DHS PREDICT project: CAIDA update

Kimberly Claffy, CAIDA

July 26-27, 2011

DHS PREDICT project: CAIDA update

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

Data collection - passive

- OC192 backbone: March 2008 - June 2011
    - 14.4 TB compressed, 26.7 TB uncompressed
    - unanonymized: 7.7 TB compressed, 14.6 TB uncompressed
    - anonymized: 6.7 TB compressed, 12.1 TB uncompressed
    - Doing cleanup towards retaining only quarterly traces
    - Released 2011 Passive Dataset
- Problems:
    - Hardware failures at collection sites - solved (for now)
    - Hardware failures on our (new) data servers
    - Working with vendors to remediate
    - Need another sysadmin to keep up
- Plans:
    - 2011 annual dataset in progress (now includes Jan -Jun)
    - 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 2.85 TiB compressed, 5.5 TiB uncompressed
  - the previous months - backed up on tape (samqfs)
    - current: 2009/12/01 - 2011/07/05
    - 53 TB (compressed), 100 TB (uncompressed)
    - received new NSF award “CRI-Telescope: A Real-time Lens into Dark Address Space of the Internet”
- OC48 traces:
  - 964.5 GB (compressed), 1.7 TB
  - unanonymized: 815.7 GB (compressed), 1.5 TB (uncompressed)
  - anonymized: 148.8 GB (compressed), 285.2 GB (uncompressed) (in PREDICT)

Data collection - active

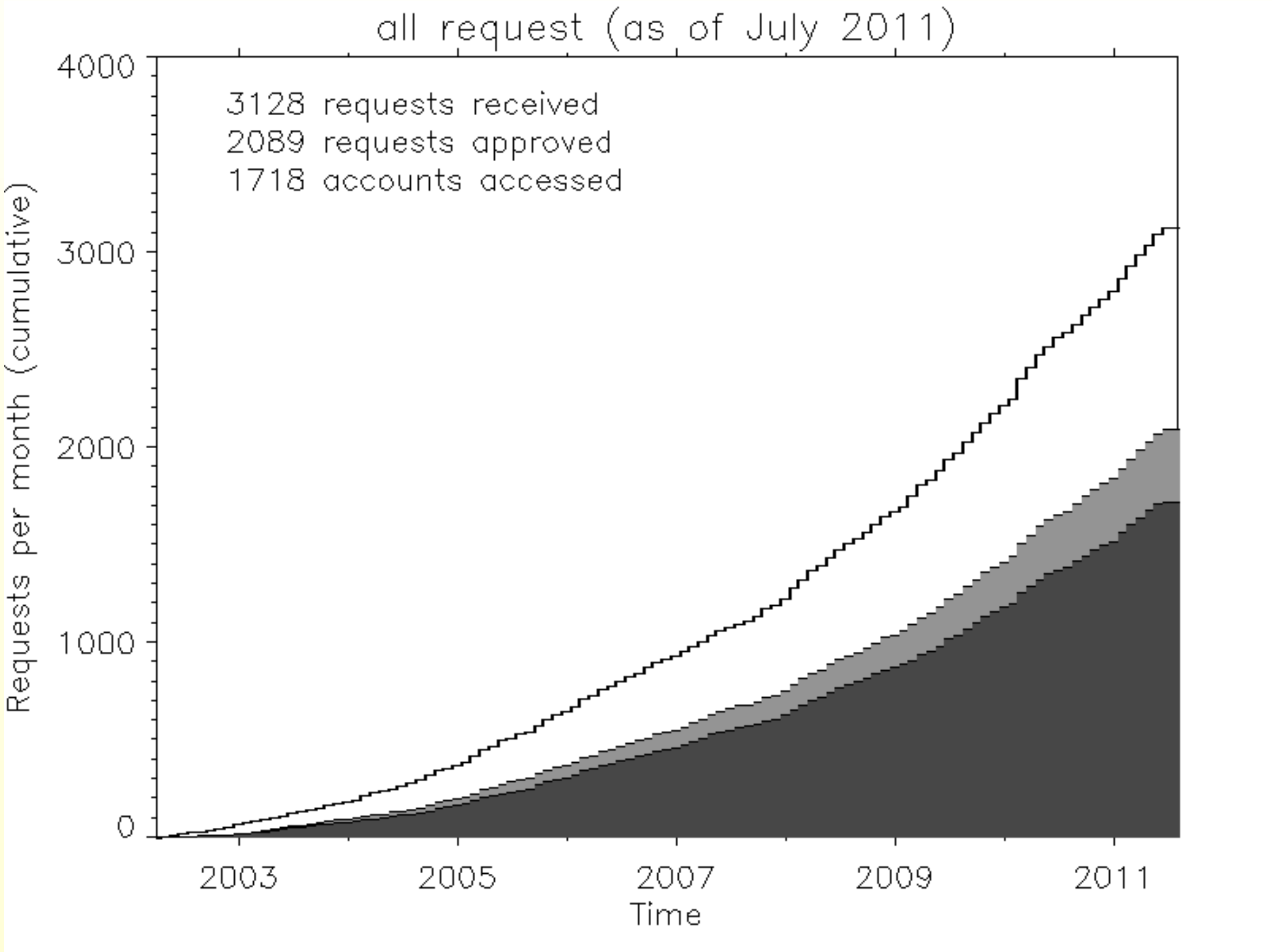
- old skitter data (in PREDICT):
    - 1.47 TB (compressed), 4.02 TB (uncompressed)
        - discontinued in February 2008
- current Ark data:
    - IPv4 topology: 1.5 TB (compressed), 4.8 TB (uncompressed)
    - IPv6 topology: 1.5 GB (compressed), 5.1 GB (uncompressed)
    - 54 monitors in 30 countries, 27 IPv6 capable
    - continues to expand
- data curation:
    - create derivative data sets
    - aggregate in 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/
- NSF award to curate/analyze/annotate IPv6 data (expected October 1, 2011)

