SPAMMING BOTNETS: SIGNATURES AND CHARACTERISTICS

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

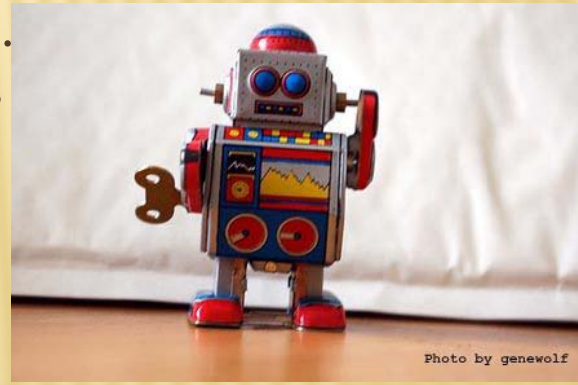
Microsoft Research, Silicon Valley

+Microsoft Corporation

Presenter: Bo Feng

INTRODUCTION

- It is generally acknowledged that botnets have become a significant part of the Internet, albeit increasingly hidden. The main drivers for botnets are recognition and financial gain. Recently, botnets have been widely used for sending spam emails at a scalable level and its popularity is still ascending.
- The call for detecting and blacklisting botnets is becoming increasingly urgent. And various of actions are implemented to deal with this headache. However, previous works have seen their limitations due to the evolve of botnets as well as its behavior. Consequently, we are waiting for the method that can not only detect and thwart botnets right now but also be applicable in the future.

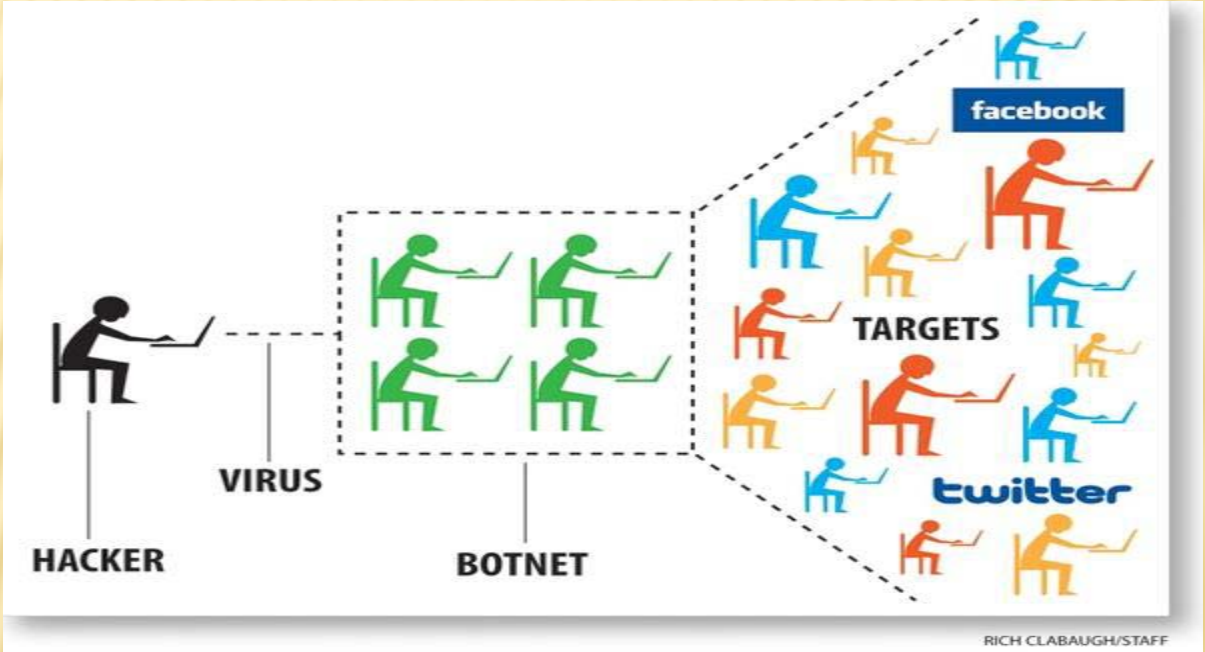


BOTNETS IN SPAMMING

- Due to the phenomenon that *large email servers (HOTMAIL)* have been the popular targets of botnet spam attacks. We try to focus on that part and develop an novel idea in detecting *botnets spamming and their membership*. Right now, we developed a system AutoRE—Automatic URL Regular Expression.

