CLEANSE: Cross-Layer Large-Scale Efficient Analysis of Network Activities to SEcure the Internet

Wenke Lee

CLEANSE Project Background

- NSF CyberTrust "Large Team" award
  - Sept. 2008 thru Sept. 2012
- Team:
  - Georgia Tech: Lee, Ahamad, Feamster, Giffin, Huo
  - Michigan: Jahanian, Bailey
  - UNC Chapel Hill: Reiter, Monrose
  - Internet Systems Consortium (ISC): Vixie
  - SRI International: Porras, Yegneswaran

CLEANSE: Project Background (cont'd)

- Industry partners:
  - Arbor Networks
  - Castlecops
  - Comcast
  - Damballa
  - Google
  - IBM/ISS
  - IronPort
  - Outblaze
  - Time Warner Cable

CLEANSE: Motivations

- Attacks are now for economic and political gains
  - E.g., spam, phishing, identity/information theft, DDoS
  - Manipulate applications to victimize users
    - We call them *layer-8* attacks
- Conventional defenses are also at layer-8
  - Symptoms- and application-specific, malleable
- Layer-8 attacks are launched from botnets
  - We should focus on the behavior and infrastructure common to all attacks at the lower layers

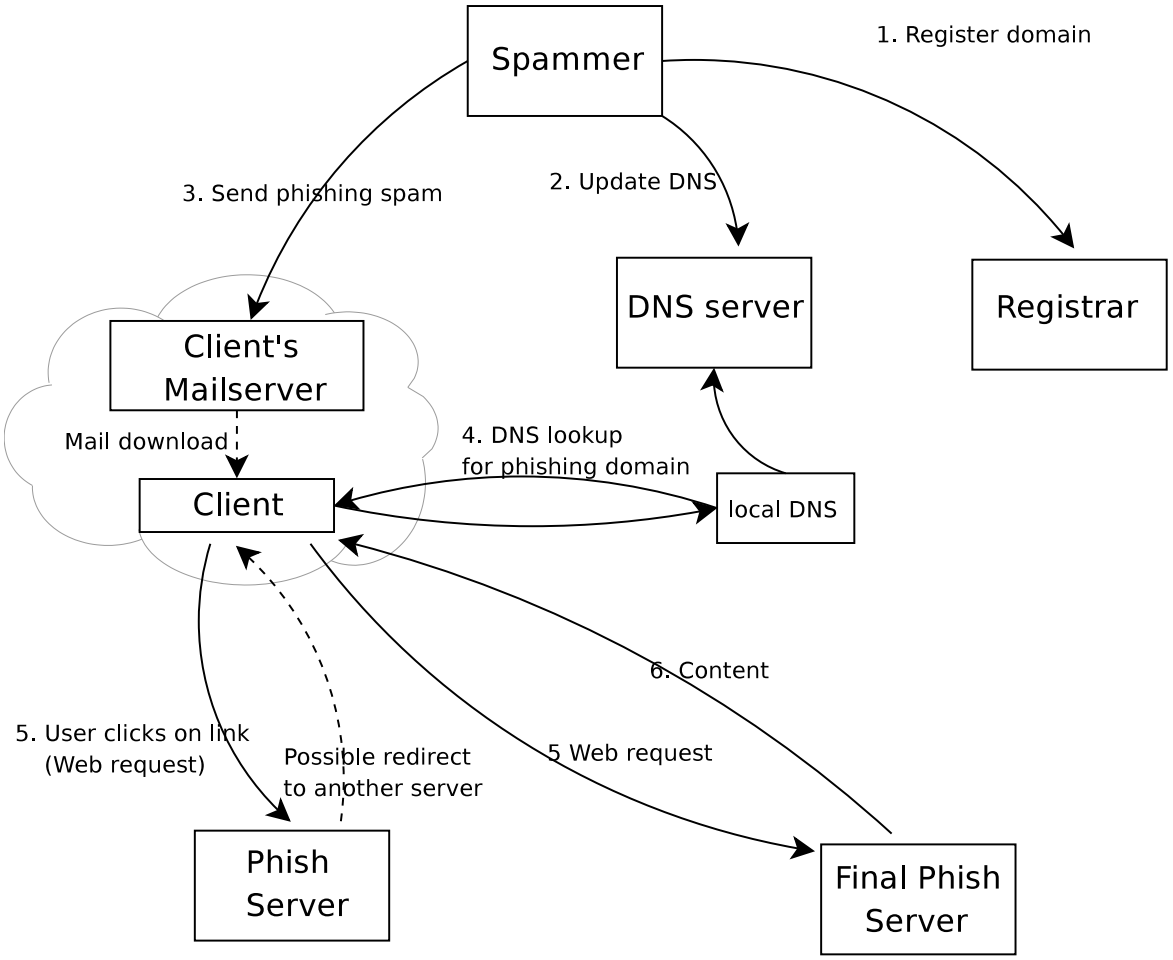
Botnets Rely on the Internet

- A botnet must use Internet protocols/services for efficiency, robustness, and stealth
  - Look-up services (e.g., DNS, P2P DHT)
    - Find C&C servers and/or peers
  - Hosting services (Web servers and proxies)
    - Storage and distribution/exchange of attack-related data
  - Transport (e.g., BGP)
    - Route (or hide) attack from bots to victims

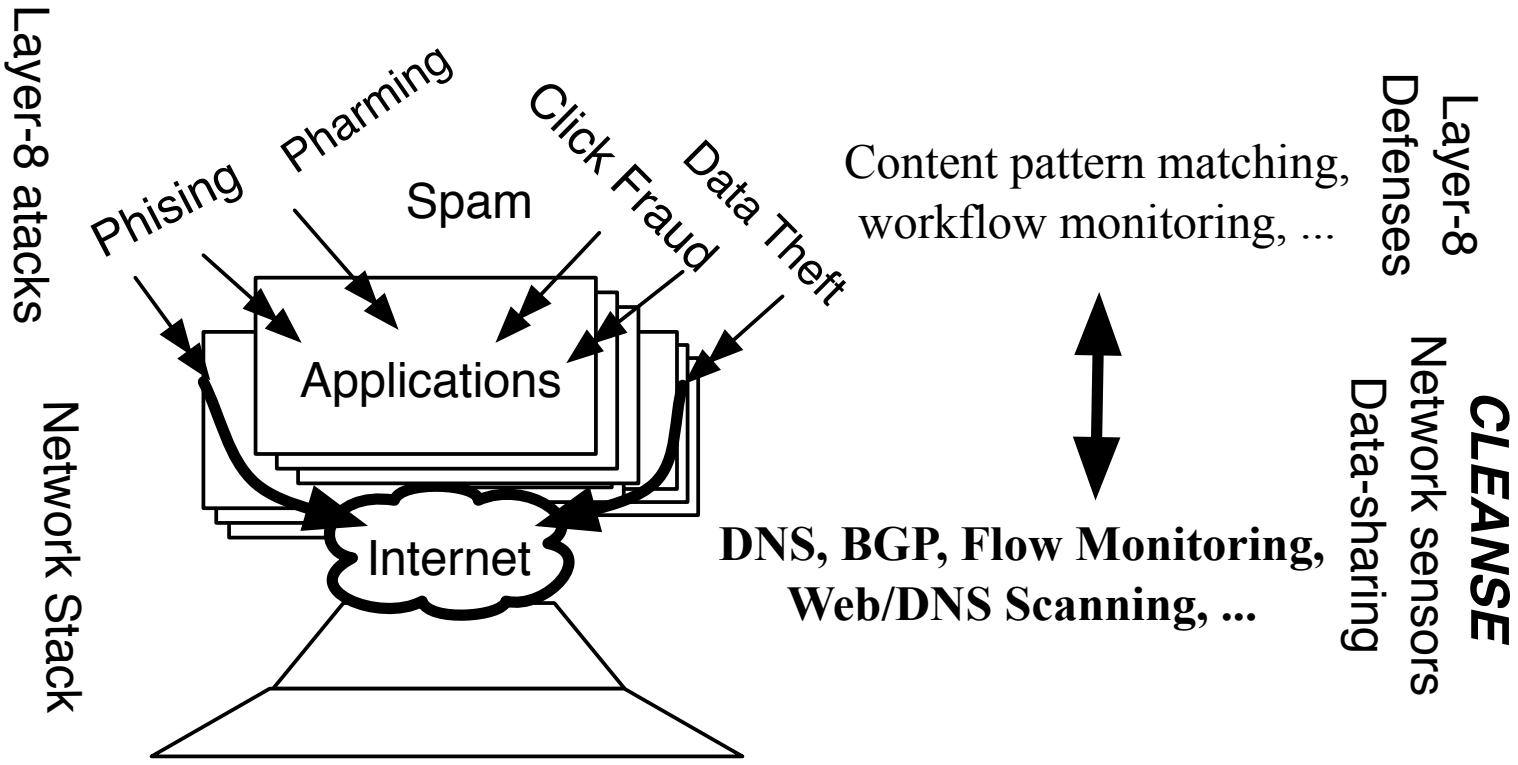
Botnets Rely on the Internet (cont'd)

