2013-ENST-0075

EDITE - ED 130

Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Sécurité de Réseaux »

présentée et soutenue publiquement par

Jelena ISACENKOVA

le 5 Décembre 2013

Analyse de Campagne Massives de Courrier Électronique collectées grâce à un Filtre Anti-spam basé sur le Principe de Défi-réponse

Directeur de thèse : **Refik MOLVA**
Co-encadrement de la thèse : **Davide BALZAROTTI**

Jury

M. Hervé DEBAR, Professeur, RST, Télécom SudParis	Rapporteur
Mme Maryline LAURENT, Professeur, R3S, Télécom SudParis	Rapporteur
M. Lorenzo CAVALLARO, Ass. de Professeur, ISG, Royal Holloway University	Examinateur
M. Magnus ALMGREN, Ass. de Professeur, Chalmers University	Examinateur

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

THÈSE

Analysis of Bulk Email Campaigns using a Challenge-Response Anti-Spam System

Jelena Isacenkova

A doctoral dissertation submitted to:

TELECOM ParisTech

in partial fulfillment of the requirements for the degree of:

DOCTOR (Ph.D)

Specialty : Network Security

Jury :

Reviewers:		
Prof. Hervé DEBAR	-	Télécom SudParis, Paris
Prof. Maryline LAURENT	-	Télécom SudParis, Paris
Examiners:		
Ass. Prof. Lorenzo CAVALLARO	-	Royal Holloway University, United Kingdom
Ass. Prof. Magnus ALMGREN	-	Chalmers University, Gothensburg, Sweden
Supervisors:		
Prof. Refik MOLVA	-	EURECOM, France
Ass.Prof. Davide BALZAROTTI	-	EURECOM, France

Abstract

The spam phenomenon exists already for over one and a half centuries, with the first copies of it being sent over the telegraph lines. Although the first *electronic* spam appeared in the 1980s, in 2002 email spam accounted to only 9% of the email traffic. But since that year, spam business took off and moved to the next level: the release of spamming tools, malware attachments, and massive abuses of open-relays led to a rapid increase of spam that reached 72% of the email traffic in 2004. The numbers continued to increase in the next years, reaching its pick in 2010 – 89% of emails were spam.

On the other side, today user mailbox also started receiving large amounts of other types of bulk emails. According to the Email Intelligence Report by ReturnPath published in 2012, newsletters and automated notifications messages summed up to 42% of inbox messages. At the same time, 16% of emails containing advertisements or marketing information were flagged as spam. At a first glance, many people would consider this “side-effect” as an advantage. But while a conventional definition of spam encompasses *unsolicited* messages sent in *bulk* – large number of copies, it is not that obvious to make a difference between unsolicited emails.

Therefore, while most of the existing research studies the efficiency of anti-spam techniques and their enhancements, this thesis focuses on the thin line that separates the two categories: those few cases in which the existing techniques fail. In particular, we limit our study to the often overlooked area of *gray emails*, i.e., those ambiguous messages that cannot be clearly categorized one way or the other by automated spam filters.

We approach the study of gray area as bulk emails, by focusing on the analysis of email campaigns. We propose a three-phase approach based on message clustering, classification, and graph-based refinement that is based only on email headers data. Our technique is able to automatically classify 50% of the gray emails with only 0.2% of false positives. Moreover, we demonstrate that by using a graph-based refinement method, legitimate email campaigns can be often identified based only on sender information.

During the study of gray area, we identified three email campaign categories – commercial, newsletters, and botnet – for which our classification method works well. To identify *419 scam* campaigns, an advanced fee fraud primarily based on confidence, we propose instead a technique based on the phone numbers. The reasoning behind the idea is that scam as a business requires a communication channel for getting in contact with the victim, thus phone numbers play a particular role in 419 scam. We next rely on this insight to identify and characterize 419 scam campaigns by describing several illustrative examples that demonstrate the diversity of such campaigns and their international geographic distribution.

As part of this thesis was conducted inside a commercial anti-spam company, MailInBlack, specialized in email management based on a Challenge-Response (CR)

anti-spam filtering system, we conducted most of our experiments with the datasets available within the company. In fact, this kind of anti-spam filtering system provides a specific vantage point that is especially convenient for studying the gray area phenomenon. The quarantined area of the CR system is a good approximation of a gray area as it already excludes most of the personal user messages and spam emails.

Contents

List of Publications

Journals

- Isacenkova, Jelena; Thonnard, Olivier; Costin, Andrei; Balzarotti, Davide; Francillon, Aurelien "Inside the SCAM Jungle: a Closer Look at 419 Scam Email Operations", Eurasip Journal on Information Security, 2013. *Under submission*

Conferences and Workshops

- Isacenkova, Jelena; Balzarotti, Davide "Measurement and evaluation of a real world deployment of a challenge-response spam filter", IMC 2011, ACM SIGCOMM Internet Measurement Conference, November 2-4, 2011, Berlin, Germany.

- Isacenkova, Jelena; Thonnard, Olivier; Costin, Andrei; Balzarotti, Davide; Francillon, Aurelien "Inside the SCAM jungle: A closer look at 419 scam email operations", IWCC 2013, International Workshop on Cyber Crime (co-located with the 34th IEEE Symposium on Security and Privacy (IEEE S&P 2013), May 24, 2013, San Francisco, CA, USA.

- Costin, Andrei; Isacenkova, Jelena; Balduzzi, Marco; Francillon, Aurelien; Balzarotti, Davide "The role of phone numbers in understanding cyber-crime", PST 2013, 11th International Conference on Privacy, Security and Trust, July 10-12, 2013, Tarragona, Catalonia, Spain.

- Isacenkova, Jelena; Balzarotti, Davide "Shades of Grey: A Closer Look at Emails in the Gray Area", AsiaCCS 2014, 9th ACM Symposium on Information, Computer and Communications Security, June 4-6, 2014, Kyoto, Japan. *Under submission*

1
Introduction

1.1 Motivation and Objectives

Spam, according to its definition, refers to unsolicited messages that are sent in bulk [7]. Although the term was born in the 1980s (from a Monthy Python sketch where it was used by a crown of Vikings), based on this definition, the first spam message was delivered in 1864 when an unsolicited message was sent over the telegraph line promoting special investment offers to a targeted audience of wealthy Americans [78]. The first *electronic* spam message was dispatched within a military computer network (ARPANET) by Gary Turk who advertised new computers to 400 people. The critical tipping point in spam history occurred in 1994, changing the commercial advertisement business forever. During a commercial spam scandal of L. Canter and M. Siegel, the practice of sending unsolicited emails was defended by lawyers, who called their critics “anti-free speech zealots” [78].

Before this event, email spam was more of an annoyance – receiving pranks, chain letters, offensive messages [57] – while after it rapidly evolved into massive commercial mailing businesses run from corporate mail servers. This also started the never ending battle to protect the user mailboxes. Spammers took another step forward in 1997, when the lines of rather innocent spammers joined more deviant senders. At this time spamming was still largely a “work from home” occupation [75], when all in a sudden, some people started to abuse dynamic dial-up internet protocol addresses that were reassigned with new ones after a reconnection. Subsequently, as a defense, receiving mail servers started blocking, connections from dial-up IPs. This served as a ground for creating Real-time Blackhole Lists (RBL) of IP addresses used to block incoming traffic from spammers and misbehaving sources.

In 1999, Hall [88] noticed that spam emails are mostly near-duplicate messages that could be recognized by using message fingerprints that can be shared with other users [135]. This worked well until around 2002, when the spam business moved to the next level: the release of “Ratware” spamming tools (e.g. DarkMailer,

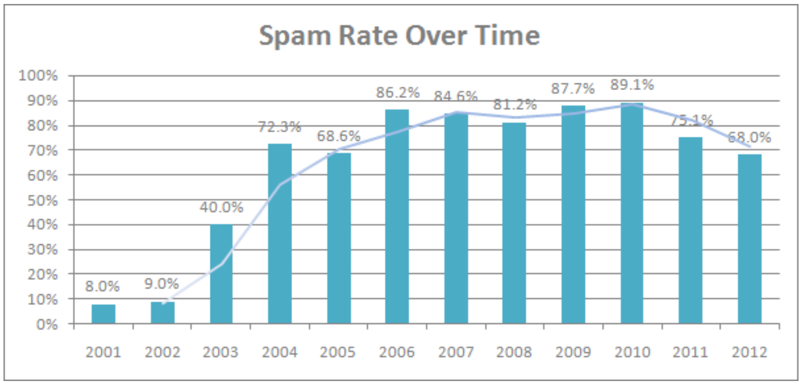


Figure 1.1: Spam trend over time by EmailTray [68]

SenderSafe) rapidly increased the number of spammers, and, most importantly, let spammers generate randomized content [150]. At the same time spammers started using open-relay for proxying their emails, taking advantage of different software (including Sendmail v.5) that was configured by default as an open-relay. One of the consequences of the appearance of such tools was the release of Sobig.a virus in January 2003 that was designed to send spam to a list of automatically downloaded email addresses.

2004 was the year when the first botnets were born, like Bagle and Bobax [75]. Botnet architecture was built on a distributed computing model relying on the network of infected personal computers that were initially used to send spam (today they are used to also perform other malicious actions). In 2007 a distributed spamming tool called Reactor Mailer was released, quickly followed by the appearance of several well-known spamming botnets as Storm, Cutwail, and Srizbi, responsible for sending billions of spam messages per day.

Today, despite the considerable effort and the large amount of proposed solutions to detect and filter unsolicited emails, spam still accounts for 66% of the emails on the Internet, according to the Intelligence Security Report [154] published by Symantec in 2013. Figure 1.1 reports the email spam rates over the last 11 years, showing how spam rapidly increased from 11% of total emails to 89.1% of them in 2010 – reaching its highest rates in the history of spam. Despite the recent take-downs of several large botnets, spam still costs as much as $20.5 billion annually in decreased productivity as well as in technical expenses [23].

Nowadays, many anti-spam filters provide a surprisingly high protection against large-scale unsolicited email campaigns. However, as spammers have improved their techniques to increase the chances of reaching their targets, anti-spam solutions have also become more aggressive in flagging suspicious emails. In 2011, despite the deployed filters, 19% of email messages delivered to a corporate email user's inbox were spam [95]. By 2012, the number had dropped to 15% [96].

On one side, this arms race has lead to a steady improvement in the detection rate. On the other, the number of false positives has also increased, with serious consequences for the users whenever an important message is erroneously flagged as spam. An Email Intelligence Report [141] published by Return Path pointed out that 16% of emails in 2012 containing advertisements or marketing information were flagged as spam and, therefore, never reached user mailboxes. At a first glance, many people would consider this "side-effect" as an advantage. However, it has been estimated that only one-third of users consider such messages as spam, while two-thirds prefer to receive unsolicited commercial emails from already known senders [59]. A more recent report shows that despite the overloaded mailboxes, consumers still read 18% of subscribed marketing emails, and continue to sign up for email offers and mailing lists [141], with the result of newsletters and automated notifications summing up to 42% of inbox messages (however, it is impossible to estimate how many of the messages were solicited). For these reasons, it is a well known fact that most people regularly check their spam folder to verify that no important messages have been misclassified by the anti-spam filters.

Unfortunately, this process is very time-consuming. Antispam solutions are not helping in this direction: they provide almost no additional information to help users in quickly identifying marketing emails, newsletters, or "borderline" cases that may be interesting for the users. In contrary, they put together harmless marketing and newsletter emails next to suspicious content emails, like phishing, scam, and other tricks used by miscreants.

Naturally, when users skim through their spam messages looking for something that looks legitimate, they need *to take a decision* about which email can be trusted, which one is just annoying, and which one can pose a real security threat. Unfortunately, several studies showed that most users are very bad in taking this kind of security-related decisions [127], making it one of the reasons why we need automated spam filters in the first place. For example, a user survey conducted in 2010 by the Messaging Anti-Abuse Working Group [115] reported that 57% of people who have accessed spam messages admitted to *have done so intentionally*, because they were unsure whether the suspicious message was spam or not. According to our data, after the obvious spam and legitimate emails have been eliminated, users still manually check on average five to six messages per day. On average, 1.5% of these messages have an attachment with 9% of them being malicious. However, some messages are solicited as proved by the fact that users read and whitelist an average of 1.5 messages per day. Our data also confirm the belief that normal users often intentionally open emails with malicious attachments, and hence perform poorly when telling spam and ham apart.

While most of the existing research studies the efficiency of anti-spam techniques and their enhancements, often using very specific data feeds, this thesis focuses on the thin line that separates the two categories: those few cases in which the existing techniques fail. In particular, we limit our study to the often overlooked area of *gray emails* [172], i.e., those ambiguous messages that cannot be clearly categorized one way or the other by automated spam filters. We start from the assumption that spam filters are good in detecting most of the spam, and if the filter has "good reasons" to believe that a message is unsolicited or it contains malicious content (e.g., by employing an antivirus, a black list, or by matching a signature of a known scam message), there would be no reason for most users to double-check that decision. On the other end of the spectrum, we have legitimate user messages that we also assume are correctly classified as ham. And in the middle, there is a small class of messages that is hard to classify automatically and that is often misplaced in the user's mailbox or spam folder [142]. Finally, the fact that this is an important problem was confirmed recently by Google that, after we conducted our study, has announced the release of inbox tabs [8] – inbox emails grouped into categories, e.g., social networks, promotions, forums.

Experimental Environment

Traditional anti-spam solutions are based on two common techniques: filtering emails based on their content, or filtering them based on their senders. The first category includes content-based text classification techniques [40, 64, 145, 146] that aim at finding (often using supervised learning) the tokens commonly associated to spam messages. The second category includes detection methods based on some properties of the sender [89, 129, 138], of his reputation [26, 167], or of the domain from which the email is delivered [60, 77, 167].

Even though these two categories cover most of the widely adopted techniques, one notable exception is the Challenge-Response (CR) filter [71, 128] – a solution based on the observation that the large majority of good emails are delivered from senders that are already known to, and trusted by, the recipient. Hence suggesting that in general the first contact between email users happens much less often then the communications between already known contacts. The name of the approach comes from the fact that, whenever the sender of an email is unknown (i.e., not yet in the user's personal whitelist), the system temporarily quarantines the email and automatically sends back a message to the sender, asking him/her to solve a simple *challenge* to verify his/her legitimacy. This technique somehow changes the traditional approach of treating incoming emails, shifting the delivery responsibility from the recipient to the sender of the message.

Part of this thesis was conducted inside a commercial anti-spam company, MailInBlack, specialized in email management based on a CR anti-spam filtering system. Therefore, most of the experiments conducted within this thesis use the email datasets available within the company (Chapter 3 and Chapter 4). This was a great advantage, but also came with some limitations. For example, we had limited access to the email content limiting our analysis to only email headers data and few other anonymized information.

Moreover, although the CR filter has some clear advantages compared with others anti-spam solutions, it was also subject to many controversies and critiques [30, 5] for its possible negative impacts. Therefore, we first conducted a study to analyze the impact and measure the effectiveness of a real-world deployment of a CR filter.

Problem Statement

Under these premises, the objective of this thesis is to:

- **Evaluate** the *impact* and *effectiveness* of a Challenge-Response filter as an email anti-spam filter;
- **Investigate** the content of the *gray area* with the goal of *reducing the burden* for email *users*, and **proposing** methods to automatically distinguish *email campaigns*;
- **Propose** a method to identify *Nigerian* scam email *campaigns.*

1.2 Outline and contributions

Even though the email gray area was already identified as the most problematic email subset by Yih et al. [172], a detailed analysis of the previous research on the subject identified that little is known about it. This section presents our methods and contributions on the analysis of the gray area and on other related topics. Note that our experiments were carried out in respect to the privacy of the data provided by the company and its limitations.

In the first part of the thesis, we start by measuring and evaluating the performance of the challenge-response system as an anti-spam filter. Our goal is to provide real-world figures and statistics and to help to shed some light on some of the myths related to CR anti-spam techniques. For that reason, we evaluate the effectiveness and measure the impact of a real-world deployment of a challenge-based anti-spam solution. To achieve the objectives, we analyze the behavior of CR systems from three different perspectives:

- From the end user's point of view, to measure how this technique affects the delivery of both normal messages to the end user's mailbox;
- From the server administrator's point of view, focusing on some of the problems of maintaining a CR installation in a real company;
- From the Internet point of view, to measure the amount and the impact of backscattered messages and misdirected challenges.

The study concludes that in general the CR systems provide an excellent anti-spam protection (99.9% detection), and its side effects have comparably low impact and costs. We estimated that the quarantined emails – pending emails to which a challenge is sent – constitute 30% of all incoming emails. Around 6% of quarantined emails are later delivered to the user inbox, suggesting that the majority of the emails in this area stay unnoticed or are irrelevant. In fact, this kind of anti-spam filtering system provides a vantage point to study the *gray area* emails phenomenon. The quarantined area is a good approximation of a gray area as it already excludes most of the personal user messages and spam emails. A subsequent approximate evaluation of email similarity in this area demonstrated that emails can be grouped into campaigns, where some have very dynamic characteristics and few authenticated emails (the ones in which the sender solved the CR challenge), and some have more static characteristics with higher authentication rates. This insight led us to the second part of this thesis – the investigation of the content of the *gray area*.

In order to ensure email user privacy and avoid analyzing personal user emails, which are not the focus of our study, we approach the study of *gray area* as bulk emails, by focusing on the analysis of email campaigns [172]. In Chapter 4 we present a three-phase approach based on message clustering, classification, and graph-based refinement. Our technique was able to automatically classify half of the gray emails, which constitutes to 15% of the total system emails being classified additionally, and with only 0.2% of false positives – a measure commonly used in spam filtering as misclassified ham messages can be very expensive due to their costs of loss to the users. Additionally, among the classified emails we found five times less emails that reached user inboxes than in unclassified emails, suggesting that most of the unique user emails stayed unclassified. Our analysis unveiled the most and the least predictive email campaign attributes. Moreover, we demonstrated that by using a graph-based refinement method, legitimate email campaigns can be often identified based only on sender information.

During the study of gray area, we identified four email campaign categories – commercial, newsletters, botnet, and scam/phishing. Our classification method works well on all campaigns except on scam/phishing campaigns. This is due to the fact that these campaigns have some common traits with legitimate ones. Therefore, in Chapter 5 we propose a novel approach to look at scam campaigns. Our technique relies on the phone numbers as some cyber crime schemes [52, 62, 74, 123, 94] tend to rely on it as a communication channel for gaining profit or getting in contact with the victim. Although phone numbers are used in several different cyber scam and fraud activities, they play a particular role in Nigerian scam. Hence, we first evaluate the use of phone numbers in spam emails and different cyber schemes (Chapter 5), and then demonstrate that the phone numbers are especially used by the Nigerian scammers, comparing them with email addresses provided by scammers (Chapter 6). Based on this finding, we try to use this feature to characterize Nigerian scam campaigns, describing several illustrative examples demonstrating the diversity of modus operandi of such campaigns and their geographical distribution (Chapter 6). The analysis shows that there are rather few large campaigns, and the ones we identified often have some connections to Nigeria as a country suggesting that such cyber criminals tend to form distributed groups of criminals.

This thesis makes the following contributions:

- We **study** and **measure** the effectiveness of challenge-response anti-spam filters;
- We **study** user behavior when treating gray emails;
- We provide an **analysis** of the *gray area* and its email campaigns using only email headers. This includes an overview of message categories, a comparison of the campaign attributes, and an analysis of the user behavior;
- We propose a novel **methodology** for the identification of email campaigns based on a supervised learning algorithm;
- We introduce a new **methodology** for the identification of Nigerian scam campaigns, which is otherwise impossible using only email headers;
- We **evaluate** the role played by phone numbers advertised by different spammers in emails, and other types of cyber crime.

The thesis is organized as follows: first, we start by discussing the background of email spam in Chapter 2 and by presenting the state of the art. Then in Chapter 3, we evaluate the Challenge-Response anti-spam filter using real-world data collected at a commercial anti-spam company. As a next step, in Chapter 4, we move to the analysis of the gray area emails and propose methods for identifying email campaigns using only email headers, and methods for identifying campaign categories, concluding that they are inapplicable for phishing and Nigerian scam type of email campaigns. Following our limitations, in Chapter 5 we evaluate a new feature – phone numbers, which are sometimes used in the email content – using which we identify groups of criminals, we also evaluate it on other types of cyber crimes. Next in the Chapter 6, relying on the acquired knowledge, we apply a multi-dimensional analysis tool, TRIAGE [158], on a public dataset of Nigerian scam emails and identify email campaigns with the help of phone numbers and email addresses as primary attributes, and characterize the diversity of Nigerian scam campaigns with examples; finally, we summarize the thesis and list possible future work in Chapter 7.

2
Background and Related Work

In this Chapter we define *spam* and the thin line that separates it from legitimate emails. We discuss the existing legislative efforts in fighting spam, and also present the main bulk email categories.

The second part of the Chapter is dedicated to the technical approaches to detect spam. We start from reviewing the existing spam filtering methods and techniques. We then emphasize the importance of the email dataset used for studying spam. Finally, we conclude the Chapter with a discussion of the techniques for email campaign identification and analysis.

2.1 Definition of Email Spam

Despite the fact that the research community has been concerned with spam for at least the part 15 years [57], it is very difficult to judge how much we succeed in combating it. Already back in 1998 [57] we knew that there are two ways of addressing the problem of spam: through technical means, or regulatory means. Since, the society has tried both approaches. But as much as we fight it, it keeps coming back in new, evolved forms, just like in a cat and mouse game. This section discusses the open issue of defining spam, especially in respect to gray area emails, and also describes the regulatory efforts that have been applied so far and their consequences. We conclude by providing our definition for spam that will be used in the rest of this thesis.

Although before 1998 [78] spam was simply a nuisance to the users, nowadays the problem grew to be more sophisticated. There are two main issues with email spam and spam in general: (i) it is annoying to the users and to ISPs as it heavily overloads both with mostly useless emails, and (ii) it became an instrument for conducting various types of cyber crimes, thus posing an increasing security threat to the users. The reason spam still exists is because it is actually effective (therefore profitable) and it requires low financial investments. But what especially escalates the problem is that it is so easy to fake one's identity, or create a new one. Email servers still lack reliable and automated mechanisms to evaluate their sender, and to verify their compliance with regulations to send newsletters or commercial emails.

The conventional definition of *spam* states that: i) it is an unwanted (unsolicited) email; ii) it is sent to a massive number of email recipients, in bulk. However, this raisesthe question of how to classify direct marketing emails. Are they also spam?

While it is possible to quantify the number of sent emails, it is very subjective to decide what is solicited for the recipients. Direct Marketing Association proposed to call spam only certain categories of spam, *"like porn and scams, send fraudulently"*, but this was perceived negatively by the community (and especially Spamhaus [25]) as an attempt to legalize some spam. According to the definition of spam given by Spamhaus, spam refers to messages sent *"as part of a larger collection of messages"* and where *"the recipient has not granted verifiable permission for the message to be sent"* [25]. Another way to put it is *"spam is about consent, not about the content"*.

Although the characteristic of *bulk* is still feasible to quantify and measure by setting up some thresholds, the other characteristic, *being unsolicited*, of spam email is almost impossible to define [165, 42]: the classical case of I-know-it-when-I-see-it. Unfortunately, recent techniques applied by spammers make the recognition of solicited emails by the ordinary user much more challenging. For example, the Blackhole spam sends messages almost identical to legitimate notification emails with malicious payloads [126].

At this point, automated spam filters become very important as they are much more efficient in identifying technical security threats than untrained users. On the other hand, the weakness of automated filtering is that it performs badly against social engineering attacks, and exhibit an average performance in predicting what users want to receive in their mailboxes (e.g., the new job offers, or Amazon alerts), and what they do not want. The solution to this is to use personalized email filters [50, 173]. But the core problem lies in the automatic indistinguishability of *spam* and *non-spam* email [165]. Here we deliberately avoid the term of email legitimacy as it is subjective to current anti-spam laws (CAN-SPAM [4] in US, E-Privacy Directive [31] in EU) that keep email spamming practice within the law boundaries. However, one could argue that we should provide some regulations for e-marketers for them to be able to legally run advertisement campaigns, which is a common practice for example via post mail.

In the next section, to better understand the nature of solicited bulk emails, we discuss in more details how email marketing campaigns operate today.

2.1.1 Email As a Marketing Tool

Email-based marketing is a common practice for advertisement and sales, and it is used both to maintain communication with current customers, as well as to acquire new ones. Unfortunately, when the user inbox started to get overloaded with various types of bulk messages, mailbox maintenance became highly time-consuming. Mail filters were introduced to protect users from unsolicited emails, starting a multi-million industry and a battle that is far from being over. In fact, even though direct mail marketing has a higher response rate (3.4%) than email marketing (0.12%) [61], the low cost of emails still makes electronic messages a very attractive solution for marketers.

Today due to an elevated email delivery complexity, marketers often use professional marketing tools in order to maximize their campaign delivery rates. These tools help to clean customer lists from non-existing emails, to deal with recipient complaints, to avoid hitting spam traps, and even provide a detailed campaign delivery statistics (e.g. MailChimp [10]). Thus, running email marketing campaigns became intricate, and more solicited bulk emails fall into the spam folder, forcing users to regularly check it manually. In fact, this folder often contains messages that cannot be clearly categorized by automated spam filters, the so called *gray emails*. Part of them are solicited bulk emails, and another part are unsolicited bulk emails. This second category contains both harmless message and emails that can result in a computer infection or steak sensitive personal data. However, users appear to be ineffective in distinguishing dangerous emails [115, 127]: the majority of users (70%) decide on email class based on the sender field and subject line.

Therefore, distribution of legitimate marketing campaigns became a business where specialized companies provide professional email marketing tools, and also sell categorized email lists as a service for marketers looking for new clients. The collection of such lists is officially legalized: when users subscribe to some services and fill out a form, they might by choice or by a default opt-in share with third-parties their contact information and be categorized corresponding to their topics of interest. Hence, at some point users do agree to receive advertisements. What recipients cannot do is to verify at any given point of time in which opt-in lists they are registered and which are their categories of interest. To compare, in phone marketing campaigns, marketers can only contact people that are located in the special lists for marketing, or in publicly available lists, hence having a certain level of control over their level of exposure to marketers. To compensate, marking and newsletters emails have to provide operational unsubscribe option, but not a way to be removed from the lists.

As we see, the line between the solicited and unsolicited emails might be very thin. Due to the complexity of the prospective client list creation, it is also very difficult to verify them, especially automatically. Next, we discuss what are the existing legislative laws that address this issue.

2.1.2 Bulk Email Legitimacy

In every country the definition of bulk commercial email legitimacy might differ. Here we present the two definitions that are used in US and in European Union.

E-Privacy Directive [31] is a directive applied in the European Union for digital data protection and privacy. Article 13 of this directive prohibits unsolicited commercial emails

- *"Prior to explicit consent of the recipients is obtained before such communications are addressed to them"*.

Thus a prior permission needs to be granted by the recipient to receive commercial content via email. However, there are also two exceptions permitting commercial email communications:

- Existing customer relationships;
- Marketing of similar products and services (Article 13(2)).

The latter means that if for example a user is a customer of a car insurance service, other car insurance companies in competition have the legal right to send the same user commercial emails promoting their products or services. Although from a regulatory perspective this definitions looks pretty straight forward, from the perspective of an automated anti-spam filter the automatic validation of these clauses is difficult.

The US CAN-SPAM Act [4] published in 2003 permits unsolicited emails communications if the message complies with:

- *Unsubscribe*: Unsubscribe mechanism is provided, 10 days delay to unsubscribe request, opt-out lists are used for compliance, not to be abuse to collect user emails;
- *Content*: The messages uses well-formed and relevant From and Subject headers, provide legitimate physical address of the advertiser, and labels the adult content;
- *Behavior*: No usage of open relays, email harvesting, header spoofing, and other illegal behaviors are tolerated.

It also excludes "transactional or relationship messages". The Act was perceived with a lot of criticisms from the anti-spam communities, stating that the act instead of prohibiting the spam, was actually approving the practice [48]. The Coalition Against Unsolicited Commercial Email (CAUSE) commented on the Act [47]: *"This* *legislation fails the most fundamental test of any anti-spam law, in that it neglects to actually tell any marketers not to spam. Instead, it gives each marketer in the United States one free shot at each consumer's e-mail inbox, and will force companies to continue to deploy costly and disruptive anti-spam technologies to block advertising messages from reaching their employees on company time and using company resources. [...] In addition, the law's weak provisions are further crippled by limiting enforcement to overworked regulatory and law enforcement agencies, rather than giving consumers legal tools with which to protect their own inboxes. [...] [The Act] only makes that spam slightly more truthful."*

As Grimes [86] estimated, the actual compliance with the Act in 2006 was very low, around 5.7%. However, later on SPAMHAUS [26] published its discontent message about the Act [2] reporting that *"during 2008 a few USA spammers honed the technique [of snowshoe spam] to a fine edge"*. Snowshoe spammers send messages from a dedicated range of static IPs, that one often leases specifically for bulk email sending. In February of 2009 snowshoe spam accounted for 20-30% of all mail server connections, and was the second largest source of spam after the botnet spam. While botnet spam was being sent from dynamic IP space promoting illegal products or services, it's rival used static IP ranges sending commercial messages in the form compliant with the US law.

Conclusions

As discussed in this section, the definition of email messages being *spam* and *not-spam* is controversial, and can be a borderline for some cases, depending on the user preferences. On one side, it is possible to quantify the amount of the same email copies being sent, on the other, it is very difficult to automatically verify if the message is solicited and is sent with the consent from the recipient. However, the *legitimate bulk senders*, like marketing companies selling as a service online marketing tools that comply with the anti-spam regulations, or companies sending newsletters to their customers, tend to fall into the area of *gray mail*, and thus are prone to classification errors, promting users to check their spam folders.

It is known that email users from time to time (some even regularly) check their spam folders to verify that no important messages went missing from their inbox, especially for notifications or other subscribed services. The main contributing factor is the lack of automatic verification of recipient subscription and its consent.

For clarity, in the analysis of gray area presented in Chapter 4 we will classify bulk emails into two categories: legitimate and spam. The first can be *solicited* (e.g. subscribed newsletters, advertisements, and notifications) or *unsolicited* (e.g. illegal advertisements) messages sent according to *legal regulations* (e.g. E-Privacy Directive [31]) and using static network infrastructure and email headers for sending bulk emails, hence be identifiable as a marketing company. The second category corresponds instead to *unsolicited malevolent* or *illegal* emails (e.g. illegal products, malware or targeted attacks, personal data and credentials theft). In other words, we will consider direct email marketing as *legitimate* emails, although different users might view it differently [59], continuing the infinite debate over the definition.

2.2 Email Campaign Categories

In this section we list and provide definitions for the existing spam and non-spam email campaign (bulk) categories. We limit ourselves to list only the main categories and recent category trends. According to statistics published by Spam-Law [23], spam categories are split as follows: commercial spam covers 36%; adult content spam corresponds to 32%; 26% contribute to financial spam; scams and fraud comprise only 2.5% of all spam email (however, phishing makes up 73% of this figure).

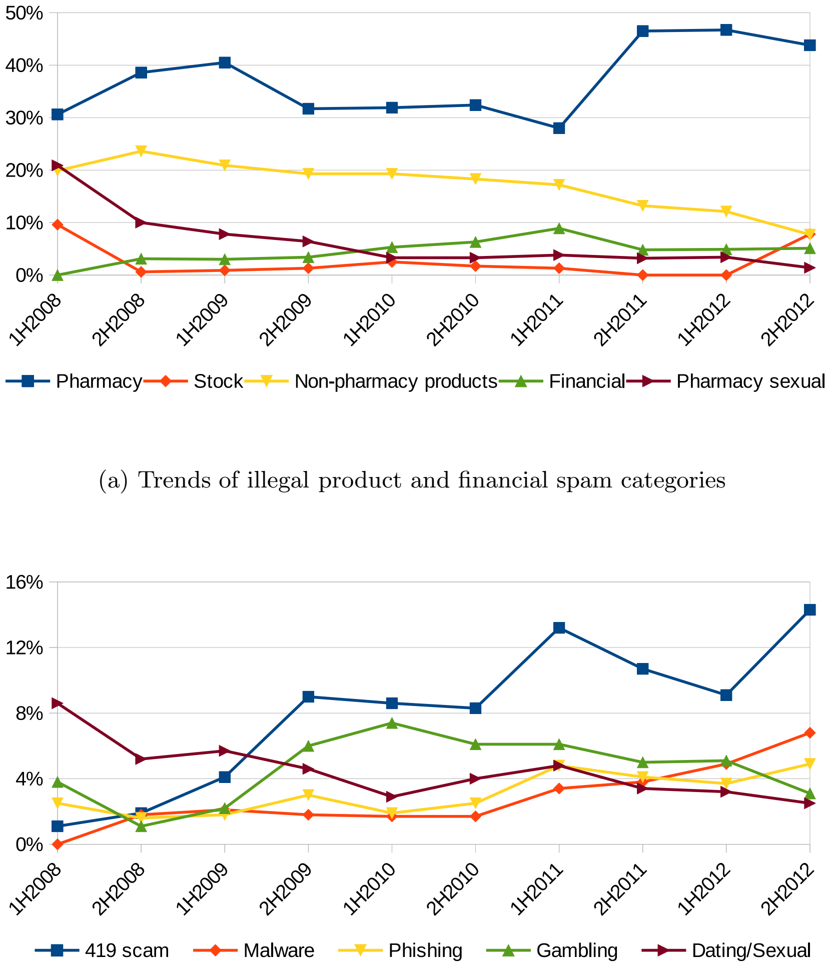
Another set of reports from Microsoft Security Intelligence team [33] present the shares of each spam category seen from 2008 until 2012. A histogram of spam category evolution is presented in Figure 2.2 where we find that the major part of spam belongs to pharmaceutics and other types of advertised products (non-pharmacy products). Figure 2.1 shows a more detailed view of the trends in some spam categories. Sexual pharmaceutical products and other publicity have strongly declined since 2008, while general pharmaceutical adds have a small increase (Figure 2.1a). Also, the stock category varies a lot, suggesting that this type of spam is connected with the current stock market situation. Figure 2.1b indicates that categories of spam with malware, phishing, and 419 scam payloads are raising since the 2008. It is especially impressive to find such high numbers of 419 scam in 2012. Finally, as for legitimate campaigns, ReturnPath Email Intelligence Report [141] states that they correspond to up to 42% of user inbox messages (newsletters and automated notifications).

Next, we list email campaign categories that could be found during our email campaign identification. We provide a short summary of the state of the art to introduce the reader with the current types of campaigns.

2.2.1 Emails Sent by Botnets

The *botnet email spam* is probably the most well-known and also the most studied category so far. This is mostly illegal spam sent from infected machines all over the world. There are two type of messages sent by botnets:

- Spreading malware aiming to infect new machines;



(a) Trends of illegal product and financial spam categories

(b) Trends of other spam categories

Figure 2.1: Spam email categories from 2008 till 2012 from Microsoft Security Intelligence Reports [33]

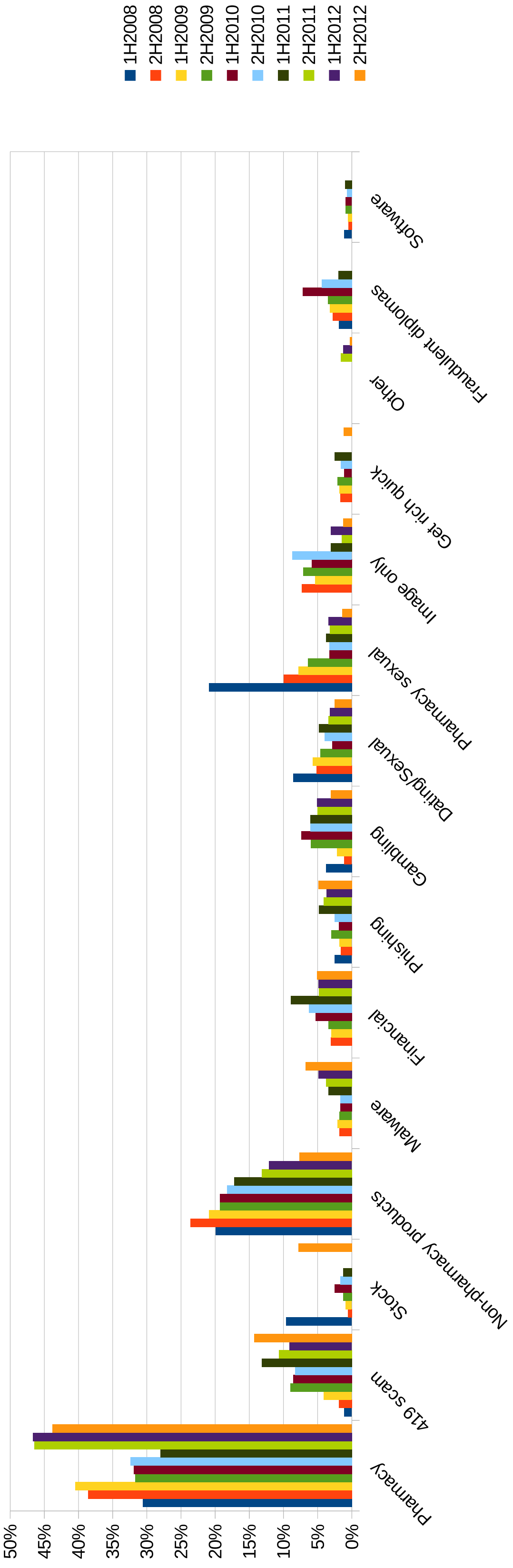


Figure 2.2: Spam email categories from 2008 till 2012 from Microsoft Security Intelligence Reports [33]

- Sending spam, e.g., illegal commercial advertisements, black market products and services.

Until 2010, botnets were responsible for most of the world email traffic (up to 89% [155]). This percentage started to drop in the past years due to several botnet take-overs, reducing the global spam shares.

Most of the previous work on spam focused on botnet-generated spam [46, 100, 107, 119]. The main techniques to identify such campaigns include: spamtraps, botnet signatures based on SMTP dialects [153], URL analysis. Actually, a large part of the anti-spam techniques are designed to fight botnet-generated spam. Thonnard et al. [160] have even studied the strategic behavior of email spam botnets using a complex multi-criteria analysis technique.

In our study of email campaigns, we show that the identification of botnet-generated campaigns based only on email headers is feasible and we demonstrate it by using an ensemble classifier. Botnet campaigns exhibit very dynamic sending patterns, and are using "dirty" email lists (with many non-existing recipients) when sending emails to multiple recipients at a time. The results of our analysis demonstrate that this behavior is very uncommon for legitimate campaigns.

Blackhole Spam

Blackhole spam is a subset of botnet-generated spam, and is a recent trend in spam campaigns. The name comes from the *Blackhole Exploit Kit* that since its release in 2010 became a popular tool on the black market. The kit is dedicated to exploit known vulnerabilities of web browsers, Adobe Acrobat Reader, Flash Player, etc. In the last years these campaigns became quite voluminous and have drawn the attention of anti-spam and anti-virus company researchers [126, 131].

In our work we also identified the presence of such campaigns. They have sending patterns very similar to the ones of conventional botnet-generated spam. However, the topics of the messages differ as they are not promoting any products; but they are trying to run phishing campaigns where emails mimic messages from known delivery companies, social networks, etc.

2.2.2 Phishing Emails

Phishing emails are disguised messages sent on behalf of some known person or some well-known organization that aim to acquire confidential information, e.g. account credentials, and credit card data. They imitate actual companies by redirecting the victims to attackers websites that also mimic the actual ones, hence fooling users into giving away their sensitive personal data. Such messagesneed to provide a way to the victim to communicate their information, thus mainly relying on URLs in the message body. These URLs are often obfuscated, redirected, and use other cloaking techniques to avoid being detected. Such attacks often use spoofed email addresses and they are sent from compromised machines or other untraceable mail servers.

The problem of phishing has been well studied in the research community [76, 174, 108, 49] and many techniques and tools were proposed to secure the users from falling into these traps. However, phishing attacks are still on the rise. For example, Stevenson [151] estimated an increase of 17% per year. Hence, phishing emails are still an important problem [109] that is actively researched.

The phishing detection techniques fall into three groups:

- Detection based on the website content by Zhang et al. [174];
- Detection based on the combination of email data and website by Fette et al. [76];
- Detection based on the email content alone.

The latter was studied by Chandrasekaran et al. [49] where authors proposed to use SVM classifier on features like language, layout, and structure of the email. As we also identify a group of phishing email campaigns, we need to point-out that our solution is limited in identifying such campaigns. That is because such campaigns exhibit sending patterns that are very similar to the ones of legitimate campaigns, and therefore they are difficult to characterize without performing an additional body content analysis. Phishing campaigns, at least in our experiments, use few machines for distributing the load of the campaign and use consistent email header information. While botnet-generated campaign are sent from a large pool of IP and countries, over longer time periods and with randomized email headers.

2.2.3 Targeted Email Attacks

This kind of attack is more damaging compared to the previous ones, and it is also more difficult to prevent. In the targeted email attacks the victims are chosen carefully. To approach the victims, a targeted attack often relies on a highly personalized and relevant to the recipient information, thus gaining their trust in order to open email attachment or follow the URL.

The attack can be directed against selected individuals, groups of people, organizations, or specific domains. As described by Thonnard et al. [159], the main characteristics of such attacks are the fact that the attackers send emails to selected recipients, with a malicious payload, and in a low volume. The attackers are also driven by different goals such as: stealing sensitive information and intellectual property; spying on their victims; or gaining control over over victim resources.

Today, targeted email attacks are still quite rare, hence are more difficult to identify and study. But, as it appeared from a recent study by Thonnard et al. [159], the attacks are organized into campaigns that are run in a "determined and patient" fashion, over long periods of time, often keeping a low profile. Additionally, the malware in attachment is different from the ones used in other types of email threats, being more complex to identify. Some of them are even using zero-day payloads with low level of obfuscation [159]. However, the victims are often tricked through a social engineering effort, with the emails constructed for particular individuals to match their interests, or by spoofing the addresses of victim's friends in the sender field.

However, Thonnard et al. [159] have tried identifying such campaigns from a preselected dataset of such emails. The authors used a multi-criteria clustering techniques in order to identify inter-connected emails through a set of features (e.g. some emails would reuse the same subject, the same sender, or the same ID number), creating a link between them. However, the biggest problem with the detection of such email is their rarity. As they are rare and difficult to identify, it is extremely difficult to detect and study them.

Amin et al. [35] proposed to use a Random Forest classifier on a set of features extracted from targeted emails to detect them. They report the most important features to be *persistent threat* and *recipient oriented* features. Similarly, Lee [110] presented a study of targeted email attacks, in which he tried to apply epidemiology prediction techniques to evaluate the odds ratio of the recipients being a victim of a targeted attack.

A particular type of targeted malicious email goes under the name of *spear phishing*. It is a subset of targeted attacks with a phishing type payload where emails are also sent in low volumes and with a short list of targeted recipients. These attackers additionally can be profit-driven, aiming at highly ranked victims (e.g. top-managers). The success of such attacks depends on their low volume and personalized threats, making them difficult to detect. Often, spearphishing emails provide some information about the victim (e.g. account name), thus gaining a level of trust, and accompany the message with an URL lure recipient into the trap.

In the context of the thesis, our proposed method is effective in identifying bulk email campaigns, but not campaigns that are sent in low volumes. Note that even the work of Thonnard et al. [159] has analyzed an already preselected dataset of targeted emails without trying to detect it. Therefore, the detection of targeted malicious messages is still an open problem.

2.2.4 Nigerian Scam Emails

Nigerian scam, also called "419 scam" as a reference to the 419 section in the Nigerian penal code, has been a known problem for several decades. Originally, the scam phenomenon started by postal mail, and then evolved into a business run via fax first, and email later. The prosecution of such criminal activity is complicated [44] and can often be evaded by criminals. As a result, reports of such crime still appear in the social media and online communities, e.g. *419scam.org* [1], exist to mitigate the risk and help users to identify scam messages.

Nowadays, 419 scam is often perceived as a particular type of *spam*. However, while most of the spam is now sent mainly by botnets and by compromised machines in bulk quantities, Nigerian scam activities are still largely performed in a manual way. Moreover, the underlying business and operation models differ. Spammers trap their victims through engineering effort, whereas scammers rely on human factors: pity, greed and social engineering techniques. Scammers use very primitive tools (if any) compared with other form of spam where operations are often completely automated. Even though today 419 scam messages are eclipsed by the large amount of spam sent by botnets, they are still a problem that causes substantial financial losses for a number of victims all around the world.

A distinctive characteristic of this particular email fraud is the communication channel set up to reach the victim: from this point of view, scammers tend to use emails and/or phone numbers as their main contacts, while other forms of spam are more likely to forward their victims to specific URLs. For instance, a previous study of spam campaigns [130] (in which scam was considered a subset of spam) indicates that 59% of spam messages contain a URL.

Nigerian scammers employ various techniques to harvest money from ingenuous victims. Tive [161] describes the tricks of Nigerian fee fraud and the philosophy of tricksters behind. Stajano and Wilson [149] studied a number of scam techniques and showed the importance of security engineering operations. Herley [91] looked into economical aspects of adversaries by trying to understand how scammers find viable victims out of millions of users, so that their business would be still profitable. A brief summary of Nigerian scam schemes was presented by Buchanan and Grant [44] indicating that Internet growth has facilitated the spread of cyber fraud. They also emphasize the difficulties of adversary prosecution - one of the main reasons why Nigerian scam is still an issue today. A more recent work by Oboh et al. [124] discusses the same problem of prosecution in a more global context taking the Netherlands as an example.

Another work by Goa et al. [80] proposes an ontology model for scam 419 email text mining demonstrating high precision in detection. A work by Pathak et al. [130] analyses email spam campaigns sent by botnets, describing their patterns and characteristics. The authors also show that 15% of the spam messages contained a phone number. A recent patent has been published by Coomer [14] on a technique that detects scam and spam emails through phone number analysis. This is the first mentioning of phone numbers being used for identifying scam. In this thesis, we will study the role of the phone numbers in various online fraud schemes and empirically demonstrate its significance in 419 scam domain. We will also look into the scam campaign characterization by relying on phone numbers and email addresses used by scammers as the main campaign linking features.

2.2.5 Snowshoe Spam

Snowshoe spammers send messages from a dedicated range of static IPs, often leased specifically for bulk email sending by ISPs. However, the difference between legitimate mailers and snowshoe spammers is very thin. Spamhaus defines these category of spam as: "*snowshoe spamming is a technique used by spammers to spread spam output across many IPs and domains, in order to dilute reputation metrics and evade filters*" [7]. Additionally, such spammers use "*many fictitious business names, fake names and identities, and frequently changing postal dropboxes and voicemail drops*". In contrast, legitimate commercial email senders, according to Spamhaus, build a reputation over time and maintain it, use small and well defined IP ranges, and identifiable domain information. As of 2009, this type of spam corresponded to 20-30% of the email traffic [7].

Unfortunately, today these is no existing valid technique to reliably identify these bulk email senders. As during our study we encounter many commercial type email campaigns, it is very difficult to tell to which bulk sender group do they belong to as most of the commercial bulk senders in our data appeared to be using dedicated IP ranges, sometimes with varying domain names.

One of the ways for marketers and commercial campaign senders to verify their IP range reputation is to check it on SenderScore [17]. The latter is a website providing IP reputation score for bulk email senders ranging between 0 (bad) and 1 (good). The score is tracked over time and calculated based on several criteria, e.g. blacklisting, complaints, spamtraps, etc. The web service is especially helpful for marketers as they can track their reputation and improve it in case of a problem. However, the queries are limited to a small number per day, and hence cannot be used in a large-scale email campaign study.

In respect to our work, we identified a large numbers of commercial email campaigns and due to the lack of their validation, we assumed that they are legitimate and reputable senders. However, a deeper study of the data and snowshoe spam is out of the scope of this thesis.

2.2.6 Marketing/Commercial Emails

Direct marketing is a old common practice for advertisement and sales used both to maintain customers and to acquire new ones. When direct marketing moved to electronic mails, email user inboxes started to get overloaded with various types of bulk messages.

By marketing emails we refer to commercial emails sent by professional marketers that try to maintain the communication with the existing customers, or try through legitimate means to acquire new customers, business offers, etc. Note that today it is possible to acquire (from legitimate companies and for a fee) prospective customer email lists that are already categorized based on the customer interests and preferences. These lists can be built in different ways. For example, during a customer subscription to receive some services from some private company or a website, customers can be asked if they wants to receive third-party commercial offers on the same topic. In this case, the company would acquire the rights to share customer acquired data with other companies specialized in that particular domain. The procedure of building such lists differs in different countries. We omit a detailed study of such rules and regulations and further refer to this subset of email campaigns as *legitimate commercial email senders* that acquire their recipient lists in a regulated fashion. However, as noted previously, there is no known automated solution that could be used to verify email (campaign) legitimacy.

In the arms race against spammers, the evolution of anti-spam filters has lead to a steady increase in the detection rates, but, as a consequence, it has also lead to an increase in the false positive rates, especially for legitimate commercial/newsletter email senders. For example, a recent Email Intelligence Report [141] pointed out that 16% of emails containing advertisements or marketing information are normally flagged as spam and, therefore, never reach user mailboxes. At a first glance, many people would consider this “side-effect” as an advantage. However, it has been estimated that only one-third of the users consider such messages spam, while two-thirds prefer to receive unsolicited commercial emails from *already known* senders [59].

