Studying Spamming Botnets Using *BotLab*

Arvind Krishnamurthy

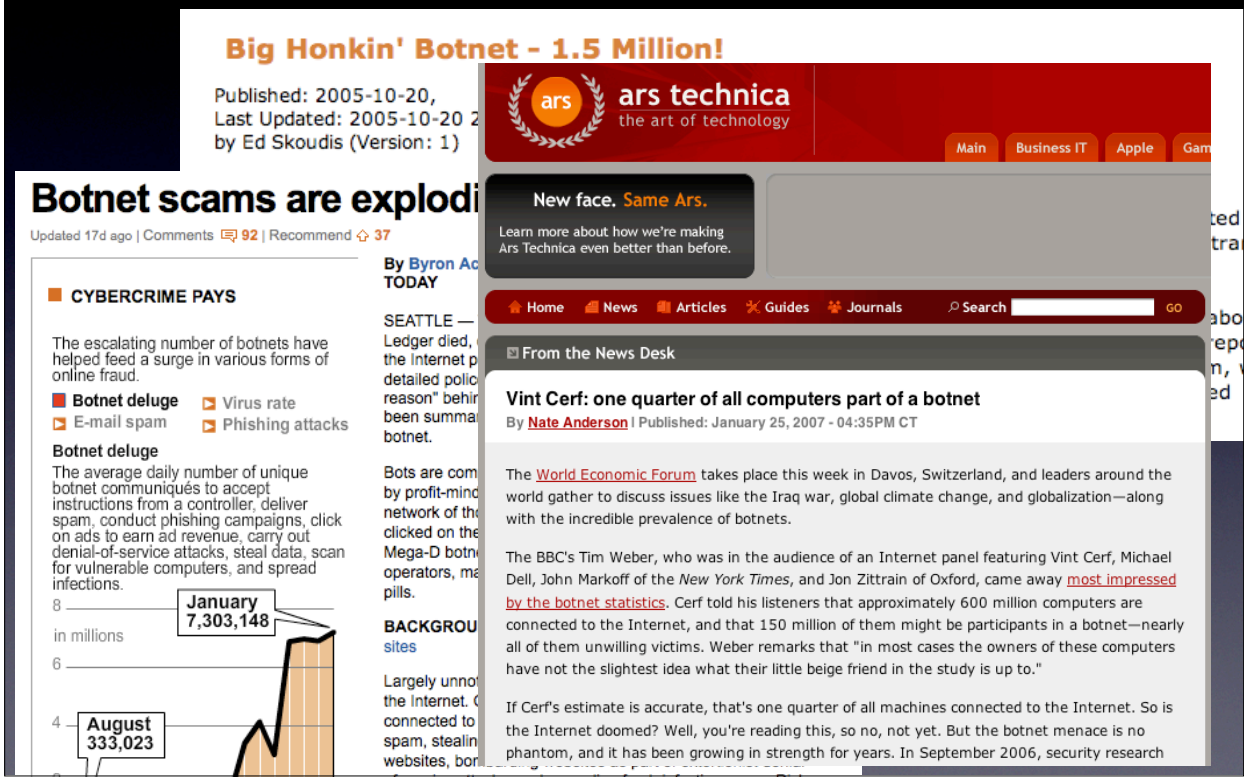
Joint work with:

