



SPAMSCATTER: CHARACTERIZING INTERNET SCAM HOSTING INFRASTRUCTURE

By

D. Anderson, C. Fleizach, S. Savage, and G.
Voelker

INTRODUCTION

- Spam, unsolicited bulk email, attained public recognition
- In 2006, industry estimated that spam messages comprise 80% over all emails
- However, money-making scams are the engine that drives such emails
- Spam is only a way to drag the user to a website (scam)

METHODOLOGY

1. Collect over one million spam messages
2. Identify the urls in spam messages and follow the links to the final destination
3. Collect all the web pages of those destinations servers
4. Perform “Image Shingling” to cluster the web pages and identify unique scams
5. Finally, probe the scam servers to characterize dynamic behaviors like availability and lifetime

SPAM FEED

- Spam Feed: All messages sent to any email address at a well-known four-letter top-level domain
- Receives over 150,000 per day collecting more than one million in total
- Any email sent is spam because no active users on the mail server for the domain
- 93% of the “From” addresses are used only once => use of random source address to defeat address-based spam blacklists

SPAMSCATTER: DATA COLLECTION

- Input: Spam Feed, then extracts sender and URLs and probes those hosts
- Sender: ping, trace-route, and DNSBL
- URLs: ping, trace-route, DNSBL, HTML page, and a screen shot of browser window
- Repeat probing of URLs every 3 hours for one week

IMAGE SHINGLING

- Depend on the webpage to identify unique scams
 - Can't depend on host IPs, a host may serve multiple scams
 - A scam may be hosted on multiple virtual servers with different IP addrs
- ⇒ Depend on scam content and identify by using image shingling algorithm

Why image shingling?

- Depend on the spam messages, but randomness is a problem
- Compare URLs, might change for the same scam to defeat URL blacklisting
- Compare HTML content, but many pages has insufficient textual information

IMAGE SHINGLING (CONTINUE)

- Compare screen shots of web browser
- Divide image into fixed memory chunks (40 x 40 pixels best trade-off between granularity and shingling performance)
- Hash each chunk to create an image shingle
- Similar if they share at least a threshold of similar images.
- By manually inspecting, found that 70% threshold minimizes false negatives and false positives of determining equivalence

RESULTS AND ANALYSIS

SUMMARY

<i>Characteristic</i>	<i>Summary Result</i>
Trace period	11/28/06 – 12/11/06
Spam messages	1,087,711
Spam w/ URLs	319,700 (30% of all spam)
Unique URLs	36,390 (11% of all URLs)
Unique IP addresses	7,029 (19% of unique URLs)
Unique scams	2,334 (6% of unique URLs)

Table 1: Summary of spamscatter trace.