To our knowledge, there are no previous studies looking into the characterization of legitimate commercial emails or campaigns, neither into their identification methods. Some studies [172, 50] refer to gray email area as to email subset that seems to be difficult to classify automatically, but authors primarily aim to identify spam.

2.2.7 Newsletter/Notification Emails

Another category of legitimate email campaigns consists of *newsletter and notification*. Users that receive such emails have in the past subscribed to some online services or community, and are regularly updated with the new content. For example, this category includes notifications from forums, social network services, mailing lists, dating website notification. The difference between commercial email campaigns and this category is that the content of notification and newsletters is typically not a commercial advertisement. Moreover, users can often customize through the website of the appropriate service how often the notification should be sent and the kind of notifications the user wants to receive. Because of this, the servers used to send these emails are often dedicated to the provided services and can be identified as such, instead of being “outsourced” to a specialized company.

Interestingly, a recent report showed that despite the mailboxes being overloaded with legitimate campaigns, consumers still read around 18% of the subscribed emails, and continue to sign up for email offers and mailing lists [141]. As a result, newsletters and automated notifications sum up to 42% of user inbox messages. However, as delivery of such emails is prone to errors, it is a well known fact that users need to regularly check their spam folder to verify that no important messages have been misclassified by the anti-spam filter. So far, we were able to find only statistical information about such type of messages. The Email Intelligence Report from Return Path [141] also reports that such email messages continuously encounter email delivery difficulties, sometimes appearing in the spam folder, or, in the worst case, being never delivered.

2.3 Spam Filtering Techniques

Spam content can reach the users via different means of digital communication, e.g.: emails, search engine results, comments on web sites, SMS, IM services and social networks. The problem of email spam has been an object of research for at least over 15 years. Even though some of the existing techniques can be applied in different contexts, it is often the case that each context requires a specific technique. In the rest of this Chapter, we will focus only on the email filtering techniques, most of which fall into two categories: content-based and sender-based filtering.

Content-based spam filtering techniques rely on signatures and machine learning algorithms, applied primarily on the email content [145, 37, 64, 40, 146]. Although content-based filters were initially very effective against spam and provided an acceptable level of protection [114], with the evolution of text obfuscation in spam bodies, the effectiveness of content based solutions has reduced over time. Hence, today anti-spam systems often combine both categories, where sender-base techniques play an important role and often represent the first line of defense of an anti-spam system.

Sender-based techniques can be especially helpful in filtering out spoofed emails (sender authentication) and emails sent from malicious IP addresses (reputation). In this section, we discuss well-known spam protection solutions, and also present some less popular, unconventional solutions.

2.3.1 Content-Based Filters

Filters in this category rely primarily on the content (body) analysis of the message. The major part of such filters are machine-learning algorithms, while others use text mining techniques, such as duplicate document detection, to build spam email signatures [51, 104]. Also, it is possible to derive email templates (with regular expressions) as demonstrated by Pitsillidis et al. [132] from the polymorphic spam messages received by monitoring emails sent by botnets.

There is a wide variety of machine learning algorithms used along with different text categorization techniques that aim at recognizing spam messages. Machine learning algorithms are usually able to achieve high detection rates (e.g., 96% [102, 106]) and require less manual intervention due to their ability to adapt and re-learn. In order to apply a machine-learning algorithm on email data, the bodies need to be first transformed to be interpretable by the machine learning algorithm. This transformation is an important process as it largely impacts the results and algorithm efficiency. Therefore, input data requirements might differ depending on the algorithm, but general text conversion steps are as follows [87]:

1. *Word extraction* (tokenization) takes place;
2. *Stemming* is performed, to *reduce* words to their root forms ;
3. *Stop words* are removed;
4. Words are *converted* to the specific format required by the machine learning algorithm. The most commonly used techniques are: bag-of-words, character n-grams and n-gram words models, tf-idf (term frequency-inverse document frequency);
5. The most prominent *features* are *selected*. This helps to improve classification accuracy of an algorithm, ignoring non-informative features [87]. The *feature selection* is performed using one of the following common techniques: information gain, document frequency, TFV (Term-Frequency Variance).

After these transformations are completed, the machine learning algorithm is applied on the transformed and selected features, providing as an output a classification model. The results of classification are often by impacted by the process of information extraction and transformation into features. Hence, there is a variety of machine learning algorithms for email classification that are combined with different transformation methods. The majority of the algorithms are applied on the message body, almost always excluding the headers due to the transformation challenges.

The most well studied algorithms tested on spam data are Bayesian [145], SVN (Support Vector Machines) [64], ANN (Artificial Neural Networks) [53], Logistic Regression [84], k-NN (nearest neighbors) [146], Artificial Immune Systems [125], and Ensemble [175] (combinations of classifiers taking the decision based on their votes). There also exist some comparative studies of these algorithms and their performance, as presented by Guzella et al. [87] and Kiran et al. [102]. However, Ensemble is one the most efficient methods reaching 96.4% classifier accuracy as shown by Kiran et al. [102] and Koprinska et al. [106]. But the actual list of existing algorithms is even longer as shown by Guzella et al. [87], thus this area continues to remain an active area of research.

In this thesis we primarily study emails by looking only to email header information. Similarly, Wu et al. [168] proposed a machine-learning algorithm that was applied to email header information and mail server logs. The study demonstrates that even without the email content spammers exhibit a specific repetitive behavior that is different from legitimate senders. The algorithm was supplied with a set of rules derived by a specialist based on observations, where a classifier was trained to identify the behavior of the spam senders – resulting in high detection rates. In our work, we also use header information in combination with a machine-learning algorithm, but we feed the classifier with the extracted features that characterize groups of similar emails – email campaigns – hence, identifying the legitimate and spam campaign sending patterns. Furthermore, we apply graph analysis metrics to improve the false positive and false negative rates.

Another comparative study of content filters was performed by Cormack et al. [55]. The study relied on several widely used open-source machine-learning spam filters that were compared by running them on the same dataset. Emails were grouped into three categories: spam, personal user emails, and advertising (e.g. advertisement, notifications, news digests, mailing list messages). As a result, most of the filters appeared to perform well on distinguishing spam and ham emails, but had difficulties to correctly classify messages from the advertising category [55]. The authors explain this defect by a low number of such messages present in the dataset. However, an annual reported by an email marketing monitoring company, Return Path, reports that in 2012 such emails corresponded to up to 42% of mailbox emails [141] and the numbers were growing. In our study of spam campaigns, we empirically demonstrate that these emails represent a considerable portion, providing methods to effectively categorize them and identify them in real-time based only on the sending characteristics.

2.3.2 Sender-Oriented Filters

Authentication

In contrast to content-based filters, sender-based filtering techniques aim at blocking spam by analyzing sender information, typically by authenticating the sender and verifying its reputation. Authentication techniques authenticate the sender either by verifying its source and domain [167, 60], or by providing a protocol to authenticate the server at each message delivery. For example, Fleizach et al. [77] suggest for mail servers to start using an authentication protocol, which unfortunately is required to be implemented at both client and server sides. The same drawback is applicable to a number of other proposed email authentication schemes, as the ones proposed by Gariss et al. [81] and Prakash [135].

However, some authentication schemes are still in use and are primarily oriented toward preventing email spoofing, a phenomenon that is very common among spammers. There are two widely deployed techniques that are mainly used to validate the sender domain name: Sender Policy Framework (SPF) [167], and DomainKeys Identified Mail (DKIM) [60]. The latter links a domain name with the message through a digital signature that can be validated by the receiving mail server. It can be applied to a person, or an organization, and the signature is located in the email headers ready for validation by the recipient. To use SPF, the owner of the IP address(es) has to publicly declare the IP ranges used to send emails for a specific domain. This allows the recipient to validate if a domain name is eligible to send an email from a specific IP (note that spammers can also register SPF records with their IP ranges and overpass the filter as it was demonstrated by Mori et al. [121] and Sipahi et al. [148]).

Domain-based Message Authentication, Reporting and Conformance (DMARC) [34] presents the most recent effort (published in 2012) to fight email spoofing, by proposing a technical specification for email authentication. While the aim is to prevent email sender spoofing (using email authentication identifiers such as SPF [167] and DKIM [60]), DMARC also supports mechanisms for forensic and aggregate report generation on rejected emails. In order to benefit from DMARC, the specification needs to be implemented by the mail server, which then needs to be configured as it has several running modes, ranging from passive to pro-active. To authenticate an incoming email, the server contacts the sender's server asking to validate the request, and vice versa, the server would be capable to respond to such demands. The difference of this specification to previous efforts is that it provides more feedback, especially about rejected/spoofed emails, and it has been supported and deployed by the 10 of the top 20 email providers in the world [143], representing 60% of world's mailboxes, thus encouraging smaller ones to also adopt this technique.

Reputation

This category covers a borad range of techniques such as sender IP reputation (e.g. Spamhaus [26] using DNS-based Blocklists (DNSBL) and Whitelists), network-level feature analysis, and sender behavior anomaly detection. IP reputation techniques [26, 99] rely on whitelists and blacklists of IP addresses that are known to either send spam or to be trusted sources. The goal is to identify misbehaving sources and add them to the DSNBLs that later are used for dropping the incoming traffic from these sources.

To identify misbehaving IPs, one of the traditional methods is to use *spam traps.* These are email addresses advertised deliberately online but hidden from humans, so that if they can be collected by automated email harvesting tools. When an email is delivered to such addresses, its source gets blacklisted. Depending on the list, the IP address blacklisting may last for different periods of time. There are also other techniques employed by blacklisting services, such as counting the number of complaints filed for a certain IP address.

Such approaches are often effective against static spammers and open-relay servers used for spam distribution. On the other hand, they might be unable to keep up with newly appearing spamming bots, since botnets dispose a large number of IP addresses and can change them quickly using a large number of different infected machines. As shown in the experiment by Ramachandran et al. [138], between 8.5% and 30% of spam senders stayed undetected by spam traps for at least 30 days. Furthermore, between 10% to 35% of spam that was delivered was undetectable by the RBL. On the other side, some blacklists started blacklisting IP clusters, e.g. BGP clusters, hence reducing their efficiency. Qian et al. [137] proposed to rely on this insight and combine it with DNS information, improving the precision of public IP-based blacklists by 50%.

Therefore, to include behavioral factors, behavior-based solutions were proposed. Pathak et al. [129] suggested to analyze sending behavior of spammers, while Ramachandran et al. [138] used behavioral blacklists to classify senders (IP addresses) based on their sending behavior. Ramachandran study is based on the observation that spammers exhibit recognizable sending patterns, based on which behavior fingerprints can be build from a set of targeted domains. Then, the new sender behavior is compared in similarity to the previously identified behavior. Authors estimated a detection improvement of 10% over the ones previously undetected emails (1.5% of total spam). In another study by Ramachandran et al. [139] of network-level behavior of spammers, the author found that sometimes spammers use "short-lived BGP routes, typically for hijacked prefixes“ to distribute spam messages. Hao et al. [89] built an automated reputation engine, called SNARE, aiming to distinguish legitimate senders from spammers based on the non-content email features. Such network-level detection techniques tend to react faster to spam campaigns than typical blacklisting services and have a lower number of false positives. However, this kind of solutions block only a part of illegitimate emails, and, therefore, again need to be used in combination with other filters. Finally, West et al. [166] built an efficient reputation model for identifying predictive behavior of spammers. The model is relying on spatial and temporal feature that can be especially useful in partial-knowledge situations. The model managed to classify up to 50% of the spam emails with low false positives that were not identified by the blacklists.

2.3.3 Other Filters

Verification of the sender SMTP protocol

Besides the more common techniques presented so far, a number of other solutions have been proposed to protect users against spam. An example is a *greylisting* technique [164, 79, 111] that temporary rejects (soft fail) or delays unknown senders expecting them to retry the initial email delivery. The *retry* functionality is a default mail server behavior that has a different time window depending on the implementation of mail server. On the other hand, spam mailers are often optimized to send as many emails as possible, and often omit this feature due to performance considerations. This characteristic of course depends on the implementation of the email client used to send emails. There are several existing implementations available of this technique [90, 63]. Also, as pointed out by Levine [111], the configuration of greylisting may vary, affecting the number of possibly lost legitimate emails.

A similar but more advanced technique, called B@bel [153], verifies email delivery mechanisms used by the sender at the SMTP level, hence identifying the client application used for delivering the email. It appears that every implementation of the SMTP communication protocol varies from one to another, therefore making it possible to build a state machine describing the "language" used by the email client. Based on that, the client might be classified as a spam sender and such communications can be blocked before even accepting the email, this makes this technique resource efficient and especially effective in identifying mail clients used by botnets.

A different approach by Esquivel et al. [72] suggests to build a collaborative SMTP server architecture that creates TCP fingerprints (OS fingerprints) of machines sending spam and shares the data with routers to reject such senders. The technique is a complement to existing sender-based filters as it identifies only 28%-59% of incoming spam.

Image recognition

A number of techniques to detect image spam was proposed by Guzella et al. [87]. Some of the techniques rely on OCR (Optical Character Recognition), while others use other content information (e.g. meta data), to make the algorithms faster, but with a lower accuracy. For example, one of more recent techniques by Hsia et al. [93] exploited hidden topics within images by using semantic analysis, and achieved detection accuracy of 92-95%.

Pay-per-email

Some researchers [163, 32] suggested to add a cost element to emails, as the core problem of spam comes from its zero cost [79]. Several different solutions were made to identify a pricing function used by the sender, where the client must compute a function before sending any message [144]. Although the enforcement of micropayments seems to be an efficient solution, some potential security issues were described by Turner [163] and other issues of logistics and policy enforcement were discussed by Gansterer et al. [79]. As a result, this category of filters was never deployed.

User Whitelists and Challenge-Response (CR)

A special category of anti-spam filters is the ones relying on email users personal *white* and *black* lists. The assumption behind this technique is that users mainly communicate with an almost static list of contacts that does not change much over time [41, 65, 71]. The challenge in this case could be solving a CAPTCHA [71] (Completely Automated Public Turing test to tell Computers and Humans Apart). One of the most wide-spread approaches for building and maintaining a list of trusted senders is based on the adoption of a challenge-response technique [117, 128, 20, 27, 18]. The solution is based on the observation that a large majority of good emails are received from already known and trusted senders. The name of the approach comes from the fact that, whenever the sender of an email is unknown (i.e., not yet in the user's personal whitelist), the system *temporarily quarantines* the email and *automatically sends back a message* to the sender, asking him/her to solve a simple challenge to verify his/her authenticity. This technique somehow changes the traditional approach of treating incoming emails, shifting the delivery responsibility from the recipient to the sender of the message.

Projects, like Internet Mail 2000 [41] and DiffMail [65, 66], provide a whitelisting feature only to the email recipient (the sender can do nothing to get his/her email delivered to the inbox) and require outgoing emails to be stored on a separate sending server, aiming to increase the burden on spammers by stocking their email on sending servers. A more flexible approach consists in providing a list of new pending emails from previously unknown senders to the user, and also to send a challenge to the original email sender, providing a possibility for him/her to solve it and move the email to the user's inbox. A user can whitelist or blacklist the sender, and also provide a chance for the sender to whitelist himself/herself. This might be important to speed up important communications and ensure the quick delivery of the original messages.

One of the main challenges with these approaches is to provide an automated way to share and maintain the users lists. Garris et al. [81] proposed a solution to this problem based on the idea of sharing the whitelist content with the user's friends on social networks. Their cryptographic solution addresses also the sender spoofing problem, and the protection of the privacy of the users during the sharing process. The main limitation of their system is the fact that it requires a large-scale adoption by many social networking users.

Although challenge-response schemes are extremely successful in blocking spam, they also have a number of limitations that makes them less attractive over other solutions [71]. Additionally, CR (Challenge-Response) solutions received a great amount of critiques from the anti-spam community [30, 5], often because of the amount of challenge emails they generate and because of the delay in delivering new emails.

To the best of our knowledge, the only empirical study that analyzes challenge-response CATPCHA-based whitelisting systems is presented by Erickson et al. [71]. The authors focus on the deployment and the usability of such system. The results of their evaluation support the usability of CR systems, but also show their limitations in coping with automatically generated legitimate emails, such as newsletters and notifications. On the other hand, the authors concluded that challenge-response systems are very successful to prevent spam and have lower false positives and false negatives rates compared to traditional content filtering techniques like SpamAssassin. Our CR measurement experiment (presented in Chapter 3) aims instead to present a comprehensive study of a real-world challenge-response antispam system, evaluating its effectiveness and its impact on the end-users, Internet, and server administration.

However, at the same time, it is important to keep in mind that such challenges could be solved automatically, as it was shown in several recent studies measuring the security level provided by CATPCHA and the cost of solving them [122, 45, 170]. Motoyama et al. [122] showed that CR schemes relying on CAPTCHAs can be solved at a low cost by the labor markets. Therefore, authors proposed to evaluate a level of authentication provided by CAPTCHA in “purely economic terms”: attacker gains versus solving the challenge price. For example, email attacks, such as a targeted email attack or a spear-phishing attack (which are low volume attacks), are considered to be much more profitable than traditional spam and could be worth the price of solving the challenges at the labor markets. In the same lines, several researchers [45, 170] demonstrated the feasibility of breaking the CAPTCHA using other methods.

2.4 Techniques for Email Campaign Identification and Analysis

As we said at the beginning of the Chapter, spam has two characteristics: it is sent *in bulk* and *is unsolicited*. Some studies try to identify spam by identifying bulk email campaigns (e.g. [112, 130, 156, 100]), since spam is by definition sent in high volume. Researchers used different approaches to study spam (mostly botnet generated emails as they correspond to the major part), thus generating different findings. In the remaining part of this section we present the techniques used to identify the campaigns and the methods to characterize them.

2.4.1 Identification

URLs

There are different methods to identify email campaigns. The most well-known are based on signature generation from the message content. Another popular approach is based on the analysis of URLs included in the messages criminals use to promote and sell underground products or services.

Li et al. [112] in 2006 were the first to cluster spam emails by URLs and their redirections, proposing a group-based clustering technique. It was very effective in the beginning, yielding 70% to 90% of spam being blocked. The authors also assumed that *"it is highly unlikely for a large group of legitimate senders to send emails with exactly the same type of signatures"*. Hence, during the experiments authors removed potential legitimate advertisement senders by whitelisting *.edu* domains, thus reducing a number of false positives.

Xie et al. [169] also used URLs to identify spam campaigns, although they reported that it appeared challenging to separate legitimate URLs from spam. They proposed a technique that required labeled data and used regular expressions to detect polymorphic URLs. The work mainly focused on URLs included in emails generated by botnets, where a spam campaign was defined as *"a targeted spam effort to a single product or service"*. The authors found that 74% of spam emails contained an URL, succeeding to classify into spam campaigns 10.8% of their initial Hotmail user labeled emails.

In the work by Pathak et al. [130] published in 2009, researchers tried to cluster spam emails sent from an open proxy into campaigns using URLs, but it proved to be a challenging task due to the evolution of techniques used by spammers, e.g. URL obfuscation. Authors were forced to perform quite a lot of manual sampling on their data, and also reported that 59% of spam contained URLs. In the study authors argued that previously proposed methods [169] to use URLs to identify spam campaigns became less efficient creating extremely high false positive rates due to a false assumption of *spam burstiness* (many spam emails sent in short burst). Their findings suggested that the problem of email campaign identification is hard, and current solutions still suffer from the high false negative rates.

Thomas et al. [156] in 2011 confirmed the challenges of spam email analysis based on URLs, and proposed a new technique to filter URLs in real-time achieving 91% accuracy. Adding additional URL features, like HTTP content and header information to already used ones like DNS and IP type information, has strongly improved the accuracy of the classification. However, for the system to continue to perform well, it required a constant re-training and new labeled datasets due to the evolution of URLs over time.

Content similarity

Another spam campaign identification method category consists of email content comparison. For example, Chowdhury et al. [51] performed a word collection statistics by generating campaign signatures and hashing token streams extracted from email bodies. Authors used an I-Match algorithm enhanced with context, relying on the lexicon; however, this technique is sensitive to text obfuscations. Kolcz et al. [103] in a follow-up work proposed improvements for the I-Match algorithm.

Calais et al. [46] proposed to use frequent pattern trees to identify spam emails belonging to the same campaigns. First, a list of features was extracted from the emails, and then the trees were build, where similar ones were grouped into campaigns. The authors used subject, URL, language, layout and message type as email similarity features. They studied 97 millions of spam emails collected from honeypots where 91% had an URL. Unfortunately, the authors reported no information about the portion on emails that was classified into campaigns, hence it is difficult to compare their method.

Probably the closest work to the one we present in Chapter 4 is the one performed by Qian et al. [136], where the authors proposed an unsupervised email campaign clustering algorithm, and recognized the problem introduced by legitimate campaigns in a real-world dataset. Qian et al. [136] identified email campaigns based on its text mining algorithm and unsupervised learning scheme. During the study, the authors assumed that the *"majority of the emails are spam belonging to some spam campaign, and that campaigns are generated from templates"*, the latter suggesting that they would be following some patterns. They used a text mining algorithm (Latent Semantic Analysis) for campaign signature generation, and two algorithms for URL and HTML signatures. The study covered over 80% of spam emails with 3.5% of false negatives, explaining the false negatives by the poor visibility of the dataset. Low visibility can lower the identification of spam campaigns and its coverage. In our study of email campaigns we reached different results. For example, we identified a group of legitimate distributed campaigns where we showed that using the thresholds applied by Qian et al. [136] the result would yield 10% of false positives compared to 0.2% rate that we achieved in our experiments. They tried to eliminate legitimate bulk emails by using keywords, and by defining an IP diversity threshold per campaign. They used a threshold of 10 unique IPs per campaign, the value below 10 being a legitimate campaign. While the latter can be efficient when dealing with campaigns sent by botnets, it is ineffective against other malevolent campaigns sent from webmail accounts. We demonstrated in our study that such campaigns tend to mimic the traits of legitimate campaigns and, therefore, it might be challenging to identify them automatically. We also presented better features that can be used to distinguish legitimate campaigns.

Judo [132] is an automated tool designed to extract spam campaign templates from botnet spam feeds. It extracts patterns from the message body, URLs, and headers transforming them into regular expression templates that can be used for spam email detection. Unfortunately, the process requires an up-to-date spam training set to identify the patterns (in this case it was a bot running in a sandbox) and build the templates.

Finally, Thonnard et al. [160] applied a multi-dimensional clustering technique in order to identify campaigns from the pre-selected set of botnet-generated email. This technique relies on several different features from the emails searching for links between emails and building them based on the weights of the selected features. Hence, in the analysis the authors showed that such campaigns were in some cases sent by several botnets, sharing the load. The same tool, TRIAGE [158], was also used for identification of other types of email campaigns from a pre-selected datasets of emails [160, 159, 56]. Our goal is to identify campaigns from the gray email dataset where the categories of email campaigns might be unknown a priori. Hence, our methods permit to identify new, previously unseen, email campaigns, thus enabling to track current spam trends along with legitimate bulk campaigns.

As we can see, most of the existing techniques used the content of the message and/or URL in order to identify spam/email campaigns. Some basic techniques were used for cleaning up the spam campaigns from legitimate ones. In our study, we were forced to use different campaigns identification methods that work only with email header data. We also propose more accurate methods to separate commercial and newsletter email campaigns from traditional spam campaigns.

2.4.2 Characterization

After the email campaigns are identified, they can be further analyzed to gain a deeper understanding about the different types campaigns (e.g. illegal, malicious, commercial, legitimate, newletters), and about the behavior of the users involved in them. Here, we discuss the accumulated knowledge of the research community concerning email campaigns, sending characteristics, and user behavior.

One of the first studies on characterization of spammers and traffic behavior was carried out by Gomes et al. [83]. The authors aimed at extracting the most prevalent characteristics of spam email traffic that would help to differentiate it from the non-spam traffic. This work is close to ours as it compared legitimate emails versus spam. However, it studies email traffic, and not email campaigns. Still, some indicative features overlap with ours. As a dataset, they used emails from a big university, already labeled by standard anti-spam tools. The analysis was performed only on the email header information. The study was based on the assumption that the transmission of ham emails is *"driven by social bilateral relationships"*, whereas spam performs an *"unilateral action"* [83]. The main findings of this study showed that:

- Spam is sent in a regular fashion: regarding time, number of emails, size of emails, and numbers of senders and recipients;
- Ham is normally sent during working hours and rarely during weekends and nights;
- The size of the spam emails is 6-8 times smaller than that of ham;
- Only 5% of ham emails have more than one recipient, whereas 15% of spam emails have more than one. Hence, this appears to be a good characteristics for separating the two classes;
- Two groups of email recipients are identified: sporadic reception, and stable, predictable reception.

In another study by Gomes et al. [82] researchers tried to identify several graph metrics that could help to characterize spammers and legitimate senders/recipients. They studied email user communication patterns by building probabilistic spam detection framework [82]. As a result of the study, the authors proposed to use a combination of graph metrics as none of them predicts the class accurately. The list of prevalent metrics is as follows:

- Spam and ham nodes have different clustering coefficients;
- Legitimate users have a higher probability of being emailed or to email someone;
- Communication reciprocity is very different, reflecting the chance that the received email will get a reply;
- Email asymmetry set (the difference of in and out edges per node) is different;
- Legitimate graphs reflect a structure of Social Networks, whereas spam is similar to technological networks.

Hence, they suggested to use these metrics together, and also along with other traditional anti-spam filters. Compared to our work on characterization of legitimate and spam campaigns, we also find that clustering coefficient metric used in graphs is indicative feature. Additionally, we also confirm that legitimate groups of campaigns have much stronger communication links and are more stable, resembling largely Social Networks. However, in our study it is difficult to draw any conclusions about spam campaigns as there are not many that appeared in the graph–based refinement step – due to the fact that we focused our study on the gray area.

Additionally, we also study user behavior as which emails they view unveiling many users being interested in spam emails, and demonstrate much higher numbers of multiple recipients per email from spammers. Some legitimate campaigns also send multi-recipient emails, but very few. Finally, we add more additional features (e.g. IP distribution, country distribution, sender email prefix and suffix, bounces and rejections, and unsubscribe headers) that permit us to achieve high classification rates and to better understand the differences between email campaign categories, spam and non-spam. Finally, we propose an approximate method to categorize spam and legitimate campaigns into categories that reflect campaign incentives (e.g. newsletter, and commercial message).

Kanich et al. [100] proposed a tool called Spamalytics used for infiltrating the Storm botnet. The authors described the spam campagins sent by Storm, measuring email sent rates, delivery rates, block rates and user click-through rates by imitating spam webpages and logging user activity on them. The authors studied the spam campaigns and evaluated their conversion rate from the marketing perspective: campaign success rates, sales (user behavior), and user infection rates. This study focused on botnet-generated campaigns and their promoted product categories. The same group of authors has purchased illegal products promoted by criminals promoting them through botnets (e.g., pharmaceuticals, adult products, and other illegal goods), and has uncovered the whole supply chain along with their operational schemes and profits [118, 119, 101].

Spamscatter [36], is a technique that follows the URLs extracted from emails and clusters graphically similar web pages. The authors studied the infrastructure of these campaigns and categories, locations, and blacklisting. The study focused mostly on the characterization of the spam URL hosting methods.

Xie et al. [169] studied the characteristics that could be used for botnet activity identification. For this purpose, they analyzed the behavior of botnets during email spam campaigns. Although they looked at ways to detect botnets, they also indicated that, studied individually, botnets did not exhibit useful insights. However, when studied in groups, the authors found some potential behavior characteristics that could help to identify them. For example, botnets send campaigns to multiple users, sometimes to non-existing users, and tend to use high connection rates as they send email in bursts.

Another study [46] looked at the spam dataset from honeypots, analyzing their abuse patterns of open relays and proxies, so the analysis excluded spam sent directly to user email accounts. In our study, we cover also Nigerian scam campaigns and phishing campaigns, showing that they have very different sending and email header patterns. The authors also looked at the patterns of abused services, like the kind of senders that use open relays. It appeared that botnets largely relied on open relays.

Pathak et al. [130] studied the characterization of three botnet spam campaigns, analyzing the botnet coordination, and workload distribution. They demonstrated that spam campaigns can also be run during longer periods: as long as 3 months. In our work we also noticed a similar behavior. They also described bots contacting mail servers in short bursts of several minutes. The authors focused on the sending patterns and load balancing between bots measuring how loaded the bots are, and how they distribute a certain spam campaign in a rather short time. Moreover, they found that some bots were using alphabetical recipient lists and dispatched the transmission over the network of bots. Another similar study of botnet campaigns by Thonnard et al. [160] studied botnet load distribution, where they also covered the time frame of a Grum botnet take-down; and show that the jobs unfinished by Grum were transmitted to other botnets.

2.5 Email Analysis Datasets

Spam evolution over time is often ignored because previous studies often rely on datasets that are several years old, thus failing to respond to the recent changes and advancements of the real spam ecosystem.

However, the study in 2009 by Pathak et al. [130] already points out that classifying spam into campaigns using URLs has become much more harder than it was in the study from 2006 by Yen et al. [171] where authors showed that the majority of the URLs in emails can be classified. Another study by Thomas et al. [157] in 2011 proposes a tool that classifies URLs in real-time, where we see that the complexity of the tool is much higher compared with one from 2006. In our study we also use a recent mixed dataset of emails that is very close to real-world spam data, incorporating commercial campaigns, few ham messages and spam.

In one of the recent measurement studies of spam feeds [133] it was shown that the type of dataset strongly impacts the results of the research. The authors recommend using "human identified feeds" as they provide good spam dataset coverage and visibility. They compare with other types of spam feeds like: botnet spam, open-relay spam, honeypot spam, and DNS blacklists.

Spam activities appear to cover only a part of the whole picture. In a number of previous spam analysis papers, researchers focused on specific spam datasets and proposed solutions based on their datasets. However, there is no measurement performed on the extent to which such techniques would be beneficial during actual deployment. Some researchers tried to estimate the impact using their universities campus email servers. However, in our measurement study of a Challenge-Response email filter (in Chapter 3), our final results based on a real-world deployed system were different from the ones reported in university campus.

Qian et al. [136] propose an unsupervised approach for email classification through detection of spam campaigns, relying on open-rely dataset and a dataset from the authors campus. In this paper, the proposed approach achieved a high detection accuracy. Unfortunately, our analysis of email campaigns shows that by applying the same thresholds, the filter generates around 10% false positives. Therefore, campus datasets might not always provide a realistic data coverage to study different spam and bulk email trends. Also the definition of a legitimate campaign is unclear in this particular work, i.e. whether commercial campaigns were considered as legitimate or not by the authors.

Finally, the closest dataset to ours was used by Yih et al. [172], where the authors studied *gray emails* constructed from *human identified feeds*. Although, in our study we have no precise human feedback about the category of the email (only whether users opened/whitelisted/blacklisted the email), our dataset consists of emails that were unrecognized by the system neither as unwanted, nor as wanted. At the same time, gray emails provide a diverse ground for studying current spamming trends as messages generated by more sophisticated spammers tend to end up in this category.

In previous studies of the gray emails phenomenon, researchers relied on “human identified feeds” as in [173, 172, 50]. And as gray emails can contain, by definition, some misclassified user personal messages or other types of legitimate messages, in an ideal case a filter should understand (learn) what message a particular user considers as good and bad. For this reason, Yih et al. [172] argued that filtering gray emails with even an optimal spam filter is a very difficult task. Moreover, email/campaign classes may be different for each user [50, 59]. Hence, gray mails are very personal, making them not trivial to process by automated filters. Therefore, Chang et al. [50] performed a study on combining user feedback with user preferences to improve the final classification results.

2.6 Summary

As shown in this Chapter, the conventional definition of spam has its limitations when trying to apply it on commercial and newsletter bulk emails. The borderline area of emails where this type of bulk messages can be found is called *gray area*. From the few existing previous efforts that tried to analyze and improve this specific area, we know that researchers are aware of the problem, although few researchers looked at this problem. This is primarily due to the specificity of the dataset that is required to study the phenomenon.

Another fundamental issue with the *gray area*, apart from the specific dataset, is that currently in research there is no method proposed to identify commercial and newsletter emails or campaigns. An analysis of the state of the art showed that there are almost no effort done in that direction.

For those reasons, we need a more systematic and reliable study of the gray area and methods to reduce it. In the rest of this thesis we approach these goals in the following way: (i) we investigate the Challenge-Response system deployed in real-world mail servers, and acquire an approximation of the gray area emails from it; (ii) we analyze the area through campaigns where we find that a big portion of emails are commercial/newsletter emails; removing them would reduce the gray area by around 50%; (iii) we notice that our email header analysis method is limited in failing to identify phishing/scam campaigns, thus we propose a different, novel solution. The latter limitation also suggests that the prior studies of email spam based only on the sending patterns are doomed to have the same limitation.

3

Evaluation of a Challenge-Response Spam Filtering System

This Chapter introduces a measurement of a real world Challenge-Response anti-spam filter. The study was conducted in collaboration with a commercial anti-spam company specializing in a Challenge-Response anti-spam filtering system. As a first step of the thesis, we measure and evaluate the performance of the filter from three perspective: (i) an Internet perspective, in which we study how much extra traffic and misdirected challenges CR filters create; (ii) a user perspective, in which we measure the level of spam protection, delivery delays and the overhead of using user whitelists; and (iii) an administrator perspective, in which we take into account the overhead of administrating this type of anti-spam filters.

3.1 Introduction

Since the first introduction of CR-based techniques, they have been considered an extremely controversial solution [30, 5]. On the one hand, they seem to be able to completely block any unsolicited email, but, on the other hand, they also have a number of side-effects that can seriously hamper their adoption on a large scale.

In particular, it is possible to group the main criticisms against CR systems around three main points. First, the *social* and *usability* issues that, on one side, are related to the efforts required from the user to maintain a proper whitelist, and, on the other, to the annoyance for the sender that has to invest time to solve a challenge in order to have his message delivered. Previous studies, in particular Erickson et al. [71], have already studied the usability of CR systems in controlled experiments. Their study concludes that such systems are very effective when accompanied with already existing anti-spoofing techniques. The authors also measure that CR solutions outperform traditional systems like SpamAssassin, generating on average 1% of false positives with zero false negatives.

The second point against CR systems concerns the fact that they can introduce a (possibly conspicuous) delay in the emails delivery due to the quarantine period applied to previously unknown senders. Finally, the last (and one of the main) critique against CR systems is due to the challenge emails sent in response to spam messages. Since unsolicited emails often contain spoofed sender addresses, the challenges are often delivered to non-existing recipients or to innocent users. These misdirected messages (often referred as "backscattered" spam) pollute the Internet with unnecessary traffic and damage other users that may receive challenges for emails they never sent. From this point of view, CR antispam filters seem to literally bounce the spam back towards other innocent users. However, supporters of the CR approach often rebut by saying that well-designed systems only send back a challenge to a few percents of the spam messages they receive. Therefore, considering the fact that real forged addresses are not too common, normal users are very unlikely to often receive misdirected challenges. Unfortunately, since both sides lack real data to support their own hypothesis, it is hard for users and companies to tell which is the truth and take a conscious decision.

To the best of our knowledge, this Chapter presents the first study on both the effectiveness and the impact of a real-world deployment of a challenge-based antispam solution. In our work we measure and analyze a large amount of data collected for a period of six months from 47 companies protected by a commercial CR-based antispam product.

In particular, we conduct our measurements to analyze the behavior of CR systems from three different perspectives:

1. From the *end user* point of view, to measure how this technique affects the delivery of both spam and normal messages to the end user's mailbox;

2. From the *server's administrator* point of view, focusing on some of the problems of maintaining a CR installation in a real company;

3. From the *Internet* point of view, to measure the amount and the impact of backscattered messages and misdirected challenges.

It is important to stress the fact that the purpose of this study is neither to attack nor to defend CR-based solutions. Instead, our goal is to provide real-world figures and statistics that can help both users and companies to take an informed decision based on our study. Our results can also help to shed some light on some of the myths related to CR antispam techniques.

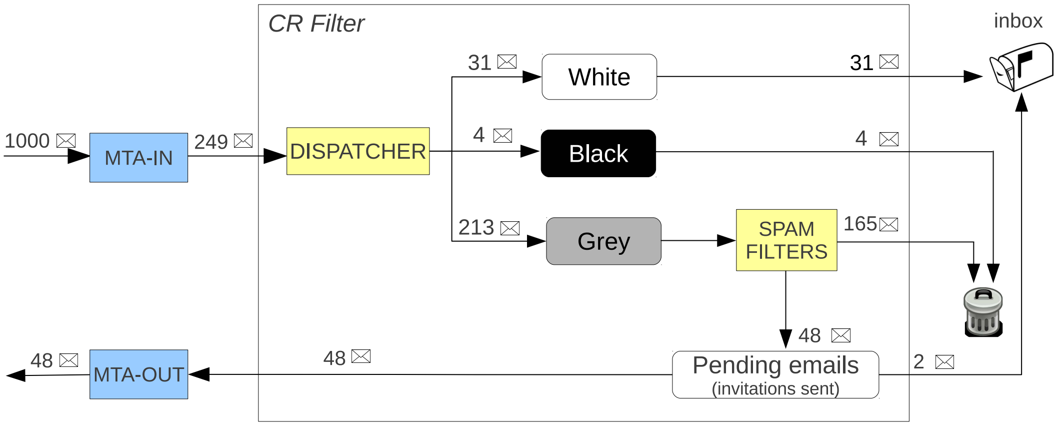


Figure 3.1: Lifecycle and distribution of incoming emails

3.2 Data Collection

In this section, we describe the dataset we used in our experiments and we provide a short overview of our data collection methodology.

3.2.1 System Overview

Our study has been carried out within a company providing an anti-spam solution based on a challenge-response technique. Figure 3.1 presents the overall system architecture and a "weighted" lifecycle of the incoming emails. The *CR filter* consists of two main components: a message dispatcher and a set of additional spam filters.

The dispatcher receives the incoming messages from the company's Incoming Mail Transfer Agent (MTA-IN) server. Some of the email servers were configured to work as *open relays*, serving emails also for a *restricted* number of domains that are different from the ones in which the systems are installed. This configuration allows the server to accept messages not targeting to, or originating from, known users in the system.

The MTA-IN server first checks if the email address is well formed (according to RFC822 [58]) and then if it is able to resolve the incoming email domain. In addition, if the server is not configured as an open relay, it also verifies that the recipient exists in the system.

Our study shows that this first layer of simple checks is responsible to drop more than 75% of the incoming messages (see Figure 3.2), while open-relay systems pass most of the messages to the next layer. These results are perfectly in line with similar values reported by the other analysis of spam delivery rate [152, 100]. The reasons behind the dropped messages are summarized in the following table:

Dropped Percentage	Reason
0.06%	Malformed email
4.19%	Unable to resolve the domain
2.27%	No relay
0.03%	Sender rejected
62.36%	Unknown Recipient

The second check point for the incoming emails is at the internal email dispatcher. This component is the core of the CR infrastructure and it is the one responsible for deciding to which category the email belongs to: white, black or gray.

The white and black spools are controlled by the user's *whitelist* and *blacklist*. Emails in the black category are dropped immediately, while emails from senders in the whitelist are delivered to the user's INBOX. Emails matching none of the previous lists fall in the gray category. These messages are then filtered with additional antispam techniques (e.g., virus scan, reverse DNS and IP blacklisting). If an email passes the filters, then dispatcher sends a challenge-response message to the original sender containing a request to solve a CAPTCHA. Otherwise, the email is considered spam and it is dropped.

Figure 3.1 also reports the average number of messages for each spool, assuming that 1,000 emails are received by the MTA-IN. The figures are computed by aggregating the data of all the monitored servers not configured as open relay.

Figure 3.3 shows that the other spam filters included in the CR engine drop on average 54% of the gray emails. Challenge messages are instead generated for 28% of emails. In the open relay cases, the engine filters have a lower performance rate, and the number of challenges sent increases by an extra 9%. This shows that, in an open relay configuration, the CR system receives more junk messages and it is more likely to reply with a challenge to illegitimate emails.

3.2.2 Whitelisting process

The process of email whitelisting involves both parties: the sender and the recipient. There exist several alternative ways for the email address to get added to a user's whitelist. In particular, the system we tested in our experiments supported the following mechanisms:

- The sender solves a challenge sent by the CR system as a response to one of his messages;
- The user authorizes the sender from the daily message digest;

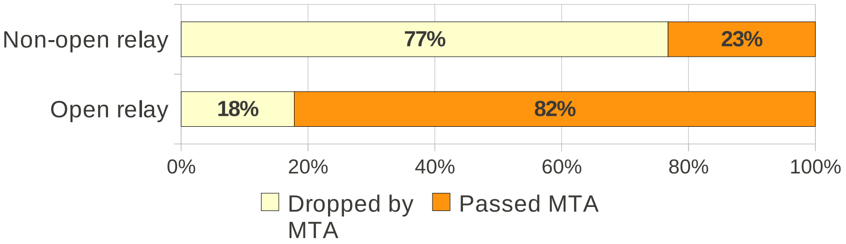


Figure 3.2: MTA-IN email treatment

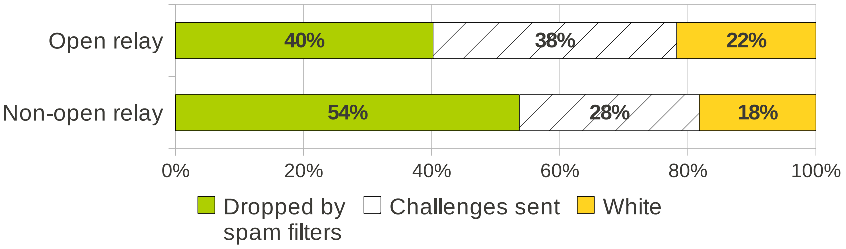


Figure 3.3: Message category at the internal email processing engine

- The address is manually added to the whitelist by the user;
- The user previously sent an email to that address.

In the general scenario, suppose that Alice sends an email to Bob, a user protected by a challenge-response system. If this is the first communication between Alice and Bob, the system temporarily stores the email in a "gray" spool and sends back a message to Alice. The message includes a link to a webpage that contains a CAPTCHA (the challenge) that Alice has to solve to get her email delivered and her address added to Bob's whitelist. After this simple authentication step, Alice's address is considered trustworthy, and the CR system will not interfere in any future communication between the two users, promptly delivering to Bob any further message coming from Alice.

If Alice does not solve the challenge, the email stays in the gray spool for a period of 30 days, after which it is dropped by the system. Bob also receives a daily digest that summarizes the quarantined messages, so that he can manually authorize them or delete them from the list.

3.2.3 General Statistics

In our experiment we collected statistical data about a commercial system deployed in 47 companies of different sizes. The monitoring period lasted for 6 months, between July and December 2010. For some of the servers we had access to the

General Statistics

Number of Companies	47	Challenge Sent	4,299,610
Open Relays	13	Emails Whitelisted from digest	55,850
Users protected by CR	19,426	Solved CAPTCHAs	150,809
Total incoming emails	90,368,573	Messages Dropped because of:	
Messages in Gray spool	11,590,532	reverse DNS filter	3,526,506
Messages in Black spool	349,697	RBL filter	4,973,755
Messages in White Spool	2,737,978	Antivirus filter	267,630
Total Messages Dropped at MTA	75,690,366	Total Msgs Dropped by filters	7,290,922

Daily Statistics

Emails (per day)	797,679	Challenges sent (per day)	53,764
Messages in White Spool (per day)	31,920	Total number of days	5,249

Table 3.1: Statistics of the collected data

data for the entire time frame, while for other companies our collection was limited to a shorter period of time (with a minimum of 2 months).

In total we collected statics for 90 millions of incoming emails. All the results were sanitized to protect both the end users and the companies privacy. In particular, we never got access to the message bodies and we stored only aggregated figures obtained from the automated analysis of the email headers.

The data collection was performed on a daily basis by analyzing the logs of the MTAs and of the challenge-response engines. In addition, information about the solved CAPTCHAs was collected by analyzing the access logs of the web-server serving the challenges. The extracted information was stored in a Postgres database and later analyzed and correlated by a number of Python scripts.

Table 4.1 shows some general statistics about the dataset we collected. Each company's server was configured to protect certain users with the challenge-response system, while protecting other accounts by traditional anti-spam techniques. In this paper we limit our analysis to the 19,426 users protected by the CR solution (this number includes normal users as well as administrative accounts and other rarely used email addresses). The table also shows the total number of the messages that we analyzed, the breakdown in the different spools (white, black, and gray), and some statistics about the effectiveness of the other spam filters included in the system (discussed in more details in Section 3.5).

Finally, since the number of days in which we were able to collect data varies between companies (for a total of 5,249 analyzed days), the table also report some *daily* statistics.

3.3 The Internet Point of View

In this section we focus on the consequences of adopting CR spam filters from a global point of view. In particular, we present an evaluation of the amount of challenge emails sent out by a challenge-response system during normal operation.

These *backscattered messages* are often criticized for two main reasons: the fact that misdirected challenges can be delivered to innocent users, and the fact that a large amount of useless messages are poured into the Internet, thus increasing the global traffic and overloading third parties email servers.

In the rest of the section we provide real-world measurements to estimate the impact of these two phenomena.

3.3.1 Email Backscattering

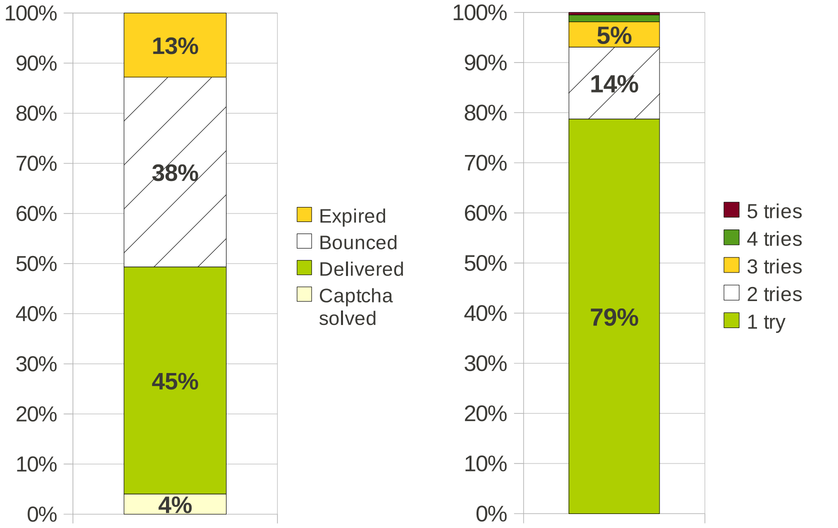
From an external point of view, a challenge response system can be approximated by a black box that receives emails from the Internet and separates them in three categories: some (the *white* set) are delivered to the users Inbox, while others (the *black* set) are immediately flagged as spam and discarded. The remaining messages (the *gray* set) are the ones for which the system is unable to take a decision. Therefore, for each email in this set, the system sends back to the sender another email containing a challenge to be solved. In this simplified model, a challenge-response system can be seen as a software that receives a certain amount of emails, and "reflects" a fraction of them back to the senders. This fraction, that we call *Reflection Ratio* $\mathcal{R}$, is an important parameter of a CR system.

By using the numbers in Figure 3.1, it is easy to compute the average reflection ratio: $\mathcal{R} = 48/249 = 19.3\%$ for the emails reaching the CR filter (or, $\mathcal{R} = 48/1000 = 4.8\%$ if we consider all the emails reaching companies' MTA-INs).

Understanding the Reflection Ratio

Is 19.3% a good value for $\mathcal{R}$? If not, what would be a reasonable value?
Unfortunately, it is very hard to answer these questions since it is not clear how to estimate which is an acceptable range for the reflection ratio.

To explain why, let us consider two extreme cases. In the first case, the CR system does not contain any other spam detector or blacklist mechanism. Therefore, the amount of challenges it sends is roughly the same as the amount of spam it receives, currently estimated between 80 and 90% [155] of the total email traffic. Values of $\mathcal{R}$ close to this range are obviously unacceptable, since, from a global point of view, the system would just act as a spam multiplier.



(a) Challenge delivery status distribution

(b) Tries required to solve CAPTCHA

Figure 3.4: Challenge statistics

In the second scenario, the CR system has been carefully configured and it has been associated with another perfect spam detector. In this case, the system never replies to spam and only sends back challenges to legitimate messages whose senders are not already in the recipients whitelist. In this case (represented by very low values of $\mathcal{R}$) the system does not generate any backscattered emails. Therefore, it may seem to be the final goal to reach in a perfect CR system.

Unfortunately, a very low value of $\mathcal{R}$ also corresponds to a completely useless system. In fact, if the internal spam filter can already distinguish good messages from spam, there is no need to add a challenge response system on top of it. In other words, in order to be useful a CR system has to be able to "substantially" reduce the amount of spam received by the users. However, this can only happen if the system sends back an equivalent "substantial" number of backscattered messages.

To conclude, the reflection ratio is a good indicator of the amount of challenges generated by a CR system. At the same time, it is important to be extremely careful to use this value alone to draw conclusions about the quality of such systems.

