Research in Botnet Detection and Malware Analysis

Wenke Lee

College of Computing

Georgia Institute of Technology

Botnets

Individual Machines Used to Be
Targets ---
Now They Are Resources

- Bot (Zombie)
  - Software Controlling a Computer Without Owner Consent
  - Professionally Written; Self-propagating; 10% of Internet
- Bot Armies (Botnets)
  - Networks of Bots Controlled by Criminals
  - Key Platform for Fraud and other For-Profit Exploits

Botnet Epidemic

- More Than 90% of All Spam
- All Denial of Service (DDOS) Attacks
- Clickfraud
- Phishing & Pharming Attacks
- Key Logging & Data/Identity Theft
- Key/Password Cracking
- Anonymized Terrorist & Criminal Communication

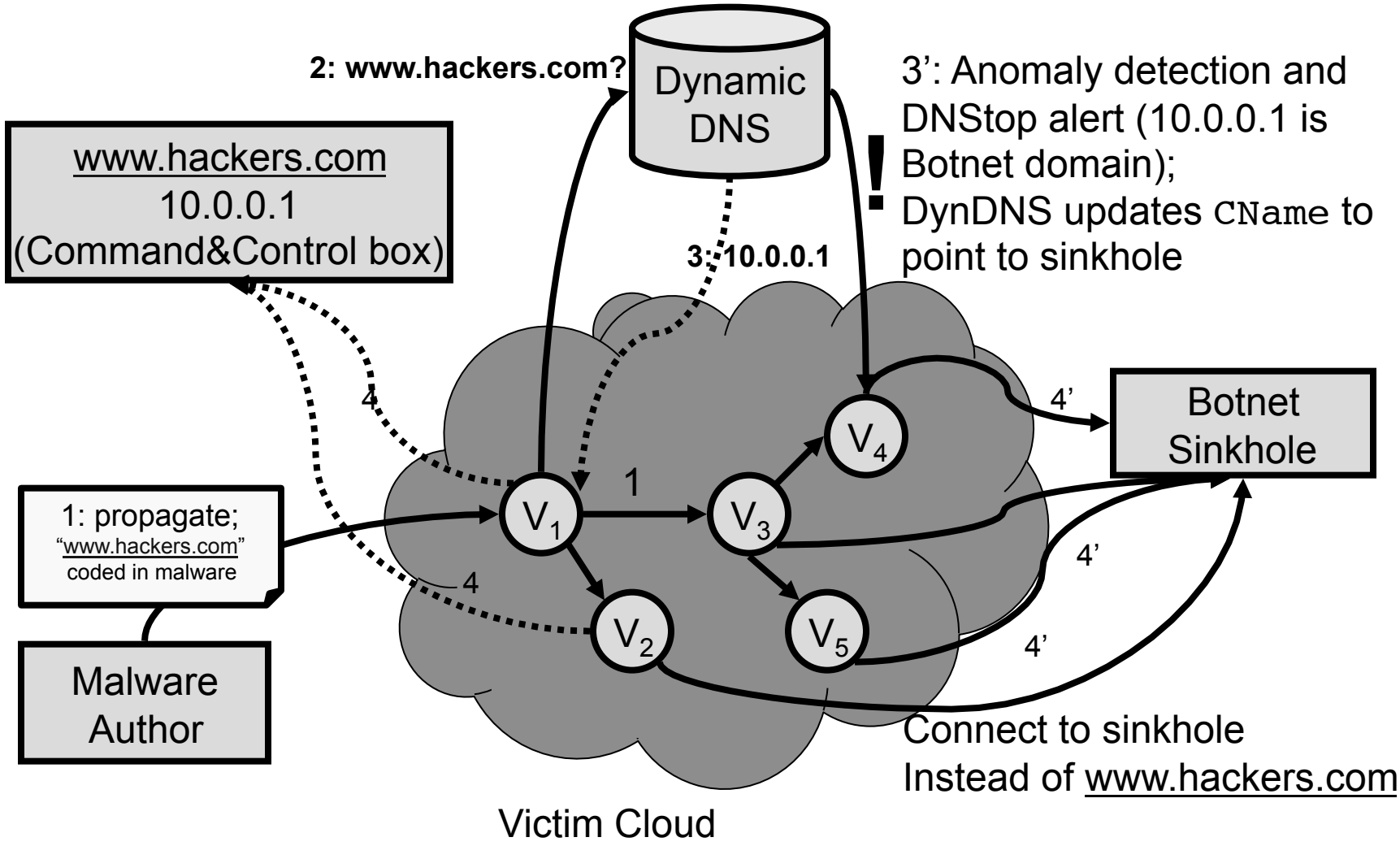
Example: Bots as Targeted Spyware

- Sub-sample of Aerospace Bots
  - Total: 272 bots
  - 32.35%: Communication Center, China Aerospace
  - 10.66%: National Aeronautics and Space Association
  - 5.88%: PARQUE DE MATERIAL AERONAUTICO DE LAGOA SANTA
  - 5.51%: Scientific Research Department of China Aerospace
  - 5.15%: No. 1 Institute of China Aerospace Corporation
  - 4.78%: Marketing Department of China Aerospace Fifth Academy (Ministry of Defense)
  - 4.78%: Communication Station of China Aerospace Seventh A
  - 4.04%: Communication Station of China Aerospace Fifth Academy
  - …

