Digital Investigation 24 (2018) S131–S141

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2018 Europe — Proceedings of the Fifth Annual DFRWS Europe

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

Houssem Eddine Bordjiba*, ElMouatez Billah Karbab, Mourad Debbabi

Concordia University, Canada

Keywords:
Telephony abuse
Telephony complaints
Vishing
Spoofing
Telephony abuse campaigns

A B S T R A C T

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

Introduction

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

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

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

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

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

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

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

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

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

Our contributions are threefold:

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

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

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

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

Dataset

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

Framework architecture

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