3.3.2 Misdirected Challenges

So far, we focused on the amount of challenges generated by a CR system. However, this value only measures the *amount* and not the real *impact* of the generated emails. In fact, not all the challenges are the same. Some of them reach the real senders and, despite being a little nuisance, could be tolerated as an acceptable price to pay for fighting spam. We refer to them as *legitimate challenges.* A second class of them is directed to non-existing addresses, and, thus, constitutes garbage traffic on the network. Finally, some misdirected challenges are delivered to existing spoofed email addresses, therefore reaching other innocent users. This category is much more harmful, and it is often referred to as *backscatter spam* (note that not all the backscattered *messages* are spam).

In order to distinguish the three categories of challenges, we analyzed the status of the challenge delivery in the servers' logs. In the systems under analysis, we found that only 49% of the challenges were successfully delivered to the destination servers. The remaining 51% were either bounced, or expired after many unsuccessful attempts (see Figure 3.4a). In the bounced set, a small portion has been stopped because the server that sent the challenges has been temporarily blacklisted (the problem will be discussed in more details in Section 3.5), while the large majority (71.7%) has been bounced due to the fact that the recipient address did not exist. This value provide a reasonable estimation of the amount of challenges that belong to the second category.

Another piece of the puzzle can be found by measuring the number of challenges that were actually solved. Previous work [71], conducted in a controlled environment, estimated that about 50% of the challenges were never solved. Unfortunately, our study shows a completely different picture. According to the web servers' logs of the companies we analyzed, on average 94% of the CAPTCHA URLs included in the delivered challenges were never even opened. The remaining were either solved (4%) or were visited by the user but not solved (0.25%). Figure 3.4b also shows the average number of attempts required to solve the CAPTCHAs. The fact that we never observed more than five attempts support the fact that probably there are still no real cases of attack against CR systems based on trying to automatically solve the challenges.

So far, we estimated the legitimate challenges to be at least 4% and the ones sent to non-existing recipients to be around 36.6% (71.7% of the 51% of undelivered messages). The third category, i.e., the backscattered spam, can instead be approximated with the number of challenges correctly delivered but never solved, i.e. somewhere between 0 and 45 %.

By combining the percentage of backscattered spam with the reflection ratio we presented before, we obtain the *Backscattered Ratio* β, i.e., the ratio of incoming emails for which the CR system sends back a misdirected challenge to the wrong user. In our experiments, we obtain, in the worst case, $\beta = 8.7\%$ (at the CR filter) or 2.1% (at the MTA-IN).

However useful, these figures must be considered only approximate upper bounds. For example, it is possible that challenge messages get dropped by some spam filter after being successfully delivered, or that real users ignore or intentionally decide to not solve a particular challenge. Finally, there are automatically generated emails (notifications from websites, mailing lists, receipts of purchase, ...) to take into account. When a user expects to receive such messages, he should either use an email address not protected by the CR system (functionality provided by the commercial product we have evaluated), or manually add the address to the whitelist.

Unfortunately, this is not always the case. In fact, we measured that around 2% of the message addresses in the gray spool have been whitelisted manually by the users from the daily digest. In other words, the challenge was not delivered or it was not solved, but the user recognized that the message was not spam and he manually added the sender to his whitelist to allow future communications.

3.3.3 Traffic Pollution

The reflection ratio only measures the number of messages, without taking into account their size. Therefore, it is not a very accurate indicator to estimate the amount of traffic generated by a challenge response system. For that purpose, we need to extend the previous definition by introducing the *ReflecteD Traffic ratio* $\mathcal{R}_T$, that represents the ratio between the amount of traffic received by the system and the amount of email traffic generated for the challenges.

To measure this new value, we deployed to all the servers a new sensor that extracts from the email headers the total size of the incoming messages and the total size of the outgoing challenges. Over a month period, the average ratio we measured at the CR filter was $\mathcal{R}_T = 2.5\%$. Unfortunately, we could not get a similar measure at the entrance of the MTA-IN servers. However, since the number of messages at MTA-IN is in average four times bigger than at the CR filter (see Figure 3.1), we can estimate that a large scale deployment of challenge-response spam filters would increase the email traffic on the Internet of around 0.62%.

3.3.4 Data Variability

In previous sections we reported aggregated figures for a number of different factors that can be used to estimate the "external" impact of a CR system.

In order to preserve the companies' privacy, each value was obtained by combining together the data collected from all the monitored installations. However, it is interesting to investigate what the variance of those values is, and if the size of the company affects in some way the presented indicators. For instance, it could be the case that CR filters work better for small companies, but fail to scale well to larger installations.

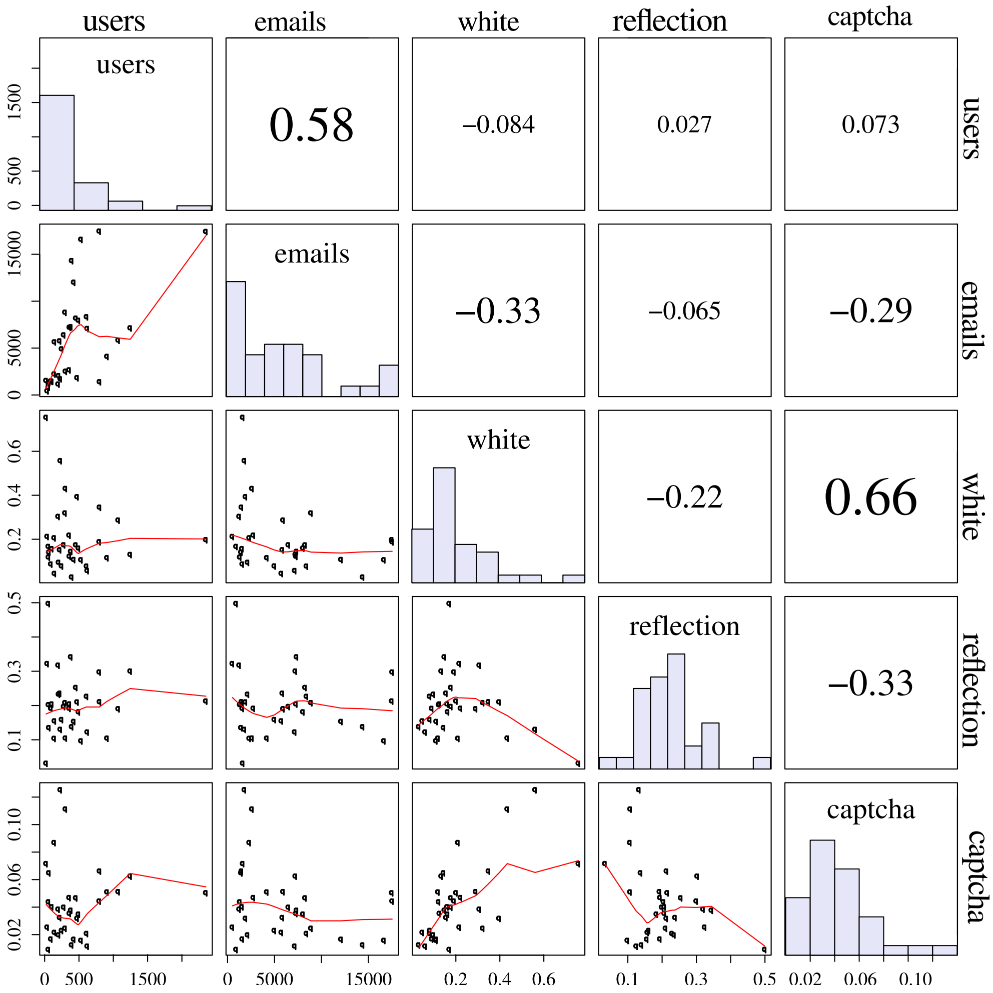


Figure 3.5: Histograms and correlations between different dimensions. Graphs on the diagonal represent the data histogram. Below the diagonal are the plots of the correlation between every pair of variables, summarized by the correlation values reported above the diagonals.

Figure 3.5 shows a scatter plot of five variables: the number of protected accounts (`users`), the amount of emails received daily (`emails`), the percentage of emails delivered in the white spool (`white`), the reflection ratio at the CR filter (`reflection`), and the percentage of challenges solved (`captcha`).

This graph represents a very efficient way to convey a large amount of information about the five variables. On the diagonal, it shows the histograms of the values of each variable. For example, the first element on the diagonal shows that most of the companies have less than 500 users, with few exceptions that have more than 2,000 users. Some values have a very high variability, such as the percentage of white emails that varies from less than 10% to over 70%. However, the two main coefficients we have measured in this Section, i.e. the reflection ratio and the percentage

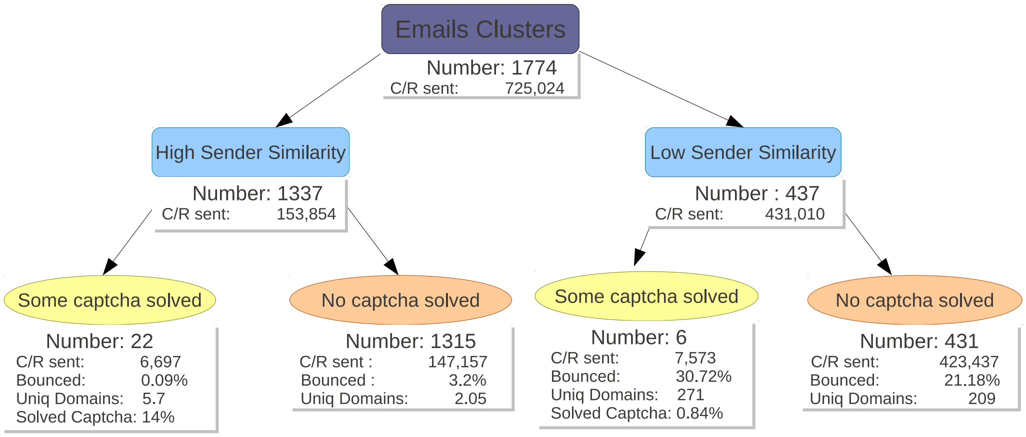


Figure 3.6: Spam clustering statistics

of solved challenges, seem to stay constant between different installations. The percentage of solved challenges only varies between 2% and 12%, and the reflection ratio stays in the range of 10% to 25%.

In Figure 3.5, the plots below the diagonals show the correlation between every pair of variables, while the upper part of the graph reports the corresponding correlation values (the font size allows to immediately focus on the higher values). Notably, the percentage of challenges sent by a CR system (`reflection`) is not correlated to the size of the companies (`users`) or to the amount of emails received. Not surprisingly, a small inverse correlation exists instead with the percentage of white emails. In other words, servers that receive a large amount of white emails (and therefore a lower amount of spam), tend to send less challenges and vice versa.

The rate of solved challenges (`captcha`) shows more correlations with other values, and in particular it is also strongly correlated with the white percentage. However, as the histogram shows, the variability of the `captcha` variable is so small that it can be considered almost a constant between the different installations.

3.4 The User Point of View

Despite the backscattering phenomenon described in the previous section, CR systems are often considered one of the most effective ways to protect users from spam. In theory, if the challenge-response system is properly configured, these systems should be able to achieve a 100% detection rate, thus blocking all unsolicited emails. However, previous studies [71] that confirmed this value were conducted on prototype systems evaluated in a controlled environment.

In this section we measure if this is actually the case in a real-world installation, and we evaluate the real impact for the end users protected by a CR system. In particular, we measure the delay introduced in the legitimate emails delivery, and the amount of spam that is able to reach the final users despite the CR filter. In addition, we also measure the change rate of the users' whitelists, one of the foundations of this kind of antispam solution.

3.4.1 Spam Protection

The main purpose of a CR system is to block all automatically generated emails coming from addresses previously unknown to the recipient. The first obvious consequence of this approach is that CR solutions are ineffective by design against targeted attacks, i.e., attacks in which the attacker manually composes a malicious message to target a particular individual. In fact, if the attacker receives back the challenge message, he can easily solve it and have his message delivered to the recipient. However, a recent Symantec report [155] estimated that only one out of 5,000 spam messages contains a targeted attack. In addition, all the existing anti-spam mechanisms can be easily evaded by targeted attacks, and, therefore, we can consider this threat beyond reach of all existing anti-spam solutions.

Unfortunately, targeted attacks are not the only ones that can pass through a CR filter. By studying a dataset of bounced challenges, we noticed that a large number of messages had the same subject and the same size. Per se, this is not surprising. However, a closer look revealed that while most of the messages were bounced or dropped by the filter, in some cases one of those emails was successfully delivered to the final user's mailbox.

To better investigate the reason behind this sporadic phenomenon, we decided to analyze the behavior, in terms of challenges and delivered messages, of a number of large spam campaigns.

For our experiment we applied standard clustering algorithms to the subject of the messages in the gray spool (i.e., the ones for which the system generated a challenge message). In particular, we put in the same cluster the messages with the same subject, limiting the analysis to the ones at least 10 words long. Finally, we discarded the clusters containing less than 50 emails. These very conservative thresholds were adopted to greatly reduce the risk of misclassification. In reality, the large majority of emails (including spam) have much shorter subjects, or they have enough variability to elude our simple comparison algorithm. However, our target was not to be able to cluster and identify all the incoming emails or all the spam campaigns, but just to identify a number of them with a low percentage of false positives.

The results obtained over a three month monitoring period are summarized in Figure 3.6. Our system identified 1,775 clusters, containing between 50 and 3696 messages each. In the next step, we divided the clusters in two categories, based on the sender email similarity. In the first group we put all the clusters where emails are

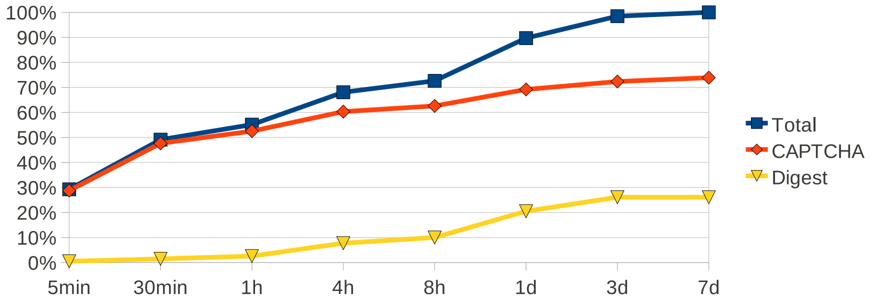


Figure 3.7: Cumulative effect of Captcha and Digest whitelisting

sent by a very limited number of senders, or in which the sender addresses are very similar to each other (for example, `dept-x.p@scn-1.com`, `dept-x.q@scn-1.com`, and `dept-x.p@scn-2.com`). These clusters are likely associated to newsletters or marketing campaigns. The second group contains instead the clusters with a very low sender similarity, i.e., the ones in which most of the emails originate from different domains and different senders' addresses. This behavior is very common in spam campaigns sent by malware infected machines.

Figure 3.6 shows that only 28 out of 1774 clusters contain at least one solved challenge. Moreover, these few clusters have very different characteristics, depending on whether they belong to the first or the second category. The ones with high sender similarity have a higher rate of solved challenges (some clusters as high as 97%) and almost no bounced emails. The clusters with low sender similarity have instead on average 31% of emails bounced because of non-existing recipient, and only one or two captchas solved each.

This second category is particularly interesting for our study. Each cluster in this group contains hundreds of emails, coming from many different domains, and often from non-existing sender addresses. However, out of these messages, sometimes one of the challenges was solved and, therefore, the email got whitelisted and delivered to the recipient's mailbox. These spam messages that are able to pass through the CR defense are likely a side effect of backscattered challenges that are sometimes erroneously delivered to innocent users. As a result, it is possible that one of these users solves a challenge for a mail he never sent. This phenomenon is, however, extremely rare. According to our measurements, we estimate that this kind of spurious spam delivery occurs ∼1 every 10,000 challenges sent. According to Table 4.1, this rate translates to an average of five spam delivery a day, over the 47 companies in our dataset. Excluding these isolated cases, CR systems are actually able to block all incoming spam messages.

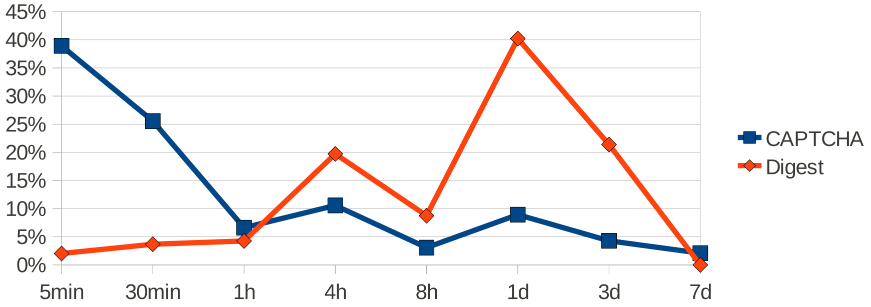


Figure 3.8: Time distribution of whitelisted messages

3.4.2 Impact on Messages Delivery

Another consequence of blocking automatically generated emails is the fact that also normal emails can get blocked and remain in the user's graylist waiting for the corresponding challenges to be solved. This can happen for two reasons: because the sender still has to solve the challenge, or because the email is sent by an automatic system and the challenge is, therefore, dropped or never delivered. In both cases, the user fails to receive a (maybe important) email.

Figure 3.7 shows the CDF of the messages that were moved from the graylist to the whitelist in the monitored servers. The two curves report the percentage of messages that were whitelisted because the sender solved the challenge, and the ones that were whitelisted manually by the user from the daily digest. According to the graph, 30% of the messages are delayed less than 5 minutes, and half of them are delivered in less than 30 minutes. However, if the challenge was not solved after 4 hours, then it is likely that it will not be solved at all (Figure 3.8). In those cases, the user has to manually select the messages from the digest, with a delivery delay that is on average between 4 hours and 3 days.

Combining the values from these figures with the number of white and whitelisted emails (31 and 2 respectively) in Figure 3.1, we can conclude that:

- 31/33 = 94% of the emails in the user's INBOX are sent from addresses already in the whitelist, and, therefore, are delivered instantly.
- Out of the remaining 6% (2/33) of the messages that are quarantined in the gray spool, half of them are delivered in less than 30 minutes because the sender solved the challenge.
- Only 0.6% (10% of the 6%) of the messages were delivered with more than one day of delay.

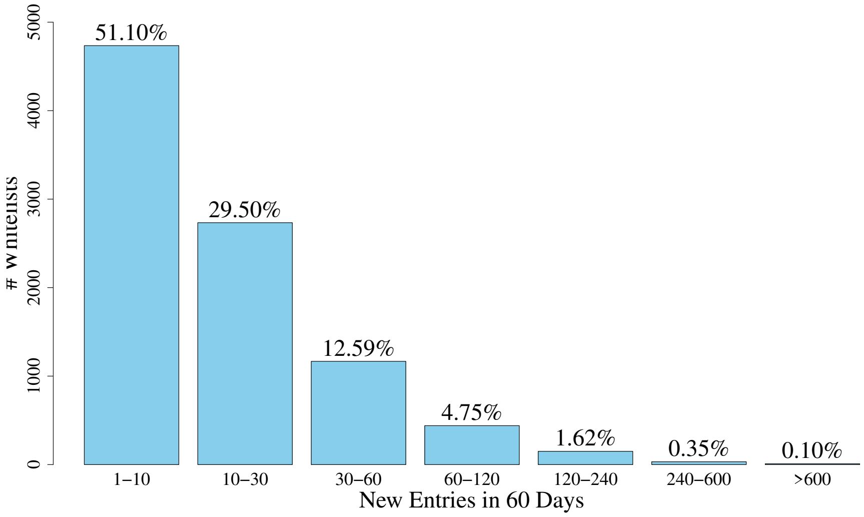


Figure 3.9: Distribution of the number of changes in users' whitelist

3.4.3 Whitelists' Change Rate

We conclude this section on the user point of view with an analysis of the rate at which the users' whitelists change over time. For this experiment we monitored the number of changes in the users' whitelists for a period of two months. Email addresses can be added to a whitelist in four different ways, two manual and two automated. A user can manually import a new entry or he can whitelist a certain address from the digest. Automatically, new entries are included when the user sends a mail to those addresses or when the senders solve the challenges generated by the CR system.

During the monitored period, 9267 whitelists were modified at least once. Out of them, only 6.8% had at least 1 new entry per day (on average), and the percentage drops even further when we look at higher changing rates (2.1% of the whitelists had at least 2 new entries per day, and 0.2% at least 5). Figure 3.9 presents a more detailed histogram of the frequency of changes. The graph shows how the large majority of the whitelists are, in fact, constantly in a steady state.

Finally, we monitored the amount of new messages present in the daily digest. This value varies greatly between users and also between different days. Figure 3.10 shows examples extracted from three different users. Some of them have constantly a large number of messages in the gray spool, while others have normally very small daily digests with anomalous peaks in conjunction to particular user behavior or unusually large amount of spam messages.

Again, a large size of the digest is at the same time a good and a bad factor for a CR system. In fact, a high number of messages means that the system is blocking

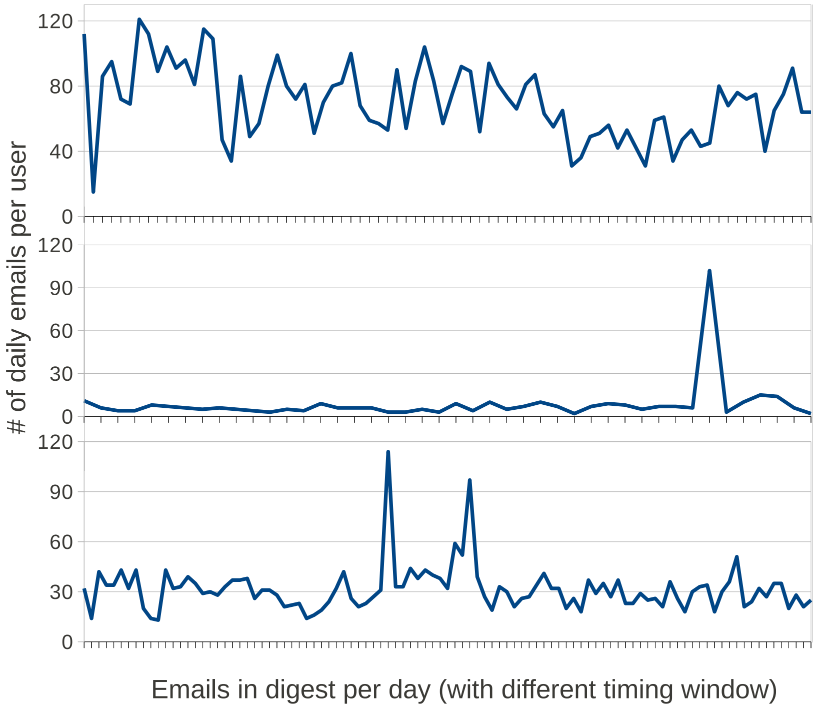


Figure 3.10: Daily pending email distribution of 3 different users

a substantial amount of spam that would have been otherwise delivered to the user (remember that these are messages that successfully pass through the antivirus, reverse DNS, and the SpamHouse blacklist). On the other side, a large digest is also a negative factor as it increases the amount of work for the user that has to manually verify its content to double-check that he did not miss any important message. Finally, this also demonstrates that the degree to which CR system works depends a lot on the interplay of users' involvement. Some recipients may diligently weed out their digest, while others may let it grow hoping for the senders to respond to the challenges.

3.5 The Administrator Point of View

In this section we analyze some of the issues related to maintaining challenge-response systems from the system administrator point of view. In particular, we focus on the effort required to maintain the challenge-response infrastructure, and on the additional antispam techniques that can be integrated in the system to reduce the backscattering effect.

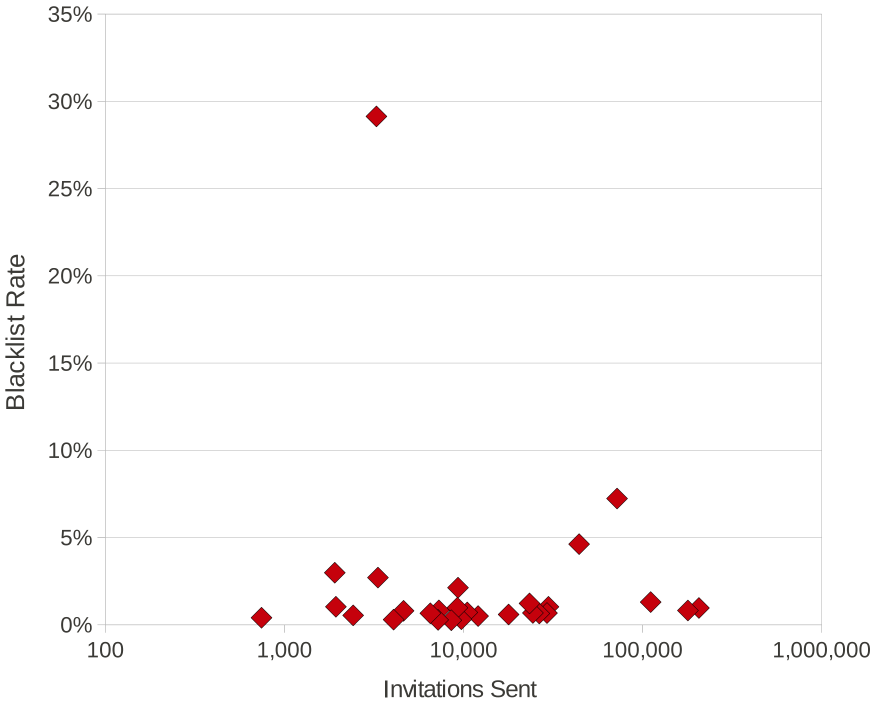


Figure 3.11: Server blacklisting rate

3.5.1 Server Blacklisting

As we already described in Section 3.4, when a CR system sends a challenge in response to a message with a spoofed sender's address, the challenge is delivered to a recipient that may not exist. As a result, these challenge-response messages can hit a spam trap [138], i.e., a set of email addresses that are maintained and distributed with the sole purpose to lure spam.

The emails collected by those traps are often adopted by popular services (e.g., SpamHaus [26], SORBS [19], SpamCop [22]) to update their blacklists. Hence, the IP used to send the challenges can itself get blacklisted as a result of the backscattered spam it sends. In order to reduce the impact of being blacklisted, one third of the systems we tested in our experiments were configured to rely on two MTA-OUT servers (with different IP addresses): one to send the challenges and another to send the outgoing user emails.

Our initial hypothesis was that the probability that a server has to get blacklisted should have been somehow proportional to the size of the email server, represented either by the number of users, or by the number of the received emails. In other words, we expected that systems sending more challenges were blacklisted more often, thus making CR solutions more difficult to maintain for large companies.

Surprisingly, our experiments proved us wrong. Using the data we collected we were able to estimate the rate at which various challenge server IPs get blacklisted. In particular, we followed two parallel approaches. In the first, we analyzed one month of data for 32 companies, measuring the ratio between the number of challenges sent and the number of blacklist-related error messages received from the challenge-response recipients. The result, summarized on a logarithmic scale in Figure 3.11, shows that while most of the servers had no problem at all with blacklisting, some of them were often blacklisted, even for a few days in a row. However, there seems to be no relationship between the server blacklisting ratio and the number of challenges it sends.

The main problem with this approach is that the error messages received when delivering the challenges were not always very accurate, providing results that may not be completely reliable. Therefore, we decided to complement our analysis with a second technique, based on an automated script that periodically checked for the IP addresses of the CR servers in a number of services that provide an IP blacklist for spam filtering. In particular, our tool queried the blacklists provided by Barracuda [3], SpamCop [22], SpamHause [26], Cannibal [21], Orbit [12], SORBS [19], CBL [6], and Surriel [13]. The queries were performed every 4 hours for a period of 132 days (between September 2010 and January 2011).

The results of this second analysis confirm our previous finding. In more than four months, 75% of the servers never appeared in any blacklists. Few servers were blacklisted for less than one day, while the remaining four servers experienced some serious problems, appearing in at least one of the blacklists for many consecutive days (17, 33, 113, and 129 respectively). Again, between the top 3 server (according to the traffic and the number of challenges sent) none appeared in any of the blacklists during our experiment. Thus, proving again that there is no direct link between the number of times a server gets blacklisted and the server size.

3.5.2 Combining CR Systems with Other Spam Filters

Our final evaluation focuses on the combination of CR systems with other antispam solutions. As we already mentioned in Section 3.2, the product we analyzed in our experiments includes three other spam filters in order to reduce the number of useless challenges sent in response to spam messages. It employs a traditional antivirus to scan the emails, an IP blacklist provided by SpamHause [26] to filter out known spammers, and a reverse DNS lookup to exclude suspicious origins.

According to Table 4.1 and Figure 3.1, the combination of these filters was responsible for dropping 77.5% of the messages in the gray spool. One may argue if this is good enough, or if a much better result could be obtained by introducing other antispam techniques. This is a difficult question to answer, since the main drawback of adding new filters is that they also introduce false positives, to avoid which CR systems were introduced in the first place.

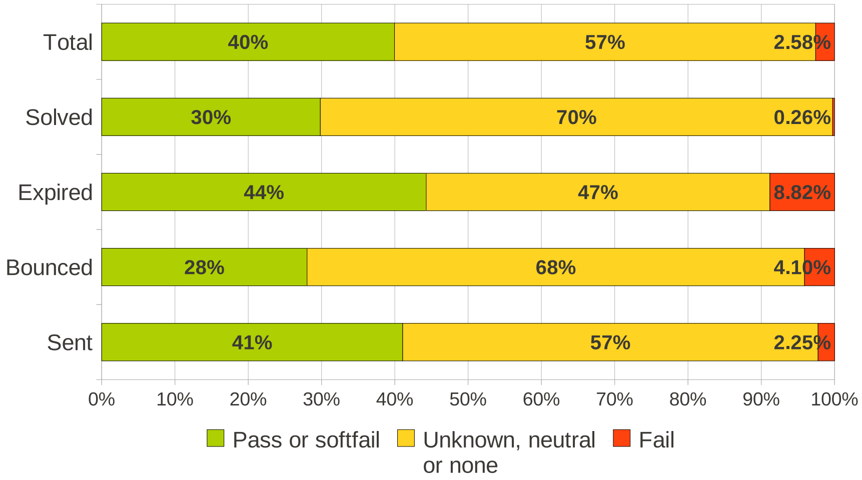


Figure 3.12: SPF validation test

However, we decided to experiment with one additional spam filter based on the verification of the Sender Policy Frawework [167] (SPF). SPF was introduced to detect source address spoofing, that is one of the main problems of CR systems. Since SPF checks were not included in the product we evaluated in this paper, we decided to evaluate the potential impact of this filter by using an offline tool to automatically test all the emails in the gray spool. Figure 3.12 shows the results of our experiment, grouped by different message categories. For instance, by dropping the emails for which the SPF test fails, it would be possible to reduce by almost 9% the challenges that cannot be delivered (`expired`), and 4.10% of the bounced ones. The overall result shows that SPF can further reduce the number of "bad" challenges by 2.5%, at the cost of loosing 0.25% of the challenges that are actually solved by the sender.

3.6 Discussion

Even though the aim of this work is neither to attack nor to defend challenge-response systems, it may be natural to ask what conclusions about this class of antispam solutions could be drawn from our measurements.

In the rest of this section we summarize the main findings we presented in the previous three sections.

Whitelist Assumptions

All approaches based on white-lists share two main assumptions: first, that the large majority of the "good" emails come from senders already in the recipient's whitelist, and, second, that these lists eventually reach a steady state where changes do not occur very often.

Both claims are supported by our experiments. In fact, over 43 companies, only 2/33 = 6.1% of the incoming emails delivered to the users' INBOX require a challenge-response phase (see Figure 3.4) and 2% require the user to manually pick the message from the digest.

The stability of the whitelists was already evaluated by Erickson et al. [71], showing that the major burden on the user is concentrated in the first three weeks, after which the number of changes drops on average to one per day. Our experiments show that, in a real deployment, there are on average 0.3 new entry per user per day (excluding new users). Only 6.8% of the users had at least one daily change in their whitelists.

Delivery Delay

Another common critique of CR systems is due to the fact that the challenge-response step introduces a delay in the incoming email delivery. This is obviously an unavoidable side-effect, but our measurements show that it also has a limited impact. In fact, according to our study, it concerns only 4.3% of incoming emails and in half of the cases the delay is below 30 minutes. Even though the remaining 2.15% may still be an unacceptable inconvenient for certain users, we believe that for most of the users it would be a reasonable price to pay to protect against spam.

Challenge Traffic

Most of the criticisms against CR systems, and most of the hassles for the system administrators, come from the challenges that are sent to the wrong recipients. If they correspond to existing email accounts, the misdirected challenges become a spam for other users. On the other hand, if the addresses do not exist, the challange may hit a spamtrap. And on top of that, they constitute useless traffic over the Internet.

Our study shows that, on average, a CR system sends back one challenge for every 21 emails it receives (see Section 3.2), accounting for a traffic increase of less than 1%. These figures depend on the amount of spam received by the server, and seems to be more or less constant between small and large servers.

Unfortunately, the large part of the challenges sent are indeed useless (only about 5% of them are solved). But, as we already explained in the paper, these challenges are "required" to justify the system. In other words, without useless challenges, it would be the CR system to be useless. Therefore, this can be considered an intrinsic and unavoidable limitation of systems based on a challenge-response approach.

Our findings confirm that the backscattered phenomenon is the main problem of solutions based on challenge-response technique. Each installation must be carefully configured in order to minimize the impact of misdirected challenges on other real users. The administrator also has to decide which other additional antispam techniques should be combined with the CR filter to maximize the benefits and, at the same time, to reduce the side effects and the risk of having the servers' IP blacklisted. However, the backscattered phenomenon is intrinsic in the behavior of a CR system and cannot be completely eliminated. From a company, the single most negative argument against the adoption of CR system is the fact that the challenge server can occasionally get blacklisted. Even worse, an attacker could intentionally forge malicious messages with the goal of forcing the server to send back the challenge to spam trap addresses, thus increasing the likelihood of getting the server IP added to one or more blacklist.

Other Limitations

This paper does not cover all aspects related to the adoption of a challenge-response system. We focused on performing a set of measurements based on real installations that were not under our direct control. Therefore, we intentionally excluded from our studies any evaluation of potential attacks against CR systems (like trying to spoof the sender address using a likely-whitelisted address).

In addition, in order to protect the users and the companies' privacy, we limited our study to the statistical information that can be automatically extracted from the headers of the messages. This ruled out other potentially interesting experiments that would have required access to the email bodies.

3.7 Conclusions

In this Chapter we presented the first measurement study of the behavior of a real world deployment of a challenge-response antispam system. The experiments lasted for a period of six months, covering 47 different companies protected by a commercial CR solution.

In particular, we *measured* the amount of challenges generated by these systems and their *impact* in terms of traffic pollution and possible backscattered messages delivered to innocent users. We then measured the amount of emails that are delayed due to the quarantine phase, and the amount of spam that is able to pass through the filter and reach the users mailboxes. Finally, we focused on a problem that is less known, i.e., the fact that the invitations sent by these systems can accidentally hit a spamtrap and cause the email server to be blacklisted.

Our findings can be used to evaluate both the effectiveness and the impact of adopting this class of techniques, and figures provided in this Chapter may help to settle the long debate between advocates and opponents of CR systems.

At the same time, we also look at the quarantined emails that already exclude the obvious spam and ham messages. In fact, these particular emails represent good approximation of a *gray area* that by definition contains messages difficult to automatically classify by the spam filters. In this Chapter, we even show that by grouping similar messages in this area we are able to identify two major classes of messages: ones with rather stable email headers and sending patterns, and others with more dynamic characteristics. This suggests that the gray area consists of a portion of very similar emails that exhibit patterns that resemble spam sent by botnet, but also includes other types of bulk email campaigns as newsletters, notification or commercial emails. The following leads us to the next researched question of the thesis – the analysis of the *gray area* and its *email campaigns*.

4
Automated Analysis of the Email Gray Area

In the previous Chapter, we showed that a Challenge-Response anti-spam system works and performs differently than other anti-spam system, having its advantages and side effects. One of the side effects is that around 30% of all incoming emails get quarantined, because they cannot be attributed by the system to any class. This particular side effect at the same time provides us a unique vantage point for building an approximate *gray area* email dataset as most of the standard email filtering has been already applied to these messages.

Intuitively, this area consists of spam emails that overpassed several existing anti-spam protection mechanism, but also from other bulk emails like subscribed newsletters, notification emails or commercial advertisements. Interestingly, the amount of the latter type is dense in this specific area as one of the effects of a CR system is that it authenticates human senders, but not automated legitimate bulk senders.

In this Chapter, we study the gray area of a CR system by adopting a three-phase approach that relies only on the information available in the email headers. The proposed method identifies email campaigns and categorizes them into campaign categories without any content analysis. We demonstrate that the gray area can be reduced by at least 50% and that identified campaigns can be actually automatically categorized into four different types: commercial, newsletters, botnet spam, and phishing/scam.

4.1 Introduction

While most of the existing research deals with the problem of efficiently and accurately distinguishing spam from ham, in this Chapter we focus on the thin line that separates the two categories. We limit our study to the often overlooked area of *gray Emails* [172], i.e., those ambiguous messages that cannot be clearly categorized one way or the other by automated spam filters.

We start our study by analyzing a real world deployment of a challenge-response antispam solution to measure the extent of this gray area and user behavior within the area. According to our data, after the obvious spam and legitimate emails have been eliminated, users still manually check on average five to six messages per day. This area is particularly dense with automated legitimate bulk messages, e.g. newsletters, notifications, etc., and also with illegal unsolicited messages, mostly distributed by botnets. Hence, users are constantly prompted to search over their quarantined emails that are mixed with spam emails, thus increasing their exposure to different cyber threats (the same is true to other anti-spam systems having a spam folder, although to a lesser extent). Our data shows that on average 1.5% of gray messages have an attachment with 9% of them being malicious. However, some of these messages also contain interesting content, as proved by the fact that users read and whitelist an average of 1.5 messages per day. We also confirm the belief that normal users are not very good in telling spam and ham apart, and they often intentionally open emails with malicious attachments in the gray area.

To analyze the gray area, we group emails into campaigns by adopting a three-phase approach that uses clustering, classification, and graph-based refinement. As a result, all the campaigns get a score that decides as to which class they most probably belong to, providing a ranking of campaigns. In previous work, other researchers relied on user feedback when studying these area, during our analysis we concluded that users tend to open or even whitelist spam emails. Thus, in our study we consider user generated data are unreliable to be used as a ground-truth.

Our *per-campaign* analysis method permits us to avoid analyzing *unique* personal user emails and to focus rather on *bulk* emails without analyzing their content, only their sheader information. Our technique is able to automatically classify up to 50% of the gray emails – reducing the gray area by half – with only 0.2% of false positives. The identified campaigns consist of illegal campaigns (spam) sent often with bad intentions, and of automated legal bulk campaigns, to which the recipient has mostly probably subscribed to.[1] We further demonstrate that there are at least four identifiable categories of email campaigns – commercial, newsletters, botnet spam, and scam/phishing, where a commercial campaigns constitutes a large fraction of the gray area and can be identified by using our graph-based refinement method. To the best of our knowledge, this is the first real-world empirical study of such emails.

Additionally, such a system could be used as a tool for monitoring the evolution of email campaigns within different categories. Moreover, as we show in this Chapter 4,

[1] Note that this is almost impossible to verify, therefore we assume that legitimate campaigns were solicited at some previous point in time by the recipient.

Table 4.1: General statistics

Mail servers	13
Active users	10,025
Total messages	11,203,905
White emails	2,806,415
Black emails	5,066,141
Gray emails	3,331,349
Challenges solved	166,279
Users whitelisted emails	42,384
Users viewed emails	104,273

the sending patterns of legal bulk senders differ from illegal campaign senders and in principle they are more static. This suggests that we could leverage sender level information of legal senders in order to create a whitelist of legal content senders, or even for identifying marketing-management companies, like MailChimp [10] or other – email campaign management and tracking services.

4.2 Data and Analysis methods

This section presents the dataset we used in our experiments and the techniques we adopted to process and analyze the email messages. Since it would be impossible to classify each email in isolation, we adopted a multi-layered approach to group them into similar campaigns (a solution proved to be effective by several previous studies [172, 136, 130]). In particular, we start by clustering them based on the message headers. We then extract a set of features based on a number of campaign attributes and we use them to train a classifier in order to predict the campaign class. Finally, we employ a graph-based refinement technique to further increase the coverage and precision of our classification. The rest of this section introduces each phase in detail.

4.2.1 Data collection

The amount and diversity of the available data is crucial in order to successfully identify email campaigns. Messages should be collected from multiple feeds, cover numerous recipients, organizations, and long periods of time [130, 133]. Our email dataset fulfills these requirements as it was collected from a commercial Challenge-Response (CR) spam system deployed in tens of different organizations. A CR filter is a software that automatically replies with a challenge (in our case a CAPTCHA) to any previously-unknown sender of incoming emails. If the sender solves the challenge, the message is delivered to the recipient; if not, it remains in a quarantined folder. Since in our study we want to focus on the borderline area that contains the emails that cannot be easily classified as legitimate or spam, we installed a sensor in the CR system to intercept any quarantined message. These emails have successfully passed through a number of antispam filters but were neither already whitelisted nor blacklisted by the recipient. In other words, these messages were not considered spam according to traditional techniques like: virus scan, reverse DNS, and DNS blacklisting. Moreover, users never had any previous conversation with the sender. Therefore, we can consider this dataset as pre-filtered from obvious ham and spam emails. Sometimes such set is referred to as a *gray zone* [50] that stores emails of uncertain class. Email categories often found in this pool include traditional spam, scam, notifications, newsletters, and commercial offers. Since this set may include also notification or personal messages, users check them manually when they look for missing messages.

We also instrumented the CR-system to collect additional information (see Table 4.1): which emails were opened by the users, and which messages where whitelisted (thus showing that the user manually classified them as legitimate). This can provide some insights on how capable users are at distinguishing harmless from harmful emails. Finally, our sensor collected the delivery status information (sent/bounced/delivered) for each challenge email sent back by the CR system.

In our experiments we relied on statistical email data that we collected from 13 companies of different sizes. The monitoring period covered 6 months, from August 2011 to January 2012. During this period around 11 million messages were delivered to the monitored mail servers (Table 4.1). The data we used in our experiments did not include the email bodies, and the headers were sanitized to protect the privacy of both the users and the companies involved.

4.2.2 Email Clustering

The task of grouping emails into campaigns has already been covered by several previous studies ([103, 132, 112, 136, 130]). Previous results were very successful in identifying email campaigns, but, unfortunately, often relied on the content of the email body. Our dataset is limited to the email headers, thus forcing us to use a different approach based only on the email subjects. The main limitation of this technique is that the email subjects have to be long enough to minimize the chances of matching different messages by coincidence.

The obvious solution for grouping similar subjects would be to apply some text mining algorithm, but our input text is short and it is important to preserve the word order. Hence, we decided to use a simple approach based on “almost exact” text matching, extended to include subjects with a variable part. The latter could be a varying phrase in the subject, a random word/id, or a user name. We use word n-grams of a decreasing length (between 70 and 8), with a sliding window

Table 4.2: Clustering statistics

Total emails	3,331,349	100%
Clusters	12,250	
Emails clustered	1,670,521	50%
- With n-grams	690,600	41%
- With exact match	979,921	59%

that permits to skip over varying parts of the subjects. Our implementation is based on an existing n-grams extraction library (Ngram Statistics Package [39]), a standard list of stop-words, and a number of custom scripts to match the extracted n-grams and assign them to clusters.

The process starts by searching for the longest n-gram (70) and then decreasing the length until enough similar matches (30 emails per cluster) were found to create a cluster. This algorithm is efficient on long subjects but it is problematic on short ones, thus limiting our analysis to subjects containing at least 10 characters and 3 words.

The results are presented in Table 4.2. In this phase we successfully clustered 50% of all emails in 12,250 clusters. Cluster sizes varied between 30 and 8,468 messages.

4.2.3 Feature-based Classification

To be able to differentiate and classify the identified clusters, we extract a set of eleven features grouped in three categories (see Table 4.3).

Group A: Features in this group reflect the similarity of a certain feature inside a cluster. The values are expressed as a range between 0 and 1, where 0 indicates high distribution (low data similarity) and 1 indicates low distribution (high data similarity) in the cluster.
The feature similarity is defined as:

$$a(C) = 1 - u/t$$

where u is the number of unique or similar feature values, and t is the number of total emails. This group contains four features measuring the similarity of sender IP prefixes and email addresses, and the similarity of the sender names. In particular, we split the email domain address into two parts: the *email prefix* and the *email suffix*. The suffixes are grouped by removing numerical differences (e.g., `abc10.com` and `abc22.com`). When similar suffixes are found, they are merged until there are no similar values left. *Email prefixes* are instead compared using a variation of the Levenshtein distance algorithm in which a threshold is computed based on the length of the email prefix itself. In this way, the similarity score is normalized to

Table 4.3: Cluster features

Group A	
Sender IPs	Distribution of network prefixes (/24)
Sender names	Distribution of email sender names
Domain of sender address	Distribution of email domain names
Prefix of sender address	Distribution of email prefixes

Group B	
Rejections	Percentage of rejected emails at MTA
White emails	Percentage of whitelisted emails
Challenges bounced	Percentage of bounced challenges
CAPTCHAs solved	Percentage of solved challenges
Unsubscribe header	Percentage of Unsubscribe headers

Group C	
Recipients per email	Normalized number of unique recipients per email
Recipient's header	Location of recipient's email: To/Cc/Bcc/Mixed
Countries	Distribution of countries based on originating IPs

account for the fact that, for example, a two-chars difference for short strings is somehow equivalent to a six-chars difference for longer ones.

Group B: Features of this group reflect the percentage of messages in a cluster that have a certain feature value. There are five features in this group: *CAPTCHA solved*, *rejections*, *white emails*, *challenges bounced*, and *unsubscribe header*. The first measures the percentage of challenges that were solved by the senders. The *challenges bounced* are instead undelivered emails as the recipient was non-existent, or did not accept email for the recipient. Whenever an email was sent to multiple recipients, we were also able to compute the percentage of *white emails* (i.e., the percentage of recipient that had already whitelisted the sender) and the percentage of incoming email *rejections* (i.e., the percentage of recipients that were rejected by the Mail Transfer Agent - normally because the corresponding addresses did not exist on the server). Finally, the *unsubscribe header* feature evaluates the percentage of emails that contained the unsubscribe header. The latter is generally used by commercial messages and notifications providing the users an option to unsubscribe from the list.

Group C: Features in this groups are computed in different ways. *Recipients per email* estimates the average number of recipients per email. The *Recipient's header* feature indicates the location of email recipient address in the email headers: *To*, *Cc*, *Bcc*, or *Mixed* when multiple locations are used in the same campaign. The *countries* feature reflects the number of countries (based on the sender IP geolocation) in the cluster.

Manual Labeling

Before performing our classification, we need to build a training set. For this reason, we randomly select 2,000 campaigns (16% of the total number of clusters) and we performed a manual classification of their messages. As a result, we identified 1,581 (79%) legitimate and 419 (21%) spam campaigns. These preliminary classification confirms that the majority of spam was already filtered out of our gray dataset.

Obviously, the result of our manual labeling process depends on the actual definition of spam that we adopt in our experiments. By definition, spam is an unsolicited email, usually sent in bulk. However, there is no way to verify if a certain email is solicited (i.e., if the recipient is subscribed to it or not). Moreover, the notion of spam is somehow subjective and it may not be the same for all the users. Most of the commercial campaigns are probably unsolicited, and therefore could be considered as spam. However, when such emails are sent by professional marketing companies according to the original country regulations, it is unclear if they should be considered legitimate or not.

In this Chapter we take a conservative approach, and flag as spam only potentially dangerous or illegal campaigns that may involve malicious, fraudulent or illegal online activities. This includes different “business models”: some emails sell illegal products, others are used to spread malware or targeted attacks, some aim at stealing personal data and credentials, and some specialize in advanced fee fraud (e.g., *419 scam*). Finally, we consider any email belonging to a commercial marketing campaign as legitimate (in the sense that general antispam filters should not block them, unless they are specifically instructed to do so by the user).

The process of manual labeling of gray email campaigns consists of a manual analysis of aggregated header information about the email campaigns. All the campaign features are only used and viewed in an aggregated form, thus never accessing any distinct email header information. A particular case is *email subject* that is a textual information and is difficult to aggregate. However, as we group emails based on subject similarity, we also keep an aggregated copy of the subject from the campaign. However, when manually labeling campaigns, we were unable to know who sent and received the emails.

Distinguishing between different classes of emails, even with the full email content, sometimes might prove to be difficult. But by looking at campaigns instead of singular messages, we access additional information (e.g., average number of recipient per email, number of originating countries, etc.) that is unavailable to the viewer when viewing only one message at a time. In the case when the subject cannot provide enough information to make a decision, aggregated email header information is used by the analyst to decide. For example, if a message comes with rather personal subject, knowing that it is a campaign of at least 30 messages helps to conclude that it is not what it looks like; or a message promoting a new product

Table 4.4: Campaign classification results

Campaign type	Manual sampling	In %	Unlabeled	In %
Legitimate	1,581	79%	8,398	81.9%
Spam	419	21%	1,852	18.1%
Total	2,000		10,250	

or services online, however, being sent in thousands of email copies, from over 30 different countries and with multiple recipient per email is also a good support for making a conclusion.

