Exploring Spamming Botnet Signatures and Characteristics

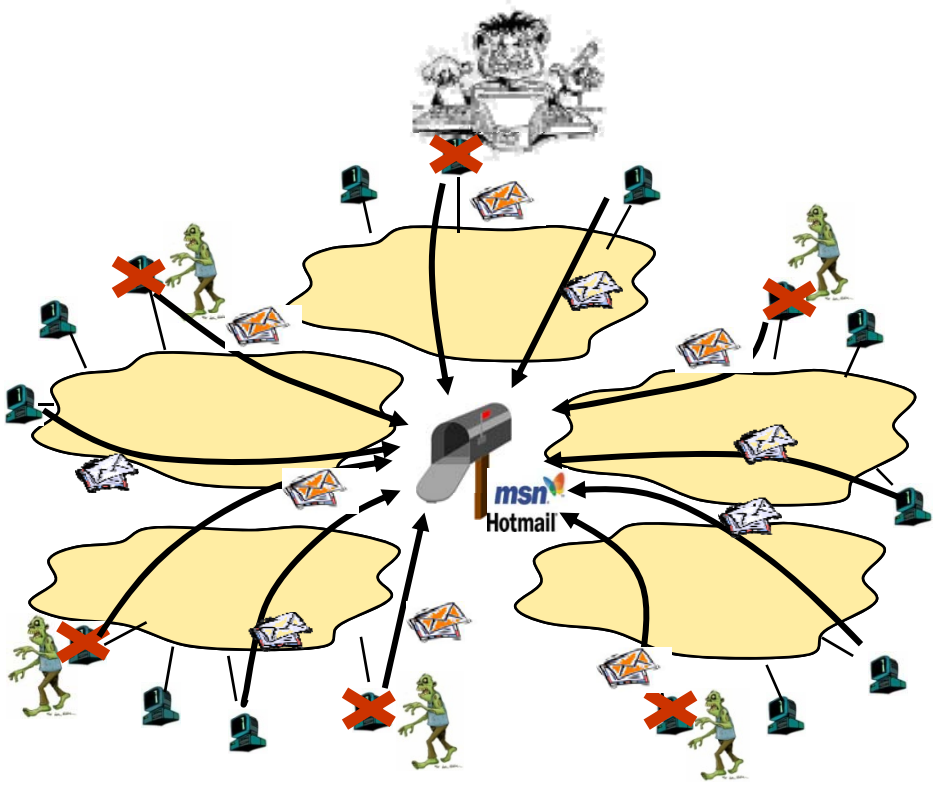
Yinglian Xie

Fang Yu, Kannan Achan, Rina Panigrahy,
Geoff Hulten, Ivan Osipkov
Microsoft Research Silicon Valley
Windows Live (Hotmail) mail, Windows Live Safety Platform

Microsoft® Research

Motivations

- Botnet attacks are prevalent
  - DDoS, spam, click fraud
  - Large-scale, hidden trail
- Can we detect them by signatures?
  - Filter future spam emails
  - Detect botnet membership
  - Understand botnet activities and trends



❖ **Focus on URLs instead of contents**
74% of spam contain at least one URL

Two Observations

- Common or similar URLs from a botnet spam campaign
  - Direct users to one or a few Websites
- Botnet attacks are distributed yet bursty
  - **Distributed:** Involve a large number of hosts
  - **Bursty:** spam sent in a short duration

➧ Identify groups of distributed hosts that sent similar URLs in a burst

Challenge I:

- Mixing spam and legitimate URLs in an email
  - Which URLs characterize a Botnet?
  - White list approach does not work

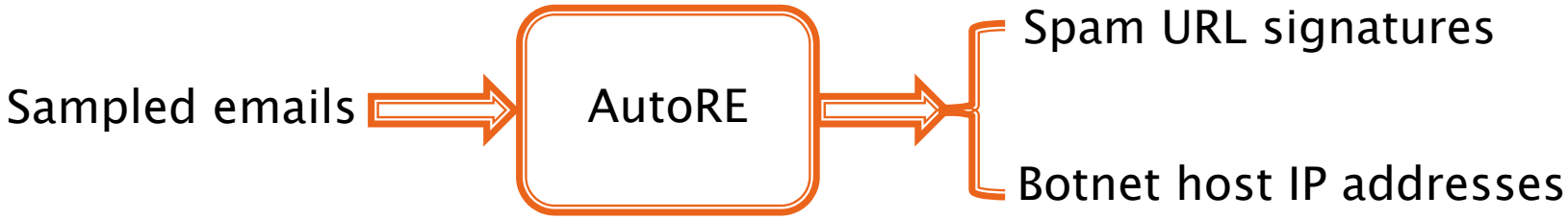
Challenge II:

- Crafting polymorphic URLs across emails
  - Insert random tails
  - Customize based on recipient
  - What are the common URL patterns?

