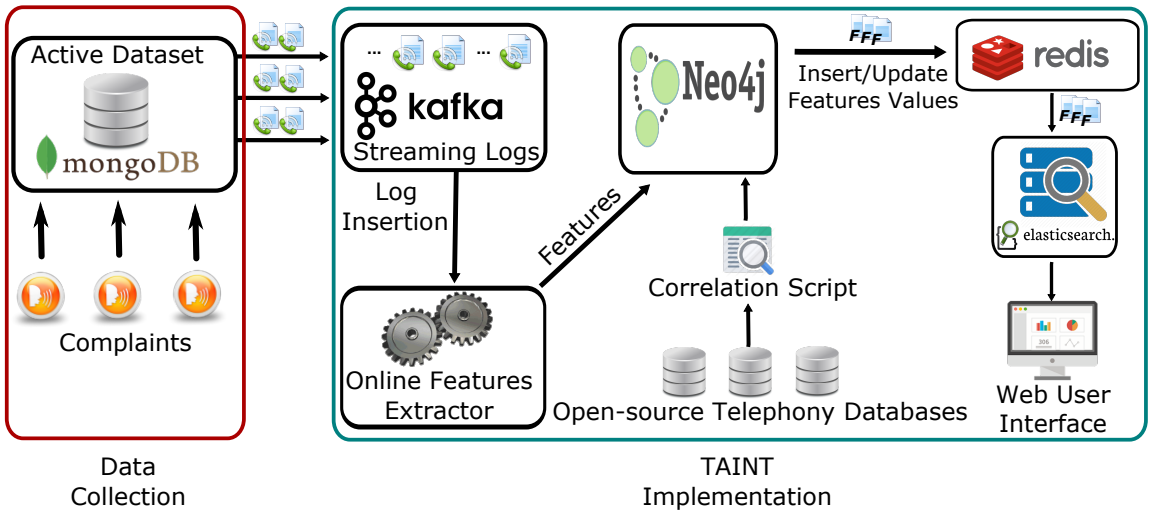

Figure 3.3: Components of TAINT Framework

Online Feature Computation

In this component, we mainly leverage the graph representation of the complaints records in the graph database – *Neo4j*. We choose to present the feature computation for **Temporal Features** and **Phone String Features**. For the rest of the features, the same techniques are used in their computation.

Temporal Features In order to extract the *temporal features*, we use *frequent item set* (part of Frequent Pattern Mining [26]). Discovery of frequent patterns is an important problem in the

area of data mining. This problem is introduced in [8] and can be formalized as follows. Let $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$ be a set of literals, called *items*. Let $\mathcal{D}$ be the transaction database where each transaction T is a non-empty itemset such that $t \subseteq \mathcal{I}$. Each transaction in the database is identified by *TID*. A set of items is called an *itemset*, and an itemset with k items is called a k-*itemset*. In our implementation, we compute the support of 1-*itemset*, where the itemset is the phone number or the caller identification. The *support* of an itemset x in $\mathcal{D}$, denoted as $\sigma(x/\mathcal{D})$, is the ratio of the number of transactions (in $\mathcal{D}$) containing x to the total number of transactions in $\mathcal{D}$.

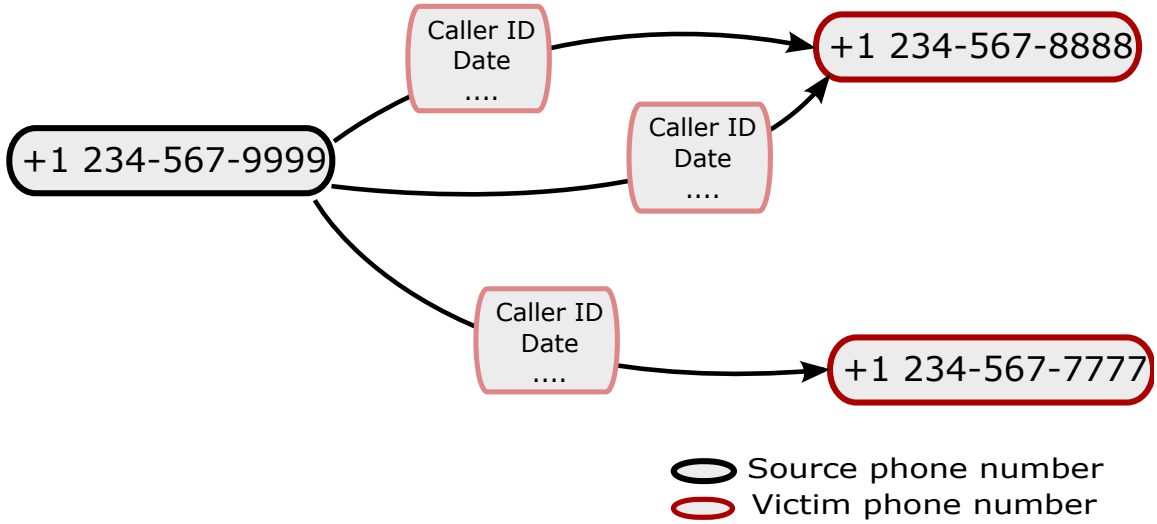

Figure 3.4: Phone Numbers Graph Structure Example

Instead of using the current implementation of *frequent item set* such as Apriori [8], FP-growth [27], and ECLAT [61], we exploit the graph structure of complaints in the database, as depicted in Fig. 3.4 for phone numbers and Fig. 3.5 for caller identifications.

In case of temporal features, for instance the daily occurrence, the itemset support is used for computing the daily occurrence by counting the edges with distinct dates. Thus, the phone number support (the daily occurrence) is the number of distinct dates divided by the total day in the life span of the collected dataset. In the example depicted in Fig. 3.5, we show how to compute the support from the graph structure. We use the same calculation method to compute the temporal feature for the phone number. Furthermore, it is also used in caller identification temporal feature extraction.

Finally, we use Scikit-learn [52] to build the machine learning classification and regression models.

