

Spamming botnets: Signatures and Characteristics

Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy,
Geoff Hulten, Ivan Osipkov
Microsoft Research, Microsoft Corporation

ACM SIGCOMM, Aug 2008

HY 558
Mar. 2010
S. Loutou

Introduction

- Botnet: A group of compromised host computers that are controlled by a small number of commander hosts
- Little effort devoted to understand the aggregate behaviors of botnets
- Previous work:
 - Zhuang et al. showed that similarity of email texts can help identify botnet-based spam campaigns
 - Li and Hsish found that spam emails with identical URLs are highly clusterable and often sent in burst

Novel Framework: AutoRE 1/2

- Identifies botnet hosts by generating botnet spam signatures from emails
- Focusing on URLs embedded in email content
- Why challenging to derive URL signatures that distinguish botnet spam from others?
 - Spam emails contain multiple URLs –some of which legitimate and general (e.g. <http://www.w3.org>)
 - Spammers add randomness into URLs to evade detection (polymorphic URLs)

Novel Framework: AutoRE 2/2

- AutoRE selects iteratively spam URLs based in the distributed and bursty property of campaigns
- Does not require labeled data or whitelists
- Outputs regular expression signatures that are more robust
- Groups emails into *spam campaigns*
- Further contribution: An in-depth analysis of identified spamming botnet characteristics and their activity trends

Key findings

- Botnets are becoming increasingly popular for spam delivery
- One botnet host is involved in multiple attacks
- Viewed individually, a botnet host does not exhibit distinctive sending patterns
- Detect botnet hosts by looking for aggregated common features from concurrent email sending activities
- Botnet attacks may have different phases

Contribution / Challenges

- Spammers add random legitimate URLs to increase legitimacy and HTML-based emails contain URLs generated by standard software
- AutoRE seeks both content prevalence and source address dispersion and ensures a low false positive rate by an iterative approach
- URL obfuscation techniques and customization of URLs to reflect recipients' email address
- AutoRE generates regular expressions

Time	URLs	Source ASes	URLs
2006-11-02	66	38	http://www.lympos.com/n/?167&carthagebolets http://www.lympos.com/n/?167&brokenacclaim http://www.lympos.com/n/?167&acceptoraudience
2006-11-15	72	39	http://shgeep.info/tota/index.html?jhjb.cvqxjby,hvx http://shgeep.info/tota/index.html?ikjija.cvqxjby,hvx http://shgeep.info/tota/index.html?ivvx_ceh.cvqxjby,hvx

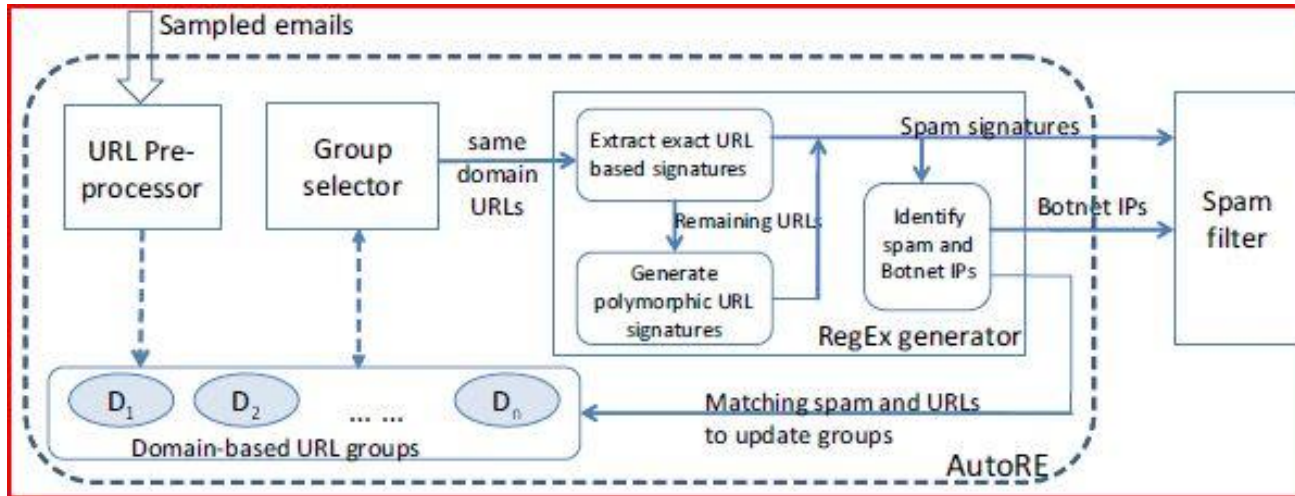
AutoRE framework 1/2

Input: a set of unlabeled email messages

Output: a set of spam signatures and a related list of botnet host IP addresses

- Seeks to discover email traffic patterns that are:
 - Bursty (emails originating from botnet hosts are sent in a highly synchronized fashion)
 - Distributed (Botnet hosts usually span a large and dispersed IP address space)

