DHS PREDICT project: CAIDA update

- Infrastructure updates
- Data collection updates
- Data set dissemination statistics
- Other activities
- Open issues
Infrastructure

- SDSC retired HPSS December 2011 and will retire SamQFS by summer 2012. SDSC will have no tape on the machine room floor.
- CAIDA working with SDSC to use Micro-condo cloud storage services (http://cloud.sdsc.edu/)
  - purchasing 25 TB of cloud storage for last 60 days of CAIDA system backups (pricing at http://rci.ucsd.edu/services/)
  - granted 20TB from SDSC Executive Team for scientific data
    - telescope (103TB)
    - packet headers (18.8TB)
    - skitter/ark topology (4TB)
Data collection - passive

- **OC192 backbone:** March 2008 - Dec 2011
  - 18.8 TB compressed, 35.7 TB uncompressed
  - unanonymized: 10.9 TB compressed, 21.3 TB uncompressed
  - anonymized: 7.9 TB compressed, 14.4 TB uncompressed
  - Doing cleanup toward retaining only quarterly traces
  - Completed 2011 Passive Datasets

- **Problems:**
  - Hardware failures at collection sites: Chicago monitors have been offline since September. Still trying to work with remote hands to troubleshoot.

- **Plans:**
  - New 2012 dataset will start with upcoming trace January 19
  - strip payload/L1/L2, transfer, anonymize, archive
  - collect 1 hour trace per month = 200-250 GB (compressed)
  - keep a quarterly sample - select the best quality
Data collection - passive

- **UCSD telescope:**
  - data from most recent 30-days (really five weeks) “live” on disk
    - typically 2.9 TiB compressed, 5.5 TiB uncompressed
  - previous month(s) - backed up on tape (now samqfs)
    - current: 2008/04/12 - 2012/01/12
    - 102 TB (compressed), 192 TB (uncompressed)
    - received new NSF award “CRI-Telescope: A Real-time Lens into Dark Address Space of the Internet” (next page)

- **OC48 traces:**
  - 964.5 GB (compressed), 1.7 TB (uncompressed)
  - unanonymized: 815.7 GB (compressed), 1.5 TB (uncompressed)
  - anonymized: 148.8 GB (compressed), 285.2 GB (uncompressed)
    (in PREDICT)
Data collection and analysis - telescope

- **Storage Transition Plan**
  - Short-term (1 month) vs Medium-Term, Long-Term

- **Telescope census questions**
  - Denial-of-service attacks
  - Specific Datasets
  - Entropy (Tanja)
  - Payload (Tanja)
  - Anomaly detection (Tanja)
  - Country-level/AS-level outages (IMC paper)
  - One-way traffic monitor (Nevil, Alistair)
Data collection - active

• old skitter data (in PREDICT):
  • 1.47 TB (compressed), 4.02 TB (uncompressed)
  • discontinued in February 2008

• current Ark data:
  • IPv4 topology: 1.8 TB (compressed), 5.8 TB (uncompressed)
  • IPv6 topology: 2.8 GB (compressed), 9.7 GB (uncompressed)
  • 57 monitors in 30 countries, 28 IPv6 capable
  • continues to expand

• data curation:
  • create derivative data sets
  • analyze/annotate -> ITDK
    • router-level topologies: nodes and links
    • host names
    • router-to-AS assignment
    • geographical information
      • http://www.caida.org/data/active/internet-topology-data-kit/
Archipelago Monitors and Data
# Requests for the data, 2011/2010/2009

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Requests</th>
<th>Approved</th>
<th>Accessed</th>
<th>Served</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backscatter</td>
<td>51/73/95</td>
<td>34/47/60</td>
<td>28/36/46</td>
<td>Feb 2003</td>
</tr>
<tr>
<td>Passive</td>
<td>275/185/233</td>
<td>210/150/179</td>
<td>170/126/157</td>
<td>Feb 2004</td>
</tr>
<tr>
<td>Topology</td>
<td>155/163/129</td>
<td>129/113/83</td>
<td>85/80/63</td>
<td>Jul 2004</td>
</tr>
<tr>
<td>Witty</td>
<td>16/16/27</td>
<td>12/13/17</td>
<td>10/11/14</td>
<td>Mar 2008</td>
</tr>
<tr>
<td>Telescope</td>
<td>29/34/37</td>
<td>22/23/21</td>
<td>18/19/17</td>
<td>Jul 2009</td>
</tr>
<tr>
<td>DNS-RTT</td>
<td>10/7/7</td>
<td>8/5/2</td>
<td>6/4/2</td>
<td>Aug 2006</td>
</tr>
<tr>
<td>DDoS</td>
<td>92/108/NA</td>
<td>62/74/NA</td>
<td>51/66/NA</td>
<td>Mar 2010</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>628/586/528</strong></td>
<td><strong>477/425/362</strong></td>
<td><strong>368/342/299</strong></td>
<td></td>
</tr>
</tbody>
</table>
Data request stats

- all requests (cumulative)

3405 requests received
2297 requests approved
1868 accounts accessed
Data request stats (cont)

- All requests (monthly)
  - spike (40 requests) in first month of DDoS dataset

![Graph showing data request stats]

- 3405 requests received
- 2297 requests approved
- 1568 accounts accessed
Data Set Popularity

- **Most popular - OC192 and OC48 traces**
  - requested 693 times, accessed 454 times (since 2009)
  - who used it: 259.edu, 141.cn, 43.uk, 28.com (since 2004)...
    - and 56 more domains
    - of 839 total accounts: 270 from U.S.