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?jxvsgfhjb.cvqxjby,hvx http://shgeep.info/tota/indexx.html?ijkkjija.cvqxjby,hvx http://shgeep.info/tota/indexx.html?ibdivvx_ceh.cvqxjby,hvx

AutoRE Overview

Automatic Botnet spam signature generation



How we address challeng

- Mixing legitimate and spa

Require no labeled data or white list

→ "distributed" and "bursty" spam activity

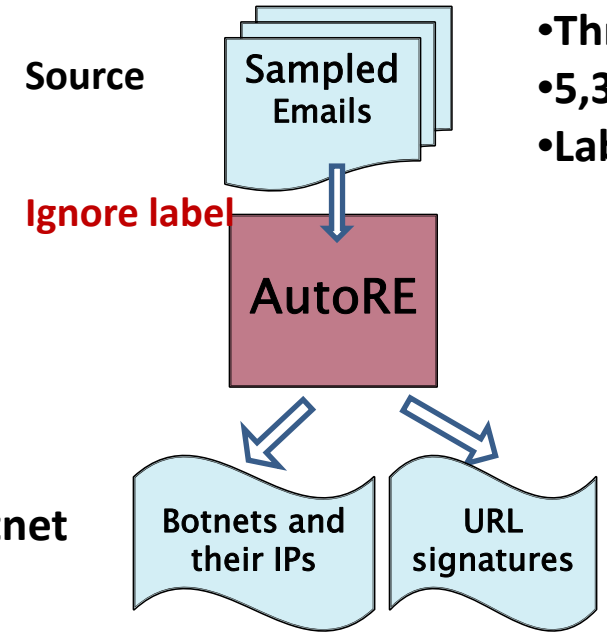
- Porlymorphic URLs

→ Two

FP rate 10x – 30x lower than keywords
FN rate10x lower than complete URLs

• **regular expressions**

Experiment Results



- Three months of sampled emails
- 5,382,460 email messages
- Labeled with spam or non spam

- Identified 7,721 botnet spam campaigns
- Span 5,916 ASes
- 340,050 botnet IP addresses
- 0.5% false positive rate

- 16% -18% of spam not captured by well known blacklists
- 0.2% false positive rate

	Nov-06	Jun-07	Jul-07
Num. of Botnets	1,748	2,426	3,547
Num. of Spam	145,510	234,685	200,271
Num. of IPs	100,293	131,234	113,294

Example Spam URL Signatures

http://deaseda.info/ego/zoom.html?QjQRPxbZf**.cVQXjbY,hVX**
http://deaseda.info/ego/zoom.html?giAfS**.cVQXjbY,hVX**
http://deaseda.info/ego/zoom.html?QNVRcjgVNSbgfSR**.XRW,hVX**
http://deaseda.info/ego/zoom.html?afRZQ**.XRW,hVX**
http://deaseda.info/ego/zoom.html?YcGGA**.XRW,hVX**

http://deaseda.info/ego/zoom.html?*{4,18}.[a-zA-Z]{3,7},hvx

http://arfasel.info/hums/jasmine.html? UbSjWcjHC**.cVQXjbY,hVX**
http://apowefe.info/hums/jasmine.html? PSeYVNfS**.cVQXjbY,hVX**
http://carvalert.info/hums/jasmine.html? SfLWVYgRIBH**.XRW,hVX**

