Characterization of Scam-Host Connectivity

Le Nolan, Robert Beverly, Joel Young

Motivation

1. On-line scams (pharmacy sales, phishing sites) continually evolve

2. Most recently, using multiple levels/types of indirection (HTTP, DNS)

3. Existing passive traffic analysis techniques rely on IP addresses, communication structure, redirection patterns, etc – can be evaded

4. Traffic characteristics should be agnostic to evasion

Current Research

"RB-Seeker: Auto-detection of Redirection Botnets"

- Parsing wget logfiles on redirection
- Netflow information
  - Short inter-flow duration, small flow size, short flow duration
- DNS log correlation
- Sequential Hypothesis Testing

"BotMiner: Clustering Analysis of Network Traffic for Protocol and Structure Independent Botnet Detection"

- No prior knowledge of botnets
- Clustering C&C communication
- C-plane Monitor: Net flow, A-plane Monitor: outbound traffic
- C-plane Clustering: finding clusters in monitor logs

"Behavioral Clustering of HTTP-Based Malware and Signature Generation Using Malicious Network Traces"

- Structural similarities between HTTP-based malware
- Automatically generating network signatures

Facts

1. Prior work finds significant redirection and traffic proxying by botnets

2. Scam content hosted by bot CDNs and by countries with poor connectivity

Hypothesis

Transport-layer traffic analysis of intermediate and landing pages reveal poor connectivity?

How connected are scam servers?

Scam Connectivity “Quality”

1. We’re agnostic to IP, DNS names, registrars, etc.

2. Collect *Transport-layer* traffic features that reveal:
   - Asymmetric bandwidth
   - Busy bots and/or poorly connected hosts

3. More detailed than NetFlow-style statistics:
   - Retransmits (in/out)
   - RSTs/FINs (in/out)
   - Congestion Window (min, zero)
   - 3WHS and per-segment RTT variance
   - Packet inter-arrival jitter

NetFlow Vs SpamFlow

NetFlow
IP Destination Address
IP Source Address
Source Port
Destination Port
Layer 3 Protocol Type
Class of Service
Router or Switch Interface

SpamFlow
Timestamp of first packet observed in flow
Number of packets from source and MTA
TCP segment retransmission from source and MTA
TCP reset segments from source and MTA
TCP segments with FIN bit set, from source and MTA
Number of times the congestion window went to zero
Minimum congestion window over flow life
Maximum flow idle time
RTT of the TCP three-way handshake
Inter-packet arrival variance
Per-segment RTT variance
Flow duration
TCP SYN window size
TCP SYN packet size
Fragment IP Bit
Arriving IP TTL

