Journal of Network and Computer Applications 43 (2014) 56–65

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Detecting spamming activities in a campus network using incremental learning

JuiHsi Fu*, PoChing Lin, SingLing Lee

Department of Computer Science and Information Engineering, National Chung Cheng University, 168 University Road, Minhsiung Township, 62162 Chiayi, Taiwan, ROC

ARTICLE INFO

Article history:
Received 1 July 2013
Received in revised form
24 February 2014
Accepted 10 March 2014
Available online 30 March 2014

Keywords:
Spamming host
SMTP session
Incremental learning
Failure information

ABSTRACT

Most spam filters deployed on the receiver side are good at curbing email spam for end users, but help little to crack down the spamming sources. This work is intended to nip the spamming hosts in the bud. We collected the logs of SMTP sessions initiated from the hosts in the campus for half a year, and analyzed the activities of the hosts with *the rates of successful deliveries and various types of failure messages* in the sessions as the features. We use *an incremental passive-aggressive learning algorithm* to efficiently adapt the classifier to the latest spamming activities for detecting the spamming hosts. The detection accuracy can reach 93.5% after the classifier is adjusted in just few rounds. This design will be useful for the network administrators to reliably detect and crack down the internal spamming hosts.

1. Introduction

It is common to see overwhelming delivery of unsolicited email, namely *spam*, due to the extremely low cost of email delivery. Around 90% of email messages are reportedly spam (Messaging Anti-Abuse Working Group, 2011), which not only wastes Internet bandwidth and the storage space of email service providers, but also annoys or even harms users. Even though most users just ignore spam, the overall profits are still large enough to support the misbehavior due to the huge spam volume (Kanich et al., 2009). Moreover, spammers can exploit more hosts for efficient spam distribution by sending spam embedded with malware or harmful links for "drive-by download" attacks, and the compromised hosts will become part of a *botnet* (Xie et al., 2008), a large group of infected hosts known as bots under control of a botmaster. It was reported that botnets are responsible for approximately 83% of global spam (Prince, 2011).

The common countermeasure is filtering out spam messages for end users as many as possible by the spam-filtering functions provided by email service providers, mail clients or mail proxies. Such functions on the receiver side have been intensively studied in the past years (Zhang et al., 2004; Cormack and Lynam, 2007; Hulten et al., 2004). Even though they can filter out spam with high accuracy, the spamming traffic still consumes Internet bandwidth, and the spamming hosts are still alive. Therefore, it is essential to also detect and nip spamming hosts in the bud on the sender side (Ramachandran et al., 2007; Zhao et al., 2009; Stringhini et al., 2011; Duan et al., 2012; Ehrlich et al., 2010; Las-Casas et al., 2013; Lam and Yeung, 2007; Tseng and Chen, 2009b; Clayton, 2004). If a spamming host can be cracked down as early as possible, the spam from it can be all eliminated thereafter. Two questions immediately arise to effectively detect spamming hosts: (1) *What are the reliable features to identify spamming hosts, which may attempt to evade the detection*? (2) *How is the detection model built from a possibly huge dataset of SMTP logs, and adapted to the latest spamming activities*?

To shed light on the activities of spamming hosts and the reliable features to detect them, we deploy the Bro network intrusion detection system (NIDS) (Paxson, 1998) to monitor the SMTP traffic in a large university campus. Bro is good at parsing application protocol messages and tracking network activities of interest. The Bro monitoring host is deployed in the computer center to observe the SMTP sessions initiated from the internal hosts in the campus to the external, and logs them for detection. We are interested in detecting only the spamming hosts in the campus, rather than those from the external because we do not have the authority to crack down the latter even if they are detected. Over the period of SMTP traffic monitoring for half a year, we learned from the Bro logs that *the spamming hosts were likely to receive failure messages in the responses from external mail servers* when they attempted to initiate the SMTP sessions. We classify various types of failure messages, and study the importance of the features based on *the rates of successful email deliveries and the failures, as well as the other related features*, to infer suspicious spamming hosts.

We use an innovative *incremental passive-aggressive* (*IPA*) *learning algorithm* (Fu and Lee, 2012) to build an adaptive classifier from

* Corresponding author.
E-mail addresses: fjh95p@cs.ccu.edu.tw (J. Fu), pclin@cs.ccu.edu.tw (P. Lin), singling@cs.ccu.edu.tw (S. Lee).

http://dx.doi.org/10.1016/j.jnca.2014.03.010
1084-8045/

the logs over the long period and adapt the detection to constantly changing activities of spamming hosts. Since spamming botnets may fade away or be taken down (Hickins, 2011), new species of spamming botnets may sprout up, or spamming hosts may constantly update the spam messages and the lists of target email addresses (Kreibich et al., 2008), *the detection model to identify spamming hosts is required to be periodically adjusted to recognize the latest spamming activities*. Moreover, the logs of SMTP activities over a long period, say half a year, can be huge. Incremental learning (Utgoff, 1989; Mohamed et al., 2007; Du et al., 2009) can solve this problem by *incrementally* using subsets of data, instead of the whole dataset. It can serve the detection purpose for building a statistical classification model based on the previously observed activities, and incrementally adjusting the model for high efficiency after newly labeled data are received. Thus, it is an essential algorithm to deal with the huge dataset of SMTP logs. We formulate a simple constrained optimization problem for each potential update of the linear classifier, and then the candidate classifier is the solution derived from the Lagrange multipliers. Particularly, a closed-form solution is derived to obtain the efficient update steps. It is presented through experimental results that the method can efficiently adjust the classifier, and the spamming activities can be recognized even though they are rare.