Requests for the data, 2011/2010/2009

Dataset	Requests	Approved	Accessed	Served since
Backscatter	33/73/95	22/47/60	16/36/46	Feb 2003
Passive	154/185/233	122/150/179	99/126/157	Feb 2004
Topology	92/163/129	75/113/83	56/80/63	Jul 2004
Witty	9/16/27	7/13/17	6/11/14	Mar 2008
Telescope	13/34/37	11/23/21	10/19/17	Jul 2009
DNS-RTT	7/7/7	5/5/2	4/4/2	Aug 2006
DDoS	58/108/NA	38/74/NA	30/66/NA	Mar 2010
Total	**364/586/528**	**280/425/362**	**221/342/299**	

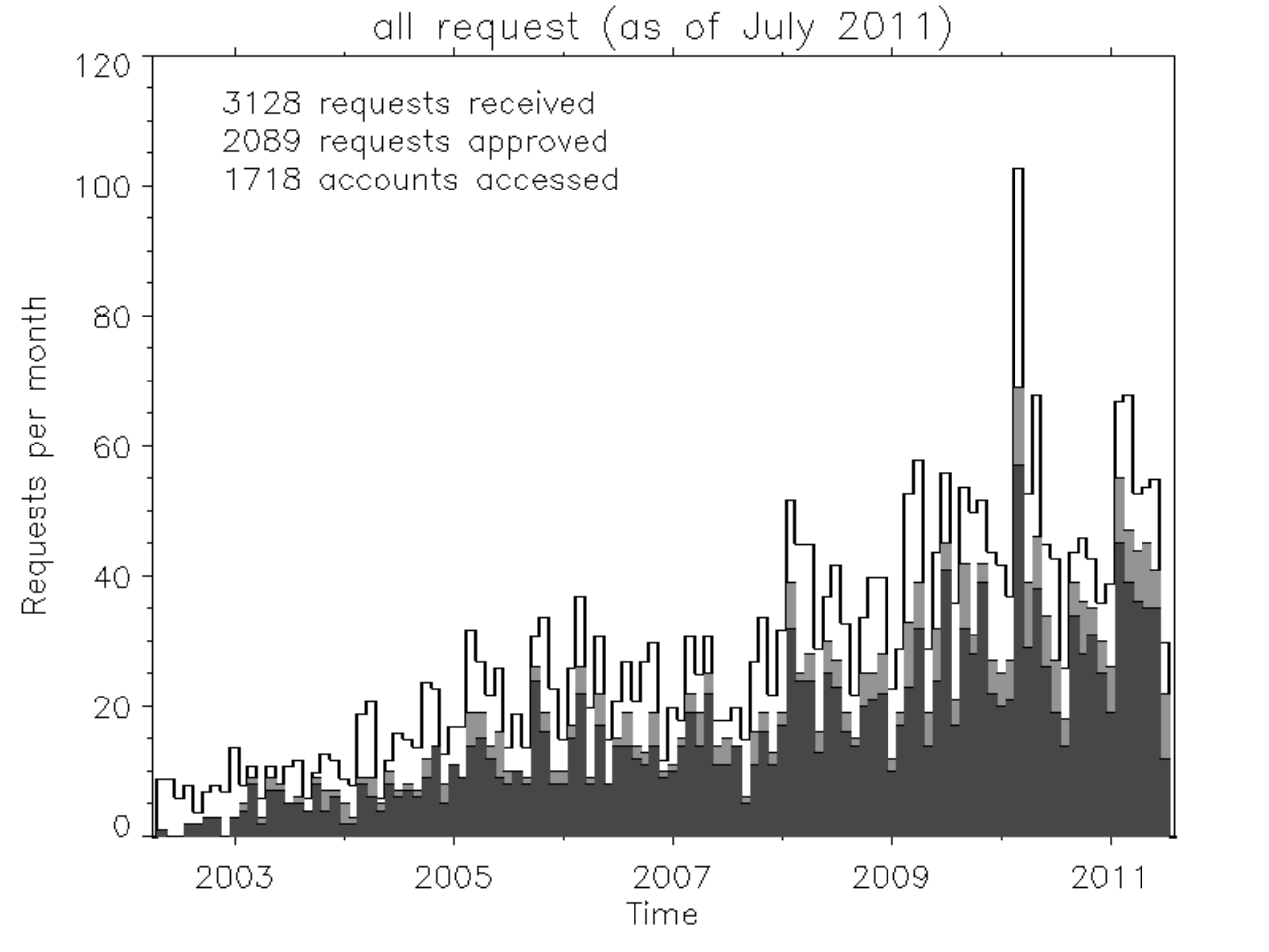
Data request stats

- all requests (cumulative)



Data request stats (cont)

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



Data Set Popularity

caida

- 1st best - OC192 and OC48 traces
  - requested 572 times, accessed 382 times (since 2009)
  - who used it: 236 .edu, 127 .cn, 38 .uk, 29 .com (since 2004) …
    - and 52 more domains
    - of 719 total accounts: 265 from U.S.
- 2nd best - topology data
  - requested 384 times, accessed 199 times (since 2009)
  - who used it: 240 .edu, 111 .cn, 38 .uk, 29 .com, 27 .kr, 23 .jp (since 2004) …
    - and 51 more domains
    - of 731 total accounts: 272 from U.S.

Data availability

- PREDICT (OC48 traces, topology from skitter, telescope)
- Derived data sets are publicly available (i.e., AS-links)
  - sample use: http://semilattice.net/projects/map-of-the-internet/
- Academics who sign AUP (OC192, topology from Ark, telescope)
- Commercial researchers
  - a small sample of 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
- topology
  - Ark statistics: http://www.caida.org/projects/ark/statistics/index.xml
  - path dispersion (AS and IP), path length distribution, RTT distribution, RTT vs. distance, median RTT per country, ...

