Spamming Botnets: Signatures and Characteristics

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

Presented by Hongyu Gao

Mar. 04 2009

Outline

- Motivation
- Introduction
- Design of AutoRE
- Experimental Results
- Spamming Botnet Characteristics
- My Comments

Motivation

- Botnets have been widely used for sending spam emails at a large scale.
  - Detecting and blacklisting individual bots is difficult.
  - Little effort has been devoted to understanding the aggregate behaviors of botnets.

Introduction

- Botnet
  - A group of compromised host computers (bots)
  - Controlled by a small number of commander hosts (bot masters)

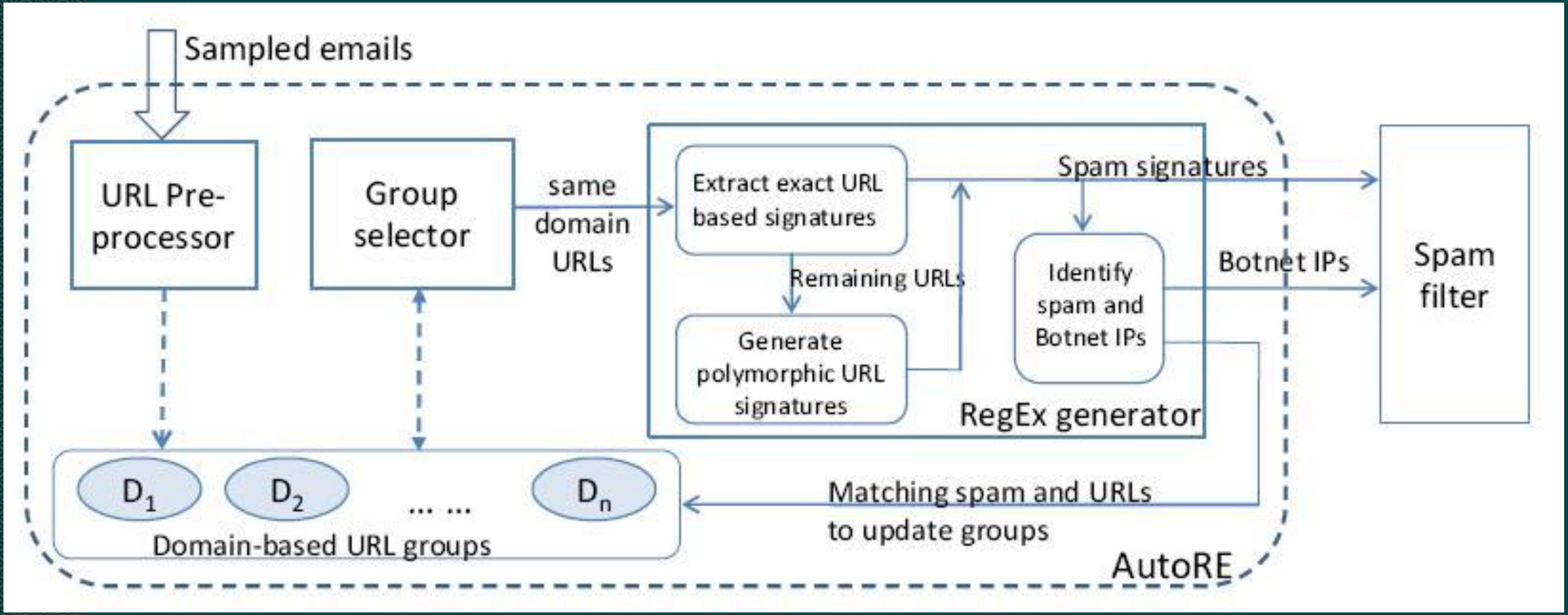
Introduction, cont'd

- High level idea
  - 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