AutoRE framework 2/2



- **URL preprocessor**: extracts URLs and other relevant fields from input emails and groups them according to Web domains
- **Group selector**: Selects URL groups with the highest degree of burstiness in sending time and feeds them to:
- **RegEx generator**: Extracts signatures by processing one group at a time

URL Pre-Processing

- Given a set of emails, AutoRE extracts:
URL string, Source server IP address, Email sending time
- Partitions URLs into groups based on their **Web domains** (Motivation: Emails originating from the same spam campaign tend to advertise the same product/service from the same domain)

URL Group Selection

- Which group best characterizes an underlying spam campaign?
- Group Selector selects the URL group that exhibits the strongest temporal correlation across a large set of distributed senders (*bursty property*)
- *How?* Discrete time signal S that represents the number of distinct source IP addresses that were active during a time window w .

Signature Generation & Botnet Identification

- Input: a set of URLs pertaining to the same domain

Output: 2 types of signatures

- Complete URL based signatures
- Regular expression signatures. More generic and powerful, can detect spam emails with polymorphic URLs

Signature Generation & Botnet Identification: **Signature Criteria**

- **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
- **Specific:** Quantified using an information entropy metric pertaining to the probability of a random URL string matching the signature

Automatic URL regular Expression Generation 1/5

- Input: A set of polymorphic URLs from the same Web domain
- Construct a keyword-based signature tree
- Generating candidate regular expressions
- Evaluating the quality of them (specific enough)

Automatic URL regular Expression Generation 2/5 : Signature Tree Construction 1/2

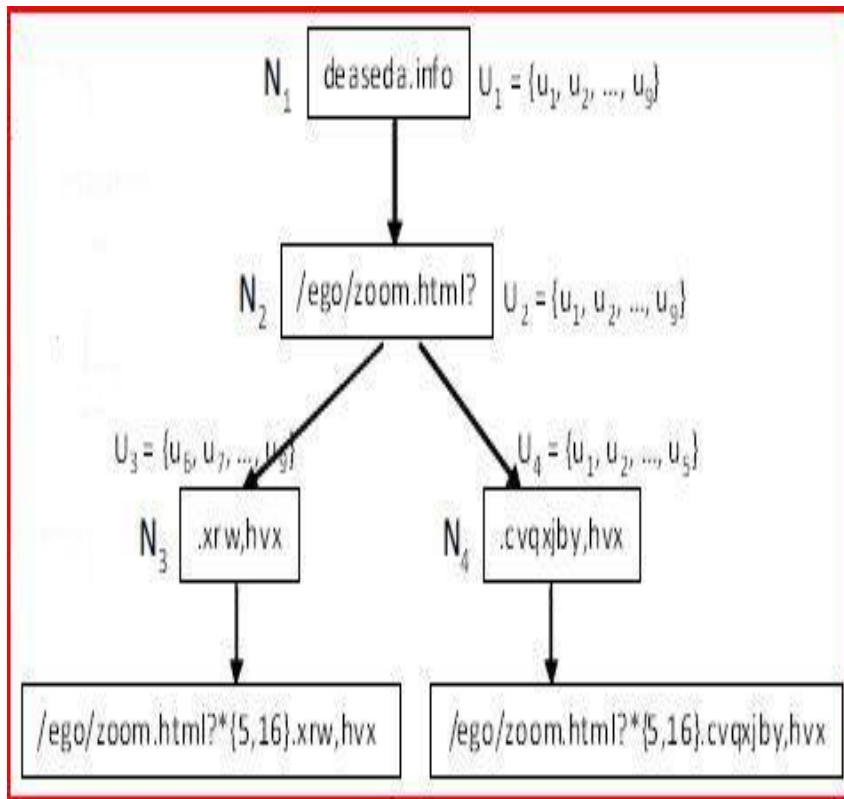
- Determine a candidate set of substrings
- Start with the most frequent substring that is both bursty and distributed
- Expand the signature by including more substrings to obtain a more specific signature
- Finally: Keyword-based signature tree where:

node \leftrightarrow substring

root \leftrightarrow domain name

path from root to a leaf \leftrightarrow signature

Automatic URL regular Expression Generation 3/5 : Signature Tree Construction 2/2



- Given a parent node
 - Look for the most frequent substring
 - If combining root satisfies the preset AS, create a new child node
- Multiple signatures?
- Correspond to different campaigns
 - Map to one campaign but occur with enough significance

Automatic URL regular Expression Generation 4/5:

Regular Expression Generation

- Given the keyword-based signatures
- **Detailing:** Returns a domain-specific regular expression using a keyword-based signature as input. Increase the quality of URL signatures to reduce false positive rates
- **Generalization:** Returns a more general domain-agnostic regular expression by merging very similar domain-specific expressions. Increase the coverage of botnet spam detection.