- Layer-8 attacks by botnets produce observable service violations and anomalies
  - E.g., anomalous DNS look-ups
  - Bot means non-human; bot*net* means large-scale and coordinated
    - Large-scale and coordinated botnet activities are different from legitimate/normal human-generated activities

The Anatomy of a Phishing Attack



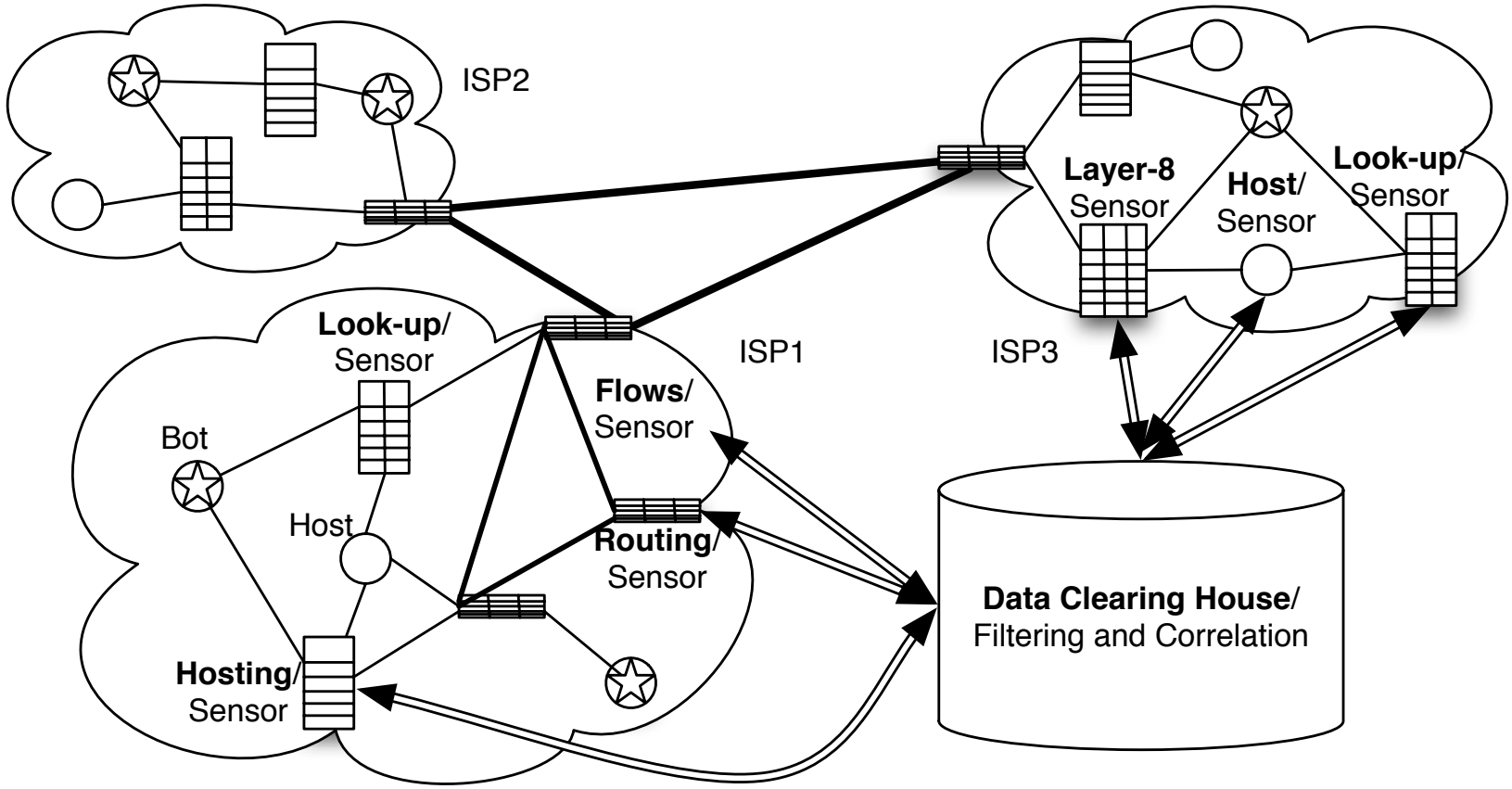
CLEANSE: Monitor Lower-Layers to Detect Layer-8 Attacks



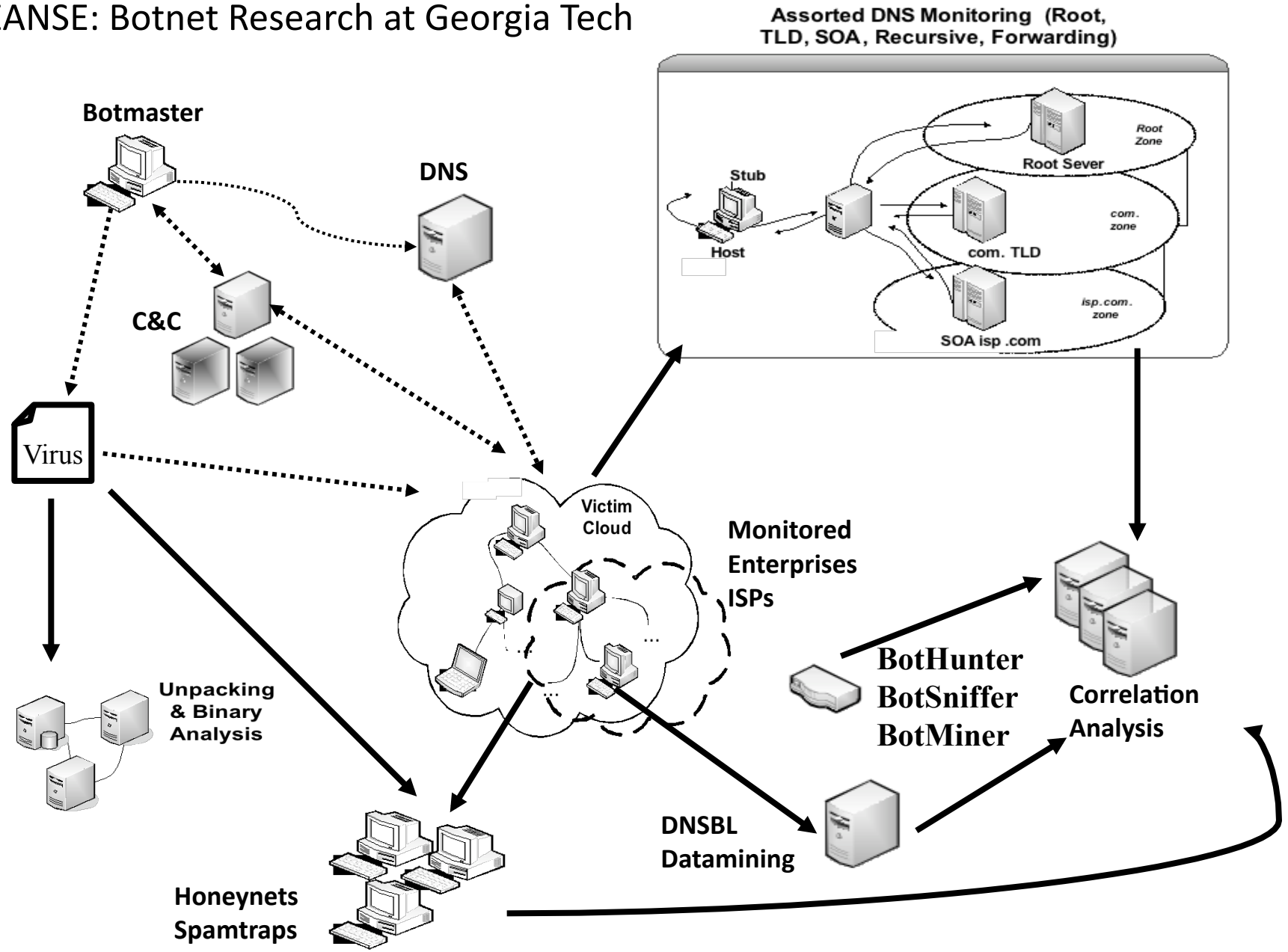
CLEANSE: Research Overview

- Control-plane monitoring
  - Anomaly detection algorithms for core networks services such as DNS and BGP
    - E.g., recursive and passive DNS monitoring, DNS cache inspection, rouge DNS scanners
- Data-plane monitoring
  - Flow-based anomaly detection algorithms
    - Traffic sampling and clustering
  - Host-based monitoring
    - Virtual machine monitoring, malware binary and script analysis
- Improved security auditing capabilities
  - On host and network/router
  - Achieve scalability while maintaining accuracy