- **2nd most popular - topology data**
  - requested 447 times, accessed 228 times (since 2009)
  - who used it: 256.edu, 119.cn, 41.uk, 31.kr, 29.com, 26.jp
    (since 2004)...
    - and 52 more domains
    - of 785 total accounts: 191 from U.S.
Data availability

- **PREDICT (OC48 traces, topology from skitter, telescope)**
- Derived data sets publicly available (e.g., AS-links)
- Academics (non-commercial) who sign AUP (OC192, topology from Ark, telescope)
- Commercial researchers
  - a small sample of CAIDA data to entice interest
  - join CAIDA, various membership levels are offered
Data statistics - online

- Aggregated, (near) real time
- OC192 backbone
  - report generator
  - http://www.caida.org/data/realtime/passive/?monitor=equinix-chicago-dirA
- Telescope
  - report generator
  - http://www.caida.org/data/realtime/passive/
- topology
  - path dispersion (AS and IP), path length distribution, RTT distribution, RTT vs. distance, median RTT per country, ...
Meta-data for packet traces

• **OC192 data: 2008-2011**
  • an hour-long trace every month
  • usually, 3rd Thursday, 13:00 - 14:00 UTC

• **OC48 data: 2002-2003**

• **Statistics:**
  • Date, start time, stop time
  • Numbers of IPv4, IPv6, unknown packets
  • Transmission rate in pkts/s, bits/s
  • Link utilization (%)
  • Average packet size
  • Graph of packet size distribution (IPv4 and IPv6)

http://www.caida.org/data/passive/trace_stats/
Recent PREDICT-related publications


- Erin Kenneally *A Refined Ethical Impact Assessment Tool and a Case Study of its Applications*, submitted to WECSR 2012.

- kc claffy, *Tracking IPv6 Evolution: Data we have and Data We Need*, ACM SIGCOMM CCR V. 41, p. 43-48, 2011.


Recent PREDICT-related publications

  • national level outages in Egypt and Libya
  • data used: public BGP, UCSD telescope, Ark (little bit)
  • analyzed methods used for traffic blocking, duration, testing

  • cross-analyzed multiple databases
  • used available ground truth data (PlanetLab, French networks, Tier 1 provider)
  • Ark RTT data
non-CAIDA publications using PREDICT-related CAIDA data (that we know of)

- total : 129
- backscatter : 15
- passive-oc48: 52
- passive-2007: 8
- witty : 12
- itdk : 9
- skitter : 51
requests for PREDICT CAIDA data in 2011

- April (2): ITDK 2003 and skitter (topology)
- July (1): ITDK 2003 and skitter (topology)
Recent blogs

- **kc claffy, network neutrality: the meme, its cost, its future**

- **kc claffy, underneath the hood: stewardship vs. ownership of the Internet**

- **kc claffy, My third FCC TAC meeting - the most exciting meeting yet**

- **kc claffy, in response to NTIA on IANA functions**
Phase II Data Sets

- UCSD telescope: near Real-Time Telescope Dataset (RTTD)
- topology: Ark data (ongoing)
  - IPv4 Routed /24 Topology dataset
  - IPv4 Routed /24 DNS Names dataset
  - IPv6 Routed Topology dataset

- OC192 backbone: 2007-2011
Preparations for Phase II

• We are ready to go!
• New MOAs signed
• Data descriptions submitted
  • Prepared and reviewed meta-data
• reviews/refinement of CAIDA AUPs
  • work still in progress
CAIDA Master AUP

- 4 categories of data - different levels of sensitivity
  - real-time telescope data
  - passive traces
  - active traces
  - derived topology

- Document proliferation
  - 7 data request forms
  - 22 data set web pages
  - 22 README files

- Master AUP 1.0 for all CAIDA data sets
  - Factor out common conditions
  - Remove inconsistencies
  - Supplemental provisions for special data (e.g., RT telescope)

- Will publish for community use
Other activities

• flood of digital data

• DMPTool
  • helpful but does not add much for our purposes

• no ready-to-use guidelines
  • NSF-required Data Management Plan
  • who bears the cost?
  • how much is the cost?
    • thousands of $ per TB per year - commercial clouds
    • $390 per TB per year - SDSC preferred rate
    • $3,000 per TB to store forever - Princeton offer
  • NSF position: communities should develop acceptable guidelines
    • what to store?
    • for how long?
New hires

• Ark system administrator: Parisa Nahavandi
• Telescope research programmer: Alistair King
• Visiting Scholar: Tanja Zseby, working with telescope group
• Postdocs: Matthew Luckie (IPv6, topology) and Alberto Dainotti (telescope)
CAIDA Marketing Efforts

• **Web site**
  - Annual reports, Program Plan, Project web page, blogging

• **Publications, Presentations, Workshops**

• **Related Projects**
  - NSF funded SDCI
    • reduce burden on contributors
    • convert from proprietary format to open source
    • expand relevance to cyber security

  - NSF funded CRI - telescope research
    • support “near real-time”, “bring code to the data” model
    • develop automated triggers and alerts
    • curate custom data sets upon request

  - NSF funded IRNC - International Research Network data
    • deploy Ark and passive monitors on IRNC links
    • new measurement functionality: DNSSEC, IPv6
    • prototype “international bureau of internet statistics” report