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

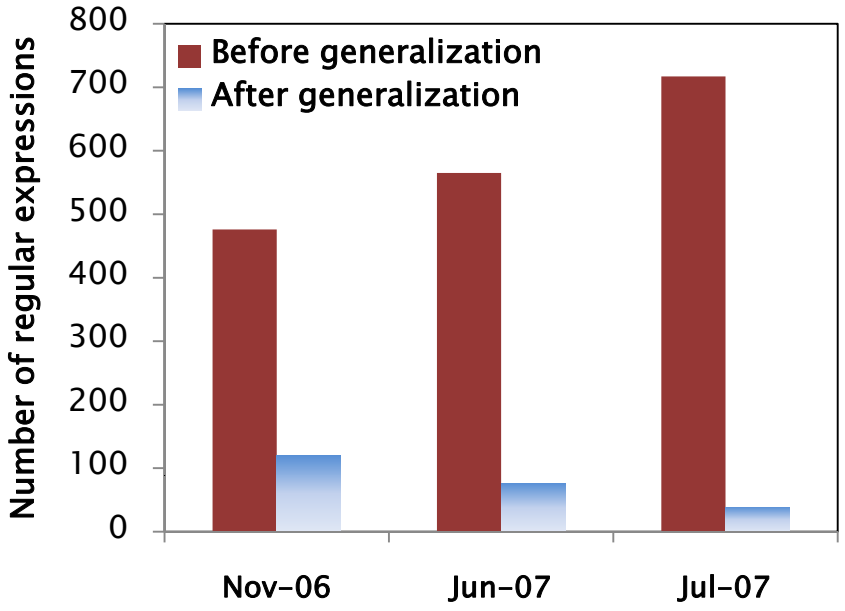
http://www.mezir.com/n/?167& acetyleneassign
http://www.aferol.com/n/?167& commonwealcirce
http://www.bedremf.com/n/?167& deprivationalign
http://www.mokver.www/n/?167& bleachbeneficial

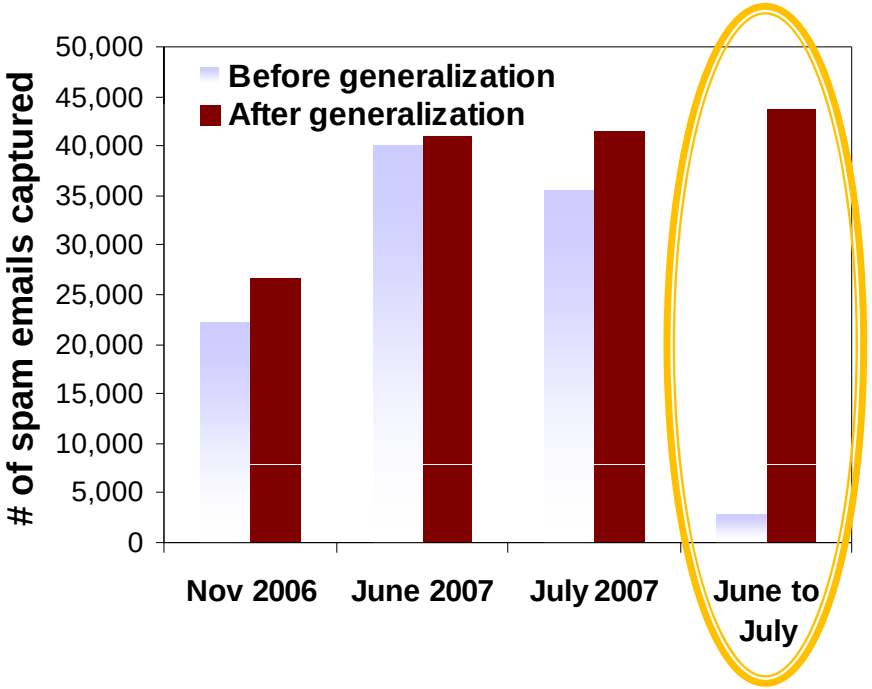
http://[^\/]*/n/?167&[a-zA-Z]{9,27}

Signature Domain Generalization

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}





Source of Botnets

- Botnet IP addresses are less than 0.5% of all IPs
- Their spam emails:
  - 4%–6% of all detected spam