AutoRe: Signature Based Botnet Identification

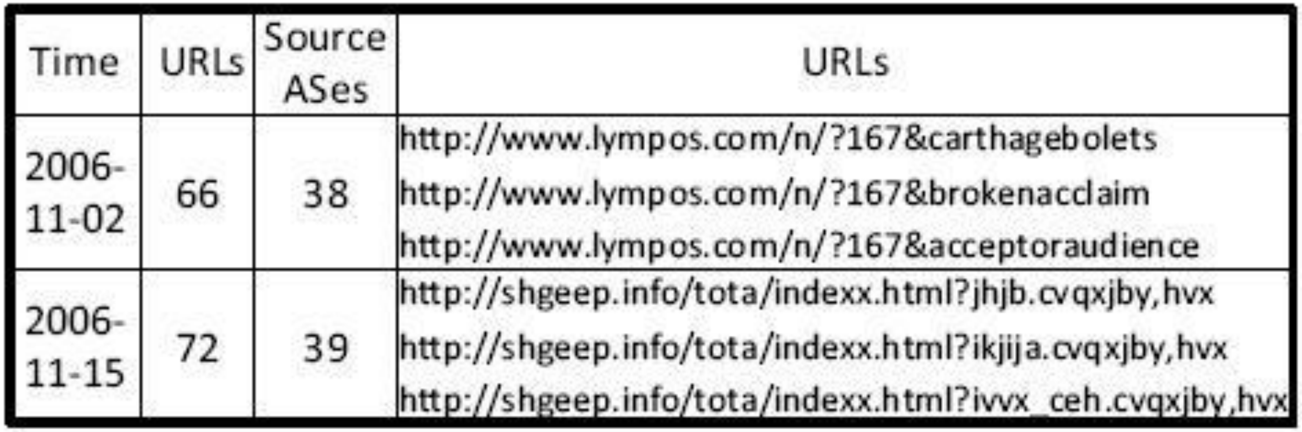
- A completely automatic tool
- Take as input a group of emails
- Produce a set of spam URL signatures and a list of botnet host IP addresses
- Three modules:
  - URL preprocessor
  - Group selector
  - RegEx generator

AutoRe: Signature Based Botnet Identification, Cont'd



URL Pre-processing

- Extract URL string, source server IP address and email sending time
- Partition URLs into groups beased on their Web domains



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

Figure 2: Examples of polymorphic URLs.

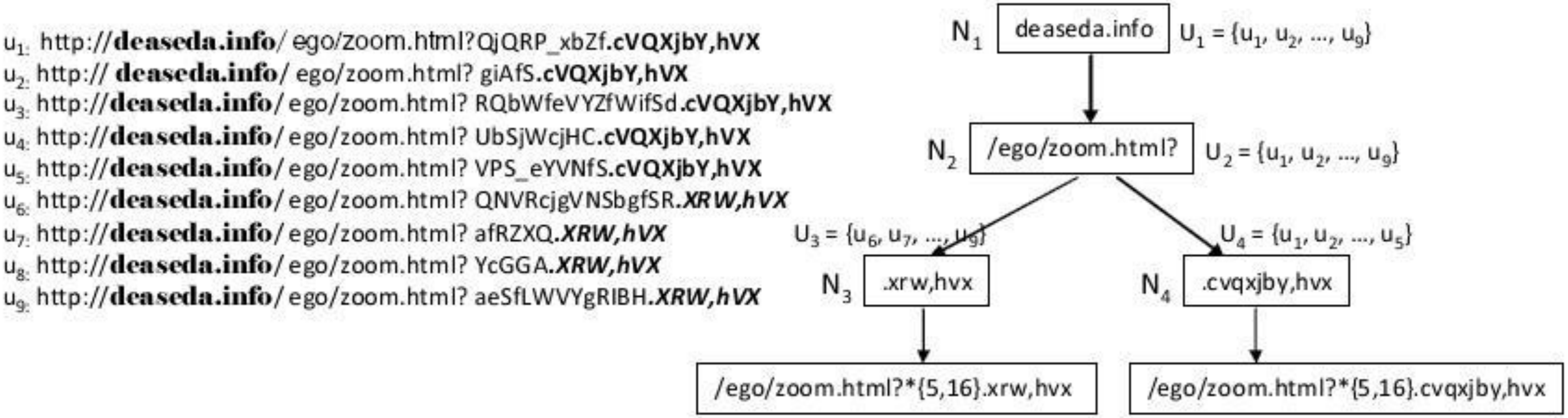
URL Group Selection

- Assume the bursty property of botnet email traffic
- Construct n time window
- $S_i(k)$ 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