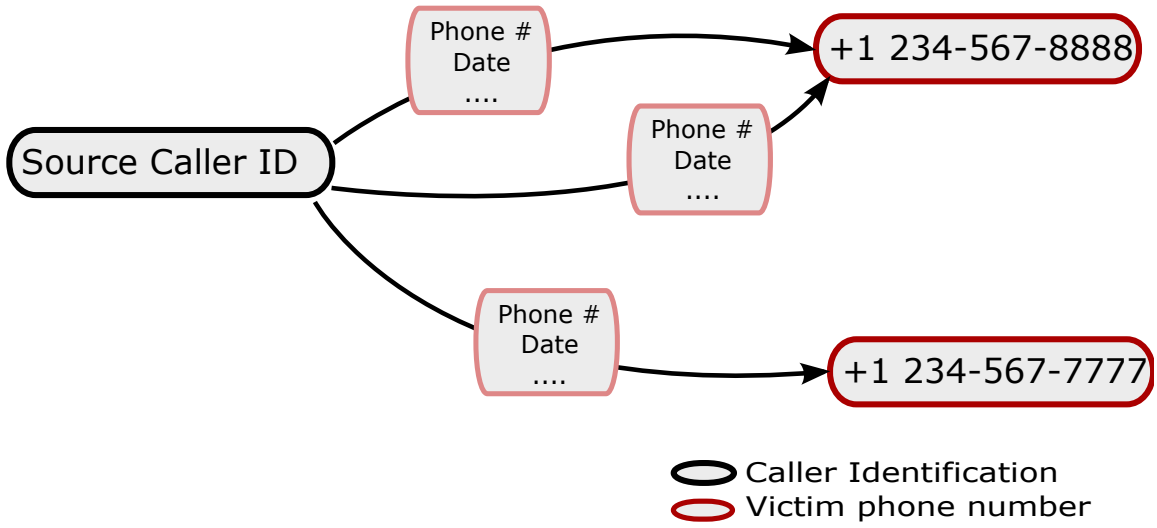

Figure 3.5: Caller Identification Graph Structure Example

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.

3.4.2 Front-end

The Front-end of TAINT Framework is a dashboard implemented using *Joomla* [18] and an interactive data analysis using *Kibana* [29], a sophisticated web front-end and part of *Elasticsearch* [54] 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. 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 Figure 3.6.

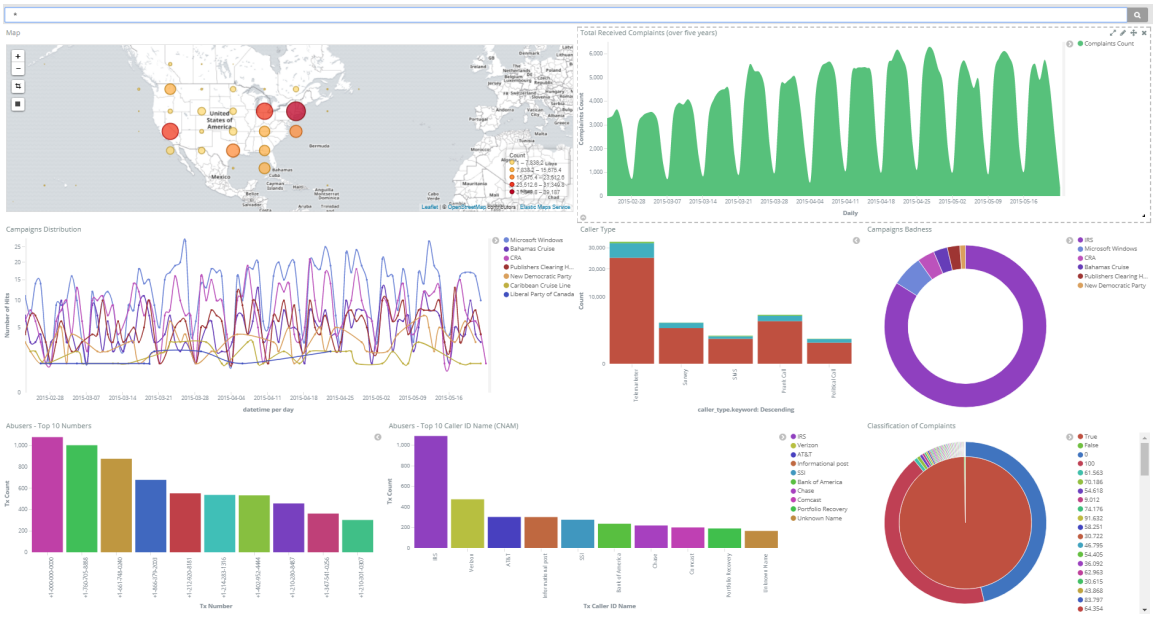

Figure 3.6: Screenshot of the Near real-time Monitoring Web Interface

Chapter 4

Results and Evaluation

In this section, we present the evaluation and provide some findings of our telephony abuse analysis framework. In this pursuit, we first present the results and evaluation of the classifiers and algorithms on our dataset. 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.

4.1 Evalution of the Algorithms

In this section, we discuss the results of the SVM classifier and the linear regression model for badness ranking. We applied many classification algorithms for this project and only the one that provided the best results was presented in the result subsection. Many experimental setups and a collection of evaluation criteria have been defined, such as, the False Positive Rate (FP_rate), the True Positive rate (TP_rate), and the precision. The precision and TP_rate measures are extensively used in classification. Precision is the percentage of tuples labeled as positive and that are a true positive (TP). It can be understood as a measure of correctness. On the other hand, TP_rate is the percentage of positive tuples and can be considered as a measure of completeness [26]. Whereas, the FP_rate is dedicated for the misclassified data. The equations of the mentioned criteria are provided in the following:

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

$$precision = \frac{TP}{TP + FP} \tag{3}$$

and

$$TP \text{ rate} = \frac{TP}{TP + FN} \tag{4}$$

and

$$FP \text{ rate} = \frac{FP}{FP + TN} \tag{5}$$

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 experiments we have used SVM. As provided in Table 4.1, we can see that the execution of SVM using our selected features gave us a very high accuracy with 98.85% and F_measure with 0.98.

Furthermore, the execution of Linerar regression for badness scoring gave also a very promising results with mean squared error of 6.4.

Run	*TP_rate*	*FP_rate*	*Precision*	*F_measure*	*Accuracy*
Support Vector Machine	0.989	0.002	0.989	0.988	98.851 %

Table 4.1: Results of the Classification Model on The Complaints Dataset

4.2 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 1^{st} of January, 2009 to 30^{th} of December 2016.

4.2.1 General statistics

(1) ***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 4.2.