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

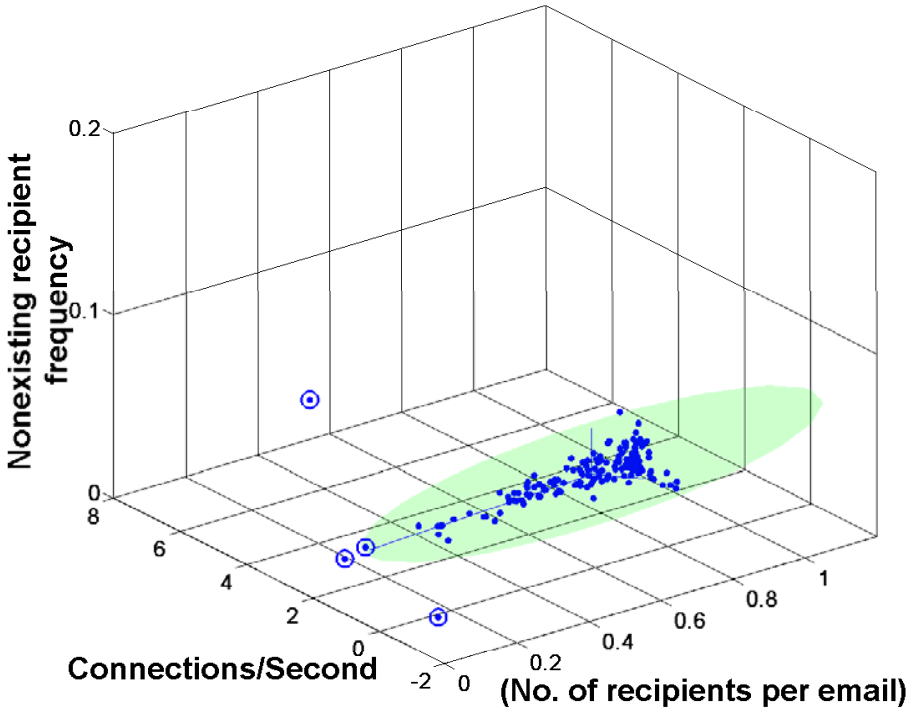
Email Sending Characteristics

- Email content similarity
  - Destination Web pages
    - Over 90% at least 75% similar
  - Features: # of URLs, # of domains, email length
    - Over 95% with negligible std
  - Email text
    - Over 60% less than 50% similar
- Sending behavior similarity
  - Over 90% with sending time std <20 hours
  - Over 50% with sending time std < 1.8 hour
  - Clustering botnet sending patterns