Signature Tree Construction

- The root node is set to the domain name
- Start with the most bursty and distributed substring
- Incrementally expand the signature tree
- Until no eligible substring remains
- The path from root to leaf defines a keyword-based signature

Signature Tree Construction, Cont'd



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***

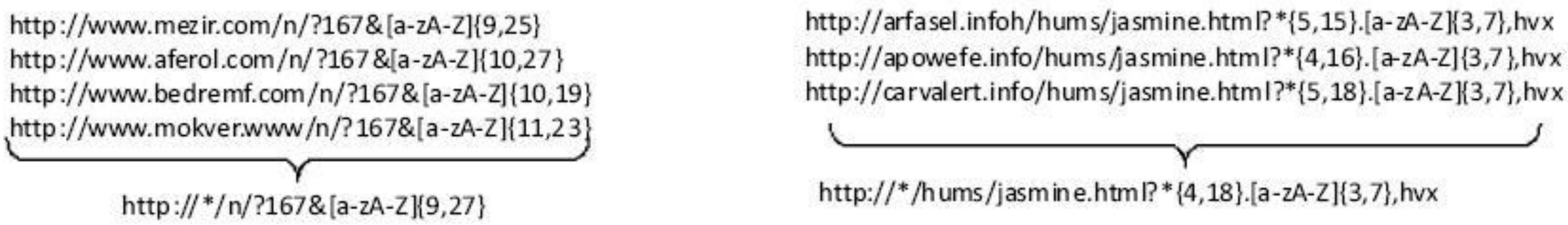
N_1
deaseda.info
$U_1 = \{u_1, u_2, \ldots, u_9\}$
N_2
/ego/zoom.html?
$U_2 = \{u_1, u_2, \ldots, u_9\}$
$U_3 = \{u_6, u_7, \ldots, u_9\}$
$U_4 = \{u_1, u_2, \ldots, u_5\}$
N_3
.xrw,hvx
N_4
.cvqxjby,hvx
/ego/zoom.html?*{5,16}.xrw,hvx
/ego/zoom.html?*{5,16}.cvqxjby,hvx

Figure 5: Example input URLs and the keyword-based signature tree constructed by AutoRE.

Regular Expression Generation

- The detailing process
  - Given the keyword-based signatures, apply a set of predefined rules to generate regular expressions for the substring between keywords.
- The generalization process
  - Takes the generated regular expressions and further groups them.

Regular Expression Generation, Cont’d



```
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}
```

