Digital Investigation 12 (2015) S12–S21

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2015 Europe

Spam campaign detection, analysis, and investigation☆

Son Dinh [a, *], Taher Azeb [a], Francis Fortin [b], Djedjiga Mouheb [a], Mourad Debbabi [a]

[a] *NCFTA Canada & Concordia University, 1455 de Maisonneuve Blvd West, Montreal, QC H3G 1M8, Canada*

[b] *Centre international de criminologie comparée, École de criminologie, Université de Montréal, Montreal, QC H3C 3J7, Canada*

Keywords:
Spam
Spam campaign
Spam analysis
Characterization
Frequent pattern tree

ABSTRACT

Spam has been a major tool for criminals to conduct illegal activities on the Internet, such as stealing sensitive information, selling counterfeit goods, distributing malware, etc. The astronomical amount of spam data has rendered its manual analysis impractical. Moreover, most of the current techniques are either too complex to be applied on a large amount of data or miss the extraction of vital security insights for forensic purposes. In this paper, we elaborate a software framework for spam campaign detection, analysis and investigation. The proposed framework identifies spam campaigns on-the-fly. Additionally, it labels and scores the campaigns as well as gathers various information about them. The elaborated framework provides law enforcement officials with a powerful platform to conduct investigations on cyber-based criminal activities.

Introduction

Electronic mail, or most commonly known as *email*, is ubiquitous and so is its abusive usage. *Spam emails* affect millions of users, waste invaluable resources and have been a burden to the email systems. For instance, according to Symantec Intelligence Report, the global ratio of spam in email traffic is 71.9% (Symantec Intelligence Report, 2013). Furthermore, adversaries have taken advantage of the ability to send countless emails anonymously, at the speed of light, to carry on vicious activities (e.g., advertising of fake goods or medications, scams causing financial losses) or even more severely, to commit cyber crimes (e.g., child pornography, identity theft, phishing and malware distribution). Consequently, spam emails contain priceless cyber security intelligence, which may unveil the world of cyber criminals.

Spam data has been used extensively in various studies to detect and investigate cyber threats such as botnets (Zhuang et al., 2008; Xie et al., 2008; John et al., 2009; Pathak et al., 2009; Thonnard and Dacier, 2011; Stringhini et al., 2011) and phishing attacks (Fette et al., 2007; Bergholz et al., 2008; Moore et al., 2009; Bergholz et al., 2010). Moreover, the accessibility of various data sources, which can be correlated to complement their incompleteness, has brought new opportunities and challenges to researchers. Unfortunately, most studies either use a single data source or work on a static spam data that is collected during a specific time frame (Pitsillidis et al., 2012). More importantly, spammers have been constantly modernizing their arsenals to defeat the anti-spam efforts. Spamming techniques have evolved remarkably from simple programs to sophisticated spamming software, which disseminate template-generated spam through a network of compromised machines. Botnet-disseminated spam emails are usually orchestrated into large-scale campaigns and act as the pivotal instrument for several cyber-based criminal activities. As a consequence, it is critical to perform an analysis of spam data, especially *spam campaigns*, for the

☆ The authors gratefully acknowledge support from the Canadian Radio-television and Telecommunications Commission (CRTC).

* Corresponding author.
E-mail address: tsondt@gmail.com (S. Dinh).

http://dx.doi.org/10.1016/j.diin.2015.01.006

purpose of cyber-crime investigation. Nevertheless, given the stunning number of spam emails, it is implausible to analyze them manually. Therefore, cyber-crime investigators need automatic techniques and tools to accomplish this task.

In this research, we aim at elaborating methodologies for spam campaign detection, analysis and investigation. We also emphasize the importance of correlating different data sources to reveal spam campaign characteristics. More precisely, we propose a framework that: (1) Consolidates spam emails into campaigns. (2) Labels spam campaigns by generating related topics for each campaign from Wikipedia data. (3) Correlates different data sources, namely *passive DNS*, malware, WHOIS information and geolocation, to provide more insights into spam campaign characteristics. (4) Scores spam campaigns based on several customizable criteria.

