Kopis: Detecting Malware Domains at the Upper DNS Hierarchy

Manos Antonakakis

Damballa Inc.

Georgia Institute of Technology, College of Computing

Usenix Security, August 2011, San Francisco, CA.

Credits

Roberto Perdisci, Wenke Lee,
Nick Vasiloglou and David Dagon

- Outline
  - Motivation & Contribution
  - Notation & Data Sources
  - Kopis' Statistical Components
    - System overview & statistical features
  - Results
    - Detection results
    - The rise of IMDDOS
  - Conclusions

Motivation

- IP-based (dynamic or not) blocking techniques cannot keep up with the number of IP addresses that the C&C domains use.
- DNSBL-based technologies cannot keep up with the volume of new domain names botnets use every day.
- Malware families utilize numerous domains for discovering the "up-to-date" C&C address.
- There is a time gap between the day the malware is released and the day the security community analyzes it.
- The daily DNS lookup signal for malware related domain names is different than normal sites:
  - Infection/remediation/OS failures of the infected machine(s) causes it to vary over time
  - Really hard to control the malware propagation phase
- ***Can we learn anything, by statistically modeling the DNS resolution patterns?***

Motivation

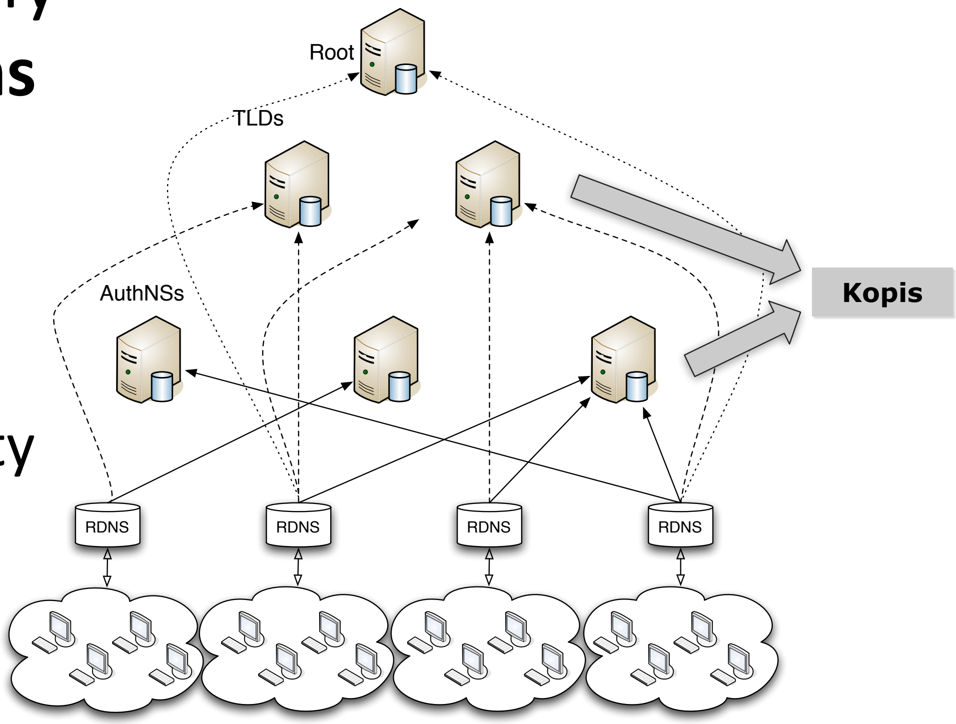
- IP-based (dynamic or not) blocking techniques cannot keep up with the number of IP addresses that the C&C domains use.
- DNSBL-based technologies cannot keep up with the volume of new domain names botnets use every day.
- Malware families utilize numerous domains for discovering the "up-to-date" C&C address.
- There is a time gap between the day the malware is released and the day the security community analyzes it.
- The daily DNS lookup signal for malware related domain names is different than normal sites:
  - Infection/remediation/OS failures of the infected machine(s) causes it to vary over time
  - Really hard to control the malware propagation phase
- ***Can we learn anything, by statistically modeling the DNS resolution patterns?***

Motivation

- IP-based (dynamic or not) blocking techniques cannot keep up with the number of IP addresses that the C&C domains use.
- DNSBL-based technologies cannot keep up with the volume of new domain names botnets use every day.
- Malware families utilize numerous domains for discovering the “up-to-date” C&C address.
- There is a time gap between the day the malware is released and the day the security community analyzes it.
- The daily DNS lookup signal for malware related domain names is different than normal sites:
  - Infection/remediation/OS failures of the infected machine(s) causes it to vary over time
  - Really hard to control the malware propagation phase
- ***Can we learn anything, by statistically modeling the DNS resolution patterns?***