RESULTS AND ANALYSIS

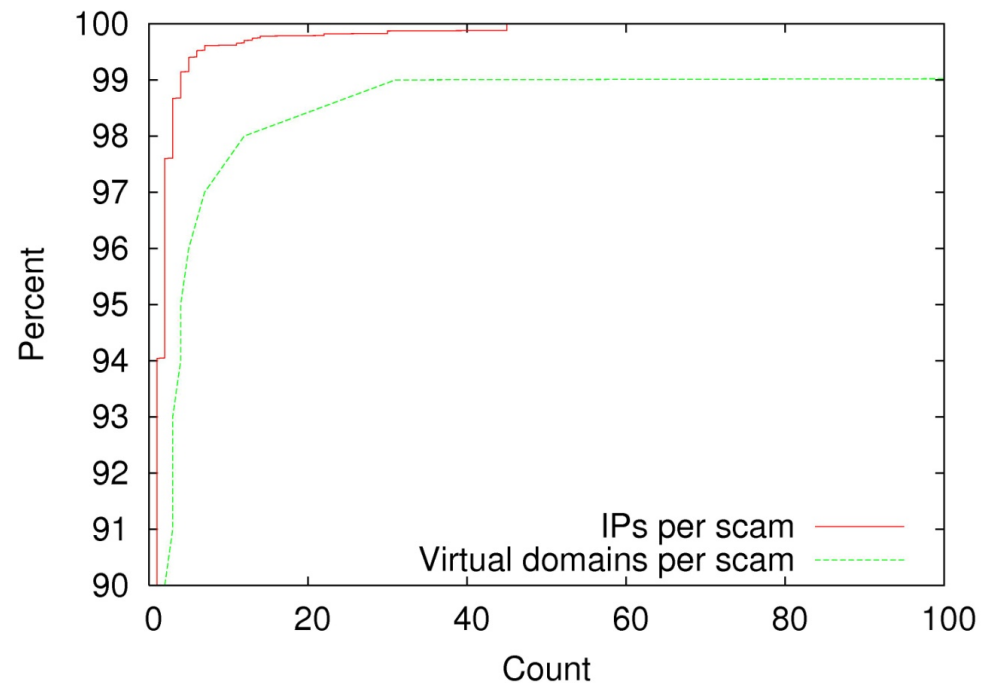
CATEGORIES

<i>Scam category</i>	<i>% of scams</i>
Uncategorized	29.57%
Information Technology	16.67%
Dynamic Content	11.52%
Business and Economy	6.23%
Shopping	4.30%
Financial Data and Services	3.61%
Illegal or Questionable	2.15%
Adult	1.80%
Message Boards and Clubs	1.80%
Web Hosting	1.63%

Table 2: Top ten scam categories.

DISTRIBUTED INFRASTRUCTURE

- *To what extent is the distributed infrastructure for scams?*
- Scams may use multiple hosts for fault-tolerance, resilience of blacklisting, and for load balancing
- Most scams are not distributed:
94% used only one IP



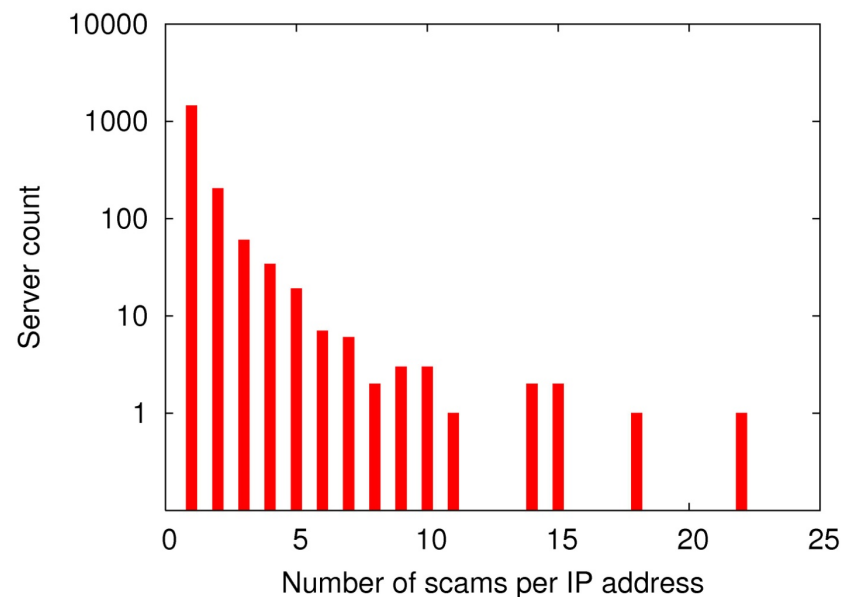
DISTRIBUTED INFRASTRUCTURE

<i>Scam category</i>	<i># of domains</i>	<i># of IPs</i>
Watches	3029	3
Pharmacy	695	4
Watches	110	3
Pharmacy	106	1
Software	99	3
Male Enhancement	94	2
Phishing	91	14
Viagra	90	1
Watches	81	1
Software	80	45

Figure 6: The ten largest virtual-hosted scams and the number of IP addresses hosting the scams.