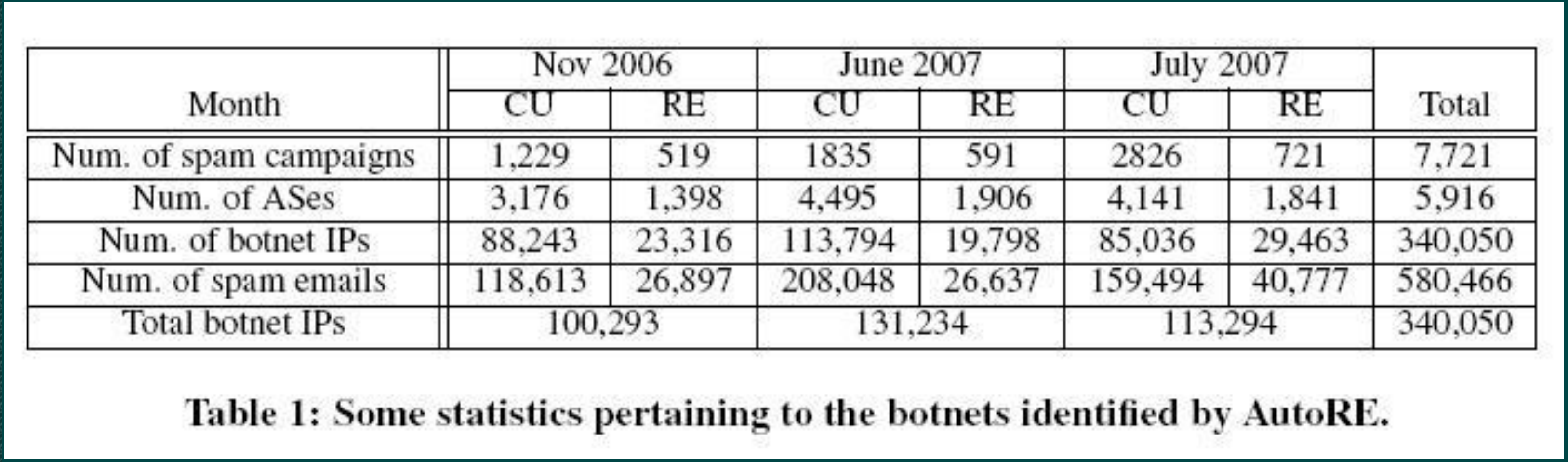
```
http://arfasel.infoh/hums/jasmine.html?*{5,15}.[a-zA-Z]{3,7},hvx
http://apowefe.info/hums/jasmine.html?*{4,16}.[a-zA-Z]{3,7},hvx
http://carvalert.info/hums/jasmine.html?*{5,18}.[a-zA-Z]{3,7},hvx

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

Figure 6: Generalization: Merging domain-specific regular expressions into domain-agnostic regular expressions.

Evaluation

- Emails were sampled from Nov. 2007, Jun. 2007 and Jul. 2007 (sampling rate 1:25000)

