BOTMINER

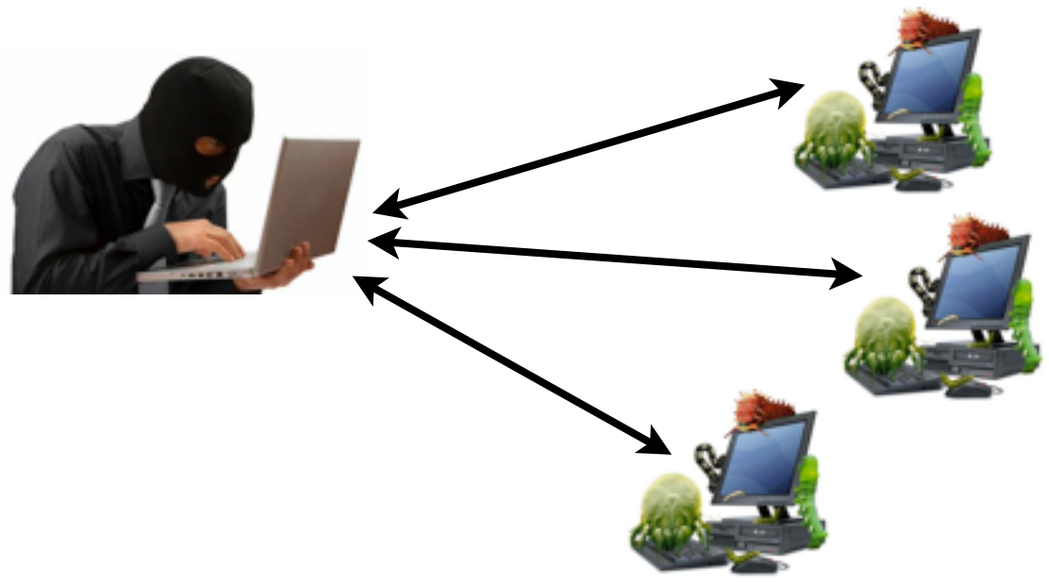
CLUSTERING ANALYSIS OF NETWORK TRAFFIC FOR PROTOCOL- AND STRUCTURE-INDEPENDENT BOTNET DETECTION

Guofei Gu, Roberto Perdisci, Junjie Zhang, Wenke Lee

Georgia Tech Information Security Center

Presenter: Roberto Perdisci

Botnets



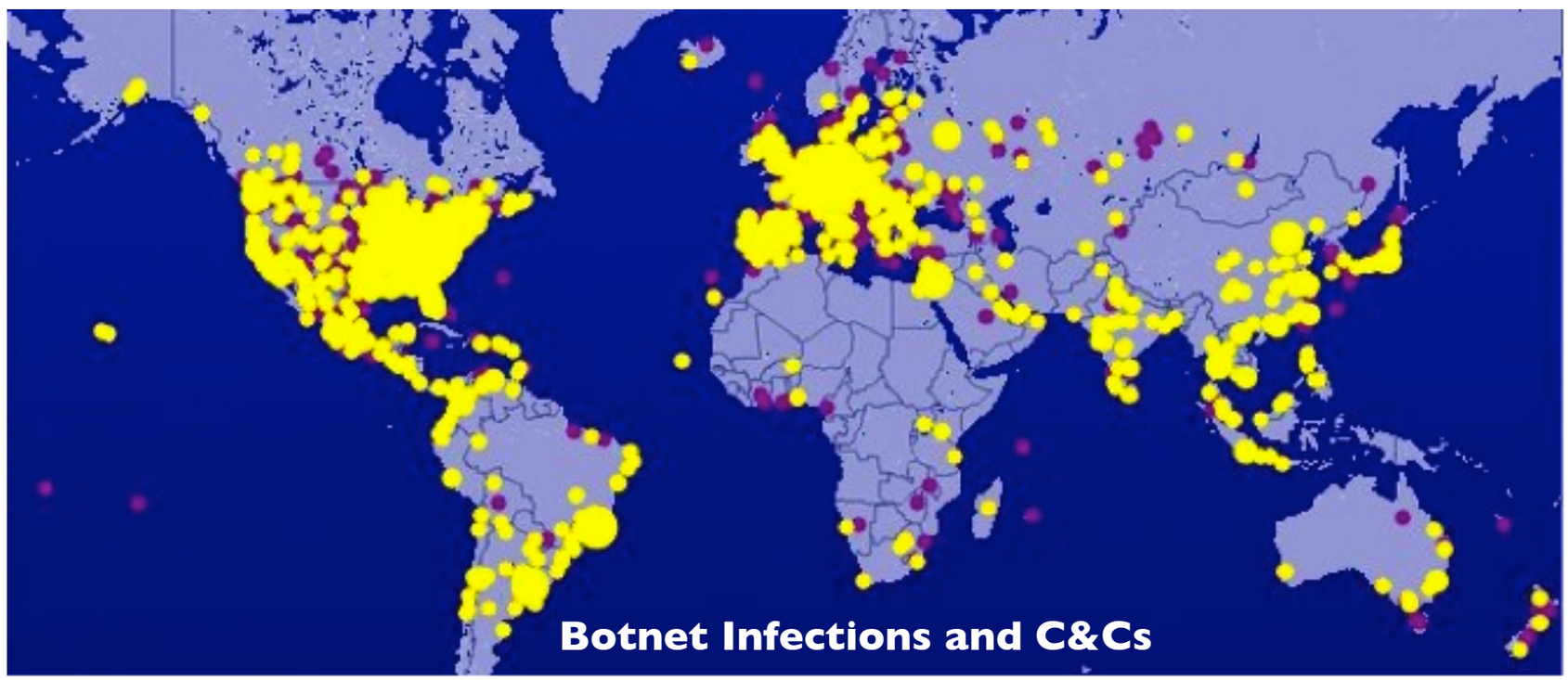
- Bot
  - an instance of malware that runs on a compromised machine, without the owner's consent
  - the bot connects to a Command and Control (C&C) channel and waits to receive commands
- Botnet
  - A group of bot-compromised machines controlled by a *botmaster*
  - Bots in the same botnet receive commands from the same botmaster and respond/act in a (loosely) coordinated way

