

BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic

Guofei Gu, Junjie Zhang, and Wenke Lee
College of Computing
Georgia Institute of Technology

Roadmap

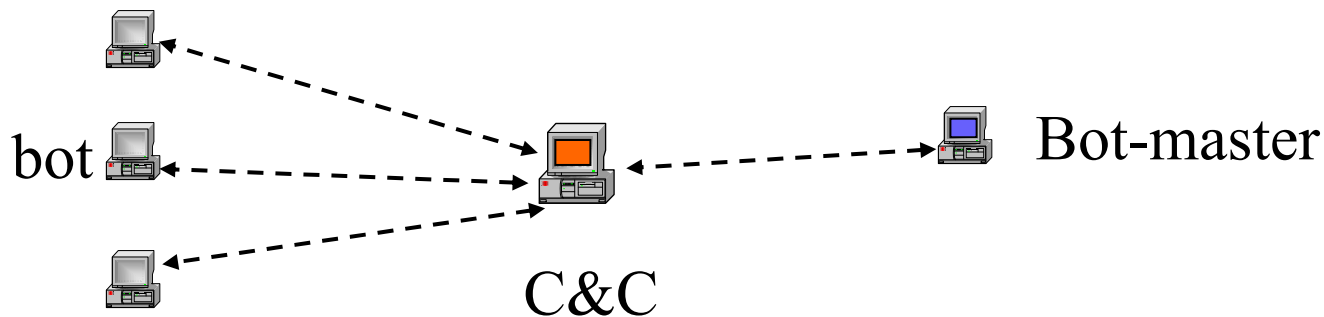
- Introduction
- BotSniffer
 - Motivation
 - Architecture
 - Algorithm
 - Experimental Evaluation
- Summary

Botnets: Big Problem

- “Attack of zombie computers is growing threat”
(New York Times)
- “Why we are losing the botnet battle”
(Network World)
- “Botnet could eat the internet”
(Silicon.com)
- “25% of Internet PCs are part of a botnet”
(Vint Cerf)

What are Bots/Botnets?

- Bot (Zombie)
 - Compromised computer controlled by botcodes (malware) without owner consent/knowledge
 - Professionally written; self-propagating
- Botnets (Bot Armies)
 - Networks of bots controlled by criminals
 - Key platform for fraud and other for-profit exploits



Botnet Epidemic

- More than 95% of all spam
- All distributed denial of service (DDoS) attacks
- Click fraud
- Phishing & pharming attacks
- Key logging & data/identity theft
- Distributing other malware, e.g., spyware/adware