Meta-data for packet traces

- OC192 data: 2008-2010, Jan-June 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 publications

- kc claffy, *Tracking IPv6 Evolution: Data we have and Data We Need,* ACM SIGCOMM CCR V. 41, p. 43-48, 2011.
- kc claffy, *The 3rd Workshop on Active Internet Measurements (AIMS-3) Report,* ACM SIGCOMM CCR V. 41, p. 37-42, 2011.
- Marina Fomenkov and kc claffy, *Internet Measurement Data Management Challenges,* presented at the Workshop on Research Data Lifecycle Management, July 2011.

Recent publications

- A. Dianotti, C. Squarcella, E. Aben, kc claffy, M. Chiesa, M. Russo, A. Pescape *Analysis of country-wide Internet outages caused by censorship,* accepted to IMC 2011.
  - national level outages in Egypt and Libya
  - data used:
    - public BGP
    - CAIDA telescope
    - Ark (could have done more)
  - analyzed methods used for traffic blocking, duration, testing
- B. Huffaker, M. Fomenkov, kc claffy *Geocompare - a comparison of public and commercial geolocation databses,* CAIDA tech report, 2011.
  - cross-analyzed multiple databases
  - used available ground truth data (PlanetLab, French networks, Tier 1 provider)
  - Ark RTT data

Recent blogs

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

  http://blog.caida.org/best_available_data/2011/07/25/my-third-fcc-tac-meeting-the-most-exciting-yet/

- kc claffy, *Exhausted IPv4 address architectures*

  http://blog.caida.org/best_available_data/2011/05/03/exhausted-ipv4-address-architectures/

- kc claffy, *CAIDA participation in IPv6 day*

  http://blog.caida.org/best_available_data/2011/06/05/caida-participation-in-ipv6-day/.