Automatic URL regular Expression Generation 5/5:

Signature Quality Evaluation

- Problem to face: The generalization process may produce overly general signatures
- Quantify the probability of a random string matching a signature: *Entropy Reduction*

$$P(e) = \frac{2^{B_e(u)}}{2^{B(u)}} = \frac{1}{2^{B(u)-B_e(u)}} = \frac{1}{2^{D(e)}}$$

- $D(e)=B(u)-B_e(u)$: *Entropy reduction*

More specific signature -> larger D(e)

Datasets and Results 1/2

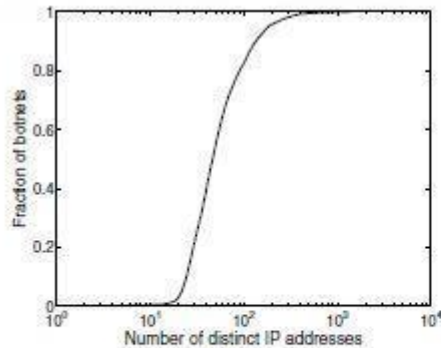
- Dataset collected in Nov 2006, Jun 2007, Jul 2007. 5.382.460 sampled emails

Month	Nov 2006		June 2007		July 2007		Total
	CU	RE	CU	RE	CU	RE	
Num. of spam campaigns	1,229	519	1835	591	2826	721	7,721
Num. of ASes	3,176	1,398	4,495	1,906	4,141	1,841	5,916
Num. of botnet IPs	88,243	23,316	113,794	19,798	85,036	29,463	340,050
Num. of spam emails	118,613	26,897	208,048	26,637	159,494	40,777	580,466
Total botnet IPs	100,293		131,234		113,294		340,050

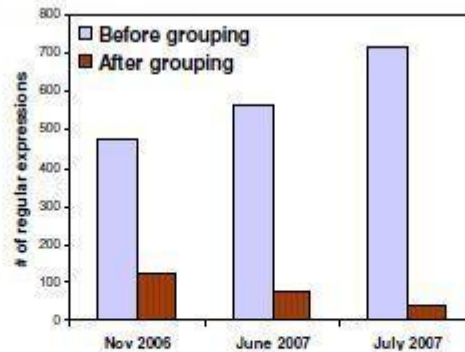
Table 1: Some statistics pertaining to the botnets identified by AutoRE.

- The majority of the campaigns belong to the CU category
- 20.4-29.7 adopted polymorphism
- Observe a steady upward trend in the number of identified campaigns
- The total number of botnet IPs/month does not increase proportionally -> Each botnet host is used more aggressively

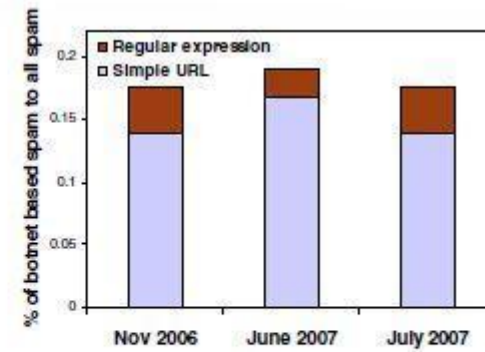
Datasets and Results 2/2



(a)



(b)



(c)

(a) Cumulative distribution of botnet size in terms of number of distinct IPs involved. (b) Number of regular expression patterns before and after generalization. (c) Percentage of spam captured by AutoRE signatures.

- No substantial difference in the shape of the distribution for the various months (a)
- Merging domain-specific regular expressions into domain-agnostic regular expressions, reduced the number of regular expressions (b)
→ Spammers use a limited number of automatic spam generation programs for polymorphic URLs
- 16-18% of the sampled emails were sent from botnet hosts (c)

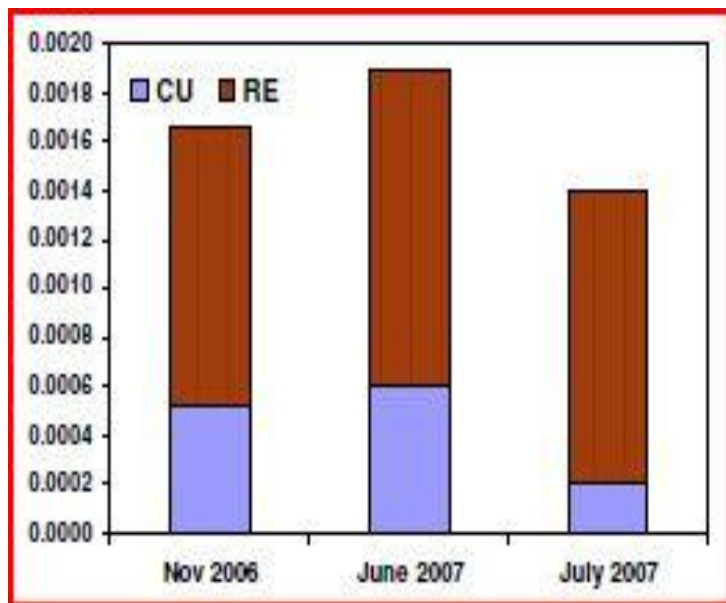
Botnet Validation 1/4 :