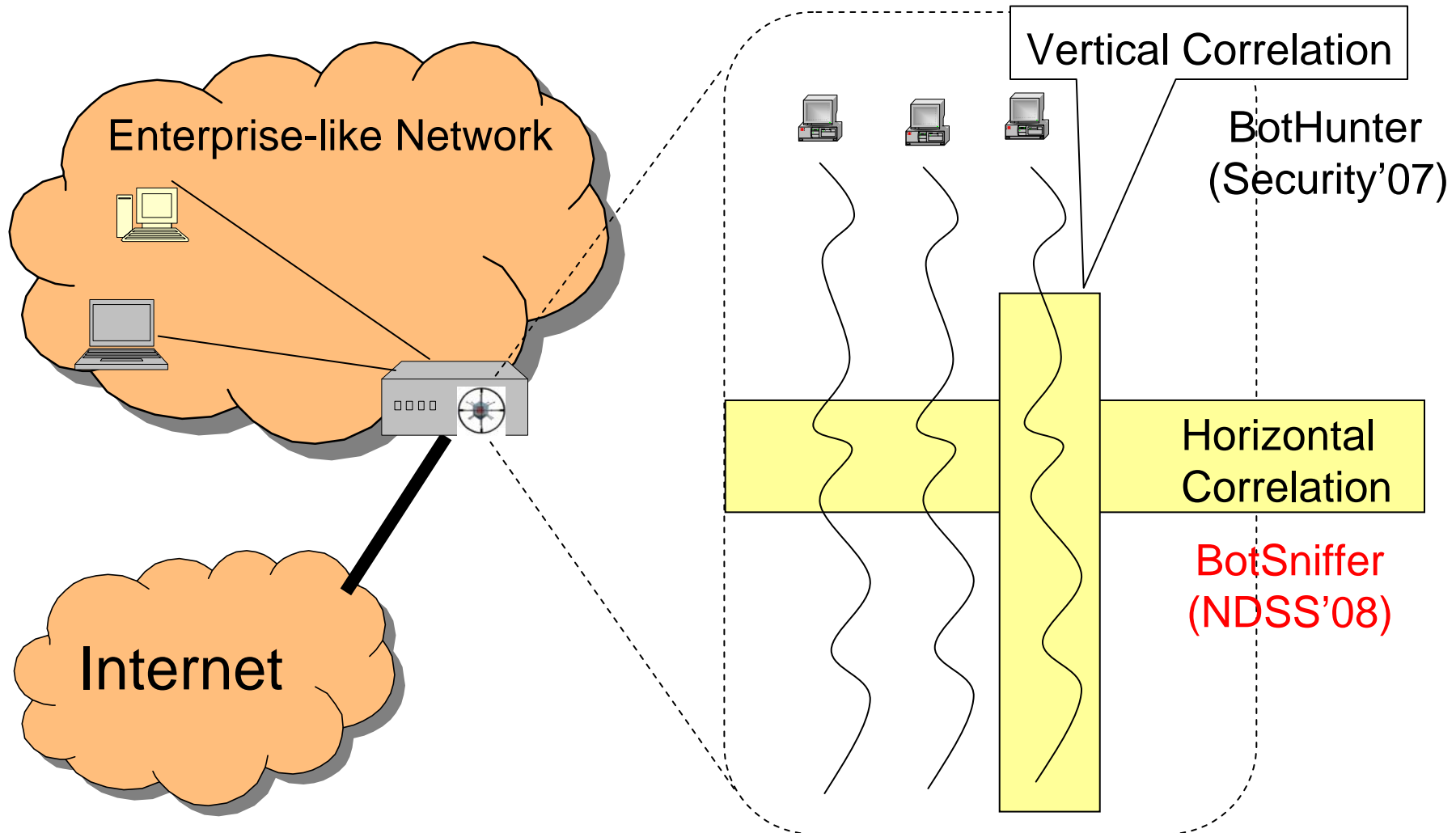
Botnet C&C Detection

- C&C is essential to a botnet
 - Without C&C, bots are just discrete, unorganized infections
- C&C detection is important
 - Relatively stable and unlikely to change within botnets
 - Reveal C&C server and local victims
 - The weakest link
- C&C detection is hard
 - Use existing common protocol instead of new one
 - Low traffic rate
 - Obscure/obfuscated communication

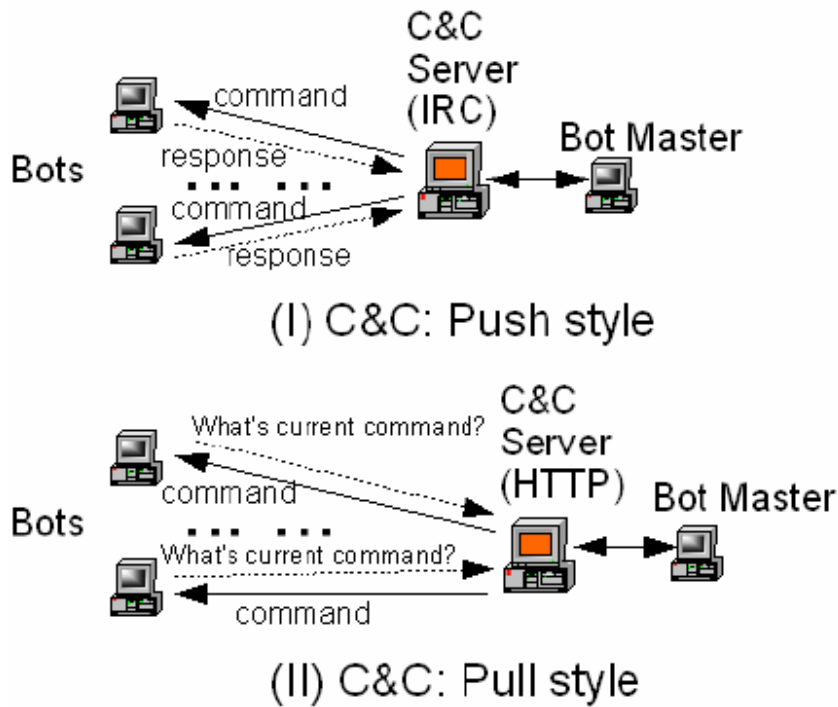
Related Work

- [Binkley, Singh 2006]: IRC-based bot detection combine IRC statistics and TCP work weight
- Rishi [Goebel, Holz 2007]: signature-based IRC bot nickname detection
- [Livadas et al. 2006]: (BBN) machine learning based approach using some general network-level traffic features (IRC botnet)
- [Karasaridis et al. 2007]: (AT&T) network flow level detection of IRC botnet controllers for backbone network (IRC botnet)
- [Gu et al. 2007]: BotHunter