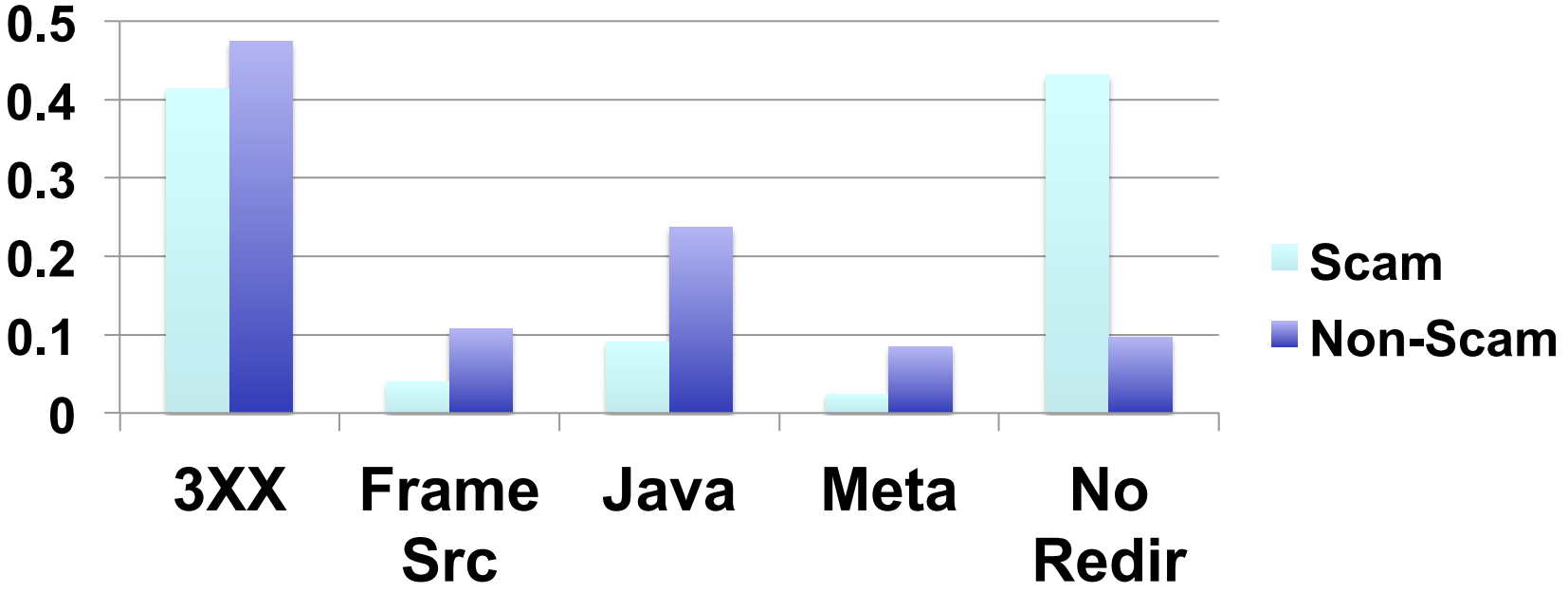
Experiment

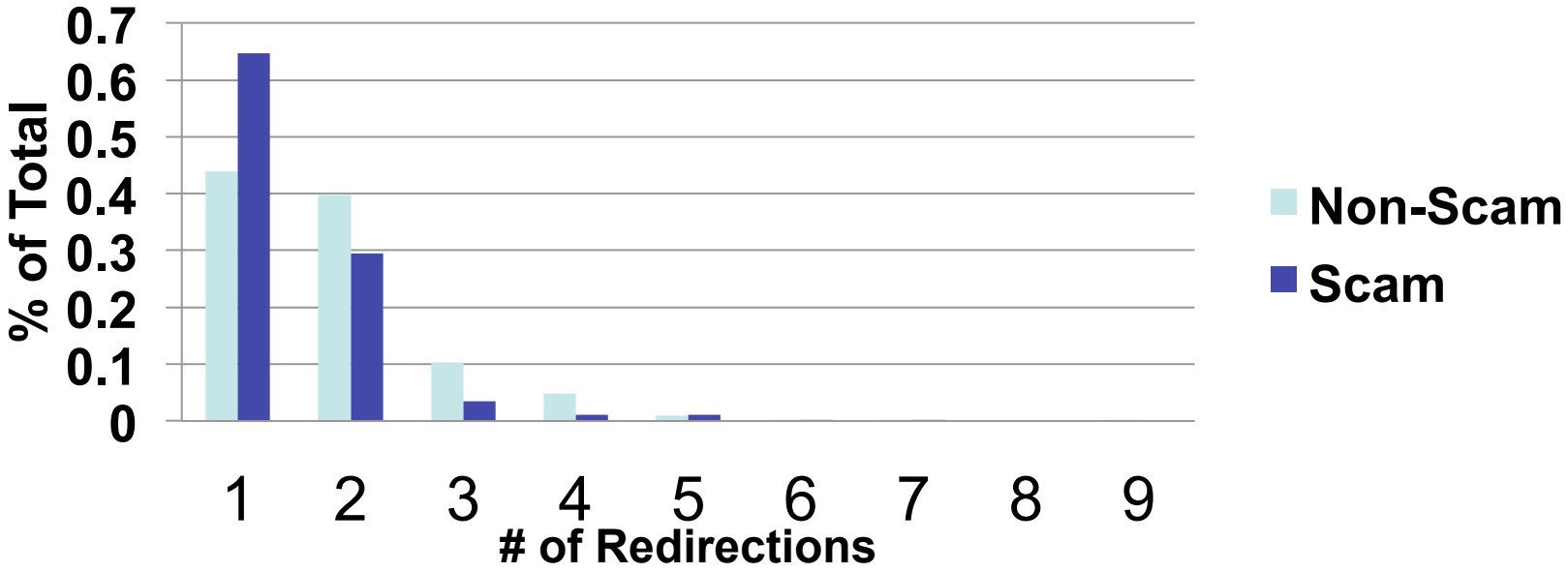
- Web-crawl: Alexa Top 10K and 35K known-scam URLs from spam sink
- Record transport layer information of each HTTP GET (including redirections):
- Find statistical discriminators between scam and non-scam hosts

Redirection Summary

- Scam URLs = 23,762, 1.45 per
- Non-Scam URLs = 3,075, 1.8 per
- Does redirection information still aid in discrimination?