Botnets for profit!

- Many cyber-crimes are perpetrated using botnets
  - Basically all the DDoS attacks
  - Send SPAM
    - >90% of all email-related Internet traffic comes from SPAM
    - >95% of all SPAM is sent using Botnets
  - Click Fraud
  - Information Theft
  - Provide infrastructure for Phishing attacks
  - Massive exploits (e.g., SQL injection attacks)
  - Distribute other malware
  - ...

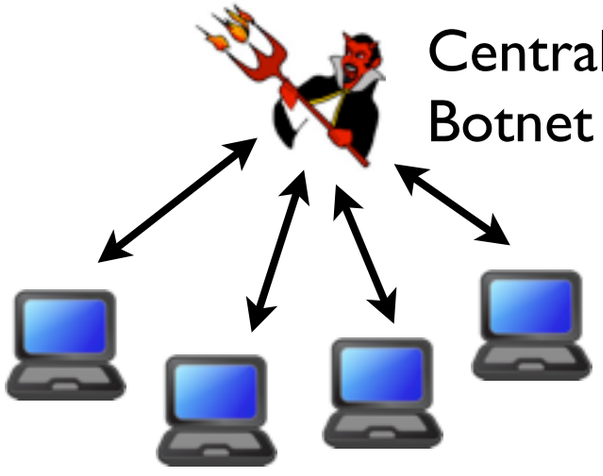
The Botnet Phenomenon

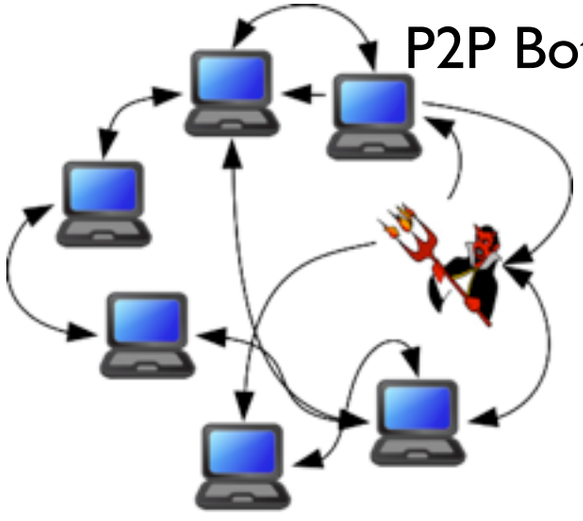
- Botnets are widespread
- Millions of computers on the Internet are bot-infected, according to some statistics

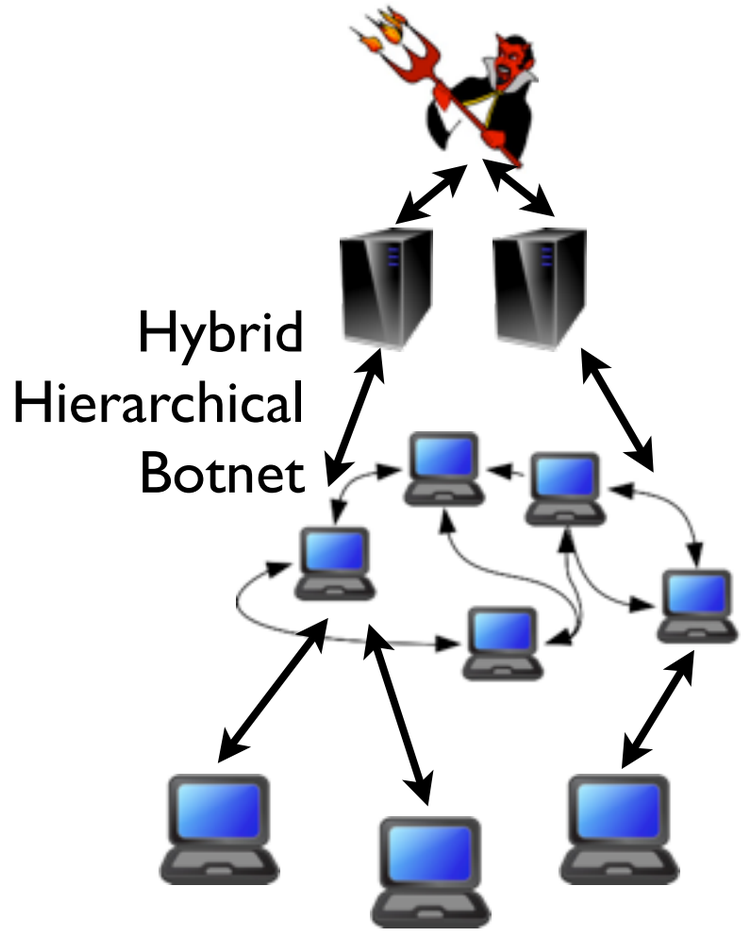


Botnet Architectures

- Centralized Botnet
  - protocols used to communicate with the C&C server : HTTP, IRC, proprietary
- P2P Botnet
  - distributed C&C, uses either known or proprietary P2P protocols to communicate with botmaster
- Hybrid/Hierarchical
  - both centralized and P2P components
  - components are organized in a hierarchy