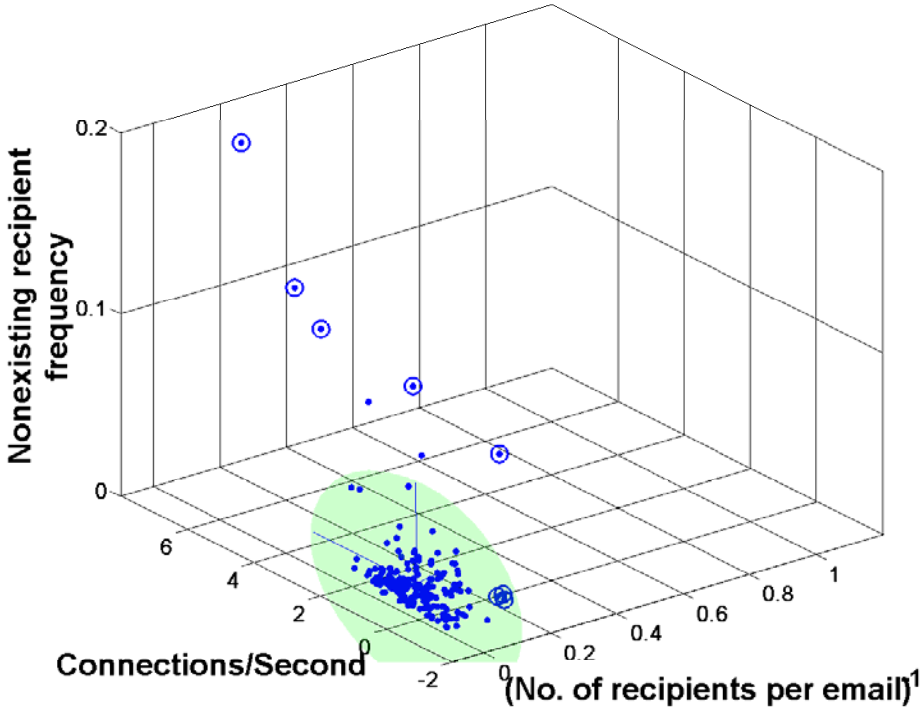
Example Per-Botnet Clusters

- 85% of spam campaigns are well clustered

162 hosts with 4 outliers

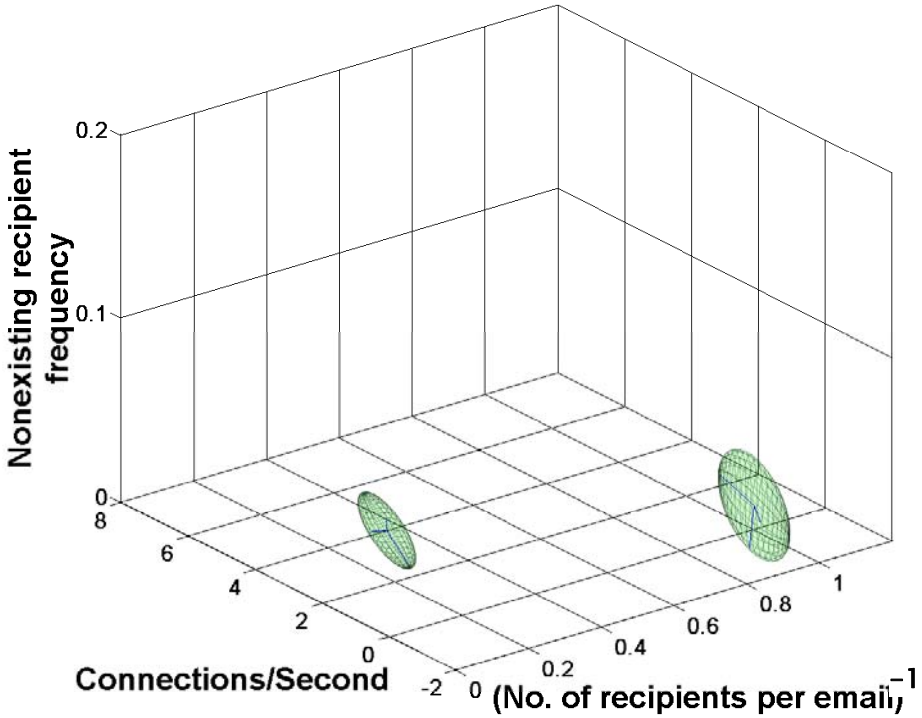


