

Spamming Botnets: Signatures and Characteristics

HY558 – Σοφία Λούτου

Botnet refers to a group of compromised host computers that are controlled by a small number of commander hosts (Command and Control Servers) and they have been widely used for sending spam emails at a large scale.

The authors of the paper are presenting a novel framework, AutoRE that identifies botnet hosts by generating botnet spam signatures from emails. Their large-scale analysis provides information about the botnets' characteristics and trends that can benefit future botnet detection. AutoRE does not require labeled data and whitelists and its output is regular expression signatures that appear to be more robust and have a low false positive rate. Furthermore AutoRE –using the generated signatures- groups the emails into spam campaigns. A campaign refers to a targeted spam effort to a single product or service.

The authors focus on URLs that are embedded in email content, as they are the most critical part of it (directing users to phishing pages or targeted products Web sites). The extraction of the URL signatures is challenging as the spam emails may contain legitimate and general URLs and other random characteristics (techniques of URL obfuscation) in order to evade detection. Thus it is critical to recognize patterns at polymorphic URLs. It should also be noticed that HTML-based emails often contain URLs generated by standard software.

The 3 basic modules of AutoRE are the following:

1)URL Preprocessor: Extracts URLs and other relevant fields (source server IP address, email sending time) from input emails and groups them according the Web domains

After preprocessing, each mail might be associated with multiple groups, as the email may contain multiple URLs pertaining to different domains. Which group best characterizes an underlying campaign?

2)Group Selector: At every iteration it selects the URL group that exhibits the strongest temporal correlation, exploring the bursty property of botnet email traffic.

3)RegEx Generator: Given a set of URLs pertaining to the same domain, it returns 2 types of signatures: complete URL based signatures (to detect spam emails that contain an identical URL string) and regular expression signatures (more generic and powerful, they can detect spam emails with polymorphic URLs).

The signatures should meet the criteria of being distributed (quantified using the total number of ASes spanned by the source IP addresses), bursty (quantified using the inferred duration of a botnet spam campaign) and specific (quantified using an information entropy metric pertaining to the probability of a random URL string matching the signature).

The generation of regular expressions requires as input a set of polymorphic URLs from the same Web Domain and based on them, a keyword-based signature tree is constructed. Candidate regular expressions are generated via Detailing and Generalization. Detailing returns a domain-specific regular expression using a keyword-based signature as input. This procedure increases the quality of URL signatures to reduce false positive rate. Generalization returns a more general domain-agnostic regular expression by merging very similar domain-specific expressions. In the end, the generated expressions are evaluated to ensure that they are specific enough.

The authors present measurements that reveal how powerful are the regular expressions introduced by the paper. They found scenarios where spammers sign up for many domains, so in case a domain is blacklisted they easily change to the next and interestingly the URL structures of these domains are quite similar (indicating that a fixed set of tools is used to set up servers and send emails). Furthermore the Web pages pointed to by each set of polymorphic URLs are similar to each other. Other results indicate that 90% of campaigns have standard deviation time less than 24 hours and that botnets sharing a domain-agnostic signature barely overlap with each other in most of the cases.

Some key findings of this paper concerning the botnet characteristics include the following: Botnets are becoming increasingly popular for spam delivery and one botnet host is involved in multiple attacks. Viewed individually, a botnet host does not exhibit distinctive sending patterns. Aggregated common features from concurrent email sending activities should be studied to identify patterns. Finally, botnet attacks may have different phases –usually the first phase is to find victim-computers to expand the botnet.

Authors claim that AutoRE has the potentiality to work in real time mode. They also support that it can cope with even more sophisticated obfuscation techniques by spammers.