Spamming Botnets: Signatures and Characteristics

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

Presented by Hee Dong Jung

Adapted from slides by Hongyu Gao and Yinglian Xie

Motivation

- Botnets have been widely used for sending spam emails at a large scale
- Detection and blacklisting is difficult as:
  - Each bot may send only a few spam emails
  - Attacks are transient in nature
- Little effort devoted to understanding aggregate behaviors of botnets from perspective of large email servers

Methodology

- Use email dataset from a large email service provider (MSN Hotmail)
- Focus on URLs embedded in email content
- Derive signatures for spam based on URLs
- Detect spam using signatures and find out characteristics of botnets

Methodology

- Challenges:
  - Random, legitimate URLs are added
  - URL obfuscation technique (polymorphic URLs, Redirection)

Email 1

http://www.shopping.com
http://www.w3.org/wai
http://www.psc.edu/networking/projects/tcp/
... ...
http://www.dvdfever.co.uk/co1118.shtml
... ...

Email 2

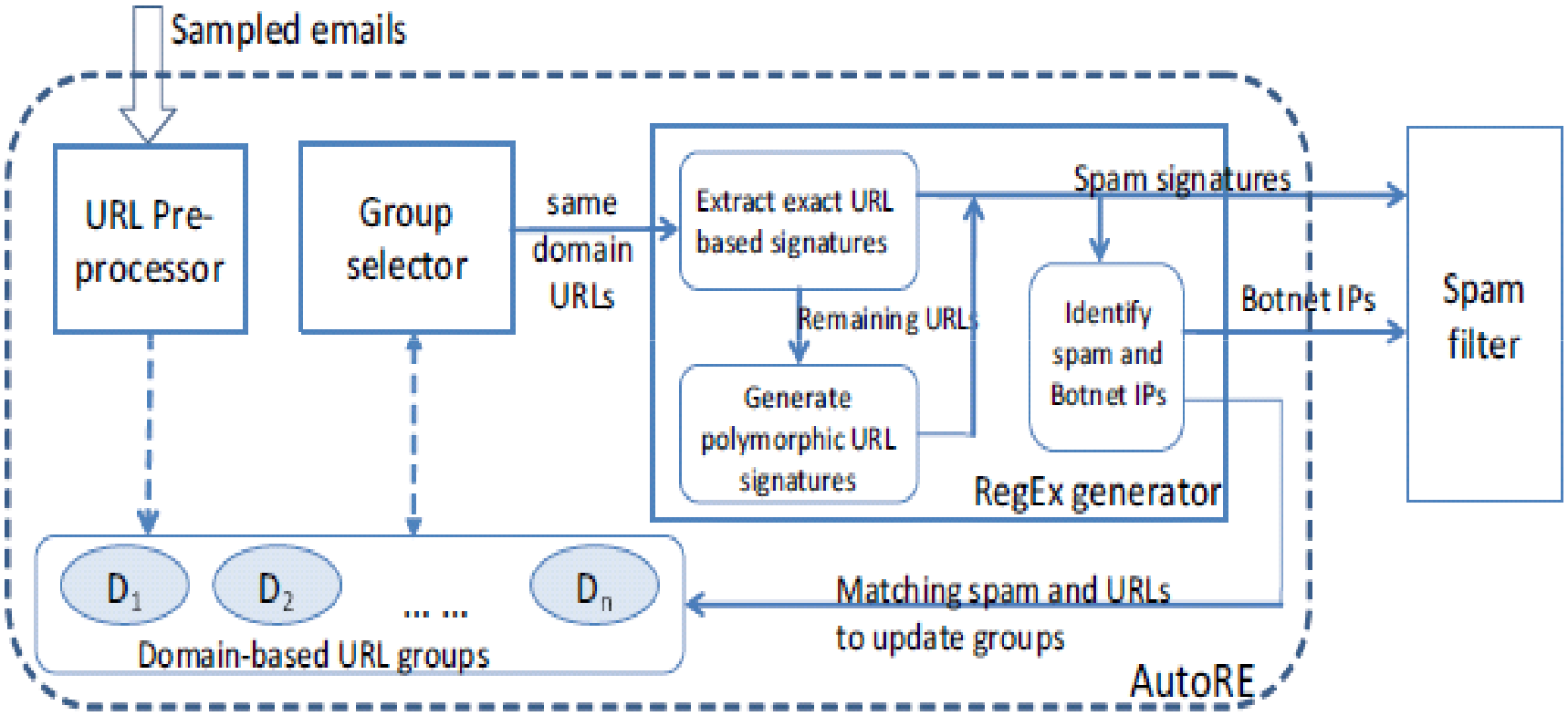
http://www.peacenvironment.net
http://www.w3.org/wai
http://www.bizrate.com
... ...
http://www.dvdfever.co.uk/co1118.shtml
... ...