John John, Alex Moshchuk, Steve Gribble

University of Washington

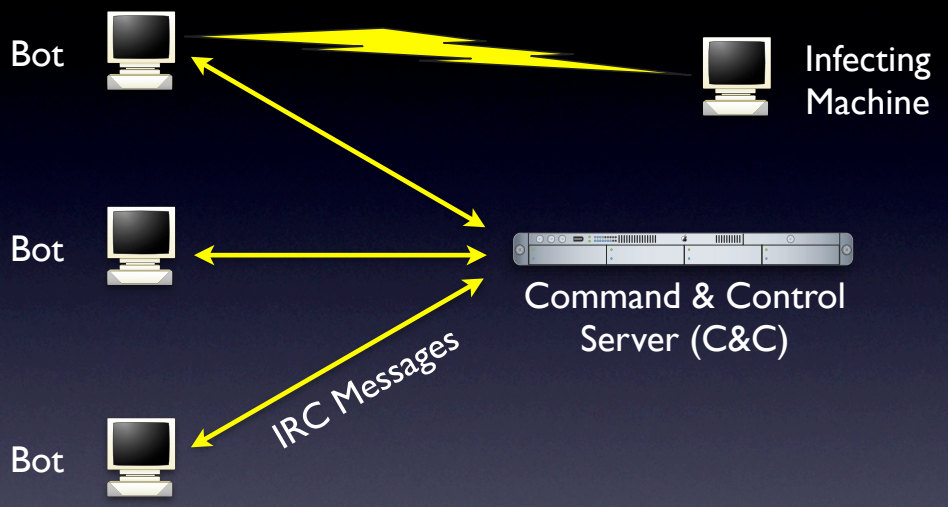
Botnets: a Growing Threat

- Increasing awareness, but there is a dearth of hard facts especially in real-time
  - Meager network-wide cumulative statistics
  - Sparse information regarding individual botnets
  - Most analysis is post-hoc

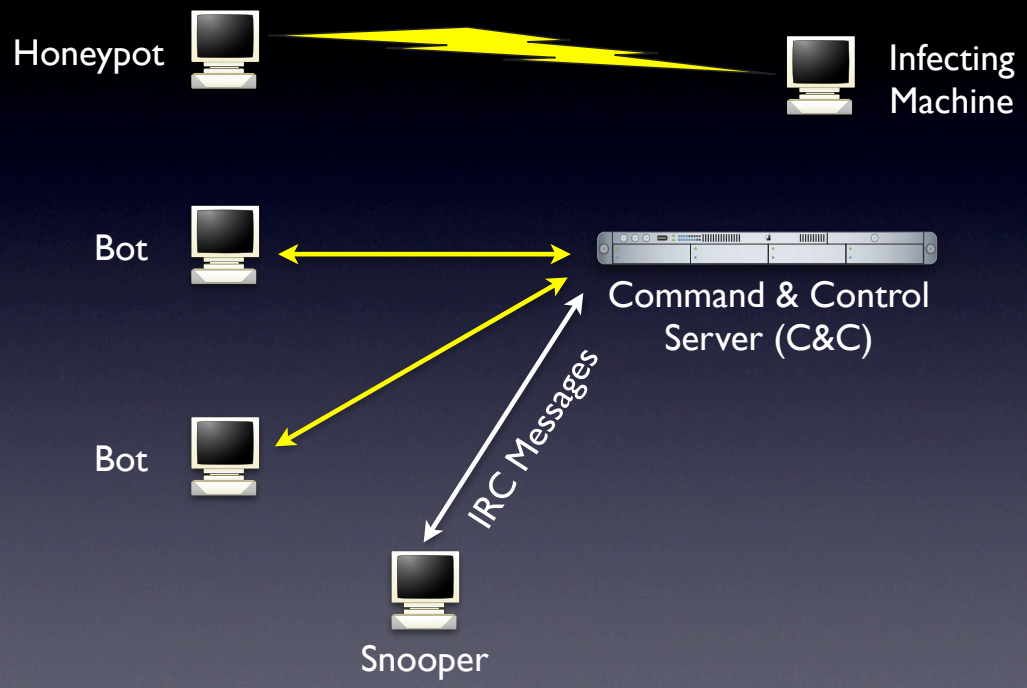


Goal is to build a *botnet monitoring platform* that can track the activities of the *most significant spamming botnets* currently operating in *real-time*

Botnet Lifecycle (Traditional View)