Botnet URL Signatures 1/2

Spam Detection

False Positive Rate

- For CU, 0.0001-0.0006
- For RE, 0.0011-0.0014



Ability to detect future spam

Month	Nov 2006			June 2007		
	CU	RE	Total	CU	RE	Total
# of spam emails	2	3	5	6,751	43,778	50529
# of non-spam emails	10	0	10	154	561	715

Number of spam and non-spam emails from July that match signatures derived from previous months

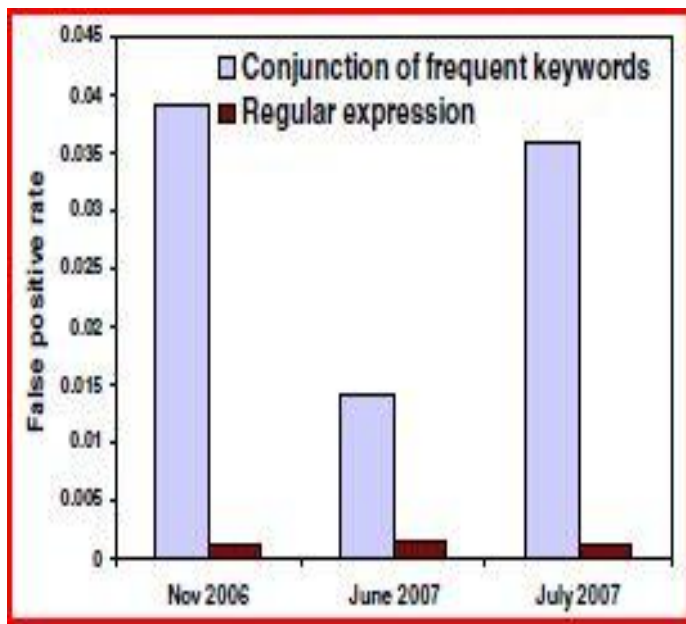
- Spam URL patterns evolve over time
- RE signatures are much more robust over time

Botnet Validation 2/4 :