Related Work

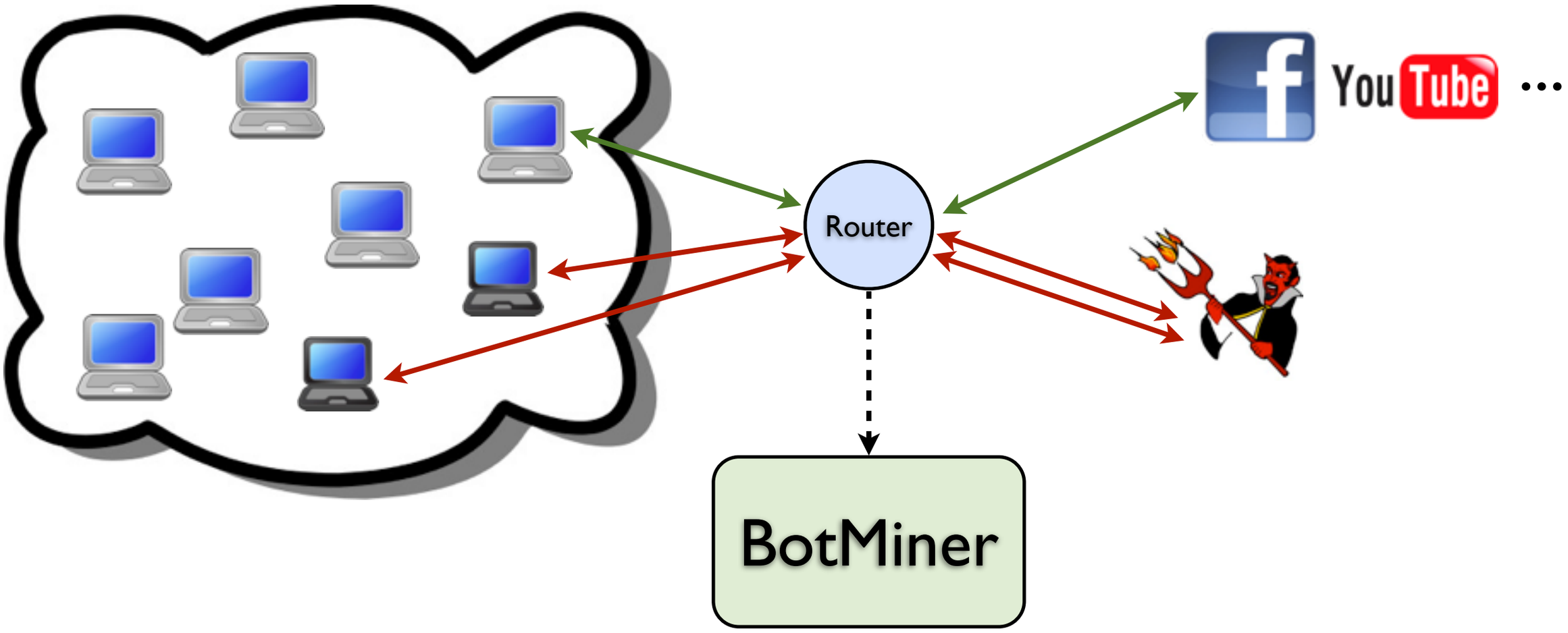
- [Rajab et al. 2006]: measuring IRC botnets
- Rishi [Goebel, Holz 2007]: signature-based IRC bot nickname detection
- [Livadas et al. 2006, Karasaridis et al. 2007]: (BBN, AT&T) network flow level detection of IRC botnets
- BotHunter [Gu et al. Security'07]: detect bots based based on a model of the infection cycle
- BotSniffer [Gu et. al NDSS'08]: spatial-temporal correlation to detect centralized botnet C&C (IRC/HTTP)
- TAMD [Yen, Reiter 2008]: traffic aggregation to detect botnets that use a centralized C&C structure

Challenges

- Packing/obfuscation prevents signature-based detection of malicious executable files
- Rootkits used to hide from sys-level analysis
- Bots evolve, and so does their behavior
- Botnets can have very flexible and diverse C&C structure
- Building a model by looking at single bots is not likely to generalize well

How does BotMiner help?