Tools for Monitoring



Botnet Operators' Response

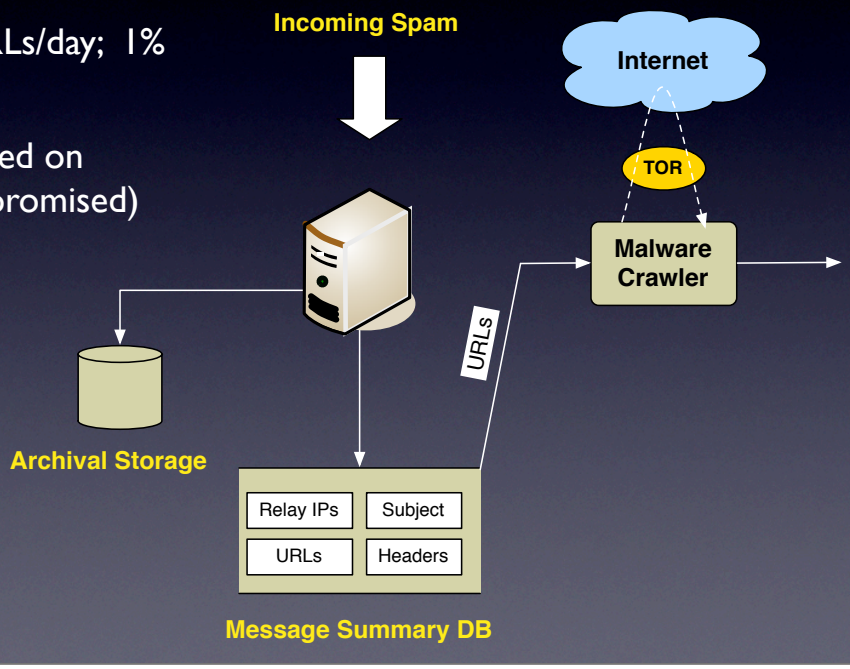
- Use *social engineering* techniques for infection
  - Cleverly crafted emails/websites induce users to download malicious programs
  - Detect virtualization techniques
- Use *customized protocols* over HTTP
- Use *dynamic adaptation*
  - Malware binaries morph every few minutes (use polymorphic packers)
  - FastFlux DNS allows for fast redirection to new C&C servers
  - Change C&C protocols as well

BotLab Design

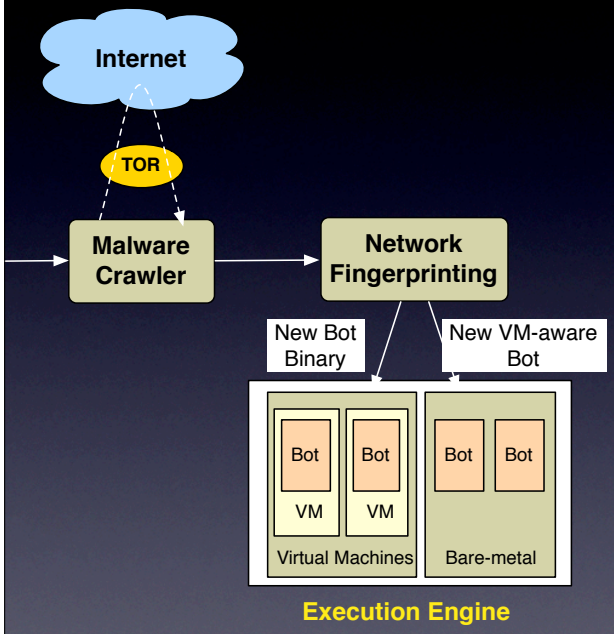
- *Active* as opposed to passive collection of binaries
- *Attribution:* run actual binaries and monitor behavior without causing harm
- *Scalably* identify duplicate binaries
- *Correlate* incoming spam with outgoing spam

Malware Collection

- Augment honeypots with active crawling of spam URLs
- 100K unique URLs/day; 1% malicious
- Most URLs hosted on legitimate (compromised) webservers