The identification of spam campaigns is a crucial step for analyzing spammers' strategies for the following reasons. First, the amount of spam data is astronomical, and analyzing all spam messages is costly and almost impossible. Hence, clustering spam data into campaigns reduces significantly the amount of data to be analyzed, while maintaining their key characteristics. Second, because of the characteristics of spam, spam messages are usually sent in bulk with specific purposes. Hence, by clustering spam messages into campaigns, we can extract relevant insights that can help investigators understand how spammers obfuscate and disseminate their messages. Labeling spam campaigns and correlating different data sources reveal the characteristics of the campaigns and therefore, significantly increase the effectiveness of an investigation. Moreover, scoring spam campaigns helps investigators concentrate on campaigns that cause more damage (e.g., malware distribution or phishing).

The remainder of this paper is organized as follows. In Section 2, we present the related work. We discuss existing techniques for detecting spam campaigns in Section 3. In Section 4, we present our framework. Experimental results are presented in Section 5. Finally, we conclude the paper in Section 6.

Related work

Several approaches for clustering spam emails into groups have been proposed in the literature:

URL-based spam email clustering

In this category, spam emails are clustered using features such as the embedded URLs or the source IP addresses of spam emails. F. Li et al. (Li and Hsieh, 2006) propose an approach for clustering spam emails based on identical URLs. The approach also uses the amount of money mentioned inside the content of the email as an extra feature of the cluster. Xie et al. (2008) propose AutoRE, a framework that detects botnet hosts by signatures that are generated from URLs embedded in email bodies. Both of these URL-based clustering approaches are very efficient in terms of performance and number of false positives. However, spammers can easily evade such techniques using dynamic source IP addresses, URL shorten services or polymorphic URLs.

Spam email clustering using text mining methods

Zhuang et al. (2008) develop techniques to unveil botnets and their characteristics using spam traces. Spam emails are clustered together into campaigns using shingling algorithm (Broder et al., 1997). The relationship between IP addresses determines if the campaigns are originated from the same botnet. In Qian et al. (2010), Qian et al. design an unsupervised, online spam campaign detection, namely *SpamCampaignAssassin* (SCA), and apply extracted campaign signatures for spam filtering. SCA utilizes a text-mining framework built on *Latent Semantic Analysis* (LSA) (Landauer et al., 1998) to detect campaigns. Nevertheless, template-generated spam and scalability render text mining ineffective.

Spam email clustering using web pages retrieved from embedded URLs

Anderson et al. (2007) propose a technique called *spamscatter* that follows embedded URLs inside spam emails, renders and clusters the websites into scams using *image shingling*. Konte et al. (2009) analyze the scam hosting infrastructure. The authors extract *scam campaigns* by clustering web pages retrieved from URLs inside spam emails using different heuristic methods. Even though URL content provides sound results, scalability is also an issue. More importantly, actively crawling embedded URLs may alert spammers and expose the spamtraps used to collect data.

Spam email clustering based on email content

Wei et al. (2008) propose an approach based on the agglomerative hierarchical clustering algorithm and the connected components with weighted edges model to cluster spam emails. Only spam emails used for advertising are tested by the authors. In Calais et al. (2008), Calais et al. introduce a spam campaign identification technique based on frequent-pattern tree (FP-Tree) (Han et al., 2000, 2004). Spam campaigns are identified by pointing out nodes that have a significant increase in the number of children. An advantage of this proposed technique is that it can detect obfuscated parts of spam emails naturally without prior knowledge. In another work (Guerra et al., 2008), Calais et al. briefly present *Spam Miner*, a platform for detecting and characterizing spam campaigns. Kanich et al. (2009) analyze spam campaigns using a parasitic infiltration of an existing botnet's infrastructure to measure spam delivery, click-through and conversion. Haider and Scheffer (2009) apply Bayesian hierarchical clustering (Heller and Ghahramani, 2005) to group spam messages into campaigns. The authors develop a generative model for clustering binary vectors based on a transformation of the input vectors.

Spam campaign detection techniques

A straightforward approach for grouping spam emails into campaigns is to calculate the distances between spam