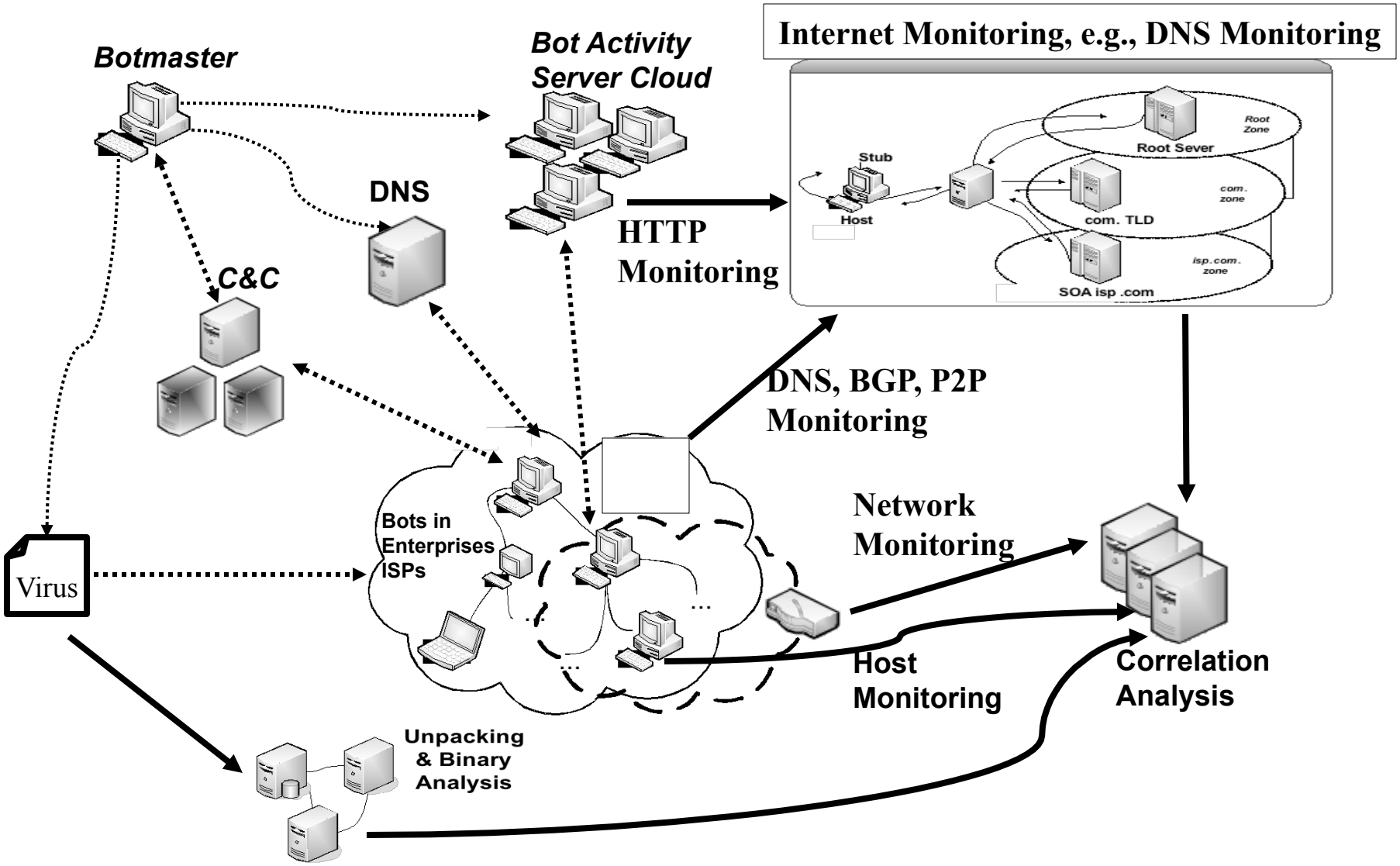
Outline

- Overview
- Recursive DNS monitoring
- Expanding and scaling up network analysis
- Analysis of network properties of KR botnet

Example: KarstNet at Georgia Tech



Research in Botnet Detection and Removal



Need Multifaceted Approach

- For example, to protect an enterprise network, we need a network appliance that uses information from:
  - Sensors on *Internet services* (e.g., DNS)
    - Servers and patterns in botnet communication
  - *Malware behavior analysis* engines