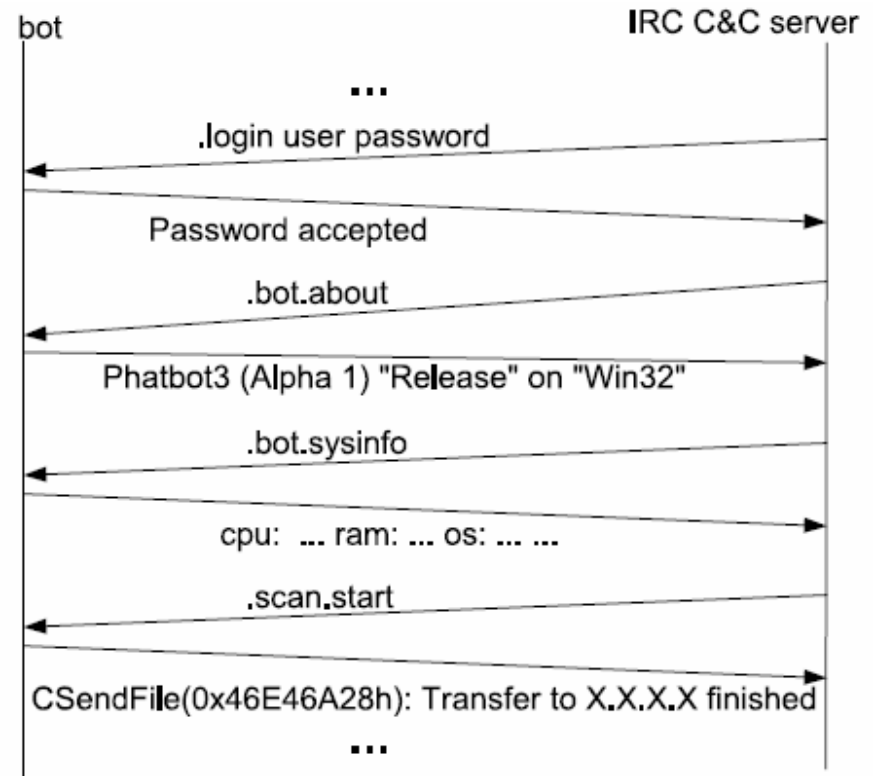
Our Approaches: General Picture



Botnet C&C Communication

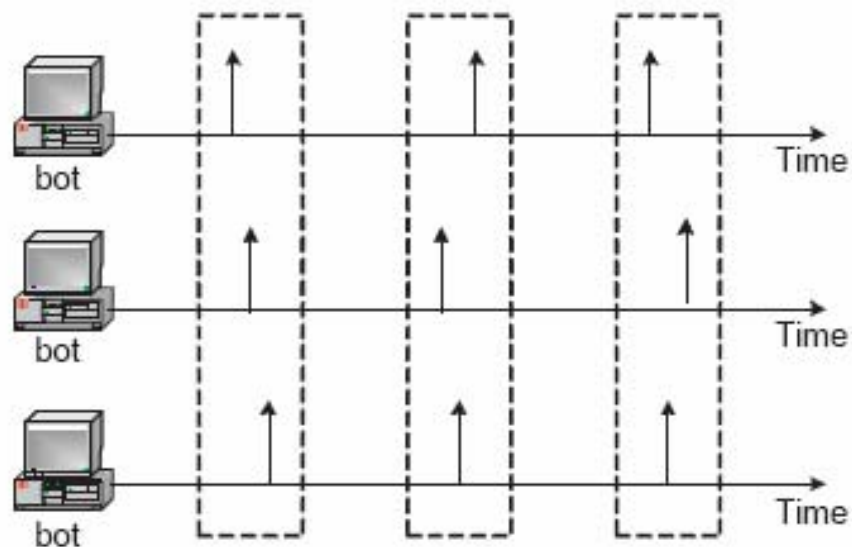


(a) Two styles of botnet C&C



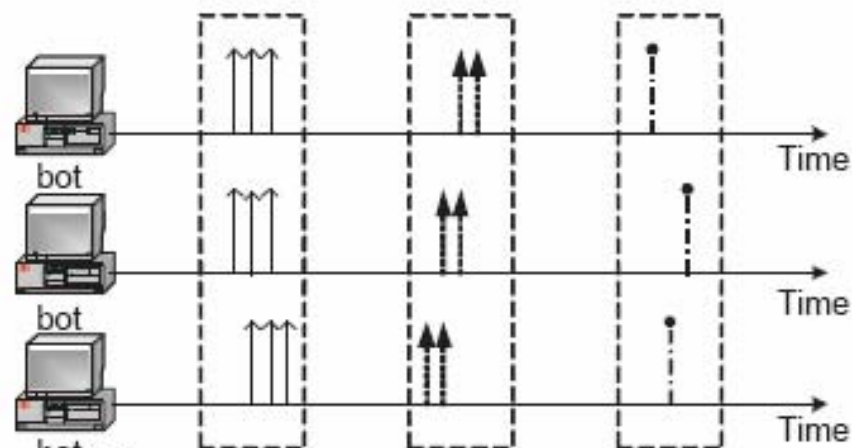
(b) An IRC-based C&C communication example




Botnet C&C: Spatial-Temporal Correlation and Similarity



Message Response (e.g., IRC PRIVMSG)