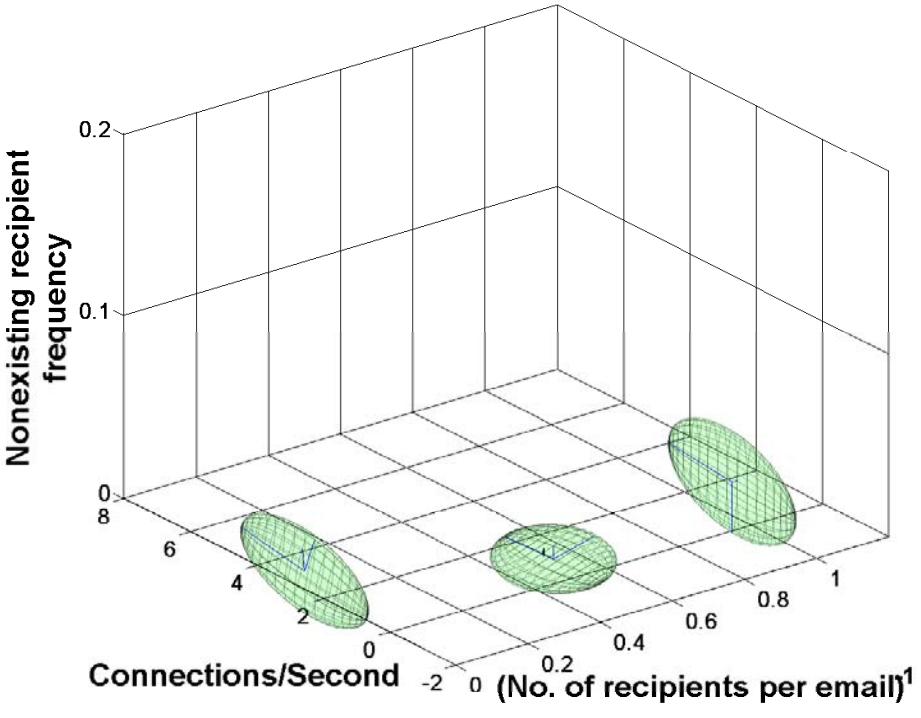
191 hosts with 9 outliers



All-Botnet Clusters

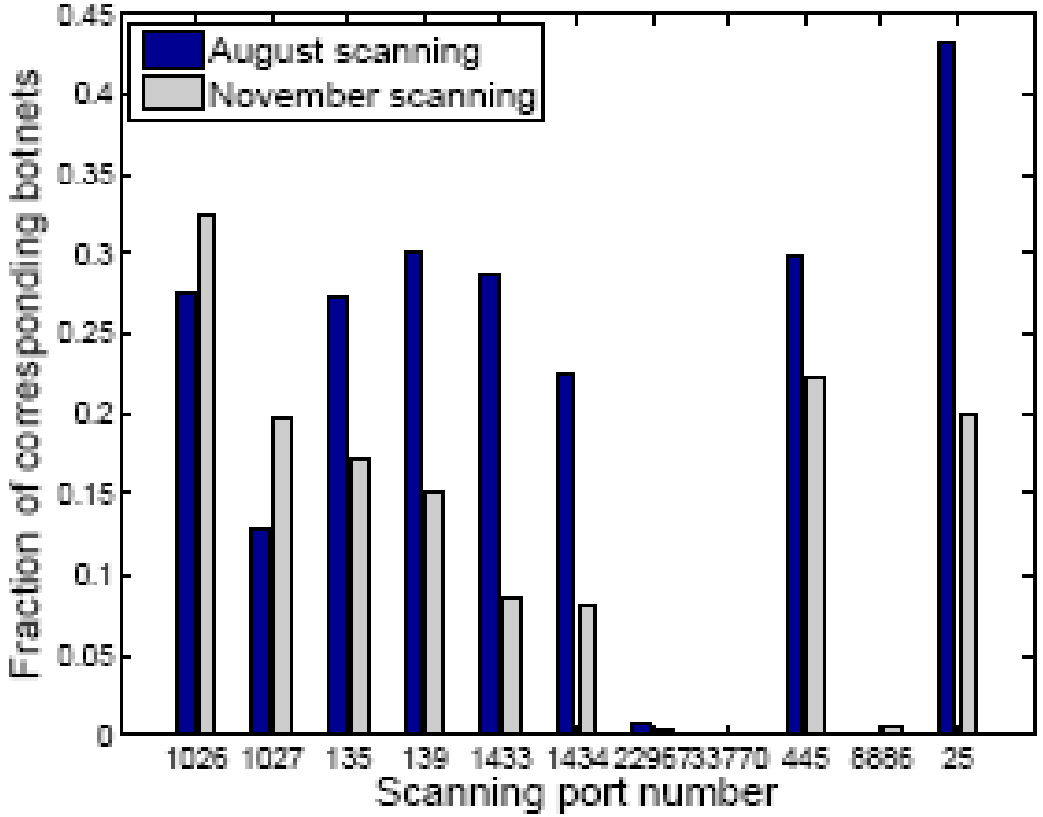
- Botnet sending activities are clusterable