Classification

Using the eleven features presented above, we trained a binary classifier to separate the legitimate clusters from the spam ones. To select a classifier we started from the results presented by Kiran et al. [102], in which the authors demonstrated that, on spam datasets, ensemble classifiers perform better than single classifiers. Based on this conclusion, for our classification task we decided to use a supervised Random Forest ensemble classifier.

We first performed a cross validation test in which we randomly split the sampled data into two groups including respectively 70% and 30% of the data. We then trained the Random Forest classifier (configured with 500 trees and three random variables per split) on the first group, and we tested the extracted model on the second one. For each cluster, the algorithm returns a score ranging between -1 (for spam) and 1 (for legitimate). A score close to zero indicates that the classifier was uncertain about the sample.

Our model achieved an accuracy rates of 97%, with 0.9% false positives (i.e., legitimate campaigns being misclassified as spam) and 10% false negatives (i.e., spam being misclassified as legitimate). These rates suggest that the set of attributes we identified are effective in separating the two types of campaigns. We also noticed that while our classifier identified well legitimate campaigns, it had a higher probability to misclassify spam campaigns. A further interpretation of this phenomenon is described in Section 4.5.

Finally, we applied the model extracted from our training set to predict the classification of the remaining unlabeled campaigns. Results are presented in Table 4.4. 10,002 (82%) of the campaigns are labeled as legitimate and 2,248 (18%) as spam.

Table 4.5: Attribute values per campaign category

Attribute	Legitimate	Spam	Grey
		Min / Mean / Max	
Countries	1 - 1.2 - 6	7 - 29 - 123	1 - 5 - 80
IPs	0.13 - 0.9 - 1	0 - 0.06 - 0.82	0 - 0.7 - 1
Sender email domain	0.2 - 0.98 - 1	0 - 0.3 - 1	0 - 0.85 - 1
Sender email prefix	0.03 - 0.98 - 1	0 - 0.09 - 1	0 - 0.81 - 1
Senders	0 - 0.98 - 1	0 - 0.3 - 1	0 - 0.8 - 1
Unsubscribe header	0 - 0.5 - 1	0 - 0 - 0.3	0 - 0.3 - 1
Bounced	0 - 0 - 1	0 - 0.1 - 1	0 - 0.1 - 0.9
CAPTCHAs	0 - 0 - 1	0 - 0 - 1	0 - 0.1 - 1
White emails	0	0	0.001
Rejections	0 - 0 - 0.4	0 - 0.23 - 1	0 - 0.1 - 0.7
Recipient per email	1 - 1 - 1.1	1 - 3 - 16	1 - 1.1 - 8
		To, *Bcc*, *Mixed* shares	
Recipient header	0.76 - 0.04 - 0.2	0.3 - 0.1 - 0.6	0.4 - 0.33 - 0.3

4.2.4 Graph-based Refinement

Although we achieved a relatively high accuracy using our classifier, we still found that for some campaigns our algorithm gave uncertain results. Luckily, the vast majority of the campaigns are located at the extremes of the classifier scores, either close to 1 (legitimate), or to -1 (spam). Campaigns become much more scarce in the range between [-0.8..0.8]. This *gray area* inside the gray area represents those cases for which we were unable to assign a definitive category.

Using these two thresholds, we can refine our classification and split the data into three classes: legitimate (77% of the total campaigns), spam (16%), and gray (6.4%). The min, average, and maximum values for each attributes in the three classes are summarized in Table 4.5. Since most of the false positives and false negatives are located in the gray area, we focused on improving the classification of those messages by using a graph-based technique.

In particular, we built a graph in which nodes represent campaigns and edges model the fact that two campaigns share a combination of sender IP address and email domain name. These links created networks of campaigns sent from the same mailing infrastructure. To avoid false connections that might appear between campaigns when they use webmail providers (spoofed or not), we removed those links from the graph.

The resulting graph contained 9,891 connected campaigns and 608 isolated subgraphs. By visually looking at the subgraphs (which is omitted due to space limitations) we noticed that the majority consisted of a predominant class (either spam

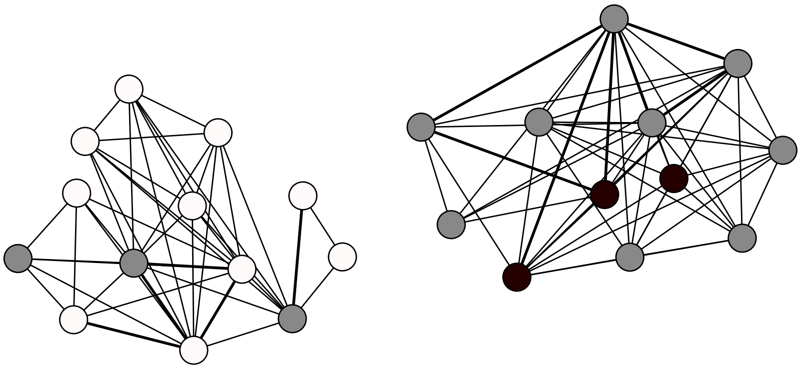


Figure 4.1: Subgraphs with mixed campaign classes: white for legitimate, gray for gray, black for spam

or legitimate nodes) sometimes intermixed with gray nodes (see an example in Figure 4.1). This seems to suggest that gray campaigns also belong to the same class as the other nodes in the same group, since they are sent using the same infrastructure.

While this approach works well for the small subgraphs, the graph also contains a Giant Component – a graph linking together 52% of all the campaigns – for which it is impossible to decide to which class it belongs to. Therefore, we apply a community finding algorithm that groups all the nodes into interconnected communities, also called groups, decomposing the Giant Component into smaller parts. We end up with 660 groups, for most of which we can accurately associate a single class. When gray campaigns are in the same group with any other class, we assign gray campaigns to the class of its group.

While this technique works well for most of the groups, some noise is still introduced in the results by the presence of loosely connected nodes. These are nodes that get erroneously connected to a group due to emails reusing the subjects of legitimate campaigns. To remove these connections, for each node we compute a graph metric called *clustering coefficient*. The coefficient for loosely connected nodes is equal to 0, whereas it approaches 1 for tightly connected nodes. As a result, we re-classify all the gray nodes with a clustering coefficient greater than zero and that belong to a group of either legitimate or spam campaigns. To decide on the class of the group we compute the mean of classifier score of all nodes in the group: groups above 0.2 are considered legitimate, and groups below this threshold are considered spam.

Using this approach we were able to properly re-classify over half of the gray campaigns (427). This reduced the false positive rate from 0.9% to 0.2% (see Table 4.6 for more information). The entire dataset is now split in ham (80%), spam (17%) and gray (2.9%) messages (an increase of 3% for legitimate campaigns and 1% for spam). Again, our method performs better with legitimate messages. This is due to legitimate campaigns forming stronger networks (reusing the same mailing infrastructure over time) than malicious campaign.

Table 4.6: Refining the campaign classification using graph analysis. Classification errors evaluated on 2,000 sampled campaigns

	RandomForest	Graph analysis
False Positives	0.9%	0.2%
False Negatives	8.6%	7.6%
Grey area	6.4%	2.9%

4.3 Attributes Role in Email Classification

In this section we analyze the characteristics of spam and legitimate campaigns and compare our findings to the ones presented in previous studies [130, 136] that also analyze spam campaigns.

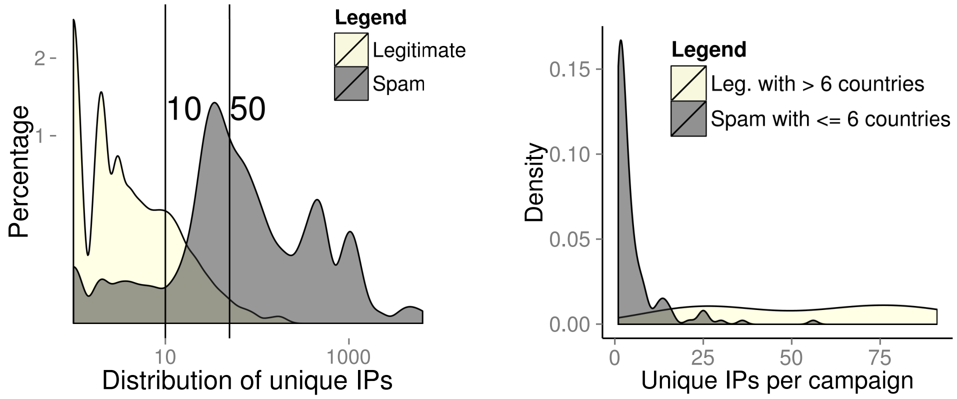
The Random Forest classifier provides some information about the relevance of each feature. Interestingly, the least important attributes are the ones in Group B, and in particular the percentage of already whitelisted emails in the cluster. The most important ones are the distributions of country and IP addresses, followed by the average number of recipients, and the sender email address similarity. The latter proved to be useful because spammers often change sender emails, while legitimate campaigns use a single or several recognizable patterns.

In particular, we found the number of originating countries to be the most indicative parameter, whereas previous research often relied on the IP address distribution (e.g. [136]).

4.3.1 The role of IPs and Geolocation

IP address variation is often regarded as a strong indicator of botnet activity and often used as a reliable metric to detect spam. However it is unclear what should be adopted as a threshold for this metric, how many different IPs should alert us of a distributed malicious activity, or how accurately we can classify email campaigns simply by looking at their IP address distribution.

In a previous study of spam campaigns, Qian et al. [136] used a threshold of 10 IPs per campaign to separate spam campaigns from legitimate. To evaluate this threshold, we apply it on our gray dataset as shown in Figure 4.2(a). The graph plots the distribution of unique IP prefixes for both spam and legitimate campaigns. Around 90% of the legitimate campaigns are indeed below the 10 IP threshold, while 90% of the spam is above - resulting in a global error rate of 9.2% (to be precise, our measure is based on /24 subnetworks and not on single IP addresses, and therefore the real error rate is much higher than 9.2%). In comparison, this error is 5 times higher than the one of our classifier.



(a) Logarithmic plot of unique IP prefix distribution

(b) Distribution of campaigns after applying 6 countries threshold on our data

Figure 4.2: IP prefix and countries distribution in campaigns

By looking at Figure 4.2 (a), we notice that above 50 IP prefixes there are few legitimate campaigns left and 99.8% of legitimate campaigns are below this threshold. However, half of the spam campaigns are located above the threshold and another half in between two thresholds (10-50). This suggest that there is not a single value that separates the two classes with a low error rate.

By looking at the IP country distribution the results improve considerably. Some legitimate campaigns have many IP prefixes, but originate from few countries. This could be the result of having the same commercial campaign being propagated by several email marketing companies. In contrast, the vast majority of spam campaigns originate from multiple IP prefixes *and* multiple countries. In fact, by using a six-countries threshold (the one used by our classifier) we misclassify only 0.4% of the legitimate and 12% of the spam campaigns - resulting in a total error rate of 2.8%. Figure 4.2(b) shows classification error results, where it is evident that mostly spam campaigns with *few IP origins* would be misclassified.

Finally, we investigate closer this group of spam campaigns with few origins. Interestingly, the classifier for most of them gave a weak score between 0 and -0.5. The graph refinement was ineffective for them, because these campaigns did not appear at all in our graph. At a closer look, these cases mainly corresponded to phishing and Nigerian scams. Several of these campaigns are sent in low volume and for short periods of time using webmail accounts, thus hiding under benign IP addresses.

Table 4.7: *To/Bcc/Mixed* recipient header distribution

	To	*Bcc*	*Mixed*
Legitimate	75%	5%	20%
Spam	30%	12%	58%
Grey	20%	53%	27%

4.3.2 Recipient-oriented attributes

The email recipient can be specified in three different headers: *To*, *Cc*, and *Bcc*. Interestingly, we found no campaigns using the *Cc* header, and some campaigns that seem to randomly change the location of the recipient over time (we categorize them as *Mixed*). We also looked at the number of recipients per incoming email and at the number of non-existing email accounts (rejected at MTA-in because of non-existent user) in a multiple recipient emails. In this section we look at these three features together, as they are often more informative when combined than when taken individually.

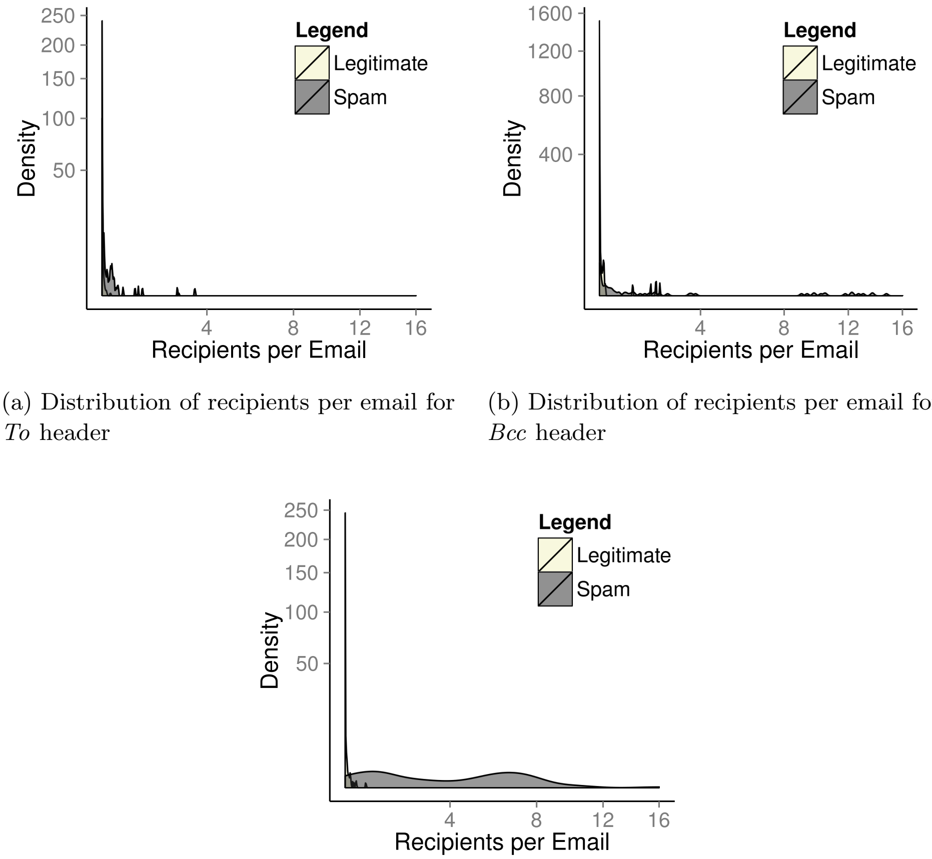
The header distribution within the campaigns is as follows: 3/4th of the legitimate campaigns use the *To* header (Table 4.7), whereas spammers often mix different headers in the same campaigns. The *Bcc* header is adopted by both campaign types, even though in a lower amount. However, it is very common between gray campaigns: in fact, half of them use exclusively this header to specify the recipient. Again, this is very common between the previously described scam campaigns.

Since the campaigns located in the gray zone often use the BCC field, they have shorter recipient lists including on average only 1.2 recipients per email. In contrast, 94% of legitimate campaigns have a single recipient (Figure 4.3 (b)), while spammers tend to include an average of at least three recipients per email.

However, these features alone cannot be used to reliably separate spam from legitimate messages. For example, 36% of spam campaigns used only one recipient per email, and in 30% of the cases specified the recipient in the *To* header (Figure 4.3 (a)). Interestingly, most of these campaigns have a high IP prefix and country distribution in these campaigns, thus we assume that they still originate from infected machines or botnets.

When some of the messages in a campaign are rejected, it is an indicator that the sender's recipients list was not properly verified or not up-to-date. Although sometimes users make typos while providing their email addresses, a higher rejection ratio along with multiple recipients is a good indicator of spammer activity, as shown on Figure 4.4.

In fact, only 1% of spam campaigns sent with two recipients per email have a rejection ratio lower than 0.1. Thus, the combination of these two characteristics perform relatively well to classify these campaigns.



(a) Distribution of recipients per email for *To* header

(b) Distribution of recipients per email for *Bcc* header

(c) Distribution of recipients per email for *Mixed* header

Figure 4.3: Recipient-oriented attributes in campaigns

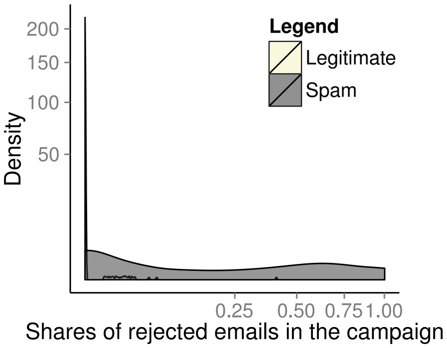


Figure 4.4: Emails rejections per campaign

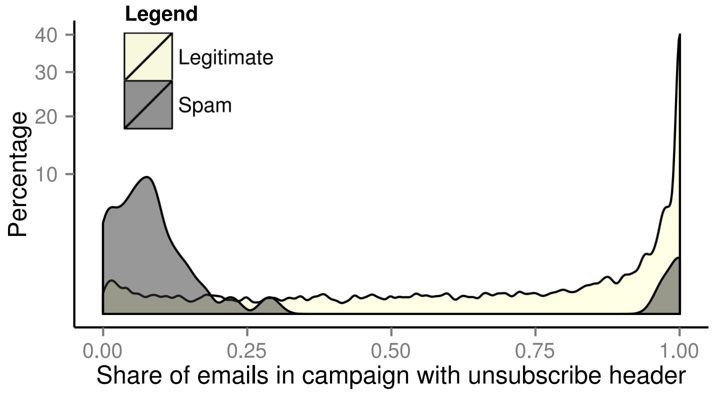


Figure 4.5: Newsletter subscription header distribution. Only the cases where the header is present are plotted

4.3.3 Newsletter subscription header

One of our features counts the presence of the *List-Unsubscribe* header in the emails. This header is used to point to a URL or email address that can be used to unsubscribe from a mailing list[2]. This header should be present in bulk emails sent regularly to a set of subscribed recipients. Another recommendation for bulk email is to use the *Precedence: bulk* header. However, since in our dataset this header was used only in few messages, we focus our study on the more common *List-Unsubscribe* header.

Figure 4.5 shows the percentage of each campaign type that uses the unsubscribe header. Only 10% of the spam campaigns adopt the header, counting only for a total of 0.6% of the spam messages. While legitimate campaigns tend to use the header in most of their emails, around half of the campaigns do not use it at all. This is due to several different email marketing companies advertising the same campaign, where some include the header, and some do not. In total, around half of the legitimate campaigns include the header (Table 4.8), and 27% of all legitimate campaigns have the header present in all messages.

In conclusion, we find it uncommon for spammers to use the *Unsubscribe* header, but at the same time legitimate campaigns use it in only half of their emails. While this attribute seems to be a good feature to identify marketing campaigns, spoofing the *Unsubscribe* header is extremely easy and it could be done in the future without adding any additional cost for spammers.

[2] In general an unsubscribe option is also included in the body of the message, but we could not check for this case since we had no access to email bodies.

Table 4.8: *Unsubscribe* header presence in campaigns

Campaigns	Header present	Missing header
Spam	225 (10%)	2,013 (90%)
Legitimate	5,064 (51%)	4,948 (49%)
Emails		
Spam	2,710 (0.6%)	482,133 (99%)
Legitimate	506,352 (43%)	668,153 (57%)

4.4 User Behavior

Our dataset also provides statistical information about which actions were performed by the users on the quarantined emails. In particular, we have information regarding the messages that were read, added to a user whitelist or blacklist, and the CAPTCHA that was later solved by the sender. These data can give us some useful insights on the ability of normal users to identify suspicious emails.

Table 4.9 presents three user action statistics. As expected, users activity involves mainly legitimate and gray campaigns. In fact, the main reason for users to go through the emails in this folder is to spot missed notifications or undelivered benign messages. However, a large fraction of users also opened spam messages, maybe attracted by some deceiving subjects. Even worse, around 6% of the spam campaigns had at least one of their messages manually whitelisted by some of the recipients. This action could be interpreted as the equivalent of clicking the “Not Spam” button provided by several webmail services.

Manually whitelisted emails include spam campaigns promoting drugs and pirated software. This may suggest two things: either users have problems in distinguishing legitimate emails from harmful, or that some users are genuinely interested in the products promoted by spammers. It is difficult to draw conclusions as both hypothesis might be true for different users, but, clearly, most of them are unaware of the security threats involved in opening malicious emails.

To measure how significant this phenomenon is, we compute that there is a 0.36% probability that a certain user whitelists a legitimate email and 0.0005% that she whitelists a spam message. These numbers may seem low, but they rapidly increase when multiplied by the number of users and the number of messages received. In total (see Figure 4.6) an average of 3.9 emails get whitelisted per legitimate campaign compared to 1.1 emails per spam campaign.

Figure 4.6 summarizes the number of user actions in each campaign, based on its classification score. Over 3,888 spam emails were opened by users during our six-month experiments, resulting in the fact that one out of five users has viewed at

Table 4.9: Campaign shares on which the actions were performed

	Viewed	Whitelisted	CAPTCHA solved
Legitimate	42%	12%	3.5%
Spam	25%	6%	0.2%
Grey	40%	17%	10%

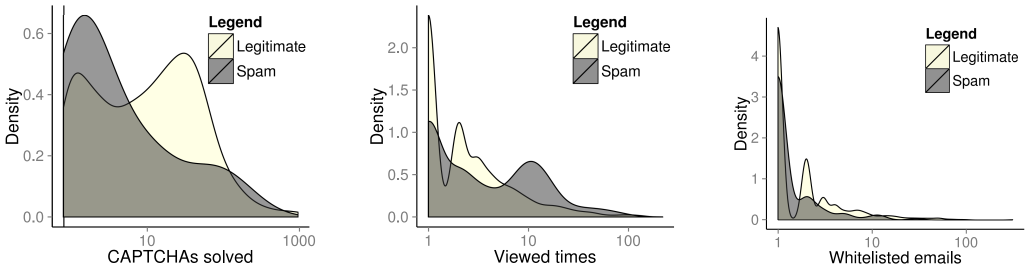
least *one spam message*, and, on average, opened 5 of them. Unfortunately, from our dataset we are unable to tell how many users downloaded attachments or followed links included in the message body.

The last question we want to answer is whether the fact that the sender solves some CAPTCHAs could be a good indicator to identify legitimate campaigns. Unfortunately, since most of the legitimate emails in the gray area are automatically generated (e.g., newsletters, online notifications, and marketing campaigns), this feature appears to be almost useless. However, still some of the spam campaigns have few CAPTCHAs solved (Figure 4.6 (a)) – probably due to challenges delivered to spoofed addresses as previously described in Chapter 3. Comparing spam with legitimate messages, the latter has more CAPTCHAs per campaign solved. But note that there are some spam campaigns with over 10 CAPTCHAs solved; they are classical scam campaigns located in the gray zone of the classifier.

Another viewpoint on the user behavior is through the viewed emails illustrated on Figure 4.6 (b). Although there might be comparatively little harm in viewing spam messages, as opposed to actually performing a click-through and being exposed to malware, this could also indicate interests of the users. As we can see in Figure 4.6 (b), legitimate campaigns rarely get more than a couple of views, while some spam campaigns have viewing rates over 10 emails per campaign. After examining them manually, it appears that these are mainly pharmaceutical campaigns, and few specific scam campaigns.

As for user whitelisted emails (shown in Figure 4.6 (c)), the distribution between classes is rather similar with fewer whitelisted emails. However, there is still a presense of spam campaigns with multiple whitelisted emails. This, again, suggests that email recipient decisions are unreliable and cannot be used as a groundtruth in our study.

To conclude, user-generated data are difficult to interpret, but overall it confirms that users are prone to make mistakes when judging emails in the gray area. They often open even potentially dangerous emails, ignoring security risks. These results are in line to what has been tested in a user study conducted by Onarlioglu et al. [127]. Finally, even the use of CAPTCHA-based challenges – whose goal is to identify human beings and to filter out most of the unwanted emails – is not a reliable campaign class indicator in the grey area.



(a) CAPTCHA solved emails (b) Viewed emails (c) Whitelisted emails

Figure 4.6: Number of user actions taken per campaign

4.5 Email Campaigns

In this section we discuss the main categories of email campaigns that we find in the gray area. First of all, we separate the spam from the legitimate ones. We then further divide the spam in two categories: the one generated by large infrastructures (likely sent by botnet or infected machines) from the smaller campaigns sent by few IPs. We also use a similar criterion to split the legitimate campaigns in two groups. On one side we have private marketing companies that send commercial campaigns and specialize in distributing legitimate advertisement in bulk emails. On the other, we have newsletters that are sent to the users subscribed to a web services or mailing lists, and notifications that are generated automatically when a user registers to a web service or performs certain online operations. Again, the first ones are delivered by large infrastructures, while the second ones are normally sent by a limited (and constant) set of IPs.

To identify these four categories in our dataset, we adopt a number of simple heuristics. As *commercial campaigns* we mark legitimate campaigns that belong to the biggest interconnected component of the graph described in Section 4.2.4. These are campaigns that are spread over many different networks and domain names. We consider the remaining scattered legitimate campaigns as *newsletters and notifications*. Botnet-generated campaigns are approximated by the spam clusters that are sent from more than six different countries and by more than 20 unique /24 IP prefixes. Finally, we manually sample over 350 of the remaining spam campaigns to identify *scam* and *phishing campaigns*.

All the categories are visualized in Figure 4.7, and the mean values of their features are summarized in Table 4.10.

4.5.1 Commercial campaigns

This is the largest category in our dataset covering 42% of the identified campaigns, with an average of 148 emails each. By looking manually at these clusters, we

Table 4.10: Feature mean values of campaign categories. Statistics of user actions are evaluated only on campaigns with actions

Attribute		Commercial	Newsletter	Botnet	Scam
Countries		1.4	1.14	28.2	2.74
Recipients per email		1.00	1.00	2.80	1.16
Recipient	*To*:	0.75	0.77	0.31	0
header (%)	*Bcc*:	0.07	0	0.12	0.83
	Mixed:	0.18	0.22	0.57	0.17
Sender email prefix		0.97	0.98	0.12	0.94
Sender email domain		0.96	0.99	0.31	0.97
IP distribution		0.84	0.94	0.08	0.86
Unique IPs		6	2	172	5
Rejections		0	0	0.24	0.02
Senders		0.97	0.98	0.34	0.95
Bounced		0.01	0.02	0.09	0.14
Unsubscribe header		0.59	0.39	0.01	0
CAPTCHAs		0.006	0.007	0	0.007
White emails		0.007	0.004	0.004	0.02
Period (days)		28	19	59	41
Viewed emails		3.6	6	7.3	2.9
Whitelisted emails		2.9	4	1.26	2.25
CAPTCHAs solved		19	26	1.7	7.6
Campaigns		5,113	3,597	2,107	150

confirm that these messages are mainly generated by professional email marketers sending both solicited and unsolicited advertisements. We were able to identify some of the main players (both national and international), and confirm that they actually run a legal business. On their websites, they repeatedly underline the fact that "they are not spammers", and that they just provide to other companies a way to send marketing emails within the boundaries of the current legislation. In fact, they also offer an online procedure for users to opt-out and be removed from future communications. These companies also use wide IP ranges to run the campaigns, probably to avoid being blacklisted. Moreover, we find quite interesting that some of these companies also provide a pre-compiled list of emails (already categorized by user interests) that can be used to acquire new clients.

Therefore, email recipients can be taken both from *cold lists* (i.e., people who are not yet customers), or from current customer lists. As a result, different marketers send many different email campaigns, thus forming a large interconnected network of campaigns (captured by our graph). As the senders also rely on cold lists, it is crucial to ensure that recipients can unsubscribe from the unsolicited advertisements. Indeed, commercial campaigns have the highest rate (59%) of *unsubscribe* headers.

On average, this class of campaigns lasts for 26 days, but some also continue for several months. Different email marketing companies are often involved in sending a single campaign, where each company is only active during a certain time frame. Also, each marketing service provider has its own dedicated range of IPs, which explains sometimes high IP variance and high geographical distribution of campaigns in this group. As a comparison, newsletters (Figure 4.7, upper-left part) use on average three times less of unique IPs than a professional marketer.

To conclude, commercial campaigns can be highly distributed, but, at the same time, they often adopt consistent email patterns with similar sender names and email addresses.

4.5.2 Newsletter campaigns

The newsletter senders rely mostly on local and small mailing infrastructure. The sender is often the actual company distributing the emails, with typically a small and fixed IP range. This category contains half of the emails of the previous one (probably because most of the legitimate mailing lists do not get into the quarantined area as they are already whitelisted by their customers) and covers around 30% of the total campaigns with an average size of 90 emails each.

A manual inspection seems to confirm that these campaigns consist mainly of notifications and newsletters sent by online services to which users have subscribed in the past. The senders are geographically localized (we encountered only one exception of a distributed newsletter campaign) and have extremely consistent sending

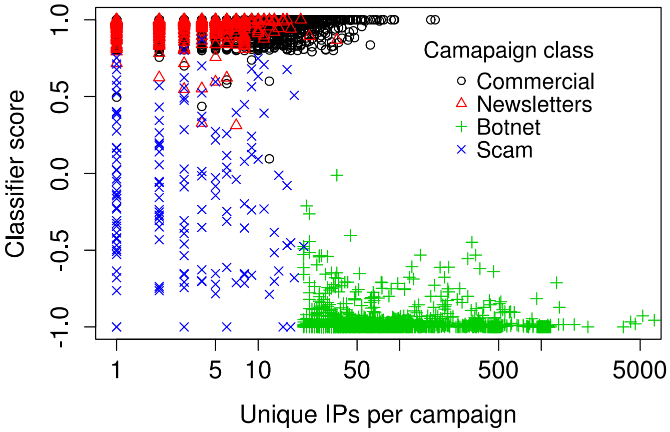


Figure 4.7: Email campaign classes distribution

patterns. Since we cluster campaigns based on their subjects, newsletters tend to last for very short periods of time. In addition, they normally use valid email recipient lists, and exhibit the lowest IP, country, and sender email address variations. Only the use of the *Unsubscribe* header seems inconsistent, as only 39% of emails use it. However, this can be explained by the fact that notification emails normally do not use this header - only newsletters are subject to optional subscription. The consistent patterns in the email headers of this category indicate that the senders are making an effort to build a reputation and successfully deliver their correspondence. Not surprisingly, this is also the category that is whitelisted the most often by the users.

4.5.3 Botnet campaigns

Unsurprisingly, Table 4.10 shows that botnet campaigns have highly dynamic attribute values, making them the easier category to identify in an automated way. This category contains the largest campaigns, but only accounts for 17% of the total campaigns (because we are focusing our study on the gray spool and most of this kind of spam is already filtered out by the antispam filters). Botnet campaigns have the highest geographical distribution as they are sent by infected computers from all over the world: 172 unique /24 networks per campaign, spread on average over 28 countries. Another prevalent characteristic is the use of multiple recipient emails sent using unverified email lists. Consequently, this leads to the highest email rejection rates (24%), and highest bounced CAPTCHA requests. The *Unsubscribe* header is rarely used, and sender email addresses have low similarities.

On average, botnet campaigns are the ones lasting the longest, with one drug-related campaign sent slowly over the entire six-months period of our experiments. Pathak et al. [130] also studied the length of *spam* campaigns, reporting a maximum length of 99 days over a dataset spanning 150 days. Our campaigns are substantially longer than that, maybe due to different datasets (we collected directly from user mail servers, not from open-relays), different email grouping methods (similar subject vs. URLs), or to changes in the behavior of spammers over time.

Despite the easily recognizable characteristics of these campaigns, users show a surprisingly high interest in these emails. This category has the highest number of email views per campaign, suggesting that users are often curious about products promoted and sold on the black market [119]. However, whitelist actions are considerably lower, suggesting that they were able to understand the nature of the campaign after reading its content.

4.5.4 Scam and Phishing campaigns

These emails contain phishing and Nigerian scam emails. They trick their victims using threatening messages or by trying to seduce them with huge monetary gains. The characteristics of this category largely resemble the ones of commercial campaigns, thus making it difficult to automatically separate these campaigns without looking at the email body. In fact, most of these campaigns belong to the gray area of our classifier. This is the reason why we needed to verify this set manually. These kind of threats are more likely to be identified by content-based detection techniques, e.g., by looking at email addresses and phone numbers (Chapter 6), or URL [130, 156] included in the body.

We find only 12,601 of such emails, with an average campaign size is 84 emails. Most of them target only few recipients (often one) at the time, located in the *Bcc* header. Phishing campaign often spoof the email addresses using well known company names (e.g. banks, eBay, Paypal), whereas Nigerian scammers rely mostly on webmail accounts (Chapter 6. In this case, many senders solve the CAPTCHA challenge - confirming that there is usually a real person behind this kind of scams. The IP addresses from where the CAPTCHAs were solved are mostly located in West-African countries, like Nigeria or Ivory Coast. None of the messages in this category include an *Unsubscribe* header.

Unfortunately, users often fell victims of this type of attacks, as they open and even whitelist messages in this campaigns.

4.6 Unclustered Area

Our campaign classification covers half of the emails in the quarantined area, with 0.2% false positive rate. One may wonder what is inside the remaining 50% that is left outside our clustering approach. Qian et al. [136] concluded that the majority of legitimate emails should not be classifiable into clusters because of the content uniqueness generated by humans. Additionally, since most of the spam and ham emails were already filtered out from our dataset, the exact proportion may be different.

We could try to approximate the content of the unclustered part by assuming that the legitimate campaign senders are always relying on a stable hosting infrastructure (as described in Section 4.2.4). In this case, for every legitimate campaign we can try to find messages sent by the same subnetwork and domain name in the unclassified email set. Using this technique, we found that 26% of the emails were sent from senders that were also responsible for legitimate campaigns. Almost 40% were sent from webmail providers. The spam set had a very low number of matches in the unclustered set, which is expected since most of these emails are sent from compromised machines that change over time.

Even though this heuristic can only provide a rough approximation of what is inside the remaining 50% of the messages, it can still be used (as part of a more complex system) to automatically separate marketing campaigns from more dangerous forms of spam.

4.7 Conclusions

In this Chapter we presented a system to identify and classify campaigns in a real-world dataset of *gray emails.* As an approximation of this email subset, we chose to use the quarantined folder of a challenge-response antispam filter, since it is already clean from obvious spam and personal user messages.

Our campaign analysis unveiled the most and the least predictive email campaign class attributes. We also demonstrated that previous techniques used for email campaign classification [136] did not provide acceptable results in our settings, confirming that the gray area contains the hardest messages to classify. Additionally, we confirmed and extended some of the findings of previous studies regarding botnet campaigns [130].

Our system could be used in different ways. First of all, it can help understanding how large commercial campaigns work, how they originate, and how they differ from other unsolicited emails. It could also serve as an input to automatically place marketing campaigns and newsletters in a separate folder, so that users can clearly differentiate these messages from other forms of spam.

The users in our study often opened botnet-generated emails and were especially prone to errors when dealing with scam and phishing messages; we believe that a separate folder dedicated to legitimate bulk emails would create an extra layer between the users and the malicious content senders inviting users first to search the bulk folder instead of spam folder. After we conducted our study, a similar solution has been implemented by Google in the Gmail Tabs [8].

Additionally, our technique could serve as an email campaign monitoring tool allowing security analysts to follow the trends of bulk email campaigns as bulk emails are known to change and evolve over the time. We demonstrated that by using a graph-based refinement method, legitimate email campaigns can be often identified based only on sender information, and can be categorized as newsletters or commercial advertisement. This is a particularly promising result in the direction of empirical study of legitimate bulk emails, and could be used for building IP whitelists of such senders, or even whitelists of marketing-management companies (like MailChimp [10]), email campaign management, and tracking services.

Finally, we also found out that our classification method works well for any campaign except scam. We believe that the latter would benefit largely from content-based email analysis. For this reason, in Chapters 5 and 6 we propose a new feature for correlating scam messages and include it in a multi-dimentional clustering tool for identifying scam campaigns.

5
The Role of Phone Numbers in Cyber Crime Schemes

During the study presented in Chapter 4, we identified four email campaign categories – commercial, newsletters, botnet, scam/phishing. We also noted that our classification method works well on all campaigns except for the scam/phishing ones. This is due to the fact that these campaigns share some common traits with legitimate emails. Such campaigns tend to be run over short periods of time, they are delivered from dedicated machine(s). Also, these campaigns are much smaller compared to other campaigns.

As in the previous Chapter, we used email sender level information for campaign identification, in this Chapter we hypothesize that the identification of these particular campaigns would benefit largely from the content-based email features. To address this, we propose the use of a new previously overlooked feature – the phone numbers. As Internet and telephones became part of everyone's modern life, several criminal activities also started relying on these technologies to reach their victims. While the use and importance of the Internet has been largely studied, previous work overlooked the role that phone numbers can play in understanding online threats. In this Chapter we aim at determining if leveraging phone number analysis can improve our understanding of the underground markets, illegal computer activities, or cyber crime in general. This knowledge could then be adopted by several defensive mechanisms, including blacklists or advanced spam heuristics.

5.1 Introduction

5.1.1 Underlying motivation

In this Chapter we look at the problem of identifying scam emails/campaigns with the goal of proposing a method to identify scam campaigns. As concluded in the previous Chapter, scam campaigns are very difficult to identify automatically based only on the email header information due to the underlying business model of scammers, who often hide behind the benign IPs of webmail providers. In order to address this specific campaign type, we focus on the content of the scam emails and study a previously overlooked feature: phone numbers. This feature might be playing an important role in this business as it provides a mean for the criminal to communicate with the victims.

However, we also intuitively assume that the same feature could be important in other types of email campaigns, or even other types of cyber crimes. Therefore, in this Chapter we first perform an empirical study of the role of phone numbers in different cyber schemes, and then look in more details at the phone numbers used by scammers performing a more detailed analysis of how scammers use their phone numbers over time. In Chapter 6, we then propose a method for identifying scam email campaigns and analyze several case studies in order to characterize the modus operandi of the scammers.

5.1.2 Phone numbers in Cyber Crime

In the current digital economy, cybercrime is ubiquitous and has become a major security issue. Every year, new attack avenues and business models arise [97, 69]. Criminals use different techniques to trap victims into various schemes and to achieve their, usually financial, goals. The used communication mechanism depends on the abuse scheme, but criminals need to have a form of interaction with their victims; for example a web page (phishing, selling counterfeit goods), an IM contact or a phone number (scams).

In many fraud schemes phone numbers play an important role. For example, criminals have been analyzed by authorities based on their phone numbers on public or underground forums [24]. In other online fraud cases, like one-click fraud [52], usage of a phone number can make the fraud appear more legitimate to a victim. Finally, scammers will often use the phone to defraud victims [149].

While the role of other features in illegal online activities has been extensively studied [113] [157] [107] [67] [54], the role of phone numbers remains relatively uncovered. The existing work is limited to the study of spam over SMS, or to phone number abuses through premium services [147] [134] [94]. However, a recent study of fraud activity in Japan [52] demonstrates that phone numbers play an important role in online fraud and can be used as a way to link and identify criminals. While there are several indications of criminals using phone numbers for their malicious activities [24], we still lack a global understanding to compare the usage and the role of the phone numbers in different criminal schemes.

5.1.3 Methodology and objectives

In this context, our research has three main objectives. First, we want to evaluate the reliability of leveraging an automated phone number analysis to improve our understanding of the underground markets, illegal computer activities and cybercriminals in general. Second, by looking at the analyzed data, we try to find various patterns associated to recurrent criminal business models. Finally, we correlate the extracted information and enrich them with a geographical HLR lookup process to automatically identify the communities responsible for Nigerian scam campaigns.

Along these three directions, we can summarize our main findings as follows:

- We present an approach, its limitations, and possible improvements for extracting phone numbers from unstructured text input;
- We study the use of phone numbers across multiple malicious online activities, with a particular focus on scam attacks. We found that while there are many overlapping numbers *within* each category, we discovered no correlation *between* datasets.;
- We show that phone numbers are a good way to detect communities of scammers and to find links between scam campaigns;
- To the best of our knowledge, we are the first to propose and use HLR lookups to verify our findings, and to study the use of phones over time of different and distributed criminal groups.

5.2 Previous Usage of the Phone Numbers in Cyber Crime

Cybercrime has become economically significant since around 2004 [120], and several research works have been conducted ever since. To this need, Fallmann et al. [73] proposed and deployed a stealthy monitoring system to capture and analyze trading information exchanged over underground Internet channels, in particular IRC and web forum marketplaces. Private forums, such as `spamdot.biz`, are often used to conduct large-scale spam operations as Stone-Gross et al. have described in [152] by taking over 16 C&C servers. Similarly, Holz et al. [92] monitored over a period of seven-months a dropzone used to collect keylogger-based stolen credentials. These works investigated the motivations and nature of these emerging underground marketplaces.

Scam is another popular technique employed by online criminals to harvest money from ingenuous victims. Stajano and Wilson, after analyzing a variety of scam techniques [149], raised the need of understanding “human factors” vulnerabilities and to take them into account in security engineering operations. One of the most popular category of scam, that goes under the name of *Nigerian/419* scam, has been extensively studied and reported, for example in [44] and [80]. Herley [91] looks into economical aspects of adversaries by trying to understand how scammers find viable victims out of millions of users, so that their business would be still profitable. Coomer [14] has recently patented a technique to use phone numbers to flag suspicious emails as either scam or spam. In comparison, our method takes an empirical approach and tries to correlate phone numbers to identify relationships between scammers and evaluate the role of phones in criminal activities. Also, it is unclear whether the patent is actually implemented in any real product.

In another scam variant, the so called "one-click" fraud, the victims click on a link presented to them, only to be informed that they just entered a binding contract and are required to pay a registration fee for a service. In [52] Christin et al. made a study on the entire business model behind these operations by analyzing over 2,000 reported incidents and correlating them using different attributes such as whois data, bank accounts, and *phone numbers*. In particular, phone numbers have been used to analyze and cluster the actors involved in the same campaign, in a similar way as we performed in our study. Dodge [62] covers several other varieties of scams over phone numbers.

Phone numbers are often used in email scams, as *premium-rate* numbers, part of fraud operations against mobile users. Porter et al. [74] analyzed 56 iOS, Android, and Symbian malware and showed that 52% of them send SMS messages to premium-rate numbers while two place *phone calls*. For example, *RedBrowser* (discovered February 2006) sends a stream of text messages, at a premium rate of $5 each to a *phone number* in Russia (as Hypponen reported in [94]). A more extensive study has been conducted by Niemela [123] who analyzed different "trojanized" and fake mobile applications that call and send SMSes to premium-rate numbers belonging to Globalstar satellite or Antarctica operators among others.

Another recent fraud that exploits telephone services for the purpose of financial rewards is *vishing* (voice phishing). Maggi [116] recently published an analysis on a real-world database of vishing attacks reported by victims through a publicly-available web application. Some papers have proposed methodologies for detecting and preventing voice-related fraud activities. Jiang et al. [98] proposed a Markov clustering-based method for detecting suspicious calls, while Enck et al. [70] used lightweight certification of applications to mitigate mobile malware at install time. Finally, Prakasam et al. [38] proposed a three step approach that first identifies emerging popular international terminating numbers, then identifies correlated foreign numbers which are contacted by the same group of mobile users, and then correlates billing information to confirm the detection results.

Last but not least, [140] describes a fully automated process of address book enrichment by means of information extraction in e-mail signature blocks. This work also confirms and emphasizes the difficulties in automated parsing of email blocks for contact details, and in particular for phone numbers.

5.3 Phone Numbers: Extraction and Quality

Phone numbers are often used, both directly and indirectly, in many cyber-criminal activities. For example, they appear in the registration of malicious domains, in the signatures of spam messages, in malware for mobile devices, and as main contact in scam and phishing campaigns. In some cases they are provided just to increase the credibility of some fake information, while in other scenarios they may represent a core component of the malicious activity itself.

At the beginning of our study we collected data from several sources related to illegal online activities. In particular, we focused on scam messages, spam messages, registration information of malicious domains (WHOIS) and Android malware. We selected those data sources because they are very likely to contain phone numbers and they are strictly related to cyber-crimes or fraud schemes.

After a first screening of the data, we observed a great variability in the quality and reliability of the collected information. To better describe this phenomenon, we classified the phone numbers along two directions: how difficult it is to extract them from raw data, and how reliable they are once they are properly extracted.

5.3.1 Extracting Phone Numbers

Properly recognizing and extracting numbers from a raw data stream proved to be quite challenging, which is consistent with results in [140]. The results mainly depend on three orthogonal factors:

How structured and easy to parse the information is

For example, WHOIS records are very easy to process and the phone number is always located inside a known and well defined field. At the other end of the spectrum, phone numbers stored in malicious binaries can be obfuscated and are, in general, very difficult to extract automatically.

How well formatted the number is

A simple regular expression can be used to extract a fully qualified number with a clearly separated international prefix (e.g., “+1 (805) 403-1234”). Unfortunately, numbers can be written in many different forms, which can be combined thus making automated parsing even harder. Phone numbers can include international prefix '+' or '00' codes, only local prefix codes, or only the phone number digits. After that, phone numbers can be grouped in variable-length groups of 2, 3 or 4 digits. Additionally, the prefixes and groups can be separated by spaces, '.', '-' or other delimiting characters, which can be country specific as well. A number without its international prefix may potentially correspond to many different numbers in different countries. Therefore, a normalization algorithm has to be used to transform the extracted number into a non ambiguous fully qualified E.164 number. When adding a country code to a candidate phone number, a *numbering plan* can be used to check if the resulting number is a valid number or not (e.g., the range is allocated and it has the correct number of digits). Unfortunately, repeating this step with too many possible country codes leads to many false positives. This is a common problem in localized cyber-crime (e.g., malicious mobile application targeting the Chinese market) because the lack of an international prefix may force the analyst to try many possibilities, thus decreasing the reliability of the collected information. Finally, short numbers (e.g., 57341) can be very challenging to detect. In fact, since the length and format are country-specific, these numbers can be easily confused with other short sequences of digits.

How noisy the data source is

This is a measure of how often the source data includes strings of digits that can be misinterpreted as phone numbers, such as identification or reference numbers, and IP addresses. This is often a problem when parsing email messages that contain several numbers mixed with text. The presence of many sequences that may resemble valid phone numbers can greatly increase the number of false positives of the automated extraction routine.

A number of heuristics can be used to improve the extraction process. For example, the immediate context of a phone number can be very useful to detect the presence of a phone number. Such context may include abbreviations or words to indicate a phone number is following (e.g., *phone, mobile, tel, fax, mobile, call, contact, line, dial, direct, ext*), combined with punctuation marks (e.g., *'.', ':'*).

The language used in the text surrounding the extracted number can also be used as a good indication of the geographic areas in which the number is supposed to be used. This is especially true for phone numbers used in scam activities, when the scammer expects the victim to call that number without ambiguity. For example, for a message written in Russian language, that includes a phone number without a full international prefix, one can try to complete the number by considering those countries where the Russian language is widely spoke, e.g., Russia '+7', Ukraine '+380', Belarus '+375', Moldova '+373'.

However, there is always a trade-off between the amount of extracted numbers and the accuracy of the results. Even by applying properly tuned heuristics, the amount of false positives when extracting poorly formatted numbers from noisy sources can be very high.

5.3.2 Phone Number Extraction Reliability

After a set of candidate numbers are extracted from the raw data, it is important to distinguish the real numbers from the fake ones. This is largely dependent on the type of activity and on the reason why the phone number was used by the attacker.

For example, numbers present in spam messages can be randomly-generated or spoofed to mimic existing phone numbers and to deceive anti-spam filters. Also, when registering a domain name there is often no validation of the authenticity of the provided numbers. However, in certain forms of cyber-crime the number has to be real and somehow controlled by the attacker. This is the case of premium numbers used in mobile malware or contact numbers used in scam campaigns.

Since distinguishing a fake or spoofed number from a real one is very hard, we decided to focus our analysis on a data source containing more reliable numbers. Unfortunately, the mobile malware dataset is very small and most of its data consists of short numbers. Therefore, in the rest of the Chapter we adopt the SCAM dataset [1] for our study.

A potential improvement to relaible phone number extraction could be achieved via *dynamic analysis validation*, i.e. calling the numbers. However, this technique is not feasible for many reasons, ranging from illegality of unsolicited calling or wardialing to financial infeasibility to call so many numbers. It is left as a separate future work.

5.4 Data Enrichment

The SCAM dataset consists of data from user reports. There are several *user reports aggregators* that cover a wide range of fraudulent activities. This information is usually reported in dedicated forums, blogs, and other online media sites. We selected the community-supported site `419scam.org` because it has a large dataset of well formatted scam reports. This dataset was manually collected, filtered and pre-processed from January 2009 to August 2012. The dataset includes metadata on each entry, i.e., the category, message headers and, for 16% of them, the corresponding original email body.

The original dataset was enriched with the service type (e.g., mobile, land line, premium) of each phone number using two different databases (so called *numbering* *plans* or *NNPC*). The first one is a free and open source XML-based database included in *libphonenumber* which derives the service type during the extraction and normalization process. The second one, is a commercial database [11] which is more complete. We use both sources to cross-check the results and detect possible discrepancies.

In our SCAM dataset, we identified in total 67,244 unique normalized phone numbers. Out of them 34,424 were UK PRS (*Premium Rate Services*) numbers (51% of total) and the rest 32,820 were non UK PRS numbers (49% of total). Out of the 32,820 non UK PRS numbers, there were 29,685 mobile phone numbers.