    - Communication and fraud activity patterns
  - *Flow-based anomaly detection* modules
    - Coordinated, non-human-initiated traffic

Recursive DNS Monitoring

RDNS Monitoring to Detect C&C Domains and Bots

- 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 yet)

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 (users can't type but bots don't type!)
  - 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

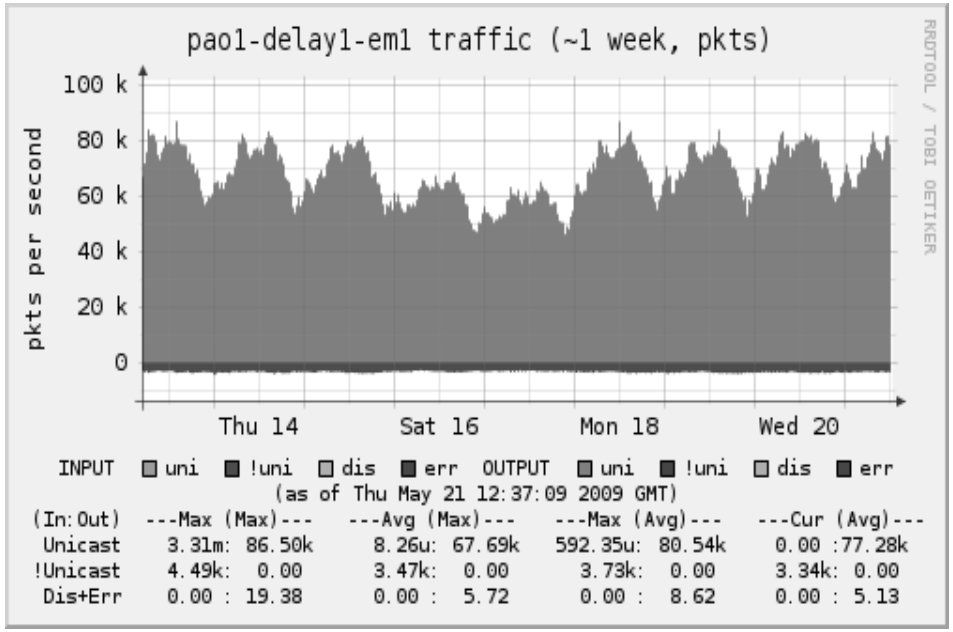
RDNS Monitoring (cont'd)