Kopis

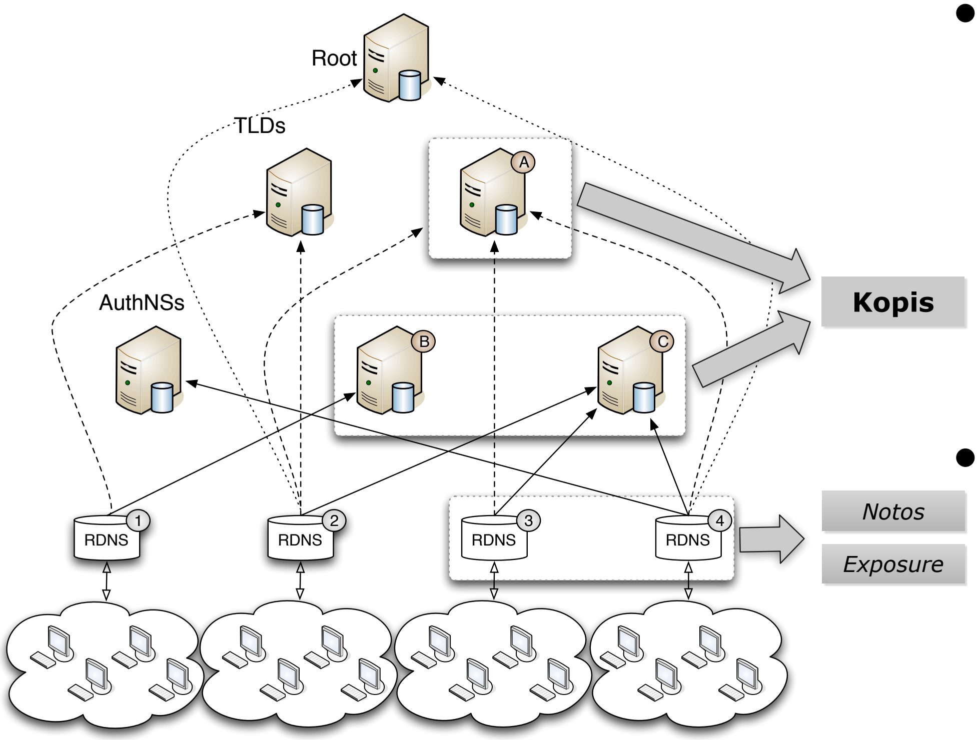
- A new approach to identify **malware-related domains**
- Observations:
  - DNS is a hierarchical, distributed database
  - We can gain global visibility per authority
- Goal: the detection of malware domains
  - On the rise
  - No need for a sample



Contributions

- Kopis can analyze large volumes of DNS messages at AuthNS or TLD servers
- Kopis introduces an alternative ***IP-reputation agnostic*** classification signal for DNS
- ***With Kopis we identify rising botnets weeks before corresponding malware is found***

Notos vs Kopis



- Notos and Exposure
  - Almost global visibility on zones
  - Partial visibility on the requesters
- Kopis
  - Global visibility on requesters
  - Focused in a specific set of zones

Difference in visibility enables different features with complementary detection abilities

Getting Familiar with the Basic Building Blocks

Basic Building Blocks

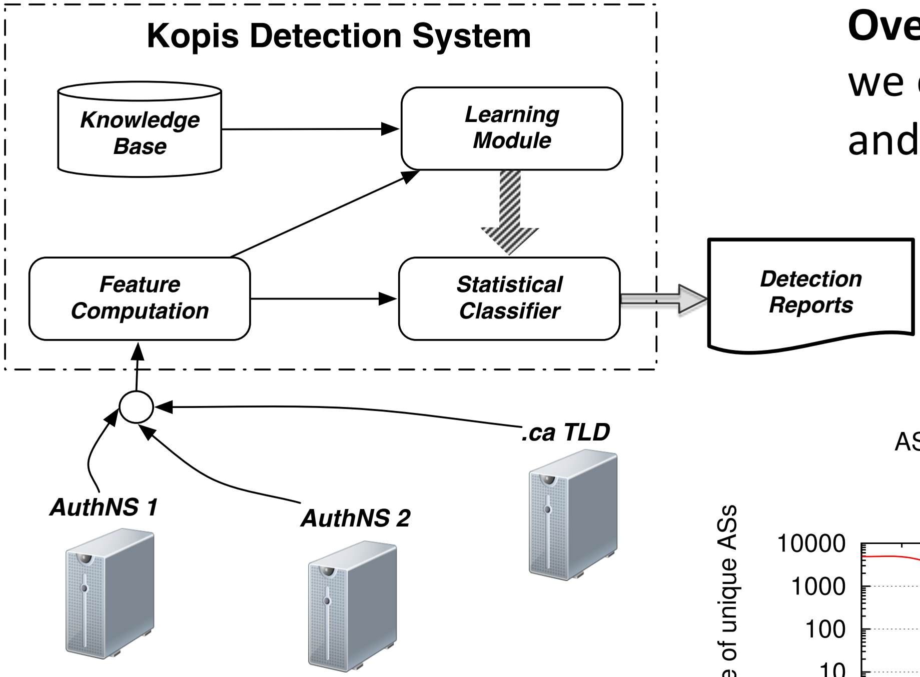
- What is a Resource Record (RR)?
  - www.example.com 192.0.32.10

- Authoritative domain name tuples?
  - **Who** is looking up **what** and **where** is pointing?

- We obtained authoritative DNS traffic from two large authoritative DNS servers (AuthNS) and the Canadian TLD (via SIE)