Finally, we collected additional information about the mobile numbers by performing an HLR lookup. HLRs are databases maintained by mobile operators containing information about the current status of a phone number – i.e., the International Mobile Subscriber Identity (IMSI), roaming status, and roaming operator. This can be very useful for our study, because this allows to know if a mobile phone number is still active and if it is roaming to a foreign country. However, HLRs are only accessible from within the SS7 telecommunication network, and therefore we had to rely on a third party commercial service [16] to query this information.

A detailed description of how HLR lookups are performed can be found in [9]. The basic idea is to contact the homing operator of a phone number pretending to be interested in initiating either an SMS or a voice call (e.g., by sending a `MAP_SEND_ROUTING_INFORMATION` message). At this point, the homing operator of the subscriber number checks the status of the mobile number and returns the details.

By performing an HLR lookup periodically for a given mobile phone number, we can get insight on the evolution of it's network status. Such status information can be used to draw conclusions about activities related to a mobile phone number. We describe the use and results of this technique in Section 5.7.

5.5 Fraud business models

In this section we summarize some of the fraud business models we observed in this work. Such models were identified using information from various sources (e.g., forums, and abused users complaints) as well as the observations we made while analyzing our datasets. While some of those business models are known, many were not well identified or were lacking empirical evidence.

5.5.1 Premium Phone Numbers

Premium phone numbers can be categorized as follows:

National Short Premium – numbers can provide high profit but are difficult to set up. However, some third party businesses offer simple point-and-click interfaces to register and configure such services;

National Premium – numbers can provide moderate to high profit, with low operational costs, and quick set up;

International Premium – numbers are complex to set up and have high operational costs. Moreover, they are blocked by some telecom operators;

UK Personal Numbering Services – UK's number ranges 070/075/076 are associated with the so called *personal numbers* allocations [28]. We detail this specific category in the next section.

5.5.2 UK Personal Numbering Services

Personal Numbering Services (PRS) (also known as *international call forwarding services* [52, 1]) are premium numbers commonly used in information services or hospital lines. However, these numbers are often abused by fraudsters as part of scams or by deceiving a victim to call a number that charges higher cost than expected. As mentioned in Section 5.4, there were 34,424 unique phone numbers in UK range of `07x` PRS numbers, which were consistent with the allocation range of UK operators [29].

Many telecom operators, some of which are only virtual operators, offer the possibility to register such numbers online. These are often offered for free: the price of communications is shared between the registrant and the operator (often retaining between 30% and 50%). In addition to this, operators can forward incoming calls to international phone numbers. This can be used as anonymization service to hide the actual geographic location of the scammer.

An interesting observation is that certain operators are used more often than others to register scam numbers. Figure 5.1 shows the distribution of phone numbers used by scammers among the providers. We observe that, in our dataset, the top 4 operators (out of 88) provide more than 90% of fraud-related UK PRS numbers. In one case, fraud-related numbers represent almost 5% of an operator allocated numbers range.

By manually comparing those and other six operators [15], we found that scammers preferred operators that:

- Have an online registration and configuration service;
- Provide an API to automate the registration process;

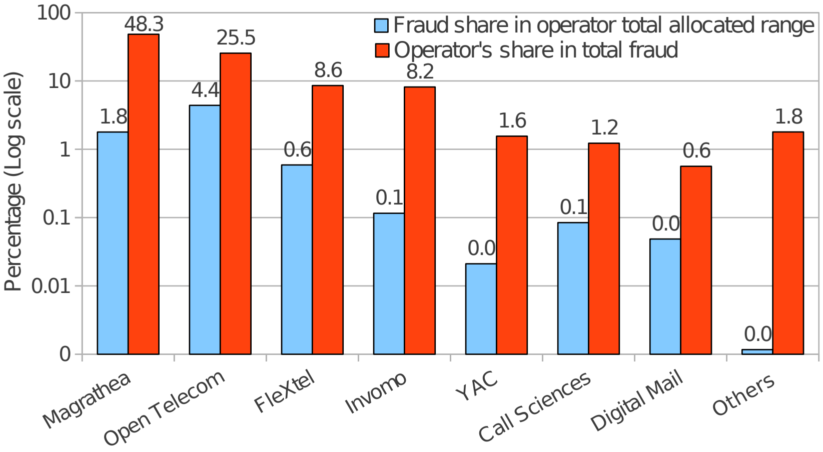


Figure 5.1: UK 07x fraud-share and fraud-vs-range allocation ratio.

- Offer cheap or free international call forwarding;
- Offer a cash back program to pay the registrant for each incoming call.

Indeed, these features are appealing to scammers and, in general, cyber-criminals that perform illegal activities.

5.6 Criminals behind the phone

In this section, we used the SCAM dataset to evaluate the use of phone numbers to identify criminals, study their behavior, and unfold the structure and the size of their networks. Scammers are known to provide real phone numbers, at which they can be reached by their victims. Therefore, this dataset is less polluted with fake or spoofed numbers, which makes our results and conclusions more reliable.

5.6.1 The SCAM Dataset

The SCAM dataset covers the period from January 2009 to August 2012 (with the exception of August 2011, which is missing from our dataset [1]). For 16% of the phone numbers, we have the original email that was used to perpetrate the scam. These emails are classified in 10 categories, three of which cover over 90% of the data: *general scam* (62%), *fake lottery* (25%) and *next of kin* (inheritance) (8%).

A first look at the relation between phone numbers and scam categories shows that scams are not evenly distributed geographically. As shown in Figure 5.2, certain types of scams rely mainly on African numbers (e.g., *new partner, orphan* scams), while others (e.g., *fake lottery, dying merchant, next of kin* scams) are almost always perpetrated by hiding behind a UK *personal number*.

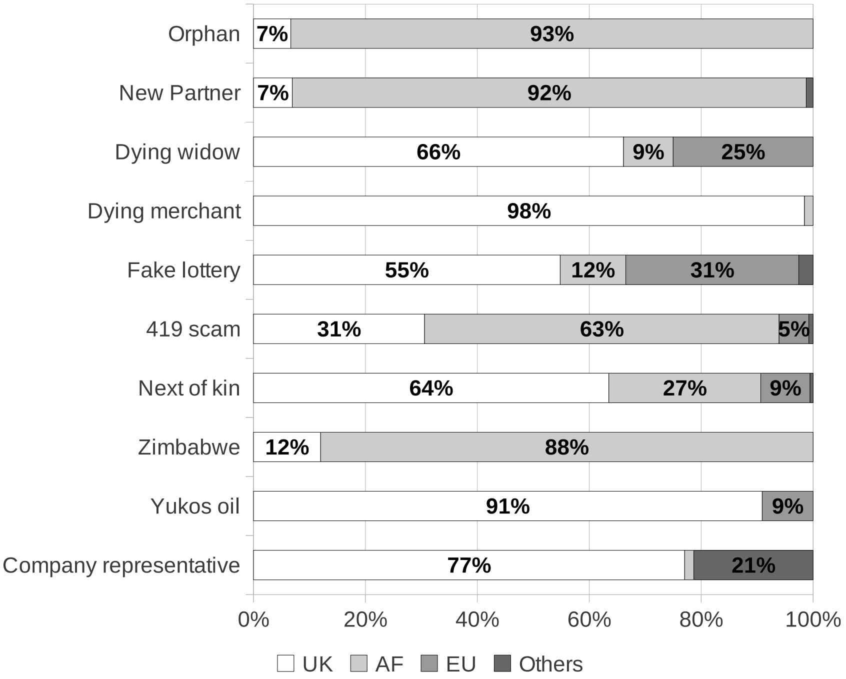


Figure 5.2: Scam email category preferences by phone number country codes.

5.6.2 Scam Communities

We first aimed at establishing relationships between phone numbers and email addresses used by scammers.

For this, we built a graph where the nodes represent either a phone number or an email address (that is used as point of contact in a scam message). The edges connecting the two types of nodes indicate that the owner of the address used that phone number in one of her scam emails. The initial graph has 34,740 nodes and 27,409 edges – 66% of nodes are emails and 34% are phone numbers. We then removed the smallest subgraphs (below 20 nodes) as they are less representative. We obtained 3,681 nodes (10.6%) and 4,360 edges (16%), consisting of 699 nodes as phone numbers and 2,982 nodes as email addresses. Globally, we identified 102 communities using the Louvain community detection algorithm [43] and 79 subgraphs.

The graph, a portion of which is shown in Figure 5.3, shows some interesting relationships. First, scammers seem to reuse a given email address to send scam messages, each message containing different phone numbers. Second, a given phone number seems to be reused in multiple scam messages or in combination with multiple different email addresses.

In particular, we observe that 37% of the phone numbers were reused by more than one scammer. Most of the largest nodes are white (phone numbers) and surrounded

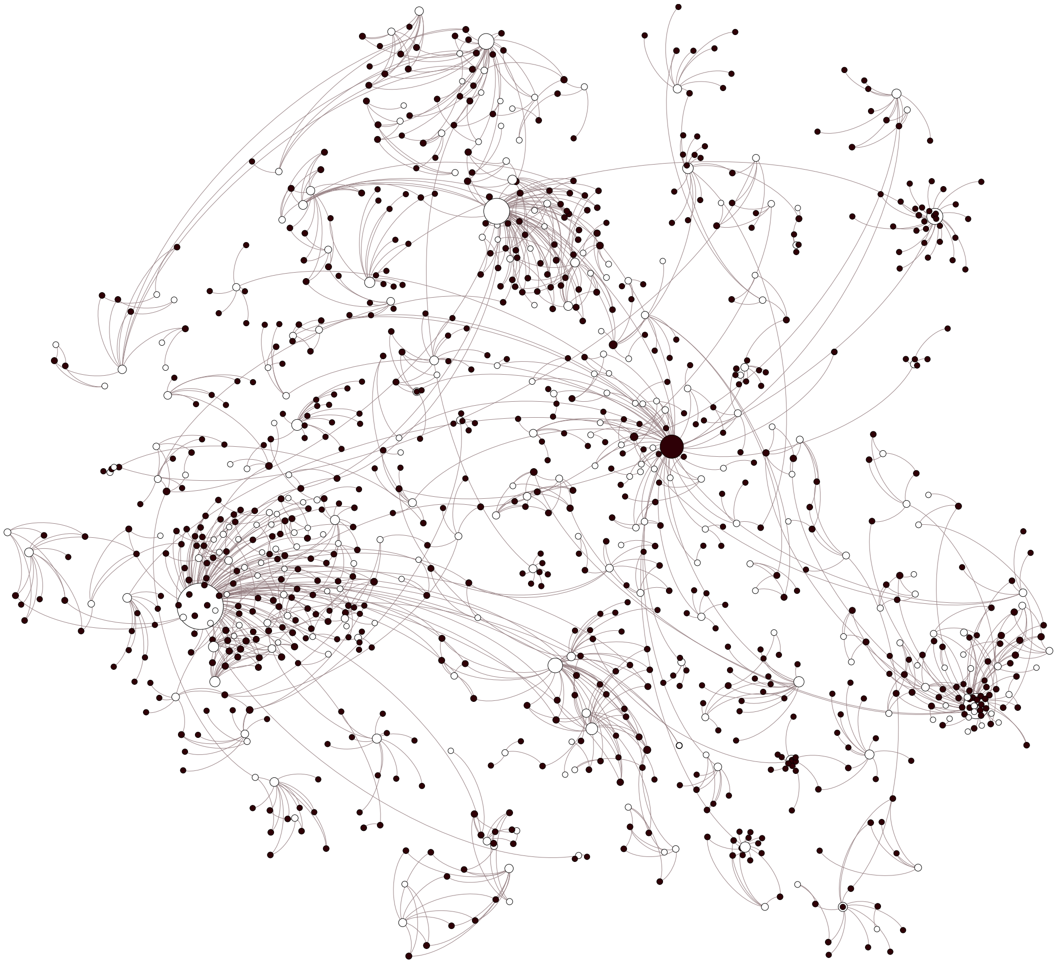


Figure 5.3: Visual relationships between phone numbers (white nodes) and email addresses (black nodes) that are used as point of contact in scam messages. The size of nodes is proportional to the number of edges.

by several small black nodes (email addresses). This suggests that phone numbers play an important role in the activities of scammers. The set of phone numbers used by scammers in their campaigns is less diverse than the email addresses. In fact, email addresses are easily blacklisted and accounts are blocked when their connection with criminal activities is discovered. Also, while email addresses are virtually free, phone numbers are usually not. This forces the scammers to continually register fresh emails for new scam campaigns. Our analysis shows that phone numbers used in scams are more stable than emails and tend to be reused over time.

By looking at the smallest subgraphs, we notice that most of them contain phone numbers registered in a single country (76%), or a country combined with UK premium numbers (10%), originating mostly from UK, Benin or Nigeria. This indicates that most of the scammers work alone, or in small groups located in a particular country. Figure 5.5 shows a real example of how scammers used four

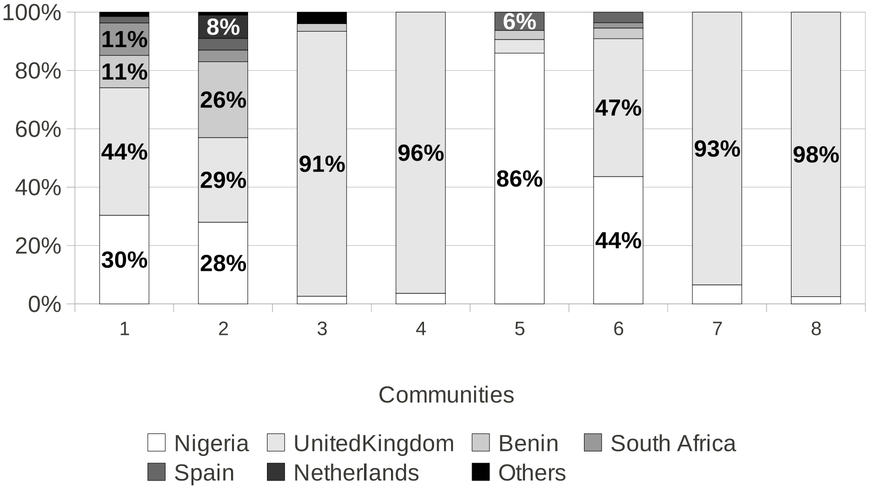


Figure 5.4: Top 8 largest communities in SCAM dataset, ordered by decreasing size from left to right.

Table 5.1: Count of SCAM phone numbers encountered in 2009-2011, reused in 2012. Includes all types of numbers.

Encounter year	Total numbers	Reused in 2012	%
2009	20,517	829	4%
2010	26,785	1,922	7%
2011	23,450	3,795	16%

Spanish mobile phone numbers in the same campaign. All the email addresses are small variations of the same person's name, probably a character that the scammers tried to impersonate.

Looking at the largest communities - densely connected sets of nodes - we see that some groups are geographically distributed over several countries. For example, Figure 5.4 shows how the eight largest communities are organized. All these communities rely on UK premium numbers (for at least 29% of their phone numbers) and on numbers from Nigerian operators. Also, these communities use cellphone numbers in several European and African countries.

5.6.3 Reusing Phone Numbers

We further tackle the question of reused phone numbers from a different angle. By looking at the SCAM dataset, which contains information on when these phone

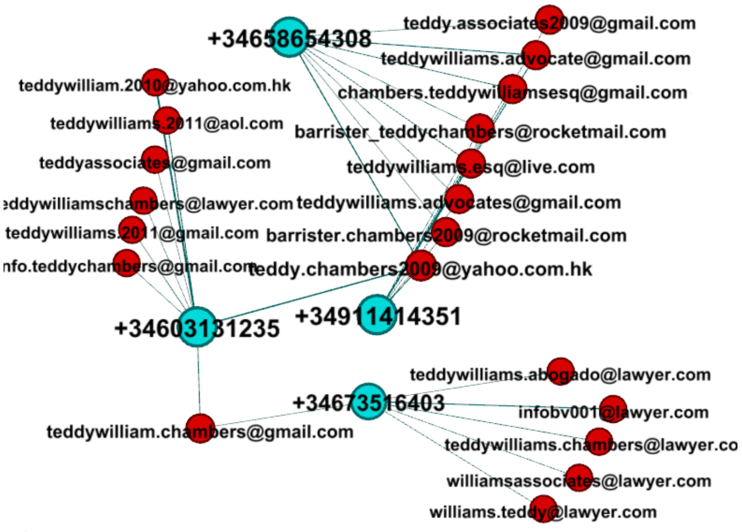

+34658654308
teddy.associates2009@gmail.com
teddywilliams.advocate@gmail.com
teddywilliam.2010@yahoo.com.hk
chambers.teddywilliamsesq@gmail.com
teddywilliams.2011@aol.com
barrister_teddychambers@rocketmail.com
teddyassociates@gmail.com
teddywilliams.esq@live.com
eddywilliamschambers@lawyer.com
teddywilliams.advocates@gmail.com
teddywilliams.2011@gmail.com
barrister.chambers2009@rocketmail.com
nfo.teddychambers@gmail.com
teddy.chambers2009@yahoo.com.hk
+34603131235
+34911414351
teddywilliams.abogado@lawyer.com
+34673516403
infobv001@lawyer.com
teddywilliam.chambers@gmail.com
teddywilliams.chambers@lawyer.co
williamsassociates@lawyer.com
williams.teddy@lawyer.com

Figure 5.5: Example of links between phone numbers and email addresses.

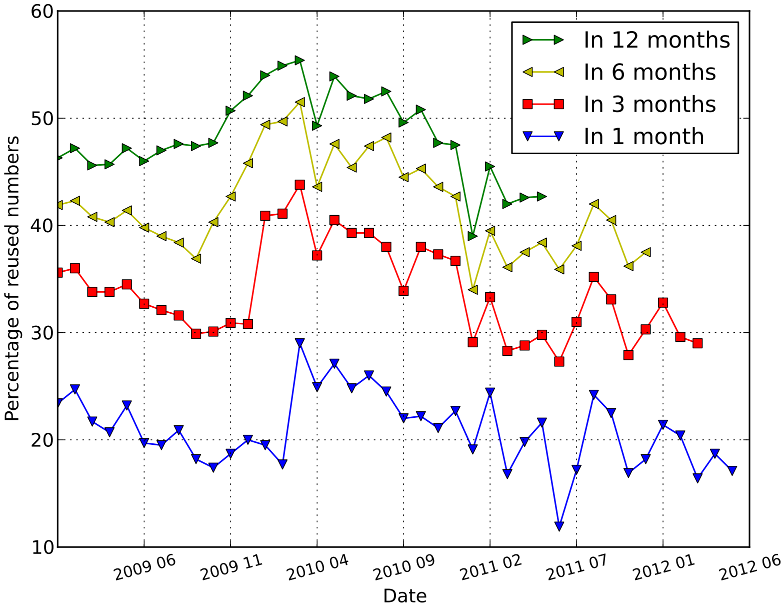


Figure 5.6: Accumulated shares of reused cellphones of scammers over time.

numbers have been used by the scammers (year and month), we understand that several of them were reused over long time periods.

Table 5.1 shows that 4% of the numbers that were in use in 2009 are still active in 2012. Figure 5.6 shows that as the period of time gets longer the amount of numbers being reused grows, from 21% (1 month) to 34% (3 months), and 48% over a year. In addition, a group of 307 phone numbers reappears yearly from 2009 to 2012. These figures do not include a detailed analysis of numbers reuse split by their type (e.g., UK PRS, mobile).

5.6.4 Discussion

The relationship between phone numbers and email addresses suggests two interesting findings. First, phones are more stable than emails and they are reused for longer periods. Therefore, phone numbers may constitute a better detection feature for the discussed threat categories. Second, even though the majority of scammers seem to operate in small groups, few communities appear to be spread over multiple countries.

However, this analysis alone is not enough to draw complete conclusions. For instance, we are still unsure how common is the phone number reuse habbit: given that 48% of phone numbers are reused within 12 months, does it mean that the remaining ones are discarded or does it mean that they are simply not reported by the website? Moreover, the fact that phones registered in different countries are used in conjunction with the same email address might be the consequence of individuals owning multiple SIM cards (e.g., collected when traveling abroad). In the next section, we introduce a dynamic phone analysis technique that helps answering these questions.

5.7 Dynamic Analysis of Scam Phone Numbers

In order to understand the organization and the dynamics behind the scam communities identified in the previous sections, we performed periodic HLR lookups (Section 5.4) of the mobile phone numbers extracted previously. With this experiment, we aim at understanding how often mobile numbers are used in other countries (i.e., roaming) and over time.

As we discussed previously, UK premium numbers (PRS) are often used by scammers to redirect calls, hiding the final call destination. We therefore had to exclude this category. We are left with 32,820 unique non-UK-PRS numbers out of which 29,685 are mobile phone numbers. Moreover, old numbers may be taken offline or assigned to a different customer. Therefore, we eventually selected the 1,333 phone numbers that were collected recently (July-August 2012).

Table 5.2: Mobile phone network status query results on 2012/08/02

Status	2012/01-06	%	2012/07	%
On the network	3,122	73%	984	84%
Replied with error	416	10%	67	6%
Turned off	734	17%	127	11%
Roaming	6	0.14%	3	0.26%

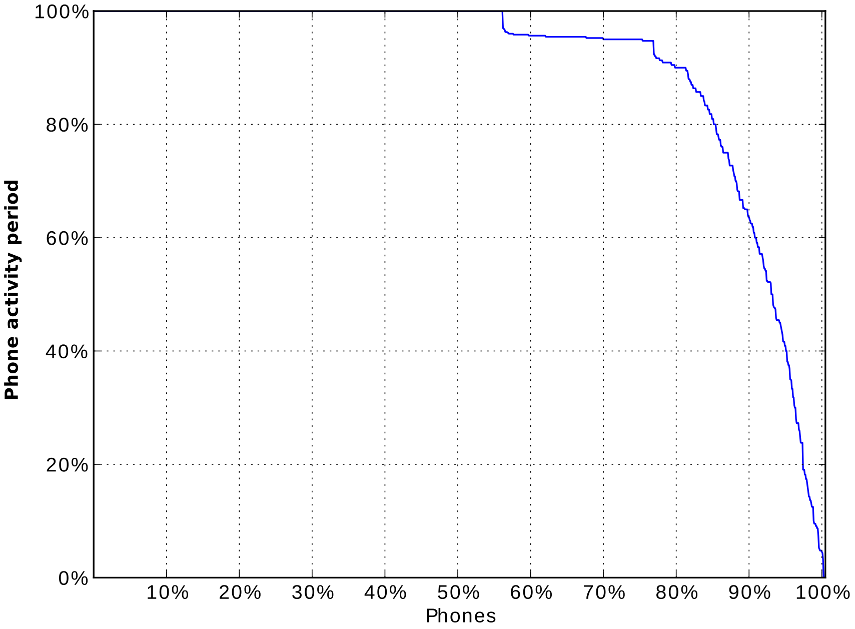


Figure 5.7: Mobile phone numbers sorted by frequency of OK status.

We verified that the selected two months period is representative of the general picture. To verify this, we performed a lookup on August 2nd, 2012 and compared the phone numbers reported in month of July 2012 with the phone numbers reported between January 2012 and June 2012. Table 5.2 shows that the population of mobile phones that were either reachable, roaming, or turned off is comparable in the two datasets, but more recently used phone numbers are more likely to be online at the time of our HLR query. This supports the fact that after a certain amount of time some phone numbers might be either discarded or replaced. Interestingly, very few numbers (only 9 in fact) were roaming in a foreign country. A first consideration is that mobile phone numbers are normally operated by criminals residing within their own countries, and not used while abroad or roaming.

That is, our first experiment consisted of doing HLR lookups for the dataset of 1,333 recently used mobile numbers. We did queries every three days and for a period

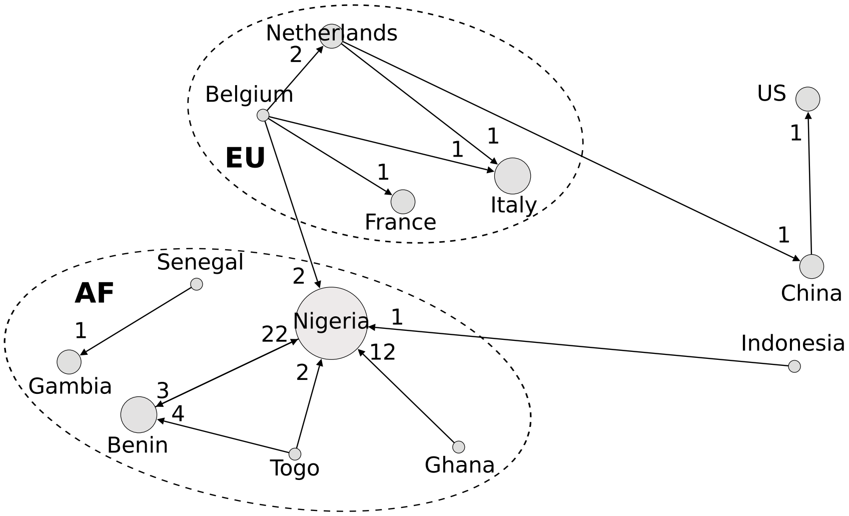


Figure 5.8: Mobile phones roaming per country. The arrow goes from the originating country to the roaming country. Edge labels indicate the number of roaming phones. The size of the node reflects the number of roaming phones in that country.

of two months. In order to appropriately choose this query window, we looked at how often the network status of a phone number is updated on average. A phone number first gets registered on the network and the HLR is updated instantly. When a phone gets turned off, the status is not updated, by default, but only when a call is received. By using one of our personal phone numbers, we determined the delay in a status change (e.g., from `OK` to `OFF`) as being 30 hours. Thus, a three days window seemed to be appropriate for our analysis.

By looking at changes in the network status attribute, we noticed that about half of the numbers have a constant `OK` status. This shows that scammers use phone numbers for long time periods by keeping them *online* most of the time. It also means that they rarely switch to new phone numbers. In fact, only 97 phones appeared to be unregistered from the network for a long time (status `Absent Subscriber`). The overall distribution of the phone availability on the network is drawn in Figure 5.7. The average scammer keeps the phone switched `ON` most of the time and only 89 numbers were `OFF` more than 75% of the time. This appears to be in-line with the business model since scammers are interested in being reached by their victims.

Finally, according to the roaming status attribute, only 50 phones were used in a different country during our evaluation (i.e., roaming). The exact roaming locations are summarized in Figure 5.8. The Figure clearly shows two clusters – one in Africa and one in Europe – with a small intersection of the two. Nigeria is still a key country for this type of business, with about 80% of the roaming belonging to it. This again supports our hypothesis that distributed groups exist and that they operate coordinated and collaboratively from multiple countries.

We then looked at the mobile operators, in order to evaluate if some of them are

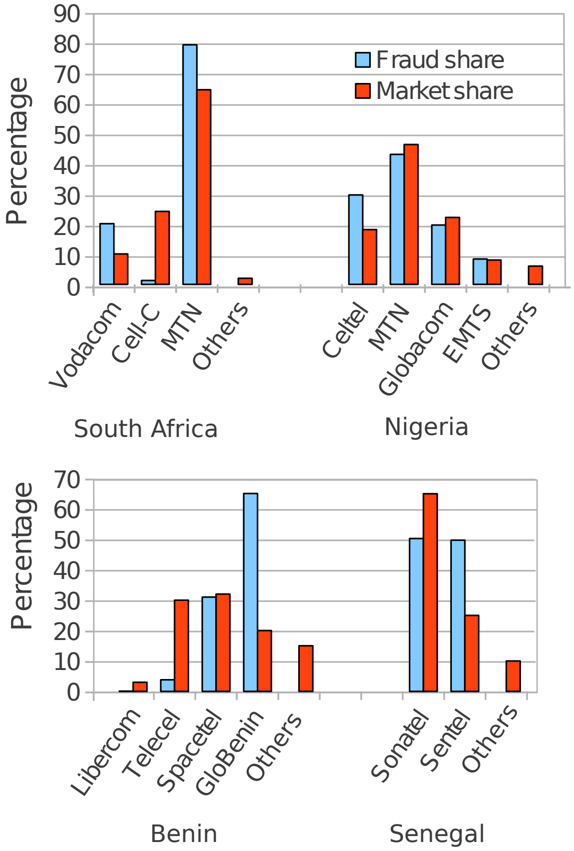


Figure 5.9: Distribution of mobile phone operators in Top 4 leading countries - market share vs. scam share.

preferred over others. We analyzed the market share of the major four countries, which contain more than 700 numbers related to scam activities: Nigeria, Benin, South Africa and Senegal. Figure 5.9 shows the difference in distribution between the market share of each operator and the “scam share” between criminals (dataset from December 2009 to December 2011). We can see that some operators seem to be less preferred by scammers (e.g., Cell-C in South Africa, Teracel in Benin), while others are clearly favored (e.g., GloBenin in Benin). The reason behind this might be due to pricing (e.g., for international calls) or stricter registration policies (e.g., strict ID checks). Like with UK PRS numbers we compared market-share and fraud-share of mobile network operators, however we did not notice any discrepancy between the two.

5.8 Conclusions

In this Chapter we analyzed the role of phone numbers in cyber-crime schemes. We collected a number of datasets and designed a technique to identify and extract phone numbers out of them. A first result is that extracting phone numbers from unstructured text is challenging and inaccurate with current tools.

We then discussed a number of common business models we observed during our experiments. Our results show that a restricted number of mobile operators are used to deliver the majority of fraud related numbers. This suggests that some operators are preferred over others by fraudsters.

For some business models changing the phone numbers of cyber criminals might be more vital for maintaining their untraceability. One option in this case would be to change the SIM cards, but this would introduce operational risks (e.g., ID checks) and other overheads. Another option would be to use Virtual Mobile Numbers that provide competitive or free pricing, laxed ID checks, and most importantly remote operation and high-level API automation.

We then focused on analyzing the role of phone numbers in 419 scam related frauds. We used HLR lookups on the scam mobile phone numbers to verify if scammers stop using their phone numbers by turning them off after publishing them, but found that many of them stay active and continue using them for long period of time (84%). This finding suggests that phone numbers would be a good feature to identify scam and to identify groups of scammers, e.g. scam campaigns. We also identified groups of scammers, created strong links between apparently unrelated actors, and analyzed their geographic distributions. One important observation is that over the period of our experiment with HLR lookups some scammers were moving across different countries (the most popular country of roaming among scammers appears to be Nigeria). Hence, in the Chapter 6, our next study of the scam campaigns will rely on the observation that the majority of the phone numbers used by scammers are mobile phone numbers used over long periods of time, and thus can be used as a feature for scam campaign identification.

Finally, while on scam messages a phone number proved to be a good identification mechanism when compared to email addresses, for traditional emails it appears to be a weak metric for identifying spam messages. This is not surprising as spammers prefer to trap their victims through engineering effort and to stay untraceable, while scammers seem to be more reluctant: they use real, therefore traceable, webmail accounts and valid phone numbers to communicate with their victims. The reuse of phone numbers is vital in the business model of scammers where trust must be established over a long period of time (e.g., when performing wire funds transfer fraud).

6

A Content-Based Approach for Nigerian Scam Campaigns

During the study of gray area in Chapter 4, we identified four email campaign categories – commercial, newsletters, botnet, scam/phishing – where our classification and categorization methods work well on all campaigns except scam/phishing campaigns. This is due to the latter having common traits with legitimate, commercial and newsletter messages.

In this Chapter, we refine the technique presented in Chapter 5 and use phone numbers to study Nigerian scam campaigns. In particular, we present some insights into a number of campaigns, showing their characteristics and relationship between each other. Finally, we describe some examples to better study criminals modus operandi.

6.1 Introduction

In this Chapter, we build upon the conclusion from the previous Chapter that phone numbers play an important role in the business of Nigerian scam. Therefore, we propose a method to identifying campaigns, then use some case studies to characterize the scam campaigns and expand our knowledge about their mode of functioning. For this purpose, we extract features such as phone numbers and email addresses appearing in scam messages. Our goal is to study how scammers orchestrate their scam campaigns by analyzing the interconnections between such features as email accounts, phone numbers and email topics used by scammers. To this aim, we use a novel multi-criteria decision algorithm to efficiently cluster scam emails that are sharing certain commonalities, even in the presence of more *volatile* features. Because of these commonalities, scam emails originating from the same scammer(s) can be grouped together, enabling us to gain insights about the scam campaigns. Additionally, we also evaluate the quality and consistency of the clustering results. For this, we perform threshold sensitivity analysis, as well as evaluate the homogeneity of clusters using compactness as a metric.

In our analysis we have identified over 1,000 different campaigns and, for most of them, phone numbers represent the cornerstone that allows us to link the different pieces together. We also discovered some larger-scale campaigns (so-called "macro-cluster"), which are made of loosely inter-connected scam operations. We believe these are likely reflecting different scam runs orchestrated by the same criminal groups, as we observe the same phone numbers or email accounts being reused across different sub-campaigns.

As demonstrated by our experiments, our methods and findings could be leveraged to *pro-actively* identify new scam operations (or variants of previous ones) by quickly associating a new scam to ongoing campaigns. We believe that this could facilitate the work of law enforcement agencies in the prosecution of scammers. Our approach could also be leveraged to improve forensic analysis and investigations of other cybercrime schemes by logging and investigating various groups of cybercriminals based on their online activities.

6.2 Dataset

In this section we describe the dataset we used for analyzing 419 scam campaigns and provide some statistics of the scam messages. There are various sources of scam often reported by users and aggregated afterwards by dedicated communities, forums, and other online activity groups. The data chosen for our analysis come from `419scam.org` – a 419 scam aggregator – as it provides a large set of preprocessed data: email bodies, headers, and some already extracted emails attributes, like the scam category and the phone numbers. Note that IP addresses data are absent. We downloaded the emails for a period spanning from January 2009 until August 2012.

In our study we also exploited the fact that the phone numbers can indicate a geographical location, typically the country where the phone is registered. Although it does not prove the origin of the message or the scammer, still it references a country of a scam operation, and improves victim's level of confidence in the received message. For example, receiving a new partnership offer from UK could seem suspicious if the phone contact has a Nigerian prefix, or a fake lottery notification with contact details originating from an African country while the victim being from Europe. Moreover, as shown in a previous Chapter, Nigerian scam mobile phone numbers are precise in indicating the country of residence of the phone owner (scammer) as few roaming cases were found. Therefore, the phone attribute is precise enough to indicate geographical origins.

Table 6.1: General statistics table

Description	Numbers
Scam messages	36,761
Unique messages	26,250
Total email addresses	112,961
Unique email addresses	34,723
Total phone numbers	41,320
Total unique phone numbers	11,768
Number of countries	12

The resulting dataset consists of 36,761 messages with 11,768 unique phone numbers. The general statistics of the data are shown in Table 6.1. A first thing to notice is that the number of email addresses is three times bigger than the number of phone numbers, emphasizing the facility to acquire mailboxes for malevolent purposes. However, still the ratio is quite low indicating rather cheap and easy access to the phone numbers.

In our dataset we did not notice any significant bursts of scam messages (verified on a monthly basis) during the three year span, suggesting that the email messages were constantly distributed over time. It is also important to note that the dataset is mostly limited to the European and African regions (with only a few Asian samples), which is due to the way the website owners are collecting and classifying the data. Nevertheless, the geographical distribution of the mentioned continents is reflected in our dataset, excluding only some minor actors.

To better understand the dataset, we look at the time during which emails and phones were advertised by scammers in scam messages. 71% of the email addresses in our dataset were used only during one day. The remaining were used for an average duration of 79 days each. Phone numbers have a longer longevity than email addresses: 51% of the phone numbers were used only for one day; the rest were used on average for 174 days (around 6 months). Hence, making it an important feature in our data clustering analysis.

Table 6.2 summarizes the phone number geographical distribution. UK numbers are twice as common as Nigerian, and three times more common than the ones from Benin, the third biggest group. Netherlands and Spain are the leading countries in Europe. Note that UK should be considered as a special case. As reported by *419scam.org* and in the previous Chapter, all UK phone numbers in this dataset belong to *personal numbering services* – services used for forwarding phone calls to other phone numbers and serving as a masking service of the real destination for the callee. In our dataset there are 44% of such phone numbers (all with UK prefix), another 44% are mobile phone numbers, 12% are fixed lines, and only less than 1% of the phones are non-existent.

Table 6.2: Phones by countries

Country	Total phones	Total in %
United Kingdom	4,499	43%
Nigeria	3,121	30%
Benin	1,448	14%
South Africa	562	5%
Spain	372	4%
Netherlands	263	3%
Ivory Coast	89	1%
China	68	1%
Senegal	47	0.5%
Togo	11	0.1%
Indonesia	1	0.01%

The initial messages are also labeled with a scam category. Around 64% of the emails are assigned to the category "419 scam" (financial fraud category). Most of the remaining emails (24%) belong to "Fake lottery" category. However, this distribution has been changing over time as shown in Figure 6.1. Especially, a big difference can be observed between 2009 and 2011, where in 2011 the "419 scam" became a dominant category. As of August 2012, there was 5 times more emails of "419 scam" than of "fake lottery" letters. This might be due to an outdated categorization process, as scam topics – like spam – may change and evolve over time. For this reason, in the next section we describe our process to automatically identify the scam topics based on the word frequencies in the messages. We also observe that most of the "fake lottery" scams are associated with European phone numbers suggesting that this category is sent to a targeted audience. In the majority of "419 scam" cases, scammers use as many Nigerian phone numbers (Figure 6.2) as of UK ones. Also notice that Benin becomes a much bigger player starting from the beginning of 2011. Hence, the geographical targets may vary with the topics and objectives persuaded by the accomplices.

6.3 Data Analysis

In this section we describe methods used for identifying groups of similar Nigerian scam emails that belong to the same campaigns and present the results. We use some metrics to evaluate the quality of the created clusters (campaigns). Finally, we extract the most repetitive keywords from the body of the scam messages in order to improve their categorization.

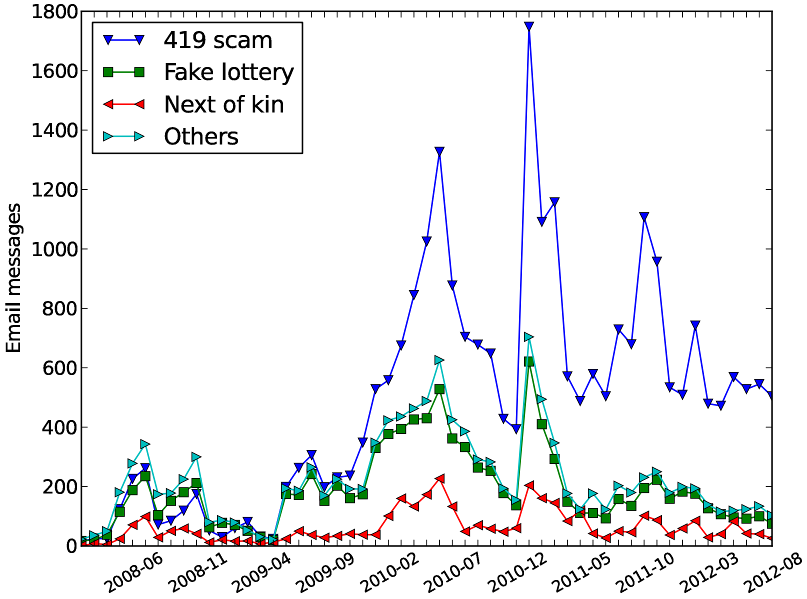


Figure 6.1: Scam email categories over time.

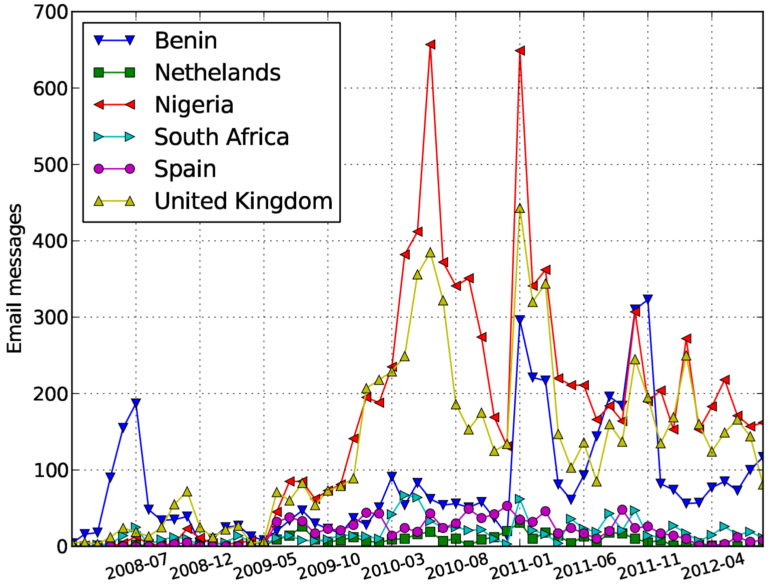


Figure 6.2: “419 scam” category phone numbers over time by countries.

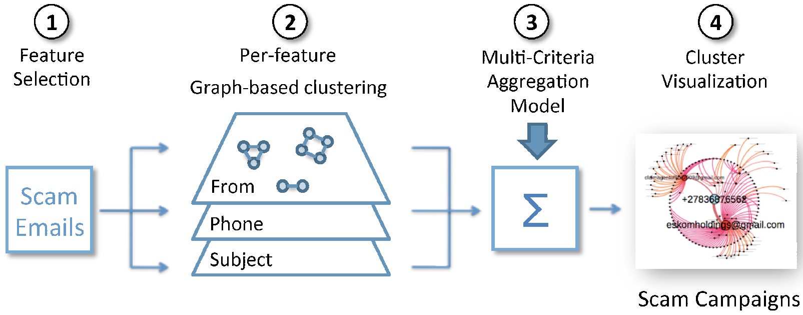


Figure 6.3: TRIAGE workflow on scam dataset.

6.3.1 Scam email clustering

To identify groups of scam emails that are likely part of a campaign orchestrated by the same group of people, we have clustered all scam messages using TRIAGE– a software framework for security data mining that takes advantage of multi-criteria data analysis to group events based on subsets of common elements (*features*). Thanks to this multi-criteria clustering approach, TRIAGE identifies complex patterns in data, unveiling even *varying* relationships among series of connected or disparate events. TRIAGE is best described as a security tool designed for intelligence extraction helping to determine the patterns and behaviors of the intruders, and highlighting "how" they operate rather than "what" they do. The framework [158] has already demonstrated its utility in the context of other security investigations, e.g. , rogue AV campaigns [56], spam botnets [160] and targeted attacks [159].

Figure 6.3 illustrates the TRIAGE workflow, as applied to our scam dataset. In step 1, a number of email characteristics (or *features*) are selected and defined as decision criteria for linking the emails. Such characteristics include the sender email address (the *from*), the email *subject*, the sending *date*, the *reply* address (as found in the email header), the *phone* number and any other *email address* found in the message itself (*email body*). In step 2, TRIAGE builds relationships among all email samples with respect to the selected features using appropriate similarity metrics. More specifically, we used various string-oriented similarity measures commonly-used in information retrieval, such as the Levenshtein similarity (for the *subject*) and the *N-gram* similarity (for features *from*, *reply*, *email body*) [105]. For features *phone* and *date*, we simply used the equality comparison method.

At step 3, the individual feature similarities are fused using an aggregation model reflecting a high-level behavior defined by the analyst, who can impose, e.g. , that *at least* k highly similar email features (out of n) are required to attribute different samples to the same campaign. The tool allows to assign different *weights* to the features, so as to give higher or lower importance to certain features. Table 6.3 shows the particular set of weights used for this analysis, in which we emphasize the importance of the phone numbers and the email subjects. The features related to the sender email addresses were given a medium importance, whereas the sending *date* was given a much lower importance.

Table 6.3: Weights of individual features ($\sum$=1)

Feature	Importance
phone	0.30
from	0.12
reply	0.18
subject	0.25
email body	0.10
date	0.05

The TRIAGE tool provides some advanced aggregation modelling capabilities, such as the *Choquet* integral – a fuzzy integral that aggregates a set of scores by taking into account importance factors assigned to individual criteria, but also *interactions* among subsets of criteria [85, 162]. This enables us to also include interactions among groups of criteria (email features), like synergies and redundancies. For this analysis, we have assigned synergies to coalitions of features involving at least the *phone* number of the scammer, so as to boost the overall similarity between emails having the same phone number, plus one additional feature in common. Inversely, some redundancy was put on certain combinations of email address-related features, such as (*reply*, *email body*), in order to diminish their redundancy effect on the overall similarity score. The definition of this aggregation model and its parameters was guided by the insights we gained previously by analyzing the role of phone numbers and email addresses in such scam operations.

As an outcome (step 4), TRIAGE identifies multi-dimensional clusters (MDC's), which in this analysis are clusters of scam emails in which any pair of emails is linked by a number of common traits. As explained in [158], a decision threshold can be chosen such that undesired linkage between attacks are eliminated, i.e. , to drop any irrelevant connection that could result from a combination of small values or an insufficient number of correlated features. The result of this sensitivity analysis is shown in Figure 6.4, which represents the total number of clusters (MDCs) found by the algorithm for increasing values of the decision threshold. The best trade-off between quality and completeness of the clustering process is usually obtained for threshold values corresponding to the maximum number of clusters [158], i.e. , we chose here to set the threshold at 0.30. Given the set of importances and interactions defined above, we can easily verify that the outcome of the Choquet aggregation will exceed this threshold for combinations of any two features involving similarities for the *phone* number and at least one other feature (besides the *date*). Any coalition of three (or more) similar features will also exceed the threshold and will lead to the formation of a cluster.

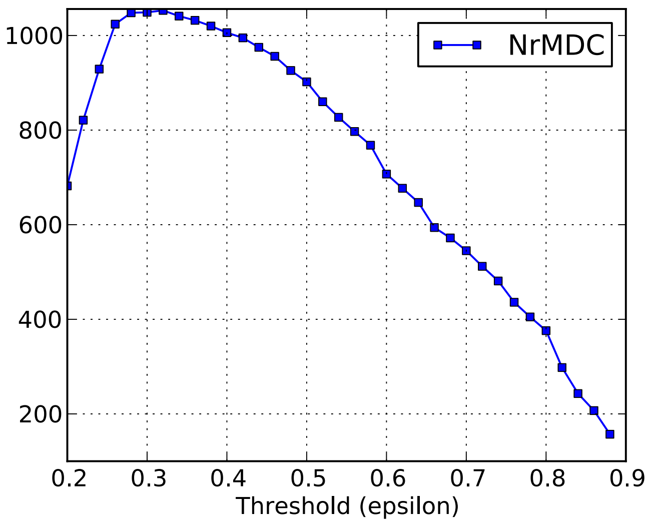


Figure 6.4: Sensitivity analysis on the decision threshold used in the TRIAGE clustering.

6.3.2 Clustering results

The TRIAGE analysis tool identified 1,040 clusters that consist of at least 5 scam emails correlated by various combinations of features. Because of the way these clusters are generated (i.e. , the multi-criteria aggregation), we hypothesize that these email clusters represent different *campaigns*, potentially organized by the same individuals – as emails within the same cluster share several common traits.

Table 6.4: Global statistics for the top 250 clusters

Statistic	**Average**	**Median**	**Maximum**
Nr emails	38	28	376
Nr from	13.9	9	181
Nr reply	6.2	5	56
Nr subjects	9.9	7	114
Nr phones	2.5	2	34
Duration (in days)	396	340	1,454
Nr dates (distinct)	27.9	22	259
Compactness	2.5	2.4	5.0

Table 6.4 provides some global statistics computed across the top-250 largest scam campaigns. In over half of these campaigns, scammers are using only two distinct phone numbers, but they still make use of more than 5 different mailboxes to get the answers from their victims. Most scam campaigns are rather *long-lived* (lasting

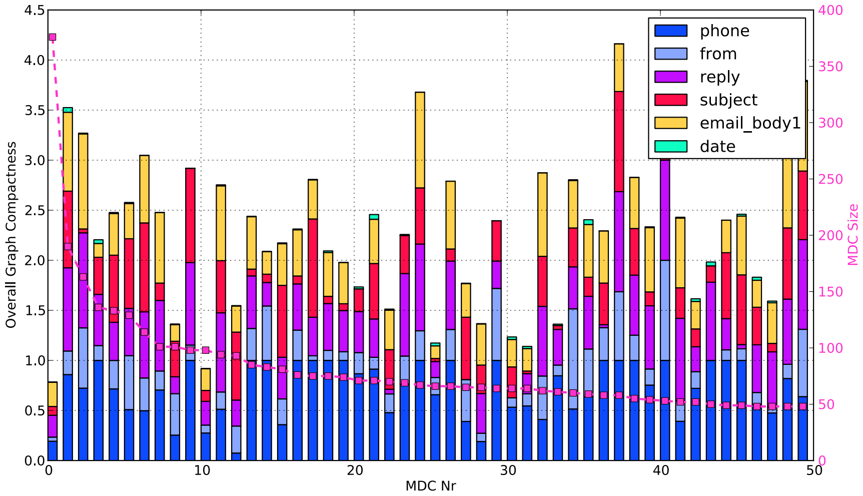


Figure 6.5: Overall compactness of the top 50 clusters (broken down by feature).

on average about a year). We note that cluster sizes are small on average indicating that there are many small, isolated campaigns and only a few dozens of messages belong to the same campaign. This might be also an artefact of the data collection process; nevertheless, we anticipate that this could also reflect the scammers' behavior who may want to stay "under the radar". Indeed, bulk amounts of the same emails would have more potential to compromise their scamming operations, as this would become visible for content-based spam filters and, hence, would get blocked on the earlier stages of email filtering.