Scan History

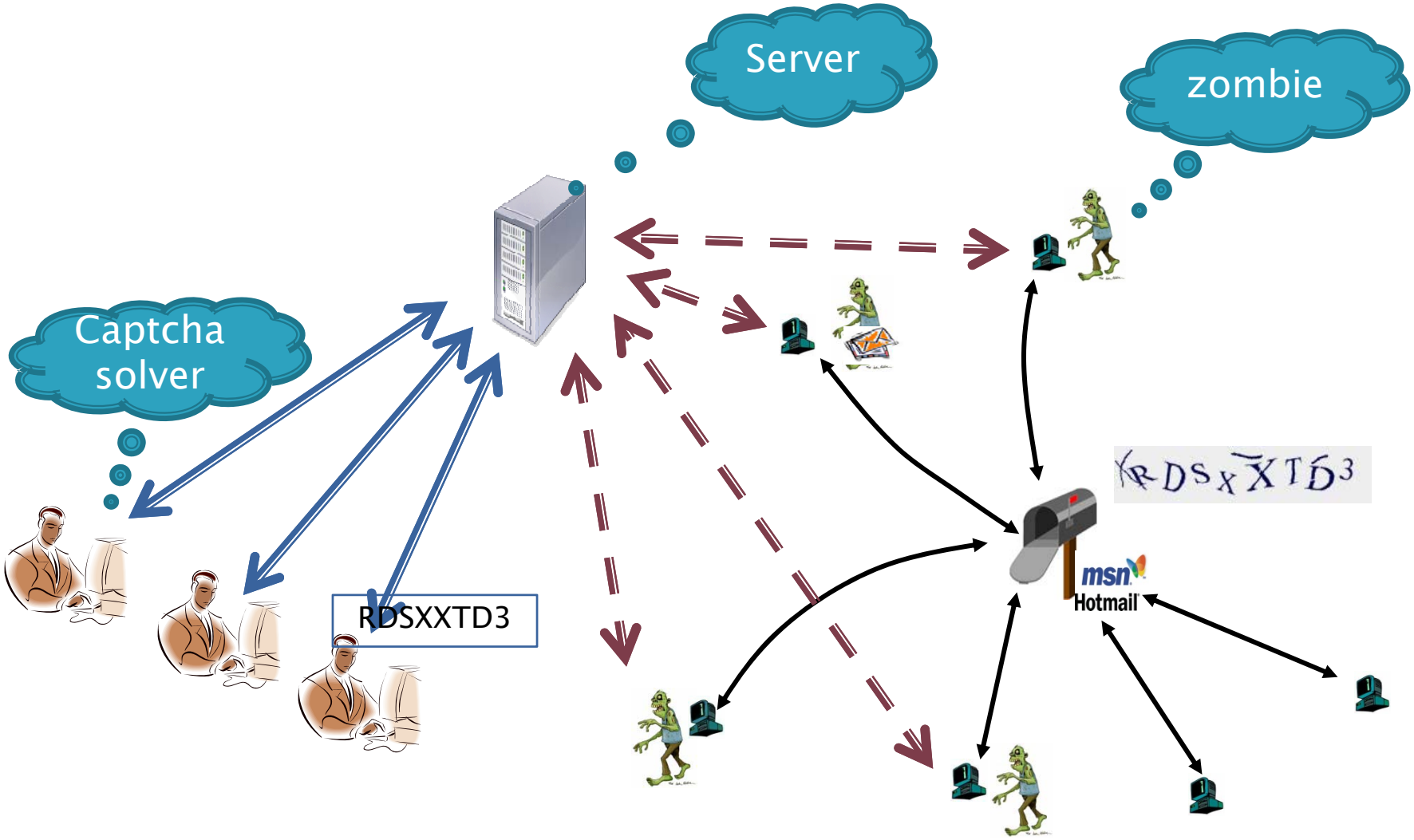
- Higher correlation in Aug than in Nov



Summary of Botnet Activities

- Botnets are more popular for sending spam
  - Number of spam campaigns doubled
  - Number of botnet IPs increased by 10%
- Spam campaigns are more stealthy
  - Adoption of polymorphic URLs increased by 50%
- Email sending patterns are clusterable when viewed in aggregate
- Botnet activities show correlation with the network telescope data

New Types of Botnet Spamming Attacks



Our Experiences

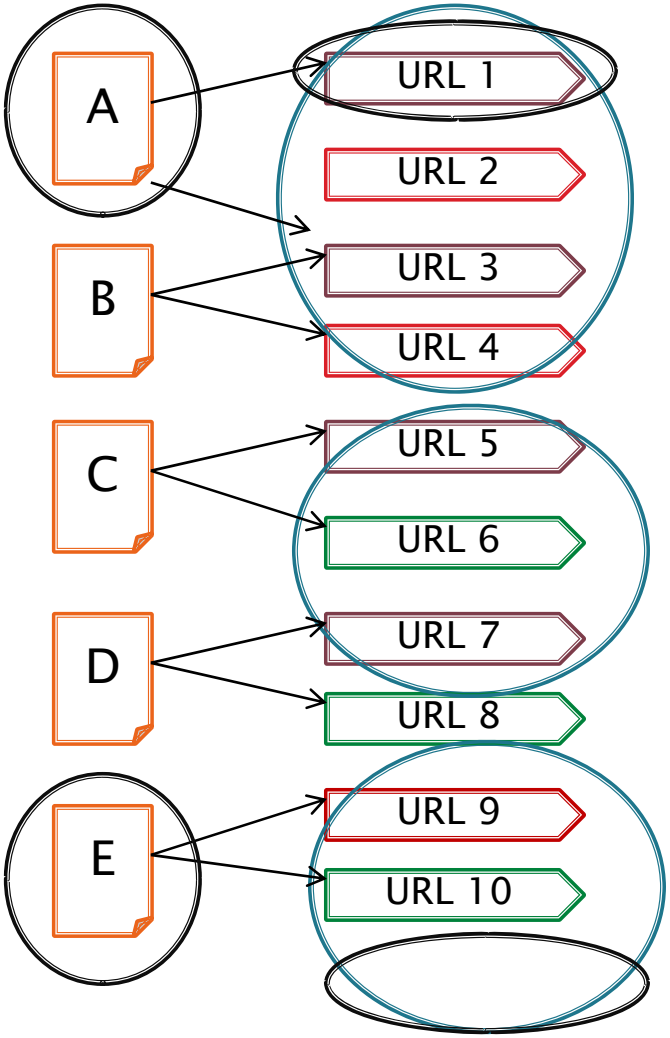
- Large volume of input data
  - Efficient algorithm
  - Distributed computing infrastructure
- Attackers are stealthy and evolving
  - Individual detection difficult
  - An aggregated view for detection
  - Robust features or attack invariant
- Rigorous evaluation challenging
  - Usually no ground truth data
  - Collaborative monitoring