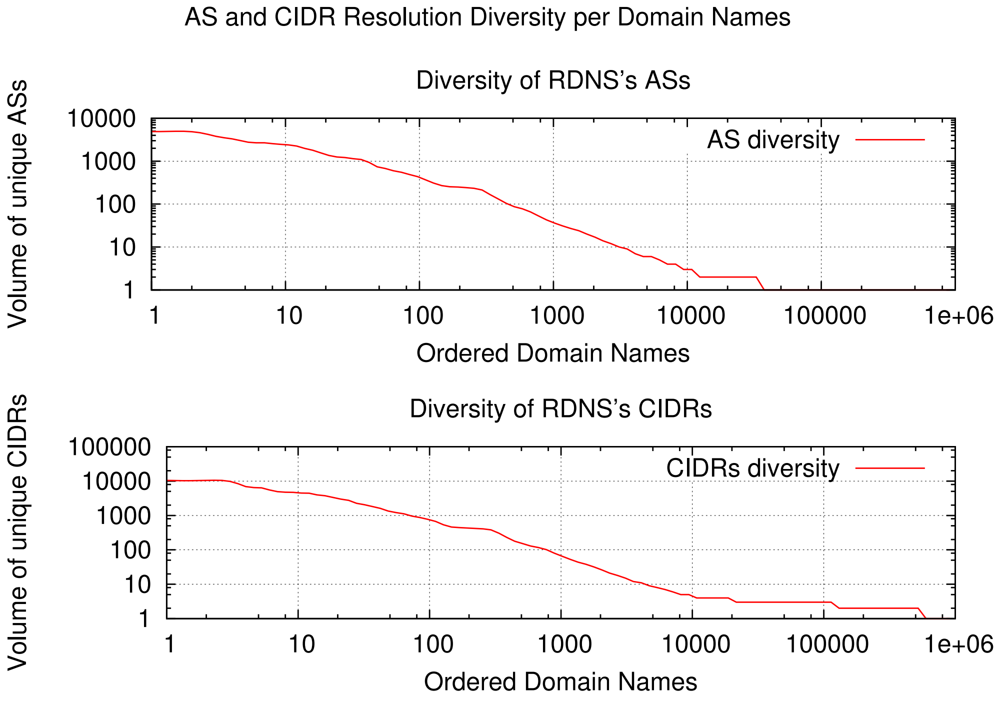
Looking into Kopis

Overview



Overview of Kopis: Based on ground truth we can model lookup patterns from benign and malware-related domain names.

Key Observation: Not all domain names are equally interesting. We chose to spend time analyzing the most interesting based on the lookup requester diversity.


AS and CIDR Resolution Diversity per Domain Names

Diversity of RDNS's ASs

Diversity of RDNS's CIDRs

Statistical Features

- Requester Diversity
  - Characterize if the machines (e.g., RDNS servers) that query a given domain name are localized or are globally distributed
- Requester Profile
  - Determine if machines resolving the domain names are from networks that historically have been prone to infections or not
- Resolved-IPs Reputation
  - Describes whether, and to what extent, the IP address space pointed to by a given domain has been historically linked with known malicious activities, or known legitimate services

Statistical Features

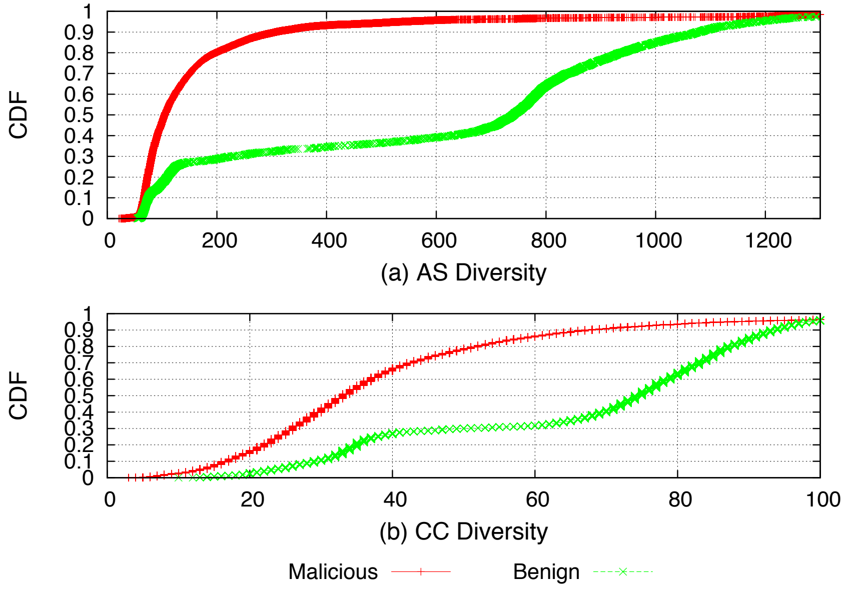
- Requester Diversity
  - Characterize if the machines (e.g., RDNS servers) that query a given domain name are localized or are globally distributed
- Requester Profile
  - Determine if machines resolving the domain names are from networks that historically have been prone to infections or not
- Resolved-IPs Reputation
  - Describes whether, and to what extent, the IP address space pointed to by a given domain has been historically linked with known malicious activities, or known legitimate services

Statistical Features

- Requester Diversity
  - Characterize if the machines (e.g., RDNS servers) that query a given domain name are localized or are globally distributed
- Requester Profile
  - Determine if machines resolving the domain names are from networks that historically have been prone to infections or not
- Resolved-IPs Reputation
  - Describes whether, and to what extent, the IP address space pointed to by a given domain has been historically linked with known malicious activities, or known legitimate services

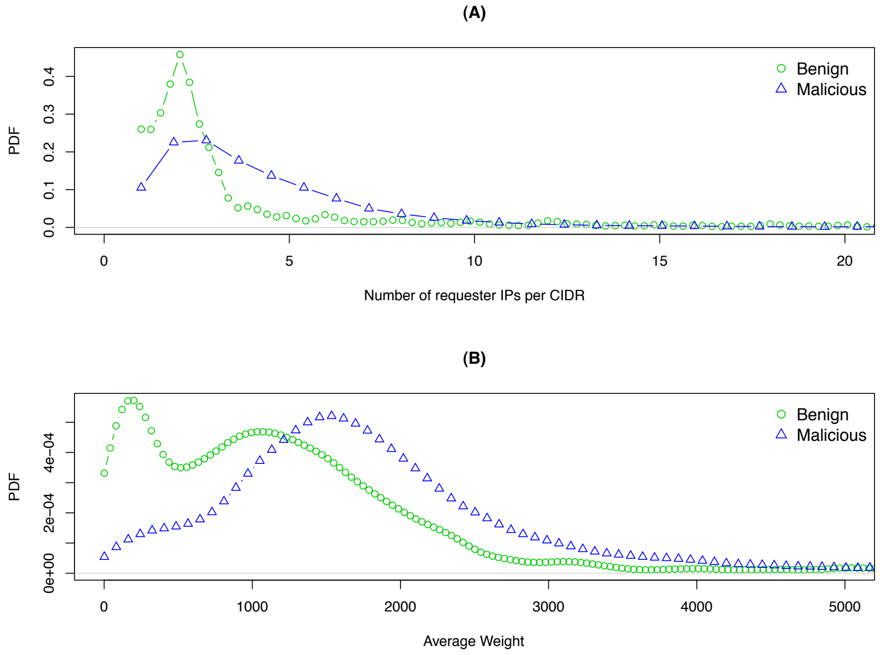
Requester Diversity

- Looking closer into the diversity of the requesters per CC and AS point of view:
  - For both features the benign domain names have a bimodal distribution.
  - Malicious domain names are spread across the spectrum.
- The malware-related domain names cover a larger spectrum of diversities:
  - This could be due to the success of the malware distribution mechanisms they employ.