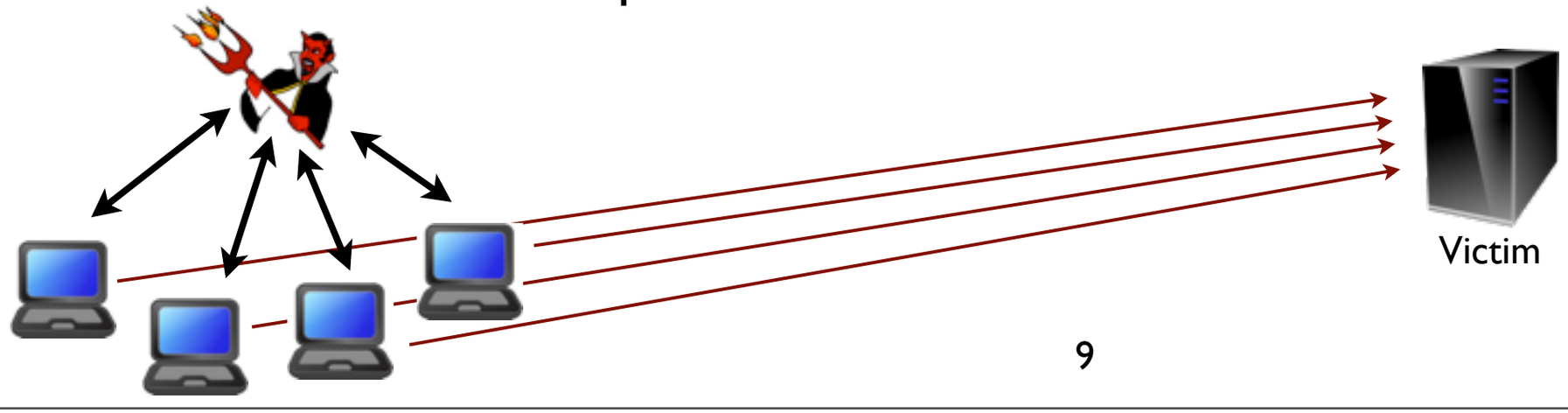
- Network-level botnet detection solution



Motivations and Intuitions

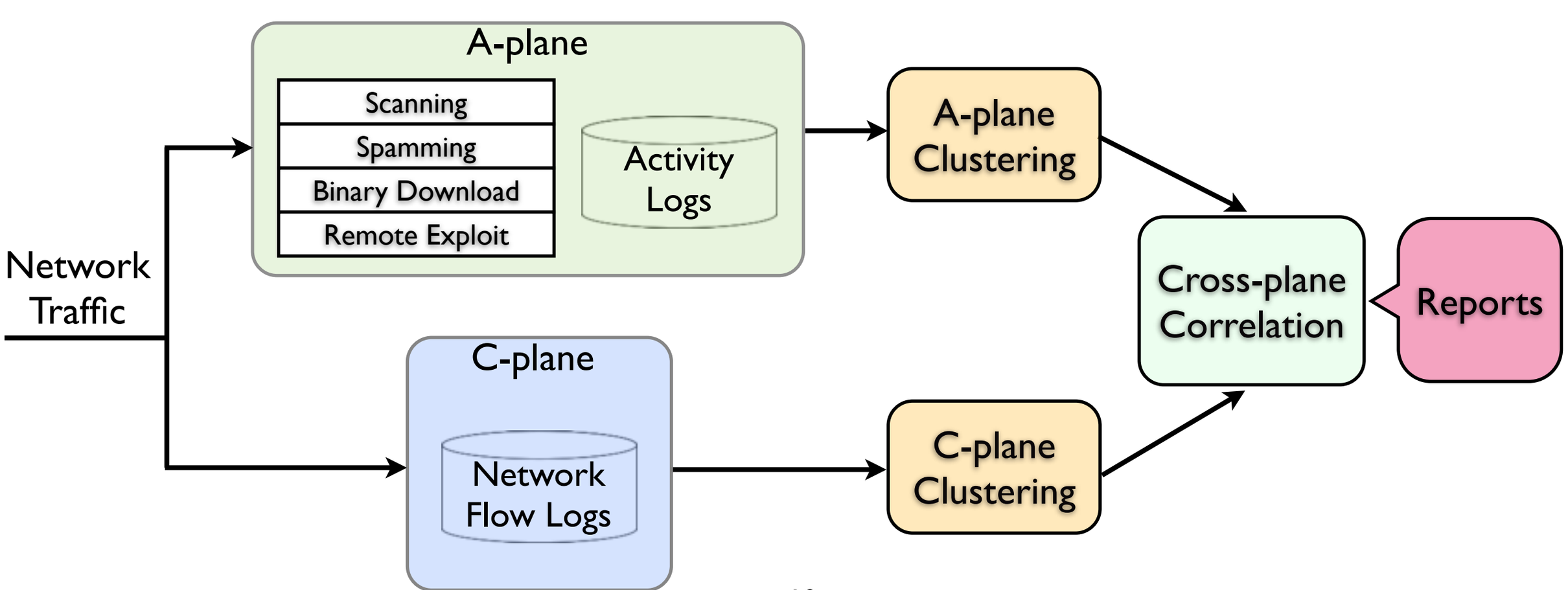
- Botnets may use different protocols and C&C infrastructure
- Communications may be encrypted
- The C&C server(s) may change frequently

- We need a protocol- and structure-independent detection approach
- BotMiner is based on characteristics that are constant in botnets
  - Bots are a long-term commodity for the botmaster
  - Bots belonging to the same botnet share the same C&C and communicate with the botmaster in a similar way
  - Bots respond to commands in similar, coordinated way