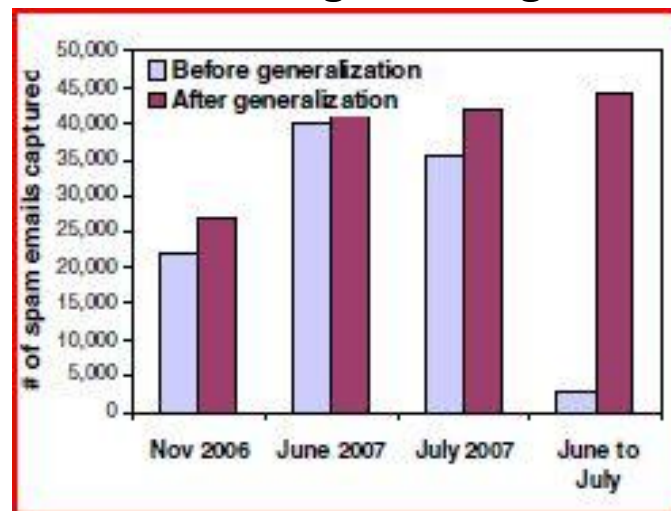
Botnet URL Signatures 2/2

Regular Expressions vs Keyword Expressions

- Their false positive rates differ dramatically



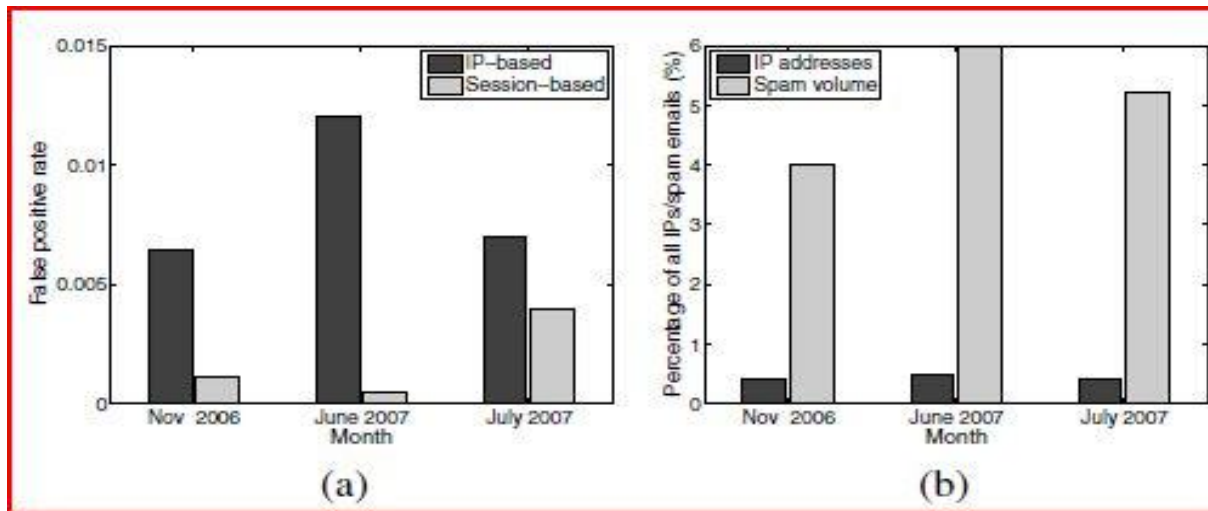
Domain Specific vs Domain Agnostic Signatures



- After generalization AutoRE can detect 9.9-20.6% more spam without affecting the false positive rates

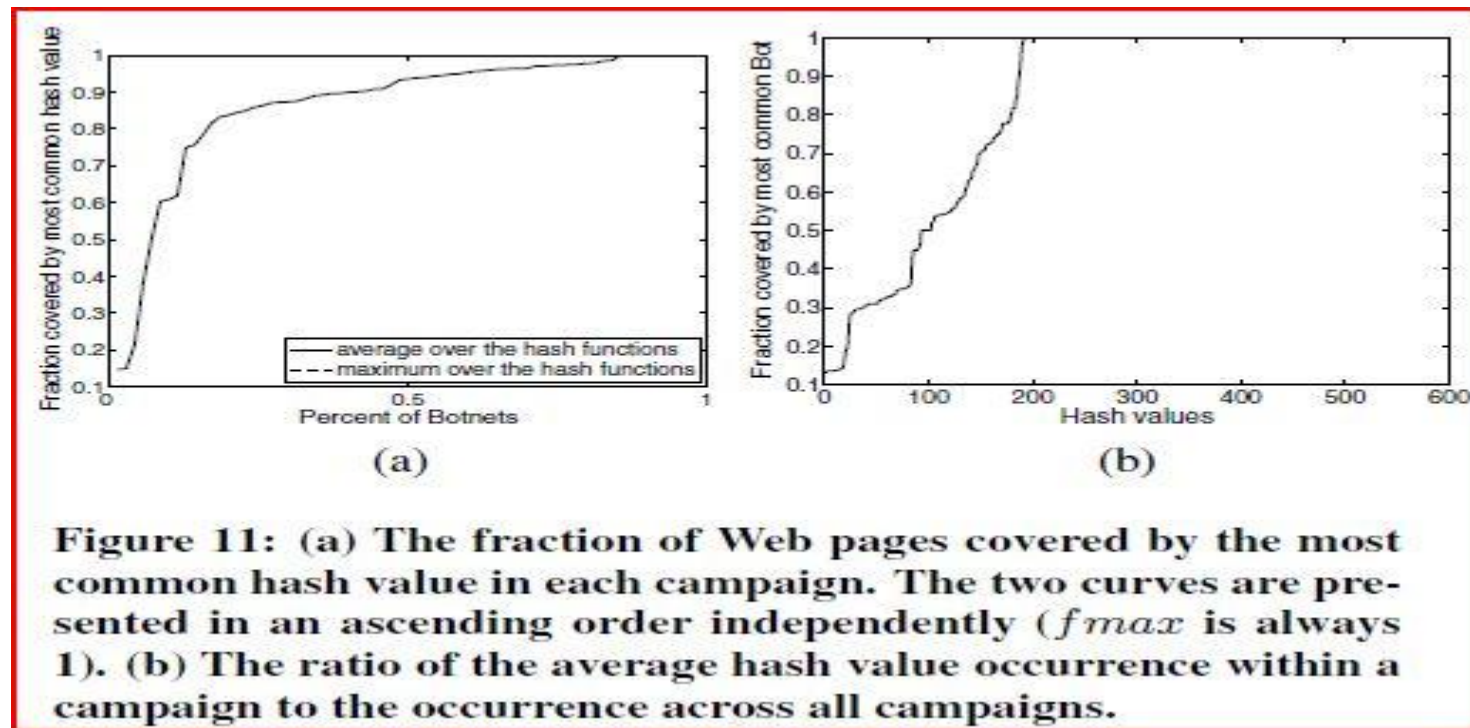
Botnet Validation 3/4 :