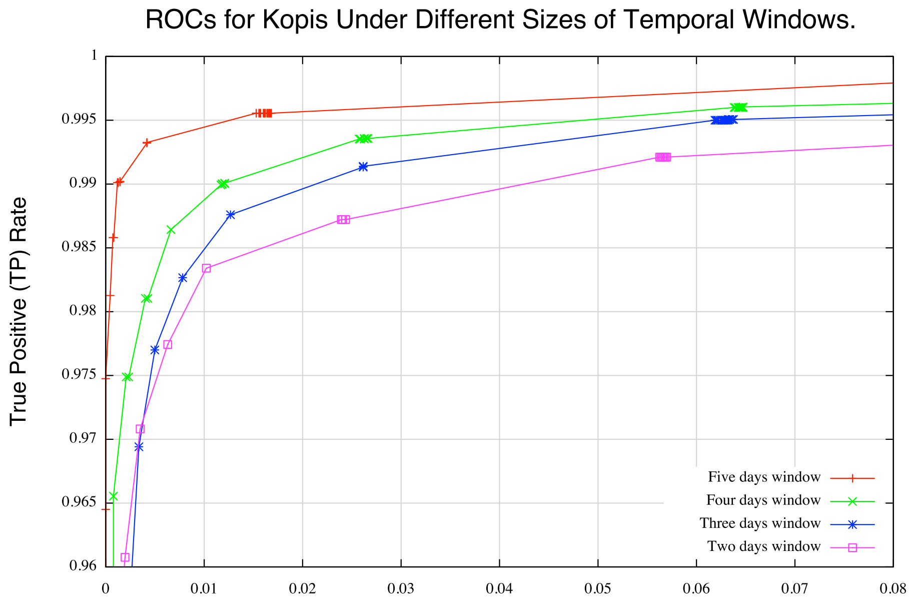
Requester Profile

- Not all querying machines have similar characteristics.
- We would like to distinguish between requesters located in ISP/small business and home networks.
- We model differently the weight to **long-lived and stable** RDNS servers.
- Evasion protection:
  - *The weighted RP features* make it significantly harder to dilute the overall classification signal that Kopis models, because DNS lookups that originate from RDNSs with low daily domain lookup spectrum will be **depreciated.**