Email 3

http://endosmosis.com/
http://www.talkway.com
http://www.bizrate.com
... ...
http://www.dvdfever.co.uk/co1118.shtml
... ...

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/indexx.html?jhjb.cvqxjby,hvx http://shgeep.info/tota/indexx.html?ikjija.cvqxjby,hvx http://shgeep.info/tota/indexx.html?ivvx_ceh.cvqxjby,hvx

AutoRE

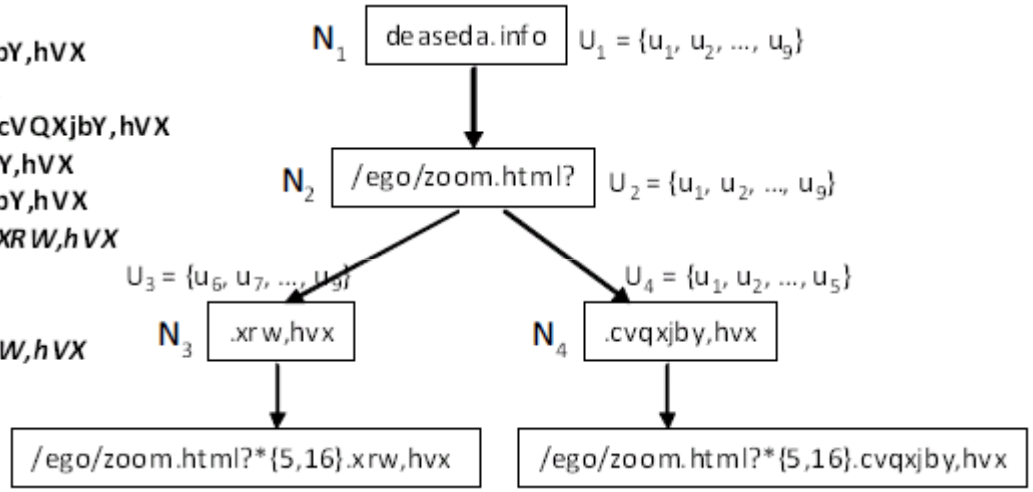


Is there a way to circumvent any of these steps?

Automatic URL Regular Expression Generation

- Signature Tree Construction

u_1: http://**deaseda.info**/ ego/zoom.html?QjQRP_xbZf.**cVQXjbY,hVX**
u_2: http:// **deaseda.info**/ ego/zoom.html? giAfS.**cVQXjbY,hVX**
u_3: http://**deaseda.info**/ ego/zoom.html? RQbWfeVYZfWifSd.**cVQXjbY,hVX**
u_4: http://**deaseda.info**/ ego/zoom.html? UbSjWcjHC.**cVQXjbY,hVX**
u_5: http://**deaseda.info**/ ego/zoom.html? VPS_eYVNfS.**cVQXjbY,hVX**
u_6: http://**deaseda.info**/ ego/zoom.html? QNVRcjgVNSbgfSR.***XRW,hVX***
u_7: http://**deaseda.info**/ ego/zoom.html? afRZXQ.***XRW,hVX***
u_8: http://**deaseda.info**/ ego/zoom.html? YcGGA.***XRW,hVX***
u_9: http://**deaseda.info**/ ego/zoom.html? aeSfLWVYgRIBH.***XRW,hVX***