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

Table 4.2: Complaints Details

4.2.2 Complaints Distribution

In this section, we present charts depicting the distribution of the source numbers. This section provides insights on the calling pattern of fraudsters. First, the distribution over time demonstrates the increasing activity of telephony abuse. Second, This will help understanding calling numbers' patterns used by the abusers as well as their evolution over time.

The chart in Figure 4.1 depict 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 30000 complaints per week in 2016. This shows that criminals are adopting more and more the telephony network infrastructure to reach and scams telephony customers.

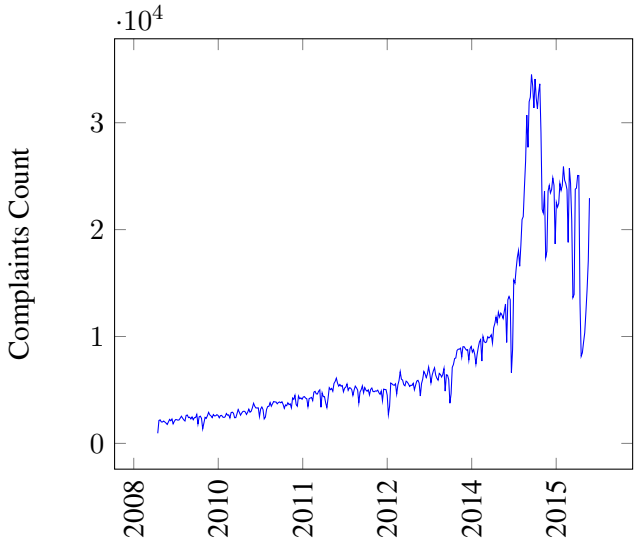

Figure 4.1: Distribution of Complaints over Time (year on the abscissa)

4.2.3 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* [19] and *North American Numbering Plan Administration* [41] 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*. It is evident that a large portion of calls come from North America, 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. In this section, we will further explore the data to illustrate the geo-location distribution of abusers. This helps in understanding the types of abuse and the distribution of telephony spammers.

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

Table 4.3: Calls Distribution Based on the Geographic Location

(1) ***Country-based Source Geo-location***

Seven percent of the received calls have been generated from a *international phone number*. There are calls, as shown in Fig. 4.3, from different countries. A big portion of the calls have been generated from the *United states*, which we expected due to the low communication cost rates - free VoIP phone calls sometimes - in the *United States* and *Canada*. There are small portions of the calls that have been generated from *Australia*, *Japan*, etc. which shows the international perspective of telephony complaints dataset. The geographic distribution of numbers also provides insights about the type of malicious calls; for instance, according to [43], long distance calls or what we call toll fraud are made to trick people to call back so that they will be charged more on their phone bills.

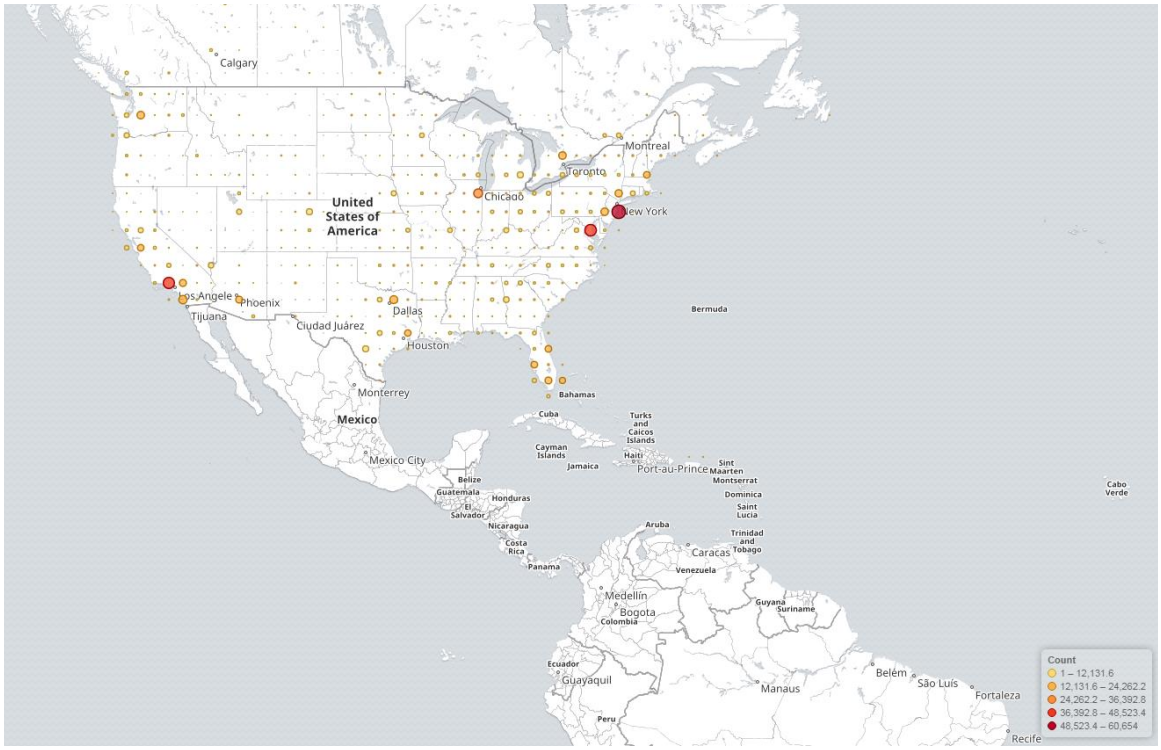

Figure 4.2: Geographic Distribution of Reported Source Phone Numbers for June 2017