Results

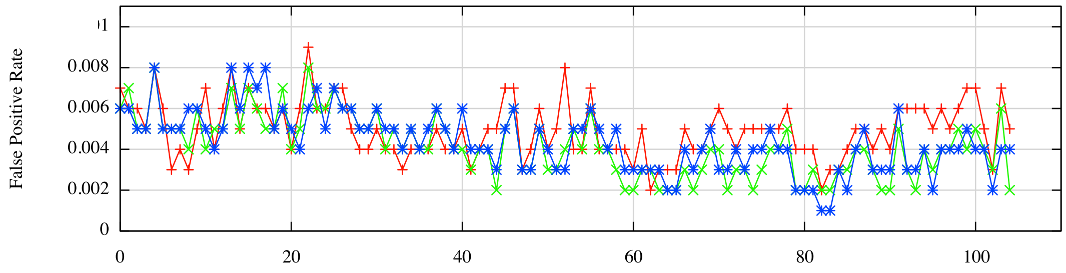
Kopis FP's and AUC

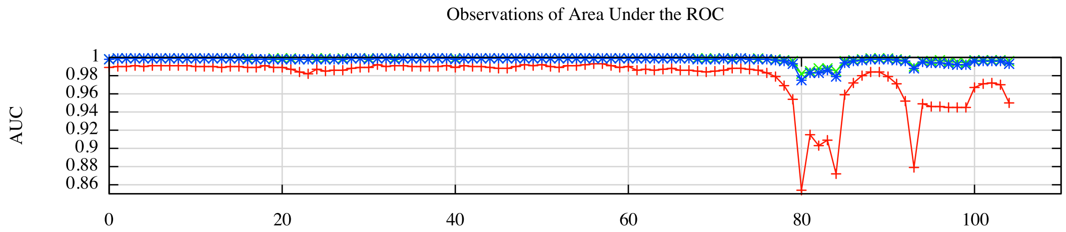


False Positive (FP) Rate

FP%=0.3% and TP%=98.4% using a five day training window

Observations of FP rates

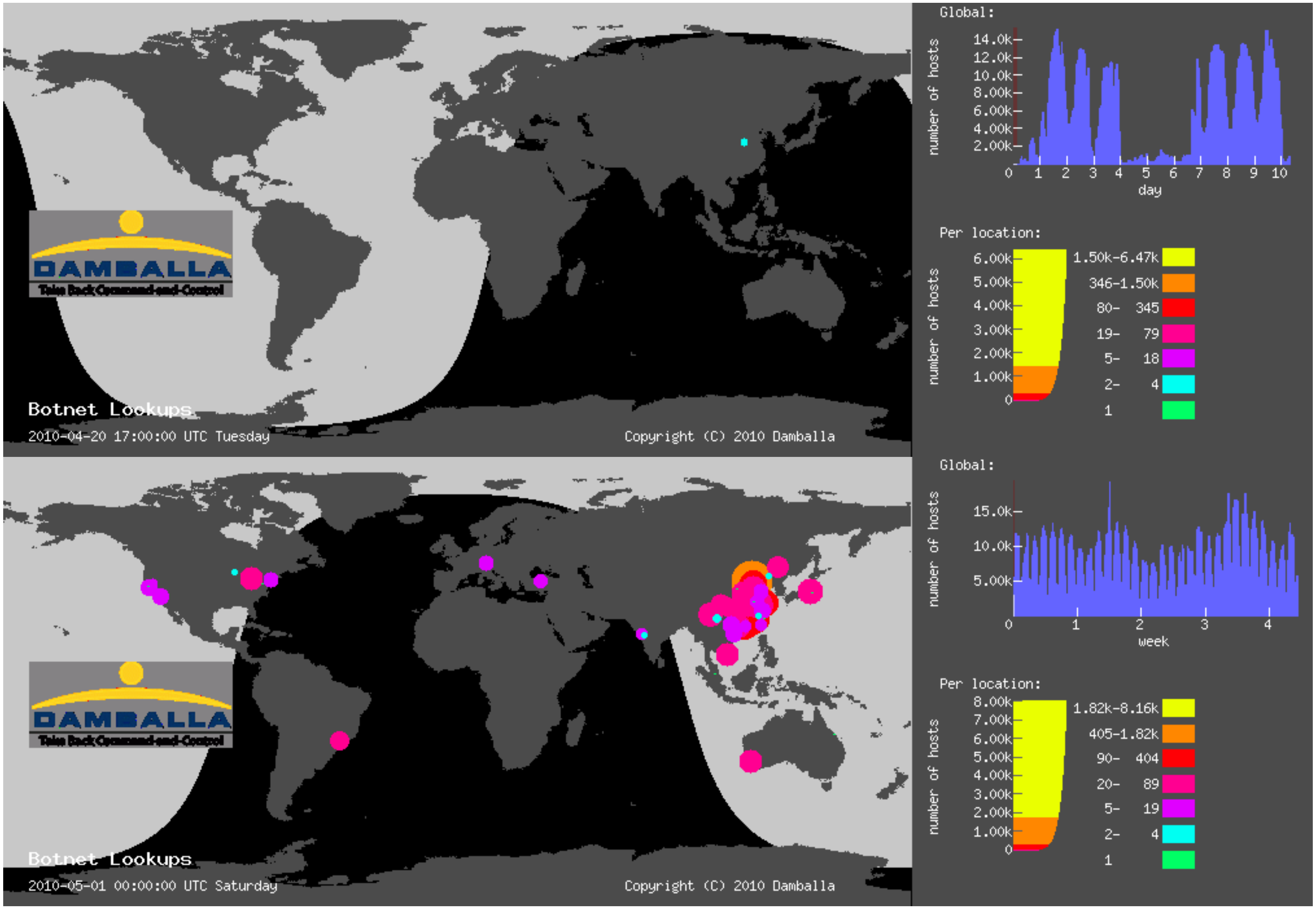




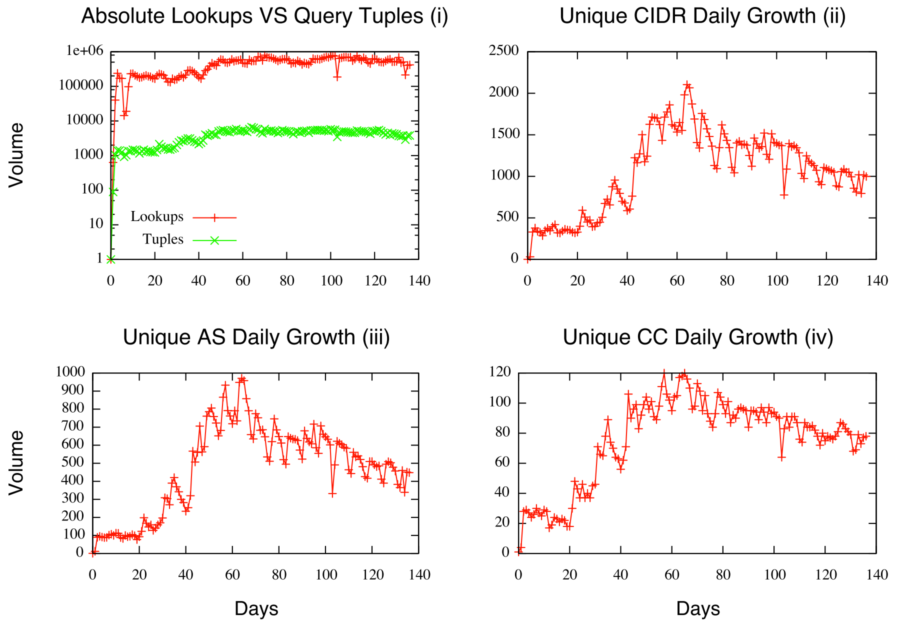
Long-term evaluation with real data shows that Kopis can reliably detect new malicious domain names, while maintain low FP rates

Looking into some botnets

The Rise of IMDDOS



Early Detection of IMDDOS


Absolute Lookups VS Query Tuples (i)

Unique CIDR Daily Growth (ii)

Unique AS Daily Growth (iii)

Unique CC Daily Growth (iv)

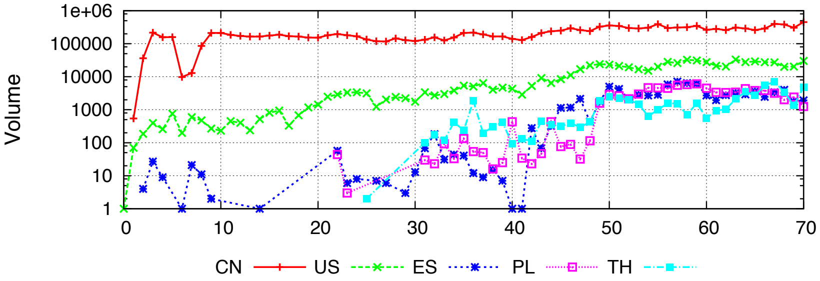
The average lookup volume every day was 438,471 with the average de-duplicated query tuples in the range of 3,883.