Botnet IP Addresses



- Determine if the identified hosts are indeed spammers
- Quantify the total amount of spam
- Very low false positive rate (a)
- The spam volume is non trivial

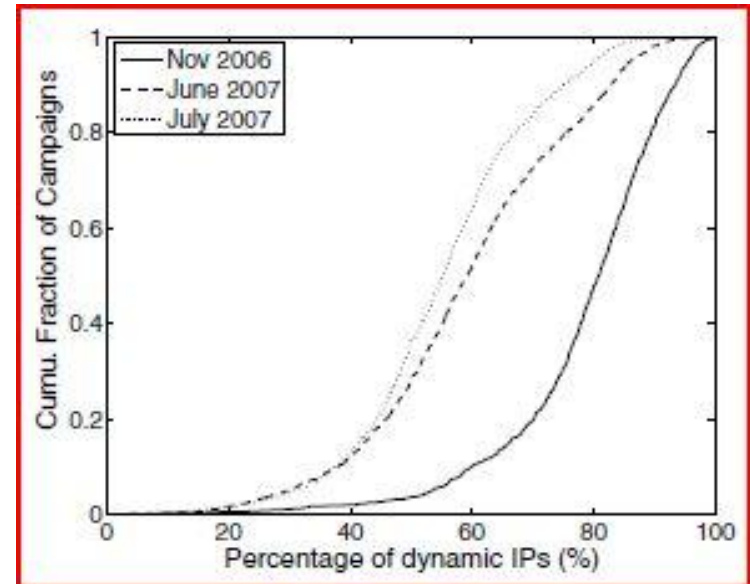
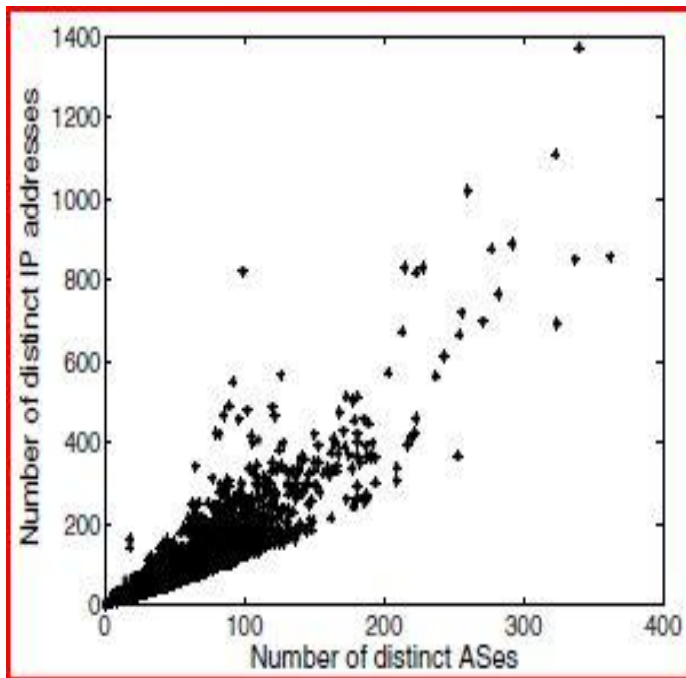
Botnet Validation 4/4 : Each campaign a group?



- Verify whether each spam campaign is correctly grouped together by computing the similarity of destination Web pages
- The web pages pointed to by each set of polymorphic URLs are similar to each other, while pages from different campaigns are different

Botnet Characteristics 1/6 :A General Perspective — Distribution of Botnet IP Addresses

- Botnet IP addresses are typically spread across a large number of ASes
- Each AS on average has only a few participating hosts