SHARED INFRASTRUCTURE

- *To what extent do multiple scams share infrastructure?*
- 38% of the scams were hosted on machines hosting at least another scam
- Ten servers hosted ten or more scams
- Top three are:
22, 18 and 15



SHARING OVER TIME

- 96% of pairs overlapped in time
- 50% of pairs of scams overlapped for at least 125 hours
- Only 10% fully overlapped each other

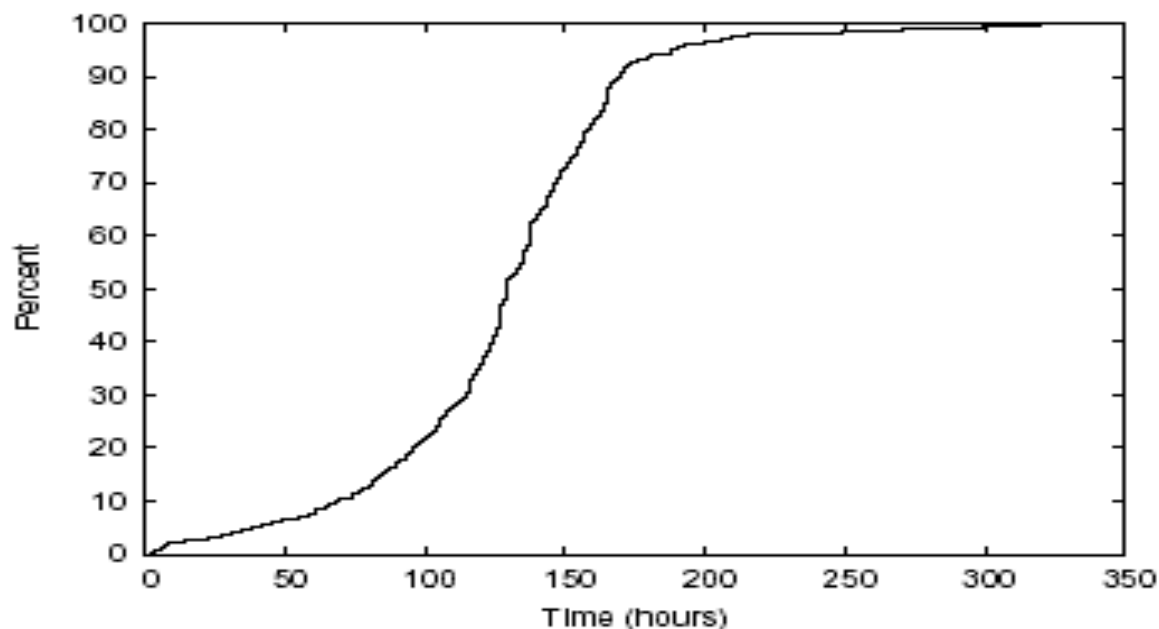


Figure 8: Overlap time for scam pairs on a server.

SHARED BETWEEN SCAM HOSTS AND RELAYS

- Overlap of 9.7 % between the scam hosts and relays
- Using blacklist queries:

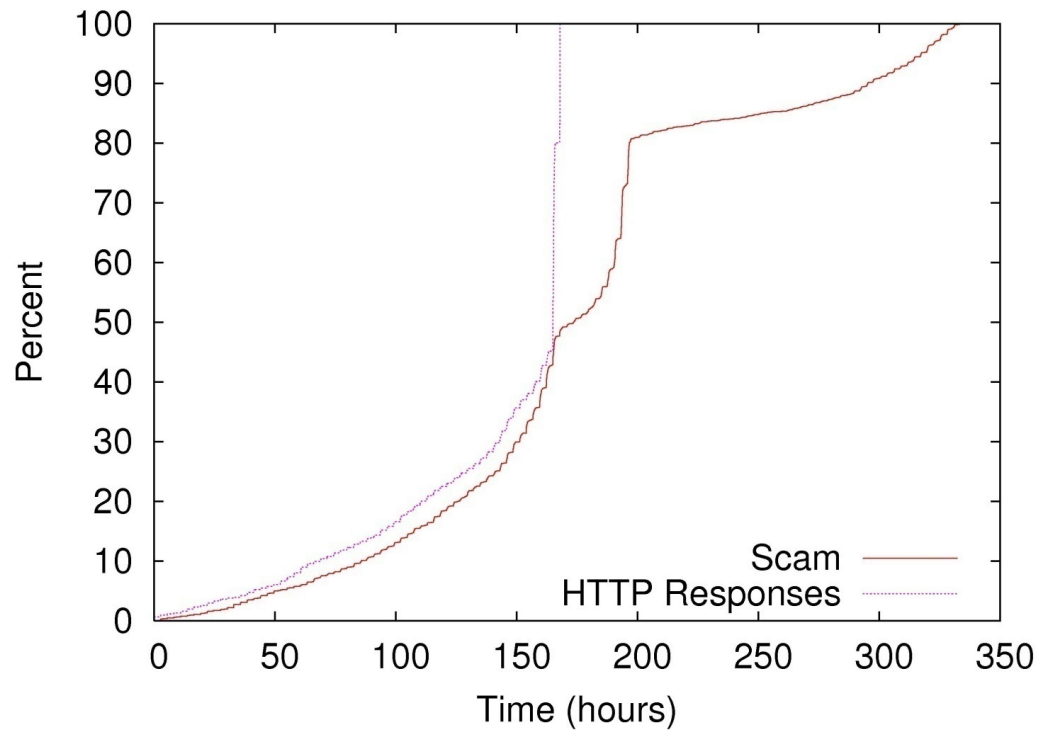
<i>Host type</i>	<i>Classification</i>	<i>% of hosts recognized</i>
Spam relay	Open proxy	72.3%
	Spam host	5.86%
Scam host	Open proxy	2.06%
	Spam host	14.9%

Table 3: Blacklist classification of spam relays and scam hosts.

LIFETIME

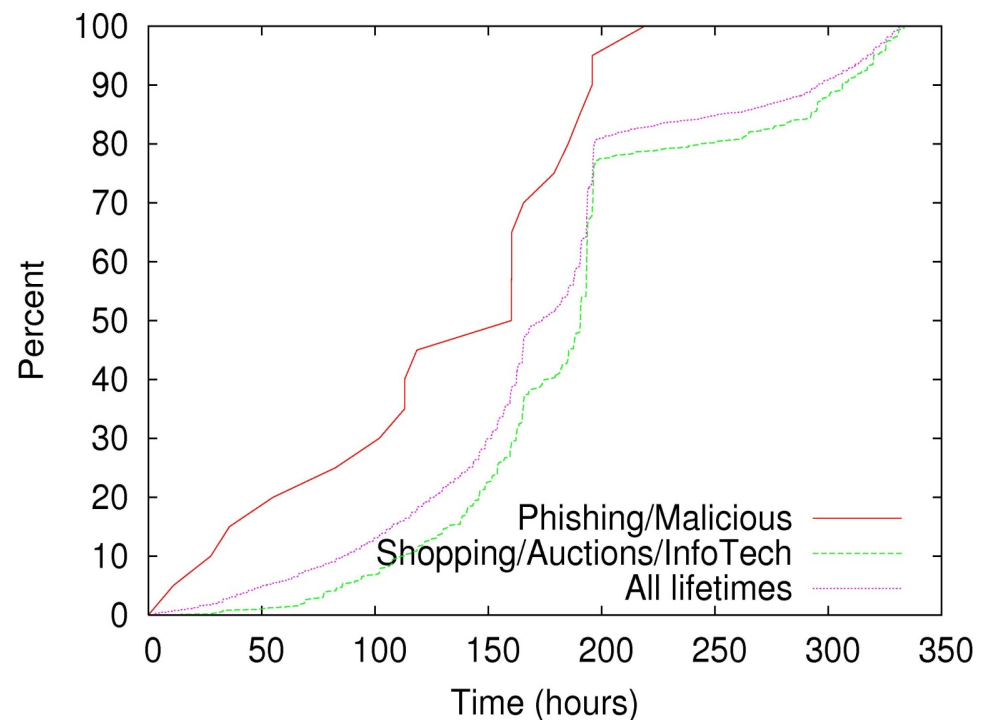
How long are scams alive?