CLEANSE: Sensors and Correlation System

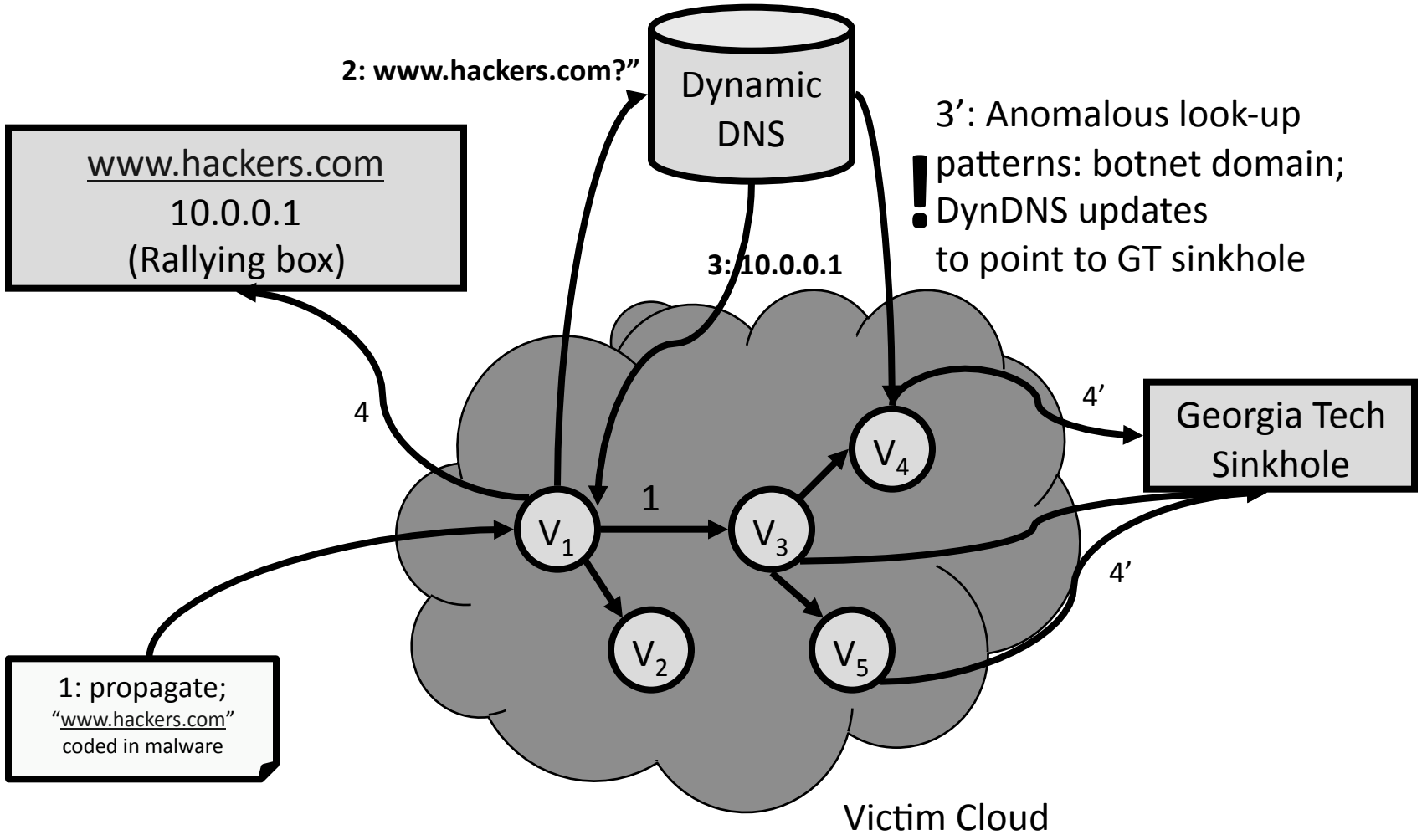


CLEANSE: Botnet Research at Georgia Tech



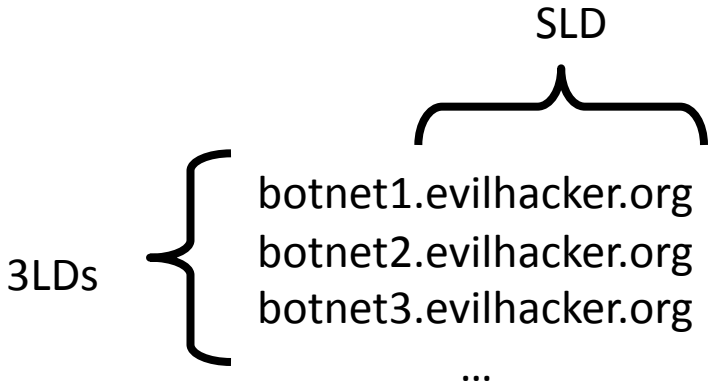
CLEANSE Research Highlights: Dynamic DNS and Recursive DNS Monitoring

Dynamic DNS Monitoring: Overview



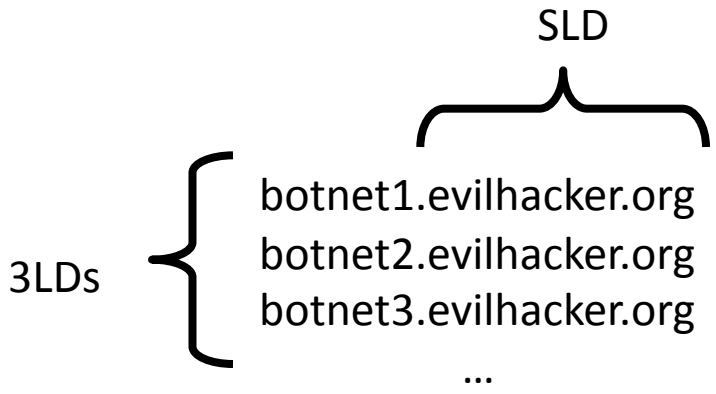
DDNS Use by Botnets

- Observation 1: hard-coded C&C domain (string)
  - Domain name purchases use traceable financial information. Multiple 3LDs can use DDNS service with one package deal
  - Thus: financial and stealthy motives for botnet authors to "reuse" SLD with numerous *similar/clustered* 3LDs