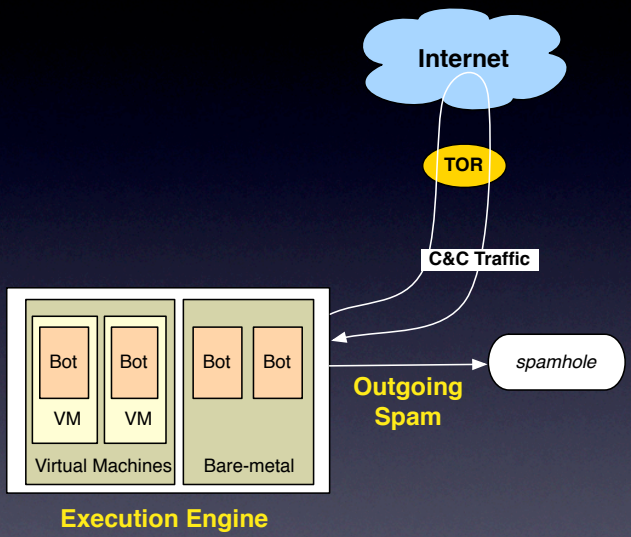


Network Fingerprinting



- Goal: find new bots while discarding old ones
- Execute binaries and generate a fingerprint, which is a sequence of flow records
- Each flow record defined by (DNS, IP, TCP/UDP)
- Execute both inside and outside of VM to check for VM detection
- Execute each binary multiple times as some bots issue random requests (e.g., Google searches)

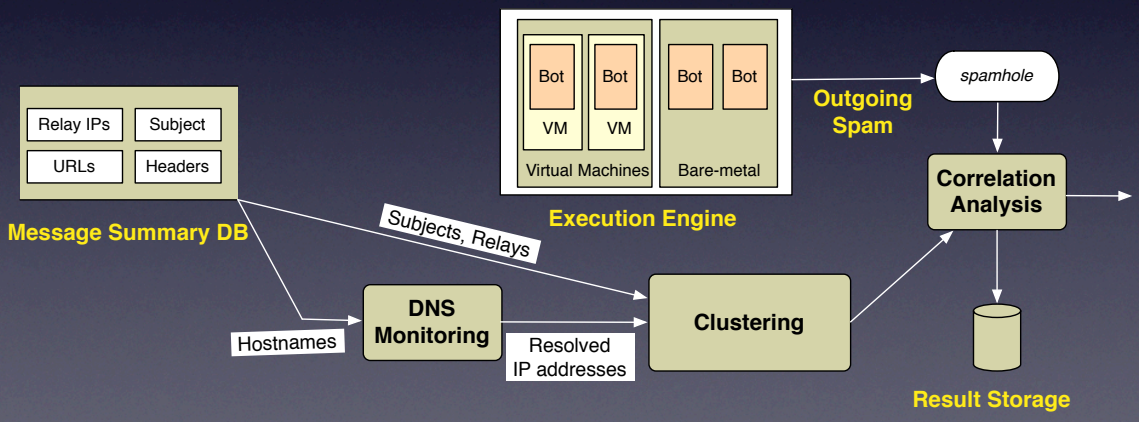
Coaxing Bots to Run



- Bots send "verification" emails before they start sending regular spam
- Some other bots spam using webservices (such as HotMail)
- C&C servers are setup to blacklist suspicious IP ranges
- Bots with 100% email delivery rate are considered suspicious
- Fortunately only O(10) botnets; so manual tweaking possible

Clustering/Correlation Analysis

- Correlate incoming spam with outgoing spam and perform attribution; identify IPs for a given botnet
- For spam that cannot be directly attributed, cluster based on source IPs and merge with an attributed set if there is overlap



Measurements

- Analysis of *outgoing spam* feed
- Analysis of *incoming spam* feed
- *Correlation* of outgoing and incoming spam feeds

Behavioral Characteristics

Botnet	C&C Discovery	C&C servers contacted over lifetime	C&C protocol	spam send rate (msgs/min)
Grum				
Kraken				
Pushdo				
Rustock				
MegaD				
Srizbi				
Storm				

Behavioral Characteristics

Botnet	C&C Discovery	C&C servers contacted over lifetime	C&C protocol	spam send rate (msgs/min)
Grum	static IP	1		
Kraken	algorithmic DNS	41		
Pushdo	set of static IPs	96		
Rustock	static IP	1		
MegaD	static DNS name	21		
Srizbi	set of static IPs	20		
Storm	p2p (Overnet)	N/A		

Behavioral Characteristics

Botnet	C&C Discovery	C&C servers contacted over lifetime	C&C protocol	spam send rate (msgs/min)
Grum	static IP	1	encrypted HTTP	
Kraken	algorithmic DNS	41	encrypted HTTP	
Pushdo	set of static IPs	96	encrypted HTTP	
Rustock	static IP	1	encrypted HTTP	
MegaD	static DNS name	21	encrypted custom protocol (port 80)	
Srizbi	set of static IPs	20	unencrypted HTTP	
Storm	p2p (Overnet)	N/A	encrypted custom	