- Regular Expression Generation
  - Detailing → Generalization

http://www.mezir.com/n/?167&[a-zA-Z]{9,25}
http://www.aferol.com/n/?167&[a-zA-Z]{10,27}
http://www.bedremf.com/n/?167&[a-zA-Z]{10,19}
http://www.mokver.www/n/?167&[a-zA-Z]{11,23}

http://*/n/?167&[a-zA-Z]{9,27}

http://arfasel.infoh/hums/jasmine.html?*{5,15}.[a-zA-Z]{3,7},hvx
http://apowefe.info/hums/jasmine.html?*{4,16}.[a-zA-Z]{3,7},hvx
http://carvalert.info/hums/jasmine.html?*{5,18}.[a-zA-Z]{3,7},hvx

http://*/hums/jasmine.html?*{4,18}.[a-zA-Z]{3,7},hvx

Datasets and Results

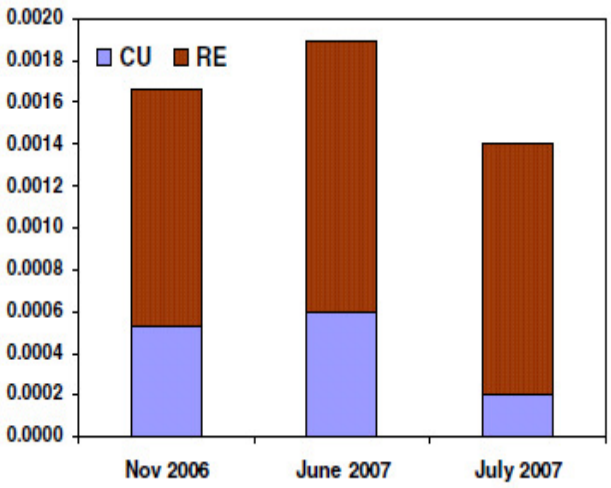
- Able to identify spam emails and related botnet hosts (IP addresses / ASes)

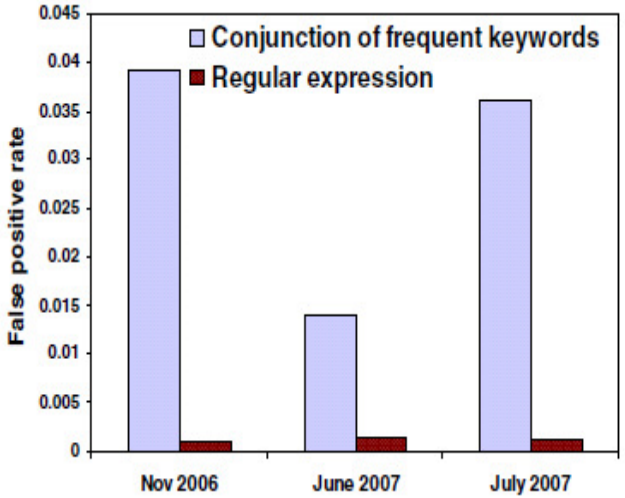
Month	Nov 2006		June 2007		July 2007		Total
	CU	RE	CU	RE	CU	RE	
Num. of spam campaigns	1,229	519	1835	591	2826	721	7,721
Num. of ASes	3,176	1,398	4,495	1,906	4,141	1,841	5,916
Num. of botnet IPs	88,243	23,316	113,794	19,798	85,036	29,463	340,050
Num. of spam emails	118,613	26,897	208,048	26,637	159,494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