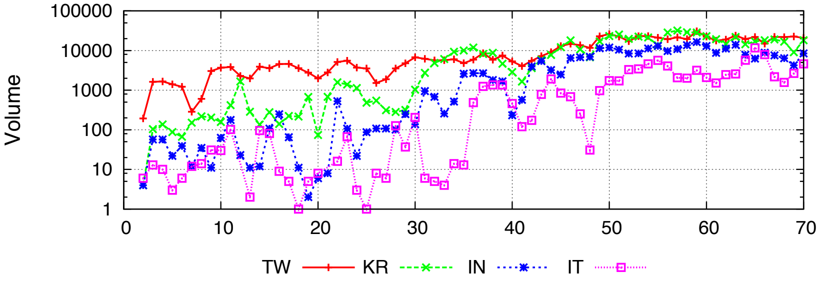
We observe that the daily growth of unique CIDRs, AS and CCs related to the RDNSs that queried the domain names used in the botnet followed the same pattern.

First big infection happened in Chinese networks in a relative short period of time.

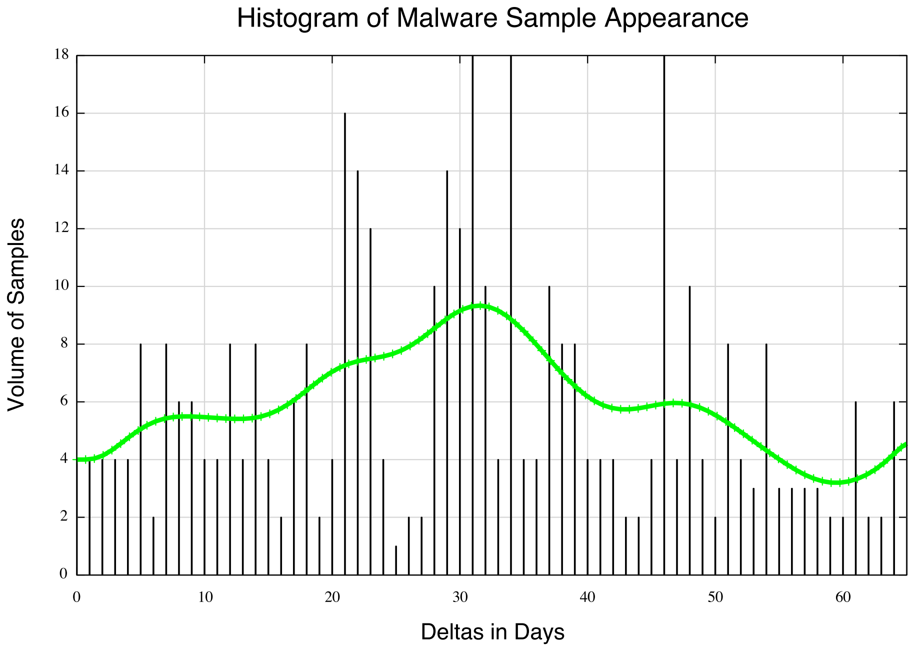
Look-ups from US networks reached 1,000 more than 20 days after the botnet was launched.

Italy, Spain and India reached the 100 daily lookups 15 days later than the beginning of this botnet.