WHAT IS AUTORE?

- AutoRE is a novel framework that identifys botnets hosts via generating botnet spam signatures from emails. This kind of framework is based upon the presupposition that botnets spamming is in a aggregated way.



- AutoRE mainly focuses on the URLs, which form the most critical part in spam emails and embedded in the email contents.

FACING CHALLENGES? WE CAN SOLVE THEM

- Here comes two challenges AutoRE has to face.

1. Spam emails often contains multiple URLs, some of which are legitimate and general. How can you distinguish between "Mr. Right "and "Mr. Wrong"?

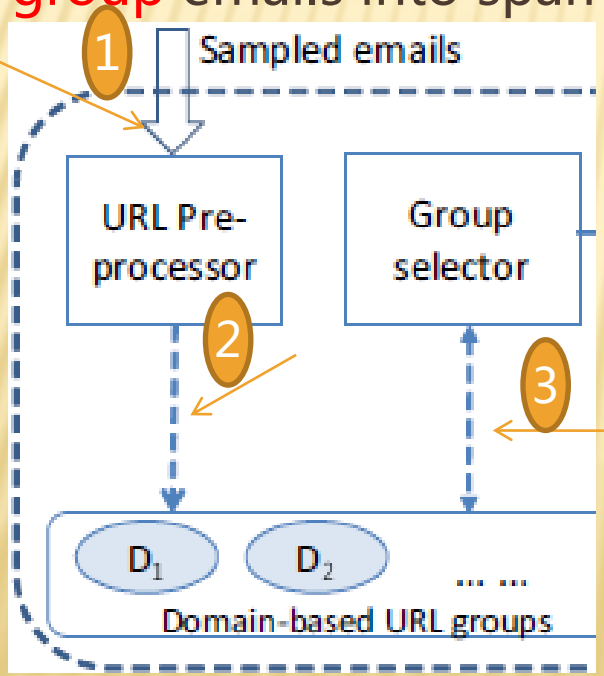
Email 1	Email 2	Email 3
http://www.shopping.com	http://www.peacenvironment.net	http://endosmosis.com/
http://www.w3.org/wai	http://www.w3.org/wai	http://www.talkway.com
http://www.psc.edu/networking/projects/tcp/	http://www.bizrate.com	http://www.bizrate.com
... ...	... ...	... ...
http://www.dvdfever.co.uk/co1118.shtml	***http://www.dvdfever.co.uk/co1118.shtml***	***http://www.dvdfever.co.uk/co1118.shtml***
... ...	... ...	... ...

2. Spammers deliberately add randomness to URLs and use URLs obfuscation techniques to *evade* detection. How can you solve this trick?

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

METHODS TO DEAL WITH THAT CHALLENGES

1. AutoRE iteratively select spam URLs that based upon the *distributed but bursty* property of botnets-based spam campaigns.
2. Further outputs domain-specific signatures.
3. Furthermore, AutoRE uses the generated spam URL signatures to group emails into spam campaigns.