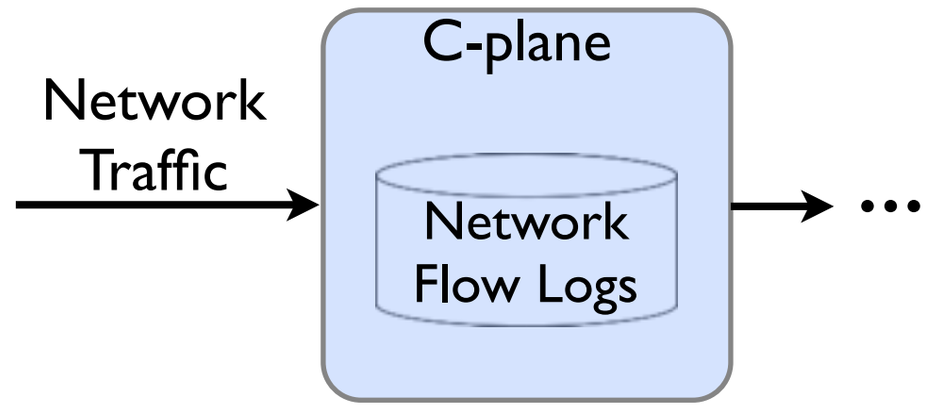
BotMiner System Architecture

- We monitor two planes
  - C-plane (C&C communications): "who is talking to whom"
  - A-plane (malicious activities): "who is doing what"

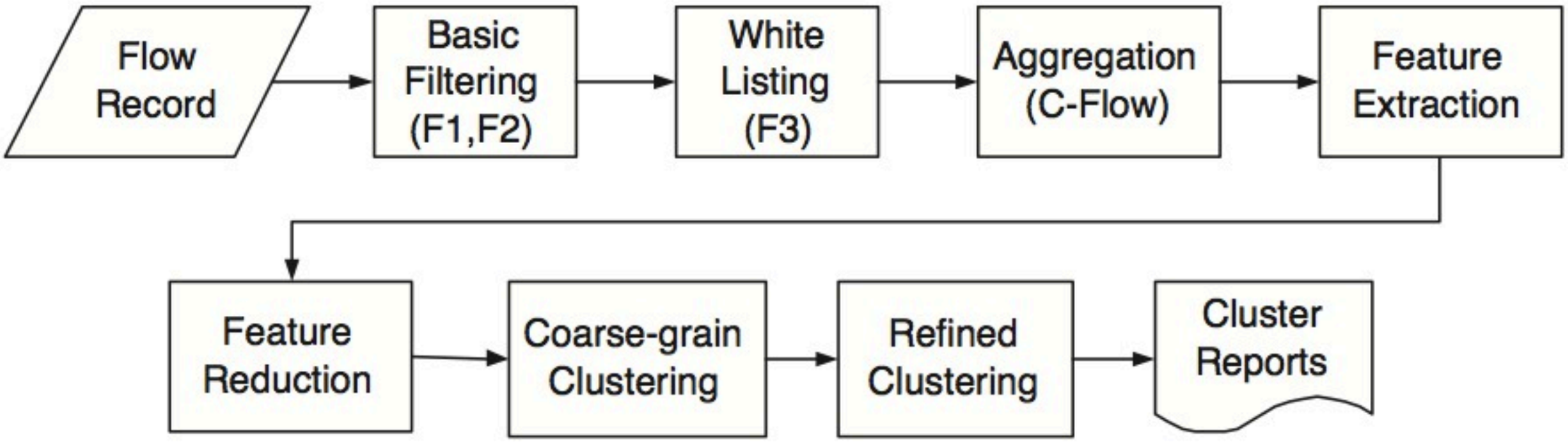


C-plane monitor

- Captures network flows (*who is talking to whom*)
  - this has to be done very efficiently to avoid packet loss at the kernel level
  - we use `fcapture` to produce short logs that record
    - start time, duration, srcIP, srcPort, dstIP, dstPort, number of packets, number of bytes transfered in both directions