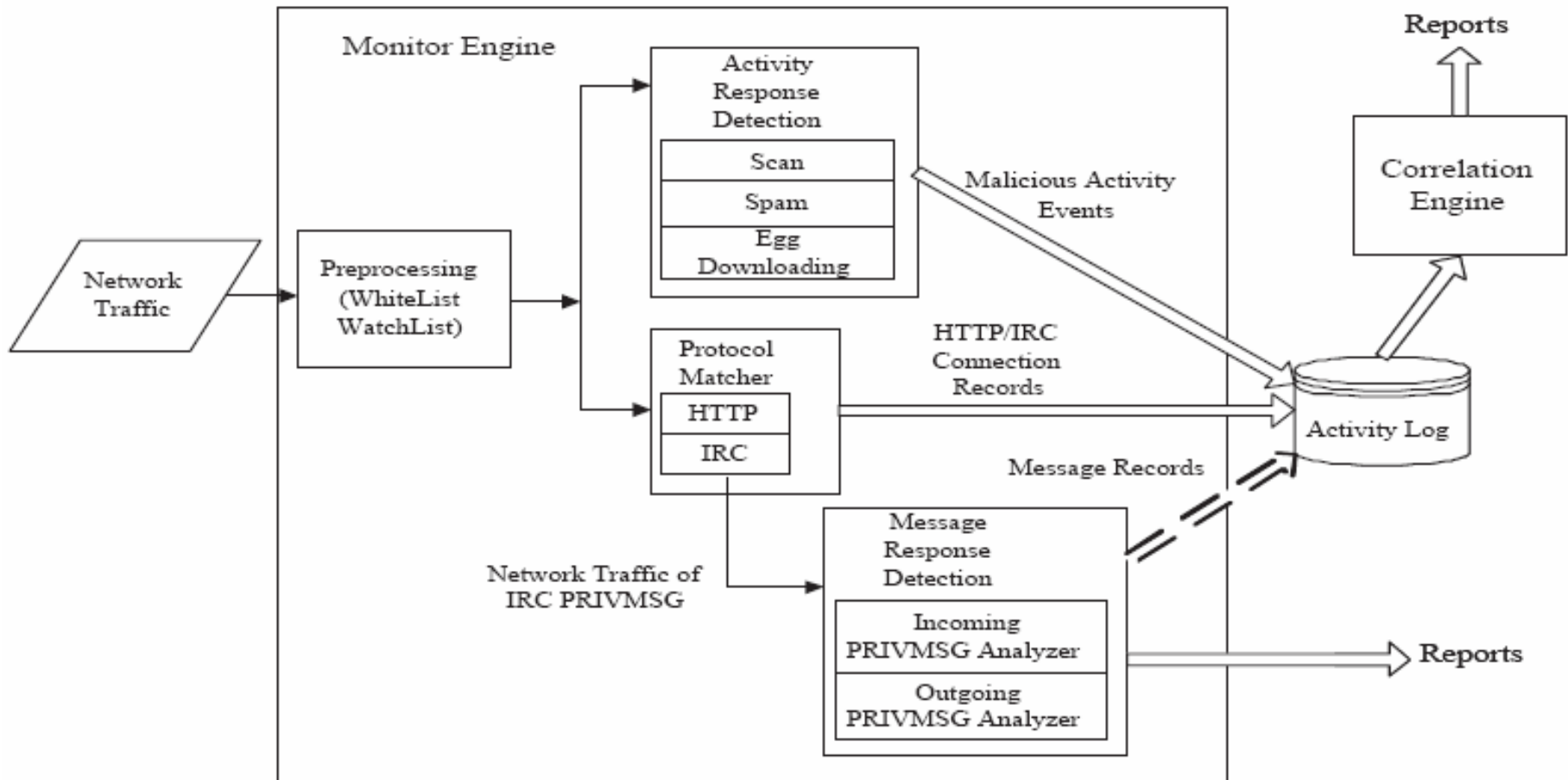
(a) Message response crowd



 Activity Response (network scanning)
 Activity Response (sending spam)
 Activity Response (binary downloading)

(b) Activity response crowd

BotSniffer Architecture



Correlation Engine

- Group clients according to their destination IP and Port pair (HTTP/IRC connection record)
- Perform a *group analysis* on spatial-temporal correlation and similarity property
 - Response-Crowd-**Density**-Check
 - Response-Crowd-**Homogeneity**-Check

Response-Crowd-Density-Check Algorithm

- Response crowd
 - a set of clients that have (message/activity) response behavior
- A **Dense** response crowd
 - the fraction of clients with message/activity behavior within the group is larger than a threshold (e.g., 0.5).
- Example: 5 clients connected to the same IRC/HTTP server, and all of them scanned at similar time (or send IRC messages at similar time)
- Accumulate the degree of suspicion
 - Sequential Probability Ratio Testing (SPRT)