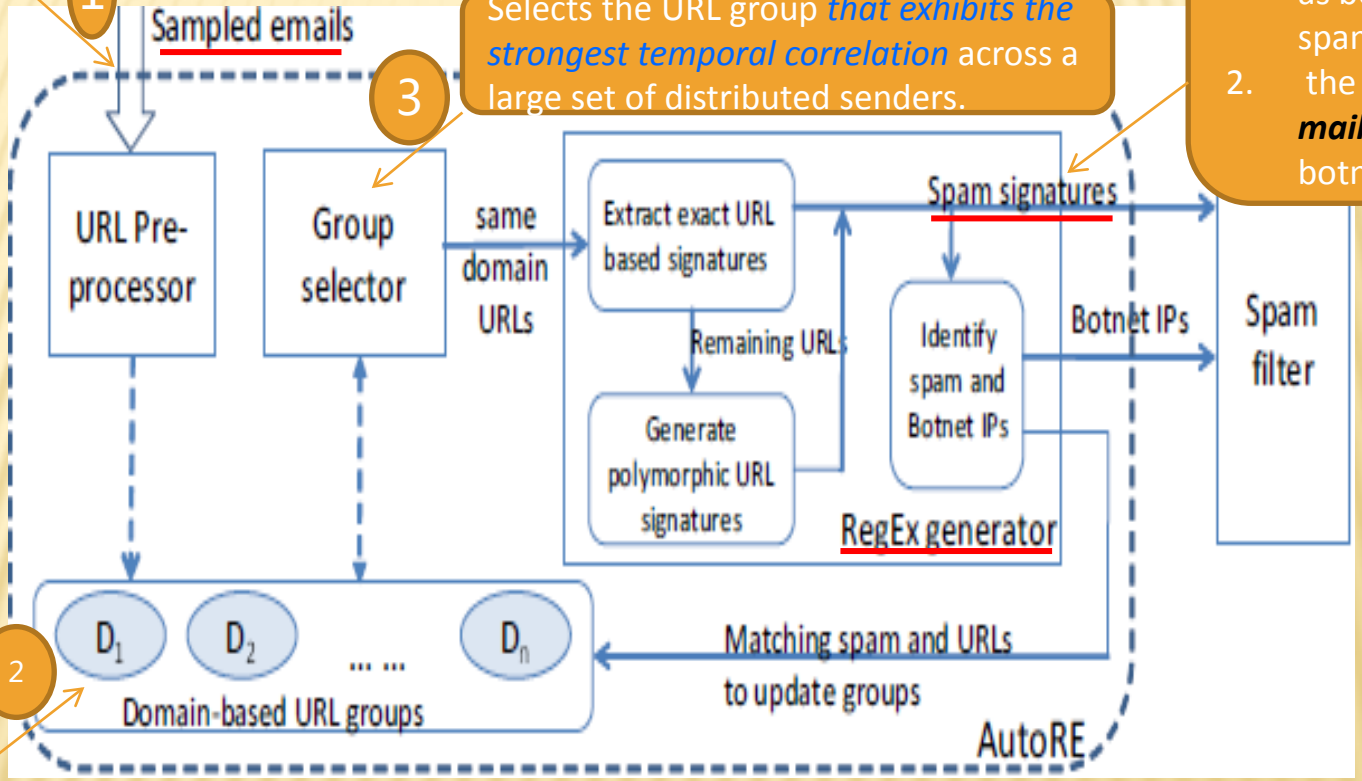


Bursty: emails originating from botnet hosts are sent in *a highly synchronized* fashion.

THE KERNEL PARTS IN AUTORE SYSTEM

Extract *URL string*, *source server IP address*, *email sending time*. After extracting them, AutoRE assigns a *unique email ID to represent the email*