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

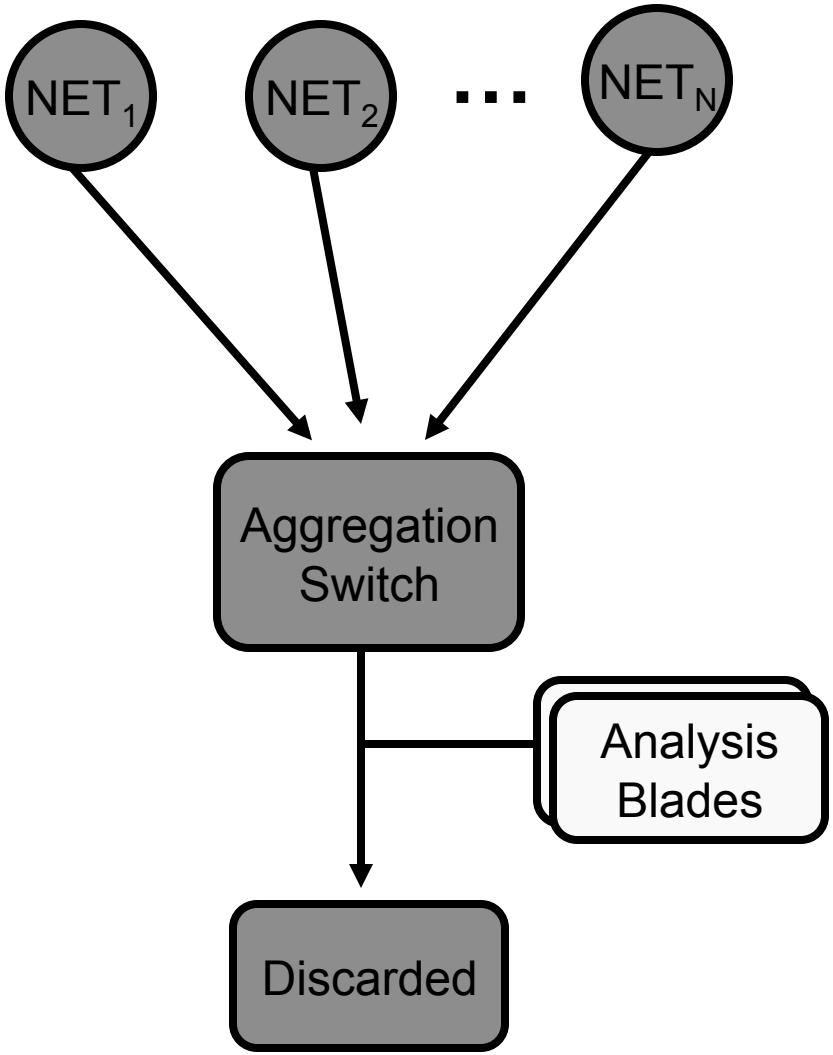
Expanding and Scaling up Network Analysis

SIE

- Security Information Exchange
- Numerous ISP, transit and educational sensor pool local data
  - Over 100MB/s of traffic
- Pooled and replayed on local analysis networks
  - Allows for real-time inspection by security analysts
  - Fine-grained control over replay allows data source to preserve and enforce policy restrictions