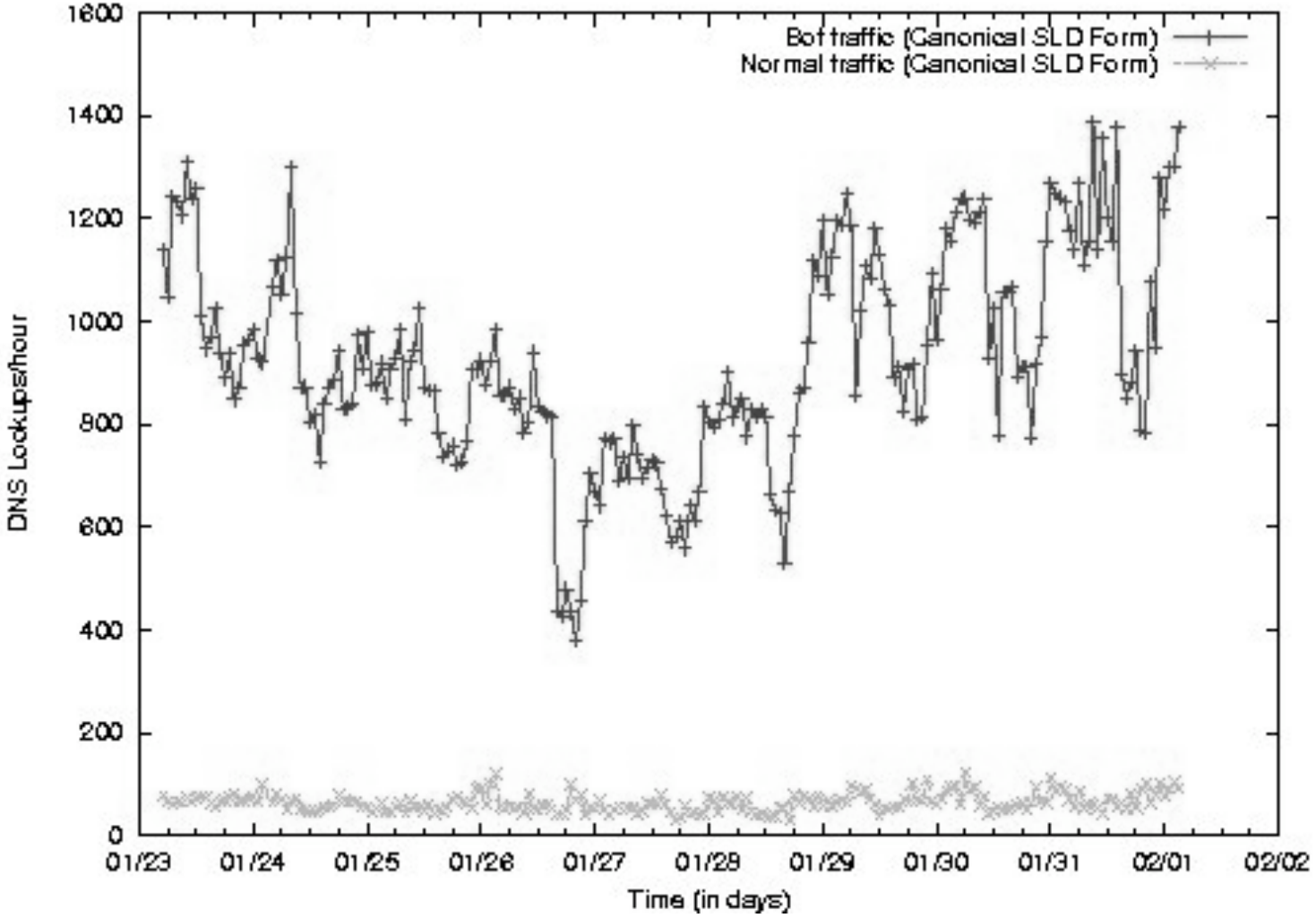


Clustered 3LD Look-ups

- Cluster the 3LDs under a SLD based on their similarities on names, and subnets of resolved IPs.
- Sum up the look-ups to all domains within a cluster



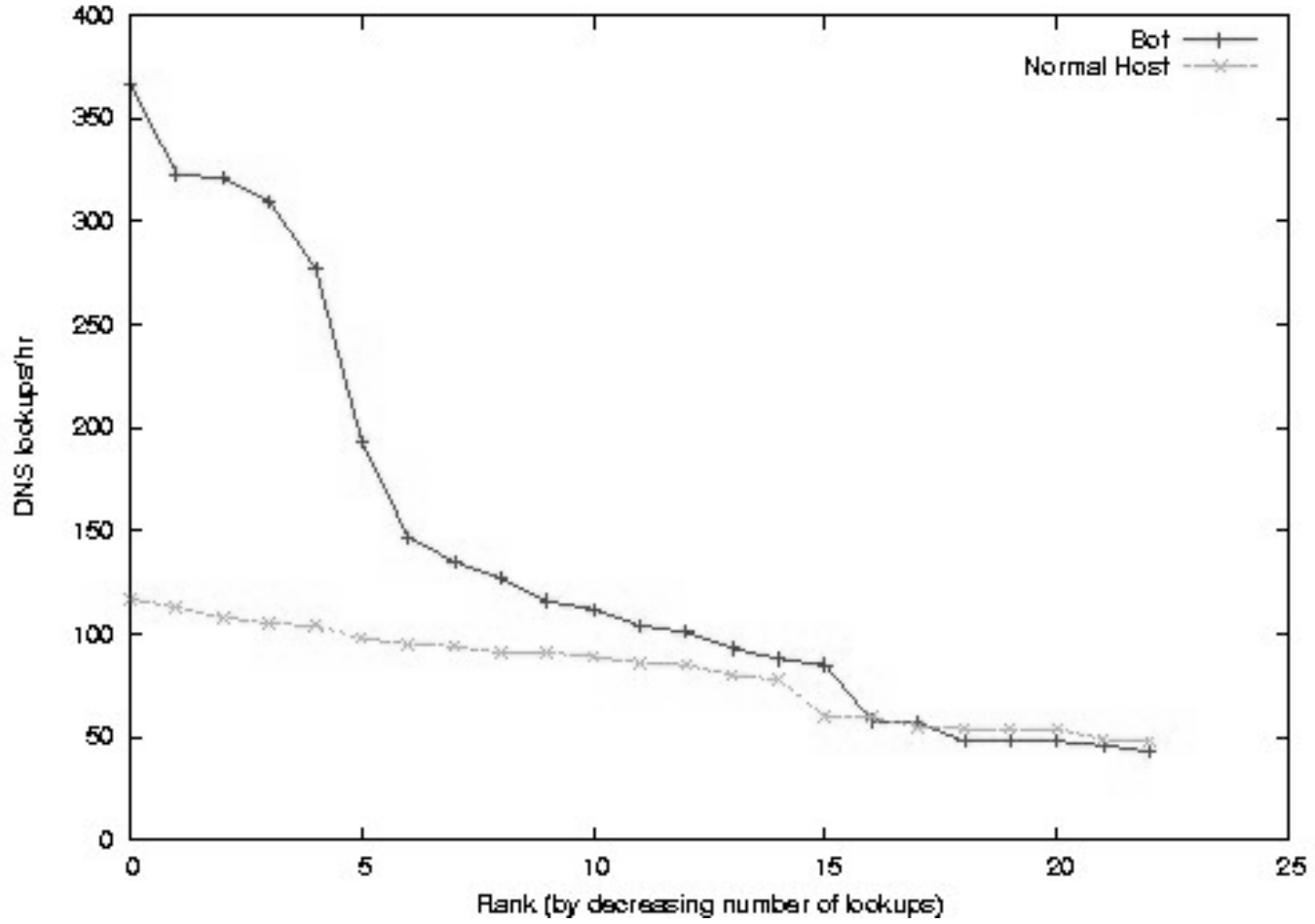
Clustered 3LD Look-ups (cont'd)



DNS Use by Botnets (cont'd)

- Observation 2: DNS look-up behavior of botnets
  - After boot, bots immediately resolve their C&C
    - Exponential arrival (spike) of bot DNS requests, because of time zones, 9 a.m./5 p.m. schedules, etc.
  - Normal DNS look-up behavior is a lot smoother
    - Human users don't all *immediately* check the same server right after boot

Look-up Arrival Rate (cont'd)