The contributions of this work are summarized as follows:

1. We characterize the spamming activities with various failure messages from the external mail servers and classify the messages in detail. The importance of each feature to detect spamming hosts is also deeply studied.
2. We apply an incremental passive-aggressive (IPA) learning algorithm to adaptively detect spamming hosts from the huge SMTP logs. The design can help network administrators to detect the spamming hosts in a campus, among other organizations, and then crack them down.

The remainder of the paper is organized as follows. Section 2 describes related work about detecting spamming hosts and incremental learning. Section 3 describes the method of analyzing various SMTP logs and detecting spamming hosts with incremental learning. Section 4 presents the detection results, as well as the case studies. Section 5 summarizes the key points in this work.

2. Related work

2.1. Detection of spam and spamming botnet

We focus on the studies dedicated to detecting spamming hosts in the related work because generic botnet detection is less relevant to this work. The studies are reviewed below with the emphasis on their limitations.

SpamTracker (Ramachandran et al., 2007) is a behavioral blacklisting algorithm to identify spamming hosts by clustering the hosts that have similar patterns of target domains in their outgoing mail messages, but spammers can easily dispatch the recipients' email addresses with different domains to the spamming hosts and confuse the detection. Stringhini et al. (2011) collected spam messages and identified the hosts for the same spam campaigns by similar spam content in them. The authors also extracted the transaction logs to group the spamming hosts destined for similar targets as the seeds, and then looked for other spamming hosts behaving similar to the seeds. In contrast, this work does not rely on any prior establishments of observed spam content or behavior, and the detection can automatically adapt to the latest spamming behavior by incremental learning.

Duan et al. (2012) redirected the outgoing messages from a campus to a spam filter, and used the sequential probability ratio test to detect the internal hosts constantly sending spam. This work does not rely on an external spam filter for two reasons. First, an SMTP session may fail due to the causes discussed in Section 3.3. Not even a spam message will be sent out if a session keeps failing in the transaction stage (e.g., because it is in the blacklist). Thus, content filtering is useless in this case. Second, a user may configure automatic forwarding on a mail server, which will forward the received mail, including spam, to an external account specified by the users (see the discussion in Section 4.2). The spam filter will see many spam messages from the mail server, and then the detection is likely to result in false positives.

Ehrlich et al. (2010) presented a method to detect spamming hosts without inspecting the mail content. They observed that the payload sizes of the SMTP requests from normal hosts are larger and much more variable than those for spamming hosts. Thus, modeling the network flows can detect the behavior associated with spamming hosts. This method does not look into the packet payloads, so a spammer can easily evade the detection by stuffing mail messages with random content (Kreibich et al., 2008) to make the distributions of payload sizes from spamming hosts and normal ones indistinguishable. Las-Casas et al. (2013) referred to features such as the number of SMTP transactions, the inter-arrival time of SMTP transactions to detect spamming hosts. Like Ehrlich et al. (2010), the spammers can manipulate the features to make them like those from normal hosts. In comparison, the features in this work are hard to manipulate, as discussed in Section 4.3.

To characterize spamming activities of email accounts, Lam and Yeung (2007) and Tseng and Chen (2009b) analyzed social interactions among email accounts, e.g., the number of email messages received (sent) and the ratio of sending/receiving email messages between users. The methods are effective in identifying spamming accounts, but tracking the social relations in a graph is not scalable because there will be probably too many tracked accounts for long tracking time. For example, Tseng and Chen (2009b) had over 600,000 nodes in the social network from a 10-day email trace. This work can handle a much larger dataset (from the logs during half a year).

In Zhu et al. (2009) proposed failure information analysis to detect generic botnets based on the observation that bots can easily incur a high rate of failures in their activities. However, not all failure messages are reliable features because ordinary activities can also incur failure messages due to improper configurations, which will result in false positives. Huang (2013) can detect botnets based on the failure models for generic network protocols. Clayton (2004) also detected spamming hosts according to failure messages, but it does not inspect the significance of different types of failures, nor is it adapted to the latest spamming activities to cope with the bulk SMTP logs. This work is inspired from the prior studies, but we examine the SMTP failure messages for detecting spamming hosts in much greater depth. This work is a significant extension of the preliminary SMTP failure analysis for spamming activities in Lin et al. (2011). Since the diverse configurations and implementations of the mail servers that the internal hosts in the campus contact result in different representations of semantically similar failure messages, we take great care to sort out various reply messages according to their semantics. We also study the significance of the failure messages as the features for identifying spamming hosts, and use the IPA algorithm to adapt the classifier to the latest spamming activities.

2.2. Incremental learning and online learning

Requests of analyzing periodic data have emerged in practical applications, including network traffic analysis (Sena and Belzarena, 2009), anomaly detection (Robertson et al., 2010) and intrusion detection (Du et al., 2009). The applications need to periodically