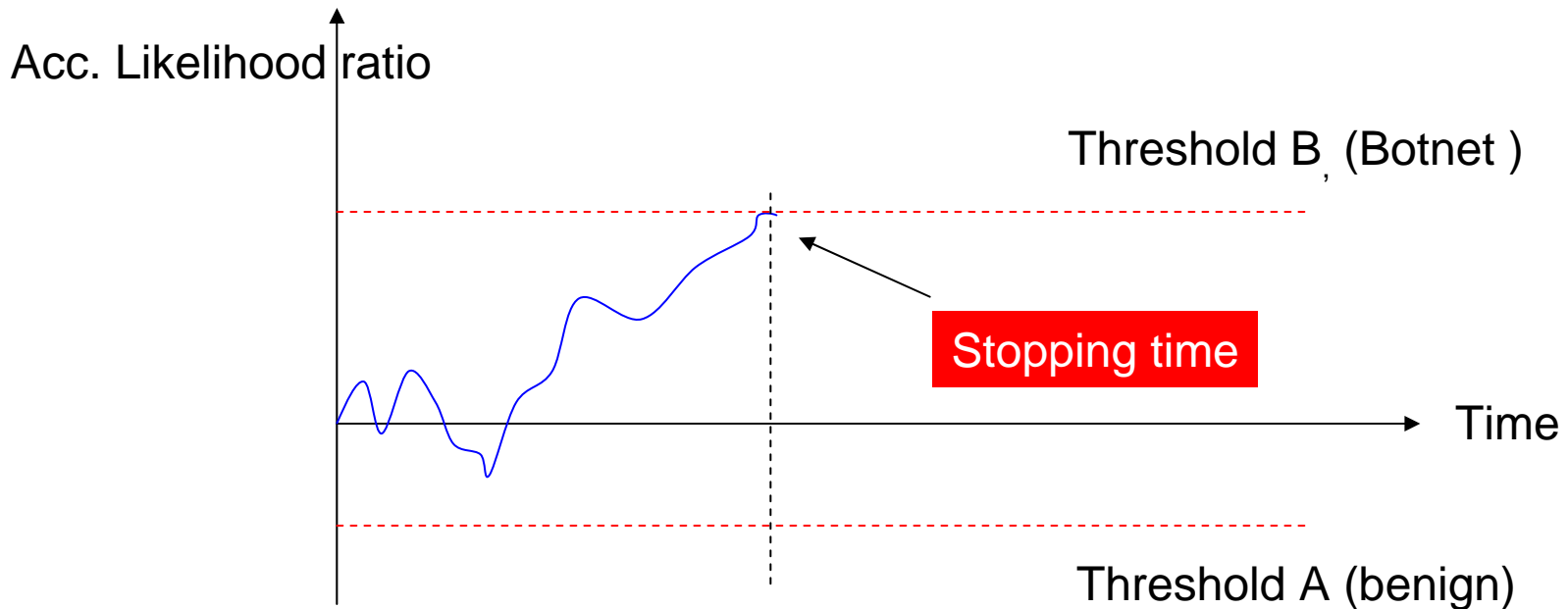
Sequential Probability Ratio Testing (SPRT)

- Each round, observe whether current crowd is dense or not ($Y=1$ or $Y=0$)
 - Hypothesis
 - $\Pr(Y=1|H_1)$ very high (for botnet)
 - $\Pr(Y=1|H_0)$ very low (for benign)
- Update accumulated likelihood ratio according to the observation Y

$$\Lambda_n = \ln \frac{\Pr(Y_1, \dots, Y_n | H_1)}{\Pr(Y_1, \dots, Y_n | H_0)} = \ln \frac{\prod_i \Pr(Y_i | H_1)}{\prod_i \Pr(Y_i | H_0)} = \sum_i \ln \frac{\Pr(Y_i | H_1)}{\Pr(Y_i | H_0)}$$

- After several rounds, we may reach a decision (which hypothesis is more likely, H_1 or H_0)

Sequential Probability Ratio Testing (cont.)



- Also called TRW (Threshold Random Walk)
- Bounded false positive and false negative rate (as desired), and usually needs only a few rounds

Response-Crowd-Homogeneity-Check Algorithm

- A **homogeneous** response crowd
 - Many members have very **similar** responses