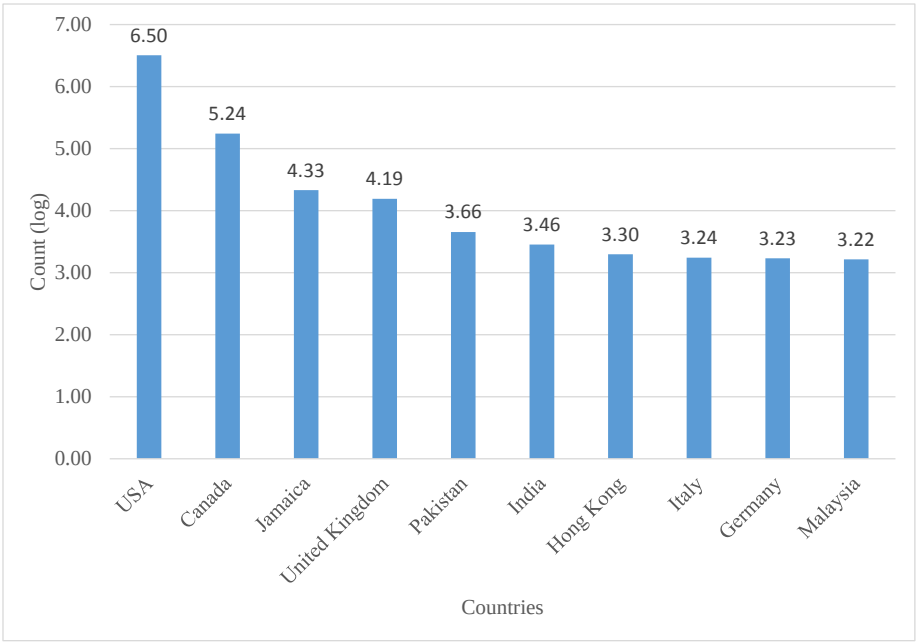

Figure 4.3: Top Source Countries

(2) ***City-based Source Geo-location***

Phone numbers in North America (Canada and United States) are composed of ten digits. The first three digits are called the Numbering Plan Area (NPA), which indicate the city of the phone number, for instance 514-xxx-xxxx is a phone number from Montreal city. In addition, other foreign conventions, such as the Chinese phone number convention, have been used to extract the city from the phone numbers. The four main cities from outside Canada are from the United States. These cities are, as shown in Fig. 4.4, *Florida City*, *New York City*, *West Liberty* and *California City*.

Figure 4.4 depicts the top cities from where the source phone calls originate. These cities are from all over the world. We notice that a large number of calls originate from big cities in North America. This is related to cities with a large population and big industrial corporations. In many cases, fraudsters spoof big companies or entities that are located in these cities.

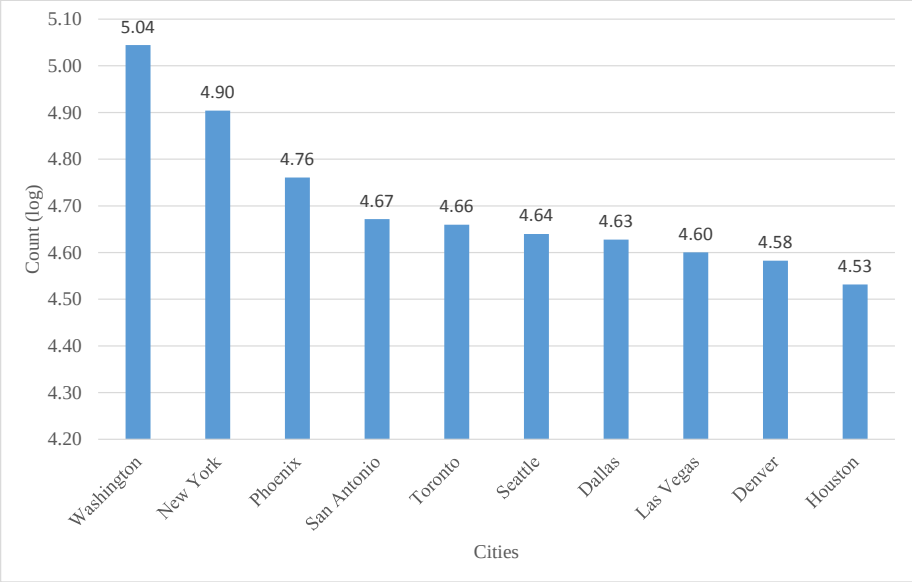

Figure 4.4: Top Source Cities

4.2.4 Phone Number Analysis

In Figure 4.5, we depict the top 10 reported phone numbers. Some of these numbers are invalid and are easy to use as spoofed numbers. Others are not assigned to any person or organization. The abusers prefer to use the later instead of assigned numbers, for which the owners are known. The large number of complaints about these sources show that these numbers either belong to criminal groups or are spoofed by such groups.

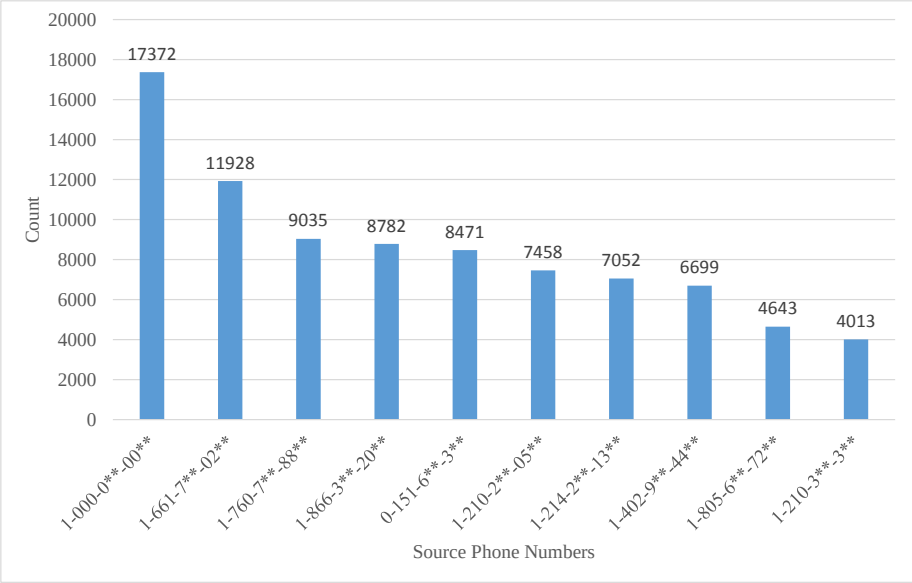

Figure 4.5: The Most Reported Phone Numbers

4.2.5 CNAM Analysis

According to our analysis, we found that telephony abusers rely mainly on the CNAM to craft their attacks. The fact that this feature can be easily set up by telephony users provides abusers with a huge advantage since, with little or no effort, they can further a variety of scams. Also due to the lack of awareness of telephony customers with respect to the reliability of this feature, many attacks have been successful.