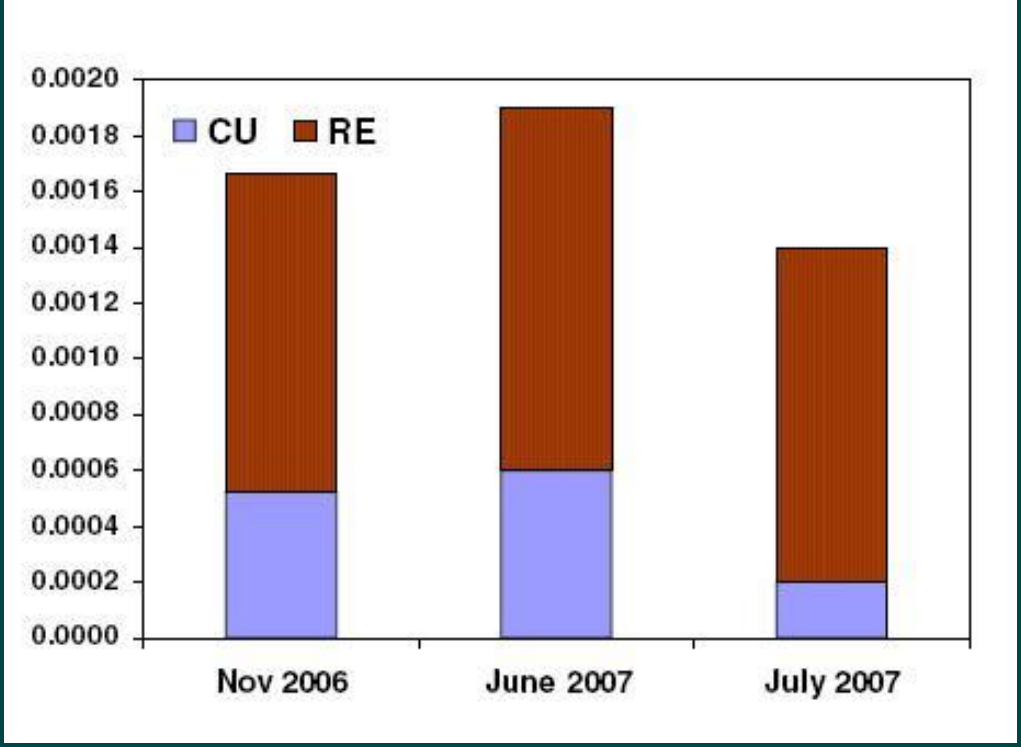


Month	Nov 2006		June 2007		July 2007		Total
	CU	RE	CU	RE	CU	RE	
Num. of spam campaigns	1,229	519	1835	591	2826	721	7,721
Num. of ASes	3,176	1,398	4,495	1,906	4,141	1,841	5,916
Num. of botnet IPs	88,243	23,316	113,794	19,798	85,036	29,463	340,050
Num. of spam emails	118,613	26,897	208,048	26,637	159,494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

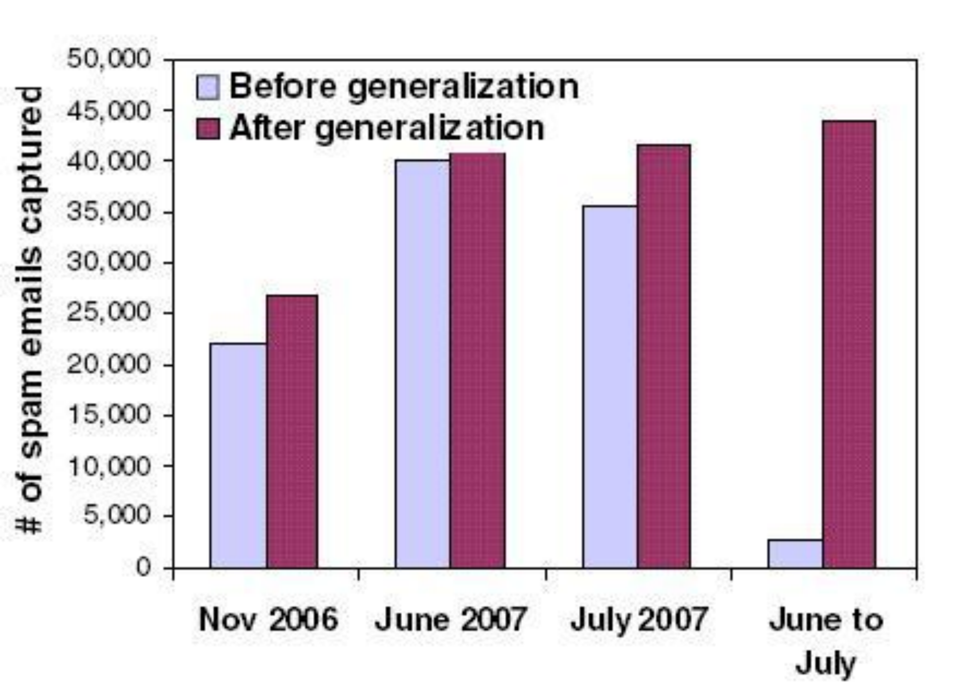
Evaluation, Cont'd

- Low false positive rate