Behavioral Characteristics

Botnet	C&C Discovery	C&C servers contacted over lifetime	C&C protocol	spam send rate (msgs/min)
Grum	static IP	1	encrypted HTTP	344
Kraken	algorithmic DNS	41	encrypted HTTP	331
Pushdo	set of static IPs	96	encrypted HTTP	289
Rustock	static IP	1	encrypted HTTP	33
MegaD	static DNS name	21	encrypted custom protocol (port 80)	1638
Srizbi	set of static IPs	20	unencrypted HTTP	1848
Storm	p2p (Overnet)	N/A	encrypted custom	20

Outgoing Spam Characteristics

- Subjects are distinguishing markers of botnets
  - 489 subjects per botnet per day with *zero overlap*
  - Across 2 months, only 0.3% overlap
- Bots are *stateless*
  - List of recipients downloaded from C&C server is randomly chosen
  - Bots can be periodically restarted to quickly obtain information on ongoing spam campaigns

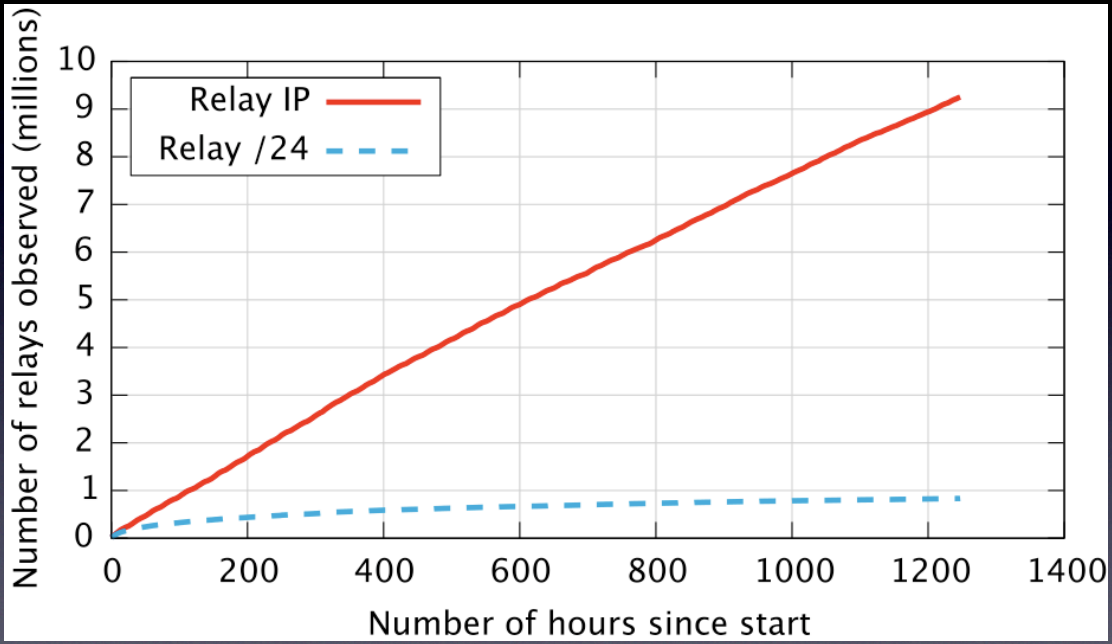
Botnet Mailing Lists

- Random fetch model allows us to estimate botnet mailing list sizes
  - As we see more of the spam feeds, there will be more duplicates in recipient email addresses
- If mailing list size is N and if bot obtains C addresses for each C&C query, then probability that an email address will appear again in the next K emails is