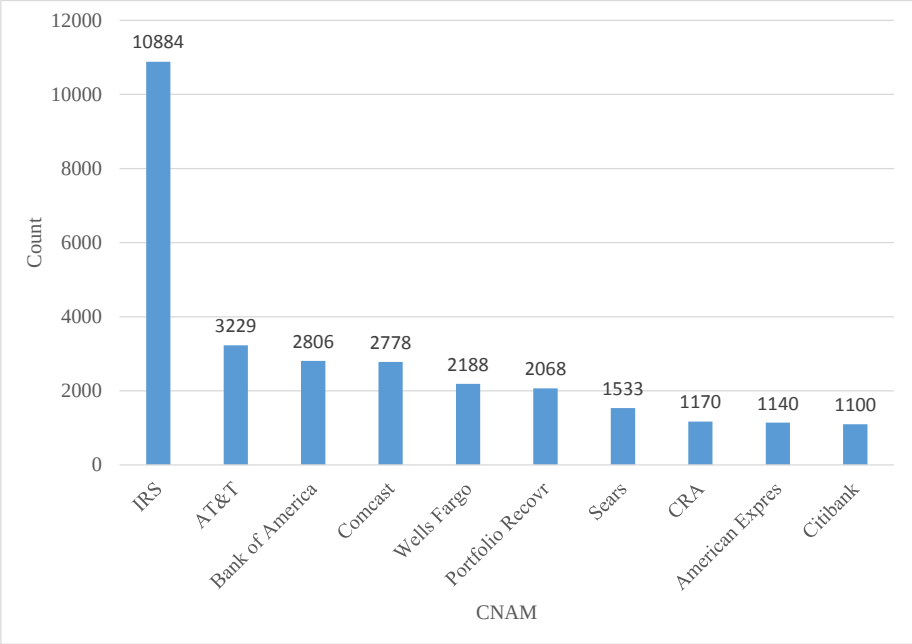

Figure 4.6: The Most Reported CNAMs

4.3 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. Figure 4.7 shows a network graph representing calling campaigns. Within this graph, we have observed that some calling numbers are shared amoung campaigns and others are specific to some campaigns. In Table 4.4, 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. We show in Figure 4.8, a word cloud, which illustrates the topics of the complaints campaigns. 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

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	1,961	2011-05-04	2016-01-30
	Air Canada	327	2013-10-10	2016-01-14
	Federal Express (FEDEX)	3,496	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	1,943	2014-12-15	2016-04-15
	Credit Card Services	5,638	2010-01-05	2016-04-20
	Reward Redemption	2,368	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?	5,638	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	4,222	2010-01-02	2016-04-21