To evaluate the quality of these clusters, we have examined their overall *compactness*, broken down by individual features. The graph compactness (C_p) is a cluster validity index that indicates how "compact" (or homogeneous) the clusters are, based on their intra-connectivity characteristics. It is a commonly-used index for evaluating the quality of a cluster, since it is easy to calculate and it reflects the average edge similarity between two objects of the cluster. Figure 6.5 depicts graphically the overall C_p for the top 50 clusters. Besides a few exceptions, most clusters have an average compactness value above 1.5, which in most cases is associated with a combination of at least 3 strongly correlated features.

Since the TRIAGE tool is keeping track of all individual links in the similarity graphs, it is also possible to compute the proportion of emails that are linked by specific combinations of features within clusters. This can be very useful to understand the reasons behind the formation of the clusters, and hence provide insights into "stable" (less *volatile*) features used by scammers when performing new campaigns.

From Table 6.5, we can observe that the top combination of features that tend to link scam emails (in 13% of the cases) involves the *phone* number, the *subject* and also all three email addresses (*from, reply, email body*) used by the scammers. To

Table 6.5: Top coalitions of features across all clusters

Coalition	**Percentage**
(phone, subject, from, reply, email_body)	13
(phone, reply, email_body)	12
(phone, subject, reply, email_body)	11
(phone, from, reply, email_body)	7
(phone, subject)	6
(phone, from)	5
(phone, reply)	4
(phone, reply, subject)	4
(phone, reply, subject, from)	4
others	33

confirm our intuition about the importance of certain features (phone numbers, and to a lesser extent, email addresses) and their effective role in identifying campaigns, we look at all similarity links within clusters. We observe that the features mainly responsible for linking scam messages in the clusters involve phone numbers (in 88% cases), followed by the *reply* email address (for 66% of the links). Not surprisingly, the *from* address (which can be easily spoofed) changes much more often and is used as linking feature in only 46% of cluster formations.

One could wonder about the longevity of these features, hence we also looked at phone numbers and email addresses from a time perspective. Figure 6.6 represents the usage of the same email addresses and phone numbers over time. The Y-axis is density of the features that indicates their distribution in time on a 100% scale. As mentioned before, many of them are used for only one day, so there is a slight concentration on the left side of the plot. However, the phone numbers are more often reused over time than email addresses. This could be explained by an easy access to new mailboxes offered by many free email providers. As for the *phone*, they probably still require some financial investment compared with emails. We checked the domain names of email addresses used in our scam dataset and found that top 100 belong to webmail providers from all over the world. This finding suggests that email messages sent from such accounts would overpass sender-based anti-spam filters that are widely deployed today. If we represent a scatter plot of *from* email addresses against phone numbers, on a per cluster basis (Figure 6.7), we find that these two parameters are uncorrelated. There are more changes performed with email addresses by scammers than with phone numbers, and even larger clusters of scammers sometimes maintain few email addresses.

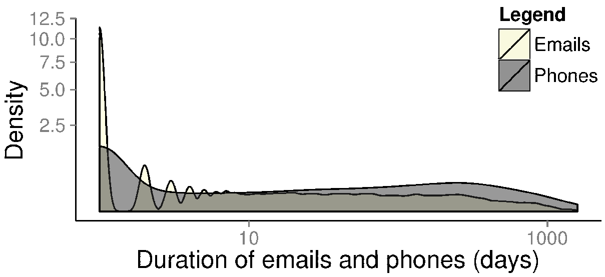


Figure 6.6: Duration of phone numbers and emails used by scammers, in days.

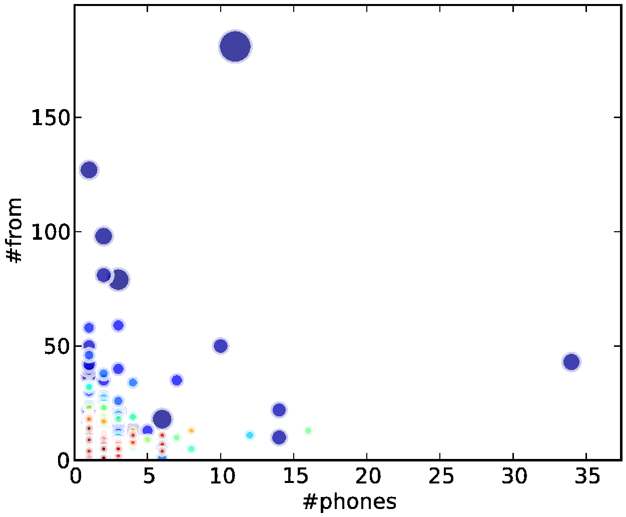


Figure 6.7: Nr of distinct *From* addresses versus the nr of phones used in the clusters. Each node represents a cluster. Node size indicates the nr of emails.

6.3.3 Clouds of words

419scam.org [1], as mentioned before, categorizes the scam emails into 10 categories. We presented their shares in the dataset in Section 6.2. Since the provided categorization is rather general, we wanted to evaluate by ourselves the scam categories present in our dataset by measuring the word frequencies in the body of the scam messages. Hence, to extract some additional knowledge from the clustered data, we create a list of the most repetitive keywords (after removing all stop words) and group them into meaningful categories. As a result, we identified three big categories within the clusters: money transfer and bank-related fraud schemes (54%), fake lottery scam (22%), and fake delivery services (11%). The rest is uncategorized and refers to 13% of the clusters. The distribution is quite similar to the one provided by the data source, except that the delivery services are separated into other categories. The so-called general “419 scam” category corresponds to messages about lost bank payments, compensations, and investment proposals. We grouped them together as it is challenging to clearly separate them due to a large number of shared keywords.

6.4 Characterization of Campaigns

This section provides deeper insights into 419 scam campaign orchestration. We present several typical scam campaigns and show the connections between clusters, which are possibly run by the same group of scammers due to multiple strong interconnections among scam emails belonging to the same cluster.

6.4.1 Scam campaign examples

Here we characterize Nigerian scam campaigns by looking at how they are operated. For this purpose we use data visualization tools to plot the clustered data in an organized manner and look at the “big picture”. Campaign graphs are likely reflecting the organization of campaigns and their maintenance over time. Interestingly, various campaigns have different operational structures and manage resources differently, as depicted by the examples in Figure 6.10.

Figures 6.8, 6.10, 6.11 and 6.12 show examples of different scam campaigns identified by TRIAGE. These diagrams were created using graph visualization tools developed in the VIS-SENSE project[1]. The graph diagrams are drawn using a circular layout that represents the various dates on which scam messages were sent. The dates are laid out starting from 9 o’clock (far left in the graph) and are growing clockwise. The other cluster nodes, which highlight other email features and their relationships,

[1] The VIS-SENSE project: `http://www.vis-sense.eu`

are drawn using a force-directed node placement algorithm. The big nodes in the graphs refer to *phone* numbers and *from* addresses. The smaller nodes represent mostly *subjects* and email addresses found in the *reply* and *from* fields, or in the message content.

Figure 6.8 is an example of a Nigerian scam campaign quite likely orchestrated by the same cyber criminals. This campaign actually consists of two sub-campaigns: first, a one-year *fake lottery* campaign located in the upper-left part of the graph (Figure 6.8); secondly, a 1,5 year campaign impersonating *ESKOM Holdings*, an electricity company in South Africa. Even though scammers changed the topic of their scam, they kept re-using the very same phone number (represented in the center of the diagram). A noteworthy aspect of this campaign, shared with other campaigns we found, is that it relies on a few *from* email addresses (*i.e.*, the bigger nodes in the figure). A set of email addresses for *reply* and *body* was used in this campaign, however, since the switch of the scam topic a set of mailboxes and subjects has also changed. Also, we observe that the load of the scam campaign is well distributed over time, and does not exhibit very high peaks on specific dates, hence keeping very low volumes of emails sent. Finally, the *from* email accounts used by scammers in this case are mostly Gmail accounts. As we have no sender IP information, we could not verify if these were spoofed or not. However, in case these are genuine email accounts, this suggests that scammers use such webmail accounts for long periods of time while staying unnoticed by the email providers.

A similar campaign, presented in Figure 6.9a, illustrates the roles of email addresses and phone numbers in Nigerian scam. This campaign, which lasted for 1,5 year, changed topic 5 times at a frequency of 1 to 2 months, which is visible in the Figure by looking at the larger subgroups placed around the circle. These shorter sub-campaigns were most probably run by the same group of scammers as the same phone number was reused over all campaigns. Inversely, we observe that the email addresses and subjects were completely changed as scammers were moving from one campaign to another. Moreover, these email addresses were often selected to match the campaign topic and subjects, probably to make the scam messages appear authentic.

While we observed a large number of such easily distinguishable campaigns, we also identified a very different, more “chaotic” type of patterns reflecting a very different modus operandi, as demonstrated with the graphical illustration in Figure 6.10a. This diagram shows a cluster representing a recent campaign of iPhone-related scams, which lasted for over 1,5 years. The communication infrastructure of the scammers operating this campaign is much more diverse – around 85 unique email accounts were used in this campaign. Moreover, it relies on a large number of “disposable” email addresses, which are typically used only once and seldom reused for long period of time. As opposed to previous examples, however, the same or quite similar subjects and *from* email address were often reused, as well as the very same two phone numbers.

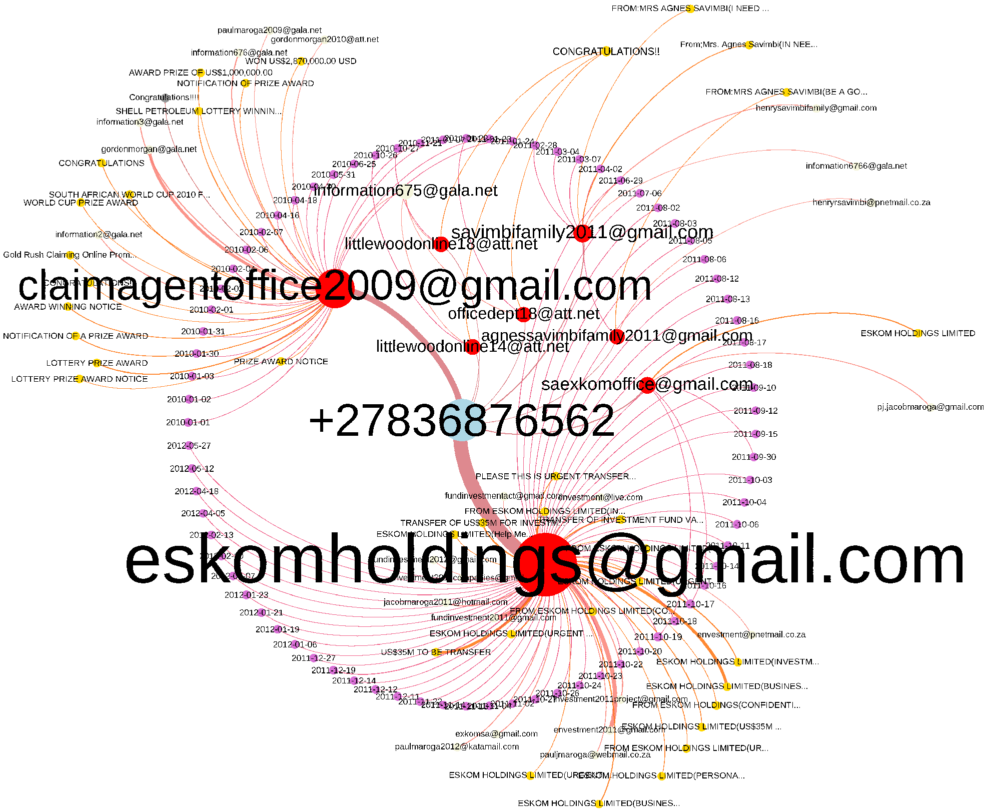


Figure 6.8: Lotteries (between 9 and 12 o'clock) and *ESKOM Holdings* impersonation.

Table 6.6: Macro-clusters, mean values of attributes

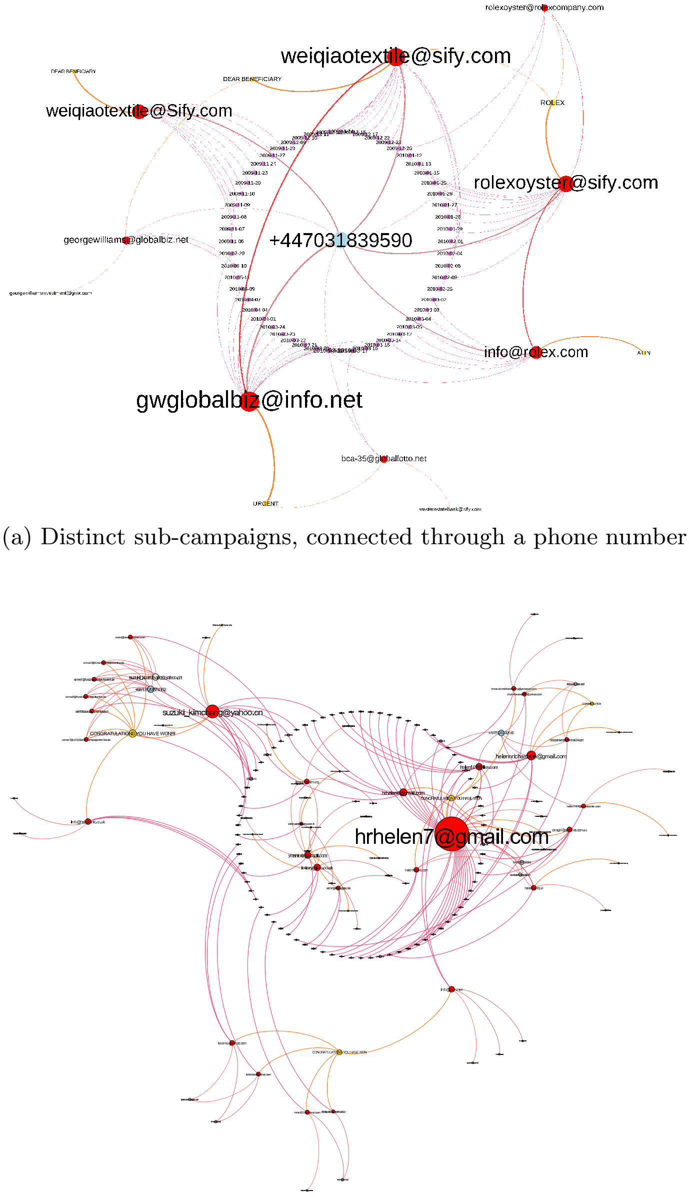
ID	Nr. of cmpg.	Tel.	Mbox	Sbj.	Dur.	Ctry	Topics
1	14	44	677	223	4 y.	4	Lottery, lost funds, investments
2	43	163	1,127	463	4 y.	7	Lottery, banks, diplomats, FBI
3	6	18	128	80	4 y.	4	Lottery
4	5	8	111	51	3,5 y.	2	Packaging, Guiness lottery, loans
5	6	7	201	96	1 y.	1	Microsoft lottery, UPS & WU delivery, lost funds
6	4	7	82	33	2 y.	1	Lottery, lost payments

Some scam campaigns still lack organization and do not always exhibit very clearly separated patterns, as illustrated by two campaigns depicted in Figures 6.9b and 6.10b. Both seem to use fewer email addresses but many different phone numbers that are changing over time. The first one is an international campaign that started with Chinese phone numbers (top-left part), then moved to UK-based anonymous proxy numbers (top-right part), and ended with Dutch-based phone numbers. Almost all analyzed features changed over time, except a single *from* address. Interestingly, most of the phone numbers were regularly switched over time.

We also note that both campaigns exploit *fake lottery* topics. For example, the second one represents a Spanish fake lottery campaign that uses topic-related email addresses (again, in order to look more legitimate), and scammers leverage up to 11 different Spanish phone numbers, which are also regularly changed over time in the campaign. In the middle of the diagram, we can see a larger node representing an email *subject* that has been reused in a large number of scam emails during this campaign, showing that scammers were probably reusing the same fake lottery email template for all these emails. However, this type of scam clusters illustrates well the challenge of identifying such dynamic campaigns, in which the links between scam emails originating from the same criminal group are constantly changing over time. This supports our choice of using a multi-dimensional clustering tool, which can not only take into account multiple features but can also identify groups of emails that are linked by *varying* sets of commonalities (*i.e.*, more *volatile* features). These complex patterns and this volatility in email attributes can also suggest that cyber criminals operate in separate groups, where each group manages its own set of mailboxes and phone number(s), however these groups are somehow federated and are collaborating with each other, for example by sharing the same email templates, same distribution lists or exchanging new scam topics.

6.4.2 Macro clusters: connecting sub-campaigns

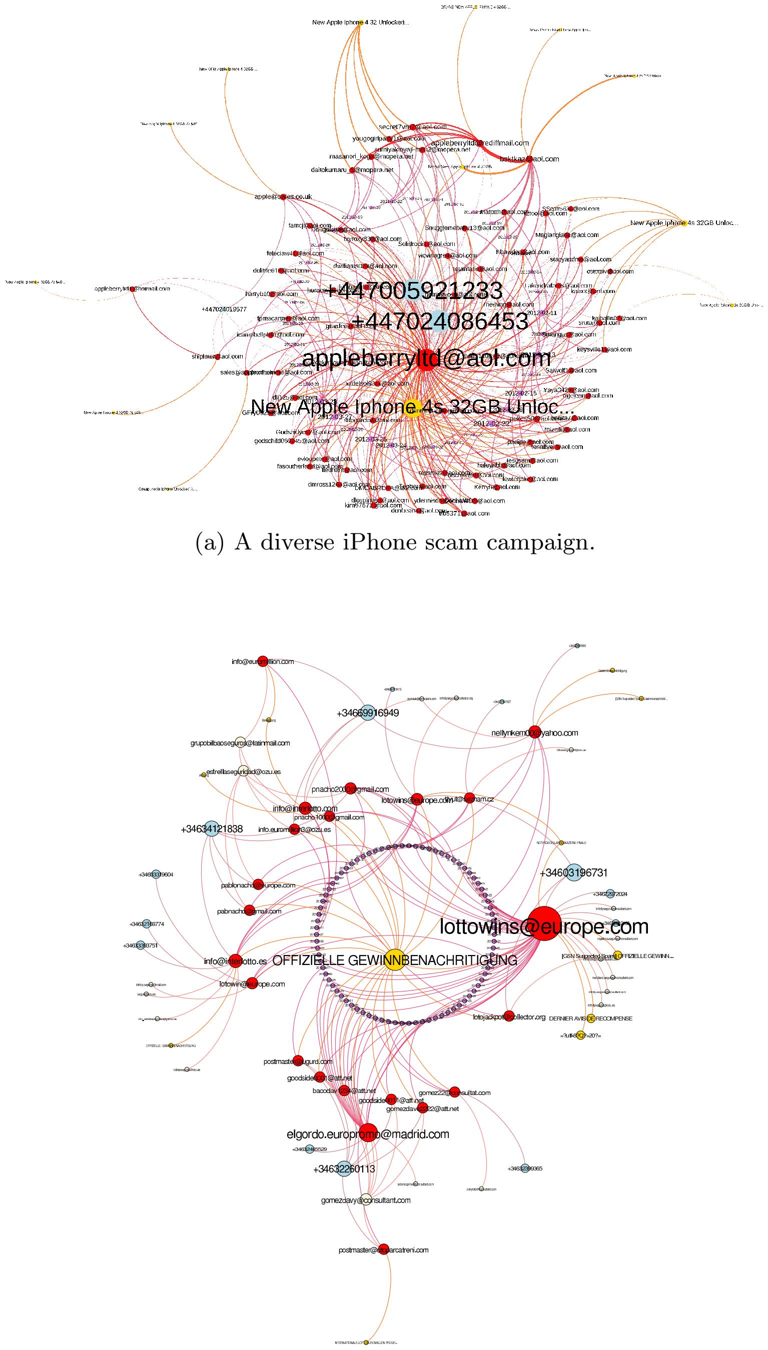
At the next step, we looked at scam campaigns from a broader perspective: by searching for loosely interconnected clusters. The goal was to pinpoint possibly



(a) Distinct sub-campaigns, connected through a phone number.

(b) International campaign operated in China, UK and Netherlands.

Figure 6.9: Examples of other scam campaign structures.



(a) A diverse iPhone scam campaign.

(b) Spanish lottery campaign changing often phone numbers and email addresses.

Figure 6.10: Examples of other scam campaign structures.

larger-scale campaigns, which are made of weakly interconnected scam operations (i.e. different scam *runs*). For this purpose, we only used email addresses and phone numbers, since the other attributes are not considered as personally identifiable information. In fact, we looked for clusters that share at least one email address and/or phone number, and use this information to build so-called *macro-clusters*.

As a result, we identified a set 845 isolated clusters, and another set of 195 connected clusters, where the latter consists of 62 macro-clusters. The characteristics of the top 6 macro-campaigns are shown in Table 6.6. These macro-clusters are particularly interesting as they consist of a set of scam campaigns that appear to be loosely interconnected and therefore could be also orchestrated by the same cybercriminals. In fact, the links between different scam clusters were considered too weak by the clustering algorithm, because of the decision scheme and thresholds set as parameters, and thus these various scam runs were eventually grouped into separate clusters. However, these weak links can be easily recovered, and it is then up to the analyst to investigate how meaningful these interconnections really are. Indeed, we believe that it is much easier for a cyber investigator to start from a set of really meaningful scam clusters, to gradually increase the decision thresholds up to the point where she can decide herself to stop merging data clusters, as it might not be meaningful any more to attribute further different campaigns to the same group due to a lack of evidence.

Macro-clusters usually span across long time periods and exhibit various bursts of emails reflecting different campaigns, which use various topics and can even be operated in different countries. An example of a macro-campaign is illustrated in Figure 6.11, where it consists of 6 different scam campaigns of various sizes that include UK and Nigerian phone numbers. We can easily distinguish them in the diagram as they appear as separate subgroups, each one having one or two bigger nodes (representing phone numbers reused multiple times) and a tail of connected nodes representing a series of *from* email addresses. Notice that campaigns in this case are well separated with respect to phone numbers and emails, which are dedicated to each campaign (or operation), and the overlaps between campaigns are quite limited. However, there is a small node just in the center that indicates how these are interconnected (through a common *from* email address). Some contact details were also reused and we used that for grouping them together. All together, these campaigns lasted for almost 3,5 years. Over this rather long time period, scammers have sent emails using 51 distinct subjects and 8 different phone numbers. This diversity of the topics suggests that there might be some competition among them, as they try to cover different online trick schemes instead of specializing in a single one.

Another example of a macro-campaign is illustrated in Figure 6.12, which consists of 14 sub-campaigns that can be more or less identified in the diagram as separate groups revolving around different phone numbers. Each one has a few dedicated phone numbers (44 in total) and its own set of *from*, *reply* and embedded email

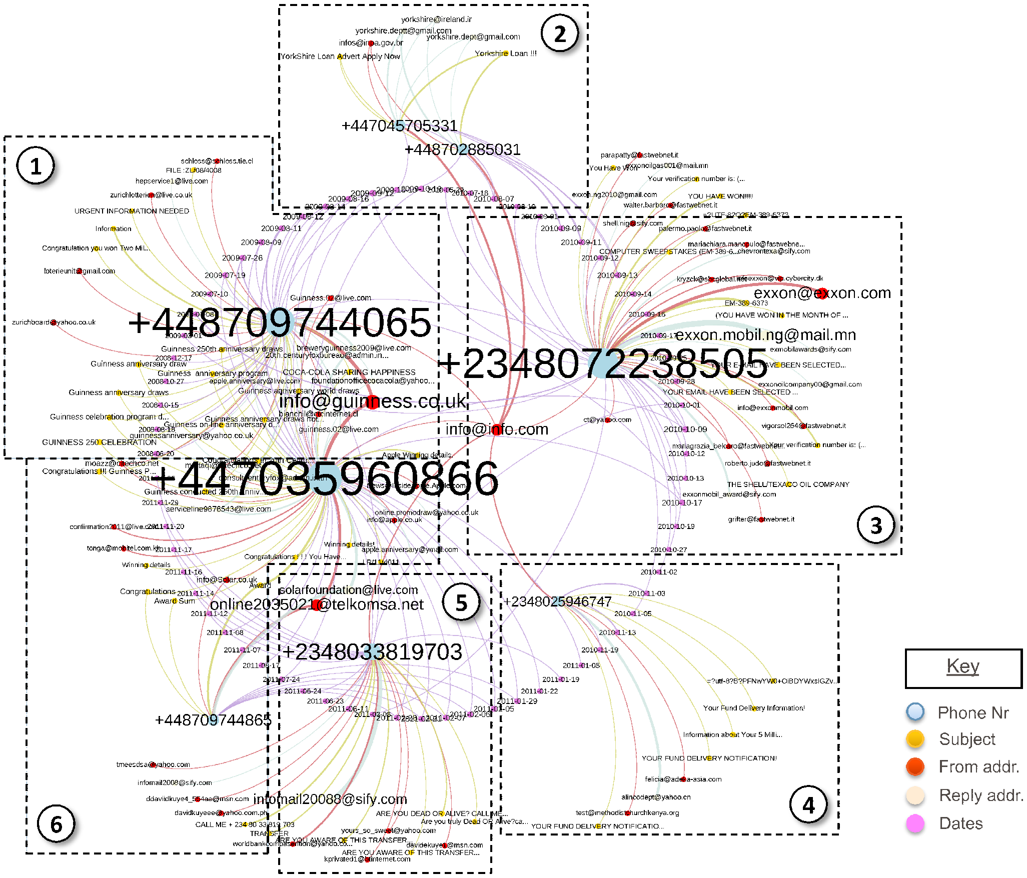


Figure 6.11: An example of macro-cluster. The nodes laid in clock-wise fashion reflect the timeline of the campaigns.

addresses. However, in this case it appears that scammers were operating these different scam runs sequentially, sometimes reusing certain resources of previous campaigns. Hence, in forensic investigations it might be necessary to look sometimes at weaker links that may possibly connect together some individuals or criminal groups that could be crime associates.

6.4.3 Geographical distribution of campaigns

To better understand how scammers operate geographically, we look at the data from a different angle. We have represented the scam email distribution per country for three subsets of our original data in Figure 6.13: (i) for the complete dataset (light grey), (ii) for scam clusters (dark grey) and (iii) for macro-clusters (black). As we can see, most campaigns identified through scam clusters originate either from African countries or from anonymized UK numbers. The difference between the light grey and dark grey bars in Figure 6.13 probably indicates a large number of stealthier or isolated scammers, as they do not form any cluster. Those quite

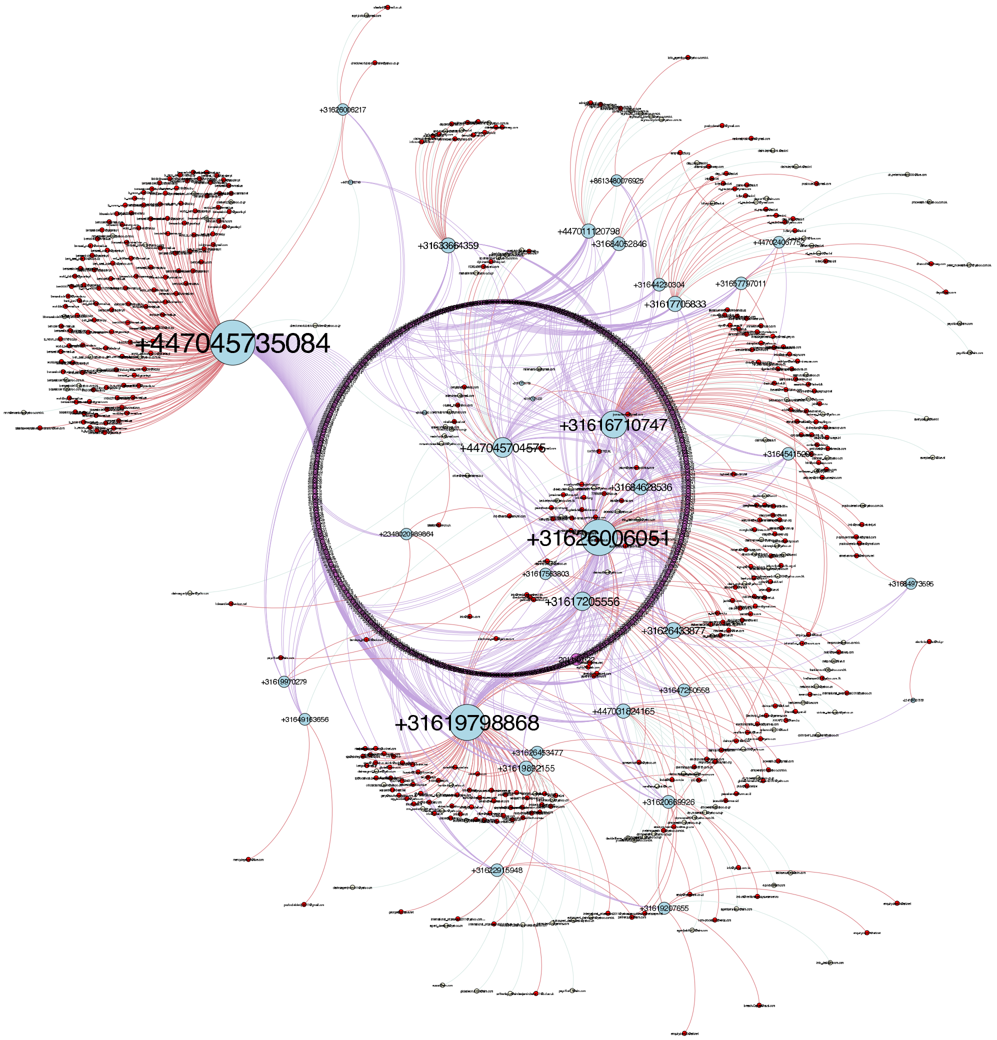


Figure 6.12: An example of macro-cluster. The nodes laid in clock-wise fashion reflect the timeline of the campaigns.

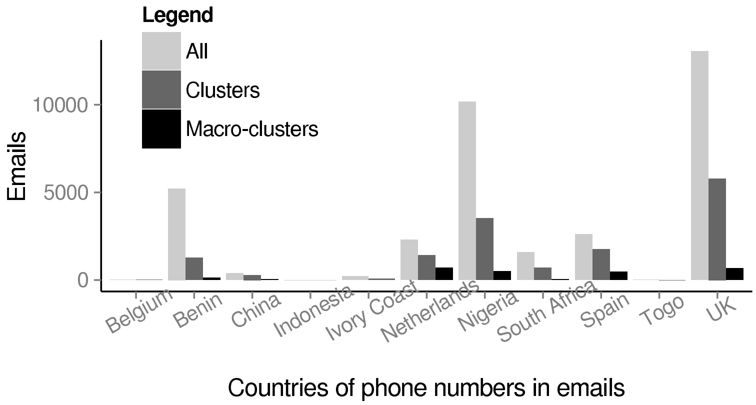


Figure 6.13: Largest macro-clusters distribution in countries.

likely refer to unorganized, opportunistic scammers, or maybe smaller gangs that operate in a loosely organised fashion.

Another interesting point is that macro-clusters (black bars) cover African and most of the European campaigns, forming bigger clusters potentially pointing to large organized groups of accomplices. Organizing such macro-campaigns might be more expensive and difficult to operate, requiring more people to coordinate in various locations and using different languages. Yet, these macro-campaigns are likely to be much more profitable, especially for the top-level leaders of these gangs.

We next look in more detail into the specific origins of macro-campaigns. Figure 6.14 shows the country distribution, versus phone number count, for the top 6 macro-campaigns. The last three campaigns are almost exclusively based in Africa, furthermore in only one or two countries, assuming that anonymised UK phone numbers are most probably used by scammers located in Africa and hiding behind these European phone numbers. The first three campaigns are more biased towards Europe, yet with strong connections to Nigeria and Benin. From an in-depth analysis we conclude that these groups are competing in several “fake lottery”-related scam, with the second group leading the pack and covering most of the countries. In comparison to the findings from previous Chapter, we observed much less UK and Nigerian numbers per group, and confirm that large-scale scam campaigns can be distributed over several continents. Indeed, the largest macro-campaign identified (#2) seems to be orchestrated by people distributed in many countries, if we assume also that mobile phones are rarely used outside its country of origin (as highlighted in Chapter 5 during the HLR study).

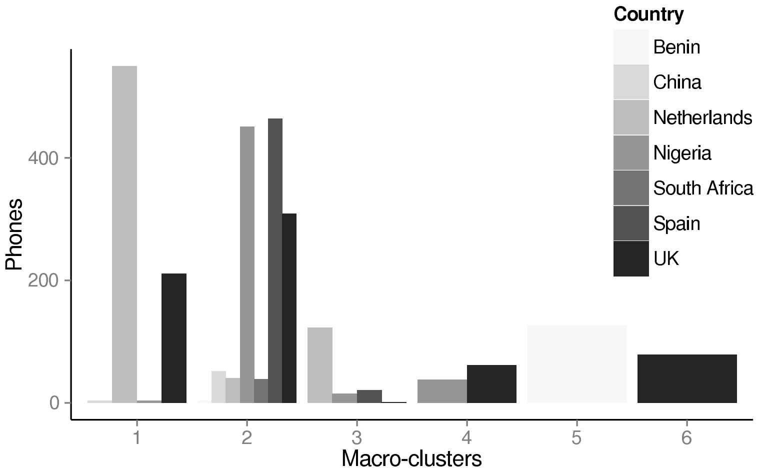


Figure 6.14: Country distribution in clustered data.

6.5 Conclusions

In this Chapter, we identified over one thousand *419 scam* campaigns with the help of a multi-dimensional clustering technique that we used for grouping similar emails. For our analysis, we then focused on the top 250 largest campaigns. Our method relies on features extracted from the email content that are very specific for Nigerian fraud: email addresses and phone numbers.

We showed that modus operandi and orchestration of such campaigns differs from traditional spam campaigns sent through botnets. Our analysis has unveiled a high diversity in scam orchestration methods, showing that scammer(s) can work on various topics within the same campaign, thus probably competing with each other over trendy scam topics. Also, the subjects within campaigns change much more often than email address or phone numbers. In general, the campaigns are diverse and are orchestrated in different ways, where some of the largest campaigns are multi-national campaigns spanning over several countries. We even identified some of them lasting for over 3 or 4 years, with some emails and phone numbers still being re-used. At the same time, scammers seem to send very low volumes of emails compared to spammers. Finally, we uncovered the existence of *macro-campaigns*, groups of loosely linked together campaigns that are probably run by the same people. We found that some of these macro-campaigns are geographically spread over several countries, both African and European.

Based on the findings, we conclude that it is challenging to identify such campaigns based only on the header data. Therefore, such campaigns require more distinct features specific to this particular kind of online fraud, like the ones studied in this Chapter. We also believe that our methods could be leveraged to improve investigations of various crime schemes – other than scam campaigns. This approach could serve forensic analysis and investigation teams to help them in studying cybercrime schemes of various groups of cybercriminals.

7 Conclusions and Perspectives

In this chapter, we summarize the investigations and contributions performed in this research thesis. We also present the answers to the general problem statements of the thesis. Finally, we propose some potential future research directions.

7.1 Research Contributions

In this work we primarily focused on studying the analysis of gray area in the email filtering system, using a real-world challenge-response (CR) system as an approximation of the gray area. We approached the problem from an empirical and analytical point of view, which enabled us to provide answers to the problem statement of this thesis.

There are three initial goals formulated in the beginning of this thesis:

1. **Evaluate** the *impact* and *effectiveness* of a Challenge-Response filter as an email anti-spam filter;
2. **Investigate** the content of the *gray area* with the goal of *reducing the burden* for email *users*, and **proposing** methods to automatically distinguish *email campaigns*;
3. **Propose** a method to identify *Nigerian* scam email *campaigns.*

The first problem was successfully addressed by performing an empirical study and analyzing 6 months of data from 47 public and private companies. Not much was previously known about the CR system's actual effectiveness in a real-world deployment, nor about its impact. In our experiments we evaluated: (i) the traffic pollution by the system due to the sent challenges; (ii) the amount of challenges generated by the system; (iii) the message delivery delay introduced by the quarantine phase; (iv) the false negative ratio as perceived by the system users; (v) the consequences of the system getting blacklisted due to hitting spamtraps. The study of the system and its quarantine area led us to the conclusion that this area is a good approximation of the gray email area, because it already excludes most of the obvious spam and ham messages.

To address the second problem, we focused on analyzing only the quarantine emails of the same CR system. We analyzed 6 months worth of data, where we intercepted around 3,3 millions quarantined emails. To perform an empirical analysis of the quarantine area, we proposed a method for email campaign identification and classification that is based only on email header information. We unveiled the most and the least class predictive attributes of the campaign, thus demonstrating that previously proposed methods would generate false positive rates as high as 10%. Furthermore, we were the first to analyze the gray area in detail and to propose a method that classifies 50% of the gray area (with a possibility to extend to 63%). Thus, the initial gray area can be reduced by half, which is an equivalent to classifying additional 15% of the total incoming emails.

While studying email campaigns in the gray area, we grouped campaigns into four categories: commercial, newsletters, botnet spam, and phishing/scam. A large amount of emails belonged to commercial and newsletters campaigns. Both categories accounted for 72% of the total campaigns. To the best of our knowledge, this was the first study that is able to automatically identify these campaign classes. However, after conducting our study, a similar solution was released by Google, Tabs in Gmail [8], that categorizes user newsletters, notifications and other commercial email content into distinctive categories.

Finally, our analysis method appeared to be inefficient against phishing/scam campaigns because they often exhibit similar behavior as commercial campaigns, and in addition use webmail accounts to hide behind webmail providers infrastructure.

Thus, the third goal of this thesis was formulated as a consequence to the latter limitation. Additionally, we concluded that in order to identify phishing/scam campaigns and accurately classify them, we need to have more descriptive features, accessible from the email content. For that reason, we proposed to use *phone numbers*, which are particularly specific to 419 scam. We empirically demonstrated, by comparing with other datasets, that this feature is especially useful in identifying 419 scam campaigns. As the phone numbers on their own can provide additional information, like country, operator, phone status (by using HLR[1] services), we enriched our experimental data with that information. We used it in the analysis of the modus operandi of scammers and their geographic distributions. We relied on a multi-dimensional clustering tool, TRIAGE [158], for grouping similar emails. Our results showed that Nigeria, as a country, plays a particularly important role

[1] Home Location Register; the database within a GSM network which stores all the subscriber data.

in the 419 scam business, and that phone numbers as a feature perform better than email addresses for identifying 419 scam campaigns.

We believe that the results of our research could be further used to improve the analysis of the gray email area, and could be especially useful for an automated identification of legitimate email campaigns, e.g., for building legitimate bulk sender whitelists. Another part of our results could serve forensic analysis and investigation teams in studying cybercrime schemes of various groups of cybercriminals.

7.2 Future perspectives

The in-depth analysis of gray area emails has opened other potential direction for research and challenges to be investigated in the future.

First, the next step could be to enrich the classifier with rejected spam emails and with commercial/newsletter emails from users in order to enlarge the coverage of both classes of email campaigns. This would further allow to reduce the gray area and to build richer legitimate bulk sender whitelists.

Second, the proposed method could be extended to also use email content data for identifying even more email campaigns, e.g. Latent Semantic Analysis by Qian [136]. This could increase the coverage of the gray area, however it is difficult to predict the amount of the possible increase.

Third, we demonstrated that by using a graph-based refinement method, legitimate email campaigns can be often be identified based only on sender information, and can be categorized as newsletters or commercial advertisement. This is a particularly promising result in the direction of empirical study of legitimate bulk emails, and could be further used to build IP whitelists of such senders, or even whitelists of marketing-management companies (like MailChimp [10]), email campaign management, and tracking services, which often rely on ranges of dedicated IP addresses.

Fourth, as it was noticed during the experiments on the quarantine area, it is especially difficult to analyze the campaigns sent from webmail providers due to the potential abuses by criminals. We reported that almost 40% of emails in the gray area were delivered from webmail provider accounts. This number is representative to pay attention to the problem of spam coming from webmail accounts. Thus, it would be interesting to look further into this phenomenon by trying to identify when webmail accounts are being abused with malicious goals and intentions.

Finally, one of the potential concerns for such whitelists would be *snowshoe spammers*. They, just like some commercial marketing companies, also rely on ranges of dedicated IP addresses, thereby could appear in the lists of legitimate bulk senders. One way to address this problem could be to add supplementary external features to the current feature set, like complaints of users filed for IP addresses, number of times the sender hit the spamtraps recently, or a historical usage (whowas) of the IP addresses.

Improvement of these different aspects could improve the currently proposed model, and could also propose new tools for identifying legitimate bulk senders, thus reducing the load from other anti-spam filters. However, the study of email spam will still continue to be ”cat and mouse game“ as spammers continuously evolve their methods of approaching the victims, thus forcing researchers to build new protection mechanisms.

Appendix A
Résumé de la thèse en Français

A.1 Introduction

Spam, selon sa définition, fait référence à des messages non sollicités qui sont envoyés en masse [7]. Historiquement, alors que son objectif demeurait le même - la publicité commerciale, l'introduction de la publicité commerciale au format numérique ouvrait simultanément de nouvelles opportunités et de nouveaux défis. Les opportunités se trouvaient dans l'automatisation et les prix très bas des annonces, tandis que de nouveaux défis introduisent de nouvelles menaces imprévues pour la sécurité des consommateurs de contenu, mettant en question la légitimité et la confiance des messages numériques. Ensuite, nous passons en revue l'historique des messages non sollicités, leur évolution et les menaces introduites et enfin la preuve de l'apparition de tendances pour l'envoi de contenu légitime en masse. Nous allons présenter le contexte de ce travail et l'énoncé du problème de thèse.

A.1.1 Contexte

Bien que le terme soit né dans les années 1980 (d'après un croquis de Monthy Python utilisé par une couronne de vikings), le premier message de spam a été envoyé en 1864 lorsqu'un message non sollicité a été envoyé via le télégramme promouvant des offres d'investissements spéciales destinés à un public ciblé de riches Américains [78]. Le premier message électronique indésirable a été envoyé sur un réseau informatique militaire (ARPANET) par Gary Turk, qui a annoncé de nouveaux ordinateurs à 400 personnes. Le point critique dans l'histoire du spam s'est produit en 1994, modifiant à jamais le secteur de la publicité commerciale. Au cours d'un scandale commercial de spam impliquant L. Canter et M. Siegel, la pratique consistant à envoyer des courriels non sollicités était défendue par des avocats, qui qualifiaient leurs critiques de "fanatiques anti-liberté de la parole" [78].

Avant cet événement, le spam par courrier électronique était plutôt une gêne - recevoir des farces, des lettres en chaîne, des messages offensants [57] - alors qu'il

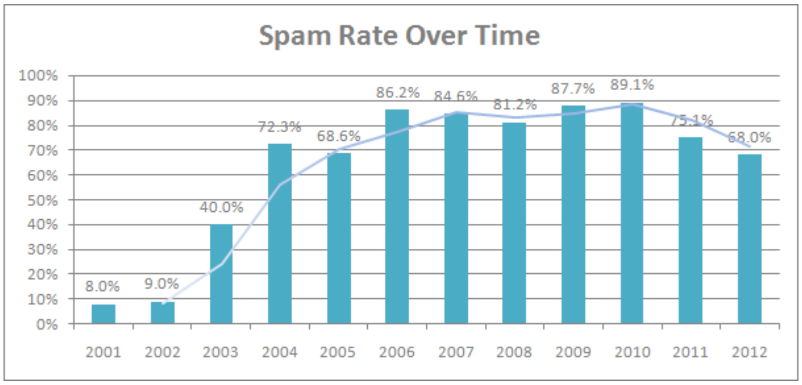


Figure A.1: Tendance du spam dans le temps d'EmailTray [68]

a rapidement évolué pour devenir de grandes entreprises de courriel commercial fonctionnant à partir de serveurs de messagerie d'entreprise. Cela a également commencé la bataille sans fin pour protéger les boîtes aux lettres d'utilisateur. Les spammeurs ont fait un autre pas en avant en 1997, lorsque les lignes de spammeurs plutôt innocents ont rejoint des expéditeurs plus déviants. À cette époque, le spam était toujours une activité liée au "travail à domicile" [75], Quand tout à coup, certaines personnes ont commencé à abuser des adresses de protocole Internet par numérotation dynamiques qui ont été réaffectées à de nouvelles adresses après une reconnexion. Par la suite, en tant que moyen de défense, les serveurs de messagerie destinataires ont commencé à bloquer les connexions à partir d'adresses IP commutées. Cela a permis de créer des listes d'adresses noires en temps réel (RBL) utilisées pour bloquer le trafic entrant provenant de spammeurs et de sources défectueuses.

En 1999, il a été remarqué [88] que les spams étaient pour la plupart des messages presque identiques pouvant être reconnus en utilisant des empreintes digitales pouvant être partagées avec d'autres utilisateurs [135]. En 2002, le secteur du spam est passé au niveau supérieur: la publication des outils de spam "Ratware" (par exemple, DarkMailer, SenderSafe) a rapidement augmenté le nombre de spammeurs et, plus important encore, a permis aux spammeurs de générer du contenu aléatoire [150]. Dans le même temps, les spammeurs ont commencé à utiliser des relais ouvert pour la transmission de leurs courriers électroniques, tirant parti de différents logiciels (y compris Sendmail v.5) configurés par défaut en tant que relais ouvert. L'apparition de tels outils a notamment eu pour conséquence la publication en janvier 2003 du virus Sobig.a, conçu pour envoyer du spam à une liste d'adresses électroniques téléchargées automatiquement.

L'année 2004 a vu naître les premiers botnets, comme Bagle et Bobax [75]. L'architecture Botnet a été construite sur un modèle informatique distribué reposant sur le réseau d'ordinateurs personnels infectés qui étaient initialement utilisés pour envoyer du spam (ils sont également utilisés aujourd'hui pour effectuer d'autres actions malveillantes). En 2007, un outil de spam distribué appelé Reactor Mailer a été mis sur le marché, suivi de plusieurs botnets de spam bien connus, tels que Storm, Cutwail et Srizbi, chargés d'envoyer des milliards de messages de spam chaque jour.

Aujourd'hui, le spam représente encore 66% des e-mails sur Internet, selon le rapport sur la sécurité des informations publié par Symantec en 2013 [154]. Malgré les récents enlèvements de plusieurs grands réseaux de zombies, le spam coûte tout de même 20,5 milliards de dollars par an en baisse de productivité ainsi qu'en coûts techniques

D'un côté, cette course aux armements a entraîné une amélioration constante du taux de détection. D'autre part, le nombre de faux positifs a également augmenté, entraînant des conséquences graves pour les utilisateurs lorsqu'un message important est signalé à tort comme courrier indésirable. Un rapport Email Intelligence [141] publié par Return Path indiquait que 16% des courriers électroniques en 2012 contenant des publicités ou des informations marketing étaient signalés comme courrier indésirable et n'atteignent donc jamais les messageries des utilisateurs. À première vue, beaucoup de gens considéraient cet "effet secondaire" comme un avantage. Cependant, on estime que seulement un tiers des utilisateurs considèrent ces messages comme du spam, tandis que deux tiers préfèrent recevoir des courriels commerciaux non sollicités d'expéditeurs déjà connus [59]. Un rapport plus récent indique que, malgré la surcharge des boîtes aux lettres, les consommateurs lisent encore 18% des courriels marketing souscrits et continuent de s'inscrire pour recevoir des offres par courrier électronique et des listes de diffusion [141]. Au final les bulletins d'information et des notifications automatisées représentent 42% des messages de la boîte de réception (il est toutefois impossible d'estimer le nombre de messages sollicités). Pour ces raisons, il est de notoriété publique que la plupart des gens consultent régulièrement leur dossier de courrier indésirable pour s'assurer qu'aucun message important n'a été classé de manière erronée par les filtres anti-spam.

Malheureusement, les solutions antispam ne sont d'aucun secours: elles ne fournissent quasiment aucune information supplémentaire pour aider les utilisateurs à identifier rapidement les courriels marketing, les newsletters ou les requêtes "à la limite" susceptibles d'intéresser les utilisateurs. Ils ont rassemblé des e-mails de marketing et de bulletin d'information inoffensifs à côté de courriels de contenu suspect, tels que le phishing, l'escroquerie et d'autres astuces utilisées par des scélérats. Naturellement, lorsque les utilisateurs parcourent leurs spams à la recherche de quelque chose qui a l'air légitime, ils doivent décider quel courrier électronique peut être approuvé, lequel est juste gênant et lequel peut constituer une véritable menace pour la sécurité. Malheureusement, plusieurs études ont montré que la plupart des utilisateurs prenaient très mal ce type de décisions liées à la sécurité [115], Ce qui en fait l'une des raisons pour lesquelles nous avons tout d'abord besoin de filtres anti-spam automatisés. Nos données confirment également la conviction selon laquelle les utilisateurs normaux ouvrent souvent intentionnellement des courriers électroniques contenant des pièces jointes malveillantes, ce qui nuit aux performances du spam et du blocage.

