

## Spamscatter: Characterizing Internet Scam Hosting Infrastructure

David S. Anderson, Chris Fleizach, Stefan Savage and Geoffrey M. Voelker

Report by Gessiou Eleni

This paper focuses on the scam infrastructure that is nourished by spam. The Spamscatter technique presented follows the links embedded in spam emails, and automatically identifies scam infrastructure using "Image Shingling" to capture graphical similarity between rendered sites. In overall, 2,000 distinct scams are indentified on 7,000 distinct servers.

The methodology used is collecting a large amount of spam emails (over 1 million in total), identifying the urls in spam messages and following the links to the final destination. Then, "Image Shingling" is performed to cluster the web pages and identify scams. In order to characterize dynamic behaviors of scams, like availability and lifetime, they probe the scam serves every 3 hours.

The algorithm of "Image Shingling" includes the following steps:

1. Compares screen shots of web browser
  2. Divides image into fixed memory chunks (40 x 40 pixels best trade-off between granularity and shingling performance)
  3. Hashes each chunk to create an image shingle
- Two scams are identified as similar, if they share at least a threshold of similar images.

The resulting findings concentrate on several aspects, such as distributed and shared infrastructure of scams, their lifetime and stability as well as their location.

Scams may use multiple hosts for fault-tolerance, resilience of blacklisting, and for load balancing. The results show that most scams (94%) are not distributed, are typically hosted on a single IP address with one domain name. The interesting part of this finding is that those few scams that are hosted on multiple IP addresses are highly distributed, indicating the concern of spammers about blacklisting.

Another question to be answered in this paper is "To what extent do multiple scams share infrastructure?". The results in this question show that 38% of the scams are hosted on machines hosting at least another scam. 96% of these pairs of scams overlap in time, in other words servers are used concurrently for different scams, but only 10% of the pairs do fully overlap each other.

Overall, scam lifetime approaches two weeks. As expected, scams hosted in multiple hosts live twice as the ones hosted in one machine. An intriguing result is that in general, malicious scams have shorter lifetime than all scams. Half of malicious scams disappear before 120 hours, while the same is happening for less than 15% of all scams. Another aspect of the study concentrates on the lifetime of spam campaigns in comparison with the one of the scams. Most spam campaigns are relatively small, 88% last 20 hours or less, while only 8% of them

last more than 2 days in contrast with the lifetime of scams that last on average one week.

The availability of scams is defined as the number of successful web page downloads divided by the number of attempts/probes. It is showed that scams have excellent availability: over 90% have an availability of 99% or higher and most of the remaining have 98% or more availability. Moreover, using passive OS fingerprinting, they show that more unix servers (43%) than windows (30%) are used for scam infrastructure and all of them have reported good link connectivity.

Finally the paper, examines the location of the scams. It is showed that scam hosts are mainly located in the US whereas the spam relays are distributed around many countries. The reason this happens is that the infrastructure of scams need to be more stable and live longer than the spam relays (which only send spam emails for a very short time), because scam hosting is a service that fundamentally depends upon user interaction to be successful.

Concluding, the authors of the paper perform some measurements of the scam infrastructure employing the Spamscluster technique for identifying clusters of scams. In overall, it is showed that more scams use one web server and their lifetime is much longer than the one of the spam emails. Finally, mapping the geographic locations of scam hosts show that their majority is concentrated in US.