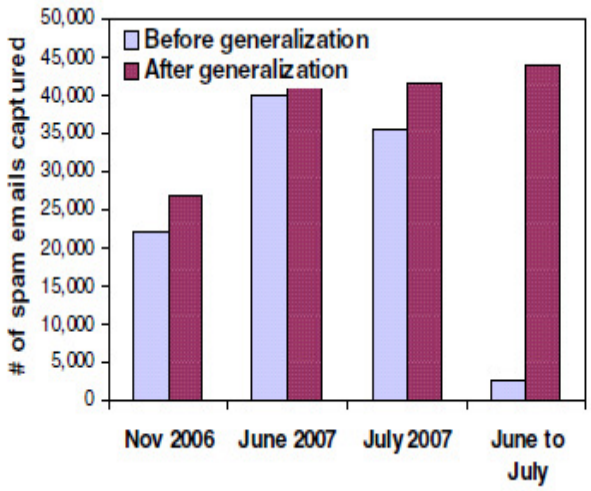
Table 1: Some statistics pertaining to the botnets identified by AutoRE.

AutoRE Performance

- Low False Positive Rate (between 0.0015 and 0.0020)
- Regular expressions reduce false positive rates by a factor of 10 to 30
- After generalization, AutoRE can detect 9.9 to 20.6% more spam without affecting false positive rates



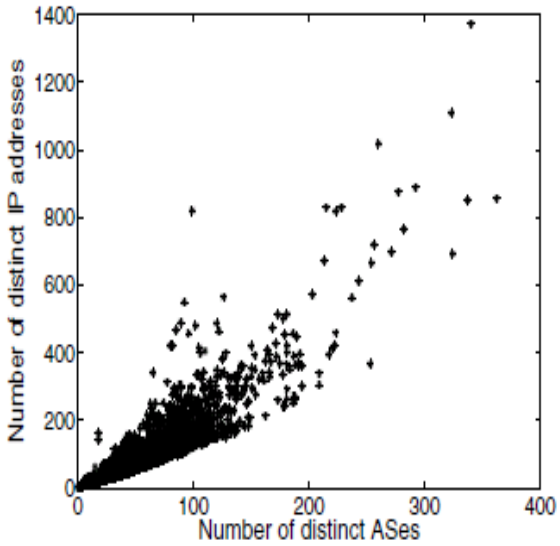


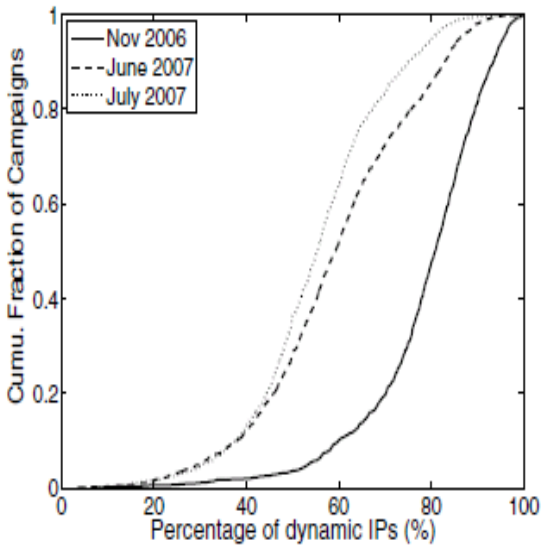


Spamming Botnet Characteristics

- Botnet IP addresses are spread across a large number of Ases
- 69% of botnet IP addresses are dynamic IPs; more than 80% of campaigns have at least half their hosts in dynamic IP ranges