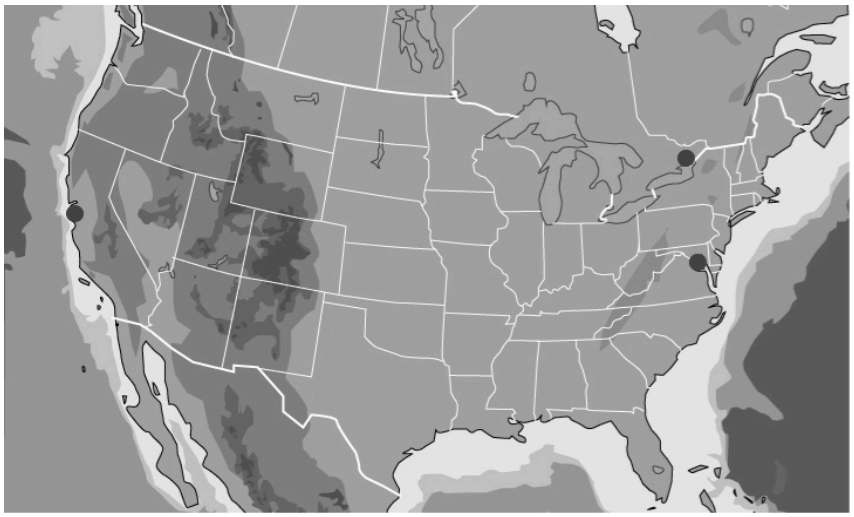
SIE Conceptual Overview



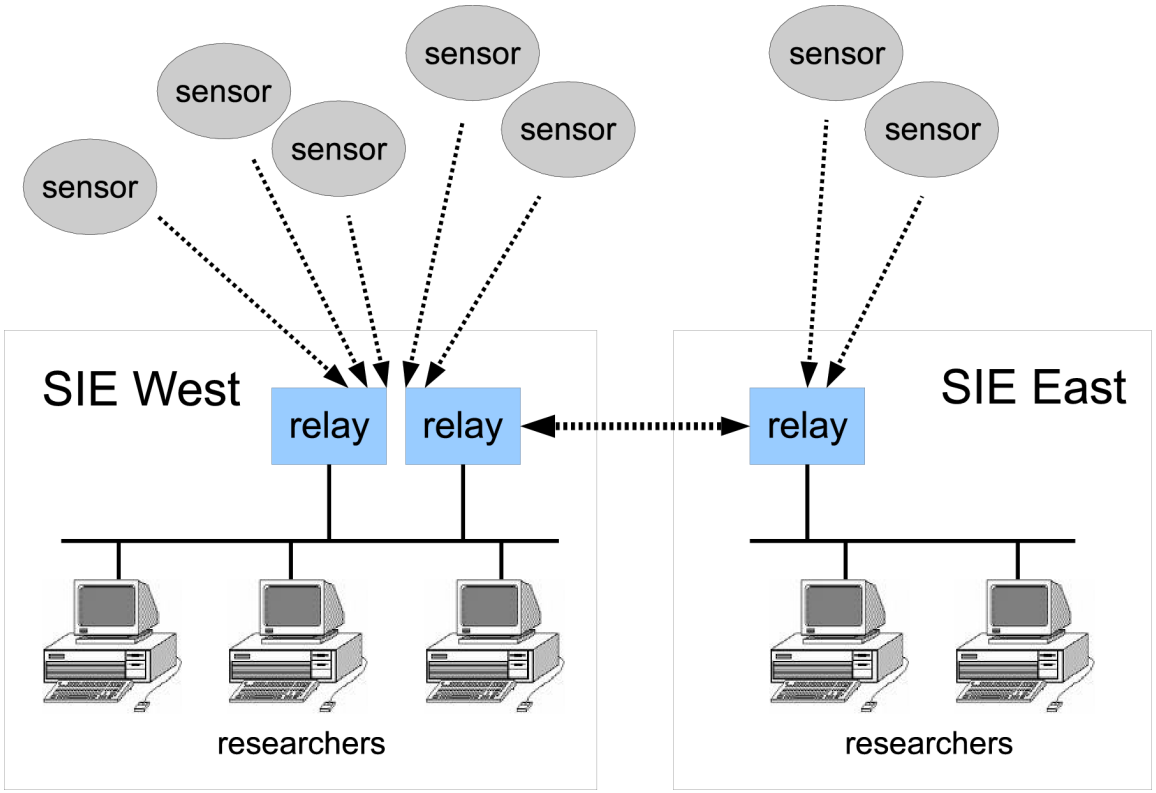
- Passive DNS and other data collected from numerous ISP, transit and academic networks
- Data rebroadcast on numerous aggregation switches, and discarded
- Blades witness traffic and output analysis

SIE Replay Switches

- Three broadcast switches:
  - Palo Alto (in production)
  - Washington DC (pending equipment arrival)
  - Ottawa (in discussion)
- A fourth at ISC
  - Used for development testing
  - Soon, traffic may outgrow pilot capacity
- Data source provide adequate coverage of N. American continent



Data Distribution Model



Real-time broadcast ensures that multiple replay switches see identical traffic

Diverse geographic analysis centers allows for choice of power, colo, transit for analysis nodes

Example: Spam Channel (ch25)