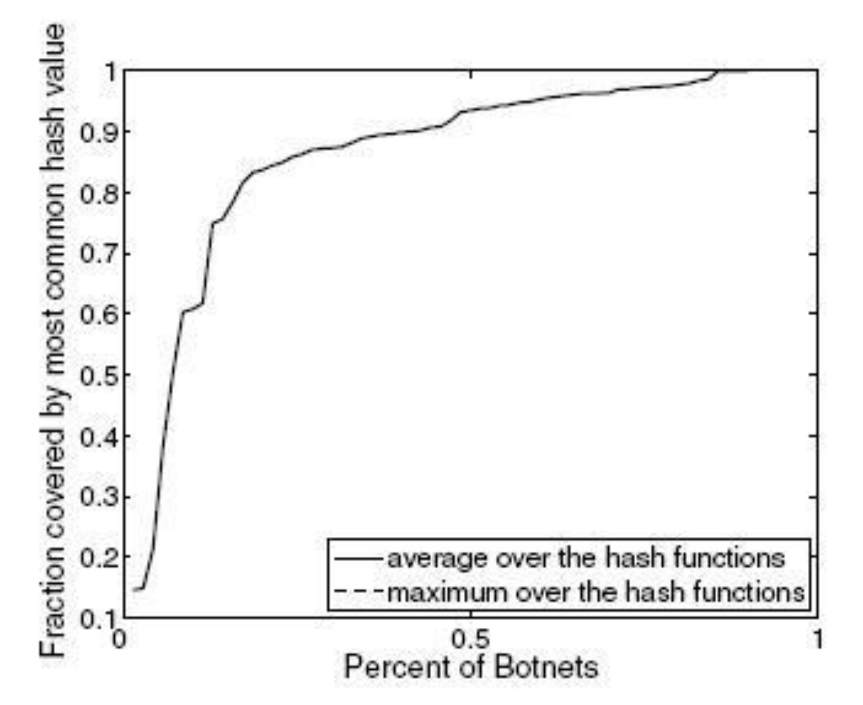
Evaluation, Cont'd

- Domain-agnostic generation improves the detection rate without affecting false positive rate.



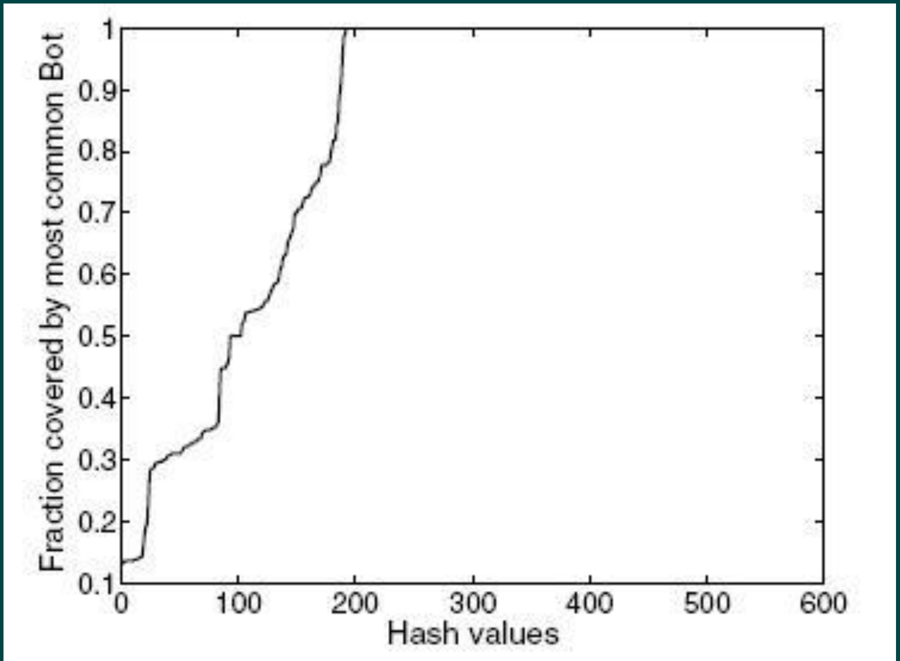
Evaluation, Cont'd

- For most spam campaigns, 90% of the destination Web pages are at least 75% similar



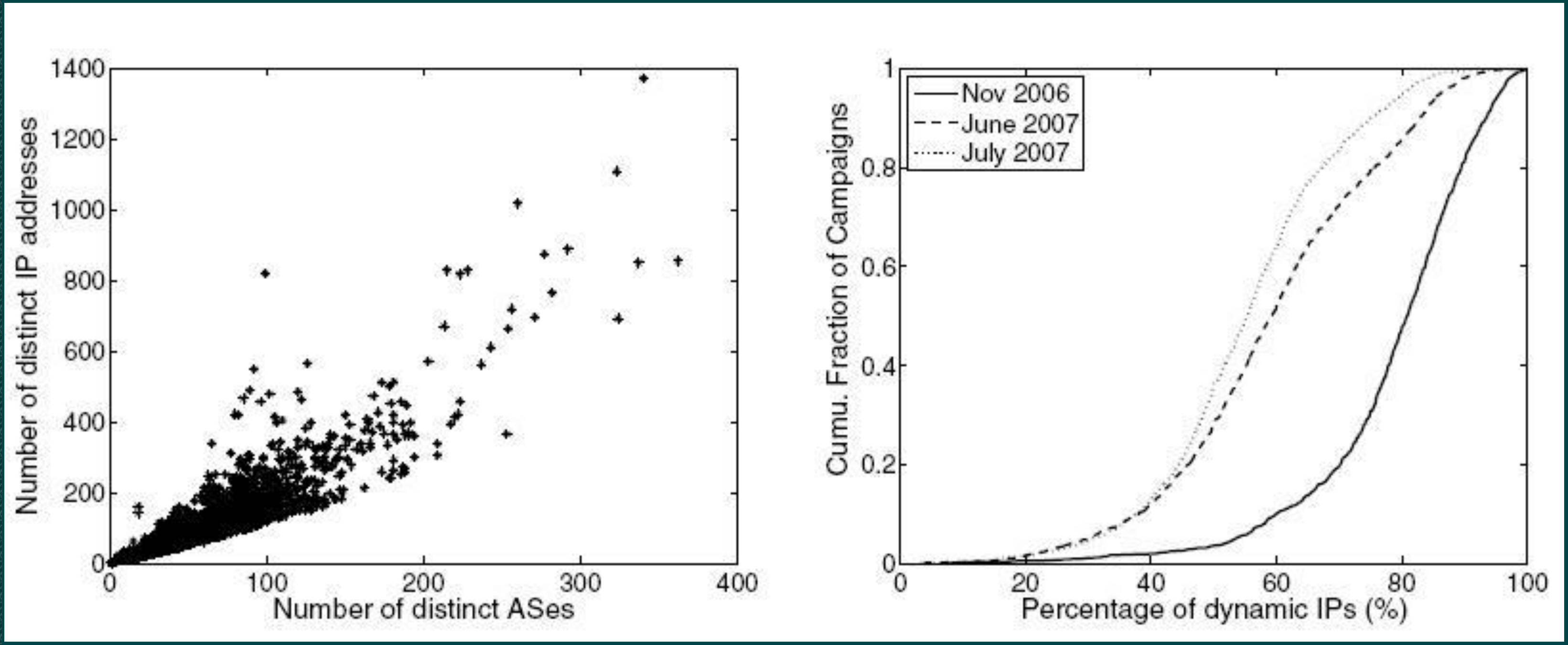
Evaluation, Cont’d

- Pages from different campaigns are different



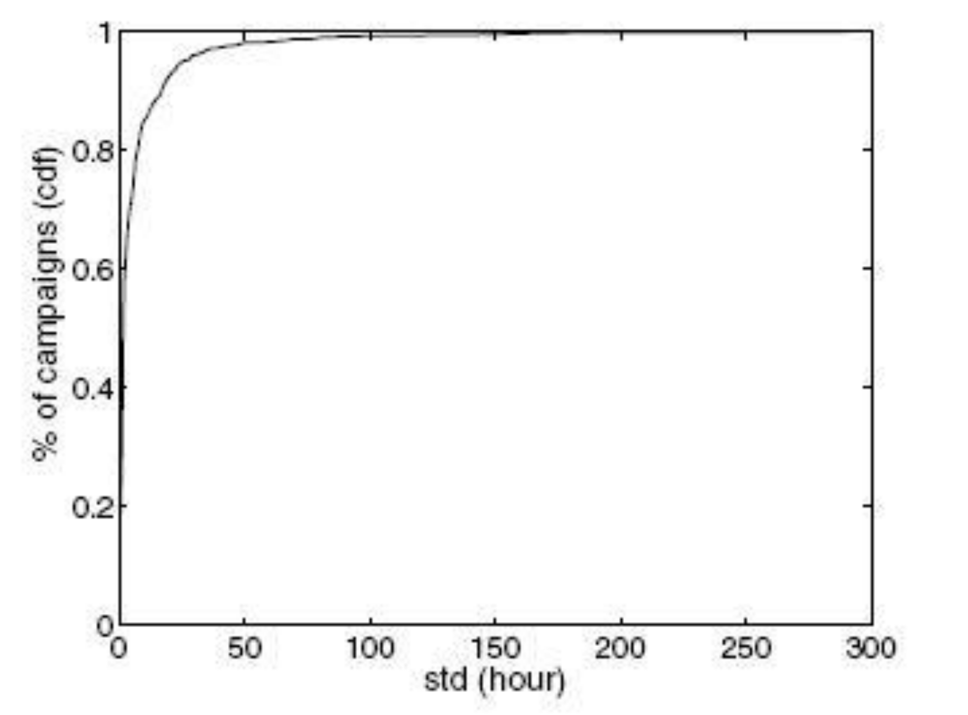
Spamming Botnet Characteristics

- Botnet IP Addresses are distributed and dynamic