Alors que la plupart des recherches existantes étudient l'efficacité des techniques anti-spam et de leurs améliorations, en utilisant souvent des flux de données très spécifiques, cette thèse se concentre sur la fine ligne qui sépare le spam du ham: les rares cas dans lesquels les techniques existantes échouent. En particulier, nous limitons notre étude au domaine souvent négligé des courriels gris [172]. C'est-à-dire aux messages ambigus qui ne peuvent pas être clairement classés d'une manière ou d'une autre au moyen de filtres anti-spam automatisés. Nous partons du principe que les filtres antispam détectent la plupart des spams et qu'ils ont de "bonnes raisons" de croire qu'un message est non sollicité ou contient un contenu malveillant (par exemple, en utilisant un antivirus, une liste noire, etc.). Ou en faisant correspondre la signature d'un message d'escroquerie connu), la plupart des utilisateurs n'auraient aucune raison de revérifier cette décision. À l'autre extrémité du spectre, nous avons des messages d'utilisateur légitimes qui, nous le supposons, sont correctement classés dans la catégorie ham. Et au milieu, il y a une petite classe de courriels difficiles à classer automatiquement et qui sont souvent mal placés dans la boîte aux lettres de l'utilisateur ou dans le dossier spam [142]. Enfin, Google a récemment confirmé qu'il s'agissait d'un problème important: après la réalisation de notre étude, nous avons annoncé la publication d'onglets de boîte de réception [142] - adresses électroniques de la boîte de réception regroupées en catégories, par exemple, réseaux sociaux, promotions et forums.

A.1.2 Contexte expérimental et énoncé du problème

Les solutions anti-spam traditionnelles reposent sur deux techniques courantes: le filtrage des e-mails en fonction de leur contenu ou leur filtrage en fonction de leurs expéditeurs. La première catégorie comprend les techniques de classification de texte basées sur le contenu [40, 64, 145, 146] qui visent à trouver (souvent à l'aide d'un apprentissage supervisé) les jetons généralement associés aux messages de spam. La deuxième catégorie comprend les méthodes de détection basées sur certaines propriétés de l'expéditeur [89, 129, 138], de sa réputation [26, 167] ou du domaine à partir duquel le courriel est livré [60, 77, 167]. Même si ces deux catégories couvrent la plupart des techniques largement adoptées, le filtre Défi-Réponse (DR) [71, 128] constitue une exception notable. Cette solution est basée sur le fait que la grande majorité des bons courriels sont envoyés par des expéditeurs déjà connu du destinataire et en qui il a confiance.

Cette technique modifie l'approche en transférant la responsabilité de livraison du destinataire à l'expéditeur du message. Lorsque l'expéditeur d'un courrier électronique est inconnu, le système le met temporairement en quarantaine et renvoie automatiquement un message à l'expéditeur, lui demandant de résoudre un simple problème afin de vérifier sa légitimité.

Cela suggère donc qu'en général, le premier contact entre utilisateurs de courrier électronique a lieu beaucoup moins souvent que les communications entre contacts déjà connus. Le nom de l'approche provient du fait que, chaque fois que l'expéditeur d'un courriel est inconnu (c'est-à-dire qu'il ne figure pas encore dans la liste blanche personnelle de l'utilisateur), le système met temporairement le courriel en quarantaine et renvoie automatiquement un message à l'expéditeur lui demandant de résoudre un simple défi pour vérifier sa légitimité. Cette technique modifie en quelque sorte l'approche traditionnelle du traitement des courriers électroniques entrants, en transférant la responsabilité de livraison du destinataire à l'expéditeur du message.

De plus, bien que le filtre DR présente des avantages évidents par rapport à d'autres solutions anti-spam, il a également fait l'objet de nombreuses controverses et critiques [30, 5] en raison de ses éventuels impacts négatifs. Par conséquent, nous avons d'abord mené une étude pour analyser l'impact et mesurer l'efficacité d'un déploiement dans le monde réel d'un filtre DR.

Une partie de cette thèse a été réalisée au sein d'une société commerciale anti-spam, Mail-InBlack, spécialisée dans la gestion de courrier électronique basée sur un système de filtrage anti-spam DR. Par conséquent, la plupart des expériences menées dans le cadre de cette thèse utilisent les ensembles de données de courrier électronique disponibles au sein de l'entreprise. Cela constituait un avantage considérable, mais comportait également certaines limitations: nous avions un accès limité au contenu du courrier électronique, ce qui limitait notre analyse aux données des en-têtes de courrier électronique.

Sous ces prémisses, l'objectif de cette thèse est de:

- Évaluer l'impact et l'efficacité d'un filtre Defi-Réponse en tant que filtre anti-spam de messagerie;
- Proposer une méthode pour identifier et analyser la zone grise du courrier électronique;
- Enquêter sur le contenu de la zone grise dans le but de réduire le fardeau des utilisateurs de courrier électronique et de proposer des méthodes pour distinguer automatiquement les campagnes par courrier électronique;
- Proposer une méthode pour identifier les campagnes de courriels frauduleux.

Notez que nos expériences ont été réalisées dans le respect de la confidentialité des données fournies par l'entreprise et de ses limites.

A.2 Évaluation d'un système de filtrage anti-spam Défi-Réponse

Même si la zone grise de courriels était déjà identifiée comme le sous-ensemble de courriels le plus problématique, une analyse détaillée des recherches précédentes sur le sujet a révélé que l'on en savait très peu sur le sujet. Nous commençons par mesurer et évaluer les performances du système défi-réponse en tant que filtre anti-spam. Notre objectif est de fournir des chiffres et des statistiques du monde réel et de contribuer à éclaircir certains mythes liés aux techniques anti-spam DR. Pour cette raison, nous évaluons l'efficacité et mesurons l'impact d'un déploiement réel d'une solution anti-spam basée sur des défis. L'étude a été réalisée en collaboration avec une société commerciale anti-spam spécialisée dans un système de filtrage anti-spam Défi-réponse. Pour atteindre les objectifs, nous analysons le comportement des systèmes DR selon trois perspectives différentes:

- Du point de vue de l'utilisateur final, pour mesurer l'incidence de cette technique sur la remise des deux messages normaux à la boîte aux lettres de l'utilisateur final;
- Du point de vue de l'administrateur du serveur, se concentrant sur certains des problèmes liés à la maintenance d'une installation de DR dans une entreprise réelle;
- Du point de vue Internet, mesurer l'ampleur et l'impact des messages rétrodiffusés et des défis mal dirigés.

A.2.1 Introduction

Depuis la première introduction des techniques basées sur la DR, elles ont été considérées comme une solution extrêmement controversée [30, 5]. ils semblent être capables de bloquer complètement tout courrier électronique non sollicité, mais ils ont également un certain nombre d'effets secondaires qui peuvent sérieusement entraver leur adoption à grande échelle. En particulier, il est possible de regrouper les principales critiques contre les systèmes de responsabilité d'entreprise sous trois points principaux.

Tout d'abord, les problèmes sociaux et d'utilisation qui, d'une part, sont liés aux efforts requis de l'utilisateur pour maintenir une liste blanche adéquate, et, d'autre part, à la gêne ressentie par l'expéditeur qui doit investir du temps pour résoudre un problème de sécurité afin que son message soit délivré.

Le deuxième point à l'encontre des systèmes DR concerne le fait qu'ils peuvent introduire un retard (éventuellement visible) dans la livraison des courriels en raison de la période de quarantaine appliquée à des expéditeurs précédemment inconnus.

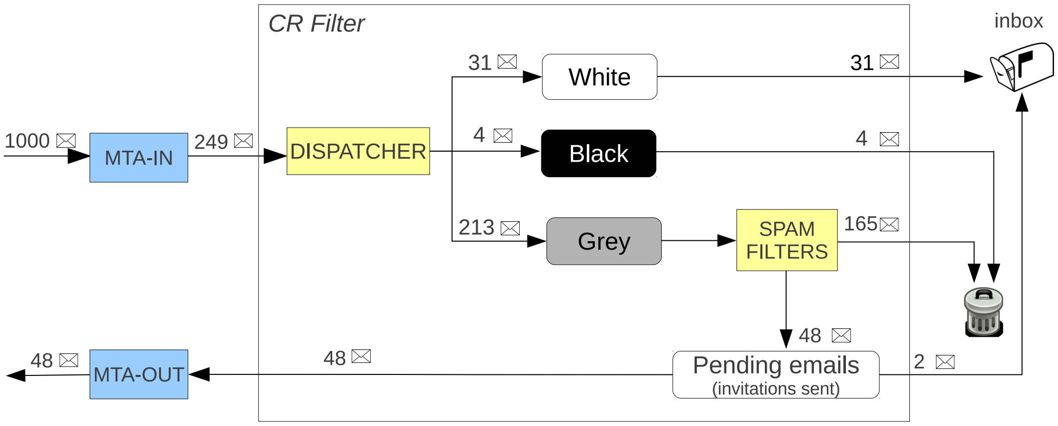


Figure A.2: Cycle de vie et distribution des courriels entrants

Enfin, la dernière (et l'une des principales) critiques à l'encontre des systèmes DR est due au défi envoyé aux courriels envoyés en réponse à des spams. étant donné que les courriers électroniques non sollicités contiennent souvent des adresses d'expéditeur falsifiées, les défis sont souvent remis à des destinataires inexistants ou à des utilisateurs innocents.

Au meilleur de notre connaissance, cette thèse présente la première étude sur l'efficacité et l'impact d'un déploiement dans le monde réel d'une solution anti-spam basée sur des défis. Dans notre travail, nous mesurons et analysons une grande quantité de données collectées pendant une période de six mois auprès de 47 entreprises protégées par un produit anti-spam commercial basé sur la DR. Nous effectuons nos mesures du point de vue de l'utilisateur final, de l'administrateur du serveur et du point de vue Internet.

Notre objectif est de fournir des chiffres et des statistiques du monde réel pouvant aider les utilisateurs et les entreprises à prendre une décision éclairée sur la base de notre étude. Nos résultats peuvent également aider à éclaircir certains des mythes liés aux techniques anti-spam CR.

La figure A.2 présente l'architecture globale du système et un cycle de vie "pondéré" des courriels entrants utilisés dans l'entreprise fournissant une solution anti-spam technique challenge-response. Le filtre DR se compose de deux composants principaux: un répartiteur de messages et un ensemble de filtres antispam supplémentaires.

Le répartiteur reçoit les messages entrants du serveur de l'agent de transfert du courrier entrant (MTA-IN) de la société. Certains des serveurs de messagerie ont été configurés pour fonctionner en tant que relais ouverts, car ils servent également des courriers électroniques pour un nombre restreint de domaines différents de ceux dans lesquels les systèmes sont installés. Cette configuration permet au serveur d'accepter des messages ne ciblant pas ou provenant d'utilisateurs connus du système.

Le serveur MTA-IN vérifie d'abord si l'adresse électronique est bien formée (conformément à RFC822), puis s'il est capable de résoudre le domaine de messagerie entrant. De plus, si le serveur n'est pas configuré en tant que relais ouvert, il vérifie également que le destinataire existe dans le système.

Notre étude montre que cette première couche de contrôles simples est responsable de l'abandon de plus de 75% des messages entrants, tandis que les systèmes à relais ouverts transmettent la plupart des messages à la couche suivante. Ces résultats sont parfaitement conformes aux valeurs similaires rapportées par l'autre analyse du taux de diffusion du spam. Le tableau suivant récapitule les raisons pour lesquelles les messages supprimés ont été supprimés: destinataire inconnu 62.36%, impossible de résoudre le domaine 4.19%, pas de relais 2.27%, courriel malformé 0,06%, expéditeur rejeté 0.03%.

Le deuxième point de contrôle pour les courriels entrants est le répartiteur de messagerie interne. Ce composant constitue le cœur de l'infrastructure DR et il incombe à chacun de décider à quelle catégorie appartient le courrier électronique: blanc, noir ou gris.

Les bobines blanche et noire sont contrôlées par la liste blanche et la liste noire de l'utilisateur. Les courriels de la catégorie noire sont immédiatement supprimés, tandis que les courriels des expéditeurs de la liste blanche sont remis à la boîte de réception de l'utilisateur. Les courriels ne correspondant à aucune des listes précédentes entrent dans la catégorie grise. Ces messages sont ensuite filtrés à l'aide de techniques antispam supplémentaires (analyse antivirus, DNS inversé et liste noire IP, par exemple). Si un courrier électronique passe les filtres, le répartiteur envoie un message de réponse à l'expéditeur d'origine contenant une demande de résolution d'un problème CAPTCHA. Sinon, le courriel est considéré comme du spam et il est supprimé.

La figure A.3 montre que les autres filtres anti-spam inclus dans le moteur de récupération d'urgence abandonnent en moyenne 54% des courriels en gris. Les messages de challenge sont à la place générés pour 28% des courriels. Dans les cas de relais ouverts, les filtres du moteur ont un taux de performance inférieur et le nombre de défis envoyés augmente de 9% supplémentaire. Cela montre que, dans une configuration de relais ouverte, le système DR reçoit plus de messages indésirables et est plus susceptible de répondre en défiant les courriels illégitimes.

A.2.1.1 Statistiques générales

Dans notre expérience, nous avons collecté des données statistiques sur un système commercial déployé dans 47 entreprises de tailles différentes. La période de surveillance a duré 6 mois, entre juillet et décembre 2010. Pour certains des serveurs, nous avons eu accès aux données pendant toute la période, tandis que pour d'autres

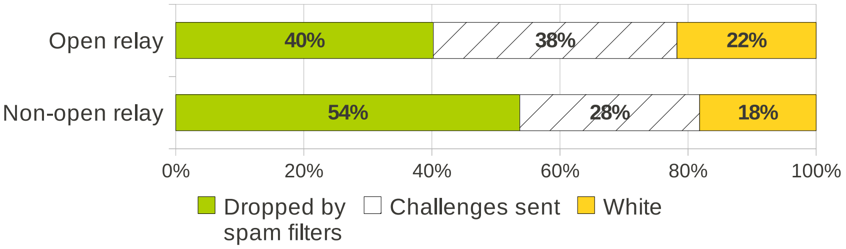


Figure A.3: Catégorie de message dans le moteur de traitement du courrier électronique interne

Statistiques Globales

Nombre de sociétés	47	Défi envoyés	4 299 610
Relais ouverts	13	Courriels Liste blanche	55 850
Utilisateurs protégés par DR	19 426	CAPTCHA résolus	150 809
Total des courriels entrants	90 368 573	Messages supprimés en raison de :	
Messages en groupe gris	11 590 532	Filtre DNS inversé	3 526 506
Messages en groupe noir	349,697	Filtre RBL	4 973 755
Messages en groupe blanc	2,737,978	Filtre antivirus et 267 630	
Nombre total de messages au MTA supprimés	75 690 366	Nombre total de messages ignorés par les filtres	7 290 922

Statistiques journalière

Courriels	797,679	Défis envoyés	53,764
Messages en blanc	31 920	Nombre total de jours	5 249

Table A.1: Statistiques des données collectées

sociétés, notre collecte a été limitée à une période plus courte (avec un minimum de 2 mois).

Au total, nous avons collecté des statistiques pour 90 millions de courriels entrants. Tous les résultats ont été nettoyés pour protéger à la fois les utilisateurs finaux et la confidentialité des entreprises. En particulier, nous n'avons jamais eu accès aux corps des messages et nous n'avons stocké que les chiffres agrégés obtenus à partir de l'analyse automatisée des en-têtes de courrier électronique.

La table A.1 affiche des statistiques générales sur l'ensemble de données que nous avons collectées. Le serveur de chaque entreprise a été configuré pour protéger certains utilisateurs avec le système de défi-réponse, tout en protégeant les autres comptes par des techniques anti-spam traditionnelles. Dans cet article, nous limitons notre analyse aux 19 426 utilisateurs protégés par la solution DR.

A.2.2 Résumé des constatations

A.2.2.1 Le point de vue Internet

Nous présentons une évaluation du nombre de courriels de défi envoyés par un système defi-réponse en fonctionnement normal. Ces messages rétrodiffusés sont souvent critiqués pour deux raisons principales: le fait que des défis mal dirigés peuvent être livrés à des utilisateurs innocents et le fait qu'un grand nombre de messages inutiles sont déversés sur Internet, augmentant ainsi le trafic mondial et surchargeant les serveurs de messagerie tiers.

Dans ce modèle simplifié, un système défi-réponse peut être considéré comme un logiciel qui reçoit un certain nombre de courriers électroniques et en "renvoie" une partie aux expéditeurs. Cette fraction, que nous appelons le taux de réflexion $\mathcal{R}$, est un paramètre important d'un système DR.

En utilisant les nombres de la figure A.2, il est facile de calculer le taux de réflexion moyen: $\mathcal{R} = 48/249 = 19.3\%$ pour les courriels atteignant le filtre CR (ou, $\mathcal{R} = 48/1000 = 4.8\%$ si l'on considère tous les courriels atteignant les MTA-IN des entreprises).

Est-ce que 19,3% est une bonne valeur pour $\mathcal{R}$? Nous concluons que le taux de réflexion est un bon indicateur de la quantité de problèmes générés par un système de DR. Dans le même temps, il est important de faire très attention à ne prendre que cette valeur pour tirer des conclusions sur la qualité de tels systèmes.

Défis mal dirigés Le nombre de défis générés ne mesure que le montant et non l'impact réel des courriels générés. En fait, tous les défis ne sont pas identiques. Certains d'entre eux atteignent les véritables expéditeurs et, même s'ils sont un peu gênants, pourraient être tolérés comme un prix acceptable à payer pour lutter contre le spam. Nous nous référons à eux comme des défis légitimes. Une deuxième classe d'entre eux est dirigée vers des adresses non existantes et constitue donc un trafic de déchets sur le réseau. Enfin, certains défis mal dirigés sont livrés aux adresses électroniques falsifiées existantes, atteignant ainsi d'autres utilisateurs innocents. Cette catégorie est beaucoup plus préjudiciable et est souvent appelée spam de rétrodiffusion.

Afin de distinguer les trois catégories de problèmes, nous avons analysé le statut de la livraison du défi dans les journaux des serveurs. Dans les systèmes analysés, nous avons constaté que seuls 49% des problèmes avaient été livrés avec succès aux serveurs de destination. Les 51% restants ont soit rebondi, soit expiré après de nombreuses tentatives infructueuses.

Une autre pièce du puzzle peut être trouvée en mesurant le nombre de problèmes réellement résolus. Les travaux antérieurs [71], menés dans un environnement contrôlé, ont estimé qu'environ 50% des problèmes n'avaient jamais été résolus.

Malheureusement, notre étude montre une image complètement différente. Selon les journaux des serveurs Web des sociétés que nous avons analysées, en moyenne, 94% des URL CAPTCHA incluses dans les défis livrés n'ont jamais été ouvertes. La troisième catégorie, à savoir les spams rétrodiffusés, peut plutôt être approchée avec le nombre de défis livrés correctement mais jamais résolus, c'est-à-dire entre 0 et 45%.

Pollution de la circulation Le taux de réflexion mesure uniquement le nombre de messages, sans tenir compte de leur taille. Par conséquent, l'estimation de la quantité de trafic généré par un système de réponse au défi n'est pas très précise. Pour cela, nous devons étendre la définition précédente en introduisant le *ReflectD Ratio de trafic* $\mathcal{R}_T$, qui représente le rapport entre la quantité de trafic reçue par le système et la quantité de trafic de messagerie générée. pour les défis. Sur une période d'un mois, le rapport moyen que nous avons mesuré au filtre CR était de $\mathcal{R}_T = 2,5\%$. Sur la base des chiffres précédents, nous pouvons estimer qu'un déploiement à grande échelle de filtres anti-spam de type challenge-response augmenterait le trafic de messagerie sur Internet d'environ 0,62%.

A.2.2.2 Le point de vue de l'utilisateur

Une autre préoccupation est que les courriels normaux peuvent être bloqués et rester dans la liste des valeurs de l'utilisateur en attendant que les problèmes correspondants soient résolus. Cela peut se produire pour deux raisons: parce que l'expéditeur doit toujours résoudre le problème ou parce que le courrier électronique est envoyé par un système automatique et que le défi est par conséquent abandonné ou jamais remis. Les données ont montré que 30% des messages sont retardés de moins de 5 minutes et que la moitié sont livrés en moins de 30 minutes. Toutefois, si le problème n'était pas résolu au bout de 4 heures, l'utilisateur devait sélectionner manuellement les messages à partir du résumé, avec un délai de livraison compris entre 4 heures et 3 jours en moyenne.

En combinant les chiffres, nous pouvons conclure que:

- 94% des courriers électroniques de la boîte de réception de l'utilisateur sont envoyés à partir d'adresses déjà inscrites dans la liste blanche, et sont donc envoyés instantanément.

- Sur les 6% restants des messages mis en quarantaine dans la file d'attente grise, la moitié d'entre eux sont livrés en moins de 30 minutes car l'expéditeur a résolu le problème.

- Seulement 0,6% (10% des 6%) des messages ont été livrés avec plus d'un jour de retard.

A.2.2.3 Point de vue de l'administrateur

En raison de l'utilisation de la DR, les messages de réponse à la demande peuvent frapper un piege à spam, c'est-à-dire un ensemble d'adresses électroniques maintenues et distribuées dans le seul but d'attirer le spam. Les courriels collectés par ces pièges sont souvent adoptés par les services populaires pour mettre à jour leurs listes noires. Par conséquent, l'adresse IP utilisée pour envoyer les défis peut elle-même être inscrite sur la liste noire à la suite du spam rétrodiffusé qu'elle envoie.

Grâce aux données que nous avons collectées, nous avons pu estimer la vitesse à laquelle différentes adresses IP de serveurs de challenge sont mises sur liste noire. Le résultat, résumé sur une échelle logarithmique dans la figure refimg: srv-black, montre que, si la plupart des serveurs n'avaient aucun problème avec la mise au noir, certains d'entre eux étaient souvent placés sur la liste noire, même pendant quelques jours d'affilés . Cependant, il semble n'y avoir aucune relation entre le ratio de liste noire de serveurs et le nombre de défis envoyés.

A.2.3 Conclusions

L'étude conclut que les systèmes DR offrent en général une excellente protection anti-spam (détection de 99,9%) et que leurs effets secondaires ont un impact et un coût comparables. Nous avons estimé que les courriels en quarantaine (courriel en attente auxquels un défi est envoyé) constituent 30% de tous les courriels entrants. Environ 6% des courriels en quarantaine sont ensuite envoyés dans la boîte de réception de l'utilisateur, ce qui suggère que la majorité des courriels de cette zone restent inaperçus ou ne sont pas pertinents. En fait, ce type de système de filtrage anti-spam fournit un point de vue privilégié pour étudier le phénomène des emails en zone grise. La zone en quarantaine est une bonne approximation de la zone grise, car elle exclut déjà la plupart des messages utilisateur personnels et des spams. Une évaluation approximative ultérieure de la similarité des courriels dans cette zone a montré que les courriels peuvent être regroupés en campagnes, certaines ayant des caractéristiques très dynamiques et peu de courriels authentifiés (ceux dans lesquels l'expéditeur a résolu le problème de la DR), tandis que d'autres ont des caractéristiques plus statiques avec des valeurs plus élevées du taux d'authentification. Cette idée nous a conduit à la deuxième partie de cette thèse - l'investigation du contenu de la zone grise.

En particulier, nous avons mesuré la quantité de défis générés par ces systèmes et leur impact en termes de pollution du trafic et de messages rétrodiffusés éventuels envoyés à des utilisateurs innocents. Nous avons ensuite mesuré la quantité de courriels retardés en raison de la phase de quarantaine et la quantité de spam pouvant passer par le filtre et atteindre les boîtes aux lettres des utilisateurs. Enfin, nous nous sommes concentrés sur un problème moins connu, à savoir le fait que les invitations envoyées par ces systèmes peuvent frapper par inadvertance un piège à spam et provoquer la mise sur liste noire du serveur de messagerie.

Nos résultats peuvent être utilisés pour évaluer à la fois l'efficacité et l'impact de l'adoption de cette classe de techniques, et les chiffres fournis peuvent aider à résoudre le long débat entre les défenseurs et les opposants des systèmes de DR.

Enfin, nous examinons également les courriels en quarantaine qui excluent déjà les messages de spam et de ham évidents. En fait, ces courriels particuliers représentent une bonne approximation d'une zone grise qui, par définition, contient des messages difficiles à classer automatiquement par les filtres anti-spam. Dans cette étude, nous montrons même qu'en regroupant des messages similaires dans ce domaine, nous sommes en mesure d'identifier deux classes principales de messages: les uns avec des en-têtes de courrier électronique et des modèles d'envoi relativement stables, et les autres avec des caractéristiques plus dynamiques. Cela suggère que la zone grise se compose d'une partie de courriels très similaires présentant des modèles ressemblant au spam envoyé par le botnet, mais inclut également d'autres types de campagnes de courriels en masse telles que des newsletters, des notifications ou des courriels commerciaux. Ce qui suit nous amène à la prochaine question de recherche de la thèse - l'analyse de la zone grise et ses campagnes par courrier électronique.

A.3 Analyse automatisée de la zone grise du courrier électronique

Nous avons montré au chapitre précédent qu'un système anti-spam de défi-réponse fonctionnait différemment des autres systèmes anti-spam en raison de ses avantages et de ses effets secondaires. L'un des effets secondaires est qu'environ 30% de tous les courriels entrants sont mis en quarantaine, car ils ne peuvent être attribués par le système à aucune classe. Cet effet secondaire particulier nous fournit en même temps un point de vue unique pour la construction d'un ensemble de données de courrier électronique approximatif dans la zone grise, car la plupart du filtrage de courrier électronique standard a déjà été appliqué à ces messages. Intuitivement, cette zone est composée de spam dépassant plusieurs des mécanismes de protection antispam existants, mais également d'autres courriels en masse, tels que des lettres d'information abonnées, des courriels de notification ou des publicités commerciales. Il est intéressant de noter que la quantité de ce dernier type est dense dans ce domaine spécifique car l'un des effets d'un système de DR est qu'il authentifie les expéditeurs humains, mais pas les expéditeurs en masse légitimes et automatisés.

A.3.1 Introduction

Nous étudions la zone grise d'un système DR en adoptant une approche en trois phases reposant uniquement sur les informations disponibles dans les en-têtes de courrier électronique. La méthode proposée identifie les campagnes par courrier électronique et les catégorise en catégories de campagne sans analyse de contenu. Nous démontrons que la zone grise peut être réduite d'au moins 50% et que les campagnes identifiées peuvent en fait être automatiquement classées en quatre types: commercial, newsletters, spam de botnet et phishing / scam. Celles-ci représentent en outre 15% de l'ensemble des courriers électroniques système, et ne représentent que 0,2% des faux positifs - une mesure couramment utilisée dans le filtrage du courrier indésirable, car les messages ham mal classés peuvent être très coûteux en raison des coûts liés à leur perte pour les utilisateurs. Enfin, nous avons démontré qu'en utilisant une méthode de raffinement basée sur un graphique, les campagnes par courrier électronique légitimes peuvent souvent être identifiées en fonction des informations de l'expéditeur. Notre méthode de classification fonctionne bien sur toutes les campagnes, sauf sur les campagnes d'escroquerie / de phishing. Cela est dû au fait que ces campagnes ont des traits communs avec les légitimes. Cette idée nous a conduits à la suite de cette thèse - la recherche de l'arnaque et en particulier les numéros de téléphone et les arnaqueurs nigérians.

Notre méthode d'analyse par campagne nous permet d'éviter d'analyser les courriels d'utilisateurs personnels uniques et de nous concentrer plutôt sur les courriers électroniques en masse sans analyser leur contenu, mais uniquement leurs informations d'entête.

Les campagnes identifiées se composent de campagnes illégales (spam) envoyées souvent avec de mauvaises intentions et de campagnes automatisées en vrac légales, auxquelles le destinataire s'est probablement souscrit.

Nous démontrons en outre qu'il existe au moins quatre catégories identifiables de campagnes par courrier électronique: campagnes commerciales, newsletters, spam de botnet et arnaques / phishing, les campagnes commerciales constituant une grande partie de la zone grise et pouvant être identifiées.

A.3.2 Ensemble de Données

Ces courriels ont passé avec succès un certain nombre de filtres antispam, mais ils n'étaient déjà ni inscrits sur la liste blanche ni sur la liste noire du destinataire. En d'autres termes, ces messages n'étaient pas considérés comme du spam selon les techniques traditionnelles telles que: analyse antivirus, reverse DNS et liste noire DNS. De plus, les utilisateurs n'ont jamais eu de conversation préalable avec l'expéditeur. Par conséquent, nous pouvons considérer que cet ensemble de données a été pré-filtré à partir des courriels de spam et de spam évidents. Parfois, cet ensemble est appelé zone grise [50] qui stocke les courriers électroniques de classe incertaine.

En particulier, nous commençons par les regrouper en fonction des en-têtes de message. Nous utilisons des mots n-grammes de longueur décroissante (entre 70 et 8), avec une fenêtre coulissante qui permet de sauter diverses parties des sujets (une liste standard de mots vides, et un certain nombre de scripts personnalisés correspondant aux mots extraits n-grammes et les affecter à des grappe).

À l'aide de nos fonctionnalités dérivées, nous avons formé un classificateur binaire afin de séparer les clusters légitimes des spams. Des études antérieures ont montré que, sur les jeux de données de courriels indésirable, les classificateurs d'ensemble fonctionnaient mieux que les classificateurs uniques. Sur la base de cette conclusion, pour notre tâche de classification, nous avons décidé d'utiliser un classificateur supervisé d'ensemble de forêts aléatoires. Notre modèle a atteint un taux de précision de 97%, avec 0,9% de faux positifs (les campagnes légitimes étant classées à tort dans le spam) et 10% de faux négatifs (le spam étant à tort classifié comme légitimes). Ces taux suggèrent que l'ensemble des attributs que nous avons identifiés est efficace pour séparer les deux types de campagnes.

Enfin, nous utilisons une technique de raffinement basée sur des graphes pour augmenter davantage la couverture et la précision de notre classification.

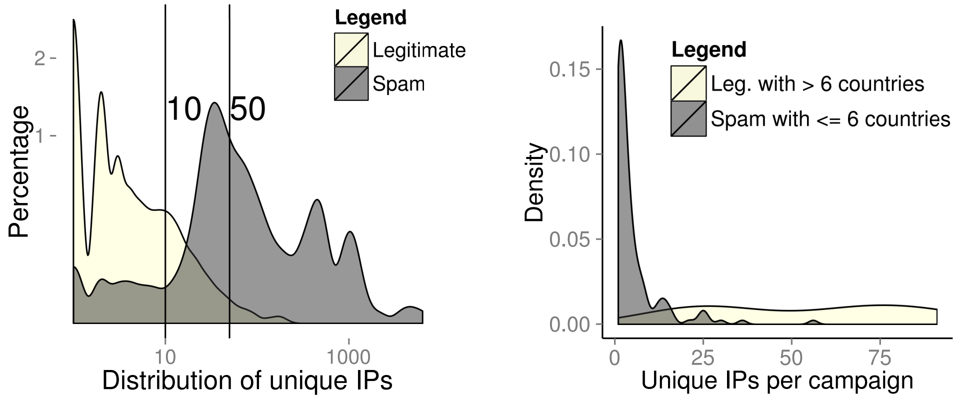
A.3.3 Résumé des résultats

A.3.3.1 Rôles des attributs dans la classification des courriels

Dans cette section, nous analysons les caractéristiques des campagnes de spam et des campagnes légitimes et comparons nos résultats à ceux présentés dans les études précédentes. Fait intéressant, les attributs les moins importants sont ceux du groupe 2, et en particulier le pourcentage de courriels déjà inscrits sur la liste blanche dans le cluster (ce qui confirme nos conclusions décrites plus loin sur la capacité des utilisateurs à juger les courriers électroniques dans la zone grise). Cependant, les plus importants sont les distributions des adresses de pays et IP, suivies du nombre moyen de destinataires et de la similarité des adresses courriels de l'expéditeur. Nous avons constaté que le nombre de pays d'origine était le paramètre le plus indicatif, alors que les recherches précédentes reposent souvent sur la variation des adresses IP en tant qu'indicateur puissant de l'activité du botnet et souvent utilisées comme mesure fiable pour détecter le spam [136]. La figure A.4 le montre en chiffres.

En examinant la répartition des pays IP, les résultats s'améliorent considérablement. Certaines campagnes légitimes ont de nombreux préfixes IP, mais proviennent de quelques pays. Cela pourrait être le résultat de la propagation de la même campagne commerciale par plusieurs sociétés de marketing par courrier électronique. En revanche, la grande majorité des campagnes de spam proviennent de plusieurs préfixes IP et de plusieurs pays.

Enfin, nous étudions de plus près ce groupe de campagnes de spam ayant peu d'origine. Le raffinement du graphique était inefficace pour eux non plus. À y regarder de plus près, ces cas correspondaient principalement à du phishing et à



(a) Logarithmique de la distribution de préfixe IP unique

(b) Répartition des campagnes après application du seuil de 6 pays sur nos données

Figure A.4: PrÃ©fixe IP et distribution des pays dans les campagnes

des escroqueries nigérianes. Plusieurs de ces campagnes sont envoyées en faible volume et pendant de courtes périodes à l'aide de comptes de messagerie Web, se cachant ainsi sous des adresses IP bénignes.

Les fonctionnalités axées sur le destinataire ne peuvent à elles seules être utilisées pour séparer de manière fiable le courrier indésirable des messages légitimes. Lorsque certains des messages d'une campagne sont rejetés, cela signifie que la liste de destinataires de l'expéditeur n'a pas été vérifiée correctement ou n'est pas à jour. Bien que les utilisateurs fassent parfois des fautes de frappe en fournissant leurs adresses courriels, un taux de rejet plus élevé ainsi que plusieurs destinataires constituent un bon indicateur de l'activité des spammeurs.

A.3.3.2 Interaction de l'utilisateur avec le résumé des courriels

La question à laquelle nous souhaitons répondre est de savoir si le fait que l'expéditeur résolve certains CAPTCHAs pourrait être un bon indicateur pour identifier des campagnes légitimes. Les données générées par l'utilisateur confirment que les utilisateurs sont susceptibles de commettre des erreurs lorsqu'ils jugent des courriels dans la zone grise. Ils ouvrent souvent même des courriels potentiellement dangereux, ignorant les risques de sécurité. Ces résultats sont conformes à ce qui a été testé dans une étude sur les utilisateurs menée par Onarlioglu et al. [127].

A.3.3.3 Campagnes de courriels

Dans cette section, nous discutons des principales catégories de campagnes de courriels que nous trouvons dans la zone grise. Tout d'abord, nous séparons le spam

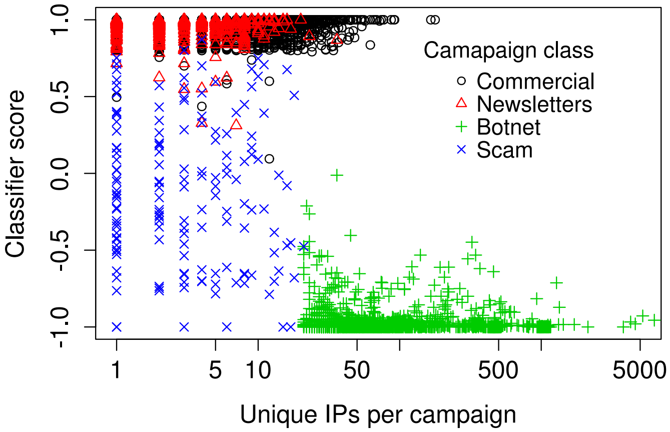


Figure A.5: Classification des campagnes

des courriels légitimes. Nous divisons ensuite le spam en deux catégories: celle générée par les grandes infrastructures (probablement envoyées par un botnet ou des machines infectées) des campagnes plus petites. Nous divisons les campagnes légitimes en deux groupes. D'un côté, nous avons des sociétés de marketing privées qui envoient des campagnes commerciales et se spécialisent dans la distribution de publicités légitimes dans des courriels en masse. D'autre part, nous avons des newsletters qui sont envoyées aux utilisateurs abonnés à des services Web ou à des listes de diffusion et à des notifications. Encore une fois, les premières sont fournies par de grandes infrastructures, tandis que les secondes sont normalement envoyées par un ensemble limité (et constant) d'adresses IP. La figure A.5 montre la distribution des campagnes classées.

Les campagnes commerciales constituent la plus grande catégorie de notre jeu de données et couvrent 42% des campagnes identifiées. Nous confirmons que ces messages sont principalement générés par des spécialistes du marketing par courriels professionnels qui envoient des publicités sollicitées ou non sollicitées. Sur leurs sites Web, ils soulignent à plusieurs reprises le fait qu'ils ne sont pas des spammeurs et qu'ils fournissent simplement à d'autres sociétés un moyen d'envoyer des courriels marketing dans les limites de la législation en vigueur. En fait, ils offrent également une procédure en ligne permettant aux utilisateurs de se retirer et d'être retirés des communications futures.

Les lettre d'information s'appuient principalement sur une infrastructure de courriel locale et réduite. En comparaison avec les campagnes commerciales, elles utilisent en moyenne trois fois moins d'adresses IP uniques. L'expéditeur est souvent la société qui distribue les courriers électroniques, avec généralement une plage d'adresses IP petite et fixe. Il couvre environ 30% du total des campagnes avec une taille moyenne de 90 courriels chacune. Les expéditeurs sont localisés géographiquement et ont des schémas d'envoi extrêmement cohérents. Ils utilisent généralement des listes de destinataires de courrier électronique valides et présentent les variations d'adresse IP, de pays et d'expéditeur les plus faibles. Seule l'utilisation de l'en-tête Unsubscribe semble incohérente, 39% seulement des courriels l'utilisant. Sans surprise, c'est également la catégorie la plus souvent sélectionnée sur liste blanche par les utilisateurs.

Les campagnes via Botnet ont des valeurs d'attribut très dynamiques, ce qui en fait la catégorie la plus facile à identifier de manière automatisée. Cette catégorie contient les campagnes les plus importantes, mais ne représente que 17% du total des campagnes. Ces campagnes ont la plus grande distribution géographique car elles sont envoyées par des ordinateurs infectés du monde entier: 172 réseaux uniques / 24 réseaux par campagne, répartis en moyenne sur 28 pays. Malgré les caractéristiques facilement reconnaissables de ces campagnes, les utilisateurs manifestent un intérêt étonnamment élevé pour ces courriels. Cette catégorie affiche le plus grand nombre de consultations de courrier électronique par campagne, ce qui suggère que les utilisateurs sont souvent curieux des produits mis en avant.

Les campagnes de fraude et de phishing trompent leurs victimes en utilisant des messages menaçants ou en tentant de les séduire avec des gains financiers énormes. Les caractéristiques de cette catégorie ressemblent en grande partie à celles des campagnes commerciales. Il est donc difficile de séparer automatiquement ces campagnes sans consulter le corps de la messagerie électronique (qui constitue le principal inconvénient de cette étude). Ce type de menace est plus susceptible d'être identifié par des techniques de détection basées sur le contenu. Les adresses IP à partir desquelles les CAPTCHA ont été résolues sont principalement situées dans des pays d'Afrique de l'Ouest, comme le Nigéria ou la Côte d'Ivoire.

A.3.4 Conclusions

Nous avons présenté un système permettant d'identifier et de classer les campagnes dans un ensemble de données réelles de courriels gris. En tant qu'approximation de ce sous-ensemble de courriels, nous avons choisi d'utiliser le dossier de quarantaine d'un filtre antispam de type défi-réponse, car il est déjà exempt de tout spam évident et de tout message personnel. Notre analyse de campagne a dévoilé les attributs de classe de campagne de courrier électronique les plus et les moins prédictifs.

Notre système pourrait être utilisé de différentes manières. Tout d'abord, cela peut aider à comprendre le fonctionnement des grandes campagnes commerciales, leur origine et leurs différences avec les autres courriers électroniques non sollicités. Cela pourrait également servir d'entrée pour placer automatiquement les campagnes marketing et les newsletters dans un dossier séparé, afin que les utilisateurs puissent différencier clairement ces messages des autres formes de spam.

Les utilisateurs de notre étude ouvraient souvent des courriels générés par des botnet et étaient particulièrement enclins aux erreurs lorsqu'ils traitaient de messages d'escroquerie ou de phishing; nous pensons qu'un dossier séparé dédié aux courriels en masse légitimes créerait une couche supplémentaire entre les utilisateurs et les expéditeurs de contenu malveillants, invitant les utilisateurs à rechercher d'abord dans le dossier en vrac plutôt que dans le dossier spam. Après avoir mené notre étude, une solution similaire a été mise en œuvre par Google dans les onglets Gmail [8].

De plus, notre technique pourrait servir d'outil de surveillance des campagnes par courrier électronique, permettant aux analystes de la sécurité de suivre les tendances des campagnes par courriel en masse, car les courriels en masse sont appelés à évoluer et à évoluer au fil du temps. Nous avons démontré qu'en utilisant une méthode de raffinement basée sur des graphes, les campagnes d'e-mails légitimes peuvent souvent être identifiées uniquement en fonction des informations sur l'expéditeur, et peuvent être classées comme des lettres d'information ou des publicités commerciales. Il s'agit là d'un résultat particulièrement prometteur dans le sens d'une étude empirique des courriers électroniques en masse légitimes. Il pourrait être utilisé pour créer des listes blanches IP de tels expéditeurs, voir des sociétés d'entreprises de gestion du marketing, des campagnes de gestion de campagnes par courrier électronique et des services de suivi.

Enfin, nous avons également découvert que notre méthode de classification fonctionnait bien pour toutes les campagnes, à l'exception de la fraude. Nous pensons que ce dernier bénéficierait largement d'une analyse de courrier électronique basée sur le contenu. C'est pourquoi nous proposons dans les chapitres suivants une nouvelle fonctionnalité permettant de corréler les messages frauduleux et de l'inclure dans un outil de classification multidimensionnel permettant d'identifier les campagnes frauduleuses.

A.4 Le rôle des numéros de téléphone dans les programmes de cybercriminalité

A.4.1 Introduction

Dans le chapitre précédent, nous avons constaté que notre méthode de classification fonctionnait bien pour toutes les campagnes, à l'exception des campagnes d'arnaque / phishing. En outre, ces campagnes sont beaucoup plus petites que les autres. Ce chapitre est une continuation de l'étude des courriels en zone grise et plus précisément de ceux qu'il nous a été demandé d'identifier à l'aide de méthodes analytiques basées uniquement sur les informations de l'en-tête de l'email.

Ce travail est basé sur un ensemble de données différent contenant du contenu de courrier électronique. Nous émettons l'hypothèse que l'identification de ces campagnes bénéficierait largement des fonctionnalités de messagerie basées sur le contenu. Nous nous concentrons sur le contenu des courriels frauduleux et étudions une fonctionnalité auparavant négligée: les numéros de téléphone. Cette fonctionnalité joue peut-être un rôle important dans ce secteur, car elle permet aux criminels de communiquer avec les victimes.

Notre objectif est de déterminer si l'utilisation de l'analyse du numéro de téléphone peut améliorer notre compréhension des marchés souterrains, des activités informatiques illégales ou de la cybercriminalité en général. Cette connaissance pourrait ensuite être adoptée par plusieurs mécanismes de défense, y compris des listes noires ou des méthodes heuristiques de spam avancées. Dans de nombreux cas de fraude, les numéros de téléphone jouent un rôle important. Par exemple, les autorités ont analysé les criminels sur la base de leurs numéros de téléphone sur des forums publics ou souterrains [24]. Dans d'autres cas de fraude en ligne [52], l'utilisation d'un numéro de téléphone peut rendre la fraude plus légitime pour la victime. Enfin, les fraudeurs utilisent souvent le téléphone pour escroquer les victimes [149].

Notre étude a trois objectifs principaux. Premièrement, nous souhaitons évaluer la fiabilité de l'utilisation d'une analyse automatisée des numéros de téléphone pour améliorer notre compréhension des marchés souterrains, des activités informatiques illégales et des cybercriminels en général. Deuxièmement, en examinant les données analysées, nous essayons de trouver différents modèles associés à des modèles commerciaux criminels récurrents. Enfin, nous corrélons les informations extraites et les enrichissons d'un processus de recherche géographique des HLR afin d'identifier automatiquement les communautés responsables des campagnes d'escroquerie au Nigéria.

A.4.2 Données expérimentales

A.4.2.1 Numéros de téléphone

Après un premier examen des données, nous avons observé une grande variabilité dans la qualité et la fiabilité des informations collectées. Pour mieux décrire ce phénomène, nous avons classé les numéros de téléphone dans deux directions: la difficulté de les extraire à partir de données brutes et leur fiabilité une fois bien extraits. Cependant, nous avons conclu qu'il était très difficile de reconnaître et d'extraire correctement les nombres d'un flux de données brutes, ce qui concorde avec les résultats obtenus dans [140].

A.4.2.2 Données de courrier électronique

Nous avons sélectionné le site *419scam.org*, financé par la communauté, car il contient un grand ensemble de données contenant des rapports frauduleux bien formatés. Cet ensemble de données a été collecté, filtré et pré-traité manuellement de janvier 2009 à août 2012. Il comprend des métadonnées sur chaque entrée.

Dans notre ensemble de données SCAM, nous avons identifié au total 67 244 numéros de téléphone normalisés uniques. Parmi eux, 34 424 étaient des numéros PRS du Royaume-Uni (51% du total) et les 32 820 restants des numéros PRS non britanniques (49% du total). Sur les 32 820 numéros PRS non britanniques, il y avait 29 685 numéros de téléphones mobiles.

Enfin, nous avons collecté des informations supplémentaires sur les numéros de téléphone mobile en effectuant une recherche HLR. Les HLR sont des bases de données gérées par des opérateurs de téléphonie mobile contenant des informations sur l'état actuel d'un numéro de téléphone. Cela permet de savoir si un numéro de téléphone mobile est toujours actif et s'il est en itinérance dans un pays étranger. En effectuant périodiquement une recherche HLR pour un numéro de téléphone mobile donné, nous pouvons avoir un aperçu de l'évolution de l'état de son réseau.

A.4.3 Résumé des résultats

A.4.3.1 Modèles commerciaux de fraude

Les numéros de téléphone Premium souvent utilisés par les fraudeurs peuvent être classés en quatre catégories: numéros abrégés nationaux, services nationaux premium, services internationaux et services de numérotation personnelle au Royaume-Uni.

En comparant manuellement ces opérateurs et ceux des six autres opérateurs, nous avons constaté que des escrocs préféraient les opérateurs qui:

- Ont un service d'enregistrement et de configuration en ligne;
- Fournit une API pour automatiser le processus d'inscription;
- Offre un renvoi d'appel international bon marché ou gratuit;
- Offre un programme de remise en argent pour payer l'inscrit pour chaque appel entrant.

A.4.3.2 Analyse dynamique des numéros de téléphone frauduleux

Afin de comprendre l'organisation et la dynamique des communautés frauduleuses identifiées dans les sections précédentes, nous avons effectué des recherches périodiques HLR (section 5.4) parmi les numéros de téléphone mobiles précédemment extraits. Avec cette expérience, nous visons à comprendre la fréquence à laquelle les numéros de téléphone mobile sont utilisés dans d'autres pays (c.-à-d. en itinérance) et dans le temps.

Nous avons finalement sélectionné les 1 333 numéros de téléphone collectés récemment. Nous avons vérifié que la période de deux mois sélectionnée est représentative de la situation générale. La population de téléphones mobiles joignables, itinérants ou désactivés est comparable dans les deux jeux de données, mais les numéros de téléphone récemment utilisés sont plus susceptibles d'être en ligne au moment de notre requête HLR. Cela confirme le fait qu'après un certain temps, certains numéros de téléphone risquent d'être supprimés ou remplacés. Fait intéressant, très peu de personnes (seulement 9 en réalité) erraient dans un pays étranger.

En examinant les modifications de l'attribut d'état du réseau, nous avons constaté qu'environ la moitié des chiffres avaient un état OK constant. Cela montre que les fraudeurs utilisent des numéros de téléphone pendant de longues périodes en les maintenant en ligne la plupart du temps. L'escroc moyen maintient le téléphone allumé la plupart du temps et seuls 89 numéros étaient éteints plus de 75% du temps. Enfin, selon l'attribut d'itinérance, seuls 50 téléphones ont été utilisés dans un pays différent au cours de notre évaluation (c.-à-d. l'itinérance). Cela montre clairement deux clusters - un en Afrique et un en Europe - avec une petite intersection des deux. Le Nigéria est toujours un pays clé pour ce type d'affaires, avec environ 80% de l'itinérance qui lui appartient. Cela confirme encore notre hypothèse selon laquelle des groupes répartis existent et fonctionnent de manière coordonnée et en collaboration à partir de plusieurs pays.

A.4.3.3 Criminels derrière le téléphone

Nous avons utilisé le jeu de données SCAM pour évaluer l'utilisation des numéros de téléphone afin d'identifier les criminels, étudier leur comportement et déplier la structure et la taille de leurs réseaux. Les fraudeurs sont connus pour fournir de vrais numéros de téléphone, auxquels leurs victimes peuvent les joindre. Par conséquent, cet ensemble de données est moins pollué par des nombres falsifiés ou usurpés, ce qui rend nos résultats et nos conclusions plus fiables.