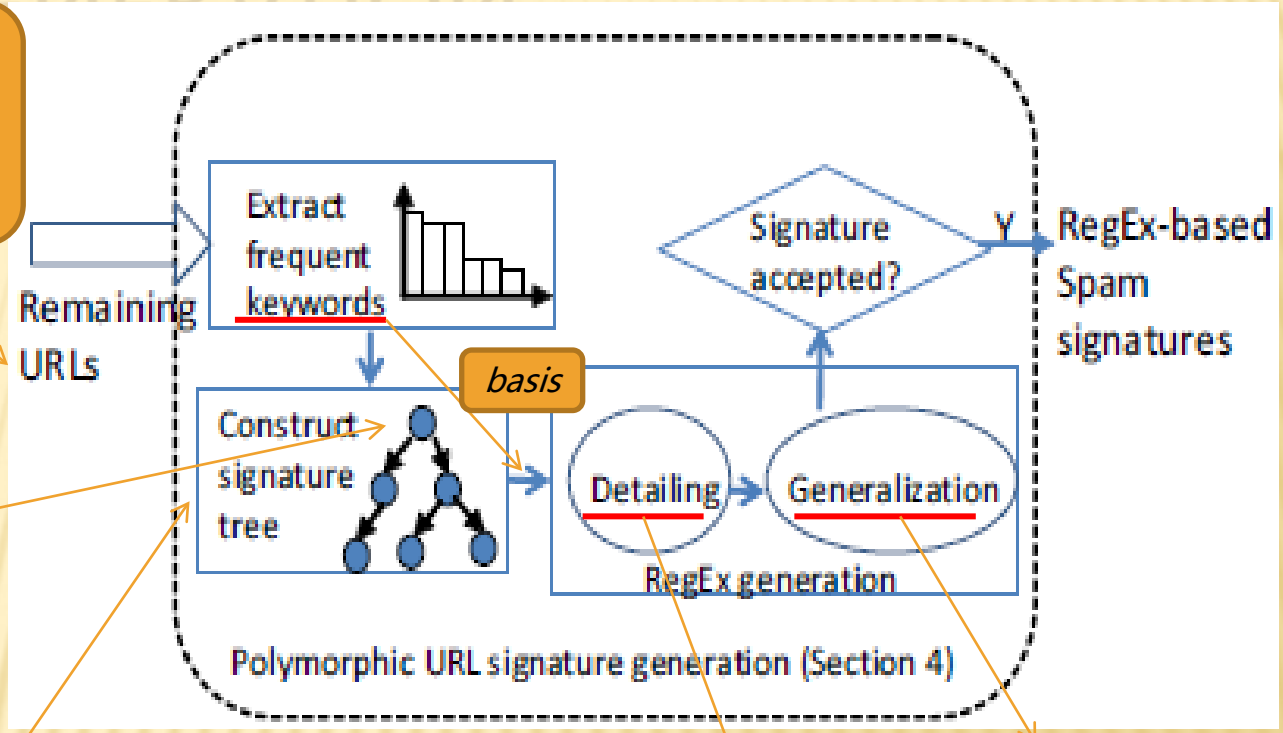
It characterizes:
1. the set of ***matching emails*** as botnet-based spam
2. the ***originating mail servers*** as botnet hosts.



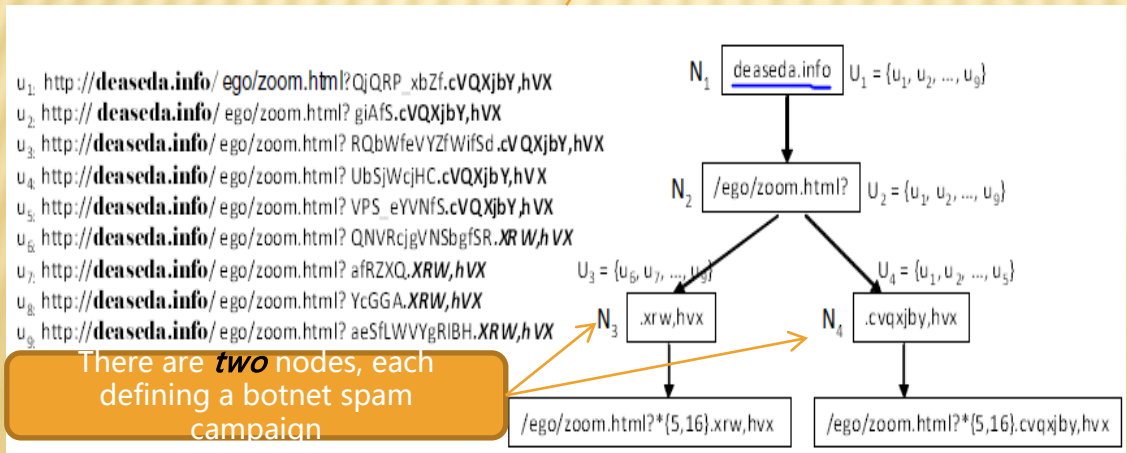
URL preprocessor then *partitions URLs into groups* based on their Web domains

AUTORE GENERATION

A set of *polymorphic* URLs from the same Web domain.



Start with the most frequent substring that is both bursty and distributed as the root, and then incrementally expand the signature by including more substrings and the following nodes.