$$1 - (1 - C/N)^{K/C}$$

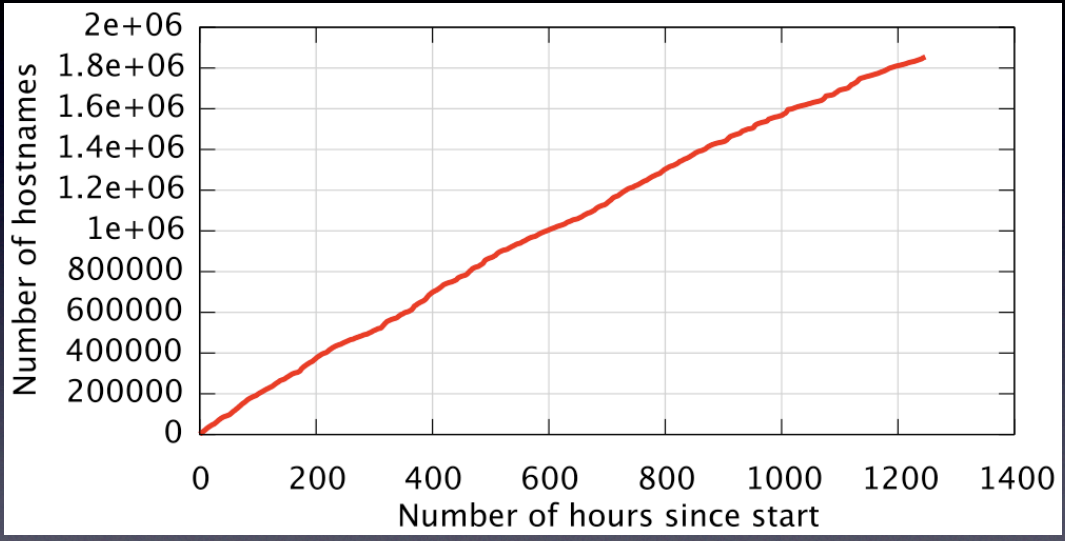
- Some mailing list sizes: MegaD's is 850 million, Rustock's is 1.2 billion, Kraken's is 350 million
- Overlap between mailing lists is small (less than 28%)

Incoming Spam: Source IPs



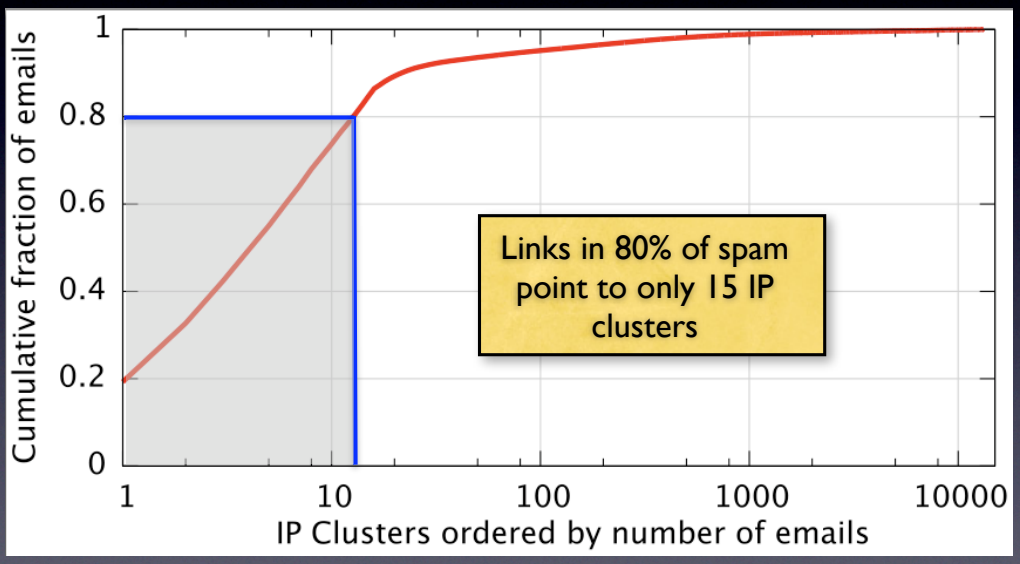
Spam is sourced by a changing set of IPs