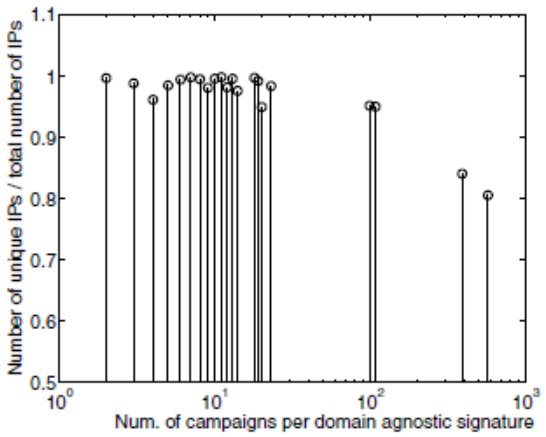
AS description	AS Number	Number of bot IPs
Korea Telecom	4766	15757
Verizon Internet service	19262	11426
France Telecom	3215	11303
China 169-backbone	4837	9960
Chinanet-backbone	4134	8113

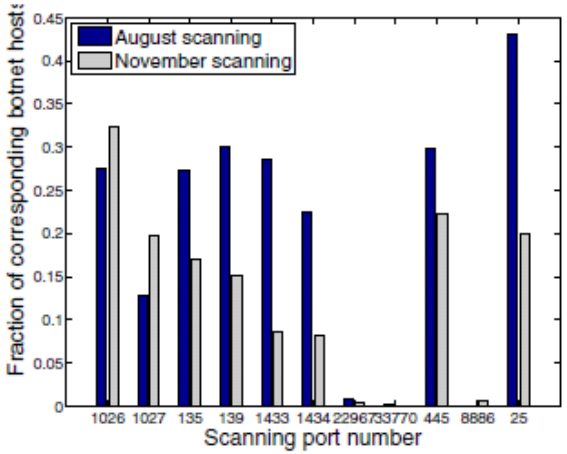




Spamming Botnet Characteristics

- Comparison of Different Campaigns
  - It is uncommon for different spam campaigns to overlap
- Correlation with Scanning Traffic
  - Amount of scanning traffic in Aug is higher than in Nov, when botnet IPs were used to send spam
  - Suggests that botnets could have different phases





Discussion and Conclusion

- AutoRE has potential to work in real-time mode
- Leverages bursty and distributed features of botnet attacks for detection
- Major Findings
  - Botnet hosts are widespread across Internet, with no distinctive sending patterns when viewed individually
  - Existence of botnet spam signatures and feasibility of detecting botnet hosts using them
  - Botnets are evolving and getting increasingly sophisticated

Discussion Points

- Do you think “Bursty” and “Distributed” properties represent the spam emails?
  - Are there other properties that should be considered?
- When would this URL based approach not work?

Thank you

Questions?

AutoRE

- Framework for automatically generating URL signatures
- Takes set of unlabeled email messages, produces 2 outputs:
  - Set of spam URL signatures
  - Related list of botnet host IP addresses
- Iteratively selects spam URLs based on distributed yet bursty property of botnets-based spam campaigns
- Uses generated spam URL signatures to group emails into spam campaigns

Group Selector (backup)

- Explores the bursty property of botnet email traffic
- Construct n time windows
- $S_i(k)$ is defined as the total number of IP addresses that sent at least one URL in group i in window k
- URL groups with sharp spikes are higher ranked

Automatic URL Regular Expression Generation (backup)

- Signature Quality Evaluation
  - Quantitatively measures quality of signature and discards signatures that are too general
  - Metric: entropy reduction
    - Leverages on information theory to quantify probability of a random string matching a signature
    - Given a regular expression e, let $B_e(u)$ and $B(u)$ denote expected # bits to encode a random string u with and without signature
    - Entropy reduction $d(e) = B(u)-B_e(u)$ reflects probability of arbitrary string with expected length allowed by e and matching e, but not encoded using e

Botnet Validation

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

Spamming Botnet Characteristics

- For each campaign, standard deviation (std) of spam email sending time is computed
  - 50% of campaigns have std less than 1.81 hours
  - 90% of campaigns have std less than 24 hours and likely located at different time zones
- For each campaign, host sending patterns are generally well-clustered
  - Number of recipients per email
  - Connection rate
- Botnet hosts do not exhibit distinct sending patterns for them to be identified