There are ***two*** nodes, each defining a botnet spam campaign

Returns a *domain-specific* regular expression. It is important to increase the quality of URL signatures

Returns a general *domain-agnostic* regular expression by ***merging*** domain-specific regular expression

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}

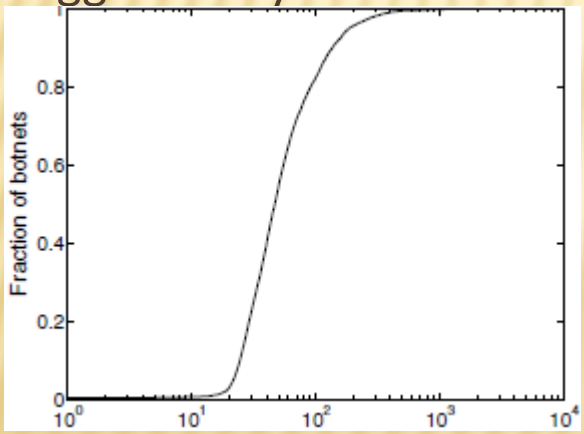
DATASETS AND RESULTS

Do these statistics mean that CU is better than RE?

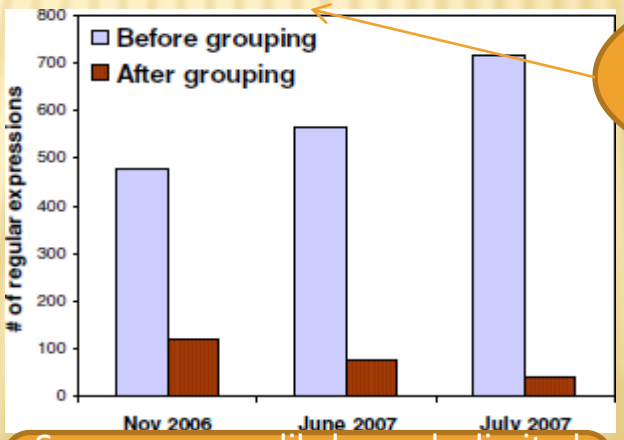
Month	Nov 2006		June 2007		July 2007		Total
	CU	RE	CU	RE	CU	RE	
Num. of spam campaigns	1,229	519	1835	591	2826	721	7,721
Num. of ASes	3,176	1,398	4,495	1,906	4,141	1,841	5,916
Num. of botnet IPs	88,243	23,316	113,794	19,798	85,036	29,463	340,050
Num. of spam emails	118,613	26,897	208,048	26,637	159,494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

CU: set of **complete URL based signatures** RE: set of **regular expression signatures**.

Through the table, we can see that the spam volume increased significantly by around 50% from NOV 2006 to JULY 2007. However, the increment of botnet IPs is not that rapidly. This might suggest that each botnet host is used more aggressively.



Number of distinct IP address. Based on *sample emails*

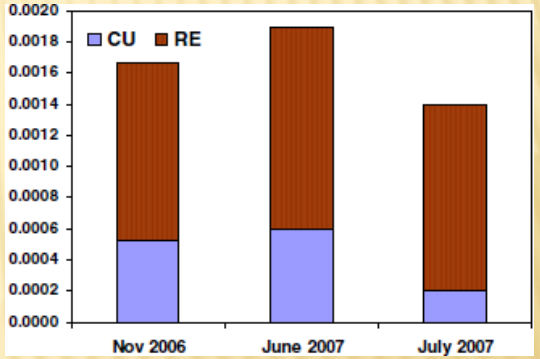


From domain-specific regular expressions into domain-agnostic regular expressions.

Spammers very likely used a limited number of automatic spam generation programs for generating polymorphic URLs.

EVALUATION

The possibility of CU is less than RE



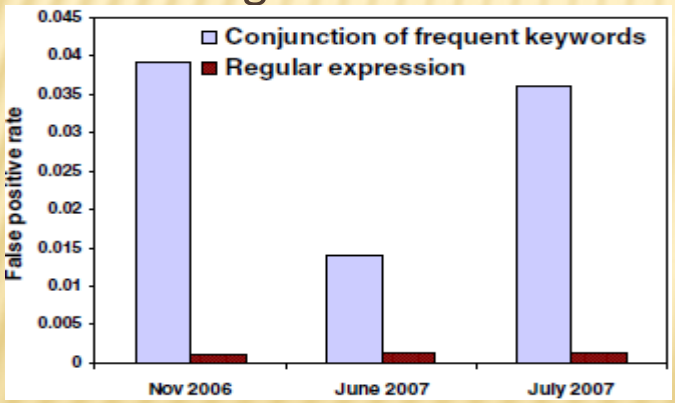
1. *False Positive Rate (FPR)*:

   The FPR is very slow based upon number in the table.