Table 4.4: Subset of the Detected Calling Campaigns

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

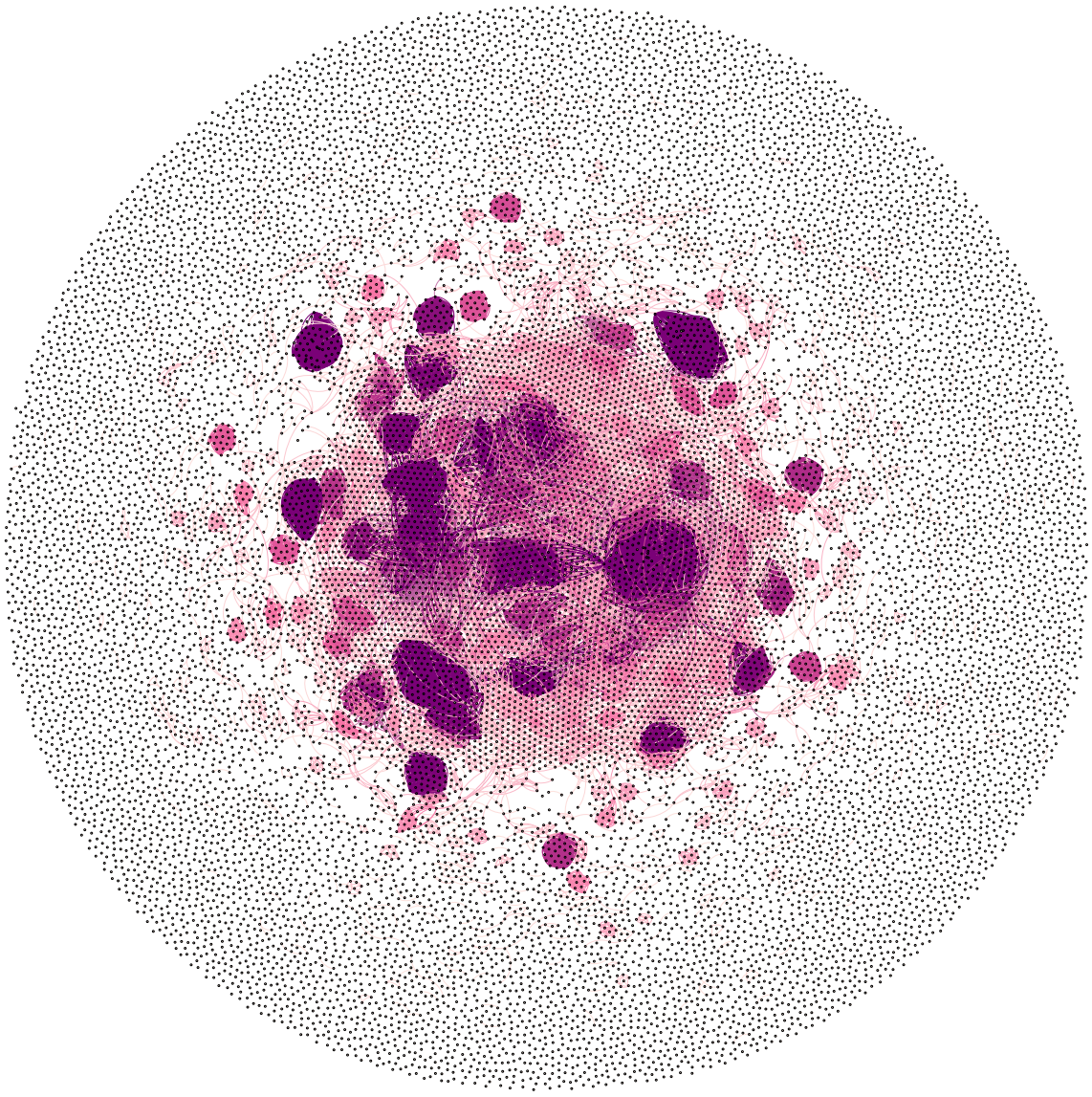

Figure 4.7: Graph of the Detected Campaigns in 2016

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

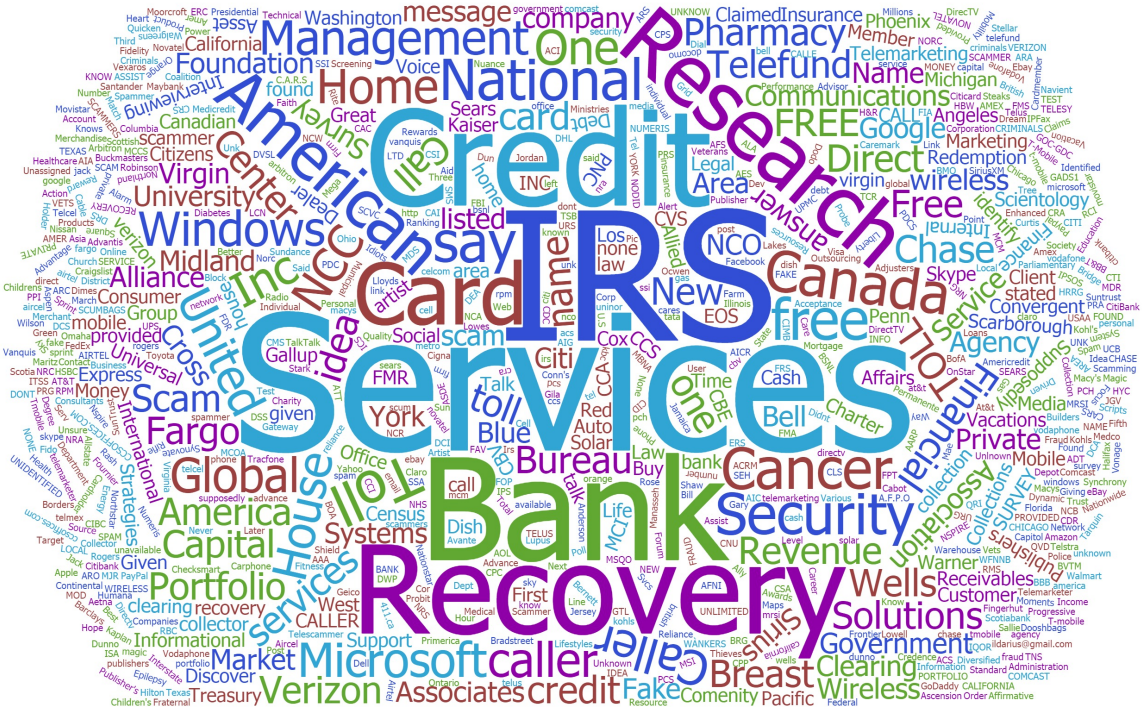

Figure 4.8: Campaigns Topics Word Cloud

spoofing their caller identification to be one of the entities mentioned previously. In the Tables 4.5, we give the top used phone numbers that were calling along with their badness score as well as the top CNAM used in this campaign. This is used to detect the patterns of this campaign. We found that the first 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 time. 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, 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

Phone Numbers	Count	Badness Score
+1-213-**1-**63	679	92.02
+1-877-**8-**13	349	87.23
+1-206-**5-**33	265	83.45
+1-800-**0-**84	237	67.26
+1-716-**6-**18	213	82.75
+1-206-**8-**49	213	94.21
+1-202-**4-**62	198	92.16
+1-202-**4-**62	188	76.73
+1-914-**5-**62	174	69.28
+1-800-**9-**40	122	68.19

Spoofed CNAM	Count
IRS	10884
Internal Revenu	199
US Treasury	51
Dpt Legal Affair	17
IRS officer	14
IRS Investg dpt	9
IRS Investigate	9
IRS Audit Dpt	9
Teasury Dpt	8
IRS Washington	8

Table 4.5: Top 10 Phone Numbers and CNAM used in the IRS Calling Campaign

Phone Numbers	Count	Badness Score
1-800-9**-10**	571	82.32
1-613-6**-24**	133	78.04
1-800-9**-22**	123	79.16
1-866-8**-58**	108	64.32
1-877-3**-03**	64	45.67
1-844-3**-79**	61	56.59
1-877-3**-16**	56	49.08
1-613-9**-96**	51	53.03
1-604-6**-49**	21	13.12
1-613-7**-62**	20	15.88

Spoofed CNAM	Count
CRA	1170
Canada Reven Ag	110
Revenue Canada	24
Canadian Rvn Ag	8
Canada Revenue	8
Gvt of Canada	3
Canada Reven S	3
CRA Vancouver	3
CRA Canada	3
Offices of CRA	2

Table 4.6: Top 10 Phone Numbers and CNAM used in the CRA Calling campaign

voicemail for the investigation from a set of numbers provided in Table 4.6, 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. The criminals used phone numbers and spoofed Caller Identification Name (CNAM) presented in Table 4.6 to craft their attacks.

Phone Numbers	Count	Badness Score
1-661-7**-02**	101	85.86
1-206-4**-06**	92	87.48
1-855-2**-33**	64	82.72
1-210-2**-05**	53	76.44
1-210-3**-03**	52	73.32
1-212-7**-34**	47	72.69
1-000-0**-00**	41	76.70
1-516-4**-68**	34	49.39
1-646-6**-34**	26	43.12
1-416-3**-11**	17	20.23

Spoofed CNAM	Count
Microsoft	200
Microsoft Wind	153
Windows	99
Win Tech Supp	24
Win Tech Dpt	18
Windows Supp	18
Mcft Win Supp	16
Microsoft Sec	12
WindowsMicrosft	10
Win Ser Center	10

Table 4.7: Top 10 Phone Numbers and CNAM used in the Microsoft Scamming Campaign

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.

Microsoft Windows Scam. 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. The phone numbers from where the calls were coming and the Caller Identifications that were used in these calls are presented in Tables 4.7.