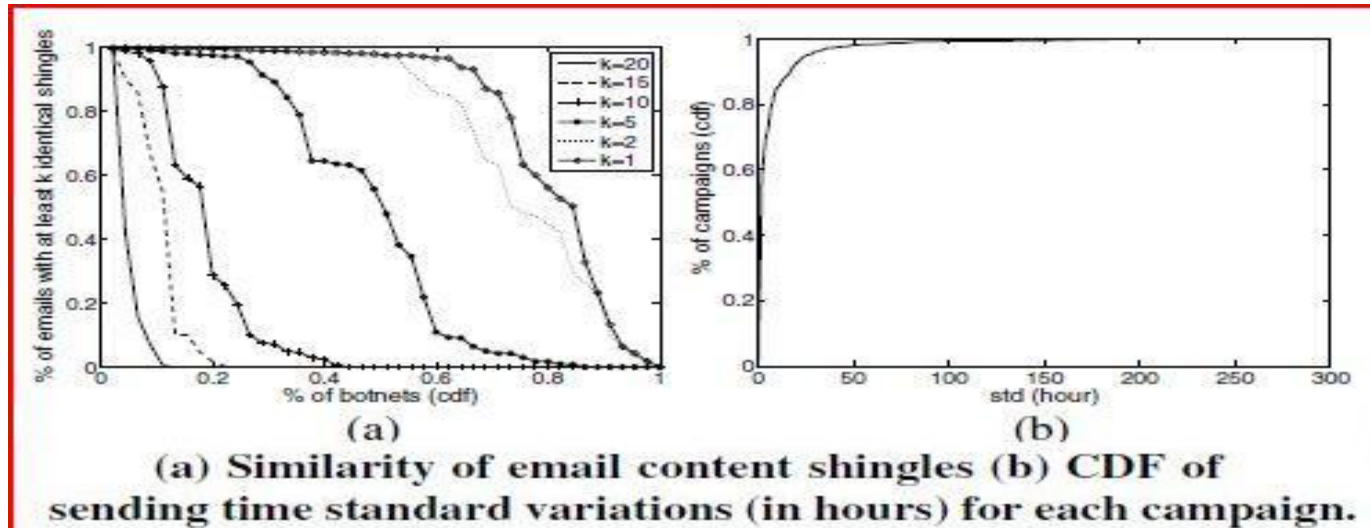


- Dynamic IP based hosts are popular targets for infection botnets
- The spam emails from botnets are switching from dynamic IP ranges to static IP addresses!

Botnet Characteristics 2/6:A General Perspective — Spam Sending Patterns

- Features to describe the sending patterns
 1. Number of recipients per email
 2. Connections per second
 3. Nonexisting recipient frequency
- (1)&(3), reflect the aggressiveness
(2), provides a measure on the amount of traffic destined to invalid addresses (scanning to obtain valid addresses)
- When viewed individually, botnet hosts do not exhibit distinct patterns

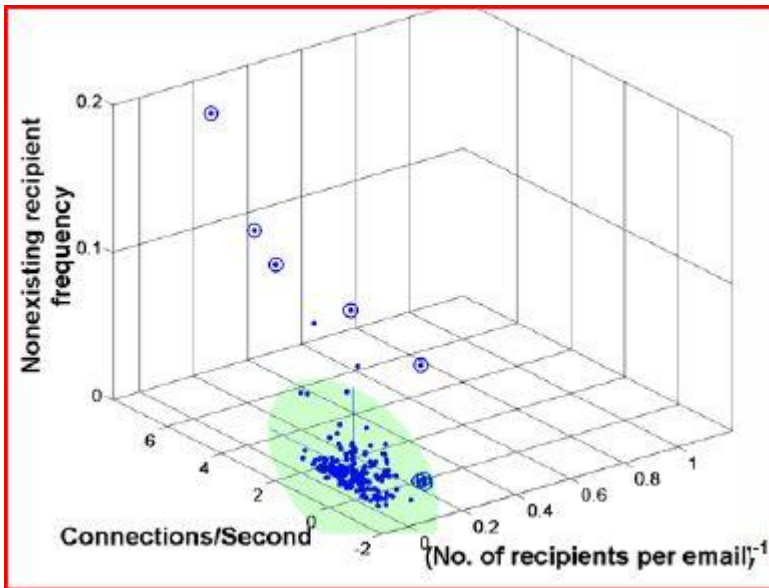
Botnet Characteristics 3/6 :An Individual Perspective — Similarity of Email Properties and Sending Time



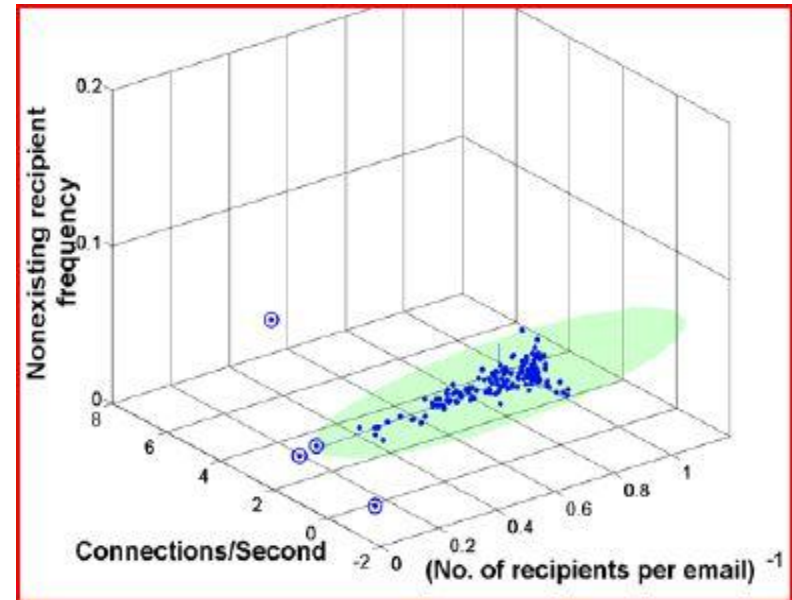
- The contents of the emails are quite different even though their target web pages are similar (a)
- 50% of the campaigns send the emails almost simultaneously (triggered by one command). The rest start sending whenever online.