Redirection Summary



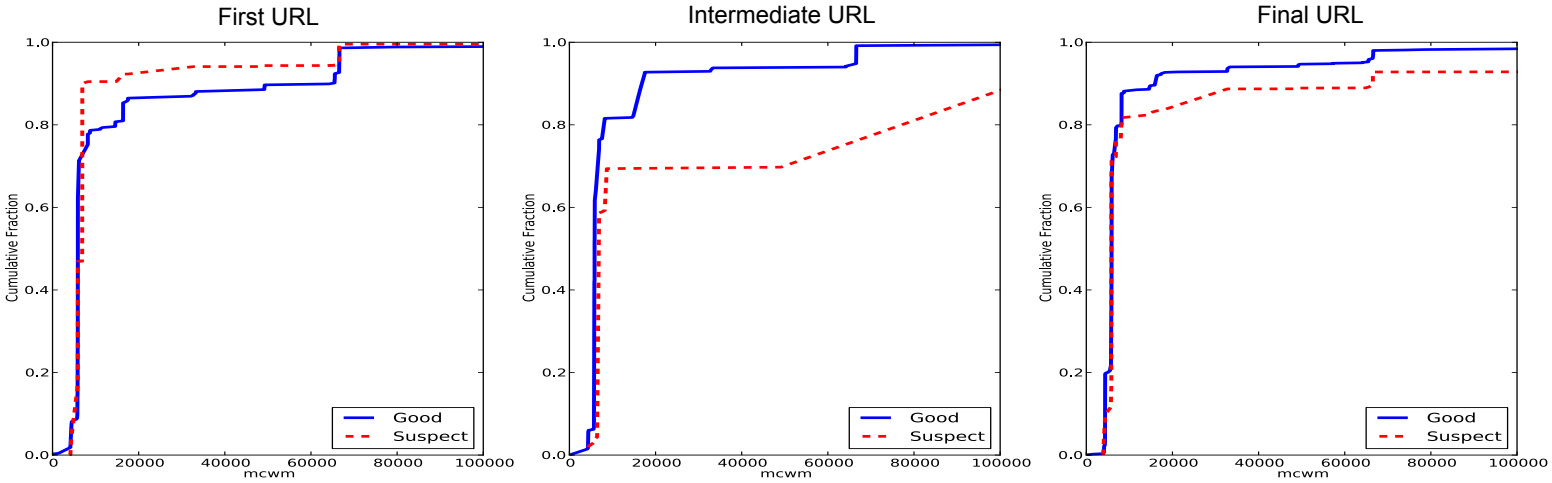


Transport-Layer Features

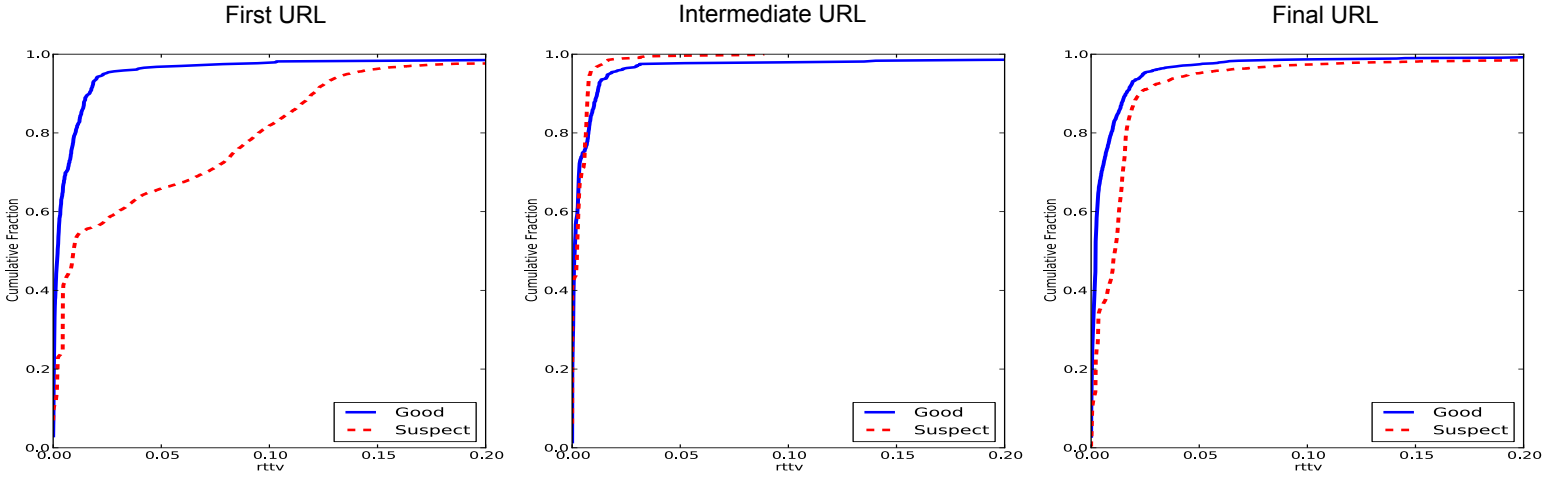
- Very different distributions (scam/non-scam) depending on redirection stage (initial, intermediate, terminal)
- Confirms previous observations that bots perform redirection

Transport-Layer Features

Minimum Congestion Window over Flow Lifetime



Estimated RTT Variance over Flow Lifetime



Classification

•Using data with 50% “good”, 50% “scam”:

Method	Acc	Sens	Spec	PPV	NPV
Bayes	0.760	0.715	0.808	0.795	0.731
SVM	0.874	0.816	0.935	0.929	0.830
Decision Tree	0.937	0.943	0.931	0.934	0.940

Future Work (Flow Analysis)

- Cookies and cookie behavior for scam infrastructure
- PHP and JavaScript code injection redirection
- Probing from multiple locations
- P2P Hosts Vs Spam Hosts similarities / differences
- Detect human behavior simulations

Future Work (Other)

- Use of legitimate in-line images
- Appearance of same URL / token across multiple emails/user
- URLs if Exif of jpgs
- Life of URLs