Phone Numbers	Count	Badness Score
1-888-8**-88**	4	19.01
1-888-6**-11**	4	18.85
1-888-2**-56**	4	15.06
1-866-4**-47**	4	14.23
1-800-9**-49**	4	16.45
1-727-5**-73**	4	25.29
1-920-6**-08**	3	12.87
1-888-9**-73**	3	15.69
1-888-7**-62**	3	21.31
1-866-3**-20**	3	23.28

Spoofed CNAM	Count
Credit Card Ser	24
US Bank	6
VISA CARD SER	4
FIA Card Ser	4
OA Credit Card	3
Barclays card S	3
Card Member Ser	3
Discover Card Ser	2
redit Card Ser	2
Card Monitoring	2

Table 4.8: Top 10 Phone Numbers and CNAM used in the Credit Card Services Scamming Campaign

Credit Card Service Scams. We identified ten credit card services fraud campaigns resulting in 5,638 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. The criminals used phone numbers and spoofed caller identification presented in Tables 4.8 to craft their attack.

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

Phone Numbers	Count	Badness Score
1-216-6**-55**	25	42.39
1-800-2**-82**	23	28.63
1-800-6**-05**	6	29.12
1-800-5**-01**	6	19.19
1-800-2**-12**	6	16.81
1-515-8**-46**	5	8.23
1-502-6**-90**	5	9.19
1-973-4**-79**	4	8.57
1-954-2**-80**	4	13.05
1-913-2**-02**	4	12.61

Spoofed CNAM	Count
Bahamas Cruise	20
Florida Bahamas	6
VCS Bahamas	4
Royal Bahamas	3
Cruise line	3
Free Bahamas	3
Caribbean	2
Travelwise	2
Sunshine Resort	2
Smart Travel	2

Table 4.9: Top 10 Phone Numbers and CNAM used in the Bahamas Cruise Trip Scam Campaign

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, depicted in tables 4.9, 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.

Chapter 5

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 thesis, we presented TAINT, an automatic framework for the near-real-time collection 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 geo-location distribution as well as the underlying campaigns by exploring the similarities among the abuse incidents. The performance of TAINT components was satisfying as the classification model reached 98%, and the regression model had an acceptable mean squared error of 6.4. In addition, TAINT detected $1,519$ 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 tool generates valuable intelligence that can be used to reduce this type of abuse.

We have observed that TAINT has some limitations that could be addressed in future work. The following presents theses limitations and suggests their mitigation.

- We implemented in TAINT a component which extracts temporal features: the daily occurrence, the days in a week occurrence, and the hourly occurrence from the phone numbers and caller identification. However, these features can be leveraged for the detection of the different types of telephony threats, such as TDoS, telemarketing, and debt collectors and SPIT. Such technique would be extremely advantageous for law enforcement investigators and telephony service operators to deal with the most serious threats more effectively.

- In our work, TAINT relied on the Caller Identification (CNAM) feature, and the text of the complaint filled by victims to detect the different calling campaign. However, by further manual analysis and experimentation, we found that some telephony abusers are not part of the same campaign although they use the same schema and techniques to scam their victims, such as, the IRS scam. Consequently, if we would add another data source with audio content to TAINT, then it would detect the different campaigns more efficiently.

- We divided the detected campaigns into four categories: fraud, political, scam, and telemarketing calling campaigns. Nevertheless, A more elaborated telephony threat campaigns taxonomy would help in providing more refined intelligence.

Bibliography

[1] Irs nebraskans lost 56000 to telephone scam. Available at: http://nebraskaradionetwork.com/2016/02/17/irs-nebraskans-lost-56000-to-telephone-scam/. Accessed on: 13 September 2016.

[2] Mongodb. Available at: https://www.mongodb.org/. Accessed on: 21 January 2015.

[3] Complain About a Telemarketing Call. Available at: https://www.lnnte-dncl.gc.ca/plt-cmp-eng. Accessed on: 08 July 2015.

[4] What Is a Toll-Free Number and How Does it Work? Available at: https://www.fcc.gov/consumers/guides/what-toll-free-number-and-how-does-it-work. Accessed on: 12 July 2017.

[5] Collection Agencies. Available at: https://www.ic.gc.ca/eic/site/oca-bc.nsf/eng/h_ca02149.html. Accessed on: 19 February 2015.

[6] Collection agency fined 500K over automated phone calls. Available at: http://www.cbc.ca/news/canada/collection-agency-fined-500k-over-automated-phone-calls-1.2251782. Accessed on: 11 March 2015.

[7] Collection agency harassed debt-free Canadians. Available at: http://www.cbc.ca/news/canada/collection-agency-harassed-debt-free-canadians-1.1153123. Accessed on: 05 Juin 2015.

[8] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[9] Mustaque Ahamad, Dave Amster, Michael Barrett, Tom Cross, George Heron, Don Jackson, Jeff King, Wenke Lee, Ryan Naraine, Gunter Ollmann, et al. Emerging cyber threats report for 2009. 2008.

[10] VFTC Complaint Assistant. Available at: https://www.ftccomplaintassistant.gov. Accessed on: 20 Juin 2016.

[11] Marco Balduzzi, Payas Gupta, Lion Gu, Debin Gao, and Mustaque Ahamad. Mobile telephony threats in asia.

[12] Marco Balduzzi, Payas Gupta, Lion Gu, Debin Gao, and Mustaque Ahamad. Mobipot: Understanding mobile telephony threats with honeycards. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, pages 723–734. ACM, 2016.

[13] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[14] FTC Issues FY 2016 National Do Not Call Registry Data 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.

[15] Pisit Chanvarasuth. Knowledge on phishing and vishing: An empirical study on thai students.

[16] Junaid Chaudhry and Shafique Ahmed Chaudhry. Secure calls and caller id spoofing countermeasures. 2016.

[17] Yun Chi, Haixun Wang, Philip S. Yu, and Richard R Muntz. Catch the moment: Maintaining closed frequent itemsets over a data stream sliding window. *Knowl. Inf. Syst.*, 10(3):265–294, October 2006.

[18] Joomla CMS. Available at: https://www.joomla.ca/. Accessed on: 14 Juin 2015.

[19] Canadian Numbering Administration Consortium. Available at: http://www.cnac.ca/ . Accessed on: 11 January 2015.

[20] Redis data structure store. http://redis.io. Accessed on: 16 February 2015.

[21] V Karamchand Gandhi. An overview study on cyber crimes in internet. *Journal of Information Engineering and Applications*, 2(1):1–5, 2012.

[22] Shuangping Gong, Yonghui Dai, Jun Ji, Jinzhao Wang, and Hai Sun. Emotion analysis of telephone complaints from customer based on affective computing. *Computational intelligence and neuroscience*, 2015:15, 2015.