- Bots may used spam to propagate
  - Analysis of SIE's spam channel used for detection
- Preprocessing packetizes into envelope, headers, URLs (python scripts)
- Spam types:
  - spam traps
  - "this is spam" reports/submissions
  - spamassassin-scored email
- Good starting point for analysis
  - Malware, phishing, bots

isc/email.proto

```
package nmsg.isc;

enum EmailType {
    unknown = 0;
    spamtrap = 1;       // email sent to a spamtrap
    rej_network = 2;    // rejected by network or SMTP (pre-DATA) checks
    rej_content = 3;    // rejected by content filter (including domain blacklists)
    rej_user = 4;       // classified by user as spam
}

message Email {
    optional EmailType  type = 8;
    optional bytes      headers = 2;    // SMTP headers
    optional bytes      srcip = 3;      // remote client IP
    optional bytes      srchost = 4;    // remote client PTR, if known
    optional bytes      helo = 5;       // HELO/EHLO parameter
    optional bytes      from = 6;       // MAIL FROM parameter (brackets stripped)
    repeated bytes      rcpt = 7;       // RCPT TO parameter(s) (brackets stripped)
    repeated bytes      bodyurl = 9;    // URL(s) found in decoded body
}
```

Example: Spam Channel

- The isc/email.proto is an nmsg format defined for the purposes of spam analysis
  - Used to track bots/botnets and associated URLs
- Key design points
  - One merely identifies the useful components of spam sensor data (date, srcIP, body URLs, etc.)
  - The sensors present a real-time view of these tuples
- In contrast, other sharing mechanism are inadequate for botnet detection
  - Sharing complete message mboxes is slow (batch-based)
  - Sharing DNSBL zone abstractions loses data (IP/date only)

How to Get Involved

- Contact:
  - info@sie.isc.org
- Tools available:
  - https://sie.isc.org/
- Network operators are urged:
  - Become involved in SIE, as a sensor or to analyzed data
  - Run your own local SIE system, if policy restrictions apply to your data

Analysis of Network Properties of the Korean Botnet

Network Properties of KR Botnet

- What can one see from the network about the Korean botnet attack of July 2009?
- First order information trivially identified:
  - Location of attacking hosts, ASN, etc

Geographic Properties

- Most victims participating in DDoS located in South Korea

PCt Country Code

96.67 KR
1.2109 US
0.504541 JP
0.403633 CN
0.403633 UNKWN
0.201816 DE
0.100908 TH
0.100908 NL
0.100908 IT
0.100908 HU
0.100908 FI
0.100908 EU

Geographic Properties

- Normally, victims are located is highly diverse countries
- A localized infected population suggests specific properties about the infection vector
  - E.g., a language-specific element may be involved
  - Host-based analysis may later confirm this, but at the zero-hour, we infer this much from the network properties of malware

Geographic Properties

- Geographic details can also assist in obtaining a binary sample, if local networks can assist in this
- Victim Geo Information also assists in remediation, if a network signature can be generated (e.g., port behavior)
- A sampling of botnet victims demonstrated:

Percent	Organization
42.7851	HANARO-AS Hanaro Telecom Inc.
26.1352	KRNIC-ASBLOCK-AP KRNIC
2.11907	FCABLE-AS Qrix, Inc.
1.71544	HANVITIAB-AS-KR Hanvit I&B
1.41271	DREAMPLUS-AS-KR DreamcityMedia
1.31181	VITSSEN-AS-KR TBROAD ABC BROADCASTING CO.,LTD.
1.31181	GINAMHANVIT-AS-KR hanvit ginam broadcasting comm.

