DHS PREDICT IMPACT project: CAIDA update

Bradley Huffaker, 
kc claffy, PI
Marina Fomenkov, Co-PI

UCSD, La Jolla, CA
2-3 February 2016
DHS IMPACT project: CAIDA update

• Data collection activities
  • Ongoing measurements
  • Data storage status
  • Data dissemination statistics
  • Aftereffects of changing CAIDA data access policies in 2014

• Other activities
  • Enabling commercial use of CAIDA data
  • Legal issues with providing CAIDA data to foreign military entities
  • Ethical issues in Cyber Research

• Open issues
  • Contributing new data to IMPACT
  • Re-orienting CAIDA data access policies toward IMPACT
Ongoing Measurements: Ark platform

- Concurrent data collection
  - IPv4 and IPv6 topology
  - spoofer
  - targeted congestion measurements

- Ark Platform (as of Jan 2016) - 132 monitors
  - 54 IPv6 enabled
  - 85 Raspberry PIs

=> Accelerated infrastructure growth
   All new monitors - Raspberry PIs
Ark Produced Data

• Data sets
  • IPv4 Routed /24
  • ITDK (the last was in August 2015)
  • IPv4 Routed /24 DNS names

• More to come - access mode TBD
  • spoofer web reports
  • spoofer data
  • congestion data (supporting publications)
  • ...

* => current - restricted, older than 2 years - publicly available

• AS links, IPv4 - daily
• AS links, IPv6 - daily, by monitors
• IPv6 DNS names
• AS ranking

Text => publicly available
Ongoing Measurements: Internet Background Radiation

• UCSD Network Telescope
  • 10 TB in December 2014
  • 22.6 TB in December 2015
  • >500 TB archived at NERSC
  • can prepare archived samples on demand

• Gold mine for security-related data sets

• Near-real-time data
  • two months sliding window of IBR traffic

• Archived samples are fine for many research uses
  • anonymization decreases the utility
Ongoing (for now) Measurements: Passive Trace Collection

• **Passive infrastructure**
  - two monitors with taps and Endace 10GE capture cards on the Equinix link in Chicago
  - *lost* the **San Jose** link in **2014** - it was upgraded to 100GB
  - they still may find a slower link there for us to monitor

• **Data**
  - Annual “Anonymized Internet Traces” since 2008
  - Our **most popular restricted dataset**
  - about **two thirds** of all requests for restricted datasets
Data storage

- Continue using NERSC for telescope data (free of charge)
- Continue to expand CAIDA storage
  - Tried SDSC Cloud, decided that local storage/backups are more convenient for us
- Resources required:
  - Labor maintaining our own hardware
  - Buying new disk shelves
  - Buying new data servers
Data Access

• 2014 - **transitional year**

• **Made the following data sets public:**
  - Ark IPv4 topology data older than 2 years
    - including raw data and ITDKs
  - All Ark IPv6 topology data
  - Older telescope data:
    - Backscatter data 2001-2008
    - Code-Red and Witty Worm data
    - Sipscan 2011 dataset
    - other archived data

• **Announcements**
  - February 2014 - blog about topology data
  - April 2014 - email to the list of CAIDA data users about changes in access mode for topology data
  - **never announced public availability of telescope data**
    - waiting to sort the availability and access issues with PREDICT
Access to publicly available data

- Fast, efficient, very user friendly
  - User info: entirely optional
    - some users still provide their info
  - No hand shake
  - Immediately downloadable

- Simplified AUP - less than a page
  - License
  - Suggested citation format
  - Disclaimer

=> Very popular!
non-CAIDA publications using IMPACT-related CAIDA data (that we know of)

* only self reported, need to spend a week looking for them.
non-CAIDA publications using PREDICT-related CAIDA data (that we know of)

* only self reported, need to spend a week looking for them.
Restricted Dataset Requests

received/approved requests for restricted datasets

nr of requests

caida
amount **restricted** data downloaded

Amount of restricted data downloaded

- Anonymized Internet traces
- UCSD Telescope
- Ark topology
- Witty
- Backscatter
- DDoS
- DNS lookups
- DNS root/gTLD RTT

TB

Amount of public data downloaded

- Anonymized Internet traces
- UCSD Telescope
- Ark topology
- AS Relationships
- Ark AS links
- Skitter topology
- Other public data
- Published data supplements

TB

The number of users downloading restricted data is shown in the graph. The categories include Anonymized Internet traces, UCSD Telescope, Ark topology, Witty, Backscatter, DDoS, DNS lookups, and DNS root/gTLD RTT. The graph displays the growth over years from 2005 to 2016.
number of users downloading **public data**

Number of users downloading public data

- Anonymized Internet traces
- UCSD Telescope
- Ark topology
- AS Relationships
- Ark AS links
- Skitter topology
- Other public data
- Published data supplements
Recent publications

Recent publications


Potential New Data Sets

• **CAIDA Internet Census 2013**
  • **category: Address Space Allocation Data**
  • **sub-category: TBD**
  • **derived from both active and passive measurements**
  • **classifies /24 address blocks as:**
    • IETF-reserved
    • observed as used
    • routed unused (unobserved)
    • unrouted assigned
    • available
  • **includes AS announcing in BGP the longest network prefix containing this address block**
  • **access mode: TBD**

• **spoofer: project in progress, data sets TBD**
Other Security-related New Data Sets

- scanning data, 2008-2015
  - IP, protocol, dst port of scanners
  - hourly rates of packets, flows
  - destinations
  - scanning strategy

- backscatter data
  - attack on Spamhaus 2013 (TCP backscatter)
  - BitTorrent index poisoning attacks

- attacks using DNS open resolvers

- botnet data: ZeroAccess and Sality
  - Command and Control packets
Other Security-related New Data Sets (cont.)

- Details and Methodology in

- Access: restricted
  - Anonymization would decrease the research utility
Ideas for IMPACT?

- fund meta-data research:
  - ascertain researchers data needs
  - access the impact of data age on data usefulness
  - disclosure control policies and methods

- continue experimenting with data access modes

- continue ethics/policy activities

- expand Data Host activities?

- change metrics

- there is (always) room for more marketing efforts
  - cf. typical advertising strategies: “How did you learn about our product”? (web, newspaper, TV, Facebook, friends…)

- Make CAIDA data accessible exclusively via IMPACT portal?