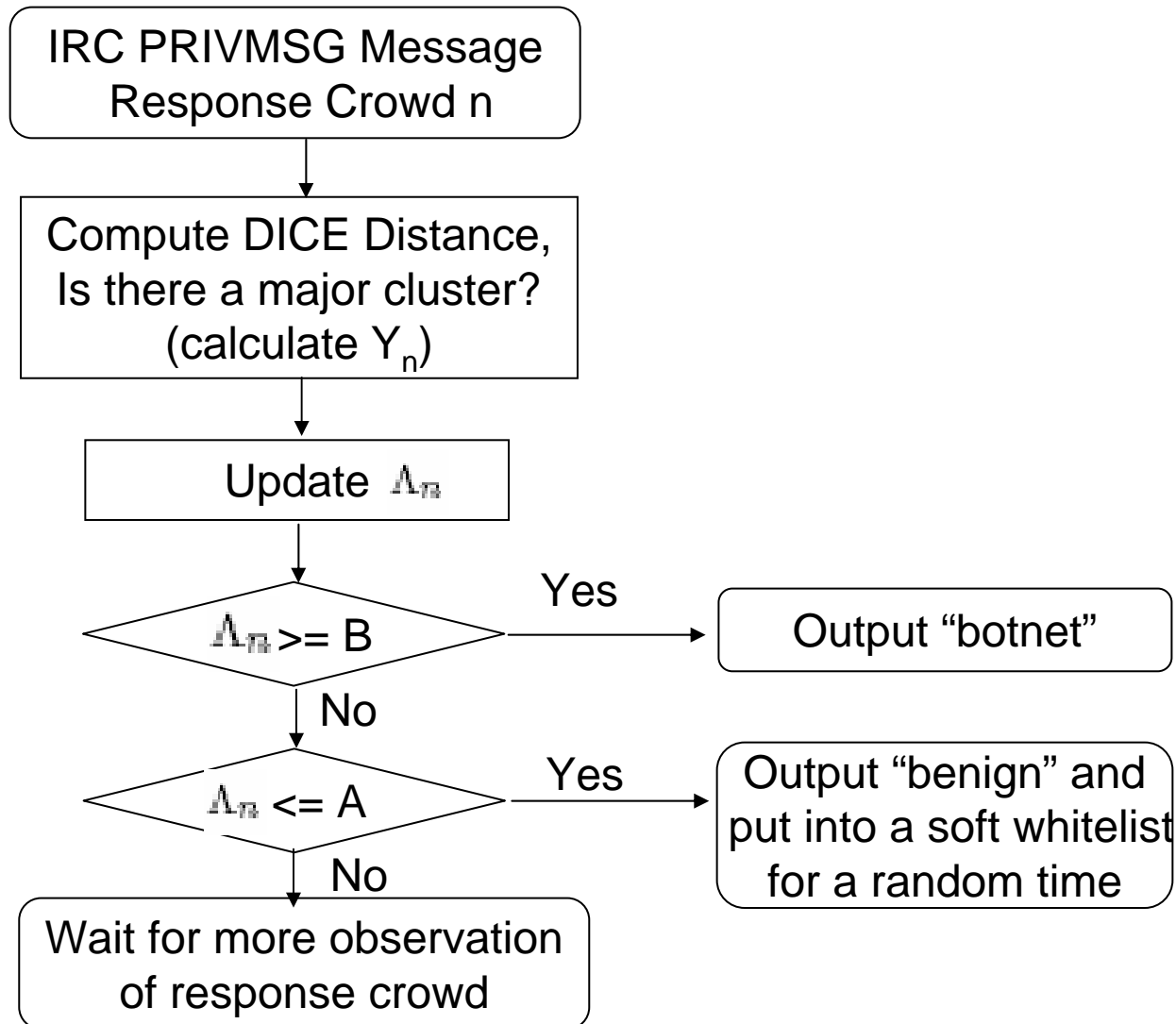
- Similarity is defined

- Message response
 - Similar payload (DICE distance)

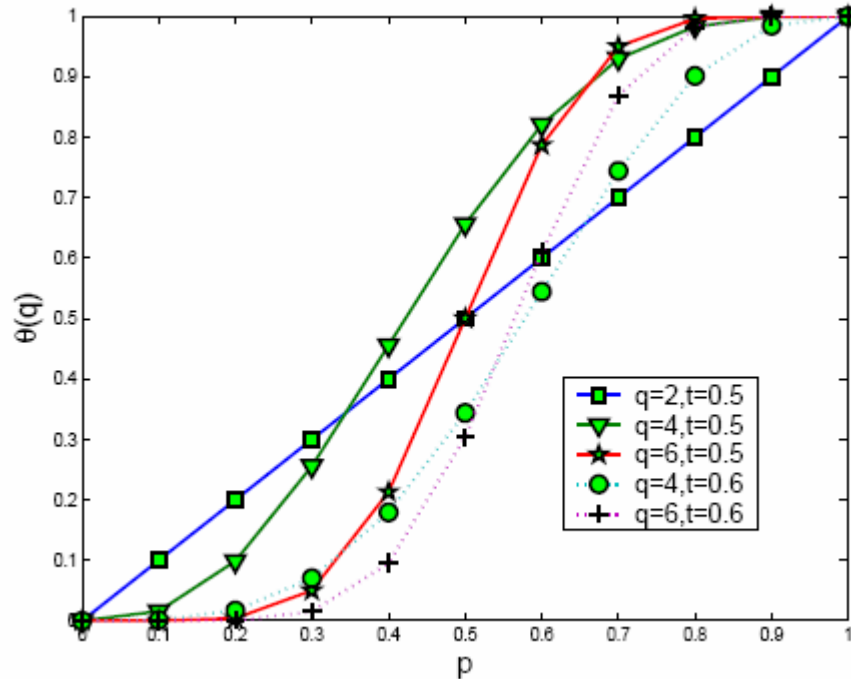
$$Dice(X, Y) = \frac{2|ngrams(X) \cap ngrams(Y)|}{|ngrams(X)| + |ngrams(Y)|}$$

- E.g., “abcde” and “bcdef”, common 2-grams: “bc,cd,de”, DICE distance is $2*3/(4+4)=6/8=0.75$
- Activity response (examples)
 - Scan same ports
 - Download same binary
 - Send similar spams