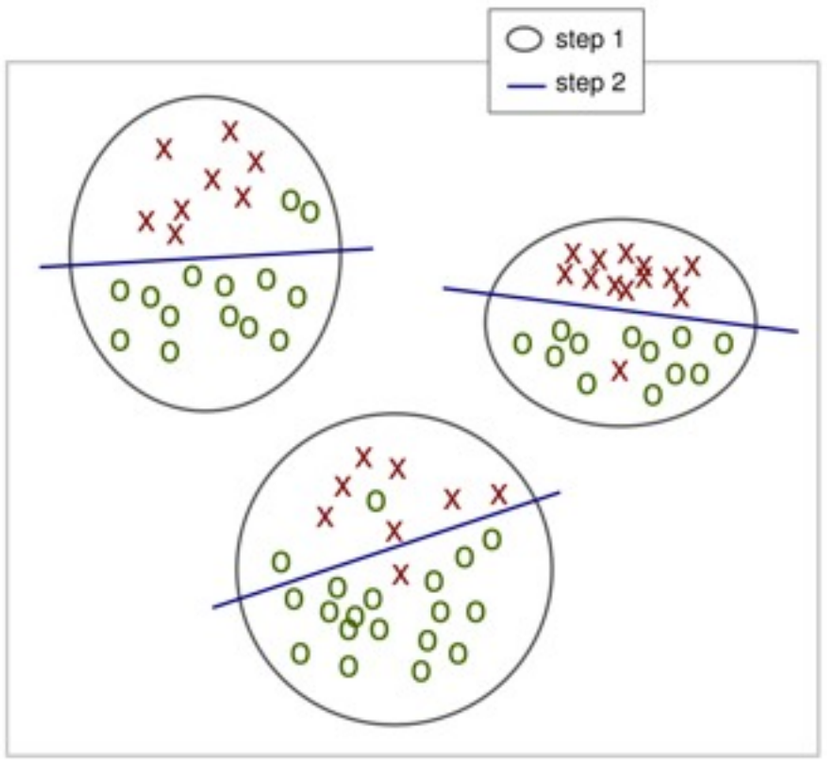
C-plane Clustering



- C-flow = $\{f_i\}$, aggregates the f_i observed in one epoch E that have srcIP, dstIP, and dstPort in common
- C-flow features: FPH, PPF, BPP, BPS
- We want to group together similar C-flows (find similar communication patterns)

C-plane clustering

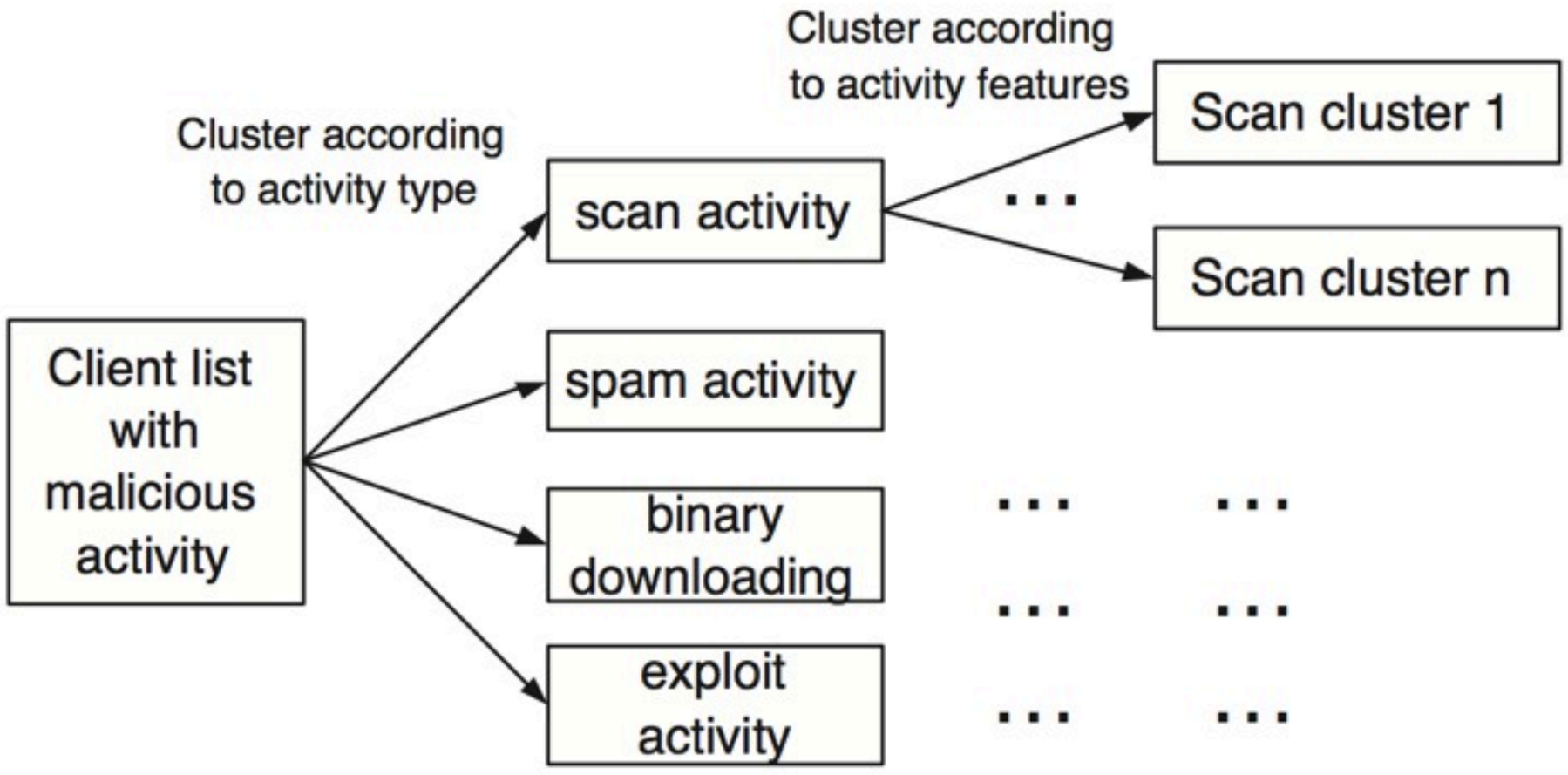
- Performed in two steps using X-means
  - coarse-grained clustering on entire dataset, but reduced feature set
  - finer-grained clustering on multiple smaller clusters using all the features



- Reduced feature set
  - avg, std-dev of each feature
- Full feature set
  - 13 bins per feature to approximate their distribution