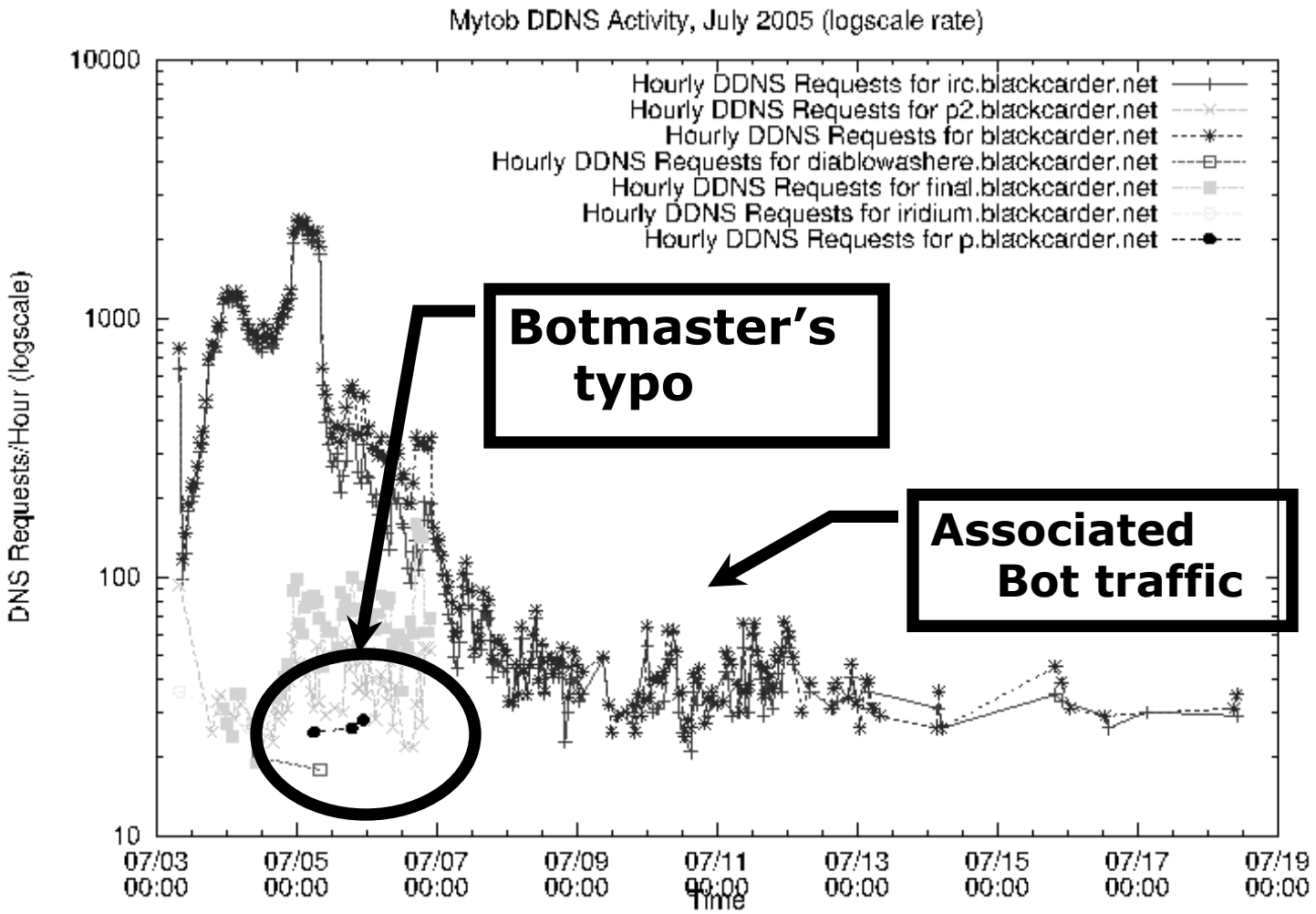
DNS Properties

- In some cases, the DNS resolution behavior of attacking bots can be used to identify origins
  - But do all bots use DNS? In ShadowServer's 2-year study of 18M samples shows almost all samples used DNS
    - Exceptions would be P2P botnets

DNS Properties (Example)

- Authority DNS monitoring can, in some cases, yield actionable information
- E.g., the early resolution of domains can indicate an origin of control
  - Unique C&C domains present a small amount of resolution traffic
- One example in Mytob/Zotob botnet

DNS Properites



DNS Properties

- In the KR Botnet attack, however, the hosts involved in the DDoS resolved numerous popular sites to generate a DDoS

DNSBL Properties

- A few victims had previous DNSBL listings
  - Out of 991 sampled IPs, 359 had prior DNSBL listings
  - This immediately suggests a naïve victim base, or a simplistic attack vector (since sophisticated attacks would recruit victims with less extensive DNSBL histories).

Conclusion

- Botnets: the source of the most serious and damaging attacks
- Challenges:
  - Botnet activities are not attacks in the traditional sense
  - Bots are stealth
    - They are valuable resources to the bot masters
- Need multifaceted approach, at the minimum:
  - Monitor the web/internet infrastructures (e.g., DNS and Web hosting)
  - Malware/script analysis
  - Monitor host and network activities

Credits

- David Dagon
- Roberto Perdisci
- Monirul Sharif
- Andrea Lanzi
- Jon Giffin
- Nick Feamster

Thank You!