- Amogh Dhamdhere, *Model for Internet Evolution Predicts Consolidation in Tier-1 Transit Market*

  http://blog.caida.org/best_available_data/2011/07/15/

  model-for-internet-evolution-predicts-consolidation-in-tier-1-transit-market/.

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
- topology: updated ITDK 2010
- OC192 backbone: 2007-2011

Preparations for Phase II

- Submitted data sets descriptions
- Extensive reviews of documents
- bi-weekly phone calls
  - (how did they become weekly?)
  - Organization Referring Letter
  - Data Host MOA
  - Data Provider MOA
  - Researcher MOA - ?
- reviews of CAIDA AUPs

Updates of CAIDA policies

- Telescope data (RTTD)
  - different from previous packaged data
  - simplified and streamlined the AUP language
  - Immediate use by postdoc A. Dainotti and his student
  - analysis of macroscopic events (e.g earthquakes) on the Internet, collaborating with RIPE-NCC on publication.
- ARK hosting sites
  - Now using updated MoC for all new hosting sites
- Passive data collection MOC
  - Recently completed
    http://www.caida.org/data/collection/aup/internet_traffic_collection_moc.xml

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
  - Sent out to PI list for feedback
- Would like to discuss having a common AUP on PREDICT portal that meets all PIs' needs

General Principles of AUPs?

- Access conditions
    - Accreditation, validation, transparency
- Use restriction
    - Purpose, probing, other
- Disclosure obligations
    - Publication, 3rd party transfer, attribution
- Enforcement
    - Compliance, attestation
- Corrections / amendments
    - Measurement error notifications
- Disposition
    - Account closure, renewal
- Policy Vehicle: AUP, MOA, MOC…

Other activities

- 20-22 April 2011 PI k claffy attended the Disclosure and Control Workshop (DCW)
- what are we protecting?
  - PII (including IP addresses)
  - organization proprietary data
  - Privacy: Individual vs. Organization
- relevant for PREDICT Best Practice documentation efforts
- let Erin summarize status tomorrow

Other activities

- 18-20 July 2011 co-PI Marina Fomenkov attended the Research Data Lifecycle Management (RDLM) workshop

- the (disastrous) flood of digital data
- 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?

CAIDA Marketing Efforts

- Web site
  - Annual reports, Program Plan, Project web page, blogging
- Publications, Presentations, Workshops
- Proposals
  - NSF funded SDCI, will start in September?
    - reduce burden on contributors
    - convert from proprietary format to open source
    - expand relevance to cyber security
  - NSF funded CRI - telescope research, will start in September?
    - support “near real-time”, “bring code to the data” model
    - develop automated triggers and alerts
    - curate custom data sets upon request
  - BAA-11-02 proposal: plans to use PREDICT
- Synergy with NSF
  - Data Management Planning
  - Broader Impact activity

Storage Update

Ark IPv4

Total stored data: 1.52 TiB
Total stored no of files: 68131
Total free space: 4.4 TiB (shared with Ark IPv6)
Yesterday growth: 1.7 GiB

Ark IPv6

Total stored data: 1.64 GiB
Total stored no of files: 5458
Total free space: 4.4 TiB (shared with Ark IPv4)
Yesterday growth: 6.2 MiB

Passive high-speed equinix traces

Total stored data: 2.85 TiB
Total stored no of files: 3898
Total free space: 16 TiB
Yesterday growth: 5.7 GiB

Live telescope data (ogma)

Total stored data: 19.9 TiB
Total stored no of files: 5104
Total free space: 19 TiB
Yesterday growth: 95.8 GiB

Long-term Telescope storage on tape:

Total stored data: 50.1 TiB
Total stored no of files: 11826
Total free space: N/A
Yesterday growth: N/A

Overall Cummulative Stats

Total stored data: 76.01 TiB
Total stored no of files: 94417
Total free space: 43.8 TiB
Yesterday growth: 100 GiB