A-plane Clustering

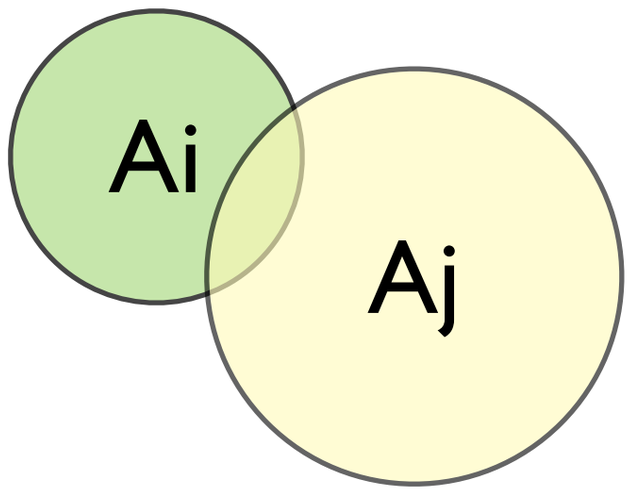
- Groups hosts that perform similar suspicious activities

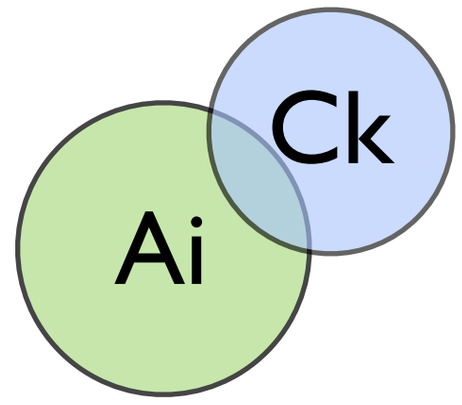


Cross-plane Correlation

- Botnet score *s(h)*
  - higher if *h* was involved in multiple suspicious activities
  - higher if there is a large overlap between activity clusters containing *h* and communication clusters containing *h*
  - *s(h)* > threshold => *h* is likely a bot

$$s(h) = \sum_{\substack{i,j \\ j>i \\ t(A_i)\neq t(A_j)}} w(A_i)w(A_j)\frac{|A_i \cap A_j|}{|A_i \cup A_j|} + \sum_{i,k} w(A_i)\frac{|A_i \cap C_k|}{|A_i \cup C_k|}$$