Real-Time IRC Message Correlation Flow Diagram



Crowd Homogeneity: Relationship with Number of Clients



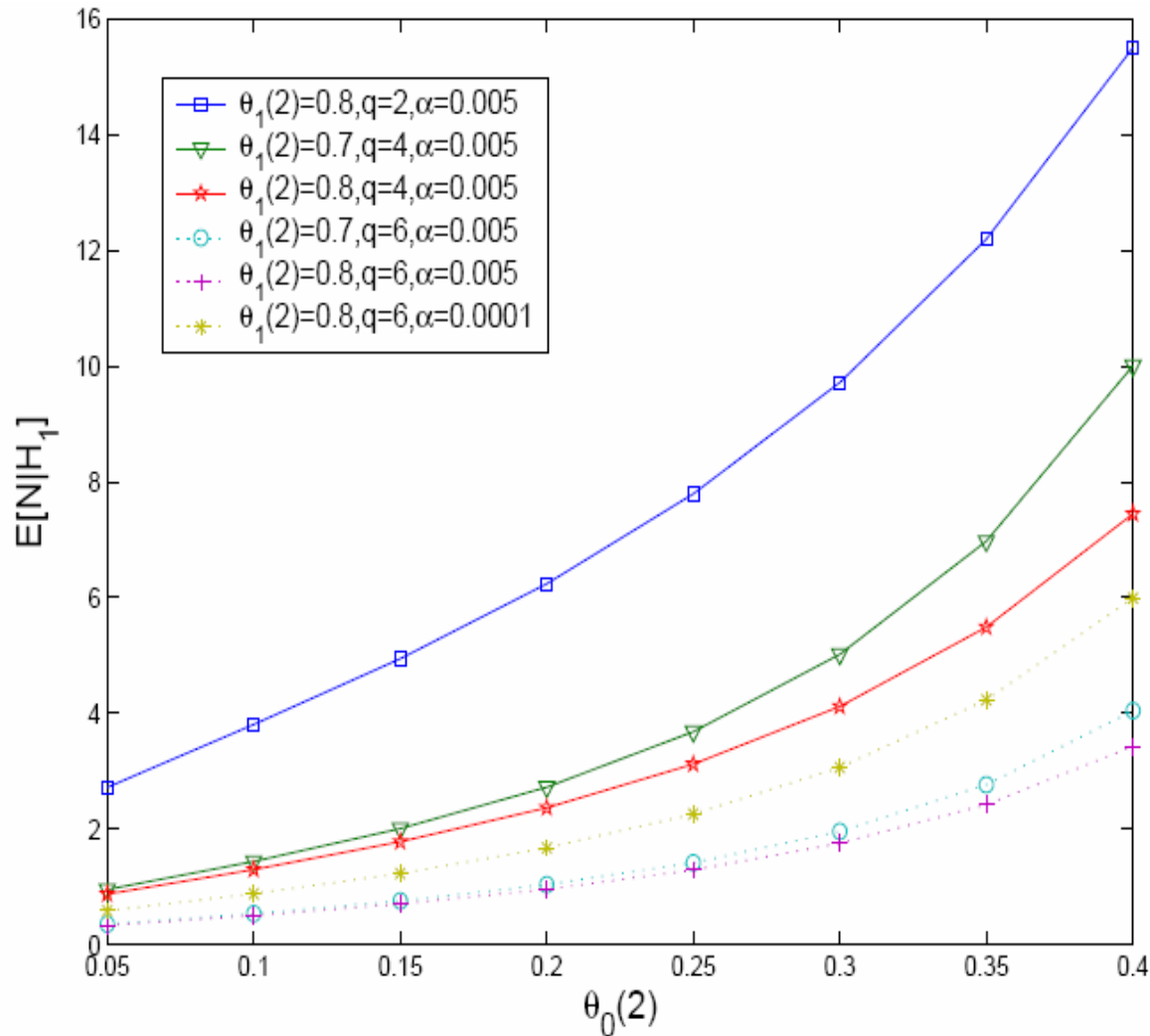
For a botnet, more clients, higher probability of crowd homogeneity

For normal IRC channel, more clients, lower probability of crowd homogeneity

q: #clients t: threshold in clustering

$P=\theta(2)$: basic probability of two clients sending similar messages

Number of Rounds Needed



Experiment

189 days' of IRC traffic

Trace	trace size	duration	Pkt	TCP flows	(IRC/Web) servers	FP
IRC-1	54MB	171h	189,421	10,530	2,957	0
IRC-2	14MB	433h	33,320	4,061	335	0
IRC-3	516MB	1,626h	2,073,587	4,577	563	6
IRC-4	620MB	673h	4,071,707	24,837	228	3
IRC-5	3MB	30h	19,190	24	17	0
IRC-6	155MB	168h	1,033,318	6,981	85	1
IRC-7	60MB	429h	393,185	717	209	0
IRC-8	707MB	1,010h	2,818,315	28,366	2,454	1
All-1	4.2GB	10m	4,706,803	14,475	1,625	0
All-2	6.2GB	10m	6,769,915	28,359	1,576	0
All-3	7.6GB	1h	16,523,826	331,706	1,717	0
All-4	15GB	1.4h	21,312,841	110,852	2,140	0
All-5	24.5GB	5h	43,625,604	406,112	2,601	0

Experiment (cont.)

BotTrace	trace size	duration	Pkt	TCP flow	Detected
B-IRC-G	950k	8h	4,447	189	Yes
B-IRC-J-1	-	-	143,431	-	Yes
B-IRC-J-2	-	-	262,878	-	Yes
V-Rbot	26MB	1,267s	347,153	103,425	Yes
V-Spybot	15MB	1,931s	180,822	147,921	Yes
V-Sdbot	66KB	533s	474	14	Yes
B-HTTP-I	6MB	3.6h	65,695	237	Yes
B-HTTP-II	37MB	19h	395,990	790	Yes

Thanks David Dagon, Fabian Monrose, and Chris Lee
 for providing some of the evaluation traces

BotSniffer Summary

- Exploiting the underlying spatial-temporal correlation and similarity property of botnet C&C (horizontal correlation)
- New anomaly-based detection algorithm
- New Botnet C&C detection system: BotSniffer
- Detected real-world botnets with a very low false positive rate