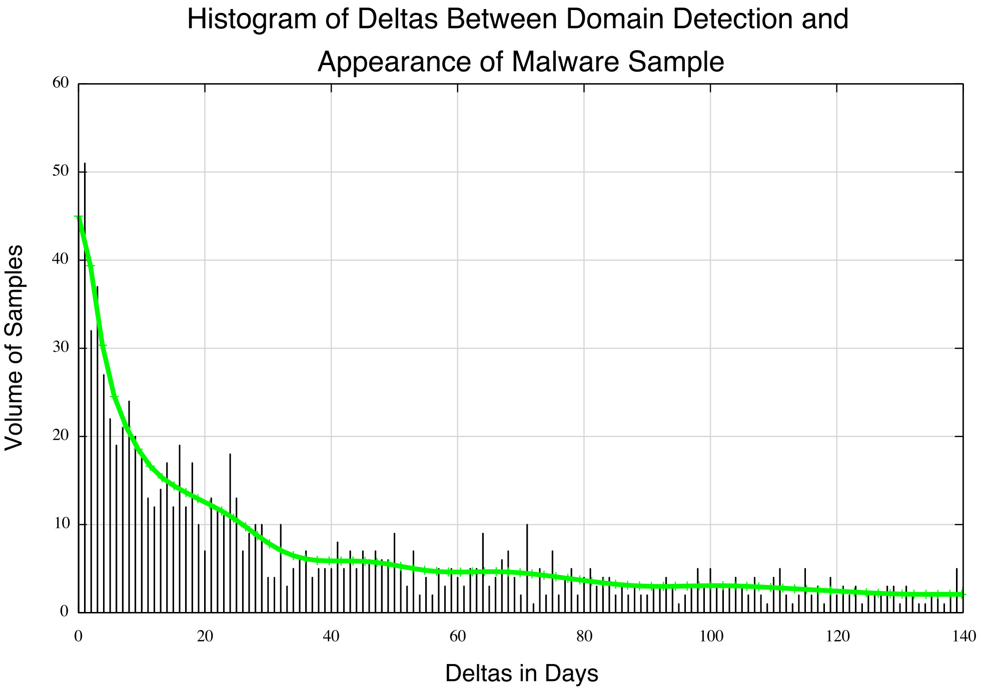




Delta between Kopis and Malware

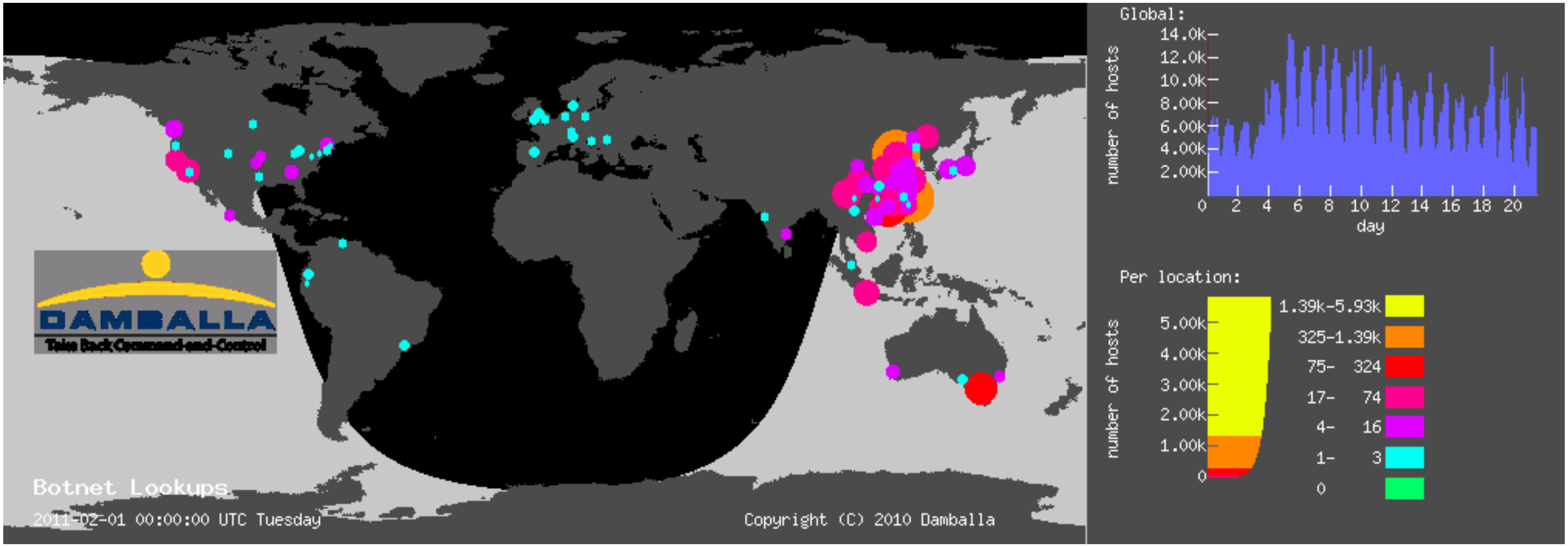


In the case of IMDDOS as malware and other variants appear in our malware feed



Using domain name from our testing dataset in 80/20 mode

H1 Botnet Discovery

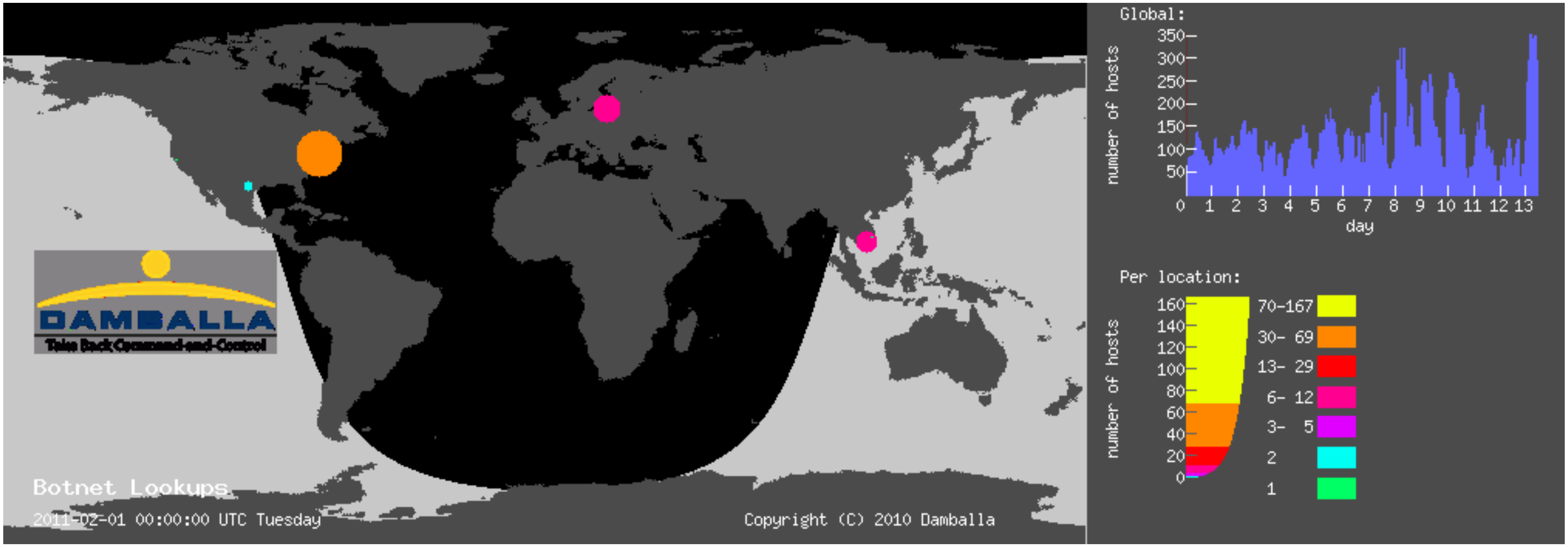


The ***H1 Botnet*** has a lower estimate of 4K infected IPs in 676 networks with a country code distribution of: ***1100 CN, 636 TW, 416 US, 244 KR, 78 HK, 69 JP, 50 FR, 45 CA etc.***

The C&Cs are hosted in (2 US, 1 SA and 1CN): VPLSNET - VPLS Inc. d/b/a Krypt Technologies. (174.139.97.122 ,98.126.115.90), SAUDINETSTC-AS (2.88.6.188), Take 2 Hosting (173.252.197.103), CHINANET-BACKBONE (118.123.12.6)

Single MD5: 9f9a**X.** Detected December 2010, malware obtained on February 2011.