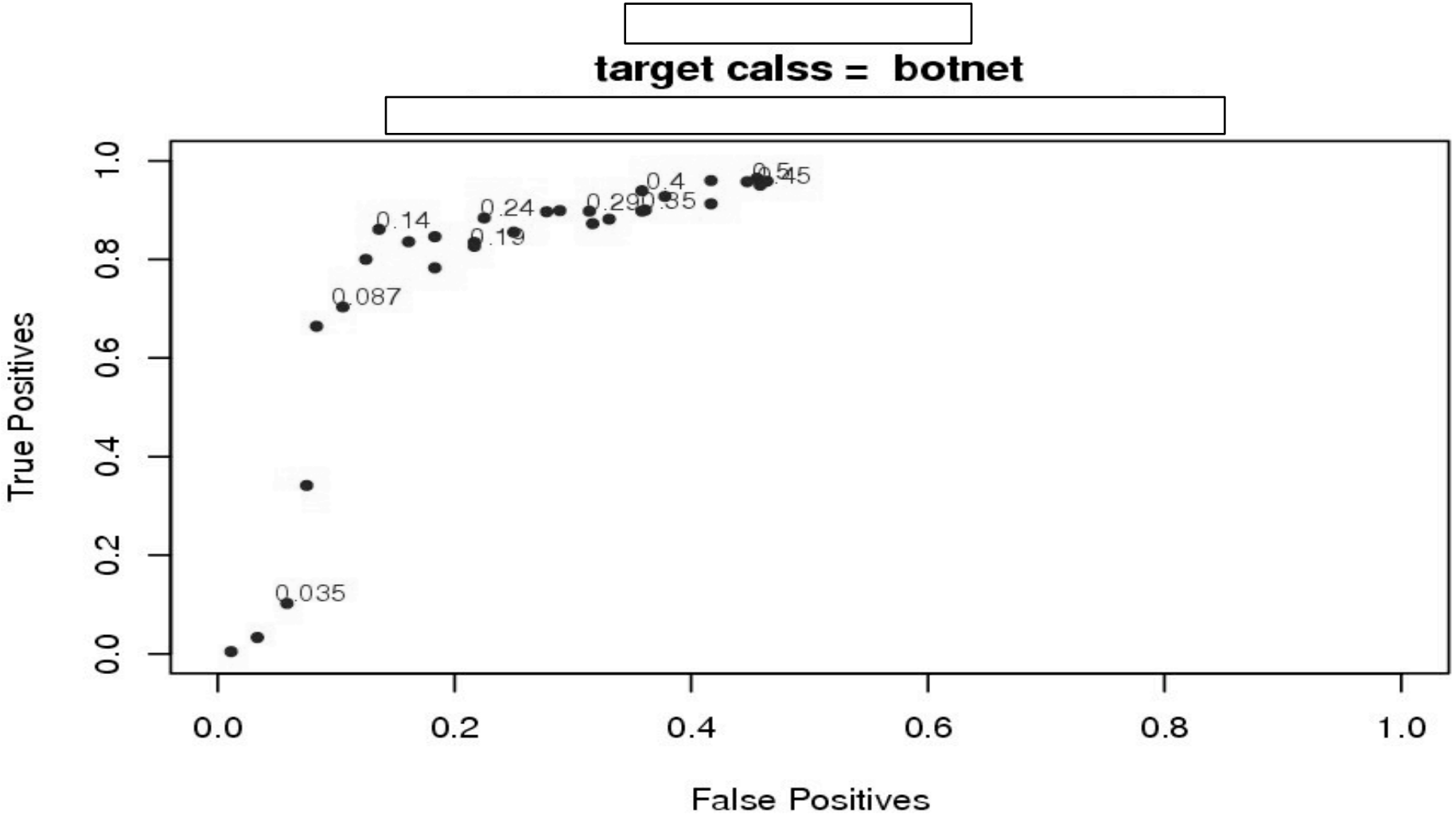
Other Observations/Features

- Fraction of TTL violations
  - E.g., more than one query to SOA within TTL
- Source IP dispersion in DNS look-ups
  - Local or global popularity of the domain
- Resolved IP dispersion
  - Distributed in many different networks?
- Number of times resolved IP changed
- …

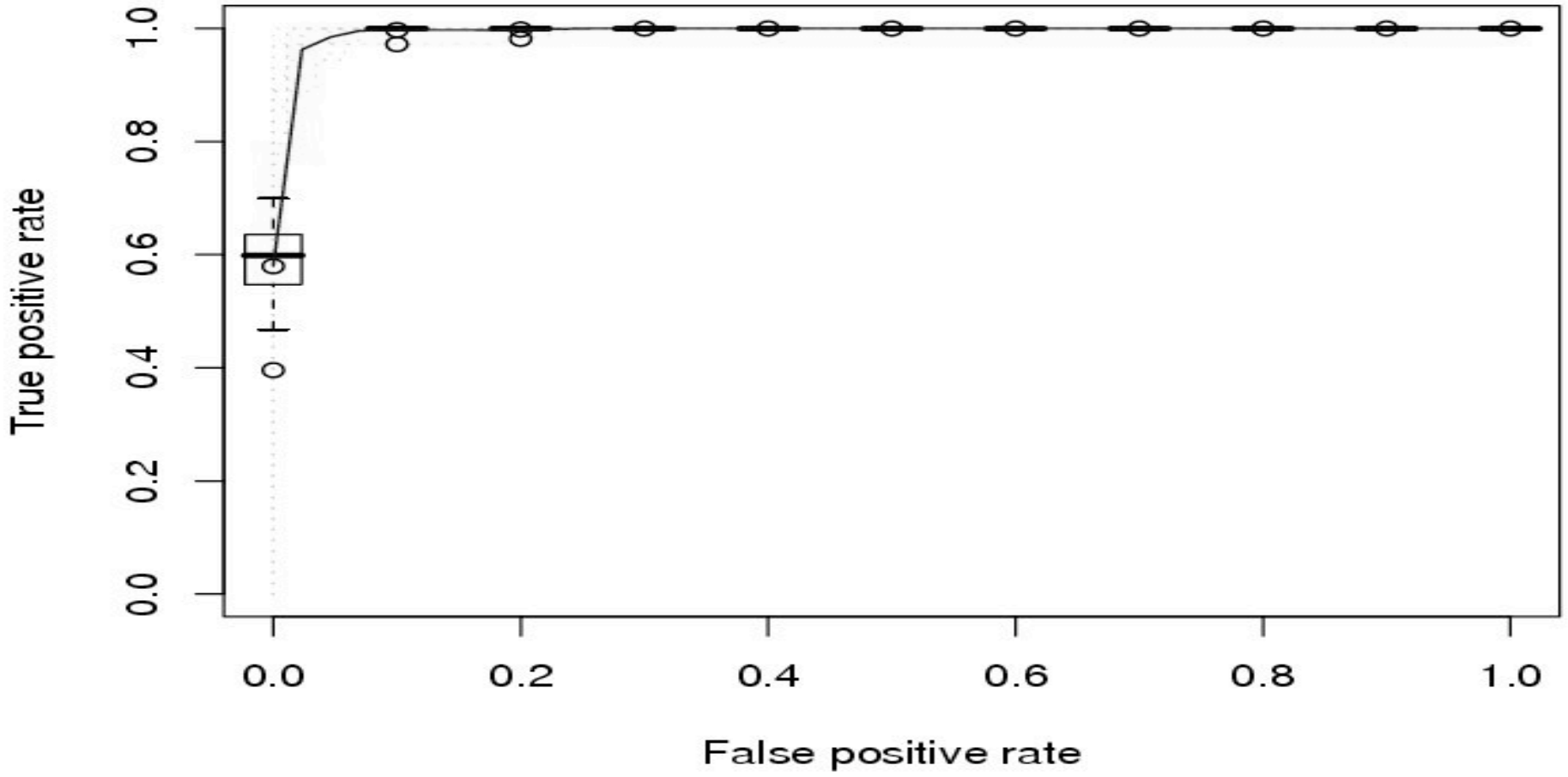
Classification

- A vector of statistical features, based on these observations, describes the look-up behavior of a DNS domain/cluser
- Use labeled training dataset to construct a statistical classifier
  - We obtained DNS query data from DDNS provider; hand identified (e.g., matched with malware analysis results, published lists, etc.) the many hundreds of botnets for ground truth to create labeled data

One-Class Classifier (cont'd)



Two-Class Classifier



RDNS Monitoring

- Analyze DNS traffic from internal hosts to a recursive DNS server(s) of the network
- Detect abnormal patterns/growth of "popularity" of a domain name
  - Identify botnet C&C domain and bots

RDNS Monitoring (cont'd)

- Common means of botnet propagation: (worm-like) exploit-based, email-based, and dry-by egg download
- Studies showed:
  - Exploit-based propagation: the number of infected machines grow exponentially in the initial phase
  - Email-based propagation: exponential or linear
  - (no known model for dry-by egg download)

Anomalous Domain Names

- Botnet-related domains usually contain random-looking (sub)strings
  - Many/most sensible domain names have been registered (for legitimate use)
  - In particular, botnet domain name 3LD often looks completely random, and the domain name tends to be very long
  - E.g. wbghid.1dumb.com, 00b24yqc.ac84562.com

Popularity Growth of the Suspicious Names

- Monitor for “new and suspicious” domain names that enjoy exponential or linear growth of interests/ look-ups
  - Train a Bloom filter for N days to record domain names being looked-up, and a Markov model of all the domain name strings
    - On the N+1 day, consider a domain “new” if it is not in the Bloom filter; and if it does not fit the Markov model, it is also “suspicious”
  - Treat the sequence of look-ups to each new and suspicious domain (on the N+1 day) as a time series
  - Apply linear and exponential regression techniques to analyze the growth of number of look-ups

RDNS Monitoring (cont'd)

- One month (2007) in a large ISP network
- ~1,500 botnet domain names
- 11% of computers on the network looked-up/ connected to these domains
  - Bots!