Overall scam lifetime approached two weeks



LIFETIME BY CATEGORY

- Malicious scams have shorter lifetime
- Over 28% of malicious scams are blacklisted.
- More than 40% of malicious scam disappear before 120 hours
- Same is true for less than 15% of all scams



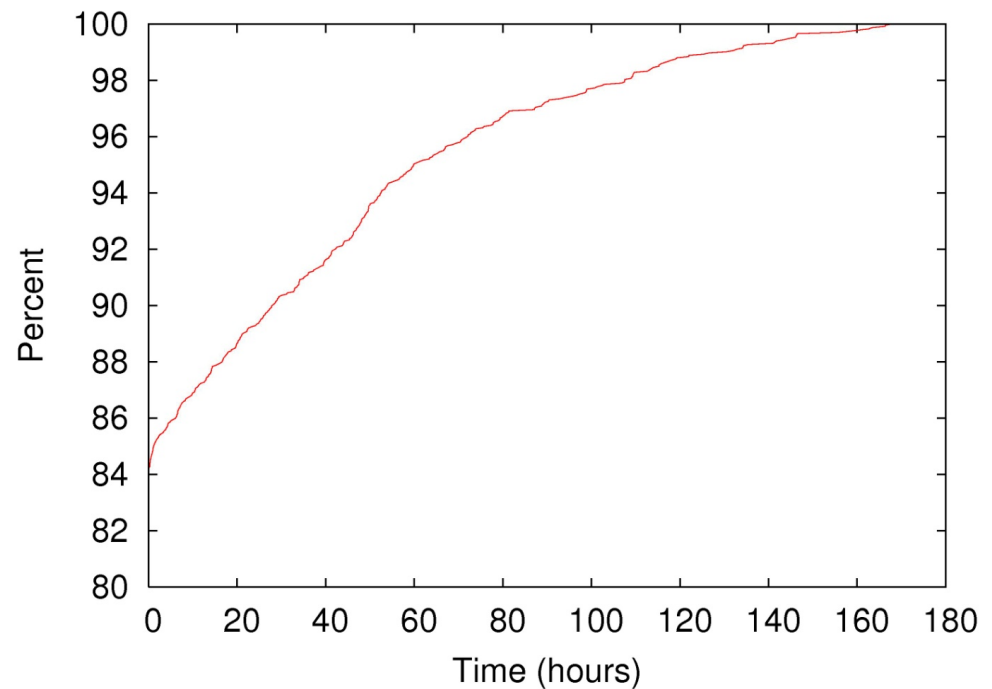
SPAM CAMPAIGN LIFETIME

How long do spam campaigns last for a scam?
137 spams messages per scam (avg)

Most spam campaigns relatively small – 88% last 20 hours or less

Only 8% last more than 2 days

Scam lifetimes longer- on average one week



STABILITY

- Availability is the number of successful web page downloads divided by the number of attempts
- Scam had excellent availability: over 90% had an availability of 99% or higher and most of the remaining had 98% or more availability
- As fingerprinted by p0f, more unix servers (43%) than windows (30%) and all of them had reported good link connectivity

LOCATION

- *Where are scam hosting servers located?*



Blue – Web servers
Red – Spam Relays

LOCATIONS

<i>Scam host country</i>	<i>% of all servers</i>
United States	57.40%
China	7.23%
Canada	3.70%
Great Britain	3.07%
France	3.06%
Germany	2.52%
Russia	1.80%
South Korea	1.77%
Japan	1.60%
Taiwan	1.53%
Other	16.32%

Table 4: Countries of scam hosts.

<i>Spam relay country</i>	<i>% of all relays</i>
United States	14.50%
France	7.06%
Spain	6.75%
China	6.65%
Poland	5.68%
India	5.42%
Germany	5.00%
South Korea	4.67%
Italy	4.44%
Brazil	3.86%
Other	30.97%

Table 5: Countries of spam relays.

CONCLUSION

- Perform some measurements of the scam infrastructure
- Employ the Spamsscatter technique for identifying the scam infrastructure
- More scams use one web server
- Life time of a scam is much longer than the spam
- Scam infrastructure more stable , longer lived , concentrated in US, compared with spam senders