Incoming Spam: Domain Names of embedded URLs



As expected, freshly registered DNS names propagated by spam

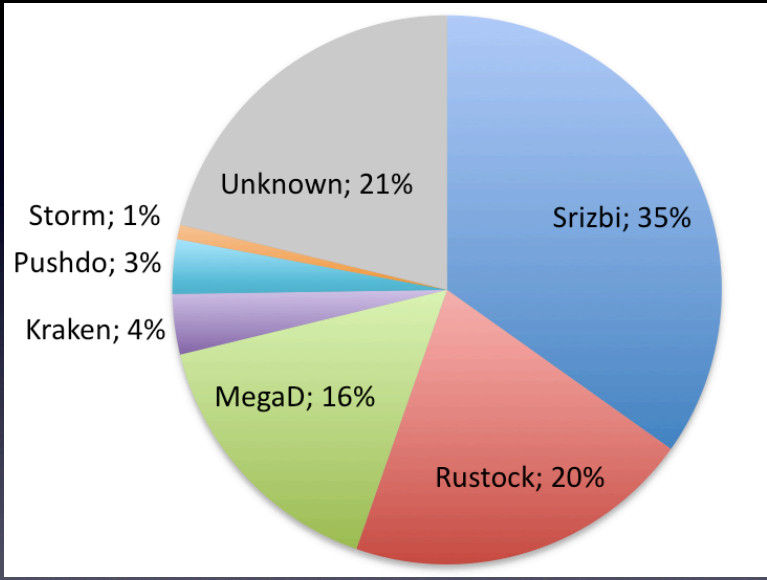
Incoming Spam: Hosting Infrastructure



Correlation Analysis

- Different botnets have different fingerprints (email subjects, recipient addresses, header formats)
- We can thus attribute incoming spam feed to specific botnets by observing the spam generated by our captive bots

Classification by Botnet



Small number of botnets source most of the spam

Spam Campaigns

	Kraken	MegaD	Pushdo	Rustock	Srizbi	Storm
Canadian Healthcare	0%	0%	0.01%	22%	3%	0%
Canadian Pharmacy	16%	28%	10%	0%	9%	6%
Diamond Watches	22%	0.1%	0%	0%	13%	0%
Downloadable Software	0%	0%	25%	0%	0%	0%
Freedom From Debt Forever!	19%	0%	0%	0%	0%	1%
Golden Gate Casino	0%	32%	0%	0%	0%	0%
KING REPLICA	0%	4%	3%	0%	15%	0%
LNHSolutions	0%	6%	0%	0%	0%	0%
MaxGain+ ... No.1 for PE	0%	0%	3%	78%	0%	0%
Prestige Replicas	7%	0%	0.3%	0%	31%	0%
VPXL - PE Made Easy	20%	8%	6%	0%	24%	55%
Unavailable	3%	22%	38%	0%	0%	24%
Other	13%	0.1%	15%	0%	5%	14%

Multiple botnets source the same spam campaign

Botnet Membership

- What fraction of the botnet members can we identify in a single day at a given location?
- Again use probabilistic analysis based on the random recipient address model
  - Let *P* is the probability that a given spam message is sent to an UW email address
  - Let *N* be the number of email messages sent by a bot over a given period
  - Then probability of UW receiving a spam message:

$$1 - e^{-N*P}$$

Botnet Membership

- Even the most gentle bots send *N = 48K* messages per day
- UW receives 2.4M messages of a total world-wide estimate of 110B messages; $P = 2.2*10^{-5}$
- Over a 24-hour uptime, probability of identifying a botnet participant is *0.65*

Applications Enabled by BotLab

- Safer browsing:
  - We found 40K malicious URLs propagated by Srizbi
  - None of them were in malware DBs (Google, etc.)
  - Further Gmail's spam filtering rate was only 21% for Srizbi.
  - BotLab can generate malware list in real-time; we have developed a Firefox plugin to check against this
- Spam filtering:
  - Developed a Thunderbird extension that compares an incoming email with the list of spam subjects and list of URLs being propagated by captive bots
  - Preliminary results are promising

Conclusions

- BotLab is an engineering exercise that pulls together many of the ideas proposed earlier
- Key components: active crawling, live execution of captive bots, network fingerprinting, and correlation
- Enables a rich set of measurements. Results include:
  - Small number of botnets generate most of the spam
  - Complex (not one-to-one) relationships between botnets, spam campaigns, and hosting infrastructures
- BotLab also promises better defenses (safe browsing, spam filtering, bot detection, etc.)