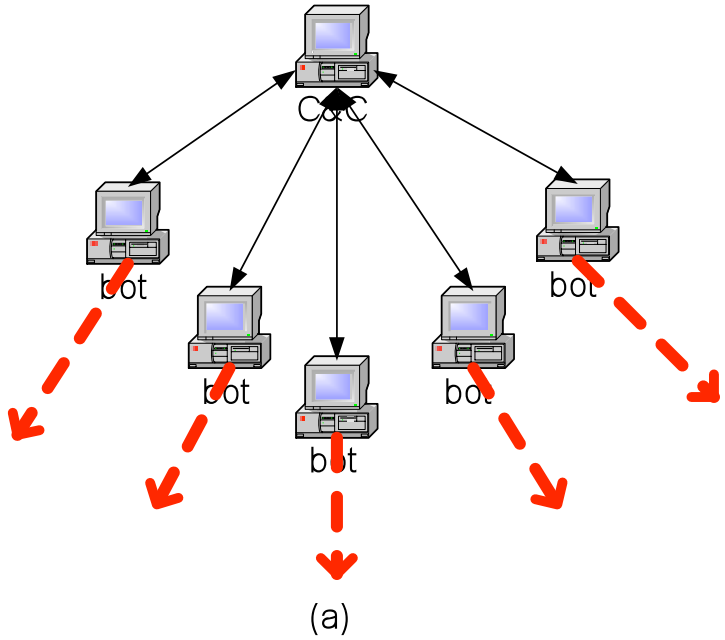
DNS Monitoring: Dealing with False Positives

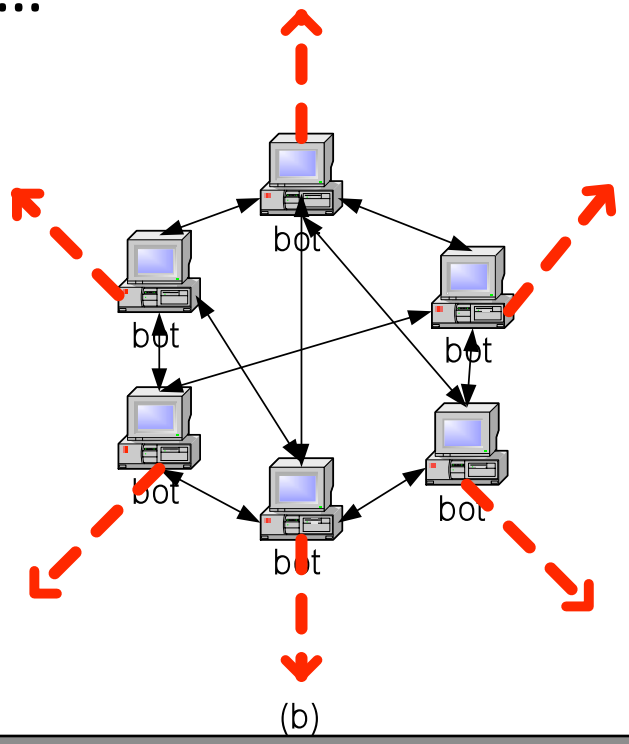
- Reviews by "abuse" analysts (ISPs)
- White-listing of known legitimate domains
- "Suspicious" domains as input to other network sensors, e.g., BotHunter
  - Analyze traffic from hosts that connect to the domains
    - Connections/activities that suggest C&C dialogues, scans, spam run, etc.

CLEANSE Research Highlights: BotMiner

Why BotMiner?

- Botnets can change their C&C content (encryption, etc.), protocols (IRC, HTTP, etc.), structures (P2P, etc.), C&C servers, infection models …



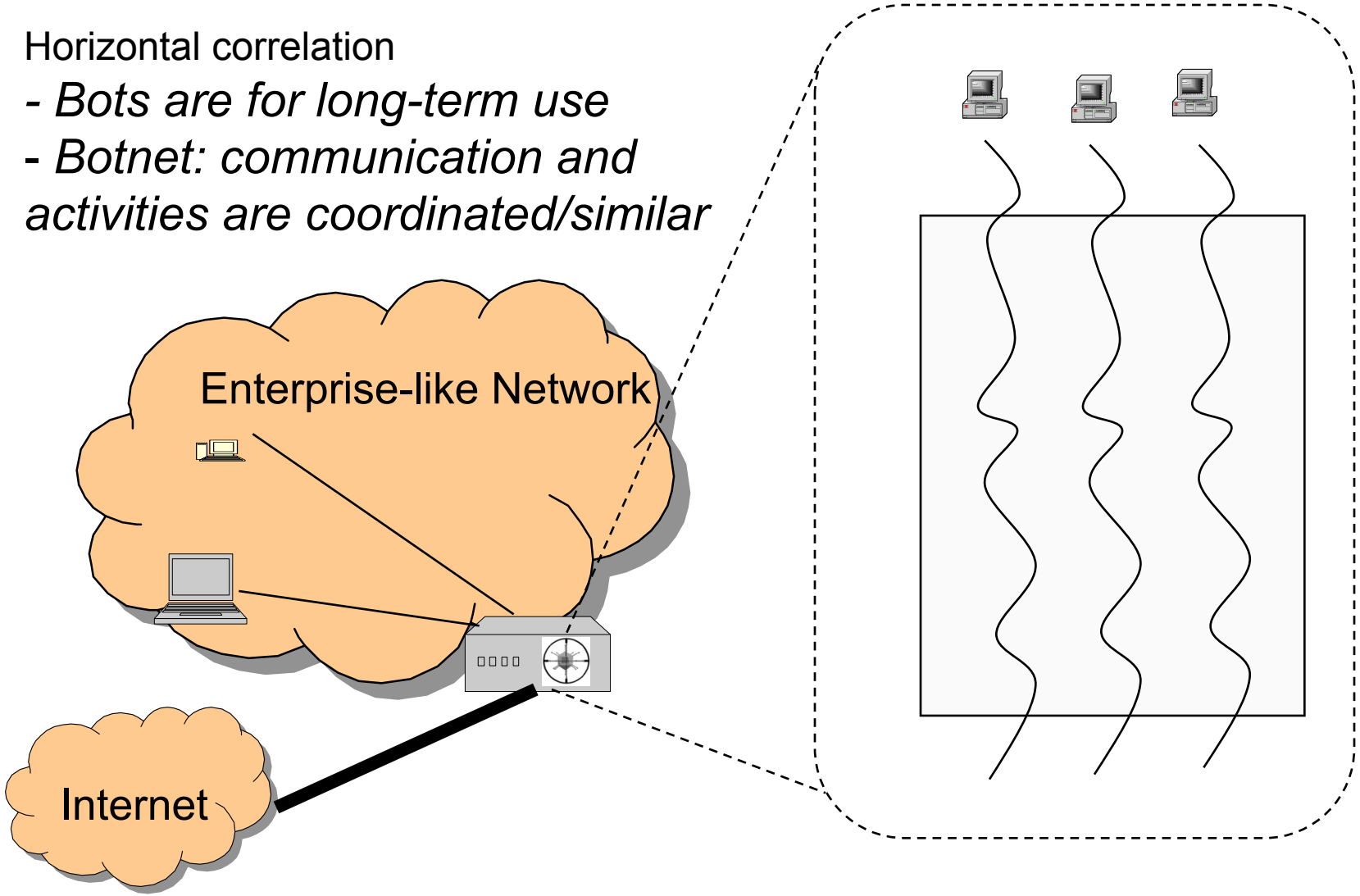


Example: Nugache, Storm, …

BotMiner: Protocol- and Structure-Independent Detection

Horizontal correlation

- *Bots are for long-term use*
- *Botnet: communication and activities are coordinated/similar*