[23] Slade E Griffin and Casey C Rackley. Vishing. In *Proceedings of the 5th annual conference on Information security curriculum development*, pages 33–35. ACM, 2008.

[24] Payas Gupta, Bharath Srinivasan, Vijay Balasubramaniyan, and Mustaque Ahamad. 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*, 2015.

[25] Srishti Gupta, Payas Gupta, Mustaque Ahamad, and Ponnurangam Kumaraguru. Exploiting phone numbers and cross-application features in targeted mobile attacks. In *Proceedings of the 6th Workshop on Security and Privacy in Smartphones and Mobile Devices*, pages 73–82. ACM, 2016.

[26] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[27] Jiawei Han and Jian Pei. Mining frequent patterns by pattern-growth. *ACM SIGKDD Explor. Newsl.*, 2(2):14–20, December 2000.

[28] Heres 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 on: 23 August 2017.

[29] Your Window into the Elastic Stack. Available at: https://www.elastic.co/products/kibana. Accessed on: 16 January 2015.

[30] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142, 1998.

[31] Apache Kafka. Available at: http://kafka.apache.org/. Accessed on: 19 February 2015.

[32] Anna Kang, Jae Dong Lee, Won Min Kang, Leonard Barolli, and Jong Hyuk Park. Security considerations for smart phone smishing attacks. In *Advances in Computer Science and its Applications*, pages 467–473. Springer, 2014.

[33] ElMouatez Billah Karbab, Mourad Debbabi, Abdelouahid Derhab, and Djedjiga Mouheb. Cypider: building community-based cyber-defense infrastructure for android malware detection. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pages 348–362. ACM, 2016.

[34] Anindita Khade and Subhash K Shinde. Detection of phishing websites using data mining techniques. In *International Journal of Engineering Research and Technology*, volume 2. ESRSA Publications, 2014.

[35] Reverse Phone Number Lookup. Available at: http://whocallme.com/ . Accessed on: 05 August 2015.

[36] Federico Maggi. Are the con artists back? a preliminary analysis of modern phone frauds. In *CIT*, pages 824–831. IEEE Computer Society, 2010.

[37] Federico Maggi, Alessandro Sisto, and Stefano Zanero. A social-engineering-centric data collection initiative to study phishing. In *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, pages 107–108. ACM, 2011.

[38] Aude Marzuoli, Hassan A Kingravi, David Dewey, Aaron Dallas, Telvis Calhoun, Terry Nelms, and Robert Pienta. Call me: Gathering threat intelligence on telephony scams to detect fraud.

[39] Najmeh Miramirkhani, Oleksii Starov, and Nick Nikiforakis. Dial one for scam: A large-scale analysis of technical support scams. In *Security and Privacy (EuroS&P), 2017 IEEE European Symposium on*, pages 235–250. IEEE, 2017.

[40] Hossen Mustafa, Wenyuan Xu, Ahmad Reza Sadeghi, and Steffen Schulz. 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*, pages 168–179. IEEE, 2014.

[41] North American Numbering Plan Administration: NANPA. Available at: http://www.nanpa.com/. Accessed on: 17 January 2015.

[42] The 20 newsgroups text dataset. Available at: http://scikit-learn.org/stable/datasets/twenty_newsgroups.html. Accessed on: 13 August 2017.

[43] Directory of Unknown Callers. Available at: http://800notes.com/. Accessed on: 15 October 2015.

[44] Debt Collector or Scammer: How to Tell the Difference? Available at: http://www.nolo.com/legal-encyclopedia/debt-collector-scammer-how-tell-the-difference.html. Accessed on: 05 March 2015.

[45] Vishing or Voice Phishing. Available at: http://www.rcmp-grc.gc.ca/scams-fraudes/vish-hame-eng.htm. Accessed on: 12 April 2015.

[46] Manjusha Pandey and Vignesh Ravi. Detecting phishing e-mails using text and data mining. In *Computational Intelligence & Computing Research (ICCIC), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

[47] Remain on IRS "Dirty Dozen" List of Tax Scams for the 2016 Filing Season Phone Scams Continue to be a Serious Threat. Available at: https://www.irs.gov/uac/newsroom/. Accessed on: 13 February 2016.

[48] Canadian Radio-television and Telecommunications Commission Unsolicited Telecommunications Rules. Available at: http://www.crtc.gc.ca/eng/trules-reglest.htm. Accessed on: 07 July 2017.

[49] Consumer Sentinel Network Reports. Available at: https://www.ftc.gov/enforcement/consumer-sentinel-network/reports . Accessed on: 19 August 2017.

[50] Merve Sahin, Aurélien Francillon, Payas Gupta, and Mustaque Ahamad. Sok: Fraud in telephony networks. In *Security and Privacy (EuroS&P), 2017 IEEE European Symposium on*, pages 235–250. IEEE, 2017.

[51] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[52] scikit-learn Machine Learning in Python. Available at: http://scikit-learn.org/stable/. Accessed on: 21 August 2016.

[53] Jaeseung Song, Hyoungshick Kim, and Athanasios Gkelias. ivisher: Real-time detection of caller id spoofing. *ETRI Journal*, 36(5):865–875, 2014.

[54] Elastic Stack. Available at: https://www.elastic.co/. Accessed on: 12 January 2015.

[55] Fight Back Against Annoying Telemarketers. Available at: http://www.callercomplaints.com/. Accessed on: 13 August 2015.

[56] Understand telemarketing rules for compliance. Available at: http://www.crtc.gc.ca/eng/info_sht/t1032.htm. Accessed on: 09 January 2015.

[57] Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn. SoK: Everyone Hates Robocalls: A Survey of Techniques against Telephone Spam. In *Proceedings of the IEEE Symposium on Security and Privacy*, May 2016.

[58] Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn. Toward authenticated caller id transmission: The need for a standardized authentication scheme in q. 731.3 calling line identification presentation. In *ITU Kaleidoscope: ICTs for a Sustainable World (ITU WT), 2016*, pages 1–8. IEEE, 2016.

[59] Jangam Upendar and Etikala Gurumohan Rao. An overview of plastic card frauds and solutions for avoiding fraudster transactions. *International Journal of Research in Engineering and Technology*, 2013.

[60] 000 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.

[61] M.J. Zaki. Scalable algorithms for association mining. *Knowledge and Data Engineering, IEEE Transactions on*, 12(3):372–390, May 2000.