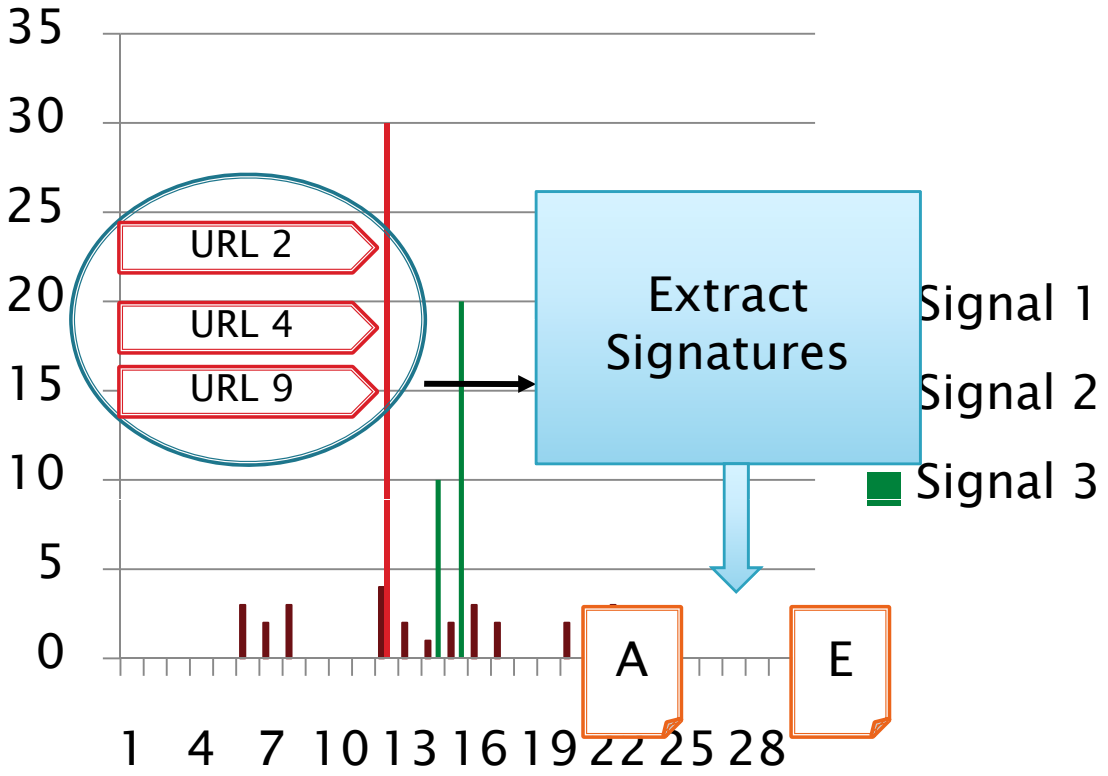
Open Questions

- Economic underpinnings of botnet attacks
- Information sharing for collaborative research
- Attack forensics and deterrence

Thanks!

Candidate URL Selection





Regular Expression Generation

- Why regular expressions?
  - Need specific signatures