Lenovo Botnet Discovery

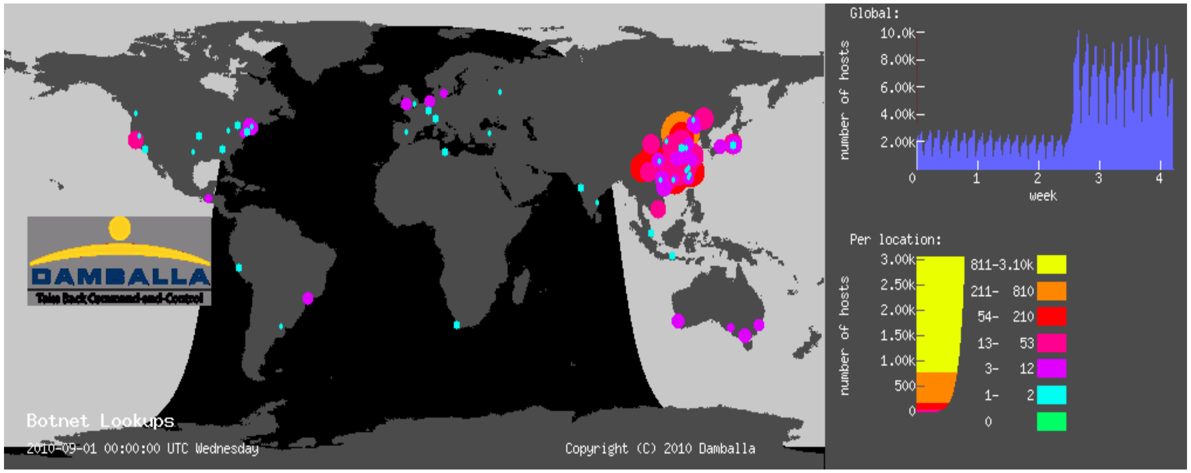


Lenovo Botnet: Hosted in CHINANET-BACKBONE 61.183.44.0/23. Nine domain names were linked with C&C activities.

~1K infected hosts in the overall (358) infected networks with distribution: 98 VN, 85 US, 65 TW, 38 FR, 12 CN, etc.

MD5 45f5**X**. Detected during end of August 2010, malware obtained on November 2010.

c0c1 Botnet (?) Discovery

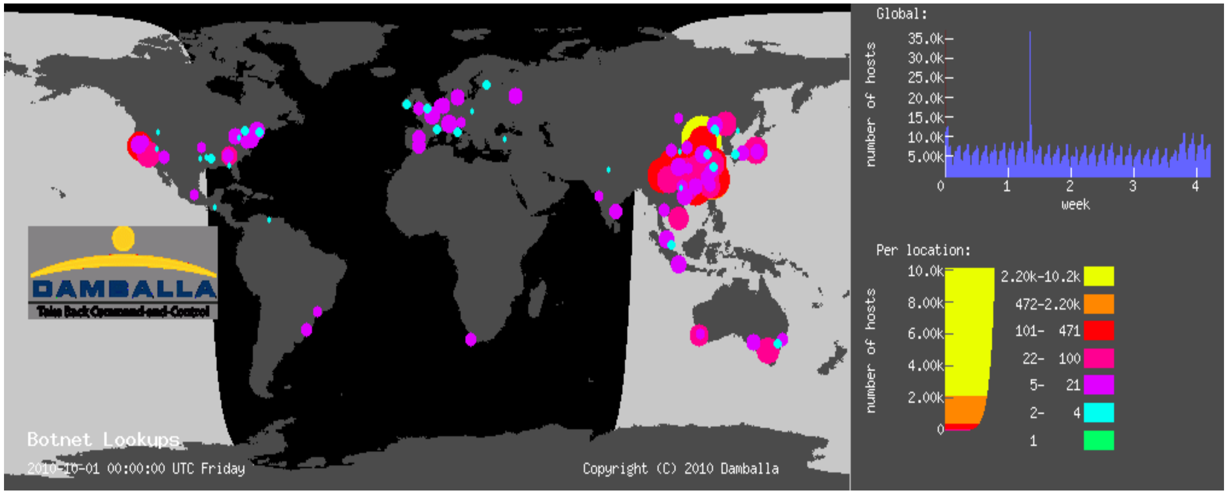


c0c1 Botnet: no MD5 yet. There are (potentially) ~15K infected IPs in 44 different counties.

The IP per CC distribution for the first month was: 1324 CN, 661 TW, 238 KR, 223 US, 91 JP, 82 ES, 66 HK, 52 FR, 34 PL, 34 D, 21 VN, 20 BR etc.

The (potential) 22 C&Cs were hosted at AS30058 FDCSERVERS (US).

The domains were suspended without any complaints. The botnet was identified in August 2010.



Looking into the Largest FP

- *Phishing campaign* detected by Kopis as malware-related domain.
- Brand hijacking and fake UGGs
- Four domain name were linked via CNAME to domains under our authorities.
- They were hosted in SHARKTECH (US->CN). {208.98.0.0/18, 70.39.105.0/24 , 174.128.229.0/24, 174.128.229.0/24}
- 25K IPs visited these domains over ***2 months*** from 193 different networks.
- Payments at
  - pay.ips.com.cn
  - Very common for fake UGGs

Conclusions

- We need additional classification signals:
    - Evasion is harder
    - Threat landscape is changing
- Malware is out there up to a couple of months before the security community finds a related MD5
- Contributions of Kopis:
    - We can detect and stop a botnet while it is on the rise **before** the responsible malware is found
    - We have the ability to measure and model key properties of malware domain names on the rise
    - Independently deployable by network operators
    - ***Early warning:*** able to detect malware domains before the malware reaches your network!
    - Low FPs and high TPs in almost all evaluation modes

Thanks! Questions?

Manos Antonakakis

manos@antonakakis.org