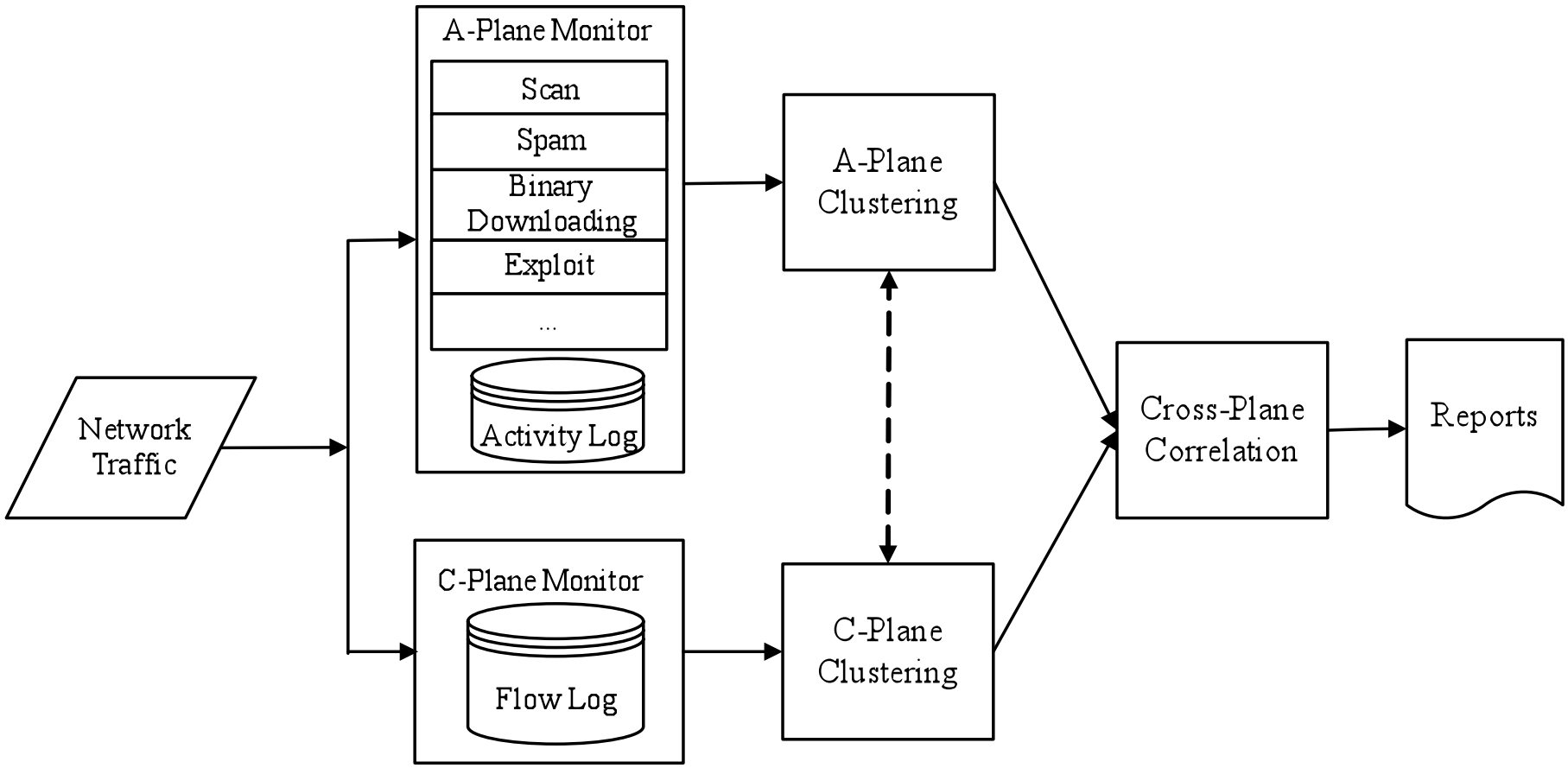


A Definition of a Botnet

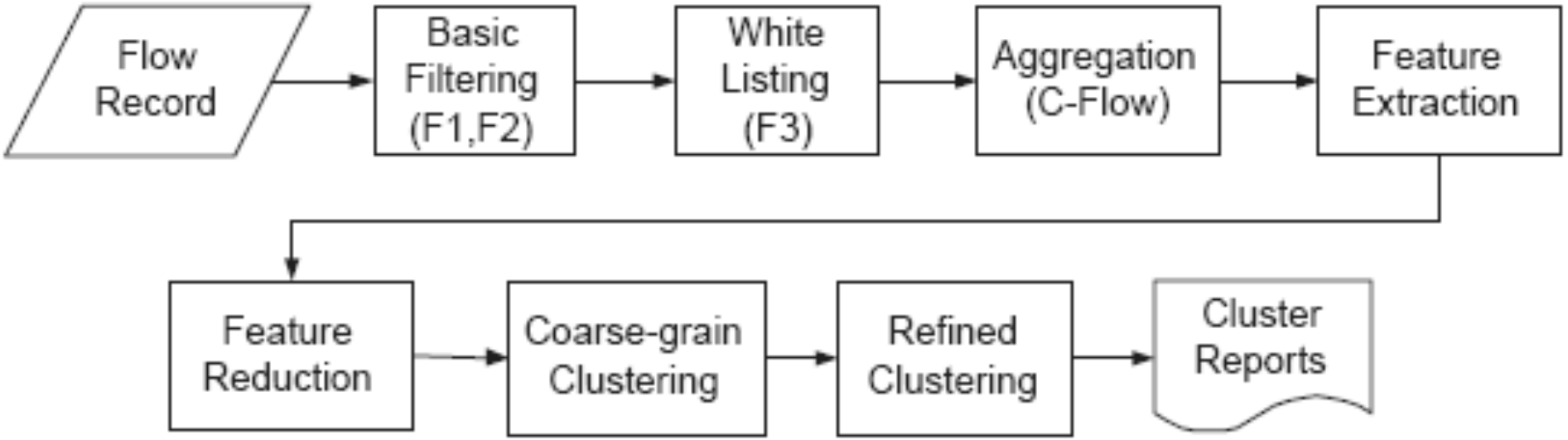
- "A coordinated group of malware instances that are controlled by a botmaster via some C&C channel"

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

BotMiner Architecture



BotMiner C-plane Clustering



- What characterizes a communication flow (C-flow) between a local host and a remote service?
  - <protocol, srcIP, dstIP, dstPort>

How to Capture "Talking in What Kind of Patterns"?

- Temporal related statistical distribution information in
  - BPS (bytes per second)
  - FPH (flow per hour)

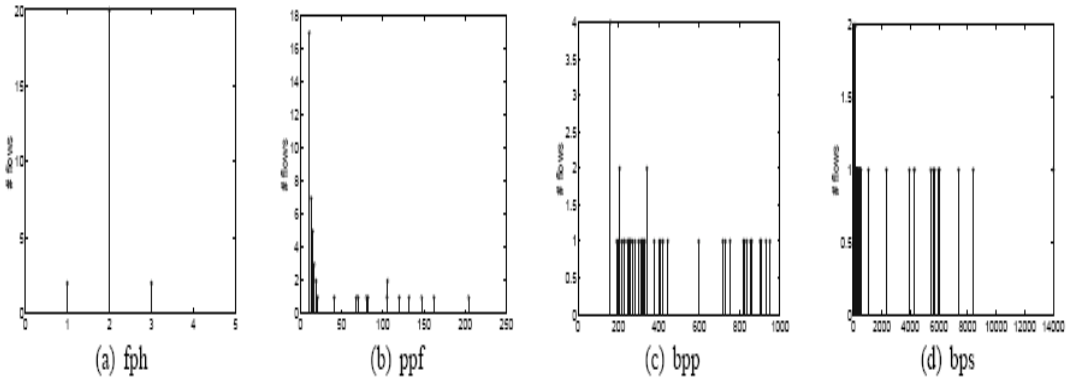


Figure 4: Visit pattern (shown in distribution) to Google from a randomly chosen normal client.