Future Work

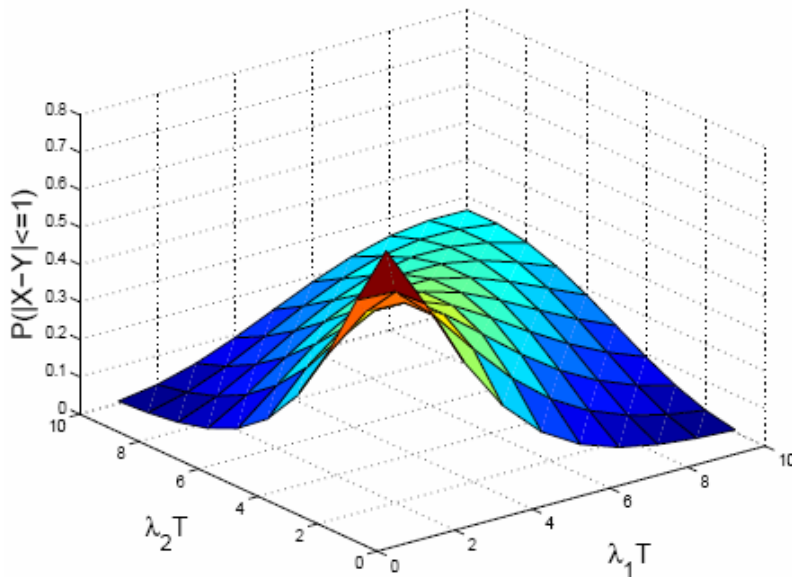
- Improving accuracy and resilience to evasion
- BotMiner: protocol- and structure-independent botnet detection technique

Thanks!

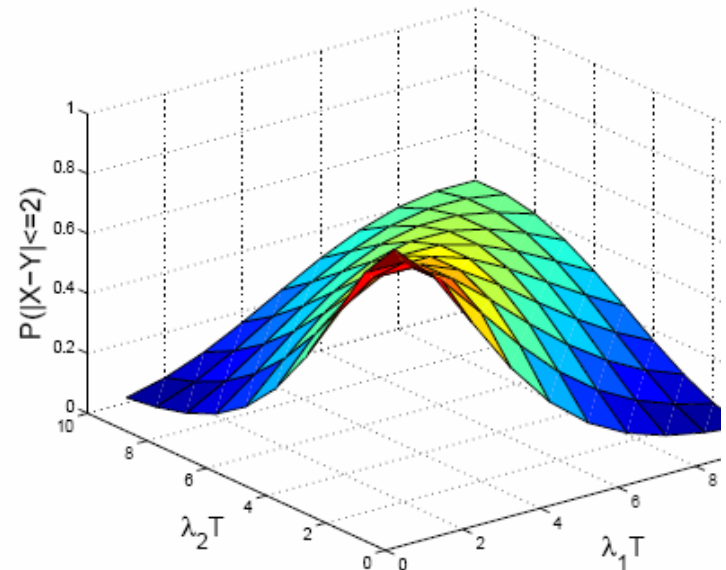
Q&A

[Http://www.cc.gatech.edu/~guofei](http://www.cc.gatech.edu/~guofei)

Probability of Having Two Similar Length Messages



(a) Probability of $P(|X - Y| \leq 1)$

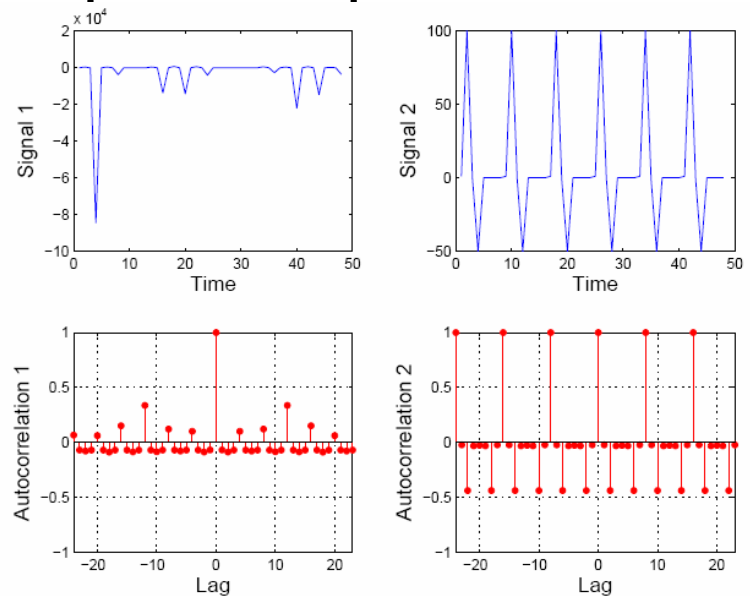


(b) Probability of $P(|X - Y| \leq 2)$

Probability of having two similar content messages are even lower

Single Client C&C Detection Under Certain Conditions

- IRC: broadcast in the channel
 - similar to the case we can monitor multiple message responses from multiple clients in the group
- HTTP: AutoCorrelation to find periodic patterns from background noise



BotSniffer Extension and Limitation

- Improving BotSniffer
 - Using activity response crowd *homogeneity*
 - Extension of suspicious C&C protocol matchers
- Possible evasion
 - Effect of encryption
 - Evasion by exploiting time window
 - Evasion by using random delay/period, injecting random noise, injecting random garbage in the packet