Botnet Characteristics 4/6 :An Individual Perspective — Similarity of Email Sending Behavior

191 botnet hosts with 9 outliers



162 botnet hosts spanning 80 ASes

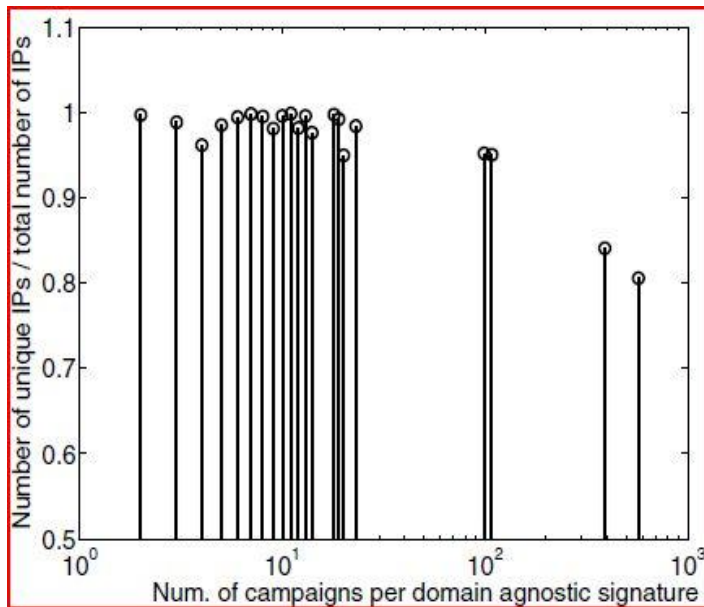


- The majority of the hosts are tightly clustered by having a similar number of recipients per email

- The hosts shared a constant connection rate in their communication with the server -> Botnet software may apply rate-control in initiating connections

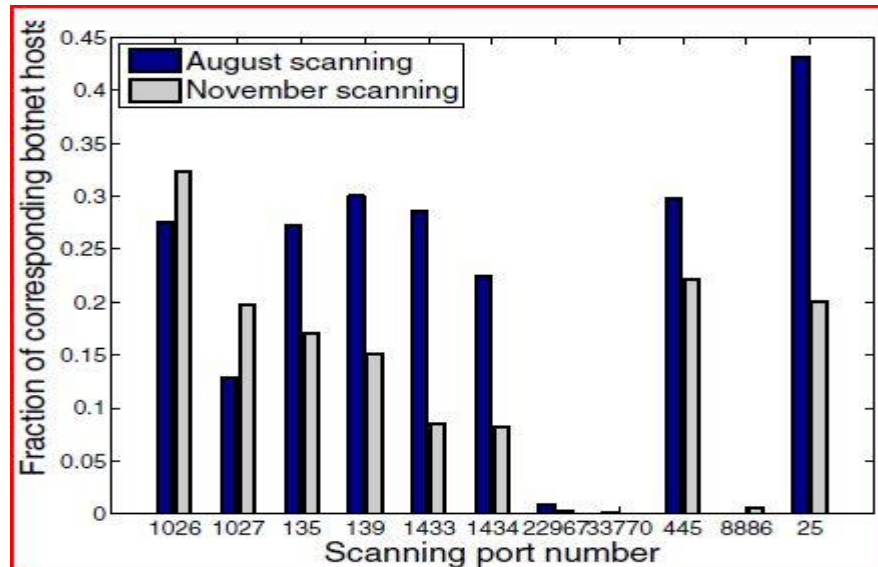
Botnet Characteristics 5/6 : Comparison of Different campaigns

- The corresponding botnets correspond to the same hosts?



Botnet Characteristics 6/6 : Correlation with Scanning Traffic

- Due to dynamic IP address assignment, using IP address as a host identifier for correlation is not robust



- The botnet attacks have different phases:
 - August: seek victim computers to expand the botnet size
 - November: launch spam attacks

Discussion 1/2

- Challenge: Work in real-time mode
- Spammers
 - Add legitimate URLs to confuse the URL selection process
 - They have no control of the sending frequency of legitimate URLs. Hence hard to select which URLs to include
 - Pollute the bursty feature by sending a spam URL from a few hosts before launching a large-scale attack
 - More robust signal processing methodology

Discussion 2/2

- No patterns in the URLs
 - Expect this scenario to be rare as the cost of registering domain makes this unattractive
- Possible case of false positive: Email flash crowd