- Spatial related statistical distribution information in
  - BPP (bytes per packet)
  - PPF (packet per flow)

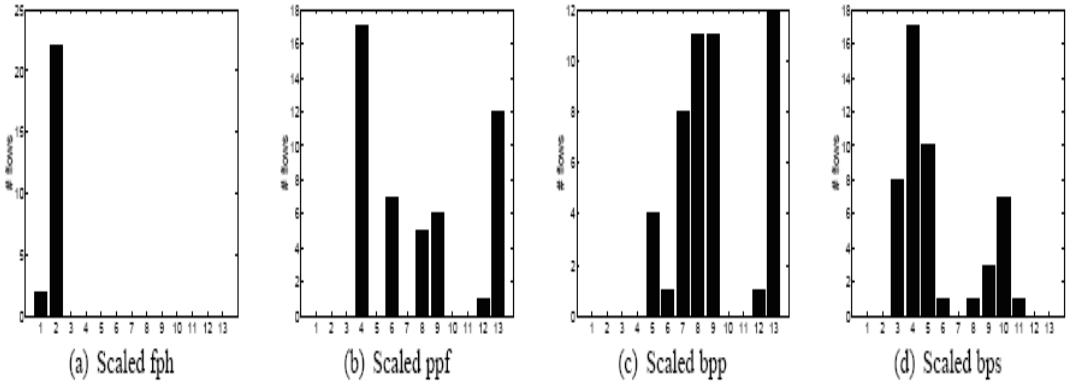
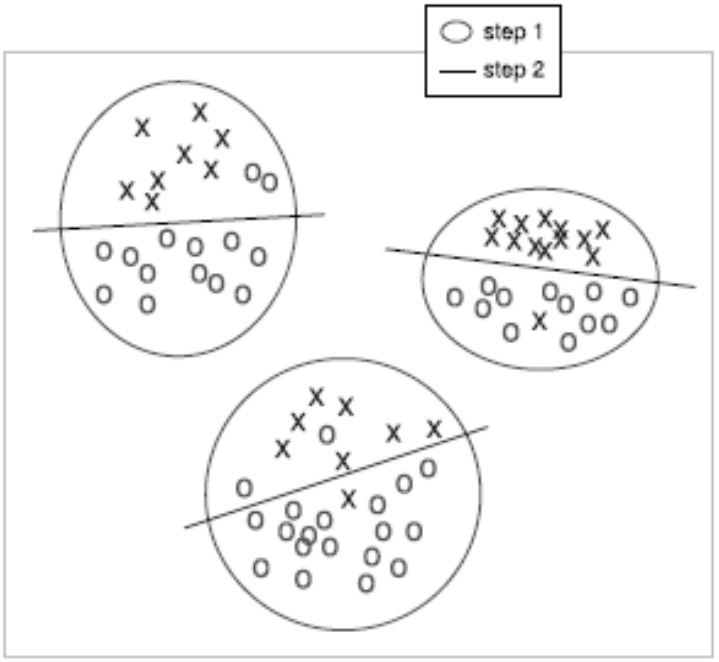


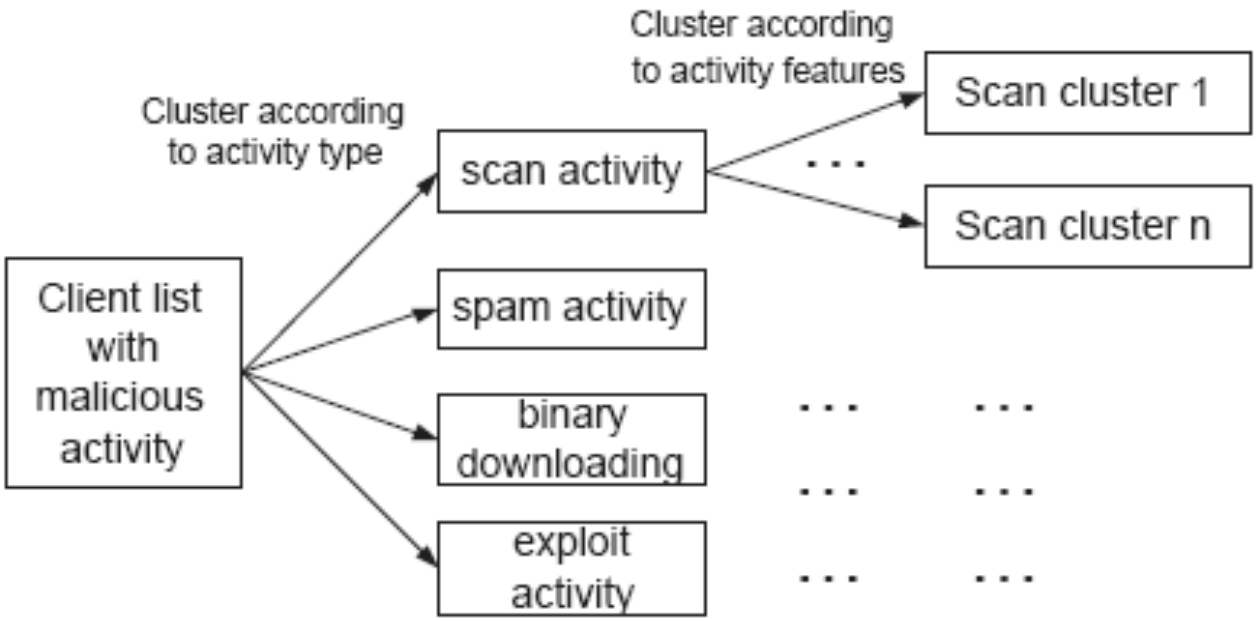
Figure 5: Scaled visit pattern (shown in distribution) to Google for the same client in Figure 4.

Two-step Clustering of C-flows

- Efficiency
- Two steps:
  - Coarse-grained clustering
    - Using reduced feature space: mean and variance of the distribution of FPH, PPF, BPP, BPS for each C-flow (2*4=8)
    - Efficient clustering algorithm: X-means
  - Fine-grained clustering
    - Using full feature space (13*4=52)



A-plane Clustering

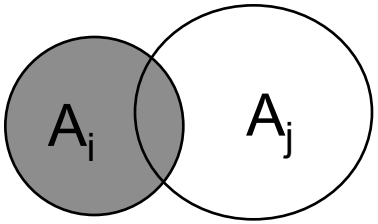


- Capture “activities in what kind of patterns”

Cross-plane Correlation

- Botnet score *s(h)* for every host *h*

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

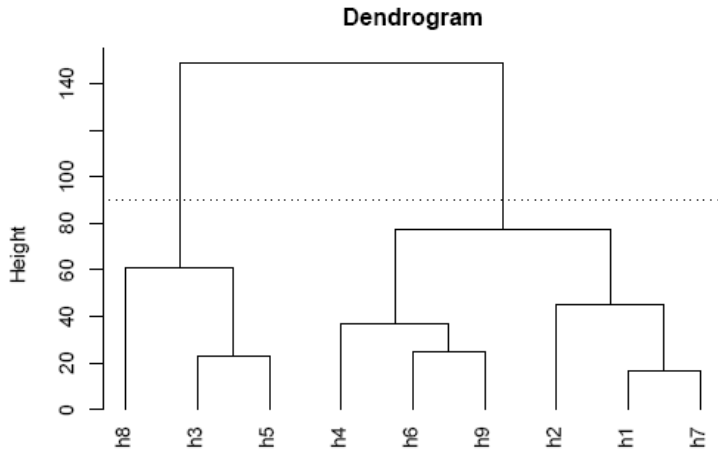


- Similarity score between host h_i and h_j

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

Two hosts in the same A-clusters and in at least one common C-cluster are clustered together

- Hierarchical clustering



Evaluation Traces

Trace	Pkts	Flows	Filtered by F1	Filtered by F2	Filtered by F3	Flows after filtering	C-flows (TCP/UDP)
Day-1	5,178,375,514	23,407,743	20,727,588	939,723	40,257	1,700,175	66,981 / 132,333
Day-2	7,131,674,165	29,632,407	27,861,853	533,666	25,758	1,211,130	34,691 / 96,261
Day-3	9,701,255,613	30,192,645	28,491,442	513,164	24,329	1,163,710	39,744 / 94,081
Day-4	14,713,667,172	35,590,583	33,434,985	600,901	33,958	1,520,739	73,021 / 167,146
Day-5	11,177,174,133	56,235,380	52,795,168	1,323,475	40,016	2,076,721	57,664 / 167,175
Day-6	9,950,803,423	75,037,684	71,397,138	1,464,571	51,931	2,124,044	59,383 / 176,210
Day-7	10,039,871,506	109,549,192	105,530,316	1,614,158	56,688	2,348,030	55,023 / 150,211
Day-8	11,174,937,812	96,364,123	92,413,010	1,578,215	60,768	2,312,130	56,246 / 179,838
Day-9	9,504,436,063	62,550,060	56,516,281	3,163,645	30,581	2,839,553	25,557 / 164,986
Day-10	11,071,701,564	83,433,368	77,601,188	2,964,948	27,837	2,839,395	25,436 / 154,294

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

Evaluation Results: False Positives

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

Evaluation Results: Detection Rate

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

Conclusion

- CLEANSE aims to develop a detection framework to secure the Internet against *large-scale* and *coordinated* layer-8 attacks by botnets and other/future forms of compromises
  - Identify the basic network services necessary for large-scale attacks
  - Develop new analysis and detection algorithms to monitor these services
  - Collaborate with government and industry