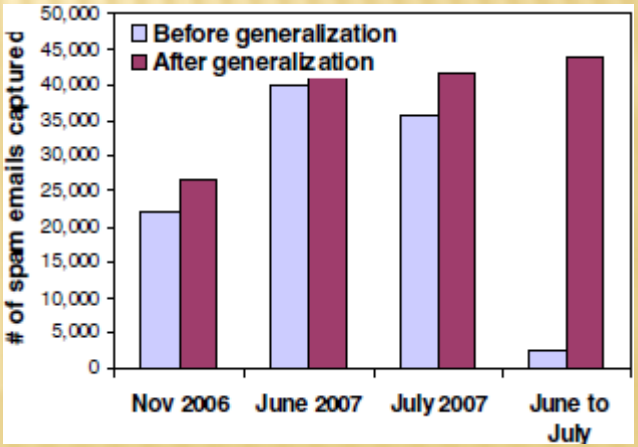
2. *Ability to Detect Future Spam:*

   Through the table, we can see that RE signatures are much more robust over time than CU signatures. Even though the FPR of CU is less than the FPR of RE.

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



By comparison, we can see that RE has a low FPR and can greatly reduce the chance of legitimate URLs matching a signature



By merging domain-specific signatures into *domain-agnostic* signatures, we are more effective in detecting future botnets spam

GET TO KNOW SPAMMING BOTNET TRAITS

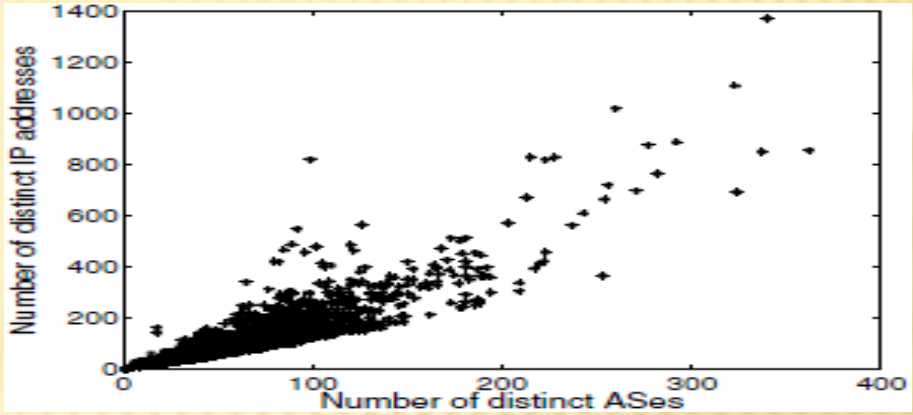
We know their traits by 2 steps:

1. Analyze their *geographic distribution* and *overall sending patterns*;——general

2. Study their *individual behavior* to examine each spam campaign;——specific

DISTRIBUTION OF BOTNETS--GENERAL

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



According to these 2 graphs, we can tell that Botnets are becoming *a global phenomenon*. It emphasizes the significance of employing a network-wide view for botnet detection and defense.

DISTRIBUTION OF BOTNETS--GENERAL

- We utilized 3 different methods to describe the sending patterns of botnets:

1. ***Number of recipients*** *per email;*
2. ***Connections*** *per second;*
3. ***Non-existing recipient*** *frequency*.

Results:

Both the sending patterns of the identified botnet hosts and other hosts are well spread in the space. This accentuates the premise that individual botnet host does not exhibit distinct sending patterns.

EACH CAMPAIGN-INDIVIDUAL

Similarity of Email Properties:

contents botnets sent are highly resembled;

Similarity of Sending Time:

botnets sent almost simultaneously;

Similarity of Email Sending Behavior

DESIRABLE FEATURES OF AUTORE

- *Low false positive rate*
- *Ability to detect stealthy botnet-based spam*
- *Ability to detect frequent domain modifications* by using *domain-agnostic signatures*.

MAIN CONTRIBUTIONS OF AUTORE

- Botnets are gaining its popularity for spam delivery and botnet host is involved in *multiple attacks*.
- It is better to detect botnets hosts by inspecting them in a *aggregated* way.
- Botnets attacks have *diverse phases*. Hence, by exploring network scanning patterns would shed light on botnets identification.