Au niveau mondial, nous avons identifié 102 communautés [43] et 79 sous-graphiques. Le graphique montre des relations intéressantes. Premièrement, les fraudeurs semblent réutiliser une adresse électronique donnée pour envoyer des messages frauduleux, chaque message contenant des numéros de téléphone différents. Deuxièmement, un numéro de téléphone donné semble être réutilisé dans plusieurs messages d'escroquerie ou en combinaison avec plusieurs adresses électroniques différentes. Nous notons en particulier que 37% des numéros de téléphone ont été réutilisés par plus d'un fraudeur.

Une autre façon de voir les communautés consiste à les classés par leur pays et leur taille. La figure 5.4 montre comment sont organisées les huit plus grandes communautés. Toutes ces communautés dépendent des numéros premium britanniques (pour au moins 29% de leurs numéros de téléphone) et des numéros des opérateurs nigérians. En outre, ces communautés utilisent des numéros de téléphone cellulaire dans plusieurs pays européens et africains.

A.4.4 Conclusions

Nous avons ensuite discuté d'un certain nombre de modèles commerciaux communs que nous avons observés au cours de nos expériences. Nos résultats montrent qu'un nombre restreint d'opérateurs de téléphonie mobile sont utilisés pour fournir la majorité des numéros liés à la fraude. Cela suggère que certains opérateurs sont préférés aux fraudeurs.

Nous avons ensuite analysé le rôle des numéros de téléphone des fraudes de type 419. Nous avons utilisé des recherches HLR sur les numéros de téléphone portables frauduleux pour vérifier si les fraudeurs cessaient d'utiliser leurs numéros de téléphone en les éteignant après leur publication, mais nous avons constaté que beaucoup d'entre eux restaient actifs et continuaient à les utiliser pendant une longue période (84%). Cette constatation suggère que les numéros de téléphone seraient un bon moyen d'identifier les arnaques et d'identifier les groupes d'arnaqueurs, par exemple les campagnes d'arnaques. Nous avons également identifié des groupes d'escrocs, créé des liens étroits entre des acteurs apparemment non liés et analysé leur répartition géographique. Une observation importante est qu'au cours de la période de notre expérience avec les recherches HLR, certains arnaqueurs se déplaçaient dans différents pays (le pays d'itinérance le plus populaire parmi les arnaqueurs semble être le Nigéria). Par conséquent, notre prochaine étude sur les campagnes d'escroquerie reposera sur l'observation selon laquelle la majorité des numéros de téléphone utilisés par des escrocs sont des numéros de téléphone mobile utilisés sur de longues périodes et peuvent donc être utilisés comme une caractéristique d'identification de l'arnaque.

A.5 Approche basée sur le contenu pour les campagnes d'escroquerie au Nigéria

A.5.1 Introduction

Au cours de l'étude de la zone grise, nous avons identifié une catégorie de campagnes de courrier électronique dans lesquelles notre méthode de classification et de catégorisation affichait une performance médiocre. Dans ce chapitre nous avons adapté notre approche en se concentrant sur le contenu des messages et utiliser les numéros de téléphone pour étudier les campagnes frauduleuses au Nigéria. En particulier, nous présentons un aperçu de plusieurs campagnes, en montrant leurs caractéristiques et leurs relations. Enfin, nous décrivons quelques exemples pour mieux étudier le modus operandi des criminels.

L'analyse montre qu'il y a assez peu de grandes campagnes et celles que nous avons identifiées ont souvent des liens avec le Nigéria en tant que pays suggérant que ces cybercriminels tendent à former des groupes de criminels répartis.

Nous utilisons un nouvel algorithme décisionnel multicritères pour regrouper efficacement les courriels frauduleux partageant certains points communs, même en présence de fonctionnalités plus volatiles. En raison de ces points communs, les courriels frauduleux provenant du même fraudeur peuvent être regroupés, ce qui nous permet de mieux comprendre les campagnes frauduleuses.

Dans notre analyse, nous avons identifié plus de 1 000 campagnes différentes et, pour la plupart, les numéros de téléphone représentent la pierre angulaire qui nous permet de relier les différentes parties. Nous avons également découvert des campagnes à plus grande échelle (appelées "macro-grappes"), qui consistent en des opérations frauduleuses faiblement interconnectées. Nous pensons qu'ils sont probablement le reflet de différentes arnaques orchestrées par les mêmes groupes criminels, car nous observons que les mêmes numéros de téléphone ou comptes de messagerie sont réutilisés dans différentes sous-campagnes.

Comme démontré par nos expériences, nos méthodes et résultats pourraient être utilisés pour identifier de manière proactive de nouvelles opérations frauduleuses (ou des variantes des précédentes) en associant rapidement une nouvelle fraude à des campagnes en cours. Nous pensons que cela pourrait faciliter le travail des organismes chargés de l'application de la loi dans la poursuite des fraudeurs. Notre approche pourrait également servir à améliorer l'analyse forensic et les enquêtes sur d'autres systèmes de cybercriminalité en enregistrant et en enquêtant divers groupes de cybercriminels sur la base de leurs activités en ligne.

A.5.2 Configuration expérimentale

Les données utilisées dans cette étude sont les mêmes que dans le chapitre précédent, provenant de l'agrégateur *419scam.org* et sont enrichies de données de numéro de téléphone supplémentaires.

L'ensemble de données résultant consiste en 36 761 messages avec 11 768 numéros de téléphone uniques. Les messages initiaux sont également étiquetés avec une catégorie d'escroquerie.

Pour identifier les groupes de courriels frauduleux susceptibles de faire partie d'une campagne orchestrée par le même groupe de personnes, nous avons regroupé tous les messages frauduleux à l'aide de TRIAGE, un framework logiciel d'exploration de données de sécurité qui tire parti de l'analyse de données multicritères pour regrouper des événements. sur des sous-ensembles d'éléments communs (caractéristiques).

1 040 grappes ont été identifiées et comprennent au moins 5 courriels frauduleux corrélés par diverses combinaisons de caractéristiques. En raison de la manière

dont ces clusters sont générés (c'est-à-dire l'agrégation multicritères), nous émettons l'hypothèse que ces clusters de courriels représentent différentes campagnes, potentiellement organisées par les mêmes individus.

Table A.2: Global statistics for the top 250 clusters

Statistique	**Moyenne**	**Médiane**	**Maximum**
Nr courriels	38	28	376
Nr de	13.9	9	181
Nr répondre	6.2	5	56
Nr sujets	9.9	7	114
Nr téléphones	2.5	2	34
Durée (en jours)	396	340	1,454
Nr dates (distinctes)	27.9	22	259
La compacité	2.5	2.4	5.0

Le tableau 6.4 fournit des statistiques globales calculées sur les 250 plus grandes campagnes frauduleuses. Dans plus de la moitié de ces campagnes, les fraudeurs n'utilisent que deux numéros de téléphone distincts, mais ils utilisent toujours plus de cinq boîtes aux lettres différentes pour obtenir les réponses de leurs victimes. La plupart des campagnes d'escroquerie durent assez longtemps.

A.5.3 Résumé des résultats

A.5.3.1 Caractérisation des campagnes

Pour la caractérisation des campagnes, nous avons utilisé deux approches principales: l'outil de visualisation de graphe spécialisé du projet VIS-SENSE [1], et l'identification des campagnes de macro-clusters en recherchant des clusters faiblement interconnectés (aidant à identifier des campagnes organisées à grande échelle). Dans ce dernier cas, nous avons uniquement utilisé des adresses électroniques et des numéros de téléphone, les autres attributs n'étant pas considérés comme des informations personnellement identifiables. Nous avons recherché des grappes partageant au moins une adresse électronique et / ou un numéro de téléphone et nous avons utilisé ces informations pour créer des macro-grappes.

La première partie des résultats décrit des exemples de campagnes identifiées étroitement liées. Ils ont tendance à avoir fonctionné pendant des périodes assez longues (un an et demi), à avoir changé de sujet au fil du temps et à partir de grappes d'adresses courriels / de numéros de téléphone réutilisés.

Nous avons ensuite expérimenté la réorganisation de grappes en macro-grappes. En conséquence, nous avons identifié un ensemble de 845 grappes isolées et un

[1] Le projet VIS-SENSE: `http://www.vis-sense.eu`

Table A.3: Macro-clusters, valeurs moyennes des attributs

ID	Nr. of cmpg.	Tel.	Mbox	Sbj.	Dur.	Ctry	Topics
1	14	44	677	223	4 y.	4	Lottery, lost funds, investments
2	43	163	1,127	463	4 y.	7	Lottery, banks, diplomats, FBI
3	6	18	128	80	4 y.	4	Lottery
4	5	8	111	51	3,5 y.	2	Packaging, Guiness lottery, loans
5	6	7	201	96	1 y.	1	Microsoft lottery, UPS & WU delivery, lost funds
6	4	7	82	33	2 y.	1	Lottery, lost payments

autre ensemble de 195 grappes connectées, cette dernière comprenant 62 macro-grappes. Les caractéristiques des 6 principales macro-campagnes sont présentées dans le tableau A.3.

Ces macro-clusters sont particulièrement intéressants car ils consistent en un ensemble de campagnes frauduleuses qui semblent être faiblement interconnectées et qui pourraient donc également être orchestrées par les mêmes cybercriminels. En fait, les algorithmes de groupement considéraient que les liens entre différents groupes d'escroquerie étaient trop faibles, en raison du schéma de décision et des seuils définis en tant que paramètres; ces différentes opérations d'escroquerie ont ensuite été regroupées dans des groupes distincts.

Nous examinons également les origines spécifiques des 6 principales macro-campagnes. Trois d'entre elles sont presque exclusivement basés en Afrique, en outre dans un ou deux pays seulement, en supposant que les numéros de téléphone britanniques anonymes sont très probablement utilisés par des fraudeurs situés en Afrique et cachés derrière ces numéros de téléphone européens. Les trois autres sont plus orientés vers l'Europe, mais entretiennent des liens étroits avec le Nigéria et le Bénin.

A.5.4 Conclusions

Nous avons identifié plus de mille campagnes frauduleuses à l'aide d'une technique de classification multidimensionnelle que nous avons utilisée pour regrouper des courriels similaires. Pour notre analyse, nous nous sommes ensuite concentrés sur les 250 plus grandes campagnes. Notre méthode repose sur des fonctionnalités extraites du contenu du courrier électronique qui sont très spécifiques à la fraude au Nigéria: adresses électroniques et numéros de téléphone.

Nous avons montré que le mode de fonctionnement et l'orchestration de telles campagnes diffèrent des campagnes de spam traditionnelles envoyées via botnet. Notre analyse a révélé une grande diversité de méthodes d'orchestration d'escroquerie, montrant que les fraudeurs peuvent travailler sur différents sujets au sein d'une même campagne, se faisant donc probablement concurrence sur des sujets d'arnaques branchés. De plus, les sujets des campagnes changent beaucoup plus souvent que les adresses électroniques ou les numéros de téléphone. En général, les campagnes sont diverses et orchestrées de différentes manières, certaines des plus grandes campagnes étant des campagnes multinationales couvrant plusieurs pays. Nous en avons même identifié certaines qui durent depuis plus de 3 ou 4 ans et dont certains courriels et numéros de téléphone étaient encore réutilisés. Dans le même temps, les fraudeurs semblent envoyer de très faibles volumes de courriels par rapport aux spammeurs. Enfin, nous avons découvert l'existence de macro-campagnes, de groupes de campagnes faiblement liées qui sont probablement dirigées par les mêmes personnes. Nous avons constaté que certaines de ces macro-campagnes sont géographiquement réparties sur plusieurs pays, africains et européens.

Sur la base des résultats, nous concluons qu'il est difficile d'identifier de telles campagnes uniquement à partir des données d'en-tête. Par conséquent, de telles campagnes nécessitent des fonctionnalités plus spécifiques à ce type de fraude en ligne, comme celles étudiées dans ce chapitre. Cette approche pourrait être utile aux équipes d'analyse forensic et d'enquêteurs pour les aider à étudier les schémas de cybercriminalité de divers groupes de cybercriminels.

A.6 Conclusions

Dans ce travail, nous nous sommes principalement concentrés sur l'étude de l'analyse de la zone grise dans le système de filtrage du courrier électronique, en utilisant un système de défi-réponse (DR) du monde réel comme approximation de la zone grise. Nous avons abordé le problème d'un point de vue empirique et analytique, ce qui nous a permis d'apporter des réponses à l'énoncé du problème de cette thèse.

Le premier objectif était d'évaluer l'impact et l'efficacité d'un filtre DR en tant que filtre anti-spam pour courrier électronique. Une étude empirique et l'analyse de 6 mois de données provenant de 47 entreprises publiques et privées ont permis de résoudre ce problème. On ne savait pas grand-chose auparavant de l'efficacité réelle du système de répression en cas de déploiement dans le monde réel, ni de son impact. Dans nos expériences, nous avons évalué: (i) la pollution du trafic par le système due aux défis envoyés; (ii) le nombre de défis générés par le système; (iii) le délai de remise des messages introduit par la phase de quarantaine; (iv) le ratio faux négatif tel que perçu par les utilisateurs du système; (v) les conséquences de la mise sur liste noire du système en raison de la découverte de spamtraps. L'étude du système et de sa zone de quarantaine nous a permis de conclure que cette zone constitue une bonne approximation de la zone grise de la messagerie, car elle exclut déjà la plupart des messages évidents de spam et de ham.

Pour étudier la zone grise, nous nous sommes concentrés sur l'analyse des courriels mis en quarantaine par le système DR. Nous avons analysé six mois de données au cours desquelles nous avons intercepté environ 3,3 millions de courriels en quarantaine. Pour effectuer une analyse empirique de la zone de quarantaine, nous avons proposé une méthode d'identification et de classification des campagnes par courrier électronique, basée uniquement sur les informations d'en-tête de courrier électronique. Nous avons dévoilé les attributs prédictifs les plus et les moins bien classés de la campagne, démontrant ainsi que les méthodes précédemment proposées généreraient des taux de faux positifs pouvant atteindre 10%. De plus, nous avons été les premiers à analyser la zone grise en détail et à proposer une méthode permettant de classer 50% de la zone grise (avec une possibilité d'extension jusqu'à 63%). Ainsi, la zone grise initiale peut être réduite de moitié, ce qui revient à classer 15% supplémentaires du total des courriers électroniques entrants.

Lors de l'étude des campagnes par courrier électronique dans la zone grise, nous avons regroupé les campagnes en quatre catégories: commercial, bulletins d'information, spam par botnet et phishing / arnaque. Un grand nombre de courriels appartenaient à des campagnes commerciales et à des lettres d'information. Les deux catégories ont représenté 72% du total des campagnes. À notre connaissance, il s'agissait de la première étude capable d'identifier automatiquement ces classes de campagnes. Cependant, après avoir mené notre étude, une solution similaire a été publiée par Google, les onglets dans Gmail, qui catégorise les bulletins d'informations, les notifications et d'autres contenus de courriels commerciaux en catégories distinctes.

Enfin, notre méthode d'analyse semblait inefficace contre les campagnes de phishing / escroquerie, car elle présente souvent un comportement similaire à celui des campagnes commerciales. De plus, elle utilise des comptes de messagerie Web pour se cacher derrière l'infrastructure des fournisseurs de messagerie Web. Ainsi, la dernière étude était une conséquence de la limitation ci-dessus. En outre, nous avons conclu que pour identifier les campagnes de phishing / arnaque et les classer avec précision, nous devons disposer de fonctionnalités plus descriptives, accessibles à partir du contenu du courrier électronique. Pour cette raison, nous avons proposé d'utiliser des numéros de téléphone, qui sont particulièrement spécifiques à 419 scam. Nous avons démontré empiriquement, en comparant avec d'autres jeux de données, que cette fonctionnalité est particulièrement utile pour identifier ces campagnes frauduleuses. Comme les numéros de téléphone eux-mêmes peuvent fournir des informations supplémentaires, telles que le pays, l'opérateur, l'état du téléphone, nous avons enrichi nos données expérimentales avec ces informations. Nous l'avons utilisé dans l'analyse du modus operandi des fraudeurs et de leurs distributions géographiques. Nous nous sommes appuyés sur un outil de classification multidimensionnel, TRIAGE [158], pour regrouper des courriels similaires. Nos résultats ont montré que le Nigéria, en tant que pays, joue un rôle particulièrement important dans le secteur des escroqueries et que les numéros de téléphone sont plus performants que les adresses électroniques pour identifier les campagnes 419 scam frauduleuses.

Nous pensons que les résultats de nos recherches pourraient être utilisés plus avant pour améliorer l'analyse de la zone grise de la messagerie et pourraient être particulièrement utiles pour l'identification automatisée des campagnes de messagerie légitimes, par exemple pour créer des listes blanches d'expéditeurs de courriels en masse légitimes. Une autre partie de nos résultats pourrait servir aux équipes d'analyse forensic et d'enquêteurs qui étudient les schémas de cybercriminalité de divers groupes de cybercriminels.

A.6.1 Avenir

L'analyse approfondie des courriers électroniques dans la zone grise a ouvert d'autres perspectives potentielles pour la recherche et les défis à relever dans le futur. Dans un premier temps, la prochaine étape pourrait consister à enrichir le classificateur avec les spams rejetés et les emails commerciaux / de newsletters des utilisateurs afin d'élargir la couverture des deux types de campagnes d'email. Cela permettrait en outre de réduire la zone grise et de créer des listes blanches plus importantes pour les expéditeurs d'envoi de courriels en masse légitimes. Deuxièmement, la méthode proposée pourrait être étendue pour utiliser également les données de contenu de courrier électronique afin d'identifier davantage de campagnes de courrier électronique. Cela pourrait augmenter la couverture de la zone grise.

Troisièmement, nous avons montré qu'en utilisant une méthode de raffinement basée sur un graphe, les campagnes par courrier électronique légitimes peuvent souvent être identifiées uniquement en fonction des informations de l'expéditeur et peuvent être classées en bulletins d'information ou en publicités commerciales. Il s'agit là d'un résultat particulièrement prometteur dans le sens d'une étude empirique des courriers électroniques en masse légitimes. Il pourrait également être utilisé pour créer des listes blanches IP de tels expéditeurs, la gestion de campagnes par courrier électronique et des services de suivi, qui reposent souvent sur des plages d'adresses IP dédiées.

Quatrièmement, comme cela a été constaté lors des expériences sur la zone de quarantaine, il est particulièrement difficile d'analyser les campagnes envoyées par les fournisseurs de messagerie Web en raison des abus potentiels des criminels. Nous avons signalé que près de 40% des courriels de la zone grise provenaient de comptes de fournisseurs de messagerie Web. Ce numéro est représentatif d'attirer l'attention sur le problème du spam provenant de comptes de messagerie Web.

Enfin, l'une des préoccupations potentielles de ces listes blanches serait celle des spammeurs de raquettes. À l'instar de certaines sociétés de marketing, elles s'appuient également sur des plages d'adresses IP dédiées. Elles pourraient donc figurer dans les listes des expéditeurs en nombre légitimes. Une façon de résoudre ce problème pourrait être d'ajouter des fonctionnalités externes supplémentaires au jeu de fonctionnalités actuel, telles que les plaintes d'utilisateurs pour des adresses IP, le nombre de fois où l'expéditeur a récemment touché le spamtraps, ou un historique de l'utilisation des adresses IP.

Bibliography

[1] 419 Scam: Fake Lottery Fraud Phone Directory. http://www.419scam.org/419-by-phone.htm. 20, 93, 95, 96, 118

[2] A Snowshoe Winter: Our Discontent with CAN-SPAM. http://www.spamhaus.org/news/article/641. 13

[3] Barracuda. http://www.barracudacentral.org/. 57

[4] CAN-SPAM Act: Controlling the Assault of Non-Solicited Pornography And Marketing Act of 2003. 10, 12

[5] Challenge-response anti-spam systems considered harmful. http://linuxmafia.com/faq/Mail/challenge-response.htm. 5, 30, 39, 139, 140

[6] Composite Blocking List. http://cbl.abuseat.org/. 57

[7] Glossary. http://www.spamhaus.org/faq/section/Glossary. 1, 21, 135

[8] Inbox tabs and category labels. http://gmailblog.blogspot.fr/2013/05/a-new-inbox-that-puts-you-back-in.html. 4, 85, 132, 153

[9] Locating mobile phones. http://events.ccc.de/congress/2008/Fahrplan/attachments/1262_25c3-locating-mobile-phones.pdf. 94

[10] MailChimp. http://mailchimp.com/. 11, 65, 86, 133

[11] NNPC - Worldwide National Numbering Plans Collection. http://bsmilano.it/aspx/ENG/MainFrameSet_ENG.aspx?Page=NumberingPlans_ENG.aspx. 94

[12] ORBITrbl. http://www.orbitrbl.com/. 57

[13] Passive Spam Block List. http://psbl.surriel.com/. 57

[14] Patent US7917655: Method and system for employing phone number analysis to detect and prevent spam and e-mail scams. http://www.patentlens.net/patentlens/patent/US_7917655/en/. 20, 90

[15] Premium Rate Services Network Operators Contact. http://www.phonepayplus.org.uk/For-Business/Setting-up-a-premium-rate-service/Network-operator-contacts.aspx. 95

[16] Routo Messaging Bulk SMS services and HLR lookups. http://www.routomessaging.com/. 94

[17] Sender Score IP Reputation. http://www.senderscore.org/. 21

[18] Sendio. http://www.sendio.com/. 29

[19] Spam and Open-Relay Blocking System. http://www.sorbs.net/. 56, 57

[20] Spam Arrest. http://www.spamarrest.com/. 29

[21] Spam Cannibal. http://www.spamcannibal.org/. 57

[22] SpamCop. http://www.spamcop.net/. 56, 57

[23] Statistics and Facts About Spam. http://www.spamlaws.com/spam-stats.html. 2, 14

[24] The Koobface malware gang exposed. http://www.sophos.com/medialibrary/PDFs/other/sophoskoobfacearticle_rev_na.pdf. 88, 154

[25] The Spam Definition and Legalization Game. http://www.spamhaus.org/news/article/9/. 10

[26] The spamhaus project. http://www.spamhaus.org/. 4, 13, 26, 56, 57, 138

[27] Total Block. http://www.totalblock.net/. 29

[28] UK Ofcom Numbering Site. http://www.ofcom.org.uk/static/numbering/index.htm. 95

[29] UK Phone Info Codes Allocations Lookup. http://www.ukphoneinfo.com/s7_code_allocations.php?GNG=70. 95

[30] Why are auto responders bad? http://spamcop.net/fom-serve/cache/329.html. 5, 30, 39, 139, 140

[31] Directive 2002/58 on Privacy and Electronic Communications, concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). , 2002. 10, 12, 13

[32] Microsoft Research Penny Black. http://research.microsoft.com/en-us/projects/PennyBlack/, 2004. 28

[33] Microsoft Security Intelligence Report. http://www.microsoft.com/security/sir/archive/default.aspx, 2008-2012. 14, 15, 16

[34] Domain-based Message Authentication, Reporting and Conformance, DMARC. http://www.dmarc.org/, 2013. 26

[35] Rohan Amin, Julie Ryan, and Johan van Dorp. Detecting Targeted Malicious Email. *Security & Privacy, IEEE*, 10(3):64–71, 2012. 19

[36] David S Anderson, Chris Fleizach, Stefan Savage, and Geoffrey M Voelker. *Spamscatter: Characterizing internet scam hosting infrastructure*. PhD thesis, University of California, San Diego, 2007. 35

[37] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras, and C.D. Spyropoulos. An evaluation of naive bayesian anti-spam filtering. 2000. 23

[38] Prakasam Appavu Siva, Jin Yu, Skudlark Ann, Jiang Nan, Hsu Wen-Ling, and Jacobson Guy. Increased Smart Device Penetration Brings Malware Vulnerability: Methods for Detecting Malware in a Large Cellular Network. 2011. 90

[39] Satanjeev Banerjee and Ted Pedersen. The Design, Implementation and Use of the Ngram Statistics Package. *ITPCL*, 2003. 67

[40] A. Bergholz, J.H. Chang, G. Paaß, F. Reichartz, and S. Strobel. Improved phishing detection using model-based features. In *Proc. of Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2008. 4, 23, 138

[41] D. Bernstein. Internet mail 2000 (IM2000). 29

[42] Enrico Blanzieri and Anton Bryl. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1):63–92, 2008. 10

[43] 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. 97, 156

[44] J. Buchanan and A. J. Grant. Investigating and Prosecuting Nigerian Fraud. *High Tech and Investment Fraud*, 2001. 20, 90

[45] Elie Bursztein, Steven Bethard, Celine Fabry, John C Mitchell, and Dan Jurafsky. How good are humans at solving CAPTCHAs? a large scale evaluation. In *Security and Privacy (SP), 2010 IEEE Symposium on*, pages 399–413. IEEE, 2010. 30

[46] Pedro Calais, Douglas EV Pires, Dorgival Olavo Guedes Neto, Wagner Meira Jr, Cristine Hoepers, and Klaus Steding-Jessen. A Campaign-based Characterization of Spamming Strategies. In *CEAS*, 2008. 17, 32, 35

[47] CAUSE. CAUCE Statement on CAN SPAM Act. `http://www.cauce.org/2004/12/cauce-statement-on-can-spam-act.html`, 2004. 12

[48] Coalition Against Unsolicited Commercial Email (CAUSE). `http://www.cauce.org/`. 12

[49] Madhusudhanan Chandrasekaran, Krishnan Narayanan, and Shambhu Upadhyaya. Phishing email detection based on structural properties. In *NYS Cyber Security Conference*, pages 1–7, 2006. 18

[50] Ming-Wei Chang, Wen-Tau Yih, and Robert Mccann. Personalized spam filtering for gray mail. *Proc. of Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2008. 10, 22, 37, 66

[51] Abdur Chowdhury, Ophir Frieder, David Grossman, and Mary Catherine McCabe. Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems (TOIS)*, 20(2):171–191, 2002. 23, 32

[52] Nicolas Christin, Sally S. Yanagihara, and Keisuke Kamataki. Dissecting one click frauds. CCS '10, pages 15–26, New York, NY, USA, 2010. ACM. 6, 88, 90, 95, 154

[53] James Clark, Irena Koprinska, and Josiah Poon. A neural network based approach to automated e-mail classification. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, pages 702–705. IEEE, 2003. 24

[54] Duncan Cook, Jacky Hartnett, Kevin Manderson, and Joel Scanlan. Catching spam before it arrives: domain specific dynamic blacklists. In *Proceedings of the 2006 Australasian workshops on Grid computing and e-research - Volume 54*, ACSW Frontiers '06, pages 193–202, Darlinghurst, Australia, Australia, 2006. Australian Computer Society, Inc. 88

[55] Gordon V Cormack and Thomas R Lynam. Online supervised spam filter evaluation. *ACM Transactions on Information Systems (TOIS)*, 25(3):11, 2007. 25

[56] Marco Cova, Corrado Leita, Olivier Thonnard, Angelos D. Keromytis, and Marc Dacier. An Analysis of Rogue AV Campaigns. In *Proceedings of the 13th international conference on Recent advances in intrusion detection*, RAID'10, pages 442–463, Berlin, Heidelberg, 2010. Springer-Verlag. 33, 112

[57] Lorrie Faith Cranor and Brian A LaMacchia. Spam! *Communications of the ACM*, 41(8):74–83, 1998. 1, 9, 135

[58] D. Crocker. RFC822: Standard for ARPA Internet Text Messages. *Retrieved April*, 7:2008, 1982. 41

[59] D. Fallows. Spam: How it is hurting email and degrading life on the Internet. Pew Internet & American Life Project, 2003. 3, 14, 22, 37, 137

[60] M. Delany. Domain-Based Email Authentication Using Public Keys Advertised. the DNS (DomainKeys)", RFC 4870, 2007. 4, 25, 26, 138

[61] Direct Marketing Association. Response Rate Report, 2012. 11

[62] M. Dodge. Slams, crams, jams, and other phone scams. *Journal of Contemporary Criminal Justice*, 17:358–368, 2001. 6, 90

[63] Emmanuel Dreyfus. Milter-greylist for Sendmail. `http://hcpnet.free.fr/milter-greylist/`. 28

[64] H. Drucker, D. Wu, and V.N. Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048–1054, 1999. 4, 23, 24, 138

[65] Z. Duan, Y. Dong, and K. Gopalan. Diffmail: A Differentiated Message Delivery Architecture to Control Spam. Technical report, 2004. 29

[66] Z. Duan, K. Gopalan, and Y. Dong. Push vs. pull: Implications of protocol design on controlling unwanted traffic. In *Proc. of USENIX Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI 2005)*, 2005. 29

[67] Eve Edelson. The 419 scam: information warfare on the spam front and a proposal for local filtering. *Computers & Security*, 22(5):392–401, 2003. 88

[68] EmailTray. Email Spam Trends at a Glance: 2001-2012. `http://www.emailtray.com/blog/email-spam-trends-2001-2012/`, 5 June, 2012. 2, 136

[69] Aaron Emigh. The Crimeware Landscape: Malware, Phishing, Identity Theft and Beyond. *J. Digital Forensic Practice*, 1(3):245–260, 2006. 88

[70] William Enck, Machigar Ongtang, and Patrick McDaniel. On lightweight mobile phone application certification. CCS '09, pages 235–245, New York, NY, USA, 2009. ACM. 90

[71] D. Erickson, M. Casado, and N. McKeown. The Effectiveness of Whitelisting: a User-Study. In *Proc. of Conference on Email and Anti-Spam*, 2008. 4, 29, 30, 39, 47, 50, 59, 138, 144

[72] Holly Esquivel, Tatsuya Mori, and Aditya Akella. Router-level spam filtering using TCP fingerprints: Architecture and measurement-based evaluation. In *Proceedings of the Sixth Conference on Email and Anti-Spam (CEAS)*, 2009. 28

[73] Hanno Fallmann, Gilbert Wondracek, and Christian Platzer. Covertly probing underground economy marketplaces. DIMVA'10, pages 101–110, Berlin, Heidelberg, 2010. Springer-Verlag. 89

[74] Adrienne Porter Felt, Matthew Finifter, Erika Chin, Steve Hanna, and David Wagner. A survey of mobile malware in the wild. SPSM '11, pages 3–14, New York, NY, USA, 2011. ACM. 6, 90

[75] Rik Ferguson. The history of the botnet. `http://countermeasures.trendmicro.eu/the-history-of-the-botnet-part-ii/`, 2010. 1, 2, 136

[76] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web*, pages 649–656. ACM, 2007. 18

[77] C. Fleizach, G.M. Voelker, and S. Savage. Slicing spam with occam's razor. *Proc. of Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2007. 4, 25, 138

[78] Dan Fletcher. A Brief History Of Spam. http://content.time.com/time/business/article/0,8599,1933796,00.html, 2009. 1, 9, 135

[79] Wilfried Gansterer, Michael Ilger, Peter Lechner, Richard Neumayer, and Jürgen Strauß. Anti-spam methods: state of the art. *Institute of Distributed and Multimedia Systems, University of Vienna*, 2005. 28, 29

[80] Yanbin Gao and Gang Zhao. Knowledge-based information extraction: a case study of recognizing emails of Nigerian frauds. NLDB'05, pages 161–172, Berlin, Heidelberg, 2005. Springer-Verlag. 20, 90

[81] Scott Garriss, Michael Kaminsky, Michael J Freedman, Brad Karp, David Mazières, and Haifeng Yu. RE: reliable email. In *Proceedings of the 3rd conference on Networked Systems Design & Implementation*, pages 22–22, 2006. 25, 29

[82] Luiz H Gomes, Rodrigo B Almeida, Luis Bettencourt, Virgilio Almeida, and Jussara M Almeida. Comparative graph theoretical characterization of networks of spam and legitimate email. *arXiv preprint physics/0504025*, 2005. 34

[83] Luiz Henrique Gomes, Cristiano Cazita, Jussara M Almeida, Virgílio Almeida, and Wagner Meira Jr. Characterizing a spam traffic. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 356–369. ACM, 2004. 33

[84] Joshua Goodman and Wen-tau Yih. Online Discriminative Spam Filter Training. In *CEAS*, 2006. 24

[85] Michel Grabisch and Christophe Labreuche. A decade of application of the choquet and sugeno integrals in multi-criteria decision aid. *Annals of Operations Research*, 2009. 113

[86] Galen A Grimes. Compliance with the CAN-SPAM Act of 2003. *Communications of the ACM*, 50(2):56–62, 2007. 13

[87] Thiago S Guzella and Walmir M Caminhas. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206–10222, 2009. 24, 28

[88] Robert J Hall. A countermeasure to duplicate-detecting anti-spam techniques. *Unknown Month, AT&T Labs Technical Report*, 99(1), 1999. 1, 136

[89] S. Hao, N.A. Syed, N. Feamster, A.G. Gray, and S. Krasser. Detecting spammers with SNARE: Spatio-temporal network-level automatic reputation engine. In *Proc. of the 18th conference on USENIX security symposium*, pages 101–118. USENIX Association, 2009. 4, 27, 138

[90] Evan Harris. The Next Step in the Spam Control War: Greylisting. `http://projects.puremagic.com/greylisting/whitepaper.html`. 28

[91] Cormac Herley. Why do NIGERIAN SCAMMERS SAY THEY ARE FROM NIGERIA? In *Proceedings of the Workshop on the Economics of Information Security*, 2012. 20, 90

[92] Thorsten Holz, Markus Engelberth, and Felix Freiling. Learning more about the underground economy: a case-study of keyloggers and dropzones. ESORICS'09, pages 1–18, Berlin, Heidelberg, 2009. Springer-Verlag. 89

[93] Jen-Hao Hsia and Ming-Syan Chen. Language-model-based detection cascade for efficient classification of image-based spam e-mail. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1182–1185. IEEE, 2009. 28

[94] Mikko Hypponen. Malware Goes Mobile. `http://www.cs.virginia.edu/~robins/Malware_Goes_Mobile.pdf`. 6, 88, 90

[95] Radicati Research Group Inc. Email Statistics Report 2011-2015, 2011. 2

[96] Radicati Research Group Inc. Email Statistics Report 2012-2016, 2012. 2

[97] M. Jakobsson and Z. Ramzan. *Crimeware: Understanding New Attacks and Defenses*. Symantec Press Series. Prentice Hall, 2008. 88

[98] Nan Jiang, Yu Jin, Ann Skudlark, Wen-Ling Hsu, Guy Jacobson, Siva Prakasam, and Zhi-Li Zhang. Isolating and analyzing fraud activities in a large cellular network via voice call graph analysis. MobiSys '12, pages 253–266, New York, NY, USA, 2012. ACM. 90

[99] J. Jung and E. Sit. An empirical study of spam traffic and the use of DNS black lists. In *Proc. of the 4th ACM SIGCOMM Conference on Internet measurement*, pages 370–375. ACM, 2004. 26

[100] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G.M. Voelker, V. Paxson, and S. Savage. Spamalytics: An empirical analysis of spam marketing conversion. In *Proceedings of the 15th ACM conference on Computer and communications security*, pages 3–14. ACM, 2008. 17, 30, 35, 41

[101] Chris Kanich, Nicholas Weaver, Damon McCoy, Tristan Halvorson, Christian Kreibich, Kirill Levchenko, Vern Paxson, Geoffrey M Voelker, and Stefan Savage. Show Me the Money: Characterizing Spam-advertised Revenue. In *USENIX Security Symposium*, 2011. 35

[102] P. Kiran and I. Atmosukarto. Spam or Not Spam–That is the Question. *Tech. rep., University of Washington*, 2009. 24, 70

[103] Aleksander Kolcz and Abdur Chowdhury. Hardening fingerprinting by context. *Proc. of Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2007. 32, 66

[104] Aleksander Kołcz, Abdur Chowdhury, and Joshua Alspector. Improved robustness of signature-based near-replica detection via lexicon randomization. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 605–610. ACM, 2004. 23

[105] Grzegorz Kondrak. N-gram similarity and distance. In *Proc. Twelfth Inter. Conf. on String Processing and Information Retrieval*, pages 115–126, 2005. 112

[106] Irena Koprinska, Josiah Poon, James Clark, and Jason Chan. Learning to classify e-mail. *Information Sciences*, 177(10):2167–2187, 2007. 24

[107] Christian Kreibich, Chris Kanich, Kirill Levchenko, Brandon Enright, Geoffrey M. Voelker, Vern Paxson, and Stefan Savage. On the spam campaign trail. In *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, LEET'08, pages 1:1–1:9, Berkeley, CA, USA, 2008. USENIX Association. 17, 88

[108] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 905–914. ACM, 2007. 18

[109] Kaspersky Lab. Kaspersky Lab Identifies Increase in Apple Phishing Scams as Cybercriminals Target Apple IDs and Financial Credentials. `http://www.kaspersky.com/about/news/virus/2013/Kaspersky_Lab_Identifies_Increase_in_Apple_Phishing_Scams_as_Cybercriminals_Target_Apple_IDs_and_Financial_Credentials`, 2013. 18

[110] Martin Lee. Who's next? Identifying risks factors for subjects of targeted attacks. In *Proc. Virus Bull. Conf*, pages 301–306, 2012. 19

[111] John R Levine. Experiences with Greylisting. In *CEAS*, 2005. 28

[112] Fulu Li and Mo han Hsieh. An Empirical Study of Clustering Behavior of Spammers and Groupbased Anti-Spam Strategies. *Proc. of Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2006. 30, 31, 66

[113] Olumide B. Longe, Victor Mbarika, M. Kourouma, F. Wada, and R. Isabalija. Seeing Beyond the Surface, Understanding and Tracking Fraudulent Cyber Activities. *CoRR*, abs/1001.1993, 2010. 88

[114] D. Lowd and C. Meek. Good word attacks on statistical spam filters. In *Proc. of Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, pages 125–132, 2005. 23

[115] MAAWG. Email Security Awareness and Usage Report, 2012. 3, 11, 137

[116] Federico Maggi. Are the Con Artists Back? A Preliminary Analysis of Modern Phone Frauds. CIT '10, pages 824–831, Washington, DC, USA, 2010. IEEE Computer Society. 90

[117] R. Mastaler. Tagged message delivery agent. `http://www.tmda.net/`. 29

[118] Damon McCoy, Hitesh Dharmdasani, Christian Kreibich, Geoffrey M Voelker, and Stefan Savage. Priceless: The role of payments in abuse-advertised goods. *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 845–856, 2012. 35

[119] Damon McCoy, Andreas Pitsillidis, Grant Jordan, Nicholas Weaver, Christian Kreibich, Brian Krebs, Geoffrey M Voelker, Stefan Savage, and Kirill Levchenko. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. *Proceedings of the 21st USENIX conference on Security symposium*, pages 1–1, 2012. 17, 35, 84

[120] Tyler Moore, Richard Clayton, and Ross Anderson. The Economics of Online Crime. *Journal of Economic Perspectives*, 23(3):3–20, Summer 2009. 89

[121] Tatsuya Mori, Kazumichi Sato, Yousuke Takahashi, and Keisuke Ishibashi. How is e-mail sender authentication used and misused? In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, pages 31–37. ACM, 2011. 26

[122] Marti Motoyama, Kirill Levchenko, Chris Kanich, Damon McCoy, Geoffrey M Voelker, and Stefan Savage. Re: CAPTCHAs-Understanding CAPTCHA-Solving Services in an Economic Context. In *USENIX Security Symposium*, volume 10, 2010. 30

[123] Jarno Niemela. Mobile Malware And Monetizing 2011. `http://www.cse.tkk.fi/fi/opinnot/T-110.6220/2011_Spring_Malware_Analysis_and_Antivirus_Technologies/luennot-files/Mobile%20Malware%20And%20Monetizing%20HUT.pdf`. 6, 90

[124] Jude Oboh and Yvette Schoenmakers. Nigerian Advance Fee Fraud in Transnational Perspective. *Policing multiple communities*, (15):235, 2010. 20

[125] Terri Oda and Tony White. Developing an immunity to spam. In *Genetic and Evolutionary Computation, GECCO 2003*, pages 231–242. Springer, 2003. 24

[126] Jon Oliver, Sandra Cheng, Lala Manly, Joey Zhu, Roland Dela Paz, Sabrina Sioting, , and Jonathan Leopando. Blackhole Exploit Kit: A Spam Campaign, Not a Series of Individual Spam Runs, 2012. 10, 17

[127] Kaan Onarlioglu, U Ozan Yilmaz, Davide Balzarotti, and Engin Kirda. Insights into user behavior in dealing with internet attacks. 2012. 3, 11, 79, 150

[128] M. Paganini. ASK: active spam killer. In *Proc. 2003 Usenix Annual Technical Conference*, 2003. 4, 29, 138

[129] A. Pathak, Y.C. Hu, and Z.M. Mao. Peeking into spammer behavior from a unique vantage point. In *Proc. of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, pages 1–9. USENIX Association, 2008. 4, 27, 138

[130] Abhinav Pathak, Feng Qian, Y. Charlie Hu, Z. Morley Mao, and Supranamaya Ranjan. Botnet Spam Campaigns Can Be Long Lasting: Evidence, Implications, and Analysis. *Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems*, pages 13–24, 2009. 20, 30, 31, 35, 36, 65, 66, 73, 83, 84, 85

[131] Samir Patil. Deciphering and mitigating Blackhole spam from email-borne threats. `http://www.virusbtn.com/conference/vb2013/abstracts/Patil.xml`, 2013. 17

[132] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G.M. Voelker, V. Paxson, N. Weaver, and S. Savage. Botnet judo: Fighting spam with itself. In *Symposium on Network and Distributed System Security (NDSS)*, 2010. 23, 32, 66

[133] Andreas Pitsillidis, Chris Kanich, Geoffrey M Voelker, Kirill Levchenko, and Stefan Savage. Taster's choice: a comparative analysis of spam feeds. *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 427–440, 2012. 36, 65

[134] Craig Pollard. Telecom fraud: the cost of doing nothing just went up. *Netw. Secur.*, 2005(2):17–19, February 2005. 88

[135] V.V. Prakash. Virpul's razor. `http://razor.sf.net`, 1999. 1, 25, 136

[136] Feng Qian, Abhinav Pathak, Yu C. Hu, Zhuoqing M. Mao, and Yinglian Xie. A case for unsupervised-learning-based spam filtering. *ACM SIGMETRICS Performance Evaluation Review*, 38(1):367–368, 2010. 32, 36, 65, 66, 73, 84, 85, 133, 149

[137] Zhiyun Qian, Zhuoqing Morley Mao, Yinglian Xie, and Fang Yu. On network-level clusters for spam detection. In *NDSS*, 2010. 27

[138] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *Proc. of the 14th ACM conference on computer and communications security*, pages 342–351. ACM, 2007. 4, 27, 56, 138

[139] Anirudh Ramachandran and Nick Feamster. Understanding the network-level behavior of spammers. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 291–302. ACM, 2006. 27

[140] Gaëlle Recourcé. Interpreting contact details out of e-mail signature blocks. In *Proceedings of the 21st international conference companion on WWW*. ACM, 2012. 90, 91, 154

[141] ReturnPath. Email Intelligence Report, Q3 2012. 3, 14, 22, 23, 25, 137

[142] ReturnPath. The TINS report. `http://landing.returnpath.com/TINS-2013-Thanks?sfdc=701000000006DIs`, 2013. 4, 138

[143] ReturnPath. Return path research finds that dmarc is effective in blocking millions of potentially fraudulent messages from 60% of the world's mailboxes. `http://tinyurl.com/l8rrj7n`, Februrary 2013. 26

[144] Rodrigo Roman, Jianying Zhou, and Javier Lopez. Protection against spam using pre-challenges. In *Security and Privacy in the Age of Ubiquitous Computing*, pages 281–293. Springer, 2005. 28

[145] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, 1998. 4, 23, 24, 138

[146] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos, and P. Stamatopoulos. A memory-based approach to anti-spam filtering for mailing lists. *Information Retrieval*, 6(1):49–73, 2003. 4, 23, 24, 138

[147] J. Shawe-Taylor, K. Howker, and P. Burge. Detection of fraud in mobile telecommunications. *Information Security Technical Report*, 4(1):16–28, 1999. 88

[148] Devrim Sipahi and G Dalkilic. Determination of SPF records for the intention of sending spam. In *Signal Processing and Communications Applications Conference (SIU), 2012 20th*, pages 1–4. IEEE, 2012. 26

[149] Frank Stajano and Paul Wilson. Understanding scam victims: seven principles for systems security. *Communications of the ACM*, 54(3):70–75, March 2011. 20, 88, 89, 154

[150] Henry Stern. A Survey of Modern Spam Tools. In *CEAS*, 2008. 2, 136

[151] Alastair Stevenson. Hackers hit 3,000 UK web users a day with phishing attacks. `http://www.v3.co.uk/v3-uk/news/2277343/hackers-hit-3-000-uk-web-users-a-day-with-phishing-attacks`, 2013. 18

[152] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna. The Underground Economy of Spam: A Botmaster's Perspective of Coordinating Large-Scale Spam Campaigns. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2011. 41, 89

[153] Gianluca Stringhini, Manuel Egele, Apostolis Zarras, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. B@ bel: leveraging email delivery for spam mitigation. *USENIX Security Symposium*, 2012. 17, 28

[154] Symantec Intelligence Report. Symantec intelligence report: July 2013, July 2013. 2, 137

[155] Symantec's MessageLabs Intelligence. Messagelabs intelligence annual security report, 2010. 17, 45, 51

[156] Kurt Thomas, Chris Grier, Justin Ma, Vern Paxson, and Dawn Song. Design and evaluation of a real-time URL spam filtering service. *IEEE Symposium on Security and Privacy*, 2011. 30, 31, 84

[157] Kurt Thomas, Chris Grier, Justin Ma, Vern Paxson, and Dawn Song. Design and evaluation of a real-time URL spam filtering service. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, SP '11, pages 447–462, Washington, DC, USA, 2011. IEEE Computer Society. 36, 88

[158] Olivier Thonnard. *A multi-criteria clustering approach to support attack attribution in cyberspace.* PhD thesis, École Doctorale d'Informatique, Télécommunications et Électronique de Paris, March 2010. 7, 33, 112, 113, 132, 162

[159] Olivier Thonnard, Leyla Bilge, Gavin O'Gorman, Seán Kiernan, and Martin Lee. Industrial espionage and targeted attacks: Understanding the characteristics of an escalating threat. In *Research in Attacks, Intrusions, and Defenses*, pages 64–85. Springer, 2012. 18, 19, 33, 112

[160] Olivier Thonnard and Marc Dacier. A strategic analysis of spam botnets operations. In *Proc. of Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '11, pages 162–171, New York, NY, USA, 2011. ACM. 17, 33, 36, 112

[161] Charles Tive. *419 scam: Exploits of the Nigerian con man.* iUniverse, 2006. 20

[162] Vicenç Torra and Yasuo Narukawa. *Modeling Decisions: Information Fusion and Aggregation Operators.* Springer, Berlin, 2007. 113

[163] David Turner and Daniel Havey. Controlling spam through lightweight currency. In *proceedings of the Hawaii International Conference on Computer Sciences*. Citeseer, 2004. 28

[164] Dan Twining, Matthew M Williamson, Miranda Mowbray, and Maher Rahmouni. Email Prioritization: Reducing Delays on Legitimate Mail Caused by Junk Mail. In *USENIX Annual Technical Conference, General Track*, pages 45–58, 2004. 28

[165] Steven J Vaughan-Nichols. Saving private e-mail. *Spectrum, IEEE*, 40(8):40–44, 2003. 10

[166] Andrew G West, Adam J Aviv, Jian Chang, and Insup Lee. Mitigating spam using spatio-temporal reputation. Technical report, DTIC Document, 2010. 27

[167] M. Wong and W. Schlitt. Sender policy framework (SPF) for authorizing use of domains in e-mail, version 1. Technical report, RFC 4408, April, 2006. 4, 25, 26, 58, 138

[168] Chih-Hung Wu. Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. *Expert Systems with Applications*, 36(3):4321–4330, 2009. 25

[169] Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten, and Ivan Osipkov. Spamming botnets: signatures and characteristics. In *ACM SIGCOMM Computer Communication Review*, volume 38, pages 171–182. ACM, 2008. 31, 35

[170] Jeff Yan and Ahmad Salah El Ahmad. A low-cost attack on a microsoft CAPTCHA. In *Proceedings of the 15th ACM conference on Computer and communications security*, pages 543–554. ACM, 2008. 30

[171] Chun-Chao Yeh and Chia-Hui Lin. Near-duplicate mail detection based on URL information for spam filtering. In *Information Networking. Advances in Data Communications and Wireless Networks*, pages 842–851. Springer, 2006. 36

[172] Wen-tau Yih, Robert McCann, and Aleksander Kolcz. Improving spam filtering by detecting gray mail. *Proc. of Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2007. 4, 5, 6, 22, 37, 64, 65, 138

[173] Seongwook Youn and Dennis McLeod. Spam decisions on gray e-mail using personalized ontologies. *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1262–1266, 2009. 10, 37

[174] Yue Zhang, Jason I Hong, and Lorrie F Cranor. Cantina: a content-based approach to detecting phishing web sites. In *Proceedings of the 16th international conference on World Wide Web*, pages 639–648. ACM, 2007. 18

[175] Vasilios Zorkadis, Dimitris A Karras, and M Panayotou. Efficient information theoretic strategies for classifier combination, feature extraction and performance evaluation in improving false positives and false negatives for spam e-mail filtering. *Neural Networks*, 18(5):799–807, 2005. 24