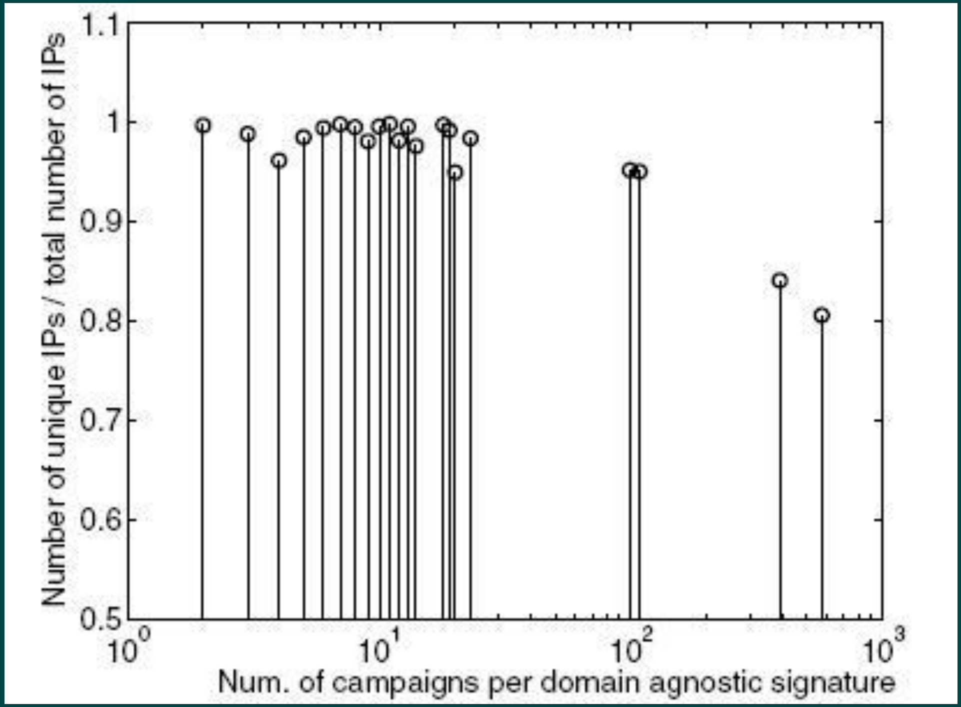
Spamming Botnet Characteristics, Cont'd

- For each campaign, the emails are sent almost simultaneously



Spamming Botnet Characteristics, Cont'd

- It is uncommon for different spam campaigns to overlap



My comments

- If the URLs are presented in image, this tool will be likely to miss them.
- This tool focuses on “bursty” and “distributed” characteristics of spamming botnets. However, if a botnet is not sending spam in a “bursty” or “distributed” way, e.g. when the botnet is small or it keeps sending spam in a long period of time, it is likely to evade the detection.

My Comments, Cont’d

- The authors assume at first the “bursty” and “distributed” nature of spamming botnets. Based on the assumption, they design a tool to detect botnets that behave in a “bursty” and “distributed” way. At last they use the detection result to prove that spamming botnets are “bursty” and “distributed”.
- The assumption can not be confirmed in this way.