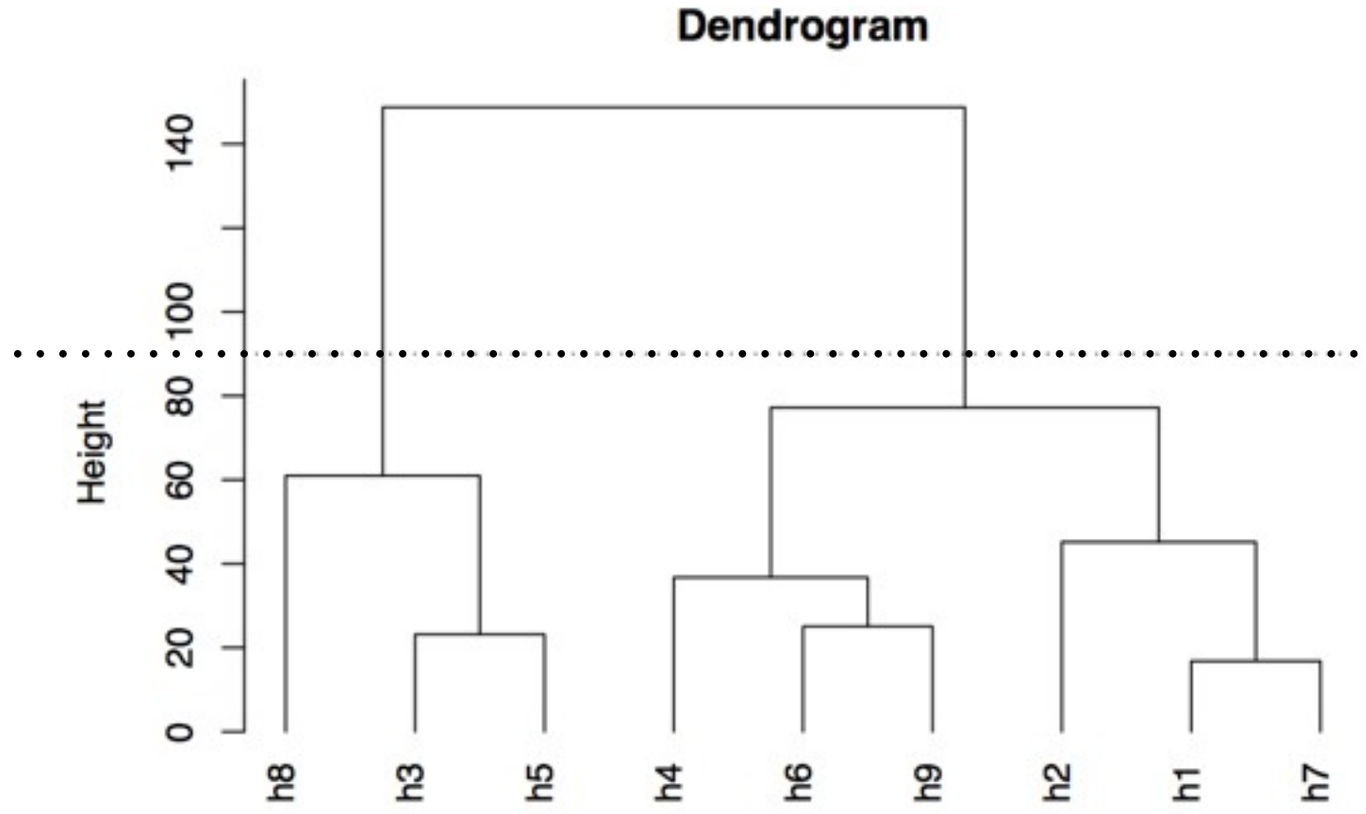


Finding Bot*nets*

- *h* = 1 0 0 1 0 0 0 1 0 1 0 1 1 0 1 1 1 1 0 1

$$sim(h_i, h_j) = \sum_{k=1}^{m_B} I(b_k^{(i)} = b_k^{(j)}) + I(\sum_{k=m_B+1}^{m_B+n_B} I(b_k^{(i)} = b_k^{(j)}) \geq 1)$$

Hierarchical Clustering



Experimental Setup

- 10-day network trace from GT-CoC
  - Considered normal traffic
  - 200-300Mbps pick
- Traffic generated by 8 different botnets

Trace	Size	Duration	Pkt	TCP/UDP flows	Botnet clients	C&C server
Botnet-IRC-rbot	169MB	24h	1,175,083	180,988	4	1
Botnet-IRC-sdbot	66KB	9m	474	19	4	1
Botnet-IRC-spybot	15MB	32m	180,822	147,945	4	1
Botnet-IRC-N	6.4MB	7m	65,111	5635	259	1
Botnet-HTTP-1	6MB	3.6h	65,695	2,647	4	1
Botnet-HTTP-2	37MB	19h	395,990	9,716	4	1
Botnet-P2P-Storm	1.2G	24h	59,322,490	5,495,223	13	P2P
Botnet-P2P-Nugache	1.2G	24h	59,322,490	5,495,223	82	P2P

Experimental Results

- Apply detection system on only legitimate traffic first

Trace	Step-1 C-clusters	Step-2 C-clusters	A-plane logs	A-clusters	False Positive Clusters	FP Rate
Day-1 (TCP/UDP)	1,374	4,958	1,671	1	0	0 (0/878)
Day-2 (TCP/UDP)	904	2,897	5,434	1	1	0.003 (2/638)
Day-3 (TCP/UDP)	1,128	2,480	4,324	1	1	0.003 (2/692)
Day-4 (TCP/UDP)	1,528	4,089	5,483	4	4	0.01 (9/871)
Day-5 (TCP/UDP)	1,051	3,377	6,461	5	2	0.0048 (4/838)
Day-6 (TCP)	1,163	3,469	6,960	3	2	0.008 (7/877)
Day-7 (TCP)	954	3,257	6,452	5	2	0.006 (5/835)
Day-8 (TCP)	1,170	3,226	8,270	4	2	0.0091 (8/877)
Day-9 (TCP)	742	1,763	7,687	2	0	0 (0/714)
Day-10 (TCP)	712	1,673	7,524	0	0	0 (0/689)

- Botnet traffic is overlaid to normal traffic (one botnet trace at a time)
- Simulates realistic scenario to measure FPs and DR

Botnet	Number of Bots	Detected?	Clustered Bots	Detection Rate	False Positive Clusters/Hosts	FP Rate
IRC-rbot	4	YES	4	100%	1/2	0.003
IRC-sdbot	4	YES	4	100%	1/2	0.003
IRC-spybot	4	YES	3	75%	1/2	0.003
IRC-N	259	YES	258	99.6%	0	0
HTTP-1	4	YES	4	100%	1/2	0.003
HTTP-2	4	YES	4	100%	1/2	0.003
P2P-Storm	13	YES	13	100%	0	0
P2P-Nugache	82	YES	82	100%	0	0

Limitations?

Limitations

- Evading C-plane clustering
  - manipulate communication patterns
  - introduce random packets (noise) to reduce similarity between C&C flows

- Evading A-plane monitoring
  - stealthy activities (e.g, slow scanning/spamming)
  - undetectable activities (e.g., send spam using Gmail, download exe from HTTPS server)

Limitations

- Experimental setup
  - relies mainly on “simulated” bots
  - although the bot-code is real, the communication with the botmaster and the external world is “artificial” for some of the traces
  - It is very hard to get real-world C&C traces for many different types of botnets...

- C-plane clustering is hard to do well in real-networks...
  - lots of traffic
  - enterprise networks use web-proxies + egress filtering (alters flow statistics)

Future Work?

Future Work

- BotMiner++, for high-speed networks
  - more efficient clustering of C-flows

- Detecting botnets regardless of whether you can observe suspicious activities
  - is it possible?
  - can we do it in a reliable way?

Why was it accepted?

Why was it accepted?

- Well motivated
  - botnets are a big problem, more research was needed!
  - clearly states the limitations of previous works
  - first work on protocol- and structure-independent botnet detection (it also covers P2P botnets)
- Well written, it walks the reader through all the components of the system
  - backs formalism with intuition/motivation
- Promising experimental results
  - Low FPs, high DR